

# frontiers

## RESEARCH TOPICS

### INVARIANT RECOGNITION OF VISUAL OBJECTS

Hosted by  
Evgeniy Bart and Jay Hegdé



frontiers in  
**COMPUTATIONAL NEUROSCIENCE**



# frontiers

## FRONTIERS COPYRIGHT STATEMENT

© Copyright 2007-2013  
Frontiers Media SA.  
All rights reserved.

All content included on this site, such as text, graphics, logos, button icons, images, video/audio clips, downloads, data compilations and software, is the property of or is licensed to Frontiers Media SA ("Frontiers") or its licensees and/or subcontractors. The copyright in the text of individual articles is the property of their respective authors, subject to a license granted to Frontiers.

The compilation of articles constituting this e-book, as well as all content on this site is the exclusive property of Frontiers. Images and graphics not forming part of user-contributed materials may not be downloaded or copied without permission.

Articles and other user-contributed materials may be downloaded and reproduced subject to any copyright or other notices. No financial payment or reward may be given for any such reproduction except to the author(s) of the article concerned.

As author or other contributor you grant permission to others to reproduce your articles, including any graphics and third-party materials supplied by you, in accordance with the Conditions for Website Use and subject to any copyright notices which you include in connection with your articles and materials.

All copyright, and all rights therein, are protected by national and international copyright laws.

The above represents a summary only. For the full conditions see the Conditions for Authors and the Conditions for Website Use.

Cover image provided by Ibbl sarl, Lausanne CH

ISSN 1664-8714

ISBN 978-2-88919-076-8

DOI 10.3389/978-2-88919-076-8

## ABOUT FRONTIERS

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

## FRONTIERS JOURNAL SERIES

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing.

All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

## DEDICATION TO QUALITY

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.

Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view.

By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

## WHAT ARE FRONTIERS RESEARCH TOPICS?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area!

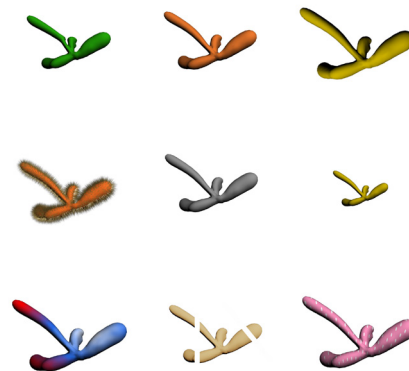
Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: [researchtopics@frontiersin.org](mailto:researchtopics@frontiersin.org)

# INVARIANT RECOGNITION OF VISUAL OBJECTS

Hosted By:

**Evgeniy Bart**, Palo Alto Research Center, USA

**Jay Hegdé**, Georgia Health Sciences University, USA



The visual system can recognize objects regardless of image variations, including those in illumination, surface color, surface texture, size, viewpoint and occlusions, as illustrated here. Is there more than one unique object in this picture, or are all objects variations of the same object?\*

Image courtesy of Jay Hegdé and Evgeniy Bart.

This Research Topic will focus on how the visual system recognizes objects regardless of variations in the viewpoint, illumination, retinal size, background, etc.

Contributors are encouraged to submit articles describing novel results, models, viewpoints, perspectives and/or methodological innovations relevant to this topic.

The issues we wish to cover include, but are not limited to, perceptual invariance under one or more of the following types of image variation:

- Object shape
- Task
- Viewpoint (from the translation and rotation of the object relative to the viewer)
- Illumination, shading, and shadows
- Degree of occlusion
- Retinal size
- Color
- Surface texture
- Visual context, including background clutter and crowding
- Object motion (including biological motion).

\*Answer: all objects in this picture are the same.

Examples of questions that are particularly interesting in this context include, but are not limited to:

- Empirical characterizations of properties of invariance: Does invariance always exist? How wide is its range and how strong is the tolerance to viewing conditions within this range?
- Invariance in naïve vs. experienced subjects: Is invariance built-in or learned? How can it be learned, under which conditions and how effectively? Is it learned incidentally, or are specific task and reward structures necessary for learning? How is generalizability and transfer of learning related to the generalizability/invariance of perception?
- Invariance during inference: Are there conditions (e.g. fast presentation time or otherwise resource-constrained recognition) when invariance breaks?
- What are some plausible computational or neural mechanisms by which invariance could be achieved?



# Table of Contents

## **05    *Invariant Recognition of Visual Objects: Some Emerging Computational Principles***

Evgeniy Bart and Jay Hegdé

## **Reviews**

## **07    *Invariant Visual Object and Face Recognition: Neural and Computational Bases, and a Model, VisNet***

Edmund T. Rolls

## **Perspectives/Opinion**

## **77    *Object Recognition can be Viewpoint Dependent or Invariant – It's Just a Matter of Time and Task***

Branka Milivojevic

## **80    *Renewing the Respect for Similarity***

Shimon Edelman and Reza Shahbazi

## **Cues to Invariance**

## **99    *Learned Non-Rigid Object Motion is a View-Invariant Cue to Recognizing Novel Objects***

Lewis L. Chuang, Quoc C. Vuong and Heinrich H. Bülthoff

## **107    *Low-Level Contrast Statistics are Diagnostic of Invariance of Natural Textures***

Iris I. A. Groen, Sennay Ghebreab, Victor A. F. Lamme and H. Steven Scholte

## **124    *Invariant Object Recognition Based on Extended Fragments***

Evgeniy Bart and Jay Hegdé

## **Models**

## **137    *Learning and Disrupting Invariance in Visual Recognition with a Temporal Association Rule***

Leyla Isik, Joel Z. Leibo and Tomaso Poggio

## **144    *Transformation-Invariant Visual Representations in Self-Organizing Spiking Neural Networks***

Benjamin D. Evans and Simon M. Stringer

## **163    *Learning View Invariant Recognition with Partially Occluded Objects***

James M. Tromans, Irina Higgins and Simon M. Stringer

## **Receptive Field Properties**

## **182    *Spatially Invariant Computations in Stereoscopic Vision***

Michel Vidal-Naquet and Sergei Gepshtein



# Invariant recognition of visual objects: some emerging computational principles

Evgeniy Bart<sup>1\*</sup> and Jay Hegdé<sup>2\*</sup>

<sup>1</sup> Palo Alto Research Center, Palo Alto, CA, USA

<sup>2</sup> Department of Ophthalmology, Vision Discovery Institute, and Brain and Behavior Discovery Institute, Georgia Health Sciences University, Augusta, GA, USA

\*Correspondence: bart@parc.com; jhegde@georgiahealth.edu

## Edited by:

Misha Tsodyks, Weizmann Institute of Science, Israel

## Reviewed by:

Misha Tsodyks, Weizmann Institute of Science, Israel

Invariant object recognition refers to recognizing an object regardless of irrelevant image variations, such as variations in viewpoint, lighting, retinal size, background, etc. The perceptual result of invariance, where the perception of a given object property is unaffected by irrelevant image variations, is often referred to as perceptual constancy (Kofka, 1935; Walsh and Kulikowski, 2010).

Mechanisms of invariant object recognition have, to a significant extent, remained unclear. This is both because experimental and computational studies have so far largely focused on understanding object recognition without these variations, and because the underlying computational problems are profoundly difficult.

The 10 articles in this Research Topic Issue focus on some of the key computational issues in invariant object recognition. There is no pretending that the articles cover all key areas of current research exhaustively or seamlessly. For instance, none of the articles in this issue address size invariance (Kilpatrick and Ittelson, 1953) or color constancy (Foster, 2011). Nonetheless, the articles collectively paint a useful pointillist picture of current research on computational principles of invariance.

## STRATEGIES OF REPRESENTING INVARIANCE

Several articles address various strategies of exploiting or representing the information in the visual image to achieve object invariance. Chuang et al. (2012) show, using psychophysical experiments, that non-rigid motion provides a cue to the invariance of dynamic objects. Groen et al. (2012) show that low-level image statistics can cue the extent to which natural textures are invariant across samples. Using electroencephalography (EEG), they also show that the differences in edge statistics predict the differences in the evoked neural responses to individual images. Using psychophysical experiments, Bart and Hegdé (2012)<sup>1</sup> show that human subjects can use small informative fragments of an image to recognize an object regardless of variations in illumination. A more radical idea is proposed by Edelman and Shadbazi (2012), who argue that representing objects by their similarity to a set of prototypes can explain many properties of the visual system, including invariance.

<sup>1</sup>Who are also the editors of this Research Topic Issue and the authors of this editorial.

## STRATEGIES OF LEARNING INVARIANCE

In a supervised setting, cues to object invariance may be provided externally (e.g., Bart and Hegdé, 2012). In unsupervised settings, finding cues to invariance is more challenging. One type of cues arises from the fact that even when an object changes in appearance, the change is generally smooth. Thus, over short, selected stretches of space and/or time, the changes in object appearance tend to be rather small, so that the visual system can, in principle, infer that the same object is changing its appearance. A theoretical approach for exploiting this contiguity is given by the continuous transformation (CT) learning (Stringer et al., 2006). A related cue arises from the fact that objects often stay in view for extended periods of time; two observations at nearby time points are therefore likely to correspond to the same object. An approach that exploits this temporal contiguity is given by the trace learning rule (Földiák, 1991).

Many articles in this issue describe models that exploit one or more of these rules to learn object invariance. The VisNet model can incorporate one or both of these strategies, depending on the particular implementation. The article by Rolls (2012) describes the various capabilities of VisNet. The article by Tromans et al. (2012) highlights the capability of VisNet to learn with clutter and occlusion. VisNet, like most neural network models, uses rate coding, in which the firing rate of a neuron determines the information coded by that neuron. The firing rate of a neuron is usually specified as a scalar, without the neuron having to actually fire spikes. The article by Evans and Stringer (2012) implements VisNet in which individual neurons actually fire spikes, and detail the merits of this implementation. The model by Isik et al. (2012) describes a different model, HMAX (also see Serre et al., 2007), that simulates many invariance properties in the primate visual system.

It is worth noting that, while it is generally thought that object invariance is represented by neurons in the higher levels of the visual pathway, such as the inferotemporal cortex, neurons in the lower levels, such as the primary visual cortex or V1, can also play key roles in implementing various aspects of invariance. The article by Vidal-Naquet and Gepshtein (2012) shows that populations of V1 complex cells, but not individual complex cells, can compute information about stereoscopic disparity in a spatially invariant fashion.

## SOME IMPORTANT CAVEATS

It is important to emphasize a few caveats about the implications of these articles for future research. First, at the perceptual level, object invariance neither is perfect nor needs to be (Bülthoff and Edelman, 1992; DiCarlo and Cox, 2007). Thus, the underlying neural mechanisms need not deliver perfect invariance. Second, not all types of invariance are equal. Some types of invariance may be more important or useful to the visual system than others, depending on the behavioral context (see Milivojevic, 2012). Third, the visual system does not necessarily have to rely on prolonged supervised learning to learn invariance. It is possible that the system is able to either learn or, alternatively, infer invariance on the fly, and without any feedback (see Rolls, 2012). Fourth, top-down factors, such as the behavioral context, play an important role in object invariance and lack thereof. This is not fully addressed by the articles in this issue, which mostly focus on bottom-up processing of invariance information. Finally, for practical reasons, current research tends to deal with invariance along the various individual stimulus parameters (e.g., viewpoint, illumination, etc.) separately from each other. But in actuality, the visual system may combine invariance across multiple visual parameters, and indeed multiple sensory modalities.

## REFERENCES

- Bart, E., and Hegdé, J. (2012). Invariant object recognition based on extended fragments. *Front. Comput. Neurosci.* 6:56. doi: 10.3389/fncom.2012.00056
- Bülthoff, H. H., and Edelman, S. (1992). Psychophysical support for a 2-D view interpolation theory of object recognition. *Proc. Natl. Acad. Sci. U.S.A.* 89, 60–64.
- Chuang, L. L., Vuong, Q. C., and Bülthoff, H. H. (2012). Learned non-rigid object motion is a view-invariant cue to recognizing novel objects. *Front. Comput. Neurosci.* 6:26. doi: 10.3389/fncom.2012.00026
- DiCarlo, J. J., and Cox, D. D. (2007). Untangling invariant object recognition. *Trends Cogn. Sci. (Regul. Ed.)* 11, 333–341.
- Edelman, S., and Shahbazi, R. (2012). Renewing the respect for similarity. *Front. Comput. Neurosci.* 6:45. doi: 10.3389/fncom.2012.00045
- Evans, B., and Stringer, S. (2012). Transform-invariant visual representations in self-organizing spiking neural networks. *Front. Comput. Neurosci.* 6:46. doi: 10.3389/fncom.2012.00046
- Földiák, P. (1991). Learning invariance from transformation sequences. *Neural Comput.* 3, 194–200.
- Foster, D. H. (2011). Color constancy. *Vision Res.* 51, 674–700.
- Groen, I. I. A., Ghebreab, S., Lamme, V. A. F., and Scholte, H. S. (2012). Low-level edge statistics predict invariance of natural textures. *Front. Comput. Neurosci.* 6:34. doi: 10.3389/fncom.2012.00034
- Isik, L., Leibo, J. Z., and Poggio, P. (2012). Learning and disrupting invariance in visual recognition with a temporal association rule. *Front. Comput. Neurosci.* 6:37. doi: 10.3389/fncom.2012.00037
- Kilpatrick, F. P., and Ittelson, W. H. (1953). The size-distance invariance hypothesis. *Psychol. Rev.* 60, 223–231.
- Kofka, K. (1935). *Principles of Gestalt Psychology*. New York, NY: Harcourt, Brace and Company.
- Milivojevic, B. (2012). Object recognition can be viewpoint dependent or invariant – it's just a matter of time and task. *Front. Comput. Neurosci.* 6:27. doi: 10.3389/fncom.2012.00027
- Rolls, E. T. (2012). Invariant visual object and face recognition: neural and computational bases, and a model, VisNet. *Front. Comput. Neurosci.* 6:35. doi: 10.3389/fncom.2012.00035
- Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., and Poggio, T. (2007). Robust object recognition with cortex-like mechanisms. *IEEE Trans. Pattern Anal. Mach. Intell.* 29, 3, 411–426.
- Stringer, S. M., Perry, G., Rolls, E. T., and Proske, J. H. (2006). Learning invariant object recognition in the visual system with continuous transformations. *Biol. Cybern.* 94, 128–142.
- Tromans, J. M., Higgins, I., and Stringer, S. M. (2012). Learning view invariant recognition with partially occluded objects. *Front. Comput. Neurosci.* 6:48. doi: 10.3389/fncom.2012.00048
- Vidal-Naquet, M., and Gepshtein, S. (2012). Spatially invariant computations in stereoscopic vision. *Front. Comput. Neurosci.* 6:47. doi: 10.3389/fncom.2012.00047
- Walsh, V., and Kulikowski, J. (eds). (2010). *Perceptual Constancy: Why Things Look as They Do*. New York, NY: Cambridge University Press.

Received: 26 July 2012; accepted: 26 July 2012; published online: 24 August 2012.

Citation: Bart E and Hegdé J (2012) Invariant recognition of visual objects: some emerging computational principles. *Front. Comput. Neurosci.* 6:60. doi: 10.3389/fncom.2012.00060

Copyright © 2012 Bart and Hegdé. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in other forums, provided the original authors and source are credited and subject to any copyright notices concerning any third-party graphics etc.



# Invariant visual object and face recognition: neural and computational bases, and a model, VisNet

Edmund T. Rolls<sup>1,2</sup> \*

<sup>1</sup> Oxford Centre for Computational Neuroscience, Oxford, UK

<sup>2</sup> Department of Computer Science, University of Warwick, Coventry, UK

## Edited by:

Evgeniy Bart, Palo Alto Research Center, USA

## Reviewed by:

Alexander G. Dimitrov, Washington State University Vancouver, USA  
Jay Hegd , Georgia Health Sciences University, USA

## \*Correspondence:

Edmund T. Rolls, Department of Computer Science, University of Warwick, Coventry CV4 7AL, UK.  
e-mail: edmund.rolls@oxcns.org

Neurophysiological evidence for invariant representations of objects and faces in the primate inferior temporal visual cortex is described. Then a computational approach to how invariant representations are formed in the brain is described that builds on the neurophysiology. A feature hierarchy model in which invariant representations can be built by self-organizing learning based on the temporal and spatial statistics of the visual input produced by objects as they transform in the world is described. VisNet can use temporal continuity in an associative synaptic learning rule with a short-term memory trace, and/or it can use spatial continuity in continuous spatial transformation learning which does not require a temporal trace. The model of visual processing in the ventral cortical stream can build representations of objects that are invariant with respect to translation, view, size, and also lighting. The model has been extended to provide an account of invariant representations in the dorsal visual system of the global motion produced by objects such as looming, rotation, and object-based movement. The model has been extended to incorporate top-down feedback connections to model the control of attention by biased competition in, for example, spatial and object search tasks. The approach has also been extended to account for how the visual system can select single objects in complex visual scenes, and how multiple objects can be represented in a scene. The approach has also been extended to provide, with an additional layer, for the development of representations of spatial scenes of the type found in the hippocampus.

**Keywords:** VisNet, invariance, face recognition, object recognition, inferior temporal visual cortex, trace learning rule, hippocampus, spatial scene representation

## 1. INTRODUCTION

One of the major problems that is solved by the visual system in the cerebral cortex is the building of a representation of visual information which allows object and face recognition to occur relatively independently of size, contrast, spatial-frequency, position on the retina, angle of view, lighting, etc. These invariant representations of objects, provided by the inferior temporal visual cortex (Rolls, 2008b), are extremely important for the operation of many other systems in the brain, for if there is an invariant representation, it is possible to learn on a single trial about reward/punishment associations of the object, the place where that object is located, and whether the object has been seen recently, and then to correctly generalize to other views, etc. of the same object (Rolls, 2008b). The way in which these invariant representations of objects are formed is a major issue in understanding brain function, for with this type of learning, we must not only store and retrieve information, but we must solve in addition the major computational problem of how all the different images on the retina (position, size, view, etc.) of an object can be mapped to the same representation of that object in the brain. It is this process with which we are concerned in this paper.

In Section 2 of this paper, I summarize some of the evidence on the nature of the invariant representations of objects

and faces found in the inferior temporal visual cortex as shown by neuronal recordings. A fuller account is provided in *Memory, Attention, and Decision-Making*, Chapter 4 (Rolls, 2008b). Then I build on that foundation a closely linked computational theory of how these invariant representations of objects and faces may be formed by self-organizing learning in the brain, which has been investigated by simulations in a model network, VisNet (Rolls, 1992, 2008b; Wallis and Rolls, 1997; Rolls and Milward, 2000).

This paper reviews this combined neurophysiological and computational neuroscience approach developed by the author which leads to a theory of invariant visual object recognition, and relates this approach to other research.

## 2. INVARIANT REPRESENTATIONS OF FACES AND OBJECTS IN THE INFERIOR TEMPORAL VISUAL CORTX

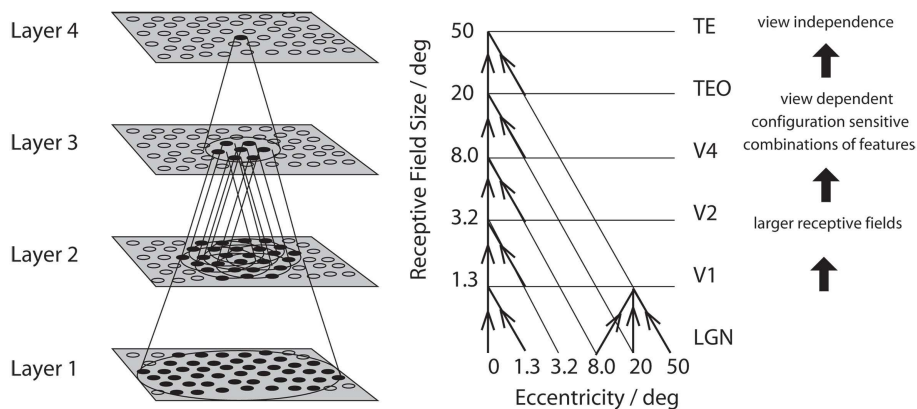
### 2.1. PROCESSING TO THE INFERIOR TEMPORAL CORTX IN THE PRIMATE VISUAL SYSTEM

A schematic diagram to indicate some aspects of the processing involved in object identification from the primary visual cortex, V1, through V2 and V4 to the posterior inferior temporal cortex (TEO) and the anterior inferior temporal cortex (TE) is shown in **Figure 1** (Rolls and Deco, 2002; Rolls, 2008b; Blumberg and Kreiman, 2010; Orban, 2011). The approximate location of

these visual cortical areas on the brain of a macaque monkey is shown in **Figure 2**, which also shows that TE has a number of different subdivisions. The different TE areas all contain visually responsive neurons, as do many of the areas within the cortex in the superior temporal sulcus (Baylis et al., 1987). For the purposes of this summary, these areas will be grouped together as the anterior inferior temporal cortex (IT), except where otherwise stated.

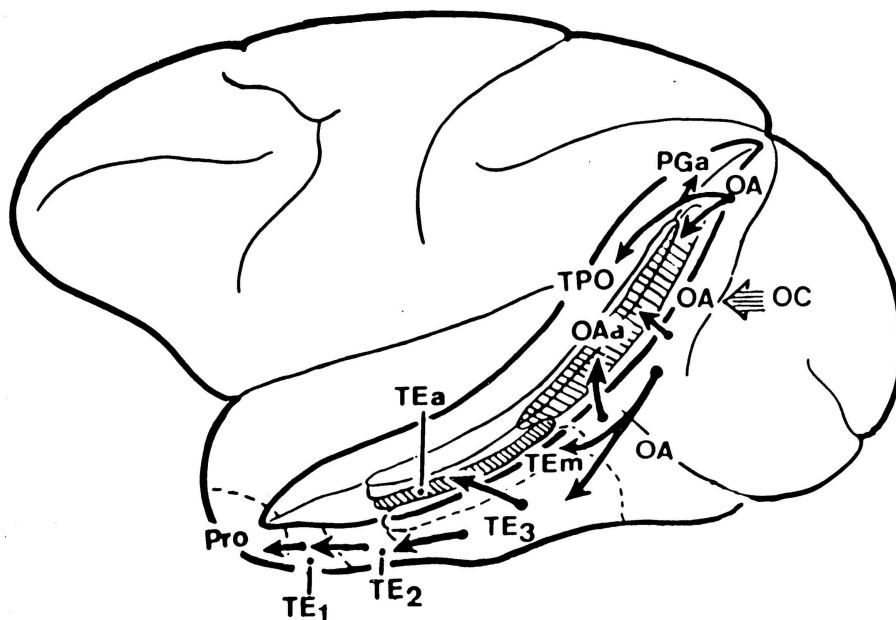
The object and face-selective neurons described in this paper are found mainly between 7 and 3 mm posterior to the sphenoid reference, which in a 3–4 kg macaque corresponds to

approximately 11–15 mm anterior to the interaural plane (Baylis et al., 1987; Rolls, 2007a,b, 2008b). For comparison, the “middle face patch” of Tsao et al. (2006) was at A6, which is probably part of the posterior inferior temporal cortex (Tsao and Livingstone, 2008). In the anterior inferior temporal cortex areas we have investigated, there are separate regions specialized for face identity in areas TEa and TEm on the ventral lip of the superior temporal sulcus and the adjacent gyrus, for face expression and movement in the cortex deep in the superior temporal sulcus (Baylis et al., 1987; Hasselmo et al., 1989a; Rolls, 2007b), and separate neuronal clusters for objects (Booth and



**FIGURE 1 | Convergence in the visual system.** Right – as it occurs in the brain. V1, visual cortex area V1; TEO, posterior inferior temporal cortex; TE, inferior temporal cortex (IT). Left – as implemented in

VisNet. Convergence through the network is designed to provide fourth layer neurons with information from across the entire input retina.



**FIGURE 2 | Lateral view of the macaque brain (left hemisphere) showing the different architectonic areas (e.g., TEm, TEa) in and bordering the anterior part of the superior temporal sulcus (STS) of the macaque (see text). The STS has been drawn opened to reveal the cortical areas inside it, and is circumscribed by a thick line.**

Rolls, 1998; Kriegeskorte et al., 2008; Rolls, 2008b). A possible way in which VisNet could produce separate representations of face identity and expression has been investigated (Tromans et al., 2011). Similarly, in humans there are a number of separate visual representations of faces and other body parts (Spiridon et al., 2006; Weiner and Grill-Spector, 2011), with the clustering together of neurons with similar responses influenced by the self-organizing map processes that are a result of cortical design (Rolls, 2008b).

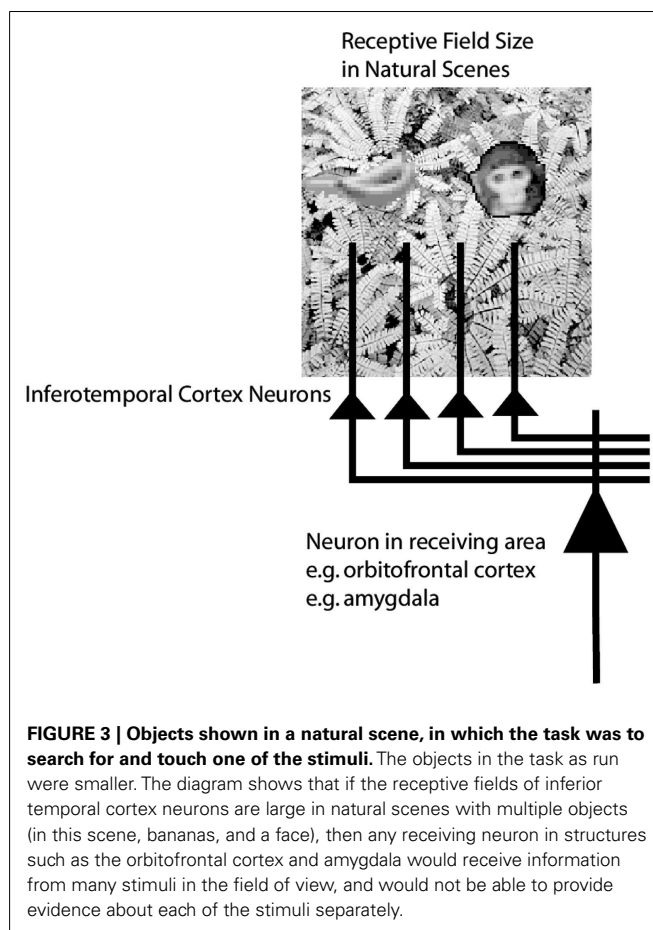
## 2.2. TRANSLATION INVARIANCE AND RECEPTIVE FIELD SIZE

There is convergence from each small part of a region to the succeeding region (or layer in the hierarchy) in such a way that the receptive field sizes of neurons (for example,  $1^\circ$  near the fovea in V1) become larger by a factor of approximately 2.5 with each succeeding stage. (The typical parafoveal receptive field sizes found would not be inconsistent with the calculated approximations of, for example,  $8^\circ$  in V4,  $20^\circ$  in TEO, and  $50^\circ$  in inferior temporal cortex Boussaoud et al., 1991; see **Figure 1**). Such zones of convergence would overlap continuously with each other (see **Figure 1**). This connectivity provides part of the basis for the fact that many neurons in the temporal cortical visual areas respond to a stimulus relatively independently of where it is in their receptive field, and moreover maintain their stimulus selectivity when the stimulus appears in different parts of the visual field (Gross et al., 1985; Tovee et al., 1994; Rolls et al., 2003). This is called translation or shift invariance. In addition to having topologically appropriate connections, it is necessary for the connections to have the appropriate synaptic weights to perform the mapping of each set of features, or object, to the same set of neurons in IT. How this could be achieved is addressed in the computational neuroscience models described later in this paper.

## 2.3. REDUCED TRANSLATION INVARIANCE IN NATURAL SCENES, AND THE SELECTION OF A REWARDED OBJECT

Until recently, research on translation invariance considered the case in which there is only one object in the visual field. What happens in a cluttered, natural, environment? Do all objects that can activate an inferior temporal neuron do so whenever they are anywhere within the large receptive fields of inferior temporal neurons (Sato, 1989; Rolls and Tovee, 1995a)? If so, the output of the visual system might be confusing for structures that receive inputs from the temporal cortical visual areas. If one of the objects in the visual field was associated with reward, and another with punishment, would the output of the inferior temporal visual cortex to emotion-related brain systems be an amalgam of both stimuli? If so, how would we be able to choose between the stimuli, and have an emotional response to one but not perhaps the other, and select one for action and not the other (see **Figure 3**).

To investigate how information is passed from the inferior temporal cortex (IT) to other brain regions to enable stimuli to be selected from natural scenes for action, Rolls et al. (2003) analyzed the responses of single and simultaneously recorded IT neurons to stimuli presented in complex natural backgrounds. In one situation, a visual fixation task was performed in which the monkey fixated at different distances from the effective stimulus.



In another situation the monkey had to search for two objects on a screen, and a touch of one object was rewarded with juice, and of another object was punished with saline (see **Figure 3** for a schematic overview and **Figure 30** for the actual display). In both situations neuronal responses to the effective stimuli for the neurons were compared when the objects were presented in the natural scene or on a plain background. It was found that the overall response of the neuron to objects was sometimes somewhat reduced when they were presented in natural scenes, though the selectivity of the neurons remained. However, the main finding was that the magnitudes of the responses of the neurons typically became much less in the real scene the further the monkey fixated in the scene away from the object (see **Figures 4** and **31** and Section 5.8.1).

It is proposed that this reduced translation invariance in natural scenes helps an unambiguous representation of an object which may be the target for action to be passed to the brain regions that receive from the primate inferior temporal visual cortex. It helps with the binding problem, by reducing in natural scenes the effective receptive field of inferior temporal cortex neurons to approximately the size of an object in the scene. The computational utility and basis for this is considered in Section 5.8 and by Rolls and Deco (2002), Trappenberg et al. (2002), Deco and Rolls (2004), Aggelopoulos and Rolls (2005), and Rolls and Deco (2006), and includes an advantage for what is at the fovea because



of the large cortical magnification of the fovea, and shunting interactions between representations weighted by how far they are from the fovea.

These findings suggest that the principle of providing strong weight to whatever is close to the fovea is an important principle governing the operation of the inferior temporal visual cortex, and in general of the output of the ventral visual system in natural environments. This principle of operation is very important in interfacing the visual system to action systems, because the effective stimulus in making inferior temporal cortex neurons fire is in natural scenes usually on or close to the fovea. This means that the spatial coordinates of where the object is in the scene do not have to be represented in the inferior temporal visual cortex, nor passed from it to the action selection system, as the latter can assume that the object making IT neurons fire is close to the fovea in natural scenes. Thus the position in visual space being fixated provides part of the interface between sensory representations of objects and their coordinates as targets for actions in the world. The small receptive fields of IT neurons in natural scenes make this possible. After this, local, egocentric, processing implemented in the dorsal visual processing stream using, e.g., stereodisparity may be used to guide action toward objects being fixated (Rolls and Deco, 2002).

The reduced receptive field size in complex natural scenes also enables emotions to be selective to just what is being fixated, because this is the information that is transmitted by the firing of IT neurons to structures such as the orbitofrontal cortex and amygdala.

There is an important comparison to be made here with some approaches in engineering in which attempts are made to analyze a whole visual scene at once. This is a massive computational problem, not yet solved in engineering. It is very instructive to see that this is not the approach taken by the (primate and human) brain, which instead analyses in complex natural scenes what is close to

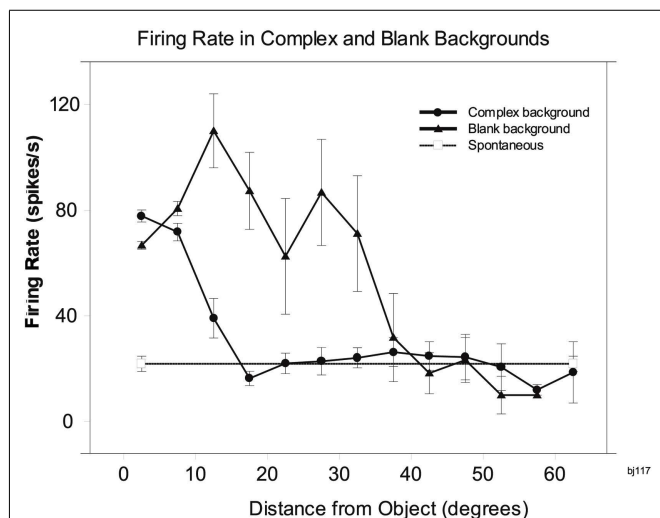
the fovea, just massively reducing the computational including feature binding problems. The brain then deals with a complex scene by fixating different parts serially, using processes such as bottom-up saliency to guide where fixations should occur (Itti and Koch, 2000; Zhao and Koch, 2011).

Interestingly, although the size of the receptive fields of inferior temporal cortex neurons becomes reduced in natural scenes so that neurons in IT respond primarily to the object being fixated, there is nevertheless frequently some asymmetry in the receptive fields (see Section 5.9 and Figure 35). This provides a partial solution to how multiple objects and their positions in a scene can be captured with a single glance (Aggelopoulos and Rolls, 2005).

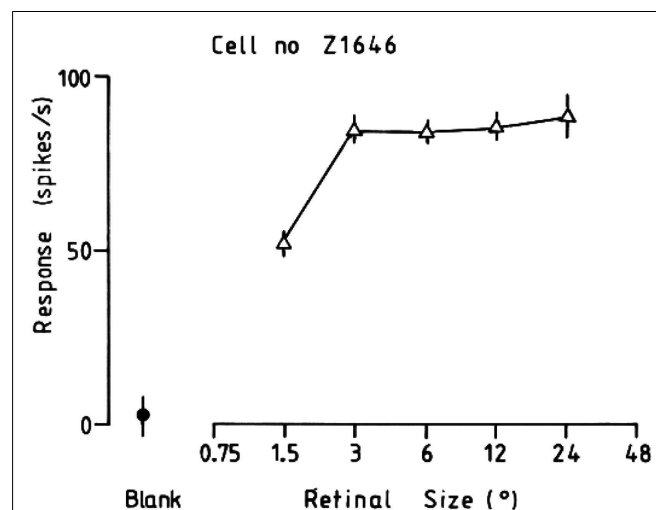
#### 2.4. SIZE AND SPATIAL-FREQUENCY INVARIANCE

Some neurons in the inferior temporal visual cortex and cortex in the anterior part of the superior temporal sulcus (IT/STS) respond relatively independently of the size of an effective face stimulus, with a mean size-invariance (to a half maximal response) of 12 times (3.5 octaves; Rolls and Baylis, 1986). An example of the responses of an inferior temporal cortex face-selective neuron to faces of different sizes is shown in Figure 5. This is not a property of a simple single-layer network (see Figure 7), nor of neurons in V1, which respond best to small stimuli, with a typical size-invariance of 1.5 octaves. Also, the neurons typically responded to a face when the information in it had been reduced from 3D to a 2D representation in gray on a monitor, with a response that was on average 0.5 of that to a real face.

Another transform over which recognition is relatively invariant is spatial-frequency. For example, a face can be identified when it is blurred (when it contains only low-spatial frequencies), and when it is high-pass spatial-frequency filtered (when it looks like a line drawing). If the face images to which these neurons respond are low-pass filtered in the spatial-frequency domain (so that they are blurred), then many of the neurons still respond when the images contain frequencies only up to 8 cycles per face. Similarly,



**FIGURE 4 |** Firing of a temporal cortex cell to an effective stimulus presented either in a blank background or in a natural scene, as a function of the angle in degrees at which the monkey was fixating away from the effective stimulus. The task was to search for and touch the stimulus. (After Rolls et al., 2003.)



**FIGURE 5 |** Typical response of an inferior temporal cortex face-selective neuron to faces of different sizes. The size subtended at the retina in degrees is shown. (From Rolls and Baylis, 1986.)

the neurons still respond to high-pass filtered images (with only high-spatial-frequency edge information) when frequencies down to only 8 cycles per face are included (Rolls et al., 1985). Face recognition shows similar invariance with respect to spatial-frequency (see Rolls et al., 1985). Further analysis of these neurons with narrow (octave) bandpass spatial-frequency filtered face stimuli shows that the responses of these neurons to an unfiltered face can not be predicted from a linear combination of their responses to the narrow bandstimuli (Rolls et al., 1987). This lack of linearity of these neurons, and their responsiveness to a wide range of spatial frequencies (see also their broad critical bandmasking Rolls, 2008a), indicate that in at least this part of the primate visual system recognition does not occur using Fourier analysis of the spatial-frequency components of images.

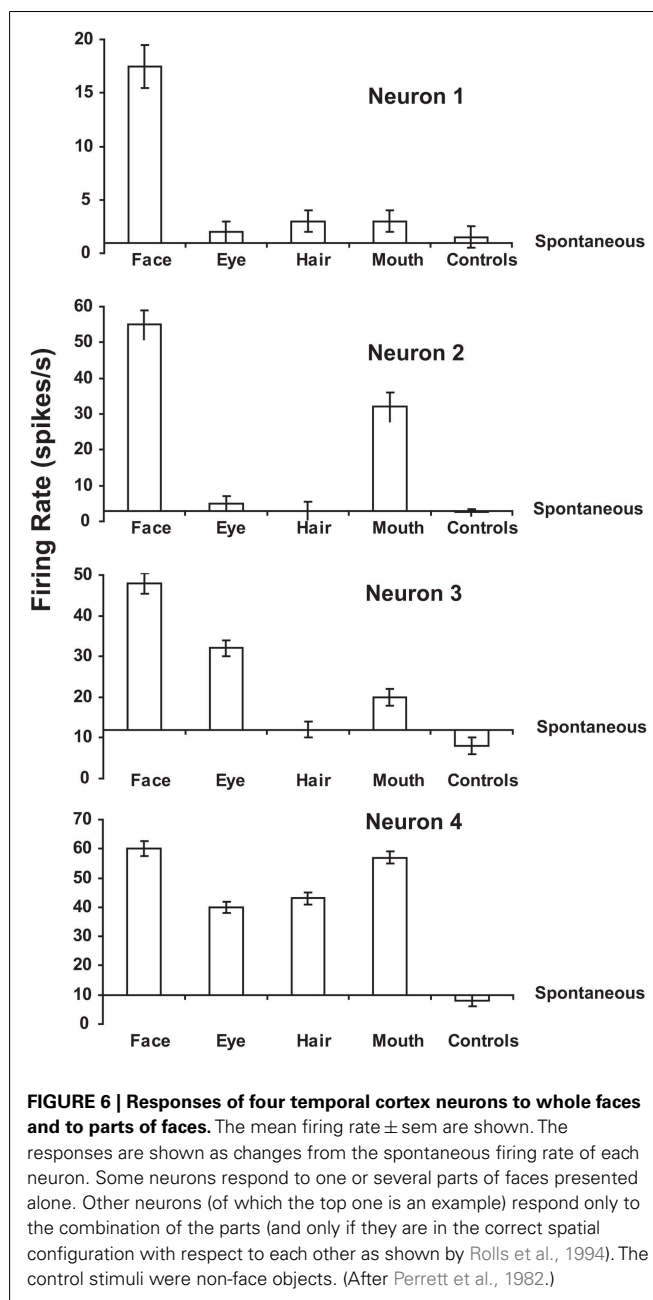
The utility of this representation for memory systems in the brain is that the output of the visual system will represent an object invariantly with respect to position on the retina, size, etc. and this simplifies the functionality required of the (multiple) memory systems, which need then simply associate the object representation with reward (orbitofrontal cortex and amygdala), associate it with position in the environment (hippocampus), recognize it as familiar (perirhinal cortex), associate it with a motor response in a habit memory (basal ganglia), etc. (Rolls, 2008b). The associations can be relatively simple, involving, for example, Hebbian associativity (Rolls, 2008b).

Some neurons in the temporal cortical visual areas actually represent the absolute size of objects such as faces independently of viewing distance (Rolls and Baylis, 1986). This could be called neurophysiological size constancy. The utility of this representation by a small population of neurons is that the absolute size of an object is a useful feature to use as an input to neurons that perform object recognition. Faces only come in certain sizes.

## 2.5. COMBINATIONS OF FEATURES IN THE CORRECT SPATIAL CONFIGURATION

Many neurons in this ventral processing stream respond to combinations of features (including objects), but not to single features presented alone, and the features must have the correct spatial arrangement. This has been shown, for example, with faces, for which it has been shown by masking out or presenting parts of the face (for example, eyes, mouth, or hair) in isolation, or by jumbling the features in faces, that some cells in the cortex in IT/STS respond only if two or more features are present, and are in the correct spatial arrangement (Perrett et al., 1982; Rolls et al., 1994; Freiwald et al., 2009; Rolls, 2011b). **Figure 6** shows examples of four neurons, the top one of which responds only if all the features are present, and the others of which respond not only to the full-face, but also to one or more features. Corresponding evidence has been found for non-face cells. For example Tanaka et al. (1990) showed that some posterior inferior temporal cortex neurons might only respond to the combination of an edge and a small circle if they were in the correct spatial relationship to each other. Consistent evidence for face part configuration sensitivity has been found in human fMRI studies (Liu et al., 2010).

These findings are important for the computational theory, for they show that neurons selective to feature combinations are part



of the process by which the cortical hierarchy operates, and this is incorporated into VisNet (Elliffe et al., 2002).

Evidence consistent with the suggestion that neurons are responding to combinations of a few variables represented at the preceding stage of cortical processing is that some neurons in V2 and V4 respond to end-stopped lines, to tongues flanked by inhibitory subregions, to combinations of lines, to combinations of colors, or to surfaces (Hegde and Van Essen, 2000, 2003, 2007; Ito and Komatsu, 2004; Brincat and Connor, 2006; Anzai et al., 2007; Orban, 2011). In the inferior temporal visual cortex, some neurons respond to spatial configurations of surface fragments to help specify the three-dimensional structure of objects (Yamane et al., 2008).



## 2.6. A VIEW-INVARIANT REPRESENTATION

For recognizing and learning about objects (including faces), it is important that an output of the visual system should be not only translation and size invariant, but also relatively view-invariant. In an investigation of whether there are such neurons, we found that some temporal cortical neurons reliably responded differently to the faces of two different individuals independently of viewing angle (Hasselmo et al., 1989b), although in most cases (16/18 neurons) the response was not perfectly view-independent. Mixed together in the same cortical regions there are neurons with view-dependent responses (for example, Hasselmo et al., 1989b; Rolls and Tovee, 1995b). Such neurons might respond, for example, to a view of a profile of a monkey but not to a full-face view of the same monkey (Perrett et al., 1985; Hasselmo et al., 1989b).

These findings of view-dependent, partially view-independent, and view-independent representations in the same cortical regions are consistent with the hypothesis discussed below that view-independent representations are being built in these regions by associating together the outputs of neurons that have different view-dependent responses to the same individual. These findings also provide evidence that one output of the visual system includes representations of what is being seen, in a view-independent way that would be useful for object recognition and for learning associations about objects; and that another output is a view-based representation that would be useful in social interactions to determine whether another individual is looking at one, and for selecting details of motor responses, for which the orientation of the object with respect to the viewer is required (Rolls, 2008b).

Further evidence that some neurons in the temporal cortical visual areas have object-based rather than view-based responses comes from a study of a population of neurons that responds to moving faces (Hasselmo et al., 1989b). For example, four neurons responded vigorously to a head undergoing ventral flexion, irrespective of whether the view of the head was full-face, of either profile, or even of the back of the head. These different views could only be specified as equivalent in object-based coordinates. Further, the movement specificity was maintained across inversion, with neurons responding, for example, to ventral flexion of the head irrespective of whether the head was upright or inverted. In this procedure, retinally encoded or viewer-centered movement vectors are reversed, but the object-based description remains the same.

Also consistent with object-based encoding is the finding of a small number of neurons that respond to images of faces of a given absolute size, irrespective of the retinal image size, or distance (Rolls and Baylis, 1986).

Neurons with view-invariant responses to objects seen naturally by macaques have also been described (Booth and Rolls, 1998). The stimuli were presented for 0.5 s on a color video monitor while the monkey performed a visual fixation task. The stimuli were images of 10 real plastic objects that had been in the monkey's cage for several weeks, to enable him to build view-invariant representations of the objects. Control stimuli were views of objects that had never been seen as real objects. The neurons analyzed were in the TE cortex in and close to the ventral lip of the anterior part of the superior temporal sulcus. Many neurons were found that responded to some views of some objects. However, for a smaller

number of neurons, the responses occurred only to a subset of the objects (using ensemble encoding), irrespective of the viewing angle. Moreover, the firing of a neuron on any one trial, taken at random and irrespective of the particular view of any one object, provided information about which object had been seen, and this information increased approximately linearly with the number of neurons in the sample. This is strong quantitative evidence that some neurons in the inferior temporal cortex provide an invariant representation of objects. Moreover, the results of Booth and Rolls (1998) show that the information is available in the firing rates, and has all the desirable properties of distributed representations, including exponentially high-coding capacity, and rapid speed of read-out of the information (Rolls, 2008b; Rolls and Treves, 2011).

Further evidence consistent with these findings is that some studies have shown that the responses of some visual neurons in the inferior temporal cortex do not depend on the presence or absence of critical features for maximal activation (Perrett et al., 1982; Tanaka, 1993, 1996). For example, neuron 4 in **Figure 6** responded to several of the features in a face when these features were presented alone (Perrett et al., 1982). In another example, Mikami et al. (1994) showed that some TE cells respond to partial views of the same laboratory instrument(s), even when these partial views contain different features. Such functionality is important for object recognition when part of an object is occluded, by, for example, another object. In a different approach, Logothetis et al. (1994) have reported that in monkeys extensively trained (over thousands of trials) to treat different views of computer generated wire-frame "objects" as the same, a small population of neurons in the inferior temporal cortex did respond to different views of the same wire-frame object (see also Logothetis and Sheinberg, 1996). However, extensive training is not necessary for invariant representations to be formed, and indeed no explicit training in invariant object recognition was given in the experiment by Booth and Rolls (1998), as Rolls' hypothesis (Rolls, 1992) is that view-invariant representations can be learned by associating together the different views of objects as they are moved and inspected naturally in a period that may be in the order of a few seconds. Evidence for this is described in Section 2.7.

## 2.7. LEARNING OF NEW REPRESENTATIONS IN THE TEMPORAL CORTICAL VISUAL AREAS

To investigate the idea that visual experience might guide the formation of the responsiveness of neurons so that they provide an economical and ensemble-encoded representation of items actually present in the environment (and indeed any rapid learning found might help in the formation of invariant representations), the responses of inferior temporal cortex face-selective neurons have been analyzed while a set of new faces were shown. Some of the neurons studied in this way altered the relative degree to which they responded to the different members of the set of novel faces over the first few (1–2) presentations of the set (Rolls et al., 1989). If in a different experiment a single novel face was introduced when the responses of a neuron to a set of familiar faces were being recorded, the responses to the set of familiar faces were not disrupted, while the responses to the novel face became stable within a few presentations. Alteration of the tuning of individual neurons in this way may result in a good discrimination over the

population as a whole of the faces known to the monkey. This evidence is consistent with the categorization being performed by self-organizing competitive neuronal networks, as described elsewhere (Rolls and Treves, 1998; Rolls, 2008b). Further evidence has been found to support the hypothesis (Rolls, 1992, 2008b) that unsupervised natural experience rapidly alters invariant object representation in the visual cortex (Li and DiCarlo, 2008; Li et al., 2011; cf. Folstein et al., 2010).

Further evidence that these neurons can learn new representations very rapidly comes from an experiment in which binarized black and white (two-tone) images of faces that blended with the background were used. These did not activate face-selective neurons. Full gray-scale images of the same photographs were then shown for ten 0.5 s presentations. In a number of cases, if the neuron happened to be responsive to that face, when the binarized version of the same face was shown next, the neurons responded to it (Tovee et al., 1996). This is a direct parallel to the same phenomenon that is observed psychophysically, and provides dramatic evidence that these neurons are influenced by only a very few seconds (in this case 5 s) of experience with a visual stimulus. We have shown a neural correlate of this effect using similar stimuli and a similar paradigm in a PET (positron emission tomography) neuroimaging study in humans, with a region showing an effect of the learning found for faces in the right temporal lobe, and for objects in the left temporal lobe (Dolan et al., 1997).

Once invariant representations of objects have been learned in the inferior temporal visual cortex based on the statistics of the spatio-temporal continuity of objects in the visual world (Rolls, 1992, 2008b; Yi et al., 2008), later processes may be required to categorize objects based on other properties than their properties as objects. One such property is that certain objects may need to be treated as similar for the correct performance of a task, and others as different, and that demand can influence the representations of objects in a number of brain areas (Fenske et al., 2006; Freedman and Miller, 2008; Kourtzi and Connor, 2011). That process may in turn influence representations in the inferior temporal visual cortex, for example, by top-down bias (Rolls and Deco, 2002; Rolls, 2008b,c).

## 2.8. DISTRIBUTED ENCODING

An important question for understanding brain function is whether a particular object (or face) is represented in the brain by the firing of one or a few gnostic (or “grandmother”) cells (Barlow, 1972), or whether instead the firing of a group or ensemble of cells each with somewhat different responsiveness provides the representation. Advantages of distributed codes include generalization and graceful degradation (fault tolerance), and a potentially very high capacity in the number of stimuli that can be represented (that is exponential growth of capacity with the number of neurons in the representation; Rolls and Treves, 1998, 2011; Rolls, 2008b). If the ensemble encoding is sparse, this provides a good input to an associative memory, for then large numbers of stimuli can be stored (Rolls, 2008b; Rolls and Treves, 2011). We have shown that in the inferior temporal visual cortex and cortex in the anterior part of the superior temporal sulcus (IT/STS), there is a sparse distributed representation in the firing rates of neurons about faces and objects (Rolls, 2008b; Rolls and Treves, 2011).

The information from a single cell is informative about a set of stimuli, but the information increases approximately linearly with the number of neurons in the ensemble, and can be read moderately efficiently by dot product decoding. This is what neurons can do: produce in their depolarization or firing rate a synaptically weighted sum of the firing rate inputs that they receive from other neurons (Rolls, 2008b). This property is fundamental to the mechanisms implemented in VisNet. There is little information in whether IT neurons fire synchronously or not (Aggelopoulos et al., 2005; Rolls and Treves, 2011), so that temporal syntactic binding (Singer, 1999) may not be part of the mechanism. Each neuron has an approximately exponential probability distribution of firing rates in a sparse distributed representation (Franco et al., 2007; Rolls and Treves, 2011).

These generic properties are described in detail elsewhere (Rolls, 2008b; Rolls and Treves, 2011), as are their implications for understanding brain function (Rolls, 2012), and so are not further described here. They are incorporated into the design of VisNet, as will become evident.

It is consistent with this general conceptual background that Krieman et al. (2000) have described some neurons in the human temporal lobe that seem to respond selectively to an object. This is consistent with the principles just described, though the brain areas in which these recordings were made may be beyond the inferior temporal visual cortex and the tuning appears to be more specific, perhaps reflecting backprojections from language or other cognitive areas concerned, for example, with tool use that might influence the categories represented in high-order cortical areas (Farah et al., 1996; Farah, 2000; Rolls, 2008b).

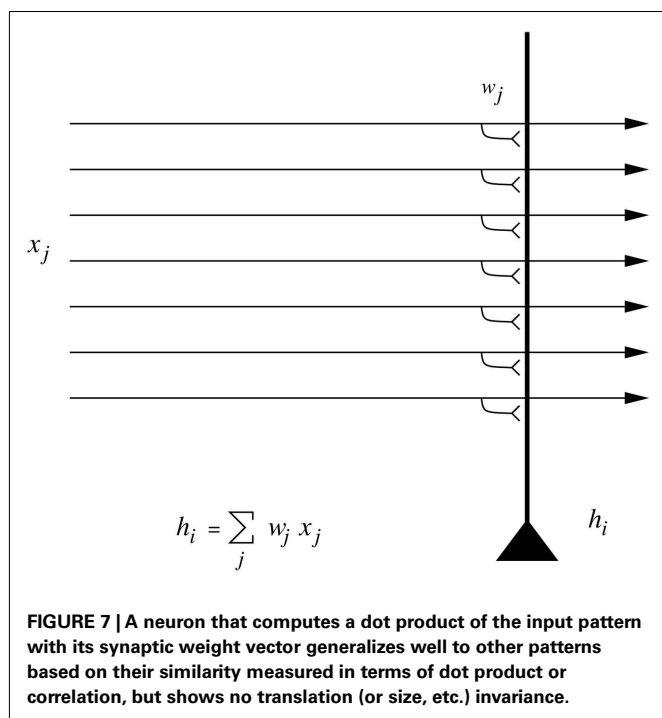
## 3. APPROACHES TO INVARIANT OBJECT RECOGNITION

A goal of my approach is to provide a biologically based and biologically plausible approach to how the brain computes invariant representations for use by other brain systems (Rolls, 2008b). This leads me to propose a hierarchical feed-forward series of competitive networks using convergence from stage to stage; and the use of a modified Hebb synaptic learning rule that incorporates a short-term memory trace of previous neuronal activity to help learn the invariant properties of objects from the temporo-spatial statistics produced by the normal viewing of objects (Wallis and Rolls, 1997; Rolls and Milward, 2000; Stringer and Rolls, 2000, 2002; Rolls and Stringer, 2001, 2006; Elliffe et al., 2002; Rolls and Deco, 2002; Deco and Rolls, 2004; Rolls, 2008b). In Sections 3.1–3.5, I summarize some other approaches to invariant object recognition, and in Section 3.6, I introduce feature hierarchies as part of the background to VisNet, which is described starting in Section 4.

I start by emphasizing that generalization to different positions, sizes, views, etc. of an object is not a simple property of one-layer neural networks. Although neural networks do generalize well, the type of generalization they show naturally is to vectors which have a high-dot product or correlation with what they have already learned. To make this clear, **Figure 7** is a reminder that the activation  $h_i$  of each neuron is computed as

$$h_i = \sum_j x_j w_{ij} \quad (1)$$

where the sum is over the  $C$  input axons, indexed by  $j$ .



Now consider translation (or shift) of the input (random binary) pattern vector by one position. The dot product will now drop to a low-level, and the neuron will not respond, even though it is the same pattern, just shifted by one location. This makes the point that special processes are needed to compute invariant representations. Network approaches to such invariant pattern recognition are described in this paper. Once an invariant representation has been computed by a sensory system, it is in a form that is suitable for presentation to a pattern association or autoassociation neural network (Rolls, 2008b).

### 3.1. FEATURE SPACES

One very simple possibility for performing object classification is based on feature spaces, which amount to lists of (the extent to which) different features are present in a particular object. The features might consist of textures, colors, areas, ratios of length to width, etc. The spatial arrangement of the features is not taken into account. If  $n$  different properties are used to characterize an object, each viewed object is represented by a set of  $n$  real numbers. It then becomes possible to represent an object by a point  $R^n$  in an  $n$ -dimensional space (where  $R$  is the resolution of the real numbers used). Such schemes have been investigated (Gibson, 1950, 1979; Selfridge, 1959; Tou and Gonzalez, 1974; Bolles and Cain, 1982; Mundy and Zisserman, 1992; Mel, 1997), but, because the relative positions of the different parts are not implemented in the object recognition scheme, are not sensitive to spatial jumbling of the features. For example, if the features consisted of nose, mouth, and eyes, such a system would respond to faces with jumbled arrangements of the eyes, nose, and mouth, which does not match human vision, nor the responses of macaque inferior temporal cortex neurons, which are sensitive to the spatial arrangement of the features in a face (Rolls et al., 1994). Similarly, such

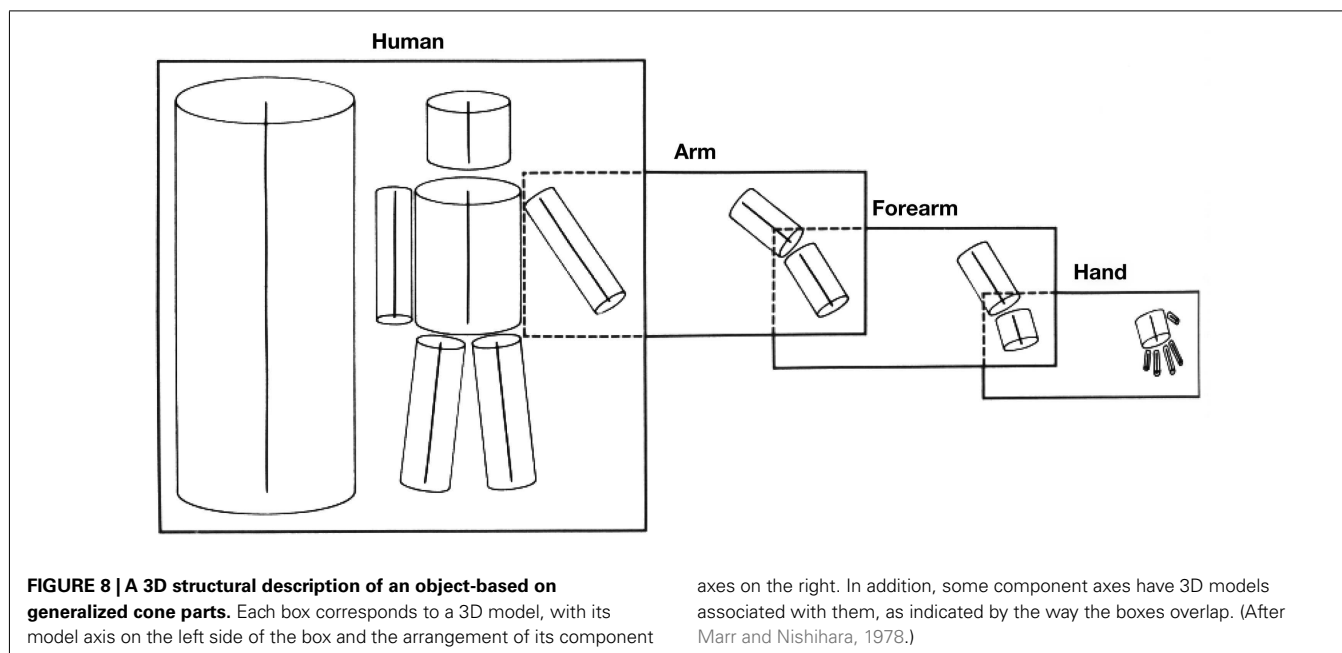
an object recognition system might not distinguish a normal car from a car with the back wheels removed and placed on the roof. Such systems do not therefore perform shape recognition (where shape implies something about the spatial arrangement of features within an object, see further Ullman, 1996), and something more is needed, and is implemented in the primate visual system. However, I note that the features that are present in objects, e.g., a furry texture, are useful to incorporate in object recognition systems, and the brain may well use, and the model VisNet in principle can use, evidence from which features are present in an object as part of the evidence for identification of a particular object. I note that the features might consist also of, for example, the pattern of movement that is characteristic of a particular object (such as a buzzing fly), and might use this as part of the input to final object identification.

The capacity to use shape in invariant object recognition is fundamental to primate vision, but may not be used or fully implemented in the visual systems of some other animals with less developed visual systems. For example, pigeons may correctly identify pictures containing people, a particular person, trees, pigeons, etc. but may fail to distinguish a figure from a scrambled version of a figure (Herrnstein, 1984; Cerella, 1986). Thus their object recognition may be based more on a collection of parts than on a direct comparison of complete figures in which the relative positions of the parts are important. Even if the details of the conclusions reached from this research are revised (Wasserman et al., 1998), it nevertheless does appear that at least some birds may use computationally simpler methods than those needed for invariant shape recognition. For example, it may be that when some birds are trained to discriminate between images in a large set of pictures, they tend to rely on some chance detail of each picture (such as a spot appearing by mistake on the picture), rather than on recognition of the shapes of the object in the picture (Watanabe et al., 1993).

### 3.2. STRUCTURAL DESCRIPTIONS AND SYNTACTIC PATTERN RECOGNITION

A second approach to object recognition is to decompose the object or image into parts, and to then produce a structural description of the relations between the parts. The underlying assumption is that it is easier to capture object invariances at a level where parts have been identified. This is the type of scheme for which Marr and Nishihara (1978) and Marr (1982) opted (Rolls, 2011a). The particular scheme (Binford, 1981) they adopted consists of generalized cones, series of which can be linked together to form structural descriptions of some, especially animate, stimuli (see Figure 8).

Such schemes assume that there is a 3D internal model (structural description) of each object. Perception of the object consists of parsing or segmenting the scene into objects, and then into parts, then producing a structural description of the object, and then testing whether this structural description matches that of any known object stored in the system. Other examples of structural description schemes include those of Sutherland (1968), Winston (1975), and Milner (1974). The relations in the structural description may need to be quite complicated, for example, “connected together,” “inside of,” “larger than,” etc.



Perhaps the most developed model of this type is the recognition by components (RBC) model of Biederman (1987), implemented in a computational model by Hummel and Biederman (1992). His small set (less than 50) of primitive parts named “geons” includes simple 3D shapes such as boxes, cylinders, and wedges. Objects are described by a syntactically linked list of the relations between each of the geons of which they are composed. Describing a table in this way (as a flat top supported by three or four legs) seems quite economical. Other schemes use 2D surface patches as their primitives (Dane and Bajcsy, 1982; Brady et al., 1985; Faugeras and Hebert, 1986; Faugeras, 1993). When 3D objects are being recognized, the implication is that the structural description is a 3D description. This is in contrast to feature hierarchical systems, in which recognition of a 3D object from any view might be accomplished by storing a set of associated 2D views (see below, Section 3.6).

There are a number of difficulties with schemes based on structural descriptions, some general, and some with particular reference to the potential difficulty of their implementation in the brain. First, it is not always easy to decompose the object into its separate parts, which must be performed before the structural description can be produced. For example, it may be difficult to produce a structural description of a cat curled up asleep from separately identifiable parts. Identification of each of the parts is also frequently very difficult when 3D objects are seen from different viewing angles, as key parts may be invisible or highly distorted. This is particularly likely to be difficult in 3D shape perception. It appears that being committed to producing a correct description of the parts before other processes can operate is making too strong a commitment early on in the recognition process.

A second difficulty is that many objects or animals that can be correctly recognized have rather similar structural descriptions.

For example, the structural description of many four-legged animals is rather similar. Rather more than a structural description seems necessary to identify many objects and animals.

A third difficulty, which applies especially to biological systems, is the difficulty of implementing the syntax needed to hold the structural description as a 3D model of the object, of producing a syntactic structural description on the fly (in real time, and with potentially great flexibility of the possible arrangement of the parts), and of matching the syntactic description of the object in the image to all the stored representations in order to find a match. An example of a structural description for a limb might be body > thigh > shin > foot > toes. In this description > means “is linked to,” and this link must be between the correct pair of descriptors. If we had just a set of parts, without the syntactic or relational linking, then there would be no way of knowing whether the toes are attached to the foot or to the body. In fact, worse than this, there would be no evidence about what was related to what, just a set of parts. Such syntactical relations are difficult to implement in any biologically plausible neuronal networks used in vision, because if the representations of all the features or parts just mentioned were active simultaneously, how would the spatial relations between the features also be encoded? (How would it be apparent just from the firing of neurons that the toes were linked to the rest of the foot but not to the body?) It would be extremely difficult to implement this “on the fly” syntactic binding in a biologically plausible network (though cf. Hummel and Biederman, 1992), and the only suggested mechanism for flexible syntactic binding, temporal synchronization of the firing of different neurons, is not well supported as a quantitatively important mechanism for information encoding in the ventral visual system, and would have major difficulties in implementing correct, relational, syntactic binding (Section 5.4.1; Rolls, 2008b; Rolls and Treves, 2011).

A fourth difficulty of the structural description approach is that segmentation into objects must occur effectively before object

recognition, so that the linked structural description list can be of one object. Given the difficulty of segmenting objects in typical natural cluttered scenes (Ullman, 1996), and the compounding problem of overlap of parts of objects by other objects, segmentation as a first necessary stage of object recognition adds another major difficulty for structural description approaches.

A fifth difficulty is that metric information, such as the relative size of the parts that are linked syntactically, needs to be specified in the structural description (Stan-Kiewicz and Hummel, 1994), which complicates the parts that have to be syntactically linked.

It is because of these difficulties that even in artificial vision systems implemented on computers, where almost unlimited syntactic binding can easily be implemented, the structural description approach to object recognition has not yet succeeded in producing a scheme which actually works in more than an environment in which the types of objects are limited, and the world is far from the natural world, consisting, for example, of 2D scenes (Mundy and Zisserman, 1992).

Although object recognition in the brain is unlikely to be based on the structural description approach, for the reasons given above, and the fact that the evidence described in this paper supports a feature hierarchy rather than the structural description implementation in the brain, it is certainly the case that humans can provide verbal, syntactic, descriptions of objects in terms of the relations of their parts, and that this is often a useful type of description. Humans may therefore, it is suggested, supplement a feature hierarchical object recognition system built into their ventral visual system with the additional ability to use the type of syntax that is necessary for language to provide another level of description of objects. This ability is useful in, for example, engineering applications.

### 3.3. TEMPLATE MATCHING AND THE ALIGNMENT APPROACH

Another approach is template matching, comparing the image on the retina with a stored image or picture of an object. This is conceptually simple, but there are in practice major problems. One major problem is how to align the image on the retina with the stored images, so that all possible images on the retina can be compared with the stored template or templates of each object.

The basic idea of the alignment approach (Ullman, 1996) is to compensate for the transformations separating the viewed object and the corresponding stored model, and then compare them. For example, the image and the stored model may be similar, except for a difference in size. Scaling one of them will remove this discrepancy and improve the match between them. For a 2D world, the possible transforms are translation (shift), scaling, and rotation. Given, for example, an input letter of the alphabet to recognize, the system might, after segmentation (itself a very difficult process if performed independently of (prior to) object recognition), compensate for translation by computing the center of mass of the object, and shifting the character to a “canonical location.” Scale might be compensated for by calculating the convex hull (the smallest envelope surrounding the object), and then scaling the image. Of course how the shift and scaling would be accomplished is itself a difficult point – easy to perform on a computer using matrix multiplication as in simple computer graphics, but not the sort of computation that could be performed easily or accurately

by any biologically plausible network. Compensating for rotation is even more difficult (Ullman, 1996). All this has to happen before the segmented canonical representation of the object is compared to the stored object templates with the same canonical representation. The system of course becomes vastly more complicated when the recognition must be performed of 3D objects seen in a 3D world, for now the particular view of an object after segmentation must be placed into a canonical form, regardless of which view, or how much of any view, may be seen in a natural scene with occluding contours. However, this process is helped, at least in computers that can perform high-precision matrix multiplication, by the fact that (for many continuous transforms such as 3D rotation, translation, and scaling) all the possible views of an object transforming in 3D space can be expressed as the linear combination of other views of the same object (see Chapter 5 of Ullman, 1996; Koenderink and van Doorn, 1991; Koenderink, 1990).

This alignment approach is the main theme of the book by Ullman (1996), and there are a number of computer implementations (Lowe, 1985; Grimson, 1990; Huttenlocher and Ullman, 1990; Shashua, 1995). However, as noted above, it seems unlikely that the brain is able to perform the high-precision calculations needed to perform the transforms required to align any view of a 3D object with some canonical template representation. For this reason, and because the approach also relies on segmentation of the object in the scene before the template alignment algorithms can start, and because key features may need to be correctly identified to be used in the alignment (Edelman, 1999), this approach is not considered further here.

We may note here in passing that some animals with a less computationally developed visual system appear to attempt to solve the alignment problem by actively moving their heads or eyes to see what template fits, rather than starting with an image on the eye and attempting to transform it into canonical coordinates. This “active vision” approach used, for example, by some invertebrates has been described by Land (1999) and Land and Collett (1997).

### 3.4. SOME FURTHER MACHINE LEARNING APPROACHES

Learning the transformations and invariances of the signal is another approach to invariant object recognition at the interface of machine learning and theoretical neuroscience. For example, rather than focusing on the templates, “map-seeking circuit theory” focuses on the transforms (Arathorn, 2002, 2005). The theory provides a general computational mechanism for discovery of correspondences in massive transformation spaces by exploiting an ordering property of superpositions. The latter allows a set of transformations of an input image to be formed into a sequence of superpositions which are then “culled” to a composition of single mappings by a competitive process which matches each superposition against a superposition of inverse transformations of memory patterns. Earlier work considered how to minimize the variance in the output when the image transformed (Leen, 1995). Another approach is to add transformation invariance to mixture models, by approximating the non-linear transformation manifold by a discrete set of points (Frey and Jojic, 2003). They showed how the expectation maximization algorithm can be used to jointly learn clusters, while at the same time inferring the transformation associated with each input. In another approach, an unsupervised

algorithm for learning Lie group operators for in-plane transforms from input data was described (Rao and Ruderman, 1999).

### 3.5. NETWORKS THAT CAN RECONSTRUCT THEIR INPUTS

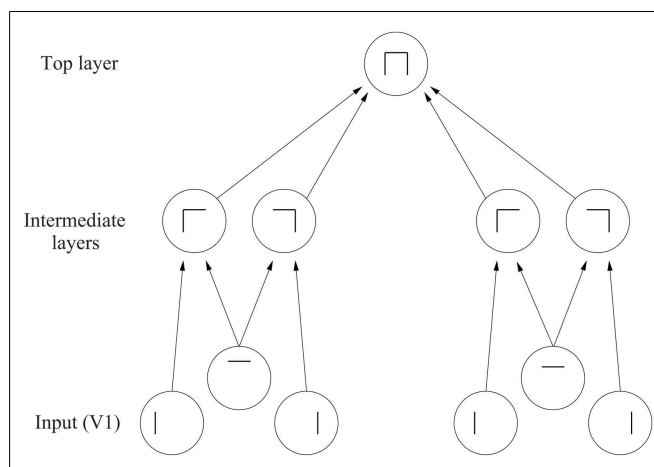
Hinton et al. (1995) and Hinton and Ghahramani (1997) have argued that cortical computation is invertible, so that, for example, the forward transform of visual information from V1 to higher areas loses no information, and there can be a backward transform from the higher areas to V1. A comparison of the reconstructed representation in V1 with the actual image from the world might in principle be used to correct all the synaptic weights between the two (in both the forward and the reverse directions), in such a way that there are no errors in the transform (Hinton, 2010). This suggested reconstruction scheme would seem to involve non-local synaptic weight correction (though see Hinton and Sejnowski, 1986; O'Reilly and Munakata, 2000) for a suggested, although still biologically implausible, neural implementation, contrastive Hebbian learning), or other biologically implausible operations. The scheme also does not seem to provide an account for why or how the responses of inferior temporal cortex neurons become the way they are (providing information about which object is seen relatively independently of position on the retina, size, or view). The whole forward transform performed in the brain seems to lose much of the information about the size, position, and view of the object, as it is evidence about which object is present invariant of its size, view, etc. that is useful to the stages of processing about objects that follow (Rolls, 2008b). Because of these difficulties, and because the backprojections are needed for processes such as recall (Rolls, 2008b), this approach is not considered further here.

In the context of recall, if the visual system were to perform a reconstruction in V1 of a visual scene from what is represented in the inferior temporal visual cortex, then it might be supposed that remembered visual scenes might be as information-rich (and subjectively as full of rich detail) as seeing the real thing. This is not the case for most humans, and indeed this point suggests that at least what reaches consciousness from the inferior temporal visual cortex (which is activated during the recall of visual memories) is the identity of the object (as made explicit in the firing rate of the neurons), and not the low-level details of the exact place, size, and view of the object in the recalled scene, even though, according to the reconstruction argument, that information should be present in the inferior temporal visual cortex.

### 3.6. FEATURE HIERARCHIES AND 2D VIEW-BASED OBJECT RECOGNITION

Another approach, and one that is much closer to what appears to be present in the primate ventral visual system (Wurtz and Kandel, 2000a; Rolls and Deco, 2002; Rolls, 2008b), is a feature hierarchy system (see Figure 9).

In this approach, the system starts with some low-level description of the visual scene, in terms, for example, of oriented straight line segments of the type that are represented in the responses of primary visual cortex (V1) neurons, and then builds in repeated hierarchical layers features based on what is represented in previous layers. A feature may thus be defined as a combination of what is represented in the previous layer. For example, after V1, features might consist of combinations of straight lines, which might



**FIGURE 9 | The feature hierarchy approach to object recognition.** The inputs may be neurons tuned to oriented straight line segments. In early intermediate-layers neurons respond to a combination of these inputs in the correct spatial position with respect to each other. In further intermediate layers, of which there may be several, neurons respond with some invariance to the feature combinations represented early, and form higher order feature combinations. Finally, in the top layer, neurons respond to combinations of what is represented in the preceding intermediate layer, and thus provide evidence about objects in a position (and scale and even view) invariant way. Convergence through the network is designed to provide top layer neurons with information from across the entire input retina, as part of the solution to translation invariance, and other types of invariance are treated similarly.

represent longer curved lines (Zucker et al., 1989), or terminated lines (in fact represented in V1 as end-stopped cells), corners, “T” junctions which are characteristic of obscuring edges, and (at least in humans) the arrow and “Y” vertices which are characteristic properties of man-made environments. Evidence that such feature combination neurons are present in V2 is that some neurons respond to combinations of line elements that join at different angles (Hegde and Van Essen, 2000, 2003, 2007; Ito and Komatsu, 2004; Anzai et al., 2007). (An example of this might be a neuron responding to a “V” shape at a particular orientation.) As one ascends the hierarchy, neurons might respond to more complex trigger features. For example, two parts of a complex figure may need to be in the correct spatial arrangement with respect to each other, as shown by Tanaka (1996) for V4 and posterior inferior temporal cortex neurons. In another example, V4 neurons may respond to the curvature of the elements of a stimulus (Carlson et al., 2011). Further on, neurons might respond to combinations of several such intermediate-level feature combination neurons, and thus come to respond systematically differently to different objects, and thus to convey information about which object is present. This approach received neurophysiological support early on from the results of Hubel and Wiesel (1962) and Hubel and Wiesel (1968) in the cat and monkey, and many of the data described in Chapter 5 of Rolls and Deco (2002) are consistent with this scheme.

A number of problems need to be solved for such feature hierarchy visual systems to provide a useful model of object recognition in the primate visual system.

First, some way needs to be found to keep the number of feature combination neurons realistic at each stage, without undergoing a combinatorial explosion. If a separate feature combination neuron was needed to code for every possible combination of  $n$  types of feature each with a resolution of 2 levels (binary encoding) in the preceding stage, then  $2^n$  neurons would be needed. The suggestion that is made in Section 4 is that by forming neurons that respond to low-order combinations of features (neurons that respond to just say 2–4 features from the preceding stage), the number of actual feature analyzing neurons can be kept within reasonable numbers. By reasonable we mean the number of neurons actually found at any one stage of the visual system, which, for V4 might be in the order of  $60 \times 10^6$  neurons (assuming a volume for macaque V4 of approximately  $2,000 \text{ mm}^3$ , and a cell density of 20,000–40,000 neurons per  $\text{mm}^3$ , Rolls, 2008b). This is certainly a large number; but the fact that a large number of neurons is present at each stage of the primate visual system is in fact consistent with the hypothesis that feature combination neurons are part of the way in which the brain solves object recognition. A factor which also helps to keep the number of neurons under control is the statistics of the visual world, which contain great redundancies. The world is not random, and indeed the statistics of natural images are such that many regularities are present (Field, 1994), and not every possible combination of pixels on the retina needs to be separately encoded. A third factor which helps to keep the number of connections required onto each neuron under control is that in a multilayer hierarchy each neuron can be set up to receive connections from only a small region of the preceding layer. Thus an individual neuron does not need to have connections from all the neurons in the preceding layer. Over multiple-layers, the required convergence can be produced so that the same neurons in the top layer can be activated by an image of an effective object anywhere on the retina (see **Figure 1**).

A second problem of feature hierarchy approaches is how to map all the different possible images of an individual object through to the same set of neurons in the top layer by modifying the synaptic connections (see **Figure 1**). The solution discussed in Sections 4, 5.1.1, and 5.3 is the use of a synaptic modification rule with a short-term memory trace of the previous activity of the neuron, to enable it to learn to respond to the now transformed version of what was seen very recently, which, given the statistics of looking at the visual world, will probably be an input from the same object.

A third problem of feature hierarchy approaches is how they can learn in just a few seconds of inspection of an object to recognize it in different transforms, for example, in different positions on the retina in which it may never have been presented during training. A solution to this problem is provided in Section 5.4, in which it is shown that this can be a natural property of feature hierarchy object recognition systems, if they are trained first for all locations on the intermediate-level feature combinations of which new objects will simply be a new combination, and therefore requiring learning only in the upper layers of the hierarchy.

A fourth potential problem of feature hierarchy systems is that when solving translation invariance they need to respond to the same local spatial arrangement of features (which are needed to specify the object), but to ignore the global position of the whole

object. It is shown in Section 5.4 that feature hierarchy systems can solve this problem by forming feature combination neurons at an early stage of processing (e.g., V1 or V2 in the brain) that respond with high-spatial precision to the local arrangement of features. Such neurons would respond differently, for example, to L, +, and T if they receive inputs from two line-responding neurons. It is shown in Section 5.4 that at later layers of the hierarchy, where some of the intermediate-level feature combination neurons are starting to show translation invariance, then correct object recognition may still occur because only one object contains just those sets of intermediate-level neurons in which the spatial representation of the features is inherent in the encoding.

The type of representation developed in a hierarchical object recognition system, in the brain, and by VisNet as described in the rest of this paper would be suitable for recognition of an object, and for linking associative memories to objects, but would be less good for making actions in 3D space to particular parts of, or inside, objects, as the 3D coordinates of each part of the object would not be explicitly available. It is therefore proposed that visual fixation is used to locate in foveal vision part of an object to which movements must be made, and that local disparity and other measurements of depth (made explicit in the dorsal visual system) then provide sufficient information for the motor system to make actions relative to the small part of space in which a local, view-dependent, representation of depth would be provided (cf. Ballard, 1990).

One advantage of feature hierarchy systems is that they can operate fast (Rolls, 2008b).

A second advantage is that the feature analyzers can be built out of the rather simple competitive networks (Rolls, 2008b) which use a local learning rule, and have no external teacher, so that they are rather biologically plausible. Another advantage is that, once trained on subset features common to most objects, the system can then learn new objects quickly.

A related third advantage is that, if implemented with competitive nets as in the case of VisNet (see Section 5), then neurons are allocated by self-organization to represent just the features present in the natural statistics of real images (cf. Field, 1994), and not every possible feature that could be constructed by random combinations of pixels on the retina.

A related fourth advantage of feature hierarchy networks is that because they can utilize competitive networks, they can still produce the best guess at what is in the image under non-ideal conditions, when only parts of objects are visible because, for example, of occlusion by other objects, etc. The reasons for this are that competitive networks assess the evidence for the presence of certain “features” to which they are tuned using a dot product operation on their inputs, so that they are inherently tolerant of missing input evidence; and reach a state that reflects the best hypothesis or hypotheses (with soft competition) given the whole set of inputs, because there are competitive interactions between the different neurons (Rolls, 2008b).

A fifth advantage of a feature hierarchy system is that, as shown in Section 5.5, the system does not need to perform segmentation into objects as part of pre-processing, nor does it need to be able to identify parts of an object, and can also operate in cluttered scenes in which the object may be partially obscured. The reason



for this is that once trained on objects, the system then operates somewhat like an associative memory, mapping the image properties forward onto whatever it has learned about before, and then by competition selecting just the most likely output to be activated. Indeed, the feature hierarchy approach provides a mechanism by which processing at the object recognition level could feed back using backprojections to early cortical areas to provide top-down guidance to assist segmentation. Although backprojections are not built into VisNet2 (Rolls and Milward, 2000), they have been added when attentional top-down processing must be incorporated (Deco and Rolls, 2004), are present in the brain, and are incorporated into the models described elsewhere (Rolls, 2008b). Although the operation of the ventral visual system can proceed as a feed-forward hierarchy, as shown by backward masking experiments (Rolls and Tovee, 1994; Rolls et al., 1999; Rolls, 2003, 2006), top-down influences can of course be implemented by the backprojections, and may be useful in further shaping the activity of neurons at lower levels in the hierarchy based on the neurons firing at a higher level as a result of dynamical interactions of neurons at different layers of the hierarchy (Rolls, 2008b; Jiang et al., 2011).

A sixth advantage of feature hierarchy systems is that they can naturally utilize features in the images of objects which are not strictly part of a shape description scheme, such as the fact that different objects have different textures, colors, etc. Feature hierarchy systems, because they utilize whatever is represented at earlier stages in forming feature combination neurons at the next stage, naturally incorporate such “feature list” evidence into their analysis, and have the advantages of that approach (see Section 3.1 and also Mel, 1997). Indeed, the feature space approach can utilize a hybrid representation, some of whose dimensions may be discrete and defined in structural terms, while other dimensions may be continuous and defined in terms of metric details, and others may be concerned with non-shape properties such as texture and color (cf. Edelman, 1999).

A seventh advantage of feature hierarchy systems is that they do not need to utilize “on the fly” or run-time arbitrary binding of features. Instead, the spatial syntax is effectively hard-wired into the system when it is trained, in that the feature combination neurons have learned to respond to their set of features when they are in a given spatial arrangement on the retina.

An eighth advantage of feature hierarchy systems is that they can self-organize (given the right functional architecture, trace synaptic learning rule, and the temporal statistics of the normal visual input from the world), with no need for an external teacher to specify that the neurons must learn to respond to objects. The correct, object, representation self-organizes itself given rather economically specified genetic rules for building the network (cf. Rolls and Stringer, 2000).

Ninth, it is also noted that hierarchical visual systems may recognize 3D objects based on a limited set of 2D views of objects, and that the same architectural rules just stated and implemented in VisNet will correctly associate together the different views of an object. It is part of the concept (see below), and consistent with neurophysiological data (Tanaka, 1996), that the neurons in the upper layers will generalize correctly within a view (see Section 5.6).

After the immediately following description of early models of a feature hierarchy approach implemented in the Cognitron and Neocognitron, we turn for the remainder of this paper to analyses of how a feature hierarchy approach to invariant visual object recognition might be implemented in the brain, and how key computational issues could be solved by such a system. The analyses are developed and tested with a model, VisNet, which will shortly be described. Much of the data we have on the operation of the high-order visual cortical areas (Section 2; Rolls and Deco, 2002; Anzai et al., 2007; Rolls, 2008b) suggest that they implement a feature hierarchy approach to visual object recognition, as is made evident in the remainder of this paper.

### 3.6.1. The cognitron and neocognitron

An early computational model of a hierarchical feature-based approach to object recognition, joining other early discussions of this approach (Selfridge, 1959; Sutherland, 1968; Barlow, 1972; Milner, 1974), was proposed by Fukushima (1975, 1980, 1989, 1991). His model used two types of cell within each layer to approach the problem of invariant representations. In each layer, a set of “simple cells,” with defined position, orientation, etc. sensitivity for the stimuli to which they responded, was followed by a set of “complex cells,” which generalized a little over position, orientation, etc. This simple cell – complex cell pairing within each layer provided some invariance. When a neuron in the network using competitive learning with its stimulus set, which was typically letters on a  $16 \times 16$  pixel array, learned that a particular feature combination had occurred, that type of feature analyzer was replicated in a non-local manner throughout the layer, to provide further translation invariance. Invariant representations were thus learned in a different way from VisNet. Up to eight layers were used. The network could learn to differentiate letters, even with some translation, scaling, or distortion. Although internally it is organized and learns very differently to VisNet, it is an independent example of the fact that useful invariant pattern recognition can be performed by multilayer hierarchical networks. A major biological implausibility of the system is that once one neuron within a layer learned, other similar neurons were set up throughout the layer by a non-local process. A second biological limitation was that no learning rule or self-organizing process was specified as to how the complex cells can provide translation-invariant representations of simple cell responses – this was simply handwired. Solutions to both these issues are provided by VisNet.

## 4. HYPOTHESES ABOUT THE COMPUTATIONAL MECHANISMS IN THE VISUAL CORTEX FOR OBJECT RECOGNITION

The neurophysiological findings described in Section 2, and wider considerations on the possible computational properties of the cerebral cortex (Rolls, 1992, 2000, 2008b; Rolls and Treves, 1998; Rolls and Deco, 2002), lead to the following outline working hypotheses on object recognition by visual cortical mechanisms (see Rolls, 1992). The principles underlying the processing of faces and other objects may be similar, but more neurons may become allocated to represent different aspects of faces because of the need to recognize the faces of many different individuals, that is to identify many individuals within the category faces.



Cortical visual processing for object recognition is considered to be organized as a set of hierarchically connected cortical regions consisting at least of V1, V2, V4, posterior inferior temporal cortex (TEO), inferior temporal cortex (e.g., TE3, TEa, and TEm), and anterior temporal cortical areas (e.g., TE2 and TE1). (This stream of processing has many connections with a set of cortical areas in the anterior part of the superior temporal sulcus, including area TPO.) There is convergence from each small part of a region to the succeeding region (or layer in the hierarchy) in such a way that the receptive field sizes of neurons (e.g.,  $1^\circ$  near the fovea in V1) become larger by a factor of approximately 2.5 with each succeeding stage (and the typical parafoveal receptive field sizes found would not be inconsistent with the calculated approximations of, e.g.,  $8^\circ$  in V4,  $20^\circ$  in TEO, and  $50^\circ$  in the inferior temporal cortex Boussaoud et al., 1991; see **Figure 1**). Such zones of convergence would overlap continuously with each other (see **Figure 1**). This connectivity would be part of the architecture by which translation-invariant representations are computed.

Each layer is considered to act partly as a set of local self-organizing competitive neuronal networks with overlapping inputs. (The region within which competition would be implemented would depend on the spatial properties of inhibitory interneurons, and might operate over distances of 1–2 mm in the cortex.) These competitive nets operate by a single set of forward inputs leading to (typically non-linear, e.g., sigmoid) activation of output neurons; of competition between the output neurons mediated by a set of feedback inhibitory interneurons which receive from many of the principal (in the cortex, pyramidal) cells in the net and project back (via inhibitory interneurons) to many of the principal cells and serve to decrease the firing rates of the less active neurons relative to the rates of the more active neurons; and then of synaptic modification by a modified Hebb rule, such that synapses to strongly activated output neurons from active input axons strengthen, and from inactive input axons weaken (Rolls, 2008b). A biologically plausible form of this learning rule that operates well in such networks is

$$\delta w_{ij} = \alpha y_i (x_j - w_{ij}) \quad (2)$$

where  $\delta w_{ij}$  is the change of the synaptic weight,  $\alpha$  is a learning rate constant,  $y_i$  is the firing rate of the  $i$ th postsynaptic neuron, and  $x_j$  and  $w_{ij}$  are in appropriate units (Rolls, 2008b). Such competitive networks operate to detect correlations between the activity of the input neurons, and to allocate output neurons to respond to each cluster of such correlated inputs. These networks thus act as categorizers. In relation to visual information processing, they would remove redundancy from the input representation, and would develop low-entropy representations of the information (cf. Barlow, 1985; Barlow et al., 1989). Such competitive nets are biologically plausible, in that they utilize Hebb-modifiable forward excitatory connections, with competitive inhibition mediated by cortical inhibitory neurons. The competitive scheme I suggest would not result in the formation of “winner-take-all” or “grandmother” cells, but would instead result in a small ensemble of active neurons representing each input (Rolls and Treves, 1998; Rolls, 2008b). The scheme has the advantages that the output neurons learn better to distribute themselves between the input patterns (cf. Bennett, 1990), and that the sparse representations formed

have utility in maximizing the number of memories that can be stored when, toward the end of the visual system, the visual representation of objects is interfaced to associative memory (Rolls and Treves, 1998; Rolls, 2008b).

Translation invariance would be computed in such a system by utilizing competitive learning to detect regularities in inputs when real objects are translated in the physical world. The hypothesis is that because objects have continuous properties in space and time in the world, an object at one place on the retina might activate feature analyzers at the next stage of cortical processing, and when the object was translated to a nearby position, because this would occur in a short period (e.g., 0.5 s), the membrane of the post-synaptic neuron would still be in its “Hebb-modifiable” state (caused, for example, by calcium entry as a result of the voltage-dependent activation of NMDA receptors), and the presynaptic afferents activated with the object in its new position would thus become strengthened on the still-activated post-synaptic neuron. It is suggested that the short temporal window (e.g., 0.5 s) of Hebb-modifiability helps neurons to learn the statistics of objects moving in the physical world, and at the same time to form different representations of different feature combinations or objects, as these are physically discontinuous and present less regular correlations to the visual system. Földiák (1991) has proposed computing an average activation of the post-synaptic neuron to assist with the same problem. One idea here is that the temporal properties of the biologically implemented learning mechanism are such that it is well suited to detecting the relevant continuities in the world of real objects. Another suggestion is that a memory trace for what has been seen in the last 300 ms appears to be implemented by a mechanism as simple as continued firing of inferior temporal neurons after the stimulus has disappeared, as has been found in masking experiments (Rolls and Tovee, 1994; Rolls et al., 1994, 1999; Rolls, 2003).

I also suggested (Rolls, 1992) that other invariances, for example, size, spatial-frequency, and rotation invariance, could be learned by a comparable process. (Early processing in V1 which enables different neurons to represent inputs at different spatial scales would allow combinations of the outputs of such neurons to be formed at later stages. Scale invariance would then result from detecting at a later stage which neurons are almost conjunctively active as the size of an object alters.) It is suggested that this process takes place at each stage of the multiple-layer cortical processing hierarchy, so that invariances are learned first over small regions of space, and then over successively larger regions. This limits the size of the connection space within which correlations must be sought.

Increasing complexity of representations could also be built in such a multiple-layer hierarchy by similar mechanisms. At each stage or layer the self-organizing competitive nets would result in combinations of inputs becoming the effective stimuli for neurons. In order to avoid the combinatorial explosion, it is proposed, following Feldman (1985), that low-order combinations of inputs would be what is learned by each neuron. (Each input would not be represented by activity in a single input axon, but instead by activity in a set of active input axons.) Evidence consistent with this suggestion that neurons are responding to combinations of a few variables represented at the preceding stage of cortical processing is that some neurons in V1 respond to combinations of

bars or edges (Shevelev et al., 1995; Sillito et al., 1995); V2 and V4 respond to end-stopped lines, to angles formed by a combination of lines, to tongues flanked by inhibitory subregions, or to combinations of colors (Hegde and Van Essen, 2000, 2003, 2007; Ito and Komatsu, 2004; Anzai et al., 2007; Orban, 2011); in posterior inferior temporal cortex to stimuli which may require two or more simple features to be present (Tanaka et al., 1990); and in the temporal cortical face processing areas to images that require the presence of several features in a face (such as eyes, hair, and mouth) in order to respond (Perrett et al., 1982; Yamane et al., 1988; Rolls, 2011b; see **Figure 6**). (Precursor cells to face-responsive neurons might, it is suggested, respond to combinations of the outputs of the neurons in V1 that are activated by faces, and might be found in areas such as V4.) It is an important part of this suggestion that some local spatial information would be inherent in the features which were being combined. For example, cells might not respond to the combination of an edge and a small circle unless they were in the correct spatial relation to each other. (This is in fact consistent with the data of Tanaka et al. (1990), and with our data on face neurons, in that some face neurons require the face features to be in the correct spatial configuration, and not jumbled, Rolls et al. (1994).) The local spatial information in the features being combined would ensure that the representation at the next level would contain some information about the (local) arrangement of features. Further low-order combinations of such neurons at the next stage would include sufficient local spatial information so that an arbitrary spatial arrangement of the same features would not activate the same neuron, and this is the proposed, and limited, solution which this mechanism would provide for the feature binding problem (Elliffe et al., 2002; cf. von der Malsburg, 1990). By this stage of processing a view-dependent representation of objects suitable for view-dependent processes such as behavioral responses to face expression and gesture would be available.

It is suggested that view-independent representations could be formed by the same type of computation, operating to combine a limited set of views of objects. The plausibility of providing view-independent recognition of objects by combining a set of different views of objects has been proposed by a number of investigators (Koenderink and Van Doorn, 1979; Poggio and Edelman, 1990; Logothetis et al., 1994; Ullman, 1996). Consistent with the suggestion that the view-independent representations are formed by combining view-dependent representations in the primate visual system, is the fact that in the temporal cortical areas, neurons with view-independent representations of faces are present in the same cortical areas as neurons with view-dependent representations (from which the view-independent neurons could receive inputs; Perrett et al., 1985; Hasselmo et al., 1989b; Booth and Rolls, 1998). This solution to “object-based” representations is very different from that traditionally proposed for artificial vision systems, in which the coordinates in 3D space of objects are stored in a database, and general-purpose algorithms operate on these to perform transforms such as translation, rotation, and scale change in 3D space (e.g., Marr, 1982). In the present, much more limited but more biologically plausible scheme, the representation would be suitable for recognition of an object, and for linking associative memories to objects, but would be less good for making actions in 3D space to particular parts of, or inside, objects, as the 3D

coordinates of each part of the object would not be explicitly available. It is therefore proposed that visual fixation is used to locate in foveal vision part of an object to which movements must be made, and that local disparity and other measurements of depth then provide sufficient information for the motor system to make actions relative to the small part of space in which a local, view-dependent, representation of depth would be provided (cf. Ballard, 1990).

The computational processes proposed above operate by an unsupervised learning mechanism, which utilizes statistical regularities in the physical environment to enable representations to be built. In some cases it may be advantageous to utilize some form of mild teaching input to the visual system, to enable it to learn, for example, that rather similar visual inputs have very different consequences in the world, so that different representations of them should be built. In other cases, it might be helpful to bring representations together, if they have identical consequences, in order to use storage capacity efficiently. It is proposed elsewhere (Rolls, 1989a,b, 2008b; Rolls and Treves, 1998) that the backprojections from each adjacent cortical region in the hierarchy (and from the amygdala and hippocampus to higher regions of the visual system) play such a role by providing guidance to the competitive networks suggested above to be important in each cortical area. This guidance, and also the capability for recall, are it is suggested implemented by Hebb-modifiable connections from the backprojecting neurons to the principal (pyramidal) neurons of the competitive networks in the preceding stages (Rolls, 1989a,b, 2008b; Rolls and Treves, 1998).

The computational processes outlined above use sparse distributed coding with relatively finely tuned neurons with a graded response region centered about an optimal response achieved when the input stimulus matches the synaptic weight vector on a neuron. The distributed nature of the coding but with fine tuning would help to limit the combinatorial explosion, to keep the number of neurons within the biological range. The graded response region would be crucial in enabling the system to generalize correctly to solve, for example, the invariances. However, such a system would need many neurons, each with considerable learning capacity, to solve visual perception in this way. This is fully consistent with the large number of neurons in the visual system, and with the large number of, probably modifiable, synapses on each neuron (e.g., 10,000). Further, the fact that many neurons are tuned in different ways to faces is consistent with the fact that in such a computational system, many neurons would need to be sensitive (in different ways) to faces, in order to allow recognition of many individual faces when all share a number of common properties.

## 5. THE FEATURE HIERARCHY APPROACH TO INVARIANT OBJECT RECOGNITION: COMPUTATIONAL ISSUES

The feature hierarchy approach to invariant object recognition was introduced in Section 3.6, and advantages and disadvantages of it were discussed. Hypotheses about how object recognition could be implemented in the brain which are consistent with much of the neurophysiology discussed in Section 2 and by Rolls and Deco (2002) and Rolls (2008b) were set out in Section 4. These hypotheses effectively incorporate a feature hierarchy system while encompassing much of the neurophysiological evidence.

In this Section (5), we consider the computational issues that arise in such feature hierarchy systems, and in the brain systems that implement visual object recognition. The issues are considered with the help of a particular model, VisNet, which requires precise specification of the hypotheses, and at the same time enables them to be explored and tested numerically and quantitatively. However, I emphasize that the issues to be covered in Section 5 are key and major computational issues for architectures of this feature hierarchical type (Rolls, 2008b), and are very relevant to understanding how invariant object recognition is implemented in the brain.

VisNet is a model of invariant object recognition based on Rolls' (Rolls, 1992) hypotheses. It is a computer simulation that allows hypotheses to be tested and developed about how multilayer hierarchical networks of the type believed to be implemented in the visual cortical pathways operate. The architecture captures a number of aspects of the architecture of the visual cortical pathways, and is described next. The model of course, as with all models, requires precise specification of what is to be implemented, and at the same time involves specified simplifications of the real architecture, as investigations of the fundamental aspects of the information processing being performed are more tractable in a simplified and at the same time quantitatively specified model. First the architecture of the model is described, and this is followed by descriptions of key issues in such multilayer feature hierarchical models, such as the issue of feature binding, the optimal form of training rule for the whole system to self-organize, the operation of the network in natural environments and when objects are partly occluded, how outputs about individual objects can be read out from the network, and the capacity of the system.

## 5.1. THE ARCHITECTURE OF VisNet

Fundamental elements of Rolls' (1992) theory for how cortical networks might implement invariant object recognition are described in Section 4. They provide the basis for the design of VisNet, and can be summarized as:

- A series of competitive networks, organized in hierarchical layers, exhibiting mutual inhibition over a short range within each layer. These networks allow combinations of features or inputs occurring in a given spatial arrangement to be learned by neurons, ensuring that higher order spatial properties of the input stimuli are represented in the network.
- A convergent series of connections from a localized population of cells in preceding layers to each cell of the following layer, thus allowing the receptive field size of cells to increase through the visual processing areas or layers.
- A modified Hebb-like learning rule incorporating a temporal trace of each cell's previous activity, which, it is suggested, will enable the neurons to learn transform invariances.

The first two elements of Rolls' theory are used to constrain the general architecture of a network model, VisNet, of the processes just described that is intended to learn invariant representations of objects. The simulation results described in this paper using VisNet show that invariant representations can be learned by the architecture. It is moreover shown that successful learning depends crucially on the use of the modified Hebb rule. The

general architecture simulated in VisNet, and the way in which it allows natural images to be used as stimuli, has been chosen to enable some comparisons of neuronal responses in the network and in the brain to similar stimuli to be made.

### 5.1.1. The trace rule

The learning rule implemented in the VisNet simulations utilizes the spatio-temporal constraints placed upon the behavior of "real-world" objects to learn about natural object transformations. By presenting consistent sequences of transforming objects the cells in the network can learn to respond to the same object through all of its naturally transformed states, as described by Földiák (1991), Rolls (1992), Wallis et al. (1993), and Wallis and Rolls (1997). The learning rule incorporates a decaying trace of previous cell activity and is henceforth referred to simply as the "trace" learning rule. The learning paradigm we describe here is intended in principle to enable learning of any of the transforms tolerated by inferior temporal cortex neurons, including position, size, view, lighting, and spatial-frequency (Rolls, 1992, 2000, 2008b; Rolls and Deco, 2002).

To clarify the reasoning behind this point, consider the situation in which a single neuron is strongly activated by a stimulus forming part of a real-world object. The trace of this neuron's activation will then gradually decay over a time period in the order of 0.5 s. If, during this limited time window, the net is presented with a transformed version of the original stimulus then not only will the initially active afferent synapses modify onto the neuron, but so also will the synapses activated by the transformed version of this stimulus. In this way the cell will learn to respond to either appearance of the original stimulus. Making such associations works in practice because it is very likely that within short-time periods different aspects of the same object will be being inspected. The cell will not, however, tend to make spurious links across stimuli that are part of different objects because of the unlikelihood in the real-world of one object consistently following another.

Various biological bases for this temporal trace have been advanced as follows: [The precise mechanisms involved may alter the precise form of the trace rule which should be used. Földiák (1992) describes an alternative trace rule which models individual NMDA channels. Equally, a trace implemented by extended cell firing should be reflected in representing the trace as an external firing rate, rather than an internal signal.]

- The persistent firing of neurons for as long as 100–400 ms observed after presentations of stimuli for 16 ms (Rolls and Tovee, 1994) could provide a time window within which to associate subsequent images. Maintained activity may potentially be implemented by recurrent connections between as well as within cortical areas (Rolls and Treves, 1998; Rolls and Deco, 2002; Rolls, 2008b). [The prolonged firing of inferior temporal cortex neurons during memory delay periods of several seconds, and associative links reported to develop between stimuli presented several seconds apart (Miyashita, 1988) are on too long a time scale to be immediately relevant to the present theory. In fact, associations between visual events occurring several seconds apart would, under *normal* environmental conditions, be detrimental to the operation of a network of the type

described here, because they would probably arise from different objects. In contrast, the system described benefits from associations between visual events which occur close in time (typically within 1 s), as they are likely to be from the same object.]

- The binding period of glutamate in the NMDA channels, which may last for 100 ms or more, may implement a trace rule by producing a narrow time window over which the *average* activity at each presynaptic site affects learning (Hestrin et al., 1990; Földiák, 1992; Rhodes, 1992; Rolls, 1992; Spruston et al., 1995).
- Chemicals such as nitric oxide may be released during high neural activity and gradually decay in concentration over a short-time window during which learning could be enhanced (Montague et al., 1991; Földiák, 1992; Garthwaite, 2008).

The trace update rule used in the baseline simulations of VisNet (Wallis and Rolls, 1997) is equivalent to both Földiák's used in the context of translation invariance (Wallis et al., 1993) and to the earlier rule of Sutton and Barto (1981) explored in the context of modeling the temporal properties of classical conditioning, and can be summarized as follows:

$$\delta w_j = \alpha \bar{y}^\tau x_j \quad (3)$$

where

$$\bar{y}^\tau = (1 - \eta) y^\tau + \eta \bar{y}^{\tau-1} \quad (4)$$

and

$x_j$ :	$j$ th input to the neuron.	$y$ :	Output from the neuron.
$\bar{y}^\tau$ :	Trace value of the output of the neuron at time step $\tau$ .	$\alpha$ :	Learning rate. Annealed between unity and zero.
$w_j$ :	Synaptic weight between $j$ th input and the neuron.	$\eta$ :	Trace value. The optimal value varies with presentation sequence length.

To bound the growth of each neuron's synaptic weight vector,  $w_i$  for the  $i$ th neuron, its length is explicitly normalized (a method similarly employed by von der Malsburg (1973) which is commonly used in competitive networks (Rolls, 2008b). An alternative, more biologically relevant implementation, using a local weight bounding operation which utilizes a form of heterosynaptic long-term depression (Rolls, 2008b), has in part been explored using a version of the Oja (1982) rule (see Wallis and Rolls, 1997).

### 5.1.2. The network implemented in VisNet

The network itself is designed as a series of hierarchical, convergent, competitive networks, in accordance with the hypotheses advanced above. The actual network consists of a series of four layers, constructed such that the convergence of information from the most disparate parts of the network's input layer can potentially influence firing in a single neuron in the final layer – see **Figure 1**. This corresponds to the scheme described by many researchers (Rolls, 1992, 2008b; Van Essen et al., 1992) as present in the primate visual system – see **Figure 1**. The forward connections to a cell in one-layer are derived from a topologically related and confined region of the preceding layer. The choice of whether a

connection between neurons in adjacent layers exists or not is based upon a Gaussian distribution of connection probabilities which roll off radially from the focal point of connections for each neuron. (A minor extra constraint precludes the repeated connection of any pair of cells.) In particular, the forward connections to a cell in one layer come from a small region of the preceding layer defined by the radius in **Table 1** which will contain approximately 67% of the connections from the preceding layer. **Table 1** shows the dimensions for VisNetL, the system we are currently using (Perry et al., 2010), which is a (16×) larger version of the version of VisNet than used in most of our previous investigations, which utilized  $32 \times 32$  neurons per layer. **Figure 1** shows the general convergent network architecture used. Localization and limitation of connectivity in the network is intended to mimic cortical connectivity, partially because of the clear retention of retinal topology through regions of visual cortex. This architecture also encourages the gradual combination of features from layer to layer which has relevance to the binding problem, as described in Section 5.4.

Modeling topological constraints in connectivity leads to an issue concerning neurons at the edges of the network layers. In principle these neurons may either receive no input from beyond the edge of the preceding layer, or have their connections repeatedly sample neurons at the edge of the previous layer. In practice either solution is liable to introduce artificial weighting on the few active inputs at the edge and hence cause the edge to have unwanted influence over the development of the network as a whole. In the real brain such edge-effects would be naturally smoothed by the transition of the locus of cellular input from the fovea to the lower acuity periphery of the visual field. However, it poses a problem here because we are in effect only simulating the small high-acuity foveal portion of the visual field in our simulations. As an alternative to the former solutions Wallis and Rolls (1997) elected to form the connections into a toroid, such that connections wrap back onto the network from opposite sides. This wrapping happens at all four layers of the network, and in the way an image on the “retina” is mapped to the input filters. This solution has the advantage of making all of the boundaries effectively invisible to the network. Further, this procedure does not itself introduce problems into evaluation of the network for the problems set, as many of the critical comparisons in VisNet involve comparisons between a network with the same architecture trained with the trace rule, or with the Hebb rule, or not trained at all. In practice, it is shown below that only the network trained with the trace rule solves the problem of forming invariant representations.

**Table 1 | VisNet dimensions.**

	Dimensions	# Connections	Radius
Layer 4	$128 \times 128$	100	48
Layer 3	$128 \times 128$	100	36
Layer 2	$128 \times 128$	100	24
Layer 1	$128 \times 128$	272	24
Input layer	$256 \times 256 \times 32$	–	–

### 5.1.3. Competition and lateral inhibition

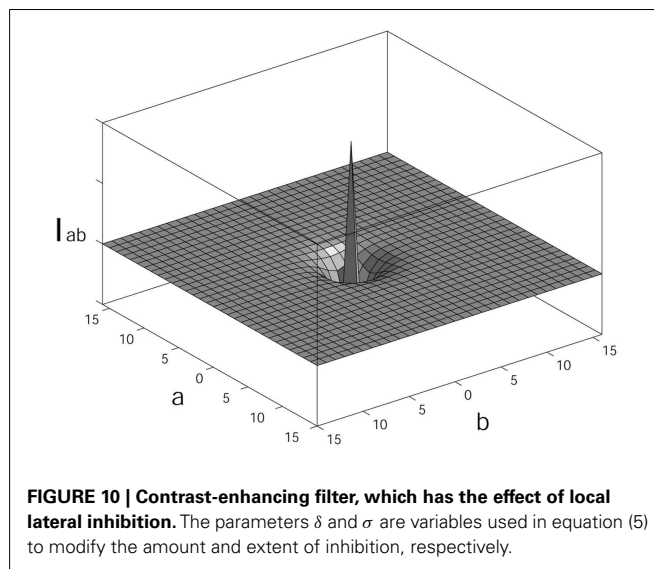
In order to act as a competitive network some form of mutual inhibition is required within each layer, which should help to ensure that all stimuli presented are evenly represented by the neurons in each layer. This is implemented in VisNet by a form of lateral inhibition. The idea behind the lateral inhibition, apart from this being a property of cortical architecture in the brain, was to prevent too many neurons that received inputs from a similar part of the preceding layer responding to the same activity patterns. The purpose of the lateral inhibition was to ensure that different receiving neurons coded for different inputs. This is important in reducing redundancy (Rolls, 2008b). The lateral inhibition is conceived as operating within a radius that was similar to that of the region within which a neuron received converging inputs from the preceding layer (because activity in one zone of topologically organized processing within a layer should not inhibit processing in another zone in the same layer, concerned perhaps with another part of the image). [Although the extent of the lateral inhibition actually investigated by Wallis and Rolls (1997) in VisNet operated over adjacent pixels, the lateral inhibition introduced by Rolls and Milward (2000) in what they named VisNet2 and which has been used in subsequent simulations operates over a larger region, set within a layer to approximately half of the radius of convergence from the preceding layer. Indeed, Rolls and Milward (2000) showed in a problem in which invariant representations over 49 locations were being used with a 17 face test set, that the best performance was with intermediate-range lateral inhibition, using the parameters for  $\sigma$  shown in Table 3. These values of  $\sigma$  set the lateral inhibition radius within a layer to be approximately half that of the spread of the excitatory connections from the preceding layer.]

The lateral inhibition and contrast enhancement just described are actually implemented in VisNet2 (Rolls and Milward, 2000) and VisNetL (Perry et al., 2010) in two stages, to produce filtering of the type illustrated in Figure 10. This lateral inhibition is implemented by convolving the activation of the neurons in a layer with a spatial filter,  $I$ , where  $\delta$  controls the contrast and  $\sigma$  controls the width, and  $a$  and  $b$  index the distance away from the center of the filter

$$I_{a,b} = \begin{cases} -\delta e^{-\frac{a^2+b^2}{\sigma^2}} & \text{if } a \neq 0 \text{ or } b \neq 0, \\ 1 - \sum_{a \neq 0, b \neq 0} I_{a,b} & \text{if } a = 0 \text{ and } b = 0. \end{cases} \quad (5)$$

This is a filter that leaves the average activity unchanged. A modified version of this filter designed as a difference of Gaussians with the same inhibition but shorter range local excitation is being tested to investigate whether the self-organizing maps that this promotes (Rolls, 2008b) helps the system to provide some continuity in the representations formed. The concept is that this may help the system to code efficiently for large numbers of untrained stimuli that fall between trained stimuli in similarity space.

The second stage involves contrast enhancement. In VisNet (Wallis and Rolls, 1997), this was implemented by raising the neuronal activations to a fixed power and normalizing the resulting firing within a layer to have an average firing rate equal to 1.0. In VisNet2 (Rolls and Milward, 2000) and in subsequent simulations



**FIGURE 10 | Contrast-enhancing filter, which has the effect of local lateral inhibition.** The parameters  $\delta$  and  $\sigma$  are variables used in equation (5) to modify the amount and extent of inhibition, respectively.

a more biologically plausible form of the activation function, a sigmoid, was used:

$$y = f^{\text{sigmoid}}(r) = \frac{1}{1 + e^{-2\beta(r-\alpha)}} \quad (6)$$

where  $r$  is the activation (or firing rate) of the neuron after the lateral inhibition,  $y$  is the firing rate after the contrast enhancement produced by the activation function, and  $\beta$  is the slope or gain and  $\alpha$  is the threshold or bias of the activation function. The sigmoid bounds the firing rate between 0 and 1 so global normalization is not required. The slope and threshold are held constant within each layer. The slope is constant throughout training, whereas the threshold is used to control the sparseness of firing rates within each layer. The (population) sparseness of the firing within a layer is defined (Rolls and Treves, 1998, 2011; Franco et al., 2007; Rolls, 2008b) as:

$$a = \frac{(\sum_i y_i / n)^2}{\sum_i y_i^2 / n} \quad (7)$$

where  $n$  is the number of neurons in the layer. To set the sparseness to a given value, e.g., 5%, the threshold is set to the value of the 95th percentile point of the activations within the layer. (Unless otherwise stated here, the neurons used the sigmoid activation function as just described.)

In most simulations with VisNet2 and later, the sigmoid activation function was used with parameters (selected after a number of optimization runs) as shown in Table 2.

In addition, the lateral inhibition parameters normally used in VisNet2 simulations are as shown in Table 3. (Where a power activation function was used in the simulations of Wallis and Rolls (1997), the power for layer 1 was 6, and for the other layers was 2.)

### 5.1.4. The input to VisNet

VisNet is provided with a set of input filters which can be applied to an image to produce inputs to the network which correspond

**Table 2 | Sigmoid parameters for the runs with 25 locations by Rolls and Milward, 2000).**

Layer	1	2	3	4
Percentile	99.2	98	88	91
Slope $\beta$	190	40	75	26

**Table 3 | Lateral inhibition parameters for the 25-location runs.**

Layer	1	2	3	4
Radius, $\sigma$	1.38	2.7	4.0	6.0
Contrast, $\delta$	1.5	1.5	1.6	1.4

to those provided by simple cells in visual cortical area 1 (V1). The purpose of this is to enable within VisNet the more complicated response properties of cells between V1 and the inferior temporal cortex (IT) to be investigated, using as inputs natural stimuli such as those that could be applied to the retina of the real visual system. This is to facilitate comparisons between the activity of neurons in VisNet and those in the real visual system, to the same stimuli. In VisNet no attempt is made to train the response properties of simple cells, but instead we start with a defined series of filters to perform fixed feature extraction to a level equivalent to that of simple cells in V1, as have other researchers in the field (Fukushima, 1980; Buhmann et al., 1991; Hummel and Biederman, 1992), because we wish to simulate the more complicated response properties of cells between V1 and the inferior temporal cortex (IT). The elongated orientation-tuned input filters used accord with the general tuning profiles of simple cells in V1 (Hawken and Parker, 1987) and in earlier versions of VisNet were computed by weighting the difference of two Gaussians by a third orthogonal Gaussian as described in detail elsewhere (Wallis and Rolls, 1997; Rolls and Milward, 2000; Perry et al., 2010). Each individual filter is tuned to spatial-frequency (0.0039–0.5 cycles/pixel over eight octaves); orientation (0–135° in steps of 45°); and sign ( $\pm 1$ ). Of the 272 layer 1 connections, the number to each group in VisNetL is as shown in Table 4. In VisNet2 (Rolls and Milward, 2000; used for most VisNet simulations) only even symmetric – “bar detecting” – filter shapes are used, which take the form of a Gaussian shape along the axis of orientation tuning for the filter, and a difference of Gaussians along the perpendicular axis.

This filter is referred to as an oriented difference of Gaussians, or DOG filter. Any zero D.C. filter can of course produce a negative as well as positive output, which would mean that this simulation of a simple cell would permit negative as well as positive firing. In contrast to some other models the response of each filter is zero thresholded and the negative results used to form a separate anti-phase input to the network. The filter outputs are also normalized across scales to compensate for the low-frequency bias in the images of natural objects.

However, Gabor filters have also been tested, also produce good results with VisNet (Deco and Rolls, 2004), and are what we implement at present in VisNetL. Following Daugman (1988) the

receptive fields of the simple cell-like input neurons are modeled by 2D-Gabor functions. The Gabor receptive fields have five degrees of freedom given essentially by the product of an elliptical Gaussian and a complex plane wave. The first two degrees of freedom are the 2D-locations of the receptive field's center; the third is the size of the receptive field; the fourth is the orientation of the boundaries separating excitatory and inhibitory regions; and the fifth is the symmetry. This fifth degree of freedom is given in the standard Gabor transform by the real and imaginary part, i.e., by the phase of the complex function representing it, whereas in a biological context this can be done by combining pairs of neurons with even and odd receptive fields. This design is supported by the experimental work of Pollen and Ronner (1981), who found simple cells in quadrature-phase pairs. Even more, Daugman (1988) proposed that an ensemble of simple cells is best modeled as a family of 2D-Gabor wavelets sampling the frequency domain in a log-polar manner as a function of eccentricity. Experimental neurophysiological evidence constrains the relation between the free parameters that define a 2D-Gabor receptive field (De Valois and De Valois, 1988). There are three constraints fixing the relation between the width, height, orientation, and spatial-frequency (Lee, 1996). The first constraint posits that the aspect ratio of the elliptical Gaussian envelope is 2:1. The second constraint postulates that the plane wave tends to have its propagating direction along the short axis of the elliptical Gaussian. The third constraint assumes that the half-amplitude bandwidth of the frequency response is about 1–1.5 octaves along the optimal orientation. Further, we assume that the mean is zero in order to have an admissible wavelet basis (Lee, 1996).

In more detail, the Gabor filters are constructed as follows (Deco and Rolls, 2004). We consider a pixelized gray-scale image given by a  $N \times N$  matrix  $\Gamma_{ij}^{\text{orig}}$ . The subindices  $ij$  denote the spatial position of the pixel. Each pixel value is given a gray-level brightness value coded in a scale between 0 (black) and 255 (white). The first step in the pre-processing consists of removing the DC component of the image (i.e., the mean value of the gray-scale intensity of the pixels). (The equivalent in the brain is the low-pass filtering performed by the retinal ganglion cells and lateral geniculate cells. The visual representation in the LGN is essentially a contrast-invariant pixel representation of the image, i.e., each neuron encodes the relative brightness value at one location in visual space referred to the mean value of the image brightness.) We denote this contrast-invariant LGN representation by the  $N \times N$  matrix  $\Gamma_{ij}$  defined by the equation

$$\Gamma_{ij} = \Gamma_{ij}^{\text{orig}} - \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N \Gamma_{ij}^{\text{orig}}. \quad (8)$$

Feed-forward connections to a layer of V1 neurons perform the extraction of simple features like bars at different locations, orientations and sizes. Realistic receptive fields for V1 neurons that extract these simple features can be represented by 2D-Gabor wavelets. Lee (1996) derived a family of discretized 2D-Gabor wavelets that satisfy the wavelet theory and the neurophysiological

constraints for simple cells mentioned above. They are given by an expression of the form

$$G_{pqkl}(x, y) = a^{-k} \Psi_{\Theta_l} \left( a^{-k} (x - 2p), a^{-k} (y - 2q) \right) \quad (9)$$

where

$$\Psi_{\Theta_l} = \Psi \left( x \cos(l\Theta_0) + y \sin(l\Theta_0), -x \sin(l\Theta_0) + y \cos(l\Theta_0) \right), \quad (10)$$

and the mother wavelet is given by

$$\Psi(x, y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{8}(4x^2 + y^2)} \left[ e^{ikx} - e^{-\frac{\kappa^2}{2}} \right]. \quad (11)$$

In the above equations  $\Theta_0 = \pi/L$  denotes the step size of each angular rotation;  $l$  the index of rotation corresponding to the preferred orientation  $\Theta_l = l\pi/L$ ;  $k$  denotes the octave; and the indices  $pq$  the position of the receptive field center at  $c_x = p$  and  $c_y = q$ . In this form, the receptive fields at all levels cover the spatial domain in the same way, i.e., by always overlapping the receptive fields in the same fashion. In the model we use  $a = 2$ ,  $b = 1$ , and  $\kappa = \pi$  corresponding to a spatial-frequency bandwidth of one octave. We now use in VisNetL both symmetric and asymmetric filters (as both are present in V1 Ringach, 2002); with the angular spacing between the different orientations set to  $45^\circ$ ; and with 8 filter frequencies spaced one octave apart starting with 0.5 cycles per pixel,

and with the sampling from the spatial frequencies set as shown in Table 4.

Cells of layer 1 receive a topologically consistent, localized, random selection of the filter responses in the input layer, under the constraint that each cell samples every filter spatial-frequency and receives a constant number of inputs. Figure 11 shows pictorially the general filter sampling paradigm.

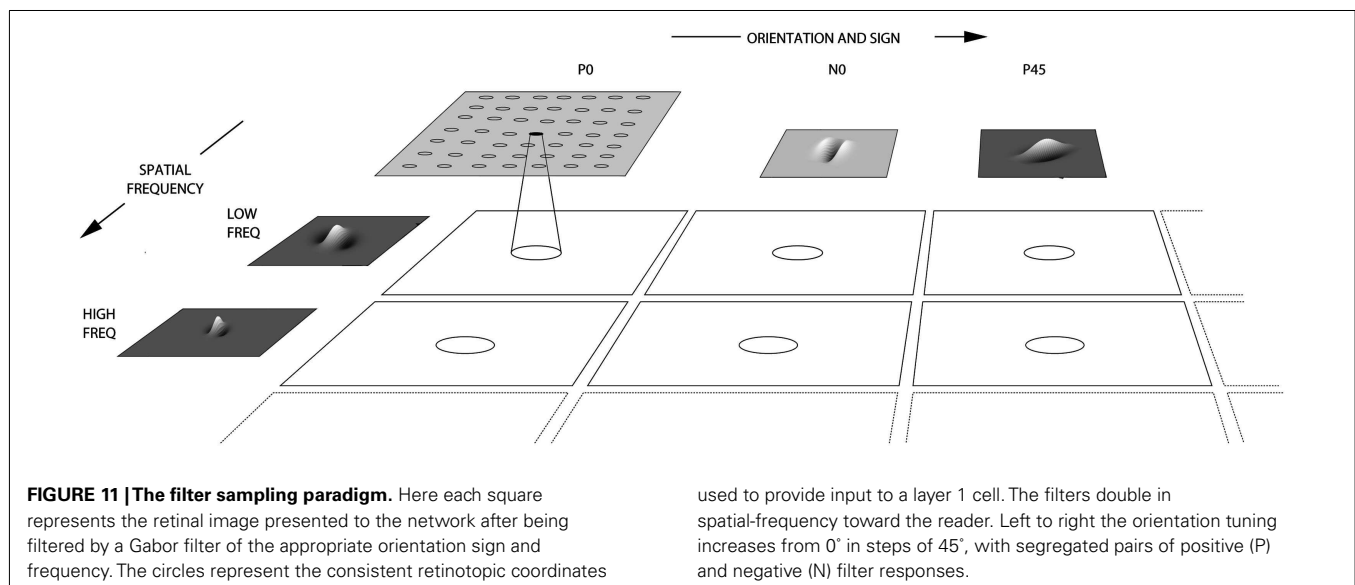
### 5.1.5. Measures for network performance

A neuron can be said to have learnt an invariant representation if it discriminates one set of stimuli from another set, across all transformations. For example, a neuron's response is translation-invariant if its response to one set of stimuli irrespective of presentation is consistently higher than for all other stimuli irrespective of presentation location. Note that we state "set of stimuli" since neurons in the inferior temporal cortex are not generally selective for a single stimulus but rather a subpopulation of stimuli (Baylis et al., 1985; Abbott et al., 1996; Rolls et al., 1997b; Rolls and Treves, 1998, 2011; Rolls and Deco, 2002; Franco et al., 2007; Rolls, 2007b, 2008b). The measure of network performance used in VisNet1 (Wallis and Rolls, 1997), the "Fisher metric" (referred to in some figure labels as the Discrimination Factor), reflects how well a neuron discriminates between stimuli, compared to how well it discriminates between different locations (or more generally the images used rather than the objects, each of which is represented by a set of images, over which invariant stimulus or object representations must be learned). The Fisher measure is very similar to taking the ratio of the two F values in a two-way ANOVA, where

**Table 4 | VisNet layer 1 connectivity.**

Frequency	0.5	0.25	0.125	0.0625	0.03125	0.0156	0.0078	0.0039
# Connections	180	45	12	7	7	7	7	7

The frequency is in cycles per pixel.



one factor is the stimulus shown, and the other factor is the position in which a stimulus is shown. The measure takes a value greater than 1.0 if a neuron has more different responses to the stimuli than to the locations. That is, values greater than 1 indicate invariant representations when this measure is used in the following figures. Further details of how the measure is calculated are given by Wallis and Rolls (1997).

Measures of network performance based on information theory and similar to those used in the analysis of the firing of real neurons in the brain (Rolls, 2008b; Rolls and Treves, 2011) were introduced by Rolls and Milward (2000) for VisNet2, and are used in later papers. A single cell information measure was introduced which is the maximum amount of information the cell has about any one stimulus/object independently of which transform (e.g., position on the retina) is shown. Because the competitive algorithm used in VisNet tends to produce local representations (in which single cells become tuned to one stimulus or object), this information measure can approach  $\log_2 N_s$  bits, where  $N_s$  is the number of different stimuli. Indeed, it is an advantage of this measure that it has a defined maximal value, which enables how well the network is performing to be quantified. Rolls and Milward (2000) showed that the Fisher and single cell information measures were highly correlated, and given the advantage just noted of the information measure, it was adopted in Rolls and Milward (2000) and subsequent papers. Rolls and Milward (2000) also introduced a multiple cell information measure, which has the advantage that it provides a measure of whether all stimuli are encoded by different neurons in the network. Again, a high value of this measure indicates good performance.

For completeness, we provide further specification of the two information theoretic measures, which are described in detail by Rolls and Milward (2000), (see Rolls, 2008b) Rolls and Treves (2011) for an introduction to the concepts). The measures assess the extent to which either a single cell, or a population of cells, responds to the same stimulus invariantly with respect to its location, yet responds differently to different stimuli. The measures effectively show what one learns about which stimulus was presented from a single presentation of the stimulus at any randomly chosen location. Results for top (4th) layer cells are shown. High information measures thus show that cells fire similarly to the different transforms of a given stimulus (object), and differently to the other stimuli. The single cell stimulus-specific information,  $I(s, R)$ , is the amount of information the set of responses,  $R$ , has about a specific stimulus,  $s$  (see Rolls et al., 1997c; Rolls and Milward, 2000).  $I(s, R)$  is given by

$$I(s, R) = \sum_{r \in R} P(r|s) \log_2 \frac{P(r|s)}{P(r)} \quad (12)$$

where  $r$  is an individual response from the set of responses  $R$  of the neuron. For each cell the performance measure used was the maximum amount of information a cell conveyed about any one stimulus. This (rather than the mutual information,  $I(S, R)$  where  $S$  is the whole set of stimuli  $s$ ), is appropriate for a competitive network in which the cells tend to become tuned to one stimulus. ( $I(s, R)$  has more recently been called the stimulus-specific

surprise (DeWeese and Meister, 1999; Rolls and Treves, 2011). Its average across stimuli is the mutual information  $I(S, R)$ .)

If all the output cells of VisNet learned to respond to the same stimulus, then the information about the set of stimuli  $S$  would be very poor, and would not reach its maximal value of  $\log_2$  of the number of stimuli (in bits). The second measure that is used here is the information provided by a set of cells about the stimulus set, using the procedures described by Rolls et al. (1997b) and Rolls and Milward (2000). The multiple cell information is the mutual information between the whole set of stimuli  $S$  and of responses  $R$  calculated using a decoding procedure in which the stimulus  $s'$  that gave rise to the particular firing rate response vector on each trial is estimated. (The decoding step is needed because the high dimensionality of the response space would lead to an inaccurate estimate of the information if the responses were used directly, as described by Rolls et al. (1997b) and Rolls and Treves (1998).) A probability table is then constructed of the real stimuli  $s$  and the decoded stimuli  $s'$ . From this probability table, the mutual information between the set of actual stimuli  $S$  and the decoded estimates  $S'$  is calculated as

$$I(S, S') = \sum_{s, s'} P(s, s') \log_2 \frac{P(s, s')}{P(s) P(s')} \quad (13)$$

This was calculated for the subset of cells which had as single cells the most information about which stimulus was shown. In particular, in Rolls and Milward (2000) and subsequent papers, the multiple cell information was calculated from the first five cells for each stimulus that had maximal single cell information about that stimulus, that is from a population of 35 cells if there were seven stimuli (each of which might have been shown in, for example, 9 or 25 positions on the retina).

## 5.2. INITIAL EXPERIMENTS WITH VisNet

Having established a network model, Wallis and Rolls (1997) following a first report by Wallis et al. (1993) described four experiments in which the theory of how invariant representations could be formed was tested using a variety of stimuli undergoing a number of natural transformations. In each case the network produced neurons in the final layer whose responses were largely invariant across a transformation and highly discriminating between stimuli or sets of stimuli. A summary showing how the network performed is presented here, with much more evidence of the factors that influence the network's performance described elsewhere (Wallis and Rolls, 1997; Rolls, 2008b).

### 5.2.1. "T," "L," and "+" as stimuli: learning translation invariance

One of the classical properties of inferior temporal cortex face cells is their invariant response to face stimuli translated across the visual field (Tovee et al., 1994). In this first experiment, the learning of translation-invariant representations by VisNet was investigated.

In order to test the network a set of three stimuli, based upon probable 3D edge cues – consisting of a "T," "L," and "+" shape – was constructed. Chakravarty (1979) describes the application of these shapes as cues for the 3D interpretation of edge junctions, and Tanaka et al. (1991) have demonstrated the existence of cells



responsive to such stimuli in IT.) These stimuli were chosen partly because of their significance as form cues, but on a more practical note because they each contain the same fundamental features – namely a horizontal bar conjoined with a vertical bar. In practice this means that the oriented simple cell filters of the input layer cannot distinguish these stimuli on the basis of which features are present. As a consequence of this, the representation of the stimuli received by the network is non-orthogonal and hence considerably more difficult to classify than was the case in earlier experiments involving the trace rule described by Földiák (1991). The expectation is that layer 1 neurons would learn to respond to spatially selective combinations of the basic features thereby helping to distinguish these non-orthogonal stimuli. The trajectory followed by each stimulus consisted of sweeping left to right horizontally across three locations in the top row, and then sweeping back, right to left across the middle row, before returning to the right hand side across the bottom row – tracing out a “Z” shape path across the retina. Unless stated otherwise this pattern of nine presentation locations was adopted in all image translation experiments described by Wallis and Rolls (1997).

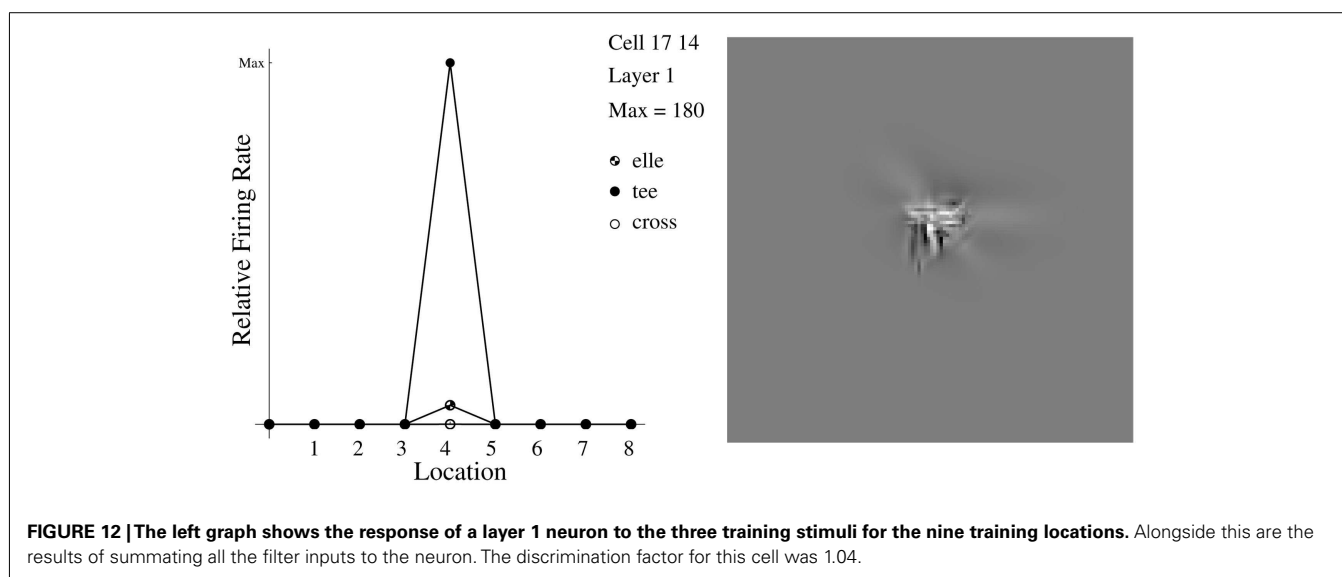
Training was carried out by permutatively presenting all stimuli in each location a total of 800 times. The sequence described above was followed for each stimulus, with the sequence start point and direction of sweep being chosen at random for each of the 800 training trials.

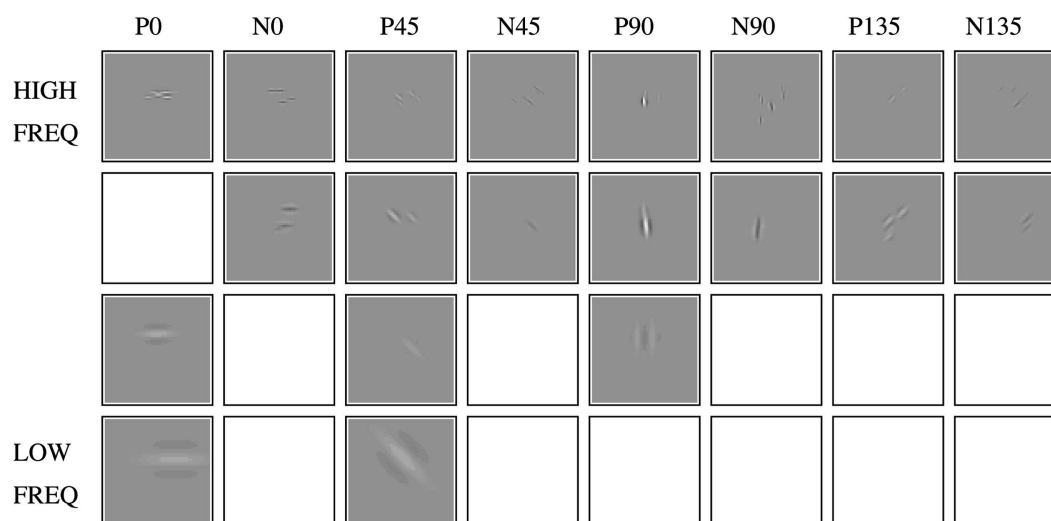
Figures 12 and 13 shows the response after training of a first layer neuron selective for the “T” stimulus. The weighted sum of all filter inputs reveals the combination of horizontally and vertically tuned filters in identifying the stimulus. In this case many connections to the lower frequency filters have been reduced to zero by the learning process, except at the relevant orientations. This contrasts strongly with the random wiring present before training (Wallis and Rolls, 1997; Rolls, 2008b). It is important that neurons at early stages of feature hierarchy networks respond to combinations of features in defined relative spatial positions, before invariance is built into the system, as this is part of the way that the binding problem is solved, as described in more detail in Section 5.4 and by

Elliffe et al. (2002). The feature combination tuning is illustrated by the VisNet layer 1 neuron shown in Figures 12 and 13.

The results for layer 4 neurons are illustrated in Figure 14. By this stage translation-invariant, stimulus-identifying, cells have emerged. The response profiles confirm the high level of neural selectivity for a particular stimulus irrespective of location. Neurons in layers 2 and 3 of VisNet had intermediate-levels of translation invariance to those illustrated for layer 1 and layer 4. The gradual increase in the invariance that the tolerance to shifts of the preferred stimulus gradually builds up through the layers.

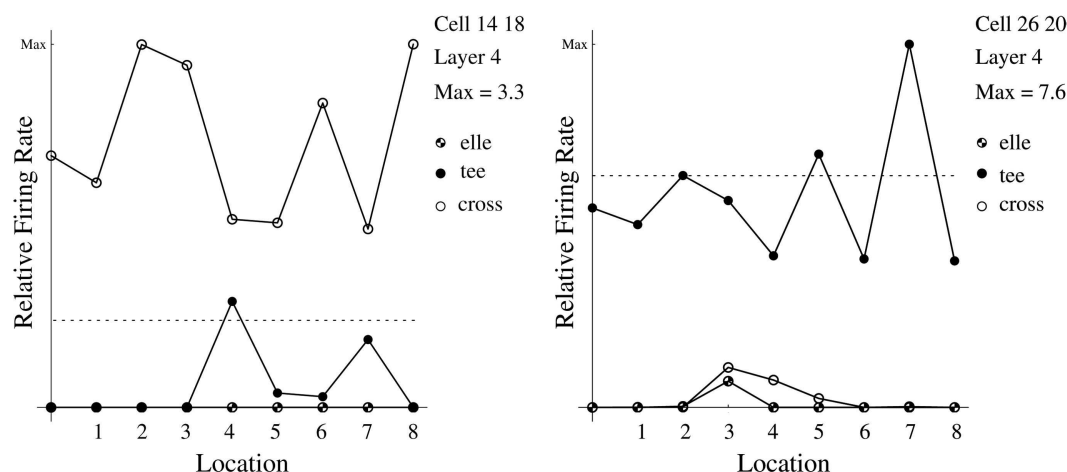
The trace used in VisNet enables successive features that, based on the natural statistics of the visual input, are likely to be from the same object or feature complex to be associated together. For good performance, the temporal trace needs to be sufficiently long that it covers the period in which features seen by a particular neuron in the hierarchy are likely to come from the same object. On the other hand, the trace should not be so long that it produces associations between features that are parts of different objects, seen when, for example, the eyes move to another object. One possibility is to reset the trace during saccades between different objects. If explicit trace resetting is not implemented, then the trace should, to optimize the compromise implied by the above, lead to strong associations between temporally close stimuli, and increasingly weaker associations between temporally more distant stimuli. In fact, the trace implemented in VisNet has an exponential decay, and it has been shown that this form is optimal in the situation where the exact duration over which the same object is being viewed varies, and where the natural statistics of the visual input happen also to show a decreasing probability that the same object is being viewed as the time period in question increases (Wallis and Baddeley, 1997). Moreover, performance can be enhanced if the duration of the trace does at the same time approximately match the period over which the input stimuli are likely to come from the same object or feature complex (Wallis and Rolls, 1997; Rolls, 2008b). Nevertheless, good performance can be obtained in conditions under which the trace rule allows associations to be





**FIGURE 13 | The connections to a single cell in layer 1 of VisNet from the filters after training in the T, L, and + stimulus set, represented by plotting the receptive fields of every input layer cell connected to the particular layer 1 cell.** Separate input layer cells have activity that represents a positive (P) or negative (N) output from the bank of filters which have different orientations in degrees (the columns) and different spatial frequencies (the rows). Here the overall

receptive field of the layer 1 cell is centered just below the center-point of the retina. The connection scheme allows for relatively fewer connections to lower frequency cells than to high-frequency cells in order to cover a similar region of the input at each frequency. The blank squares indicate that no connection exists between the layer 1 cell chosen and the filters of that particular orientation, sign, and spatial-frequency.



**FIGURE 14 | Response profiles for two fourth layer neurons – discrimination factors 4.07 and 3.62 – in the L, T, and + experiment.**

formed only between successive items in the visual stream (Rolls and Milward, 2000; Rolls and Stringer, 2001).

It is also the case that the optimal value of  $\eta$  in the trace rule is likely to be different for different layers of VisNet, and for cortical processing in the “what” visual stream. For early layers of the system, small movements of the eyes might lead to different feature combinations providing the input to cells (which at early stages have small receptive fields), and a short duration of the trace would be optimal. However, these small eye movements might be around the same object, and later layers of the architecture would benefit from being able to associate together their inputs over longer

times, in order to learn about the larger scale properties that characterize individual objects, including, for example, different views of objects observed as an object turns or is turned. Thus the suggestion is made that the temporal trace could be effectively longer at later stages (e.g., inferior temporal visual cortex) compared to early stages (e.g., V2 and V4) of processing in the visual system. In addition, as will be shown in Section 5.4, it is important to form feature combinations with high-spatial precision before invariance learning supported by a temporal trace starts, in order that the feature combinations and not the individual features have invariant representations. This leads to the suggestion that the trace rule

should either not operate, or be short, at early stages of cortical visual processing such as V1. This is reflected in the operation of VisNet2, which does not use a temporal trace in layer 1 (Rolls and Milward, 2000).

### 5.2.2. Faces as stimuli: translation invariance

The aim of the next set of experiments described by Wallis and Rolls (1997) was to start to address the issues of how the network operates when invariant representations must be learned for a larger number of stimuli, and whether the network can learn when much more complicated, real biological stimuli, faces, are used.

**Figure 15** contrasts the measure of invariance, or discrimination factor, achieved by cells in the four layers, averaged over five separate runs of the network (Wallis and Rolls, 1997; Rolls, 2008b). Translation invariance clearly increases through the layers, as expected.

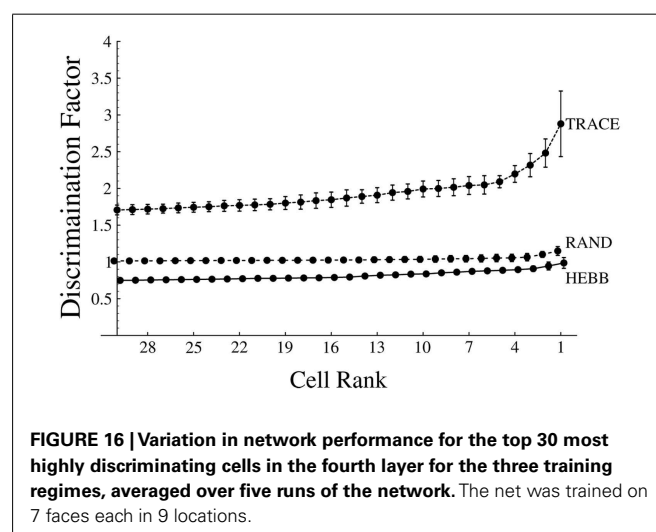
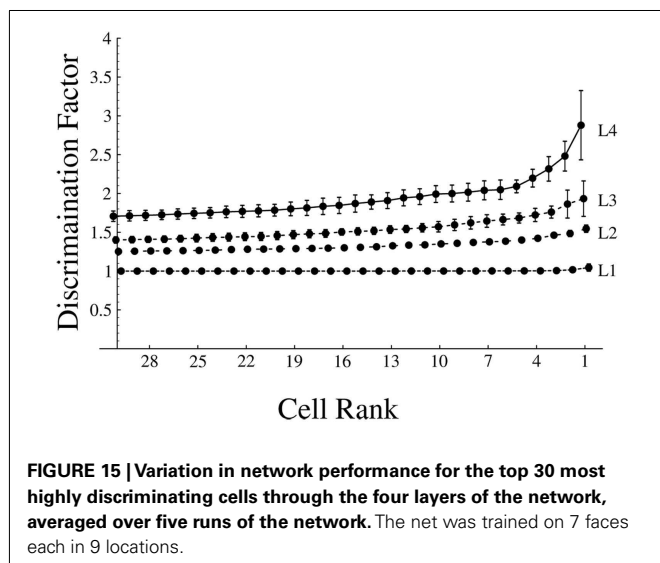
Having established that invariant cells have emerged in the final layer, we now consider the role of the trace rule, by assessing the network tested under two new conditions. Firstly, the performance of the network was measured before learning occurs, that is with its initially random connection weights. Secondly, the network was trained with  $\eta$  in the trace rule set to 0, which causes learning to proceed in a traceless, standard Hebbian, fashion. (Hebbian learning is purely associative Rolls, 2008b.) **Figure 16** shows the results under the three training conditions. The results show that the trace rule is the decisive factor in establishing the invariant responses in the layer 4 neurons. It is interesting to note that the Hebbian learning results are actually *worse* than those achieved by chance in the untrained net. In general, with Hebbian learning, the most highly discriminating cells barely rate higher than 1. This value of discrimination corresponds to the case in which a cell responds to only one stimulus and in only one location. The poor performance with the Hebb rule comes as a direct consequence of the presentation paradigm being employed. If we consider an image as representing a vector in multidimensional space, a particular image in the top left-hand corner of the input retina will tend

to look more like any other image in that same location than the same image presented elsewhere. A simple competitive network using just Hebbian learning will thus tend to categorize images by *where* they are rather than what they are – the exact opposite of what the net was intended to learn. This comparison thus indicates that a small memory trace acting in the standard Hebbian learning paradigm can radically alter the normal vector averaging, image classification, performed by a Hebbian-based competitive network.

In order to check that there was an invariant representation in layer 4 of VisNet that could be read by a receiving population of neurons, a fifth layer was added to the net which fully sampled the fourth layer cells. This layer was in turn trained in a supervised manner using gradient descent or with a Hebbian associative learning rule. (Wallis and Rolls, 1997) showed that the object classification performed by the layer 5 network was better if the network had been trained with the trace rule than when it was untrained or was trained with a Hebb rule.

### 5.2.3. Faces as stimuli: view-invariance

Given that the network had been shown to be able to operate usefully with a more difficult translation invariance problem, we next addressed the question of whether the network can solve other types of transform invariance, as we had intended. The next experiment addressed this question, by training the network on the problem of 3D stimulus rotation, which produces non-isomorphic transforms, to determine whether the network can build a view-invariant categorization of the stimuli (Wallis and Rolls, 1997). The trace rule learning paradigm should, in conjunction with the architecture described here, prove capable of learning any of the transforms tolerated by IT neurons, so long as each stimulus is presented in short sequences during which the transformation occurs and can be learned. This experiment continued with the use of faces but now presented them centrally in the retina in a sequence of different views of a face (Wallis and Rolls, 1997; Rolls, 2008b). The faces were again smoothed at the edges to erase the harsh image boundaries, and the D.C. term was removed. During the 800 epochs of learning, each stimulus was chosen at random, and



a sequence of preset views of it was shown, rotating the face either to the left or to the right.

Although the actual number of images being presented is smaller, some 21 views in all, there is good reason to think that this problem may be harder to solve than the previous translation experiments. This is simply due to the fact that all 21 views exactly overlap with one another. The net was indeed able to solve the invariance problem, with examples of invariant layer 4 neuron response profiles appearing in **Figure 17**.

Further analyses confirmed the good performance on view-invariance learning (Wallis and Rolls, 1997; Rolls, 2008b).

### 5.3. DIFFERENT FORMS OF THE TRACE-LEARNING RULE, AND THEIR RELATION TO ERROR CORRECTION AND TEMPORAL DIFFERENCE LEARNING

The original trace-learning rule used in the simulations of Wallis and Rolls (1997) took the form

$$\delta w_j = \alpha \bar{y}^\tau x_j^\tau \quad (14)$$

where the trace  $\bar{y}^\tau$  is updated according to

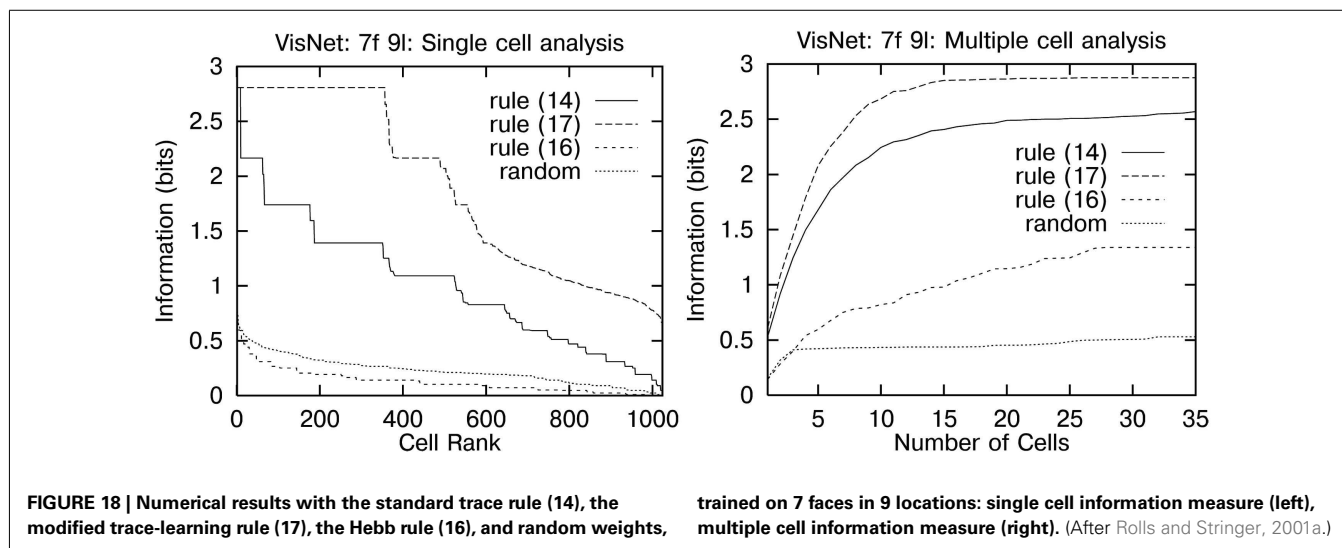
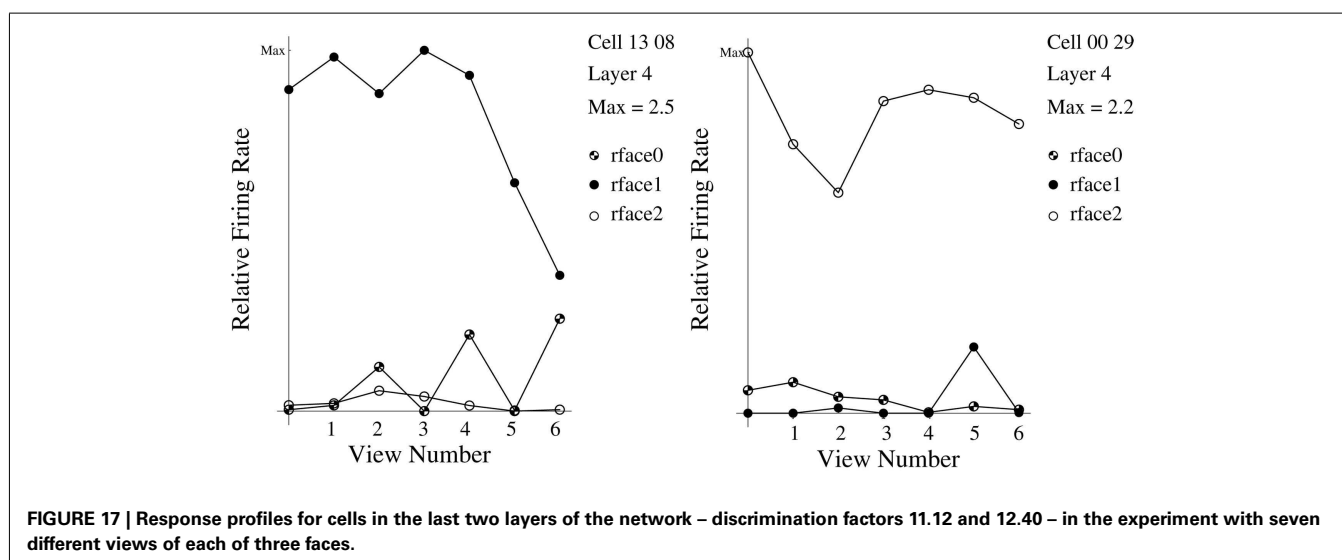
$$\bar{y}^\tau = (1 - \eta) y^\tau + \eta \bar{y}^{\tau-1}. \quad (15)$$

The parameter  $\eta \in [0, 1]$  controls the relative contributions to the trace  $\bar{y}^\tau$  from the instantaneous firing rate  $y^\tau$  and the trace at the previous time step  $\bar{y}^{\tau-1}$ , where for  $\eta = 0$  we have  $\bar{y}^\tau = y^\tau$  and equation (14) becomes the standard Hebb rule

$$\delta w_j = \alpha y^\tau x_j^\tau. \quad (16)$$

At the start of a series of investigations of different forms of the trace-learning rule (Rolls and Milward, 2000) demonstrated that VisNet's performance could be greatly enhanced (see **Figure 18**) with a modified Hebbian trace-learning rule (equation (17)) that incorporated a trace of activity from the preceding time steps, with no contribution from the activity being produced by the stimulus at the current time step. This rule took the form

$$\delta w_j = \alpha \bar{y}^{\tau-1} x_j^\tau. \quad (17)$$



The trace shown in equation (17) is in the post-synaptic term, and similar effects were found if the trace was in the presynaptic term, or in both the pre- and the post-synaptic terms. The crucial difference from the earlier rule (see equation (14)) was that the trace should be calculated up to only the preceding timestep, with no contribution to the trace from the firing on the current trial to the current stimulus. How might this be understood?

One way to understand this is to note that the trace rule is trying to set up the synaptic weight on trial  $\tau$  based on whether the neuron, based on its previous history, is responding to that stimulus (in other transforms, e.g., position). Use of the trace rule at  $\tau - 1$  does this that is it takes into account the firing of the neuron on previous trials, with no contribution from the firing being produced by the stimulus on the current trial. On the other hand, use of the trace at time  $\tau$  in the update takes into account the current firing of the neuron to the stimulus in that particular position, which is not a good estimate of whether that neuron should be allocated to invariantly represent that stimulus. Effectively, using the trace at time  $\tau$  introduces a Hebbian element into the update, which tends to build position-encoded analyzers, rather than stimulus-encoded analyzers. (The argument has been phrased for a system learning translation invariance, but applies to the learning of all types of invariance.) A particular advantage of using the trace at  $\tau - 1$  is that the trace will then on different occasions (due to the randomness in the location sequences used) reflect previous histories with different sets of positions, enabling the learning of the neuron to be based on evidence from the stimulus present in many different positions. Using a term from the current firing in the trace (i.e., the trace calculated at time  $\tau$ ) results in this desirable effect always having an undesirable element from the current firing of the neuron to the stimulus in its current position.

### 5.3.1. The modified Hebbian trace rule and its relation to error correction

The rule of equation (17) corrects the weights using a post-synaptic trace obtained from the previous firing (produced by other transforms of the same stimulus), with no contribution to the trace from the current post-synaptic firing (produced by the current transform of the stimulus). Indeed, insofar as the current firing  $y^\tau$  is not the same as  $\bar{y}^{\tau-1}$ , this difference can be thought of as an error. This leads to a conceptualization of using the difference between the current firing and the preceding trace as an error correction term, as noted in the context of modeling the temporal properties of classical conditioning by Sutton and Barto (1981), and developed next in the context of invariance learning (see Rolls and Stringer, 2001).

First, we re-express the rule of equation (17) in an alternative form as follows. Suppose we are at timestep  $\tau$  and have just calculated a neuronal firing rate  $y^\tau$  and the corresponding trace  $\bar{y}^\tau$  from the trace update equation (15). If we assume  $\eta \in (0, 1)$ , then rearranging equation (15) gives

$$\bar{y}^{\tau-1} = \frac{1}{\eta} (\bar{y}^\tau - (1 - \eta) y^\tau), \quad (18)$$

and substituting equation (18) into equation (17) gives

$$\begin{aligned} \delta w_j &= \alpha \frac{1}{\eta} (\bar{y}^\tau - (1 - \eta) y^\tau) x_j^\tau \\ &= \alpha \frac{1 - \eta}{\eta} \left( \frac{1}{1 - \eta} \bar{y}^\tau - y^\tau \right) x_j^\tau \\ &= \hat{\alpha} (\hat{\beta} \bar{y}^\tau - y^\tau) x_j^\tau \end{aligned} \quad (19)$$

where  $\hat{\alpha} = \alpha \frac{1 - \eta}{\eta}$  and  $\hat{\beta} = \frac{1}{1 - \eta}$ . The modified Hebbian trace-learning rule (17) is thus equivalent to equation (19) which is in the general form of an error correction rule (Hertz et al., 1991). That is, rule (19) involves the subtraction of the current firing rate  $y^\tau$  from a target value, in this case  $\hat{\beta} \bar{y}^\tau$ .

Although above we have referred to rule (17) as a modified Hebbian rule, we note that it is only associative in the sense of associating *previous* cell firing with the current cell inputs. In the next section we continue to explore the error correction paradigm, examining five alternative examples of this sort of learning rule.

### 5.3.2. Five forms of error correction learning rule

Error correction learning rules are derived from gradient descent minimization (Hertz et al., 1991), and continually compare the current neuronal output to a target value  $t$  and adjust the synaptic weights according to the following equation at a particular timestep  $\tau$

$$\delta w_j = \alpha (t - y^\tau) x_j^\tau. \quad (20)$$

In this usual form of gradient descent by error correction, the target  $t$  is fixed. However, in keeping with our aim of encouraging neurons to respond similarly to images that occur close together in time it seems reasonable to set the target at a particular timestep,  $t^\tau$ , to be some function of cell activity occurring close in time, because encouraging neurons to respond to temporal classes will tend to make them respond to the different variants of a given stimulus (Földiák, 1991; Rolls, 1992; Wallis and Rolls, 1997). For this reason, Rolls and Stringer (2001) explored a range of error correction rules where the targets  $t^\tau$  are based on the trace of neuronal activity calculated according to equation (15). We note that although the target is not a fixed value as in standard error correction learning, nevertheless the new learning rules perform gradient descent on each timestep, as elaborated below. Although the target may be varying early on in learning, as learning proceeds the target is expected to become more and more constant, as neurons settle to respond invariantly to particular stimuli. The first set of five error correction rules we discuss are as follows.

$$\delta w_j = \alpha (\beta \bar{y}^{\tau-1} - y^\tau) x_j^\tau, \quad (21)$$

$$\delta w_j = \alpha (\beta y^{\tau-1} - y^\tau) x_j^\tau, \quad (22)$$

$$\delta w_j = \alpha (\beta \bar{y}^\tau - y^\tau) x_j^\tau, \quad (23)$$

$$\delta w_j = \alpha (\beta \bar{y}^{\tau+1} - y^\tau) x_j^\tau, \quad (24)$$

$$\delta w_j = \alpha (\beta y^{\tau+1} - y^\tau) x_j^\tau, \quad (25)$$

where updates (21–23) are performed at timestep  $\tau$ , and updates (24) and (25) are performed at timestep  $\tau + 1$ . (The reason for adopting this convention is that the basic form of the error correction rule (20) is kept, with the five different rules simply replacing

the term  $t$ .) It may be readily seen that equations (22) and (25) are special cases of equations (21) and (24), respectively, with  $\eta = 0$ .

These rules are all similar except for their targets  $t^\tau$ , which are all functions of a temporally nearby value of cell activity. In particular, rule (23) is directly related to rule (19), but is more general in that the parameter  $\hat{\beta} = \frac{1}{1-\eta}$  is replaced by an unconstrained parameter  $\beta$ . In addition, we also note that rule (21) is closely related to a rule developed in Peng et al. (1998) for view-invariance learning. The above five error correction rules are biologically plausible in that the targets  $t^\tau$  are all local cell variables (see Rolls and Treves, 1998 and Rolls, 2008b). In particular, rule (23) uses the trace  $\bar{y}^\tau$  from the current time level  $\tau$ , and rules (22) and (25) do not need exponential trace values  $\bar{y}$ , instead relying only on the instantaneous firing rates at the current and immediately preceding timesteps. However, all five error correction rules involve decrementing of synaptic weights according to an error which is calculated by subtracting the current activity from a target.

Numerical results with the error correction rules trained on 7 faces in 9 locations are presented by Rolls and Stringer (2001). For all the results the synaptic weights were clipped to be positive during the simulation, because it is important to test that decrementing synaptic weights purely within the positive interval  $w \in [0, \infty]$  will provide significantly enhanced performance. That is, it is important to show that error correction rules do not necessarily require possibly biologically implausible modifiable negative weights. For each of the rules (21–25), the parameter  $\beta$  has been individually optimized to the following respective values: 4.9, 2.2, 2.2, 3.8, 2.2. All five error correction rules offer considerably improved performance over both the standard trace rule (14) and rule (17). Networks trained with rule (21) performed best, and this is probably due to two reasons. Firstly, rule (21) incorporates an exponential trace  $\bar{y}^{\tau-1}$  in its target  $t^\tau$ , and we would expect this to help neurons to learn more quickly to respond invariantly to a class of inputs that occur close together in time. Hence, setting  $\eta = 0$  as in rule (22) results in reduced performance. Secondly, unlike rules (23) and (24), rule (21) does not contain any component of  $y^\tau$  in its target. If we examine rules (23), (24), we see that their respective targets  $\beta \bar{y}^\tau$ ,  $\beta \bar{y}^{\tau+1}$  contain significant components of  $y^\tau$ .

### 5.3.3. Relationship to temporal difference learning

Rolls and Stringer (2001) not only considered the relationship of rule (17) to error correction, but also considered how the error correction rules shown in equations (21–25) are related to temporal difference learning (Sutton, 1988; Sutton and Barto, 1998). Sutton (1988) described temporal difference methods in the context of prediction learning. These methods are a class of incremental learning techniques that can learn to predict final outcomes through comparison of successive predictions from the preceding time steps. This is in contrast to traditional supervised learning, which involves the comparison of predictions only with the final outcome. Consider a series of multistep prediction problems in which for each problem there is a sequence of observation vectors,  $\mathbf{x}^1, \mathbf{x}^2, \dots, \mathbf{x}^m$ , at successive timesteps, followed by a final scalar outcome  $z$ . For each sequence of observations temporal difference methods form a sequence of predictions  $y^1, y^2, \dots, y^m$ , each of which is a prediction of  $z$ . These predictions are based on the

observation vectors  $\mathbf{x}^\tau$  and a vector of modifiable weights  $\mathbf{w}$ ; i.e., the prediction at time step  $\tau$  is given by  $y^\tau(\mathbf{x}^\tau, \mathbf{w})$ , and for a linear dependency the prediction is given by  $y^\tau = \mathbf{w}^T \mathbf{x}^\tau$ . (Note here that  $\mathbf{w}^T$  is the transpose of the weight vector  $\mathbf{w}$ .) The problem of prediction is to calculate the weight vector  $\mathbf{w}$  such that the predictions  $y^\tau$  are good estimates of the outcome  $z$ .

The supervised learning approach to the prediction problem is to form pairs of observation vectors  $\mathbf{x}^\tau$  and outcome  $z$  for all time steps, and compute an update to the weights according to the gradient descent equation

$$\delta \mathbf{w} = \alpha (z - y^\tau) \nabla_{\mathbf{w}} y^\tau \quad (26)$$

where  $\alpha$  is a learning rate parameter and  $\nabla_{\mathbf{w}}$  indicates the gradient with respect to the weight vector  $\mathbf{w}$ . However, this learning procedure requires all calculation to be done at the end of the sequence, once  $z$  is known. To remedy this, it is possible to replace method (26) with a temporal difference algorithm that is mathematically equivalent but allows the computational workload to be spread out over the entire sequence of observations. Temporal difference methods are a particular approach to updating the weights based on the values of successive predictions,  $y^\tau, y^{\tau+1}$ . Sutton (1988) showed that the following temporal difference algorithm is equivalent to method (26)

$$\delta \mathbf{w} = \alpha (y^{\tau+1} - y^\tau) \sum_{k=1}^{\tau} \nabla_{\mathbf{w}} y^k, \quad (27)$$

where  $y^{m+1} \equiv z$ . However, unlike method (26) this can be computed incrementally at each successive time step since each update depends only on  $y^{\tau+1}$ ,  $y^\tau$  and the sum of  $\nabla_{\mathbf{w}} y^k$  over previous time steps  $k$ . The next step taken in Sutton (1988) is to generalize equation (27) to the following final form of temporal difference algorithm, known as “TD( $\lambda$ )”

$$\delta \mathbf{w} = \alpha (y^{\tau+1} - y^\tau) \sum_{k=1}^{\tau} \lambda^{\tau-k} \nabla_{\mathbf{w}} y^k \quad (28)$$

where  $\lambda \in [0, 1]$  is an adjustable parameter that controls the weighting on the vectors  $\nabla_{\mathbf{w}} y^k$ . Equation (28) represents a much broader class of learning rules than the more usual gradient descent-based rule (27), which is in fact the special case TD(1).

A further special case of equation (28) is for  $\lambda = 0$ , i.e., TD(0), as follows

$$\delta \mathbf{w} = \alpha (y^{\tau+1} - y^\tau) \nabla_{\mathbf{w}} y^\tau. \quad (29)$$

But for problems where  $y^\tau$  is a linear function of  $\mathbf{x}^\tau$  and  $\mathbf{w}$ , we have  $\nabla_{\mathbf{w}} y^\tau = \mathbf{x}^\tau$ , and so equation (29) becomes

$$\delta \mathbf{w} = \alpha (y^{\tau+1} - y^\tau) \mathbf{x}^\tau. \quad (30)$$

If we assume the prediction process is being performed by a neuron with a vector of inputs  $\mathbf{x}^\tau$ , synaptic weight vector  $\mathbf{w}$ , and output  $y^\tau = \mathbf{w}^T \mathbf{x}^\tau$ , then we see that the TD(0) algorithm (30) is identical to the error correction rule (25) with  $\beta = 1$ . In understanding this

comparison with temporal difference learning, it may be useful to note that the firing at the end of a sequence of the transformed exemplars of a stimulus is effectively the temporal difference target  $z$ . This establishes a link to temporal difference learning (Rolls, 2008b). Further, we note that from learning epoch to learning epoch, the target  $z$  for a given neuron will gradually settle down to be more and more fixed as learning proceeds.

We now explore in more detail the relation between the error correction rules described above and temporal difference learning. For each sequence of observations with a single outcome the temporal difference method (30), when viewed as an error correction rule, is attempting to adapt the weights such that  $y^{\tau+1} = y^\tau$  for all successive pairs of time steps – the same general idea underlying the error correction rules (21–25). Furthermore, in Sutton and Barto (1998), where temporal difference methods are applied to reinforcement learning, the TD( $\lambda$ ) approach is again further generalized by replacing the target  $y^{\tau+1}$  by any weighted average of predictions  $y$  from arbitrary future timesteps, e.g.,  $t^\tau = \frac{1}{2}y^{\tau+3} + \frac{1}{2}y^{\tau+7}$ , including an exponentially weighted average extending forward in time. So a more general form of the temporal difference algorithm has the form

$$\delta w = \alpha (t^\tau - y^\tau) \mathbf{x}^\tau, \quad (31)$$

where here the target  $t^\tau$  is an arbitrary weighted average of the predictions  $y$  over future timesteps. Of course, with standard temporal difference methods the target  $t^\tau$  is always an average over *future* timesteps  $k = \tau + 1, \tau + 2$ , etc. But in the five error correction rules this is only true for the last exemplar (25). This is because with the problem of prediction, for example, the ultimate target of the predictions  $y^1, \dots, y^m$  is a final outcome  $y^{m+1} \equiv z$ . However, this restriction does not apply to our particular application of neurons trained to respond to temporal classes of inputs within VisNet. Here we only wish to set the firing rates  $y^1, \dots, y^m$  to the same value, not some final given value  $z$ . However, the more general error correction rules clearly have a close relationship to standard temporal difference algorithms. For example, it can be seen that equation (22) with  $\beta = 1$  is in some sense a temporal mirror image of equation (30), particularly if the updates  $\delta w_j$  are added to the weights  $w_j$  only at the end of a sequence. That is, rule (22) will attempt to set  $y^1, \dots, y^m$  to an *initial* value  $y^0 \equiv 0$ . This relationship to temporal difference algorithms allows us to begin to exploit established temporal difference analyses to investigate the convergence properties of the error correction methods (Rolls and Stringer, 2001).

Although the main aim of Rolls and Stringer (2001) in relating error correction rules to temporal difference learning was to begin to exploit established temporal difference analyses, they observed that the most general form of temporal difference learning, TD( $\lambda$ ), in fact suggests an interesting generalization to the existing error correction learning rules for which we currently have  $\lambda = 0$ . Assuming  $y^\tau = \mathbf{w}^T \mathbf{x}^\tau$  and  $\nabla_{\mathbf{w}} y^\tau = \mathbf{x}^\tau$ , the general equation (28) for TD( $\lambda$ ) becomes

$$\delta w = \alpha (y^{\tau+1} - y^\tau) \sum_{k=1}^{\tau} \lambda^{\tau-k} \mathbf{x}^k \quad (32)$$

where the term  $\sum_{k=1}^{\tau} \lambda^{\tau-k} \mathbf{x}^k$  is a weighted sum of the vectors  $\mathbf{x}^k$ . This suggests generalizing the original five error correction rules (21–25) by replacing the term  $x_j^\tau$  by a weighted sum  $\hat{x}_j^\tau = \sum_{k=1}^{\tau} \lambda^{\tau-k} x_j^k$  with  $\lambda \in [0, 1]$ . In Sutton (1988)  $\hat{x}_j^\tau$  is calculated according to

$$\hat{x}_j^\tau = x_j^\tau + \lambda \hat{x}_j^{\tau-1} \quad (33)$$

with  $\hat{x}_j^0 \equiv 0$ . This gives the following five temporal difference-inspired error correction rules

$$\delta w_j = \alpha (\beta \bar{y}^{\tau-1} - y^\tau) \hat{x}_j^\tau, \quad (34)$$

$$\delta w_j = \alpha (\beta y^{\tau-1} - y^\tau) \hat{x}_j^\tau, \quad (35)$$

$$\delta w_j = \alpha (\beta \bar{y}^\tau - y^\tau) \hat{x}_j^\tau, \quad (36)$$

$$\delta w_j = \alpha (\beta \bar{y}^{\tau+1} - y^\tau) \hat{x}_j^\tau, \quad (37)$$

$$\delta w_j = \alpha (\beta y^{\tau+1} - y^\tau) \hat{x}_j^\tau, \quad (38)$$

where it may be readily seen that equation (35) and (38) are special cases of equations (34) and (37), respectively, with  $\eta = 0$ . As with the trace  $\bar{y}^\tau$ , the term  $\hat{x}_j^\tau$  is reset to zero when a new stimulus is presented. These five rules can be related to the more general TD( $\lambda$ ) algorithm, but continue to be biologically plausible using only local cell variables. Setting  $\lambda = 0$  in rules (34–38), gives us back the original error correction rules (21–25) which may now be related to TD(0).

Numerical results with error correction rules (34–38), and  $\hat{x}_j^\tau$  calculated according to equation (33) with  $\lambda = 1$ , with positive clipping of weights, trained on 7 faces in 9 locations are presented by Rolls and Stringer (2001). For each of the rules (34–38), the parameter  $\beta$  has been individually optimized to the following respective values: 1.7, 1.8, 1.5, 1.6, 1.8. Comparing these five temporal difference-inspired rules it was found that the best performance is obtained with rule (38) where many more cells reach the maximum level of performance possible with respect to the single cell information measure. In fact, this rule offered the best such results. This may well be due to the fact that this rule may be directly compared to the standard TD(1) learning rule, which itself may be related to classical supervised learning for which there are well known optimality results, as discussed further by Rolls and Stringer (2001).

From the simulations described by Rolls and Stringer (2001) it appears that the form of optimization described above associated with TD(1) rather than TD(0) leads to better performance within VisNet. The TD(1)-like rule (38) with  $\lambda = 1.0$  and  $\beta = 1.8$  gave considerably superior results to the TD(0)-like rule (25) with  $\beta = 2.2$ . In fact, the former of these two rules provided the best single cell information results in these studies. We hypothesize that these results are related to the fact that only a finite set of image sequences is presented to VisNet, and so the type of optimization performed by TD(1) for repeated presentations of a finite data set is more appropriate for this problem than the form of optimization performed by TD(0).

### 5.3.4. Discussion of the different training rules

In terms of biological plausibility, we note the following. First, all the learning rules investigated by Rolls and Stringer (2001) are local learning rules, and in this sense are biologically plausible (Rolls and Treves, 1998; Rolls, 2008b). (The rules are local in that the terms used to modify the synaptic weights are potentially available in the pre- and post-synaptic elements.)

Second we note that all the rules do require some evidence of the activity on one or more previous stimulus presentations to be available when the synaptic weights are updated. Some of the rules, e.g., learning rule (23), use the trace  $\bar{y}^\tau$  from the current time level, while rules (22) and (25) do not need to use an exponential trace of the neuronal firing rate, but only the instantaneous firing rates  $y$  at two successive time steps. It is known that synaptic plasticity does involve a combination of separate processes each with potentially differing time courses (Koch, 1999), and these different processes could contribute to trace rule learning. Another mechanism suggested for implementing a trace of previous neuronal activity is the continuing firing for often 300 ms produced by a short (16 ms) presentation of a visual stimulus (Rolls and Tovee, 1994) which is suggested to be implemented by local cortical recurrent attractor networks (Rolls and Treves, 1998).

Third, we note that in utilizing the trace in the targets  $t^\tau$ , the error correction (or temporal difference-inspired) rules perform a comparison of the instantaneous firing  $y^\tau$  with a temporally nearby value of the activity, and this comparison involves a subtraction. The subtraction provides an error, which is then used to increase or decrease the synaptic weights. This is a somewhat different operation from long-term depression (LTD) as well as long-term potentiation (LTP), which are *associative* changes which depend on the pre- and post-synaptic activity. However, it is interesting to note that an error correction rule which appears to involve a subtraction of current firing from a target might be implemented by a combination of an associative process operating with the trace, and an anti-Hebbian process operating to remove the effects of the current firing. For example, the synaptic updates  $\delta w_j = \alpha(t^\tau - y^\tau)x_j^\tau$  can be decomposed into two separate associative processes  $\alpha t^\tau x_j^\tau$  and  $-\alpha y^\tau x_j^\tau$ , that may occur independently. (The target,  $t^\tau$ , could in this case be just the trace of previous neural activity from the preceding trials, excluding any contribution from the current firing.) Another way to implement an error correction rule using associative synaptic modification would be to force the post-synaptic neuron to respond to the error term. Although this has been postulated to be an effect which could be implemented by the climbing fiber system in the cerebellum (Ito, 1984, 1989; Rolls and Treves, 1998), there is no similar system known for the neocortex, and it is not clear how this particular implementation of error correction might operate in the neocortex.

In Section 5.3.2 we describe five learning rules as error correction rules. We now discuss an interesting difference of these error correction rules from error correction rules as conventionally applied. It is usual to derive the general form of error correction learning rule from gradient descent minimization in the following way (Hertz et al., 1991). Consider the idealized situation of a

single neuron with a number of inputs  $x_j$  and output  $y = \sum_j w_j x_j$ , where  $w_j$  are the synaptic weights. We assume that there are a number of input patterns and that for the  $k$ th input pattern,  $\mathbf{x}^k = [x_1^k, x_2^k, \dots]^T$ , the output  $y^k$  has a target value  $t^k$ . Hence an error measure or cost function can be defined as

$$e(\mathbf{w}) = \frac{1}{2} \sum_k (t^k - y^k)^2 = \frac{1}{2} \sum_k \left( t^k - \sum_j w_j x_j^k \right)^2. \quad (39)$$

This cost function is a function of the input patterns  $\mathbf{x}^k$  and the synaptic weight vector  $\mathbf{w} = [w_1, w_2, \dots]^T$ . With a fixed set of input patterns, we can reduce the error measure by employing a gradient descent algorithm to calculate an improved set of synaptic weights. Gradient descent achieves this by moving downhill on the error surface defined in  $\mathbf{w}$  space using the update

$$\delta w_j = -\alpha \frac{\partial e}{\partial w_j} = \alpha \sum_k (t^k - y^k) x_j^k. \quad (40)$$

If we update the weights after each pattern  $k$ , then the update takes the form of an error correction rule

$$\delta w_j = \alpha (t^k - y^k) x_j^k, \quad (41)$$

which is also commonly referred to as the delta rule or Widrow-Hoff rule (see Widrow and Hoff, 1960; Widrow and Stearns, 1985). Error correction rules continually compare the neuronal output with its pre-specified target value and adjust the synaptic weights accordingly. In contrast, the way Rolls and Stringer (2001) introduced of utilizing error correction is to specify the target as the activity trace based on the firing rate at nearby timesteps. Now the actual firing at those nearby time steps is not a pre-determined fixed target, but instead depends on how the network has actually evolved. This effectively means the cost function  $e(\mathbf{w})$  that is being minimized changes from timestep to timestep. Nevertheless, the concept of calculating an error, and using the magnitude and direction of the error to update the synaptic weights, is the similarity Rolls and Stringer (2001) made to gradient descent learning.

To conclude this discussion, the error correction and temporal difference rules explored by Rolls and Stringer (2001) provide interesting approaches to help understand invariant pattern recognition learning. Although we do not know whether the full power of these rules is expressed in the brain, we provided suggestions about how they might be implemented. At the same time, we note that the original trace rule used by Földiák (1991), Rolls (1992), and Wallis and Rolls (1997) is a simple associative rule, is therefore biologically very plausible, and, while not as powerful as many of the other rules introduced by Rolls and Stringer (2001), can nevertheless solve the same class of problem. Rolls and Stringer (2001) also emphasized that although they demonstrated how a number of new error correction and temporal difference rules might play a role in the context of view-invariant object recognition, they may also operate elsewhere where it is important for neurons to learn to respond similarly to temporal classes of inputs that tend to occur close together in time.



## 5.4. THE ISSUE OF FEATURE BINDING, AND A SOLUTION

In this section we investigate two key issues that arise in hierarchical layered network architectures, such as VisNet, other examples of which have been described and analyzed by Fukushima (1980), Ackley et al. (1985), Rosenblatt (1961), and Riesenhuber and Poggio (1999b). One issue is whether the network can discriminate between stimuli that are composed of the same basic alphabet of features. The second issue is whether such network architectures can find solutions to the spatial binding problem. These issues are addressed next and by Elliffe et al. (2002) and Rolls (2008b).

The first issue investigated is whether a hierarchical layered network architecture of the type exemplified by VisNet can discriminate stimuli that are composed of a limited set of features and where the different stimuli include cases where the feature sets are subsets and supersets of those in the other stimuli. An issue is that if the network has learned representations of both the parts and the wholes, will the network identify that the whole is present when it is shown, and not just that one or more parts is present. (In many investigations with VisNet, complex stimuli (such as faces) were used where each stimulus might contain unique features not present in the other stimuli.) To address this issue Elliffe et al. (2002) used stimuli that are composed from a set of four features which are designed so that each feature is spatially separate from the other features, and no unique combination of firing caused, for example, by overlap of horizontal and vertical filter outputs in the input representation distinguishes any one stimulus from the others. The results described in Section 5.4.4 show that VisNet can indeed learn correct invariant representations of stimuli which do consist of feature sets where individual features do not overlap spatially with each other and where the stimuli can be composed of sets of features which are supersets or subsets of those in other stimuli. Fukushima and Miyake (1982) did not address this crucial issue where different stimuli might be composed of subsets or supersets of the same set of features, although they did show that stimuli with partly overlapping features could be discriminated by the Neocognitron.

In Section 5.4.5 we address the spatial binding problem in architectures such as VisNet. This computational problem that needs to be addressed in hierarchical networks such as the primate visual system and VisNet is how representations of features can be (e.g., translation) invariant, yet can specify stimuli or objects in which the features must be specified in the correct spatial arrangement. This is the feature binding problem, discussed, for example, by von der Malsburg (1990), and arising in the context of hierarchical layered systems (Rosenblatt, 1961; Fukushima, 1980; Ackley et al., 1985). The issue is whether or not features are bound into the correct combinations in the correct relative spatial positions, or if alternative combinations of known features or the same features in different relative spatial positions would elicit the same responses. All this has to be achieved while at the same time producing position-invariant recognition of the whole combination of features, that is, the object. This is a major computational issue that needs to be solved for memory systems in the brain to operate correctly. This can be achieved by what is effectively a learning process that builds into the system a set of neurons in the hierarchical network that enables the recognition process to operate correctly with the appropriate position, size, view, etc. invariances.

### 5.4.1. Syntactic binding of separate neuronal ensembles by synchronization

The problem of syntactic binding of neuronal representations, in which some features must be bound together to form one object, and other simultaneously active features must be bound together to represent another object, has been addressed by von der Malsburg (1990). He has proposed that this could be performed by temporal synchronization of those neurons that were temporarily part of one representation in a different time slot from other neurons that were temporarily part of another representation. The idea is attractive in allowing arbitrary relinking of features in different combinations. Singer, Engel, König, and colleagues (Singer et al., 1990; Engel et al., 1992; Singer and Gray, 1995; Singer, 1999; Fries, 2005, 2009; Womelsdorf et al., 2007), and others (Abeles, 1991) have obtained some evidence that when features must be bound, synchronization of neuronal populations can occur (but see Shadlen and Movshon, 1999), and this has been modeled (Hummel and Biederman, 1992).

Synchronization to implement syntactic binding has a number of disadvantages and limitations (Rolls and Treves, 1998, 2011; Riesenhuber and Poggio, 1999a; Rolls, 2008b). The greatest computational problem is that synchronization does not by itself define the spatial relations between the features being bound, so is not just as a binding mechanism adequate for shape recognition. For example, temporal binding might enable features 1, 2, and 3, which might define one stimulus to be bound together and kept separate from, for example, another stimulus consisting of features 2, 3, and 4, but would require a further temporal binding (leading in the end potentially to a combinatorial explosion) to indicate the relative spatial positions of the 1, 2, and 3 in the 123 stimulus, so that it can be discriminated from, e.g., 312.

A second problem with the synchronization approach to the spatial binding of features is that, when stimulus-dependent temporal synchronization has been rigorously tested with information theoretic approaches, it has so far been found that most of the information available is in the number of spikes, with rather little, less than 5% of the total information, in stimulus-dependent synchronization (Franco et al., 2004; Rolls et al., 2004; Aggelopoulos et al., 2005; Rolls, 2008b; Rolls and Treves, 2011). For example, Aggelopoulos et al. (2005) showed that when macaques used object-based attention to search for one of two objects to touch in a complex natural scene, between 99 and 94% of the information was present in the firing rates of inferior temporal cortex neurons, and less than 5% in any stimulus-dependent synchrony that was present between the simultaneously recorded inferior temporal cortex neurons. The implication of these results is that any stimulus-dependent synchrony that is present is not quantitatively important as measured by information theoretic analyses under natural scene conditions when feature binding, segmentation of objects from the background, and attention are required. This has been found for the inferior temporal cortex, a brain region where features are put together to form representations of objects (Rolls and Deco, 2002; Rolls, 2008b), and where attention has strong effects, at least in scenes with blank backgrounds (Rolls et al., 2003). It would of course also be of interest to test the same hypothesis in earlier visual areas, such as V4, with quantitative,

information theoretic, techniques (Rolls and Treves, 2011). In connection with rate codes, it should be noted that a rate code implies using the number of spikes that arrive in a given time, and that this time can be very short, as little as 20–50 ms, for very useful amounts of information to be made available from a population of neurons (Tovee et al., 1993; Rolls and Tovee, 1994; Rolls et al., 1994, 1999, 2006a; Tovee and Rolls, 1995; Rolls, 2003, 2008b; Rolls and Treves, 2011).

A third problem with the synchronization or “communication through coherence” approach (Fries, 2005, 2009) is that when information transmission between connected networks is analyzed, synchronization is not produced at the levels of synaptic strength necessary for information transmission between the networks, and indeed does not appear to affect the information transmission between a pair of weakly coupled networks that model weakly coupled cortical networks (Rolls et al., 2012).

In the context of VisNet, and how the real visual system may operate to implement object recognition, the use of synchronization does not appear to match the way in which the visual system is organized. For example, von der Malsburg’s argument would indicate that, using only a two-layer network, synchronization could provide the necessary feature linking to perform object recognition with relatively few neurons, because they can be reused again and again, linked differently for different objects. In contrast, the primate uses a considerable part of its cortex, perhaps 50% in monkeys, for visual processing, with therefore what could be in the order of  $6 \times 10^8$  neurons and  $6 \times 10^{12}$  synapses involved (Rolls, 2008b), so that the solution adopted by the real visual system may be one which relies on many neurons with simpler processing than arbitrary syntax implemented by synchronous firing of separate assemblies suggests. On the other hand, a solution such as that investigated by VisNet, which forms low-order combinations of what is represented in previous layers, is very demanding in terms of the number of neurons required, and this matches what is found in the primate visual system.

#### 5.4.2. Sigma-Pi neurons

Another approach to a binding mechanism is to group spatial features based on local mechanisms that might operate for closely adjacent synapses on a dendrite (in what is a Sigma-Pi type of neuron, see Section 7; Finkel and Edelman, 1987; Mel et al., 1998; Rolls, 2008b). A problem for such architectures is how to force one particular neuron to respond to the same feature combination invariantly with respect to all the ways in which that feature combination might occur in a scene.

#### 5.4.3. Binding of features and their relative spatial position by feature combination neurons

The approach to the spatial binding problem that is proposed for VisNet is that individual neurons at an early stage of processing are set up (by learning) to respond to low-order combinations of input features occurring in a given relative spatial arrangement and position on the retina (Rolls, 1992, 1994, 1995; Wallis and Rolls, 1997; Rolls and Treves, 1998; Elliffe et al., 2002; Rolls and Deco, 2002; cf. Feldman, 1985). (By low-order combinations of input features we mean combinations of a few input features. By forming neurons that respond to combinations of a few features

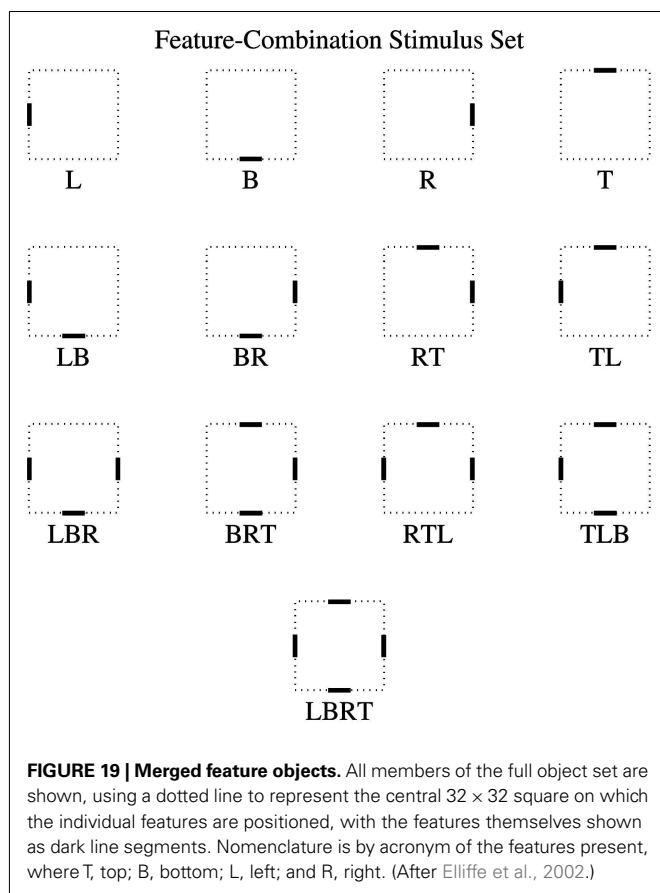
in the correct spatial arrangement the advantages of the scheme for syntactic binding are obtained, yet without the combinatorial explosion that would result if the feature combination neurons responded to combinations of many input features so producing potentially very specifically tuned neurons which very rarely responded.) Then invariant representations are developed in the next layer from these feature combination neurons which already contain evidence on the local spatial arrangement of features. Finally, in later layers, only one stimulus would be specified by the particular set of low-order feature combination neurons present, even though each feature combination neuron would itself be somewhat invariant. The overall design of the scheme is shown in Figure 9. Evidence that many neurons in V1 respond to combinations of spatial features with the correct spatial configuration is now starting to appear (see Section 4), and neurons that respond to feature combinations (such as two lines with a defined angle between them, and overall orientation) are found in V2 (Hegde and Van Essen, 2000; Ito and Komatsu, 2004). The tuning of a VisNet layer 1 neuron to a combination of features in the correct relative spatial position is illustrated in Figures 12 and 13.

#### 5.4.4. Discrimination between stimuli with super- and sub-set feature combinations

Some investigations with VisNet (Wallis and Rolls, 1997) have involved groups of stimuli that might be identified by some unique feature common to all transformations of a particular stimulus. This might allow VisNet to solve the problem of transform invariance by simply learning to respond to a unique feature present in each stimulus. For example, even in the case where VisNet was trained on invariant discrimination of T, L, and +, the representation of the T stimulus at the spatial-filter level inputs to VisNet might contain unique patterns of filter outputs where the horizontal and vertical parts of the T join. The unique filter outputs thus formed might distinguish the T from, for example, the L.

Elliffe et al. (2002) tested whether VisNet is able to form transform invariant cells with stimuli that are specially composed from a common alphabet of features, with no stimulus containing any firing in the spatial-filter inputs to VisNet not present in at least one of the other stimuli. The limited alphabet enables the set of stimuli to consist of feature sets which are subsets or supersets of those in the other stimuli.

For these experiments the common pool of stimulus features chosen was a set of two horizontal and two vertical  $8 \times 1$  bars, each aligned with the sides of a  $32 \times 32$  square. The stimuli can be constructed by arbitrary combination of these base level features. We note that effectively the stimulus set consists of four features, a top bar (T), a bottom bar (B), a left bar (L), and a right bar (R). Figure 19 shows the complete set used, containing the possible image feature combination. Subsequent discussion will group these objects by the number of features each contains: single-; double-; triple-; and quadruple-feature objects correspond to the respective rows of Figure 19. Stimuli are referred to by the list of features they contain; e.g., “LBR” contains the left, bottom, and right features, while “TL” contains top and left only. Further details of how the stimuli were prepared are provided by Elliffe et al. (2002).



To train the network a stimulus was presented in a randomized sequence of nine locations in a square grid across the  $128 \times 128$  input retina of VisNet2. The central location of the square grid was in the center of the “retina,” and the eight other locations were offset 8 pixels horizontally and/or vertically from this. Two different learning rules were used, “Hebbian” (16), and “trace” (17), and also an untrained condition with random weights. As in earlier work (Wallis and Rolls, 1997; Rolls and Milward, 2000) only the trace rule led to any cells with invariant responses, and the results shown are for networks trained with the trace rule.

The results with VisNet trained on the set of stimuli shown in **Figure 19** with the trace rule are as follows. First, it was found that single neurons in the top layer learned to differentiate between the stimuli in that the responses of individual neurons were maximal for one of the stimuli and had no response to any of the other stimuli invariantly with respect to location. Moreover, the translation invariance was perfect for every stimulus (by different neurons) over every location (for all stimuli except “RTL” and “TLBR”).

The results presented show clearly that the VisNet paradigm can accommodate networks that can perform invariant discrimination of objects that have a subset–superset relationship. The result has important consequences for feature binding and for discriminating stimuli for other stimuli which may be supersets of the first stimulus. For example, a VisNet cell which responds invariantly to feature combination TL can genuinely signal the presence of exactly that combination, and will not necessarily be activated

by T alone, or by TLB. The basis for this separation by competitive networks of stimuli which are subsets and supersets of each other is described by Rolls and Treves, 1998, Section 4.3.6) and by Rolls (2008b).

#### 5.4.5. Feature binding in a hierarchical network with invariant representations of local feature combinations

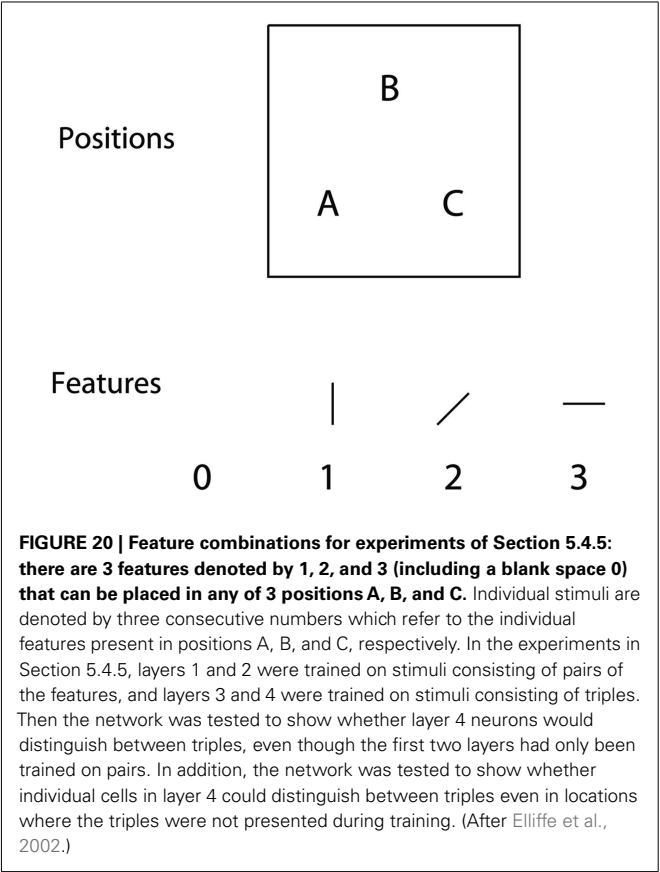
In this section we consider the ability of output layer neurons to learn new stimuli if the lower layers are trained solely through exposure to simpler feature combinations from which the new stimuli are composed. A key question we address is how invariant representations of low-order feature combinations in the early layers of the visual system are able to uniquely specify the correct spatial arrangement of features in the overall stimulus and contribute to preventing false recognition errors in the output layer.

The problem, and its proposed solution, can be treated as follows. Consider an object 1234 made from the features 1, 2, 3, and 4. The invariant low-order feature combinations might represent 12, 23, and 34. Then if neurons at the next layer respond to combinations of the activity of these neurons, the only neurons in the next layer that would respond would be those tuned to 1234, not to, for example, 3412, which is distinguished from 1234 by the input of a pair neuron responding to 41 rather than to 23. The argument (Rolls, 1992) is that low-order spatial-feature combination neurons in the early stage contain sufficient spatial information so that a particular combination of those low-order feature combination neurons specifies a unique object, even if the relative positions of the low-order feature combination neurons are not known, because they are somewhat invariant.

The architecture of VisNet is intended to solve this problem partly by allowing high-spatial precision combinations of input features to be formed in layer 1. The actual input features in VisNet are, as described above, the output of oriented spatial-frequency tuned filters, and the combinations of these formed in layer 1 might thus be thought of in a simple way as, for example, a T or an L or for that matter a Y. Then in layer 2, application of the trace rule might enable neurons to respond to a T with limited spatial invariance (limited to the size of the region of layer 1 from which layer 2 cells receive their input). Then an “object” such as H might be formed at a higher layer because of a conjunction of two Ts in the same small region.

To show that VisNet can actually solve this problem, Elliffe et al. (2002) performed the experiments described next. They trained the first two layers of VisNet with feature pair combinations, forming representations of feature pairs with some translation invariance in layer 2. Then they used feature triples as input stimuli, allowed no more learning in layers 1 and 2, and then investigated whether layers 3 and 4 could be trained to produce invariant representations of the triples where the triples could only be distinguished if the local spatial arrangement of the features within the triple had effectively to be encoded in order to distinguish the different triples. For this experiment, they needed stimuli that could be specified in terms of a set of different features (they chose vertical (1), diagonal (2), and horizontal (3) bars) each capable of being shown at a set of different relative spatial positions (designated A, B, and C), as shown in **Figure 20**.

The stimuli are thus defined in terms of what features are present and their precise spatial arrangement with respect to each other. The length of the horizontal and vertical feature bars shown in **Figure 20** is 8 pixels. To train the network a stimulus (that is a pair or triple feature combination) is presented in a randomized sequence of nine locations in a square grid across the  $128 \times 128$  input retina. The central location of the square grid is in the center of the “retina,” and the eight other locations are offset 8 pixels horizontally and/or vertically from this. We refer to the two and three feature stimuli as “pairs” and “triples,” respectively. Individual stimuli are denoted by three numbers which refer to the individual features present in positions A, B and C, respectively. For example, a stimulus with positions A and C containing a vertical and diagonal bar, respectively, would be referred to as stimulus 102, where the 0 denotes no feature present in position B. In total there are 18 pairs (120, 130, 210, 230, 310, 320, 012, 013, 021, 023, 031, 032, 102, 103, 201, 203, 301, 302) and 6 triples (123, 132, 213, 231, 312, 321). This nomenclature not only defines which features are present within objects, but also the spatial relationships of their component features. Then the computational problem can be illustrated by considering the triple 123. If invariant representations are formed of single features, then there would be no way that neurons higher in the hierarchy could distinguish the object 123 from 213 or any other arrangement of the three features. An approach to this problem (see, e.g., Rolls, 1992) is to form early on in the processing neurons that respond to overlapping combinations of features in the correct spatial arrangement, and



then to develop invariant representations in the next layer from these neurons which already contain evidence on the local spatial arrangement of features. An example might be that with the object 123, the invariant feature pairs would represent 120, 023, and 103. Then if neurons at the next layer correspond to combinations of these neurons, the only next layer neurons that would respond would be those tuned to 123, not to, for example, 213. The argument is that the low-order spatial-feature combination neurons in the early stage contain sufficient spatial information so that a particular combination of those low-order feature combination neurons specifies a unique object, even if the relative positions of the low-order feature combination neurons are not known because these neurons are somewhat translation-invariant (cf. also Fukushima, 1988).

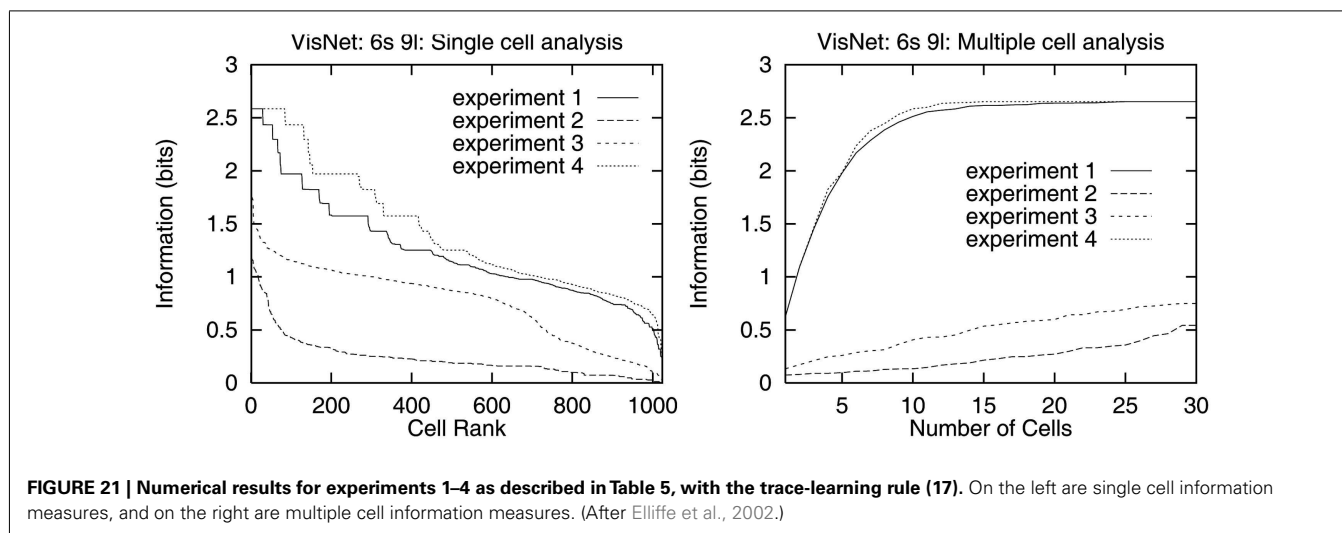
The stimuli used in the experiments of Elliffe et al. (2002) were constructed from pre-processed component features as discussed in Section 5.4.4. That is, base stimuli containing a single feature were constructed and filtered, and then the pairs and triples were constructed by merging these pre-processed single feature images. In the first experiment layers 1 and 2 of VisNet were trained with the 18 feature pairs, each stimulus being presented in sequences of 9 locations across the input. This led to the formation of neurons that responded to the feature pairs with some translation invariance in layer 2. Then they trained layers 3 and 4 on the 6 feature triples in the same 9 locations, while allowing no more learning in layers 1 and 2, and examined whether the output layer of VisNet had developed transform invariant neurons to the 6 triples. The idea was to test whether layers 3 and 4 could be trained to produce invariant representations of the triples where the triples could only be distinguished if the local spatial arrangement of the features within the triple had effectively to be encoded in order to distinguish the different triples. The results from this experiment were compared and contrasted with results from three other experiments which involved different training regimes for layers 1, 2 and layers 3, 4. All four experiments are summarized in **Table 5**. Experiment 2 involved no training in layers 1, 2 and 3, 4, with the synaptic weights left unchanged from their initial random values. These results are included as a baseline performance with which to compare results from the other experiments 1, 3, and 4. The model parameters used in these experiments were as described by Rolls and Milward (2000) and Rolls and Stringer (2001).

In **Figure 21** we present numerical results for the four experiments listed in **Table 5**. On the left are the single cell information measures for all top (4th) layer neurons ranked in order of their

**Table 5 | The different training regimes used in VisNet experiments 1–4 of Section 5.4.5.**

	Layers 1, 2	Layers 3, 4
Experiment 1	Trained on pairs	Trained on triples
Experiment 2	No training	No training
Experiment 3	No training	Trained on triples
Experiment 4	Trained on triples	Trained on triples

*In the no training condition the synaptic weights were left in their initial untrained random values.*



invariance to the triples, while on the right are multiple cell information measures. To help to interpret these results we can compute the maximum single cell information measure according to

$$\text{Maximum single cell information} = \log_2(\text{Number of triples}), \quad (42)$$

where the number of triples is 6. This gives a maximum single cell information measure of 2.6 bits for these test cases. First, comparing the results for experiment 1 with the baseline performance of experiment 2 (no training) demonstrates that even with the first two layers trained to form invariant responses to the pairs, and then only layers 3 and 4 trained on feature triples, layer 4 is indeed capable of developing translation-invariant neurons that can discriminate effectively between the 6 different feature triples. Indeed, from the single cell information measures it can be seen that a number of cells have reached the maximum level of performance in experiment 1. In addition, the multiple cell information analysis presented in **Figure 21** shows that all the stimuli could be discriminated from each other by the firing of a number of cells. Analysis of the response profiles of individual cells showed that a fourth layer cell could respond to one of the triple feature stimuli and have no response to any other of the triple feature stimuli invariantly with respect to location.

A comparison of the results from experiment 1 with those from experiment 3 (see **Table 5** and **Figure 21**) reveals that training the first two layers to develop neurons that respond invariantly to the pairs (performed in experiment 1) actually leads to improved invariance of 4th layer neurons to the triples, as compared with when the first two layers are left untrained (experiment 3).

Two conclusions follow from these results (Elliffe et al., 2002). First, a hierarchical network that seeks to produce invariant representations in the way used by VisNet can solve the feature binding problem. In particular, when feature pairs in layer 2 with some translation invariance are used as the input to later layers, these later layers can nevertheless build invariant representations of objects where all the individual features in the stimulus must occur in the correct spatial position relative to each other. This

is possible because the feature combination neurons formed in the first layer (which could be trained just with a Hebb rule) do respond to combinations of input features in the correct spatial configuration, partly because of the limited size of their receptive fields. The second conclusion is that even though early layers can in this case only respond to small feature subsets, these provide, with no further training of layers 1 and 2, an adequate basis for learning to discriminate in layers 3 and 4 stimuli consisting of combinations of larger numbers of features. Indeed, comparing results from experiment 1 with experiment 4 (in which all layers were trained on triples, see **Table 5**) demonstrates that training the lower layer neurons to develop invariant responses to the pairs offers almost as good performance as training all layers on the triples (see **Figure 21**).

#### 5.4.6. Stimulus generalization to untrained transforms of new objects

Another important aspect of the architecture of VisNet is that it need not be trained with every stimulus in every possible location. Indeed, part of the hypothesis (Rolls, 1992) is that training early layers (e.g., 1–3) with a wide range of visual stimuli will set up feature analyzers in these early layers which are appropriate later on with no further training of early layers for new objects. For example, presentation of a new object might result in large numbers of low-order feature combination neurons in early layers of VisNet being active, but the particular set of feature combination neurons active would be different for the new object. The later layers of the network (in VisNet, layer 4) would then learn this new set of active layer 3 neurons as encoding the new object. However, if the new object was then shown in a new location, the same set of layer 3 neurons would be active because they respond with spatial invariance to feature combinations, and given that the layer 3–4 connections had already been set up by the new object, the correct layer 4 neurons would be activated by the new object in its new untrained location, and without any further training.

To test this hypothesis Elliffe et al. (2002) repeated the general procedure of experiment 1 of Section 5.4.5, training layers 1 and 2 with feature pairs, but then instead trained layers 3 and 4 on the

triples in only 7 of the original 9 locations. The crucial test was to determine whether VisNet could form top layer neurons that responded invariantly to the 6 triples when presented over all nine locations, not just the seven locations at which the triples had been presented during training.

It was found that VisNet is still able to develop some fourth layer neurons with perfect invariance, that is which have invariant responses over all nine locations, as shown by the single cell information analysis. The response profiles of individual fourth layer cells showed that they can continue to discriminate between the triples even in the two locations where the triples were not presented during training. In addition, the multiple cell analysis showed that a small population of cells was able to discriminate between all of the stimuli irrespective of location, even though for two of the test locations the triples had not been trained at those particular locations during the training of layers 3 and 4.

The use of transformation rules learned by early stages of the hierarchy to enable later stages to perform correctly on transformed views never seen before of objects is now being investigated by others (Leibo et al., 2010).

#### 5.4.7. Discussion of feature binding in hierarchical layered networks

Elliffe et al. (2002) thus first showed (see Section 5.4.4) that hierarchical feature-detecting neural networks can learn to respond differently to stimuli that consist of unique combinations of non-unique input features, and that this extends to stimuli that are direct subsets or supersets of the features present in other stimuli.

Second Elliffe et al. (2002) investigated (see Section 5.4.5) the hypothesis that hierarchical layered networks can produce identification of unique stimuli even when the feature combination neurons used to define the stimuli are themselves partly translation-invariant. The stimulus identification should work correctly because feature combination neurons in which the spatial features are bound together with high-spatial precision are formed in the first layer. Then at later layers when neurons with some translation invariance are formed, the neurons nevertheless contain information about the relative spatial position of the original features. There is only then one object which will be consistent with the set of active neurons at earlier layers, which though somewhat translation-invariant as combination neurons, reflect in the activity of each neuron information about the original spatial position of the features. I note that the trace rule training used in early layers (1 and 2) in Experiments 1 and 4 would set up partly invariant feature combination neurons, and yet the late layers (3 and 4) were able to produce during training neurons in layer 4 that responded to stimuli that consisted of unique spatial arrangements of lower order feature combinations. Moreover, and very interestingly Elliffe et al. (2002) were able to demonstrate that VisNet layer 4 neurons would respond correctly to visual stimuli at untrained locations, provided that the feature subsets had been trained in early layers of the network at all locations, and that the whole stimulus had been trained at some locations in the later layers of the network.

The results described by Elliffe et al. (2002) thus provide one solution to the feature binding problem. The solution which has been shown to work in the model is that in a multilayer competitive

network, feature combination neurons which encode the spatial arrangement of the bound features are formed at intermediate layers of the network. Then neurons at later layers of the network which respond to combinations of active intermediate-layer neurons do contain sufficient evidence about the local spatial arrangement of the features to identify stimuli because the local spatial arrangement is encoded by the intermediate-layer neurons. The information required to solve the visual feature binding problem thus becomes encoded by self-organization into what become hard-wired properties of the network. In this sense, feature binding is not solved at run-time by the necessity to instantaneously set up arbitrary syntactic links between sets of co-active neurons. The computational solution proposed to the superset/subset aspect of the binding problem will apply in principle to other multilayer competitive networks, although the issues considered here have not been explicitly addressed in architectures such as the Neocognitron (Fukushima and Miyake, 1982).

Consistent with these hypotheses about how VisNet operates to achieve, by layer 4, position-invariant responses to stimuli defined by combinations of features in the correct spatial arrangement, investigations of the effective stimuli for neurons in intermediate layers of VisNet showed as follows. In layer 1, cells responded to the presence of individual features, or to low-order combinations of features (e.g., a pair of features) in the correct spatial arrangement at a small number of nearby locations. In layers 2 and 3, neurons responded to single features or to higher order combinations of features (e.g., stimuli composed of feature triples) in more locations. These findings provide direct evidence that VisNet does operate as described above to solve the feature binding problem.

A further issue with hierarchical multilayer architectures such as VisNet is that false binding errors might occur in the following way (Mozer, 1991; Mel and Fiser, 2000). Consider the output of one-layer in such a network in which there is information only about which pairs are present. How then could a neuron in the next layer discriminate between the whole stimulus (such as the triple 123 in the above experiment) and what could be considered a more distributed stimulus or multiple different stimuli composed of the separated subparts of that stimulus (e.g., the pairs 120, 023, 103 occurring in 3 of the 9 training locations in the above experiment)? The problem here is to distinguish a single object from multiple other objects containing the same component combinations (e.g., pairs). We propose that part of the solution to this general problem in real visual systems is implemented through lateral inhibition between neurons in individual layers, and that this mechanism, implemented in VisNet, acts to reduce the possibility of false recognition errors in the following two ways.

First, consider the situation in which neurons in layer  $N$  have learned to represent low-order feature combinations with location invariance, and where a neuron  $n$  in layer  $N + 1$  has learned to respond to a particular set  $\Omega$  of these feature combinations. The problem is that neuron  $n$  receives the same input from layer  $N$  as long as the same set  $\Omega$  of feature combinations is present, and cannot distinguish between different spatial arrangements of these feature combinations. The question is how can neuron  $n$  respond only to a particular favored spatial arrangement  $\Psi$  of the feature combinations contained within the set  $\Omega$ . We suggest that as the favored spatial arrangement  $\Psi$  is altered by rearranging

the spatial relationships of the component feature combinations, the new feature combinations that are formed in new locations will stimulate additional neurons nearby in layer  $N + 1$ , and these will tend to inhibit the firing of neuron  $n$ . Thus, lateral inhibition within a layer will have the effect of making neurons more selective, ensuring neuron  $n$  responds only to a single spatial arrangement  $\Psi$  from the set of feature combinations  $\Omega$ , and hence reducing the possibility of false recognition.

The second way in which lateral inhibition may help to reduce binding errors is through limiting the sparseness of neuronal firing rates within layers. In our discussion above the spurious stimuli we suggested that might lead to false recognition of triples were obtained from splitting up the component feature combinations (pairs) so that they occurred in separate training locations. However, this would lead to an increase in the number of features present in the complete stimulus; triples contain 3 features while their spurious counterparts would contain 6 features (resulting from 3 separate pairs). For this trivial example, the increase in the number of features is not dramatic, but if we consider, say, stimuli composed of 4 features where the component feature combinations represented by lower layers might be triples, then to form spurious stimuli we need to use 12 features (resulting from 4 triples occurring in separate locations). But if the lower layers also represented all possible pairs then the number of features required in the spurious stimuli would increase further. In fact, as the size of the stimulus increases in terms of the number of features, and as the size of the component feature combinations represented by the lower layers increases, there is a combinatorial explosion in terms of the number of features required as we attempt to construct spurious stimuli to trigger false recognition. And the construction of such spurious stimuli will then be prevented through setting a limit on the sparseness of firing rates within layers, which will in turn set a limit on the number of features that can be represented. Lateral inhibition is likely to contribute in both these ways to the performance of VisNet when the stimuli consist of subsets and supersets of each other, as described in Section 5.4.4.

Another way in which the problem of multiple objects is addressed is by limiting the size of the receptive fields of inferior temporal cortex neurons so that neurons in IT respond primarily to the object being fixated, but with nevertheless some asymmetry in the receptive fields (see Section 5.9). Multiple objects are then “seen” by virtue of being added to a visuo-spatial scratchpad (Rolls, 2008b).

A related issue that arises in this class of network is whether forming neurons that respond to feature combinations in the way described here leads to a combinatorial explosion in the number of neurons required. The solution to this issue that is proposed is to form only low-order combinations of features at any one stage of the network (Rolls, 1992; cf. Feldman, 1985). Using low-order combinations limits the number of neurons required, yet enables the type of computation that relies on feature combination neurons that is analyzed here to still be performed. The actual number of neurons required depends also on the redundancies present in the statistics of real-world images. Even given these factors, it is likely that a large number of neurons would be required if the ventral visual system performs the computation of invariant representations in the manner captured by the hypotheses

implemented in VisNet. Consistent with this, a considerable part of the non-human primate brain is devoted to visual information processing. The fact that large numbers of neurons and a multilayer organization are present in the primate ventral visual system is actually thus consistent with the type of model of visual information processing described here.

## 5.5. OPERATION IN A CLUTTERED ENVIRONMENT

In this section we consider how hierarchical layered networks of the type exemplified by VisNet operate in cluttered environments. Although there has been much work involving object recognition in cluttered environments with artificial vision systems, many such systems typically rely on some form of explicit segmentation followed by search and template matching procedure (see Ullman, 1996 for a general review). In natural environments, objects may not only appear against cluttered (natural) backgrounds, but also the object may be partially occluded. Biological nervous systems operate in quite a different manner to those artificial vision systems that rely on search and template matching, and the way in which biological systems cope with cluttered environments and partial occlusion is likely to be quite different also.

One of the factors that will influence the performance of the type of architecture considered here, hierarchically organized series of competitive networks, which form one class of approaches to biologically relevant networks for invariant object recognition (Fukushima, 1980; Poggio and Edelman, 1990; Rolls, 1992, 2008b; Wallis and Rolls, 1997; Rolls and Treves, 1998), is how lateral inhibition and competition are managed within a layer. Even if an object is not obscured, the effect of a cluttered background will be to fire additional neurons, which will in turn to some extent compete with and inhibit those neurons that are specifically tuned to respond to the desired object. Moreover, where the clutter is adjacent to part of the object, the feature analyzing neurons activated against a blank background might be different from those activated against a cluttered background, if there is no explicit segmentation process. We consider these issues next, following investigations of Stringer and Rolls (2000).

### 5.5.1. VisNet simulations with stimuli in cluttered backgrounds

In this section we show that recognition of objects learned previously against a blank background is hardly affected by the presence of a natural cluttered background. We go on to consider what happens when VisNet is set the task of learning new stimuli presented against cluttered backgrounds.

The images used for training and testing VisNet in the simulations described next performed by Stringer and Rolls (2000) were specially constructed. There were 7 face stimuli approximately 64 pixels in height constructed without backgrounds. In addition there were 3 possible backgrounds: a blank background (gray-scale 127, where the range is 0–255), and two cluttered backgrounds as shown in Figure 22 which are  $128 \times 128$  pixels in size. Each image presented to VisNet's  $128 \times 128$  input retina was composed of a single face stimulus positioned at one of 9 locations on either a blank or cluttered background. The cluttered background was intended to be like the background against which an object might be viewed in a natural scene. If a background is used in an experiment described here, the same background is always used,



**FIGURE 22 | Cluttered backgrounds used in VisNet simulations: backgrounds 1 and 2 are on the left and right, respectively.**

and it is always in the same position, with stimuli moved to different positions on it. The 9 stimulus locations are arranged in a square grid across the background, where the grid spacings are 32 pixels horizontally or vertically. Before images were presented to VisNet's input layer they were pre-processed by the standard set of input filters which accord with the general tuning profiles of simple cells in V1 (Hawken and Parker, 1987); full details are given in Rolls and Milward (2000). To train the network a sequence of images is presented to VisNet's retina that corresponds to a single stimulus occurring in a randomized sequence of the 9 locations across a background. At each presentation the activation of individual neurons is calculated, then their firing rates are calculated, and then the synaptic weights are updated. After a stimulus has been presented in all the training locations, a new stimulus is chosen at random and the process repeated. The presentation of all the stimuli across all locations constitutes 1 epoch of training. In this manner the network is trained one-layer at a time starting with layer 1 and finishing with layer 4. In the investigations described in this subsection, the numbers of training epochs for layers 1–4 were 50, 100, 100, and 75, respectively.

In this experiment (see Stringer and Rolls, 2000, experiment 2), VisNet was trained with the 7 face stimuli presented on a blank background, but tested with the faces presented on each of the 2 cluttered backgrounds.

The single and multiple cell information showed perfect performance. Compared to performance when shown against a blank background, there was very little deterioration in performance when testing with the faces presented on either of the two cluttered backgrounds.

This is an interesting result to compare with many artificial vision systems that would need to carry out computationally intensive serial searching and template matching procedures in order to achieve such results. In contrast, the VisNet neural network architecture is able to perform such recognition relatively quickly through a simple feed-forward computation.

Further results from this experiment showed that different neurons can achieve excellent invariant responses to each of the 7 faces even with the faces presented on a cluttered background. The response profiles are independent of location but differentiate between the faces in that the responses are maximal for only one of the faces and minimal for all other faces.

This is an interesting and important result, for it shows that after learning, special mechanisms for segmentation and for attention are not needed in order for neurons already tuned by previous learning to the stimuli to be activated correctly in the output layer. Although the experiments described here tested for position invariance, we predict and would expect that the same results would be demonstrable for size and view-invariant representations of objects.

In experiments 3 and 4 of Stringer and Rolls (2000), VisNet was trained with the 7 face stimuli presented on either one of the 2 cluttered backgrounds, but tested with the faces presented on a blank background. Results for this experiment showed poor performance. The results of experiments 3 and 4 suggest that in order for a cell to *learn* invariant responses to different transforms of a stimulus when it is presented during training in a cluttered background, some form of segmentation is required in order to separate the Figure (i.e., the stimulus or object) from the background. This segmentation might be performed using evidence in the visual scene about different depths, motions, colors, etc. of the object from its background. In the visual system, this might mean combining evidence represented in different cortical areas, and might be performed by cross-connections between cortical areas to enable such evidence to help separate the representations of objects from their backgrounds in the form-representing cortical areas.

Another mechanism that helps the operation of architectures such as VisNet and the primate visual system to learn about new objects in cluttered scenes is that the receptive fields of inferior temporal cortex neurons become much smaller when objects are seen against natural backgrounds (Sections 5.8.1 and 5.8). This will help greatly to learn about new objects that are being fixated, by reducing responsiveness to other features elsewhere in the scene.

Another mechanism that might help the learning of new objects in a natural scene is attention. An attentional mechanism might highlight the current stimulus being attended to and suppress the effects of background noise, providing a training representation of the object more like that which would be produced when it is presented against a blank background. The mechanisms that could implement such attentional processes are described elsewhere (Rolls, 2008b). If such attentional mechanisms do contribute to the development of view-invariance, then it follows that cells in the temporal cortex may only develop transform invariant responses to objects to which attention is directed.

Part of the reason for the poor performance in experiments 3 and 4 was probably that the stimuli were always presented against the same fixed background (for technical reasons), and thus the neurons learned about the background rather than the stimuli. Part of the difficulty that hierarchical multilayer competitive networks have with learning in cluttered environments may more generally be that without explicit segmentation of the stimulus from its background, at least some of the features that should be formed to encode the stimuli are not formed properly, because the neurons learn to respond to combinations of inputs which come partly from the stimulus, and partly from the background. To investigate this Stringer and Rolls (2000) performed experiment 5 in which layers 1–3 were pre-trained with stimuli to ensure that



good feature combination neurons for stimuli were available, and then allowed learning in only layer 4 when stimuli were presented in the cluttered backgrounds. Layer 4 was then trained in the usual way with the 7 faces presented against a cluttered background. The results showed that prior random exposure to the face stimuli led to much improved performance.

These results demonstrated that the problem of developing position-invariant neurons to stimuli occurring against cluttered backgrounds may be ameliorated by the prior existence of stimulus-tuned feature-detecting neurons in the early layers of the visual system, and that these feature-detecting neurons may be set up through previous exposure to the relevant class of objects. When tested in cluttered environments, the background clutter may of course activate some other neurons in the output layer, but at least the neurons that have learned to respond to the trained stimuli are activated. The result of this activity is sufficient for the activity in the output layer to be useful, in the sense that it can be read-off correctly by a pattern associator connected to the output layer. Indeed, Stringer and Rolls (2000) tested this by connecting a pattern associator to layer 4 of VisNet. The pattern associator had seven neurons, one for each face, and 1,024 inputs, one from each neuron in layer 4 of VisNet. The pattern associator learned when trained with a simple associative Hebb rule (equation (16)) to activate the correct output neuron whenever one of the faces was shown in any position in the uncluttered environment. This ability was shown to be dependent on invariant neurons for each stimulus in the output layer of VisNet, for the pattern associator could not be taught the task if VisNet had not been previously trained with a trace-learning rule to produce invariant representations. Then it was shown that exactly the correct neuron was activated when any of the faces was shown in any position with the cluttered background. This read-off by a pattern associator is exactly what we hypothesize takes place in the brain, in that the inferior temporal visual cortex (where neurons with invariant responses are found) projects to structures such as the orbitofrontal cortex and amygdala, where associations between the invariant visual representations and stimuli such as taste and touch are learned (Rolls and Treves, 1998; Rolls, 1999, 2005, 2008b, 2013; Rolls and Grabenhorst, 2008; Grabenhorst and Rolls, 2011). Thus testing whether the output of an architecture such as VisNet can be used effectively by a pattern associator is a very biologically relevant way to evaluate the performance of this class of architecture.

### **5.5.2. Learning invariant representations of an object with multiple objects in the scene and with cluttered backgrounds**

The results of the experiments just described suggest that in order for a neuron to *learn* invariant responses to different transforms of a stimulus when it is presented during training in a cluttered background, some form of segmentation is required in order to separate the figure (i.e., the stimulus or object) from the background. This segmentation might be performed using evidence in the visual scene about different depths, motions, colors, etc. of the object from its background. In the visual system, this might mean combining evidence represented in different cortical areas, and might be performed by cross-connections between cortical areas to enable such evidence to help separate the representations of

objects from their backgrounds in the form-representing cortical areas.

A second way in which training a feature hierarchy network in a cluttered natural scene may be facilitated follows from the finding that the receptive fields of inferior temporal cortex neurons shrink from in the order of 70° in diameter when only one object is present in a blank scene to much smaller values of as little as 5–10° close to the fovea in complex natural scenes (Rolls et al., 2003). The proposed mechanism for this is that if there is an object at the fovea, this object, because of the high-cortical magnification factor at the fovea, dominates the activity of neurons in the inferior temporal cortex by competitive interactions (Trappenberg et al., 2002; Deco and Rolls, 2004; see Section 5.8). This allows primarily the object at the fovea to be represented in the inferior temporal cortex, and, it is proposed, for learning to be about this object, and not about the other objects in a whole scene.

Third, top-down spatial attention (Deco and Rolls, 2004, 2005a; Rolls, 2008b) could bias the competition toward a region of visual space where the object to be learned is located.

Fourth, if object 1 is presented during training with other different objects present on different trials, then the competitive networks that are part of VisNet will learn to represent each object separately, because the features that are part of each object will be much more strongly associated together, than are those features with the other features present in the different objects seen on some trials during training (Stringer et al., 2007; Stringer and Rolls, 2008). It is a natural property of competitive networks that input features that co-occur very frequently together are allocated output neurons to represent the pattern as a result of the learning. Input features that do not co-occur frequently, may not have output neurons allocated to them. This principle may help feature hierarchy systems to learn representations of individual objects, even when other objects with some of the same features are present in the visual scene, but with different other objects on different trials. With this fundamental and interesting property of competitive networks, it has now become possible for VisNet to self-organize invariant representations of individual objects, even though each object is always presented during training with at least one other object present in the scene (Stringer et al., 2007; Stringer and Rolls, 2008). This has been extended to learning separate representations of face expression and face identity from the same set of images, depending on the statistics with which the images are presented (Tromans et al., 2011); and learning separate representations of independently rotating objects (Tromans et al., 2012).

### **5.5.3. VisNet simulations with partially occluded stimuli**

In this section we examine the recognition of partially occluded stimuli. Many artificial vision systems that perform object recognition typically search for specific markers in stimuli, and hence their performance may become fragile if key parts of a stimulus are occluded. However, in contrast we demonstrate that the model of invariance learning in the brain discussed here can continue to offer robust performance with this kind of problem, and that the model is able to correctly identify stimuli with considerable flexibility about what part of a stimulus is visible.

In these simulations (Stringer and Rolls, 2000), training and testing was performed with a blank background to avoid

confounding the two separate problems of occlusion and background clutter. In object recognition tasks, artificial vision systems may typically rely on being able to locate a small number of key markers on a stimulus in order to be able to identify it. This approach can become fragile when a number of these markers become obscured. In contrast, biological vision systems may generalize or complete from a partial input as a result of the use of distributed representations in neural networks, and this could lead to greater robustness in situations of partial occlusion.

In this experiment (6 of Stringer and Rolls, 2000), the network was first trained with the 7 face stimuli without occlusion, but during testing there were two options: either (i) the top halves of all the faces were occluded or (ii) the bottom halves of all the faces were occluded. Since VisNet was tested with either the top or bottom half of the stimuli no stimulus features were common to the two test options. This ensures that if performance is good with both options, the performance cannot be based on the use of a single feature to identify a stimulus. Results for this experiment are shown in **Figure 23**, with single and multiple cell information measures on the left and right, respectively. When compared with the performance without occlusion (Stringer and Rolls, 2000), **Figure 23** shows that there is only a modest drop in performance in the single cell information measures when the stimuli are partially occluded.

For both options (i) and (ii), even with partially occluded stimuli, a number of cells continue to respond maximally to one preferred stimulus in all locations, while responding minimally to all other stimuli. However, comparing results from options (i) and (ii) shows that the network performance is better when the bottom half of the faces is occluded. This is consistent with psychological results showing that face recognition is performed more easily when the top halves of faces are visible rather than the bottom halves (see Bruce, 1988). The top half of a face will generally contain salient features, e.g., eyes and hair, that are particularly helpful for recognition of the individual, and it is interesting that these simulations appear to further demonstrate this point. Furthermore, the multiple cell information measures confirm that performance is better with the upper half of the face

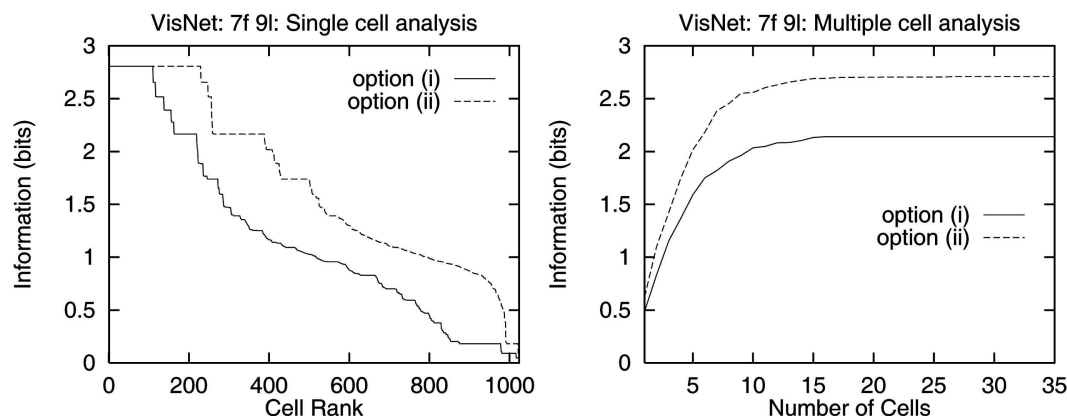
visible (option (ii)) than the lower half (option (i)). When the top halves of the faces are occluded the multiple cell information measure asymptotes to a suboptimal value reflecting the difficulty of discriminating between these more difficult images.

Thus this model of the ventral visual system offers robust performance with this kind of problem, and the model is able to correctly identify stimuli with considerable flexibility about what part of a stimulus is visible, because it is effectively using distributed representations and associative processing.

## 5.6. LEARNING 3D TRANSFORMS

In this section we describe investigations of Stringer and Rolls (2002) which show that trace-learning can in the VisNet architecture solve the problem of in-depth rotation invariant object recognition by developing representations of the transforms which features undergo when they are on the surfaces of 3D objects. Moreover, it is shown that having learned how features on 3D objects transform as the object is rotated in-depth, the network can correctly recognize novel 3D variations within a generic view of an object which is composed of previously learned feature combinations.

Rolls' hypothesis of how object recognition could be implemented in the brain postulates that trace rule learning helps invariant representations to form in two ways (Rolls, 1992, 1994, 1995, 2000). The first process enables associations to be learned between different generic 3D views of an object where there are different qualitative shape descriptors. One example of this would be the front and back views of an object, which might have very different shape descriptors. Another example is provided by considering how the shape descriptors typical of 3D shapes, such as Y vertices, arrow vertices, cusps, and ellipse shapes, alter when most 3D objects are rotated in 3 dimensions. At some point in the 3D rotation, there is a catastrophic rearrangement of the shape descriptors as a new generic view can be seen (Koenderink, 1990). An example of a catastrophic change to a new generic view is when a cup being viewed from slightly below is rotated so that one can see inside the cup from slightly above. The bottom surface disappears,



**FIGURE 23 | Effects of partial occlusion of a stimulus: numerical results for experiment 6 of Stringer and Rolls (2000), with the 7 faces presented on a blank background during both training and testing.** Training was performed with the whole face. However, during testing there

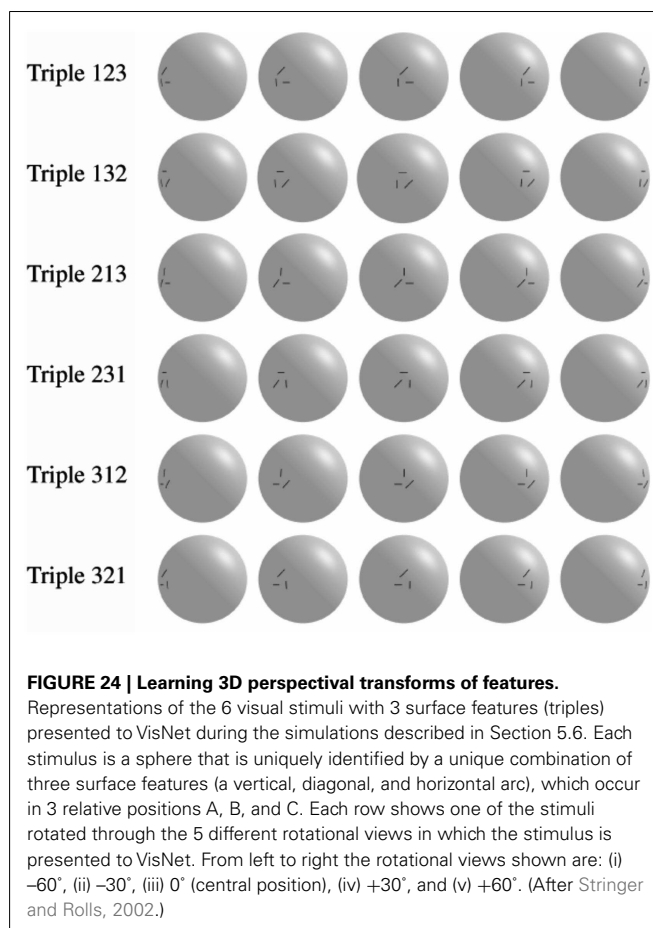
are two options: either (i) the top half of all the faces are occluded, or (ii) the bottom half of all the faces are occluded. On the left are single cell information measures, and on the right are multiple cell information measures.

the top surface of the cup changes from a cusp to an ellipse, and the inside of the cup with a whole set of new features comes into view. The second process is that within a generic view, as the object is rotated in-depth, there will be no catastrophic changes in the qualitative 3D shape descriptors, but instead the quantitative values of the shape descriptors alter. For example, while the cup is being rotated within a generic view seen from somewhat below, the curvature of the cusp forming the top boundary will alter, but the qualitative shape descriptor will remain a cusp. Trace-learning could help with both processes. That is, trace-learning could help to associate together qualitatively different sets of shape descriptors that occur close together in time, and describe, for example, the generically different views of a cup. Trace-learning could also help with the second process, and learn to associate together the different quantitative values of shape descriptors that typically occur when objects are rotated within a generic view.

We note that there is evidence that some neurons in the inferior temporal cortex may show the two types of 3D invariance. First Booth and Rolls (1998) showed that some inferior temporal cortex neurons can respond to different generic views of familiar 3D objects. Second, some neurons do generalize across quantitative changes in the values of 3D shape descriptors while faces (Hasselmo et al., 1989b) and objects (Logothetis et al., 1995; Tanaka, 1996) are rotated within-generic views. Indeed, Logothetis et al. (1995) showed that a few inferior temporal cortex neurons can generalize to novel (untrained) values of the quantitative shape descriptors typical of within-generic view object rotation.

In addition to the qualitative shape descriptor changes that occur catastrophically between different generic views of an object, and the quantitative changes of 3D shape descriptors that occur within a generic view, there is a third type of transform that must be learned for correct invariant recognition of 3D objects as they rotate in-depth. This third type of transform is that which occurs to the surface features on a 3D object as it transforms in-depth. The main aim here is to consider mechanisms that could enable neurons to learn this third type of transform, that is how to generalize correctly over the changes in the surface markings on 3D objects that are typically encountered as 3D objects rotate within a generic view. Examples of the types of perspectival transforms investigated are shown in **Figure 24**. Surface markings on the sphere that consist of combinations of three features in different spatial arrangements undergo characteristic transforms as the sphere is rotated from 0° to -60° and +60°. We investigated whether the class of architecture exemplified by VisNet, and the trace-learning rule, can learn about the transforms that surface features of 3D objects typically undergo during 3D rotation in such a way that the network generalizes across the change of the quantitative values of the surface features produced by the rotation, and yet still discriminates between the different objects (in this case spheres). In the cases being considered, each object is identified by surface markings that consist of a different spatial arrangement of the same three features (a horizontal, vertical, and diagonal line, which become arcs on the surface of the object).

We note that it has been suggested that the finding that neurons may offer some degree of 3D rotation invariance after training with a single view (or limited set of views) represents a challenge for



**FIGURE 24 | Learning 3D perspectival transforms of features.**

Representations of the 6 visual stimuli with 3 surface features (triples) presented to VisNet during the simulations described in Section 5.6. Each stimulus is a sphere that is uniquely identified by a unique combination of three surface features (a vertical, diagonal, and horizontal arc), which occur in 3 relative positions A, B, and C. Each row shows one of the stimuli rotated through the 5 different rotational views in which the stimulus is presented to VisNet. From left to right the rotational views shown are: (i) -60°, (ii) -30°, (iii) 0° (central position), (iv) +30°, and (v) +60°. (After Stringer and Rolls, 2002.)

existing trace-learning models, because these models assume that an initial exposure is required during learning to every transformation of the object to be recognized (Riesenhuber and Poggio, 1998). Stringer and Rolls (2002) showed as described here that this is not the case, and that such models can generalize to novel within-generic views of an object provided that the characteristic changes that the features show as objects are rotated have been learned previously for the sets of features when they are present in different objects.

Elliffe et al. (2002) demonstrated for a 2D system how the existence of translation-invariant representations of low-order feature combinations in the early layers of the visual system could allow correct stimulus identification in the output layer even when the stimulus was presented in a novel location where the stimulus had not previously occurred during learning. The proposal was that the low-order spatial-feature combination neurons in the early stages contain sufficient spatial information so that a particular combination of those low-order feature combination neurons specifies a unique object, even if the relative positions of the low-order feature combination neurons are not known because these neurons are somewhat translation-invariant (see Section 5.4.5). Stringer and Rolls (2002) extended this analysis to feature combinations on 3D objects, and indeed in their simulations described in this section therefore used surface markings for the 3D objects that consisted of triples of features.

The images used for training and testing VisNet were specially constructed for the purpose of demonstrating how the trace-learning paradigm might be further developed to give rise to neurons that are able to respond invariantly to novel within-generic view perspectives of an object, obtained by rotations in-depth up to 30° from any perspectives encountered during learning. The stimuli take the form of the surface feature combinations of 3-dimensional rotating spheres, with each image presented to VisNet's retina being a 2-dimensional projection of the surface features of one of the spheres. Each stimulus is uniquely identified by two or three surface features, where the surface features are (1) vertical, (2) diagonal, and (3) horizontal arcs, and where each feature may be centered at three different spatial positions, designated A, B, and C, as shown in **Figure 24**. The stimuli are thus defined in terms of what features are present and their precise spatial arrangement with respect to each other. We refer to the two and three feature stimuli as "pairs" and "triples," respectively. Individual stimuli are denoted by three numbers which refer to the individual features present in positions A, B and C, respectively. For example, a stimulus with positions A and C containing a vertical and diagonal bar, respectively, would be referred to as stimulus 102, where the 0 denotes no feature present in position B. In total there are 18 pairs (120, 130, 210, 230, 310, 320, 012, 013, 021, 023, 031, 032, 102, 103, 201, 203, 301, 302) and 6 triples (123, 132, 213, 231, 312, 321).

To train the network each stimulus was presented to VisNet in a randomized sequence of five orientations with respect to VisNet's input retina, where the different orientations are obtained from successive in-depth rotations of the stimulus through 30°. That is, each stimulus was presented to VisNet's retina from the following rotational views: (i) -60°, (ii) -30°, (iii) 0° (central position with surface features facing directly toward VisNet's retina), (iv) 30°, and (v) 60°. **Figure 24** shows representations of the 6 visual stimuli with 3 surface features (triples) presented to VisNet during the simulations. (For the actual simulations described here, the surface features and their deformations were what VisNet was trained and tested with, and the remaining blank surface of each sphere was set to the same gray-scale as the background.) Each row shows one of the stimuli rotated through the 5 different rotational views in which the stimulus is presented to VisNet. At each presentation the activation of individual neurons is calculated, then the neuronal firing rates are calculated, and then the synaptic weights are updated. Each time a stimulus has been presented in all the training orientations, a new stimulus is chosen at random and the process repeated. The presentation of all the stimuli through all 5 orientations constitutes 1 epoch of training. In this manner the network was trained one-layer at a time starting with layer 1 and finishing with layer 4. In the investigations described here, the numbers of training epochs for layers 1–4 were 50, 100, 100, and 75, respectively.

In experiment 1, VisNet was trained in two stages. In the first stage, the 18 feature pairs were used as input stimuli, with each stimulus being presented to VisNet's retina in sequences of five orientations as described above. However, during this stage, learning was only allowed to take place in layers 1 and 2. This led to the formation of neurons which responded to the feature pairs with some rotation invariance in layer 2. In the second stage, we

used the 6 feature triples as stimuli, with learning only allowed in layers 3 and 4. However, during this second training stage, the triples were only presented to VisNet's input retina in the first 4 orientations (i–iv). After the two stages of training were completed Stringer and Rolls (2002) examined whether the output layer of VisNet had formed top layer neurons that responded invariantly to the 6 triples when presented in all 5 orientations, not just the 4 in which the triples had been presented during training. To provide baseline results for comparison, the results from experiment 1 were compared with results from experiment 2 which involved no training in layers 1, 2 and 3, 4, with the synaptic weights left unchanged from their initial random values.

In **Figure 25** numerical results are given for the experiments described. On the left are the single cell information measures for all top (4th) layer neurons ranked in order of their invariance to the triples, while on the right are multiple cell information measures. To help to interpret these results we can compute the maximum single cell information measure according to

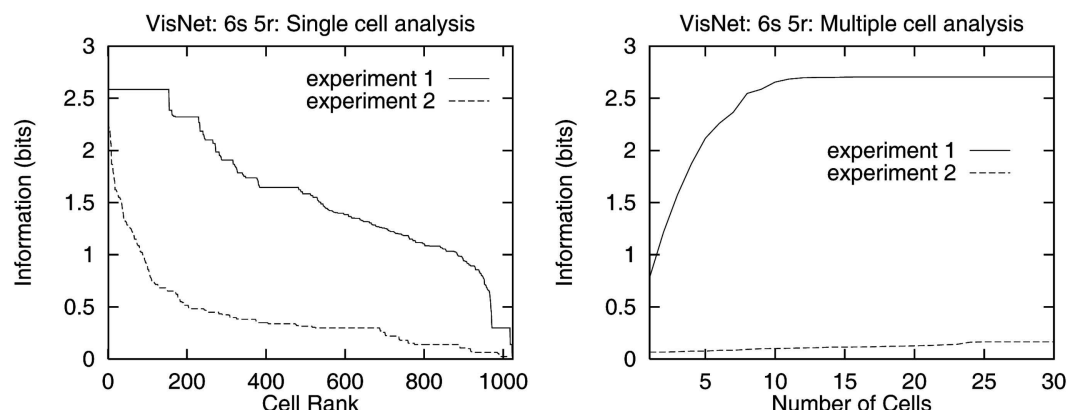
$$\text{Maximum single cell information} = \log_2(\text{Number of triples}), \quad (43)$$

where the number of triples is 6. This gives a maximum single cell information measure of 2.6 bits for these test cases. The information results from the experiment demonstrate that even with the triples presented to the network in only four of the five orientations during training, layer 4 is indeed capable of developing rotation invariant neurons that can discriminate effectively between the 6 different feature triples in all 5 orientations, that is with correct recognition from all five perspectives. In addition, the multiple cell information for the experiment reaches the maximal level of 2.6 bits, indicating that the network as a whole is capable of perfect discrimination between the 6 triples in any of the 5 orientations.

These results may be compared with the very poor baseline performance from the control experiment, where no learning was allowed before testing.

Stringer and Rolls (2002) also performed a control experiment to show that the network really had learned invariant representations specific to the kinds of 3D deformations undergone by the surface features as the objects rotated in-depth. In the control experiment the network was trained on "spheres" with non-deformed surface features; and then as predicted the network failed to operate correctly when it was tested with objects with the features present in the transformed way that they appear on the surface of a real 3D object.

Stringer and Rolls (2002) were thus able to show how trace-learning can form neurons that can respond invariantly to novel rotational within-generic view perspectives of an object, obtained by within-generic view 3D rotations up to 30° from any view encountered during learning. They were able to show in addition that this could occur for a novel view of an object which was not an interpolation from previously shown views. This was possible given that the low-order feature combination sets from which an object was composed had been learned about in early layers of VisNet previously. The within-generic view transform invariant object recognition described was achieved through the development of true 3-dimensional representations of objects based on



**FIGURE 25 | Learning 3D perspectival transforms of features.** Numerical results for experiments 1 and 2: on the left are single cell information measures, and on the right are multiple cell information measures. (After Stringer and Rolls, 2002.)

3-dimensional features and feature combinations, which, unlike 2-dimensional feature combinations, are invariant under moderate in-depth rotations of the object. Thus, in a sense, these rotation invariant representations encode a form of 3-dimensional knowledge with which to interpret the visual input from the real-world, that is able provide a basis for robust rotation invariant object recognition with novel perspectives. The particular finding in the work described here was that VisNet can learn how the surface features on 3D objects transform as the object is rotated in-depth, and can use knowledge of the characteristics of the transforms to perform 3D object recognition. The knowledge embodied in the network is knowledge of the 3D properties of objects, and in this sense assists the recognition of 3D objects seen from different views.

The process investigated by Stringer and Rolls (2002) will only allow invariant object recognition over moderate 3D object rotations, since rotating an object through a large angle may lead to a catastrophic change in the appearance of the object that requires the new qualitative 3D shape descriptors to be associated with those of the former view. In that case, invariant object recognition must rely on the first process referred to at the start of this Section (6) in order to associate together the different generic views of an object to produce view-invariant object identification. For that process, association of a few cardinal or generic views is likely to be sufficient (Koenderink, 1990). The process described in this section of learning how surface features transform is likely to make a major contribution to the within-generic view transform invariance of object identification and recognition.

## 5.7. CAPACITY OF THE ARCHITECTURE, AND INCORPORATION OF A TRACE RULE INTO A RECURRENT ARCHITECTURE WITH OBJECT ATTRACTORS

One issue that has not been considered extensively so far is the capacity of hierarchical feed-forward networks of the type exemplified by VisNet that are used for invariant object recognition. One approach to this issue is to note that VisNet operates in the general mode of a competitive network, and that the number of different stimuli that can be categorized by a competitive network is in the order of the number of neurons in the output layer (Rolls, 2008b).

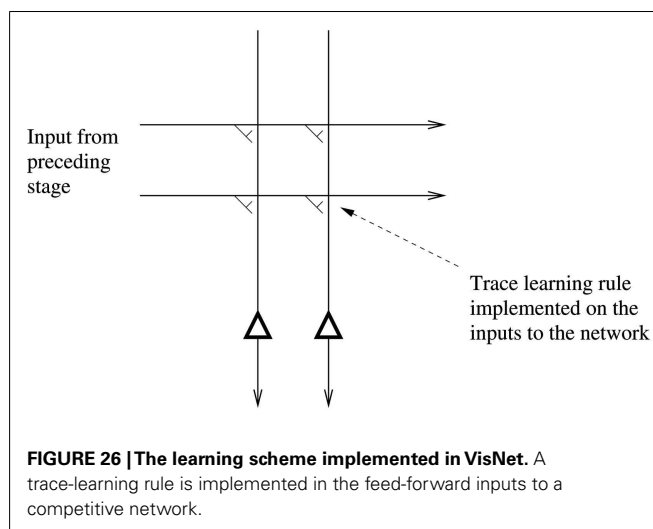
Given that the successive layers of the real visual system (V1, V2, V4, posterior inferior temporal cortex, anterior inferior temporal cortex) are of the same order of magnitude, VisNet is designed to work with the same number of neurons in each successive layer. (Of course the details are worth understanding further. V1 is, for example, somewhat larger than earlier layers, but on the other hand serves the dorsal as well as the ventral stream of visual cortical processing.) The hypothesis is that because of redundancies in the visual world, each layer of the system by its convergence and competitive categorization can capture sufficient of the statistics of the visual input at each stage to enable correct specification of the properties of the world that specify objects. For example, V1 does not compute all possible combinations of a few lateral geniculate inputs, but instead represents linear series of geniculate inputs to form edge-like and bar-like feature analyzers, which are the dominant arrangement of pixels found at the small scale in natural visual scenes. Thus the properties of the visual world at this stage can be captured by a small proportion of the total number of combinations that would be needed if the visual world were random. Similarly, at a later stage of processing, just a subset of all possible combinations of line or edge analyzers would be needed, partly because some combinations are much more frequent in the visual world, and partly because the coding because of convergence means that what is represented is for a larger area of visual space (that is, the receptive fields of the neurons are larger), which also leads to economy and limits what otherwise would be a combinatorial need for feature analyzers at later layers. The hypothesis thus is that the effects of redundancies in the input space of stimuli that result from the statistical properties of natural images (Field, 1987), together with the convergent architecture with competitive learning at each stage, produces a system that can perform invariant object recognition for large numbers of objects. Large in this case could be within one or two orders of magnitude of the number of neurons in any one-layer of the network (or cortical area in the brain). The extent to which this can be realized can be explored with simulations of the type implemented in VisNet, in which the network can be trained with natural images which therefore reflect fully the natural statistics of the stimuli presented to the real brain.

We should note that a rich variety of information in perceptual space may be represented by subtle differences in the distributed representation provided by the output of the visual system. At the same time, the actual number of different patterns that may be stored in, for example, a pattern associator connected to the output of the visual system is limited by the number of input connections per neuron from the output neurons of the visual system (Rolls, 2008b). One essential function performed by the ventral visual system is to provide an invariant representation which can be read by a pattern associator in such a way that if the pattern associator learns about one view of the object, then the visual system allows generalization to another view of the same object, because the same output neurons are activated by the different view. In the sense that any view can and must activate the same output neurons of the visual system (the input to the associative network), then we can say the invariance is made explicit in the representation. Making some properties of an input representation explicit in an output representation has a major function of enabling associative networks that use visual inputs in, for example, recognition, episodic memory, emotion and motivation to generalize correctly, that is invariantly with respect to image transforms that are all consistent with the same object in the world (Rolls and Treves, 1998).

Another approach to the issue of the capacity of networks that use trace learning to associate together different instances (e.g., views) of the same object is to reformulate the issue in the context of autoassociation (attractor) networks, where analytic approaches to the storage capacity of the network are well developed (Amit, 1989; Rolls and Treves, 1998; Rolls, 2008b). This approach to the storage capacity of networks that associate together different instantiations of an object to form invariant representations has been developed by Parga and Rolls (1998) and Elliffe et al. (2000), and is described next.

In this approach, the storage capacity of a *recurrent* network which performs, for example, view-invariant recognition of objects by associating together different views of the same object which tend to occur close together in time, was studied (Parga and Rolls, 1998; Elliffe et al., 2000). The architecture with which the invariance is computed is a little different to that described earlier. In the model of Rolls (1992, 1994, 1995), Wallis and Rolls (1997), Rolls and Milward (2000) Rolls and Stringer (2006), the post-synaptic memory trace enabled different afferents from the preceding stage to modify onto the same post-synaptic neuron (see Figure 26). In that model there were no recurrent connections between the neurons, although such connections were one way in which it was postulated the memory trace might be implemented, by simply keeping the representation of one view or aspect active until the next view appeared. Then an association would occur between representations that were active close together in time (within, e.g., 100–300 ms).

In the model developed by Parga and Rolls (1998) and Elliffe et al. (2000), there is a set of inputs with fixed synaptic weights to a network. The network itself is a recurrent network, with a trace rule incorporated in the recurrent collaterals (see Figure 27). When different views of the same object are presented close together in time, the recurrent collaterals learn using the trace rule that the different views are of the same object. After learning, presentation



of any of the views will cause the network to settle into an attractor that represents all the views of the object, that is which is a view-invariant representation of an object. (In this Section, the different exemplars of an object which need to be associated together are called views, for simplicity, but could at earlier stages of the hierarchy represent, for example, similar feature combinations (derived from the same object) in different positions in space.)

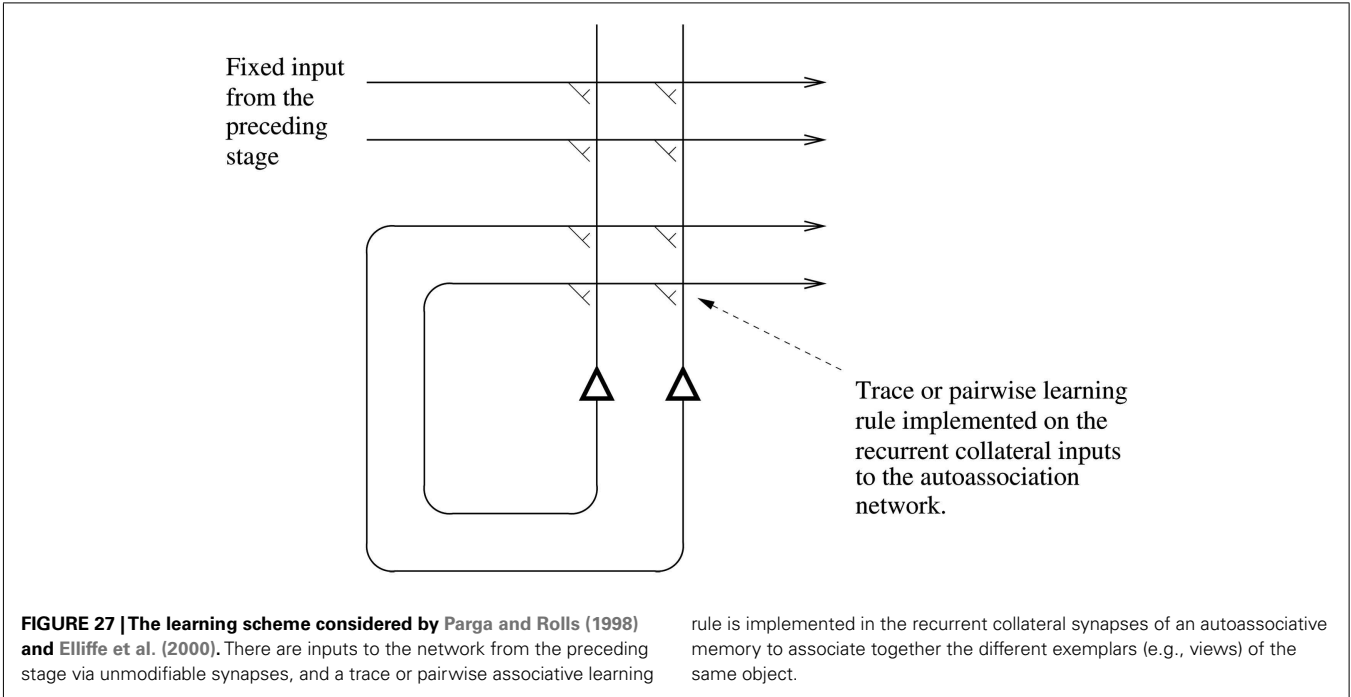
We envisage a set of neuronal operations which set up a synaptic weight matrix in the recurrent collaterals by associating together because of their closeness in time the different views of the same object.

In more detail Parga and Rolls (1998) considered two main approaches. First, one could store in a synaptic weight matrix the  $s$  views of an object. This consists of equally associating all the views to each other, including the association of each view with itself. Choosing in Figure 28 an example such that objects are defined in terms of five different views, this might produce (if each view produced firing of one neuron at a rate of 1) a block of  $5 \times 5$  pairs of views contributing to the synaptic efficacies each with value 1. Object 2 might produce another block of synapses of value 1 further along the diagonal, and symmetric about it. Each object or memory could then be thought of as a single attractor with a distributed representation involving five elements (each element representing a different view).

Then the capacity of the system in terms of the number  $P_o$  of objects that can be stored is just the number of separate attractors which can be stored in the network. For random fully distributed patterns this is as shown numerically by Hopfield (1982)

$$P_o = 0.14 C \quad (44)$$

where there are  $C$  inputs per neuron (and  $N = C$  neurons if the network is fully connected). Now the synaptic matrix envisaged here does not consist of random fully distributed binary elements, but instead we will assume has a sparseness  $a = s/N$ , where  $s$  is the number of views stored for each object, from any of which the whole representation of the object must be recognized. In this case, one can show (Gardner, 1988; Tsodyks and Feigel'man, 1988;



	$O_1v_1$	$O_1v_2$	$O_1v_3$	$O_1v_4$	$O_1v_5$	$O_2v_1$	$O_2v_2$	$O_2v_3$	$O_2v_4$	$O_2v_5$	.	.	.
$O_1v_1$	1	1	1	1	1								
$O_1v_2$	1	1	1	1	1								
$O_1v_3$	1	1	1	1	1								
$O_1v_4$	1	1	1	1	1								
$O_1v_5$	1	1	1	1	1								
$O_2v_1$						1	1	1	1	1			
$O_2v_2$						1	1	1	1	1			
$O_2v_3$						1	1	1	1	1			
$O_2v_4$						1	1	1	1	1			
$O_2v_5$						1	1	1	1	1			
.													
.													
.													

**FIGURE 28 | A schematic illustration of the first type of associations contributing to the synaptic matrix considered by Parga and Rolls (1998).** Object 1 ( $O_1$ ) has five views labeled  $v_1$  to  $v_5$ , etc. The matrix is formed by associating the pattern presented in the columns with itself, that is with the same pattern presented as rows.

Treves and Rolls, 1991) that the number of objects that can be stored and correctly retrieved is

$$P_o = \frac{k C}{a \ln (1/a)}$$

(45)

where  $C$  is the number of synapses on each neuron devoted to the recurrent collaterals from other neurons in the network, and  $k$  is a factor that depends weakly on the detailed structure of the rate distribution, on the connectivity pattern, etc., but is approximately in the order of 0.2–0.3. A problem with this proposal is that as the

number of views of each object increases to a large number (e.g.,  $>20$ ), the network will fail to retrieve correctly the internal representation of the object starting from any one view (which is only a fraction  $1/s$  of the length of the stored pattern that represents an object).

The second approach, taken by Parga and Rolls (1998) and Elliffe et al. (2000), is to consider the operation of the network when the associations between pairs of views can be described by a matrix that has the general form shown in **Figure 29**. Such an association matrix might be produced by different views of an object appearing after a given view with equal probability, and

	$O_1v_1$	$O_1v_2$	$O_1v_3$	$O_1v_4$	$O_1v_5$	$O_2v_1$	$O_2v_2$	$O_2v_3$	$O_2v_4$	$O_2v_5$	.	.	.
$O_1v_1$	1	$b$	$b$	$b$	$b$								
$O_1v_2$	$b$	1	$b$	$b$	$b$								
$O_1v_3$	$b$	$b$	1	$b$	$b$								
$O_1v_4$	$b$	$b$	$b$	1	$b$								
$O_1v_5$	$b$	$b$	$b$	$b$	1								
$O_2v_1$						1	$b$	$b$	$b$	$b$			
$O_2v_2$						$b$	1	$b$	$b$	$b$			
$O_2v_3$						$b$	$b$	1	$b$	$b$			
$O_2v_4$						$b$	$b$	$b$	1	$b$			
$O_2v_5$						$b$	$b$	$b$	$b$	1			
.													
.													
.													

**FIGURE 29 | A schematic illustration of the second and main type of associations contributing to the synaptic matrix considered by Parga and Rolls (1998) and Elliffe et al. (2000).** Object 1 ( $O_1$ ) has five views

labeled  $v_1$  to  $v_5$ , etc. The association of any one view with itself has strength 1, and of any one with another view of the same object has strength  $b$ .

synaptic modification occurring of the view with itself (giving rise to the diagonal term), and of any one view with that which immediately follows it.

The same weight matrix might be produced not only by pairwise association of successive views because the association rule allows for associations over the short-time scale of, e.g., 100–200 ms, but might also be produced if the synaptic trace had an exponentially decaying form over several hundred milliseconds, allowing associations with decaying strength between views separated by one or more intervening views. The existence of a regime, for values of the coupling parameter between pairs of views in a finite interval, such that the presentation of any of the views of one object leads to the same attractor regardless of the particular view chosen as a cue, is one of the issues treated by Parga and Rolls (1998) and Elliffe et al. (2000). A related problem also dealt with was the capacity of this type of synaptic matrix: how many objects can be stored and retrieved correctly in a view-invariant way? Parga and Rolls (1998) and Elliffe et al. (2000) showed that the number grows linearly with the number of recurrent collateral connections received by each neuron. Some of the groundwork for this approach was laid by the work of Amit and collaborators (Amit, 1989; Griniasty et al., 1993).

A variant of the second approach is to consider that the remaining entries in the matrix shown in **Figure 29** all have a small value. This would be produced by the fact that sometimes a view of one object would be followed by a view of a different object, when, for example, a large saccade was made, with no explicit resetting of the trace. On average, any one object would follow another rarely, and so the case is considered when all the remaining associations between pairs of views have a low value.

Parga and Rolls (1998) and Elliffe et al. (2000) were able to show that invariant object recognition is feasible in attractor neural networks in the way described. The system is able to store and retrieve in a view-invariant way an extensive number of objects, each defined by a finite set of views. What is implied by extensive is that the number of objects is proportional to the size of the network. The crucial factor that defines this size is the number of connections per neuron. In the case of the fully connected networks considered in this section, the size is thus proportional to

the number of neurons. To be particular, the number of objects that can be stored is  $0.081 N/5$ , when there are five views of each object. The number of objects is  $0.073 N/11$ , when there are eleven views of each object. This is an interesting result in network terms, in that  $s$  views each represented by an independent random set of active neurons can, in the network described, be present in the same “object” attraction basin. It is also an interesting result in neurophysiological terms, in that the number of objects that can be represented in this network scales linearly with the number of recurrent connections per neuron. That is, the number of objects  $P_o$  that can be stored is approximately

$$P_o = \frac{kC}{s} \quad (46)$$

where  $C$  is the number of synapses on each neuron devoted to the recurrent collaterals from other neurons in the network,  $s$  is the number of views of each object, and  $k$  is a factor that is in the region of 0.07–0.09 (Parga and Rolls, 1998).

Although the explicit numerical calculation was done for a rather small number of views for each object (up to 11), the basic result, that the network can support this kind of “object” phase, is expected to hold for any number of views (the only requirement being that it does not increase with the number of neurons). This is of course enough: once an object is defined by a set of views, when the network is presented with a somewhat different stimulus or a noisy version of one of them it will still be in the attraction basin of the object attractor.

Parga and Rolls (1998) thus showed that multiple (e.g., “view”) patterns could be within the basin of attraction of a shared (e.g., “object”) representation, and that the capacity of the system was proportional to the number of synapses per neuron divided by the number of views of each object.

Elliffe et al. (2000) extended the analysis of Parga and Rolls (1998) by showing that correct retrieval could occur where retrieval “view” cues were distorted; where there was some association between the views of different objects; and where there was only partial and indeed asymmetric connectivity provided by the associatively modified recurrent collateral connections in the network. The simulations also extended the analysis by showing that



the system can work well with sparse patterns, and indeed that the use of sparse patterns increases (as expected) the number of objects that can be stored in the network.

Taken together, the work described by Parga and Rolls (1998) and Elliffe et al. (2000) introduced the idea that the trace rule used to build invariant representations could be implemented in the recurrent collaterals of a neural network (as well as or as an alternative to its incorporation in the forward connections from one-layer to another incorporated in VisNet), and provided a precise analysis of the capacity of the network if it operated in this way. In the brain, it is likely that the recurrent collateral connections between cortical pyramidal cells in visual cortical areas do contribute to building invariant representations, in that if they are associatively modifiable, as seems likely, and because there is continuing firing for typically 100–300 ms after a stimulus has been shown, associations between different exemplars of the same object that occur together close in time would almost necessarily become built into the recurrent synaptic connections between pyramidal cells.

Invariant representation of faces in the context of attractor neural networks has also been discussed by Bartlett and Sejnowski (1997) in terms of a model where different views of faces are presented in a fixed sequence (Griniasty et al., 1993). This is not however the general situation; normally any pair of views can be seen consecutively and they will become associated. The model described by Parga and Rolls (1998) treats this more general situation.

I wish to note the different nature of the invariant object recognition problem studied here, and the paired associate learning task studied by Miyashita (1988), Miyashita and Chang (1988), and Sakai and Miyashita (1991). In the invariant object recognition case no particular learning protocol is required to produce an activity of the inferior temporal cortex cells responsible for invariant object recognition that is maintained for 300 ms. The learning can occur rapidly, and the learning occurs between stimuli (e.g., different views) which occur with no intervening delay. In the paired associate task, which had the aim of providing a model of semantic memory, the monkeys must learn to associate together two stimuli that are separated in time (by a number of seconds), and this type of learning can take weeks to train. During the delay period the sustained activity is rather low in the experiments, and thus the representation of the first stimulus that remains is weak, and can only poorly be associated with the second stimulus. However, formally the learning mechanism could be treated in the same way as that used by Parga and Rolls (1998) for invariant object recognition. The experimental difference is just that in the paired associate task used by Miyashita et al., it is the weak memory of the first stimulus that is associated with the second stimulus. In contrast, in the invariance learning, it would be the firing activity being produced by the first stimulus (not the weak memory of the first stimulus) that can be associated together. It is possible that the perirhinal cortex makes a useful contribution to invariant object recognition by providing a short-term memory that helps successive views of the same objects to become associated together (Buckley et al., 2001; Rolls et al., 2005a).

The mechanisms described here using an attractor network with a trace associative learning rule would apply most naturally

when a small number of representations need to be associated together to represent an object. One example is associating together what is seen when an object is viewed from different perspectives. Another example is scale, with respect to which neurons early in the visual system tolerate scale changes of approximately 1.5 octaves, so that the whole scale range could be covered by associating together a limited number of such representations (see Chapter 5 of Rolls and Deco (2002) and **Figure 1**). The mechanism would not be so suitable when a large number of different instances would need to be associated together to form an invariant representation of objects, as might be needed for translation invariance. For the latter, the standard model of VisNet with the associative trace-learning rule implemented in the feed-forward connections (or trained by continuous spatial transformation learning as described in Section 5.10) would be more appropriate. However, both types of mechanism, with the trace rule in the feed-forward or in recurrent collateral synapses, could contribute (separately or together) to achieve invariant representations. Part of the interest of the attractor approach described in this section is that it allows analytic investigation.

Another approach to training invariance is the purely associative mechanism continuous spatial transformation learning, described in Section 5.10. With this training procedure, the capacity is increased with respect to the number of training locations, with, for example, 169 training locations producing translation-invariant representations for two face stimuli (Perry et al., 2010). When we scaled up the  $32 \times 32$  VisNet used for most of the investigations described here to  $128 \times 128$  neurons per layer in the VisNetL specified in **Table 1**, it was demonstrated that perfect translation-invariant representations were produced over at least 1,089 locations for 5 objects. Thus the indications are that scaling up the size of VisNet does markedly improve performance, and in this case allows invariant representations for 5 objects across more than 1,000 locations to be trained with continuous spatial transformation learning (Perry et al., 2010).

It will be of interest in future research to investigate how the VisNet architecture, whether trained with a trace or purely associative rule, scales up with respect to capacity as the number of neurons in the system increases further. More distributed representations in the output layer may also help to increase the capacity. In recent investigations, we have been able to train VisNetL (i.e.,  $128 \times 128$  neurons in each layer, a  $256 \times 256$  input image, and 8 spatial frequencies for the Gabor filters as shown in **Table 4**) on a view-invariance learning problem, and have found good scaling up with respect to the original VisNet (i.e.,  $32 \times 32$  neurons in each layer, a  $64 \times 64$  input image, and 4 spatial frequencies for the filters). For example, VisNetL can learn with the trace rule perfect invariant representations of 32 objects each shown in 24 views (T. J. Webb and E. T. Rolls, recent observations). The objects were made with Blender 3D modeling software, so the image views generated were carefully controlled for lighting, background intensity, etc. When trained on half of these views for each object, with the other half used for cross-validation testing, the performance was reasonable at approximately 68% correct for the 32 objects, and having the full set of 8 spatial frequencies did improve performance.

## 5.8. VISION IN NATURAL SCENES – EFFECTS OF BACKGROUND VERSUS ATTENTION

Object-based attention refers to attention to an object. For example, in a visual search task the object might be specified as what should be searched for, and its location must be found. In spatial attention, a particular location in a scene is pre-cued, and the object at that location may need to be identified. Here we consider some of the neurophysiology of object selection and attention in the context of a feature hierarchy approach to invariant object recognition. The computational mechanisms of attention, including top-down biased competition, are described elsewhere (Rolls and Deco, 2002; Deco and Rolls, 2005b; Rolls, 2008b).

### 5.8.1. Neurophysiology of object selection and translation invariance in the inferior temporal visual cortex

Much of the neurophysiology, psychophysics, and modeling of attention has been with a small number, typically two, of objects in an otherwise blank scene. In this Section, I consider how attention operates in complex natural scenes, and in particular describe how the inferior temporal visual cortex operates to enable the selection of an object in a complex natural scene (see also Rolls and Deco, 2006). The inferior temporal visual cortex contains distributed and invariant representations of objects and faces (Rolls and Baylis, 1986; Hasselmo et al., 1989a; Tovee et al., 1994; Rolls and Tovee, 1995b; Rolls et al., 1997b; Booth and Rolls, 1998; Rolls, 2000, 2007a,b,c, 2011b; Rolls and Deco, 2002; Rolls and Treves, 2011).

To investigate how attention operates in complex natural scenes, and how information is passed from the inferior temporal cortex (IT) to other brain regions to enable stimuli to be selected from natural scenes for action, Rolls et al. (2003) analyzed the responses of inferior temporal cortex neurons to stimuli presented

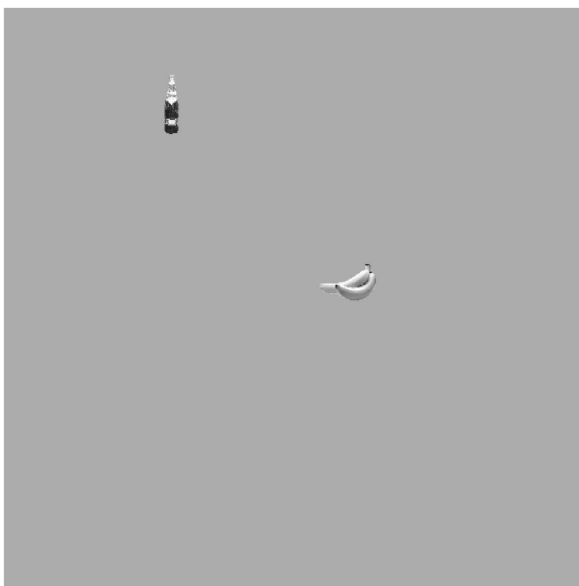
in complex natural backgrounds. The monkey had to search for two objects on a screen, and a touch of one object was rewarded with juice, and of another object was punished with saline (see **Figure 3** for a schematic illustration and **Figure 30** for a version of the display with examples of the stimuli shown to scale). Neuronal responses to the effective stimuli for the neurons were compared when the objects were presented in the natural scene or on a plain background. It was found that the overall response of the neuron to objects was hardly reduced when they were presented in natural scenes, and the selectivity of the neurons remained. However, the main finding was that the magnitudes of the responses of the neurons typically became much less in the real scene the further the monkey fixated in the scene away from the object (see **Figure 4**). A small receptive field size has also been found in inferior temporal cortex neurons when monkeys have been trained to discriminate closely spaced small visual stimuli (DiCarlo and Maunsell, 2003).

It is proposed that this reduced translation invariance in natural scenes helps an unambiguous representation of an object which may be the target for action to be passed to the brain regions that receive from the primate inferior temporal visual cortex. It helps with the binding problem, by reducing in natural scenes the effective receptive field of at least some inferior temporal cortex neurons to approximately the size of an object in the scene.

It is also found that in natural scenes, the effect of object-based attention on the response properties of inferior temporal cortex neurons is relatively small, as illustrated in **Figure 31** (Rolls et al., 2003).

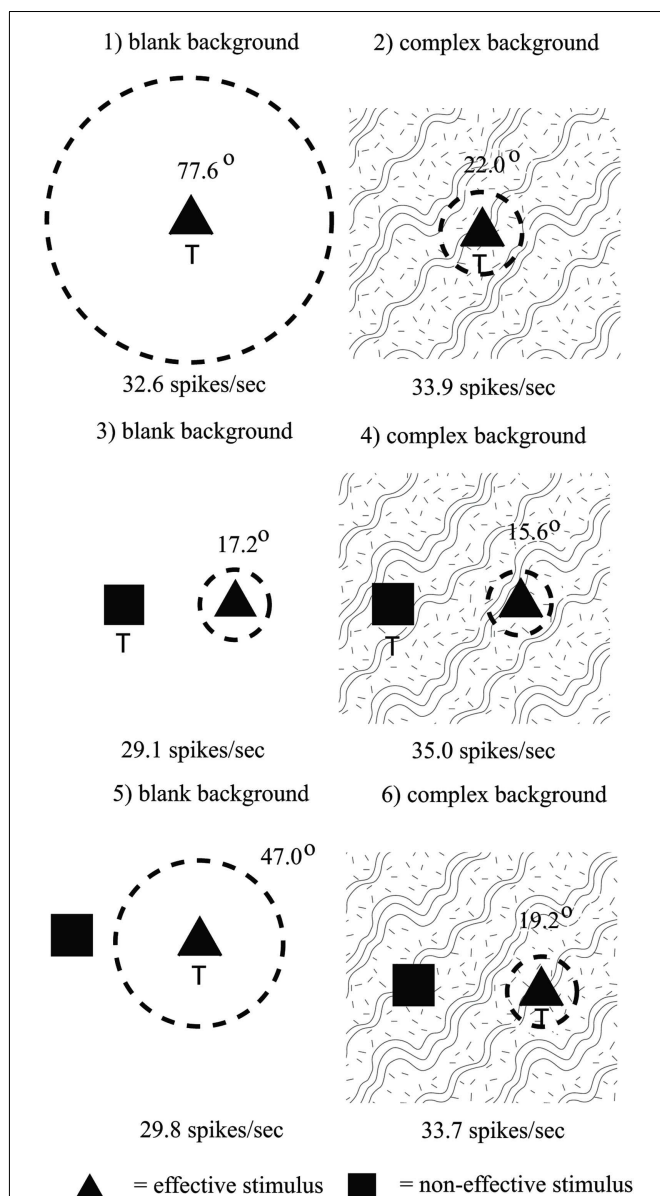
### 5.8.2. Attention and translation invariance in natural scenes – a computational account

The results summarized in **Figure 31** for 5° stimuli show that the receptive fields were large (77.6°) with a single stimulus in a blank



**FIGURE 30 | The visual search task.** The monkey had to search for and touch an object (in this case a banana) when shown in a complex natural scene, or when shown on a plain background. In each case a second

object is present (a bottle) which the monkey must not touch. The stimuli are shown to scale. The screen subtended 70° × 55° (After Rolls et al., 2003.)



**FIGURE 31 | Summary of the receptive field sizes of inferior temporal cortex neurons to a 5° effective stimulus presented in either a blank background (blank screen) or in a natural scene (complex background).**

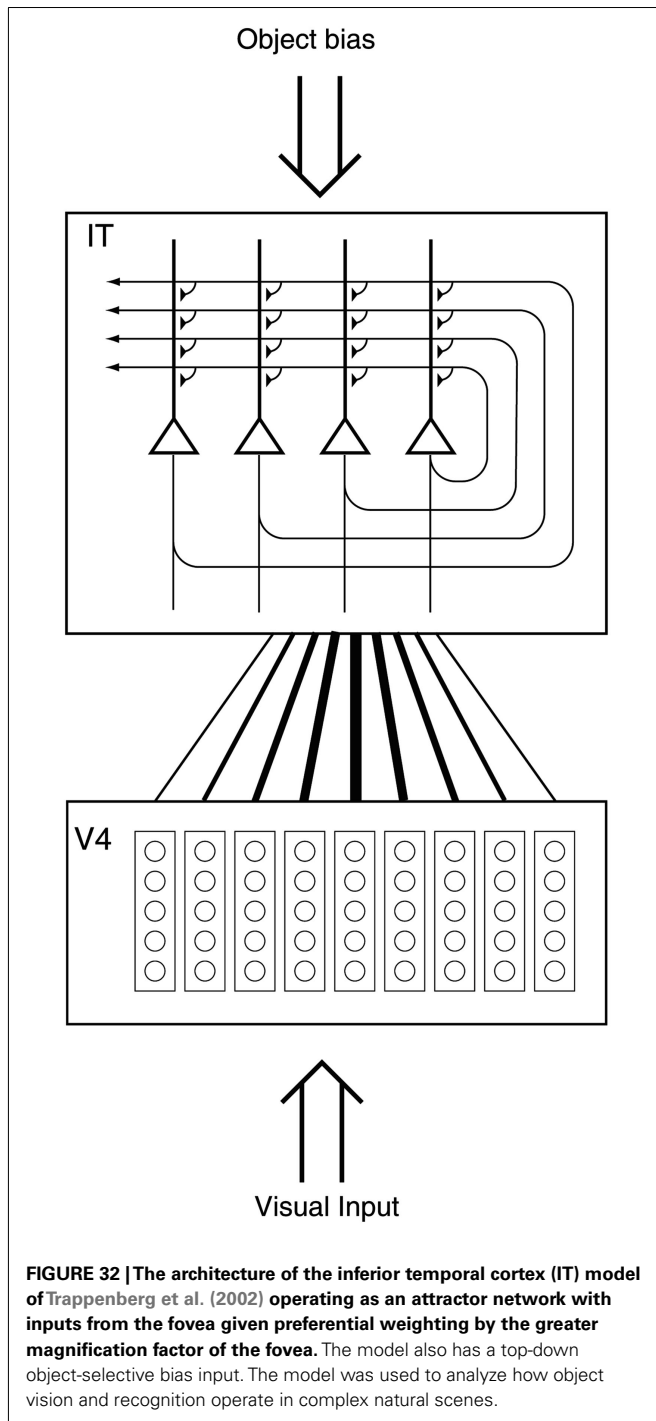
The stimulus that was a target for action in the different experimental conditions is marked by T. When the target stimulus was touched, a reward was obtained. The mean receptive field diameter of the population of neurons analyzed, and the mean firing rate in spikes/s, is shown. The stimuli subtended  $5^\circ \times 3.5^\circ$  at the retina, and occurred on each trial in a random position in the  $70^\circ \times 55^\circ$  screen. The dashed circle is proportional to the receptive field size. Top row: responses with one visual stimulus in a blank (left) or complex (right) background. Middle row: responses with two stimuli, when the effective stimulus was not the target of the visual search. Bottom row: responses with two stimuli, when the effective stimulus was the target of the visual search. (After Rolls et al., 2003.)

background (top left), and were greatly reduced in size (to  $22.0^\circ$ ) when presented in a complex natural scene (top right). The results also show that there was little difference in receptive field size or

firing rate in the complex background when the effective stimulus was selected for action (bottom right,  $19.2^\circ$ ), and when it was not (middle right,  $15.6^\circ$ ; Rolls et al., 2003). (For comparison, the effects of attention against a blank background were much larger, with the receptive field increasing from  $17.2^\circ$  to  $47.0^\circ$  as a result of object-based attention, as shown in Figure 31, left middle and bottom.)

Trappenberg et al. (2002) have suggested what underlying mechanisms could account for these findings, and simulated a model to test the ideas. The model utilizes an attractor network representing the inferior temporal visual cortex (implemented by the recurrent connections between inferior temporal cortex neurons), and a neural input layer with several retinotopically organized modules representing the visual scene in an earlier visual cortical area such as V4 (see Figure 32). The attractor network aspect of the model produces the property that the receptive fields of IT neurons can be large in blank scenes by enabling a weak input in the periphery of the visual field to act as a retrieval cue for the object attractor. On the other hand, when the object is shown in a complex background, the object closest to the fovea tends to act as the retrieval cue for the attractor, because the fovea is given increased weight in activating the IT module because the magnitude of the input activity from objects at the fovea is greatest due to the higher magnification factor of the fovea incorporated into the model. This results in smaller receptive fields of IT neurons in complex scenes, because the object tends to need to be close to the fovea to trigger the attractor into the state representing that object. (In other words, if the object is far from the fovea, then it will not trigger neurons in IT which represent it, because neurons in IT are preferentially being activated by another object at the fovea.) This may be described as an attractor model in which the competition for which attractor state is retrieved is weighted toward objects at the fovea.

Attentional top-down object-based inputs can bias the competition implemented in this attractor model, but have relatively minor effects (in, for example, increasing receptive field size) when they are applied in a complex natural scene, as then as usual the stronger forward inputs dominate the states reached. In this network, the recurrent collateral connections may be thought of as implementing constraints between the different inputs present, to help arrive at firing in the network which best meets the constraints. In this scenario, the preferential weighting of objects close to the fovea is a useful principle in enabling the system to provide useful output. The attentional object biasing effect is much more marked in a blank scene, or a scene with only two objects present at similar distances from the fovea, which are conditions in which attentional effects have frequently been examined. The results of the investigation (Trappenberg et al., 2002) thus suggest that top-down attention may be a much more limited phenomenon in complex, natural, scenes than in reduced displays with one or two objects present. The results also suggest that the alternative principle, of providing strong weight to whatever is close to the fovea, is an important principle governing the operation of the inferior temporal visual cortex, and in general of the output of the visual system in natural environments. This principle of operation is very important in interfacing the visual system to action systems,



because the effective stimulus in making inferior temporal cortex neurons fire is in natural scenes usually on or close to the fovea. This means that the spatial coordinates of where the object is in the scene do not have to be represented in the inferior temporal visual cortex, nor passed from it to the action selection system, as the latter can assume that the object making IT neurons fire is close to the fovea in natural scenes.

There may of course be in addition a mechanism for object selection that takes into account the locus of covert attention when

actions are made to locations not being looked at. However, the simulations described in this section suggest that in any case covert attention is likely to be a much less significant influence on visual processing in natural scenes than in reduced scenes with one or two objects present.

Given these points, one might question why inferior temporal cortex neurons can have such large receptive fields, which show translation invariance. At least part of the answer to this may be that inferior temporal cortex neurons must have the capability to be large if they are to deal with large objects. A V1 neuron, with its small receptive field, simply could not receive input from all the features necessary to define an object. On the other hand, inferior temporal cortex neurons may be able to adjust their size to approximately the size of objects, using in part the interactive effects involved in attention (Rolls, 2008b), and need the capability for translation invariance because the actual relative positions of the features of an object could be at different relative positions in the scene. For example, a car can be recognized whichever way it is viewed, so that the parts (such as the bonnet or hood) must be identifiable as parts wherever they happen to be in the image, though of course the parts themselves also have to be in the correct relative positions, as allowed for by the hierarchical feature analysis architecture described in this paper.

Some details of the simulations follow. Each independent module within “V4” in **Figure 32** represents a small part of the visual field and receives input from earlier visual areas represented by an input vector for each possible location which is unique for each object. Each module was 6° in width, matching the size of the objects presented to the network. For the simulations Trappenberg et al. (2002) chose binary random input vectors representing objects with  $N^{V4}a^{V4}$  components set to ones and the remaining  $N^{V4}(1 - a^{V4})$  components set to zeros.  $N^{V4}$  is the number of nodes in each module and  $a^{V4}$  is the sparseness of the representation which was set to be  $a^{V4} = 0.2$  in the simulations.

The structure labeled “IT” represents areas of visual association cortex such as the inferior temporal visual cortex and cortex in the anterior part of the superior temporal sulcus in which neurons provide distributed representations of faces and objects (Booth and Rolls, 1998; Rolls, 2000). Nodes in this structure are governed by leaky integrator dynamics with time constant  $\tau$

$$\tau \frac{dh_i^{IT}(t)}{dt} = -h_i^{IT}(t) + \sum_j \left( w_{ij}^{IT} - c^{IT} \right) y_j^{IT}(t) + \sum_k w_{ik}^{IT-V4} y_k^{V4}(t) + k^{IT-BIAS} I_i^{OBJ}. \quad (47)$$

The firing rate  $y_i^{IT}$  of the  $i$ th node is determined by a sigmoidal function from the activation  $h_i^{IT}$  as follows

$$y_i^{IT}(t) = \frac{1}{1 + \exp[-2\beta(h_i^{IT}(t) - \alpha)]}, \quad (48)$$

where the parameters  $\beta = 1$  and  $\alpha = 1$  represent the gain and the bias, respectively.

The recognition functionality of this structure is modeled as an attractor neural network (ANN) with trained memories indexed

by  $\mu$  representing particular objects. The memories are formed through Hebbian learning on sparse patterns,

$$w_{ij}^{\text{IT}} = k^{\text{IT}} \sum_{\mu} (\xi_i^{\mu} - a^{\text{IT}}) (\xi_j^{\mu} - a^{\text{IT}}), \quad (49)$$

where  $k^{\text{IT}}$  (set to 1 in the simulations) is a normalization constant that depends on the learning rate,  $a^{\text{IT}} = 0.2$  is the sparseness of the training pattern in IT, and  $\xi_i^{\mu}$  are the components of the pattern used to train the network. The constant  $c^{\text{IT}}$  in equation (47) represents the strength of the activity-dependent global inhibition simulating the effects of inhibitory interneurons. The external “top-down” input vector  $I^{\text{OBJ}}$  produces object-selective inputs, which are used as the attentional drive when a visual search task is simulated. The strength of this object bias is modulated by the value of  $k^{\text{IT-BIAS}}$  in equation (47).

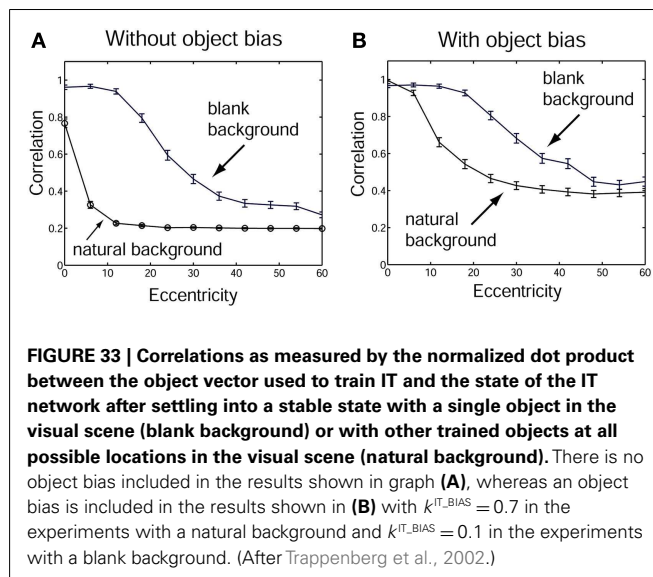
The weights  $w_{ij}^{\text{IT-V4}}$  between the V4 nodes and IT nodes were trained by Hebbian learning of the form

$$w_{ij}^{\text{IT-V4}} = k^{\text{IT-V4}}(k) \sum_{\mu} (\xi_i^{\mu} - a^{\text{V4}}) (\xi_j^{\mu} - a^{\text{IT}}). \quad (50)$$

to produce object representations in IT based on inputs in V4. The normalizing modulation factor  $k^{\text{IT-V4}}(k)$  allows the gain of inputs to be modulated as a function of their distance from the fovea, and depends on the module  $k$  to which the presynaptic node belongs. The model supports translation-invariant object recognition of a single object in the visual field if the normalization factor is the same for each module and the model is trained with the objects placed at every possible location in the visual field. The translation invariance of the weight vectors between each “V4” module and the IT nodes is however explicitly modulated in the model by the module-dependent modulation factor  $k^{\text{IT-V4}}(k)$  as indicated in **Figure 32** by the width of the lines connecting V4 with IT. The strength of the foveal V4 module is strongest, and the strength decreases for modules representing increasing eccentricity. The form of this modulation factor was derived from the parameterization of the cortical magnification factors given by Dow et al. (1981).

To study the ability of the model to recognize trained objects at various locations relative to the fovea the system was trained on a set of objects. The network was then tested with distorted versions of the objects, and the “correlation” between the target object and the final state of the attractor network was taken as a measure of the performance. The correlation was estimated from the normalized dot product between the target object vector that was used during training the IT network, and the state of the IT network after a fixed amount of time sufficient for the network to settle into a stable state. The objects were always presented on backgrounds with some noise (introduced by flipping 2% of the bits in the scene which were not the test stimulus) in order to utilize the properties of the attractor network, and because the input to IT will inevitably be noisy under normal conditions of operation.

In the first simulation only one object was present in the visual scene in a plain (blank) background at different eccentricities from the fovea. As shown in **Figure 33A** by the line labeled “blank background,” the receptive fields of the neurons were very large. The



value of the object bias  $k^{\text{IT-BIAS}}$  was set to 0 in these simulations. Good object retrieval (indicated by large correlations) was found even when the object was far from the fovea, indicating large IT receptive fields with a blank background. The reason that any drop is seen in performance as a function of eccentricity is because flipping 2% of the bits outside the object introduces some noise into the recall process. This demonstrates that the attractor dynamics can support translation-invariant object recognition even though the translation-invariant weight vectors between V4 and IT are explicitly modulated by the modulation factor  $k^{\text{IT-V4}}$  derived from the cortical magnification factor.

In a second simulation individual objects were placed at all possible locations in a natural and cluttered visual scene. The resulting correlations between the target pattern and the asymptotic IT state are shown in **Figure 33A** with the line labeled “natural background.” Many objects in the visual scene are now competing for recognition by the attractor network, and the objects around the foveal position are enhanced through the modulation factor derived from the cortical magnification factor. This results in a much smaller size of the receptive field of IT neurons when measured with objects in natural backgrounds.

In addition to this major effect of the background on the size of the receptive field, which parallels and may account for the physiological findings outlined above and in Section 5.8.1, there is also a dependence of the size of the receptive fields on the level of object bias provided to the IT network. Examples are shown in **Figure 33B** where an object bias was used. The object bias biases the IT network toward the expected object with a strength determined by the value of  $k^{\text{IT-BIAS}}$ , and has the effect of increasing the size of the receptive fields in both blank and natural backgrounds (see **Figure 33B** compared to **Figure 33A**). This models the effect found neurophysiologically (Rolls et al., 2003).

Some of the conclusions are as follows (Trappenberg et al., 2002). When single objects are shown in a scene with a blank background, the attractor network helps neurons to respond to an object with large eccentricities of this object relative to the fovea

of the agent. When the object is presented in a natural scene, other neurons in the inferior temporal cortex become activated by the other effective stimuli present in the visual field, and these forward inputs decrease the response of the network to the target stimulus by a competitive process. The results found fit well with the neurophysiological data, in that IT operates with almost complete translation invariance when there is only one object in the scene, and reduces the receptive field size of its neurons when the object is presented in a cluttered environment. The model described here provides an explanation of the responses of real IT neurons in natural scenes.

In natural scenes, the model is able to account for the neurophysiological data that the IT neuronal responses are larger when the object is close to the fovea, by virtue of fact that objects close to the fovea are weighted by the cortical magnification factor related modulation  $k^{IT-V4}$ .

The model accounts for the larger receptive field sizes from the fovea of IT neurons in natural backgrounds if the target is the object being selected compared to when it is not selected (Rolls et al., 2003). The model accounts for this by an effect of top-down bias which simply biases the neurons toward particular objects compensating for their decreasing inputs produced by the decreasing magnification factor modulation with increasing distance from the fovea. Such object-based attention signals could originate in the prefrontal cortex and could provide the object bias for the inferior temporal visual cortex (Renart et al., 2000; Rolls, 2008b).

Important properties of the architecture for obtaining the results just described are the high magnification factor at the fovea and the competition between the effects of different inputs, implemented in the above simulation by the competition inherent in an attractor network.

We have also been able to obtain similar results in a hierarchical feed-forward network where each layer operates as a competitive network (Deco and Rolls, 2004). This network thus captures many of the properties of our hierarchical model of invariant object recognition (Rolls, 1992; Wallis and Rolls, 1997; Rolls and Milward, 2000; Stringer and Rolls, 2000, 2002; Rolls and Stringer, 2001, 2006, 2007; Elliffe et al., 2002; Rolls and Deco, 2002; Stringer et al., 2006), but incorporates in addition a foveal magnification factor and top-down projections with a dorsal visual stream so that attentional effects can be studied, as shown in **Figure 34**.

Deco and Rolls (2004) trained the network shown in **Figure 34** with two objects, and used the trace-learning rule (Wallis and Rolls, 1997; Rolls and Milward, 2000) in order to achieve translation invariance. In a first experiment we placed only one object on the retina at different distances from the fovea (i.e., different eccentricities relative to the fovea). This corresponds to the blank background condition. In a second experiment, we also placed the object at different eccentricities relative to the fovea, but on a cluttered natural background. Larger receptive fields were found with the blank as compared to the cluttered natural background.

Deco and Rolls (2004) also studied the influence of object-based attentional top-down bias on the effective size of the receptive field of an inferior temporal cortex neuron for the case of an object in a blank or a cluttered background. To do this, they repeated the two simulations but now considered a non-zero top-down bias coming from prefrontal area 46v and impinging on

the inferior temporal cortex neuron specific for the object tested. When no attentional object bias was introduced, a shrinkage of the receptive field size was observed in the complex vs the blank background. When attentional object bias was introduced, the shrinkage of the receptive field due to the complex background was somewhat reduced. This is consistent with the neurophysiological results (Rolls et al., 2003). In the framework of the model (Deco and Rolls, 2004), the reduction of the shrinkage of the receptive field is due to the biasing of the competition in the inferior temporal cortex layer in favor of the specific IT neuron tested, so that it shows more translation invariance (i.e., a slightly larger receptive field). The increase of the receptive field size of an IT neuron, although small, produced by the external top-down attentional bias offers a mechanism for facilitation of the search for specific objects in complex natural scenes (Rolls, 2008b).

I note that it is possible that a “spotlight of attention” (Desimone and Duncan, 1995) can be moved covertly away from the fovea (Rolls, 2008b). However, at least during normal visual search tasks in natural scenes, the neurons are sensitive to the object at which the monkey is looking, that is primarily to the object that is on the fovea, as shown by Rolls et al. (2003) and Aggelopoulos and Rolls (2005), and described in Sections 1 and 9.

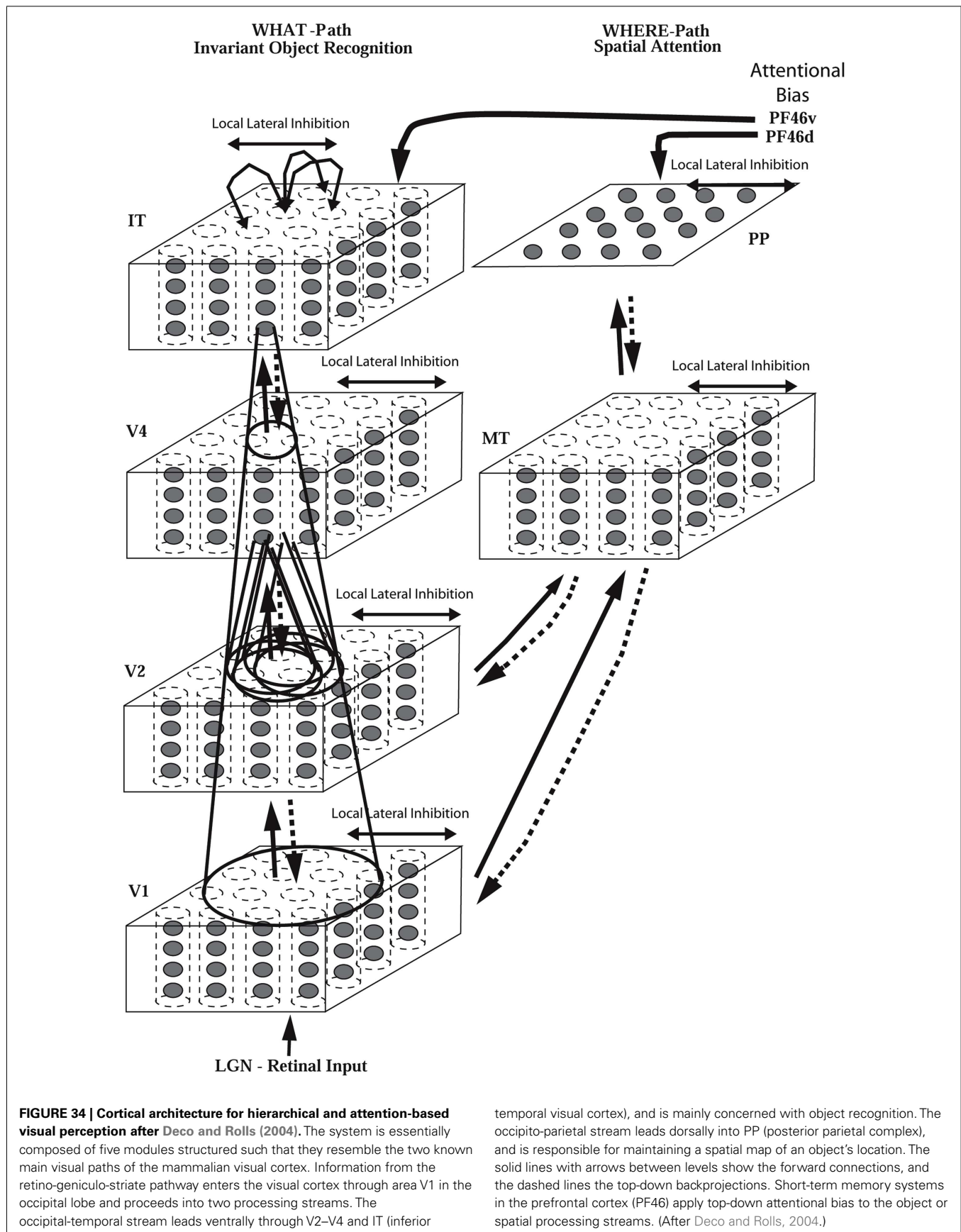
## 5.9. THE REPRESENTATION OF MULTIPLE OBJECTS IN A SCENE

When objects have distributed representations, there is a problem of how multiple objects (whether the same or different) can be represented in a scene, because the distributed representations overlap, and it may not be possible to determine whether one has an amalgam of several objects, or a new object (Mozzer, 1991), or multiple instances of the same object, let alone the relative spatial positions of the objects in a scene. Yet humans can determine the relative spatial locations of objects in a scene even in short presentation times without eye movements (Biederman, 1972; and this has been held to involve some spotlight of attention). Aggelopoulos and Rolls (2005) analyzed this issue by recording from single inferior temporal cortex neurons with five objects simultaneously present in the receptive field. They found that although all the neurons responded to their effective stimulus when it was at the fovea, some could also respond to their effective stimulus when it was in some but not other parafoveal positions 10° from the fovea. An example of such a neuron is shown in **Figure 35**. The asymmetry is much more evident in a scene with 5 images present (**Figure 35A**) than when only one image is shown on an otherwise blank screen (**Figure 35B**). Competition between different stimuli in the receptive field thus reveals the asymmetry in the receptive field of inferior temporal visual cortex neurons.

The asymmetry provides a way of encoding the position of multiple objects in a scene. Depending on which asymmetric neurons are firing, the population of neurons provides information to the next processing stage not only about which image is present at or close to the fovea, but where it is with respect to the fovea.

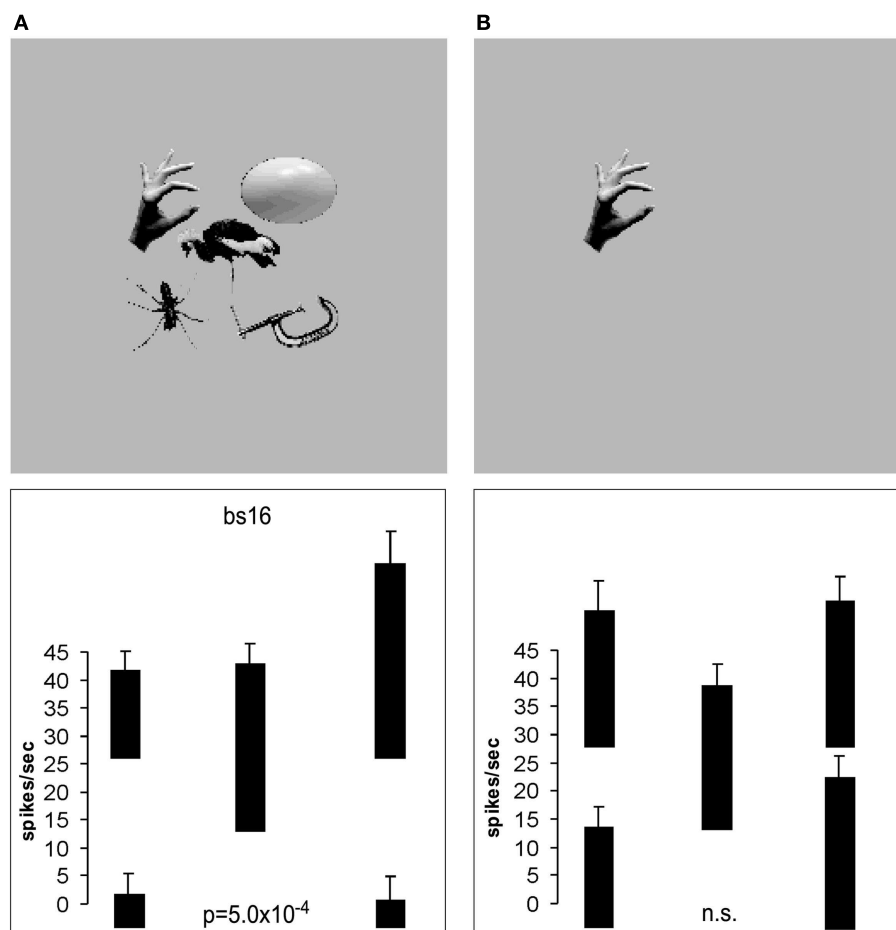
Simulations with VisNet with an added layer to simulate hippocampal scene memory have demonstrated that receptive field asymmetry appears when multiple objects are simultaneously present because of the probabilistic connectivity from the preceding stage which introduces asymmetry, which becomes revealed





**FIGURE 34 | Cortical architecture for hierarchical and attention-based visual perception after Deco and Rolls (2004).** The system is essentially composed of five modules structured such that they resemble the two known main visual paths of the mammalian visual cortex. Information from the retino-geniculo-striate pathway enters the visual cortex through area V1 in the occipital lobe and proceeds into two processing streams. The occipito-temporal stream leads ventrally through V2–V4 and IT (inferior

temporal visual cortex), and is mainly concerned with object recognition. The occipito-parietal stream leads dorsally into PP (posterior parietal complex), and is responsible for maintaining a spatial map of an object's location. The solid lines with arrows between levels show the forward connections, and the dashed lines the top-down backprojections. Short-term memory systems in the prefrontal cortex (PF46) apply top-down attentional bias to the object or spatial processing streams. (After Deco and Rolls, 2004.)



**FIGURE 35 | (A)** The responses (firing rate with the spontaneous rate subtracted, means  $\pm$  sem) of an inferior temporal cortex neuron when tested with 5 stimuli simultaneously present in the close ( $10^\circ$ ) configuration with the parafoveal stimuli located  $10^\circ$  from the fovea. **(B)** The responses of the same neuron when only the effective

stimulus was presented in each position. The firing rate for each position is that when the effective stimulus (in this case the hand) for the neuron was in that position. The p value is that from the ANOVA calculated over the four parafoveal positions. (After Aggelopoulos and Rolls, 2005.)

by the enhanced lateral inhibition when multiple objects are presented simultaneously (Rolls et al., 2008).

The information in the inferior temporal visual cortex is provided by neurons that have firing rates that reflect the relevant information, and stimulus-dependent synchrony is not necessary (Aggelopoulos and Rolls, 2005). Top-down attentional biasing input could thus, by biasing the appropriate neurons, facilitate bottom-up information about objects without any need to alter the time relations between the firing of different neurons. The exact position of the object with respect to the fovea, and effectively thus its spatial position relative to other objects in the scene, would then be made evident by the subset of asymmetric neurons firing.

This is thus the solution that these experiments (Aggelopoulos and Rolls, 2005; Rolls et al., 2008) indicate is used for the representation of multiple objects in a scene, an issue that has previously been difficult to account for in neural systems with distributed representations (Mozier, 1991) and for which “attention” has been a proposed solution.

The learning of invariant representations of objects when multiple objects are present in a scene is considered in Section 5.5.2.

#### 5.10. LEARNING INVARIANT REPRESENTATIONS USING SPATIAL CONTINUITY: CONTINUOUS SPATIAL TRANSFORMATION LEARNING

The temporal continuity typical of objects has been used in an associative learning rule with a short-term memory trace to help build invariant object representations in the networks described previously in this paper. Stringer et al. (2006) showed that spatial continuity can also provide a basis for helping a system to self-organize invariant representations. They introduced a new learning paradigm “continuous spatial transformation (CT) learning” which operates by mapping spatially similar input patterns to the same post-synaptic neurons in a competitive learning system. As the inputs move through the space of possible continuous transforms (e.g., translation, rotation, etc.), the active synapses are modified onto the set of post-synaptic neurons. Because other transforms of the same stimulus overlap with previously learned



exemplars, a common set of post-synaptic neurons is activated by the new transforms, and learning of the new active inputs onto the same post-synaptic neurons is facilitated.

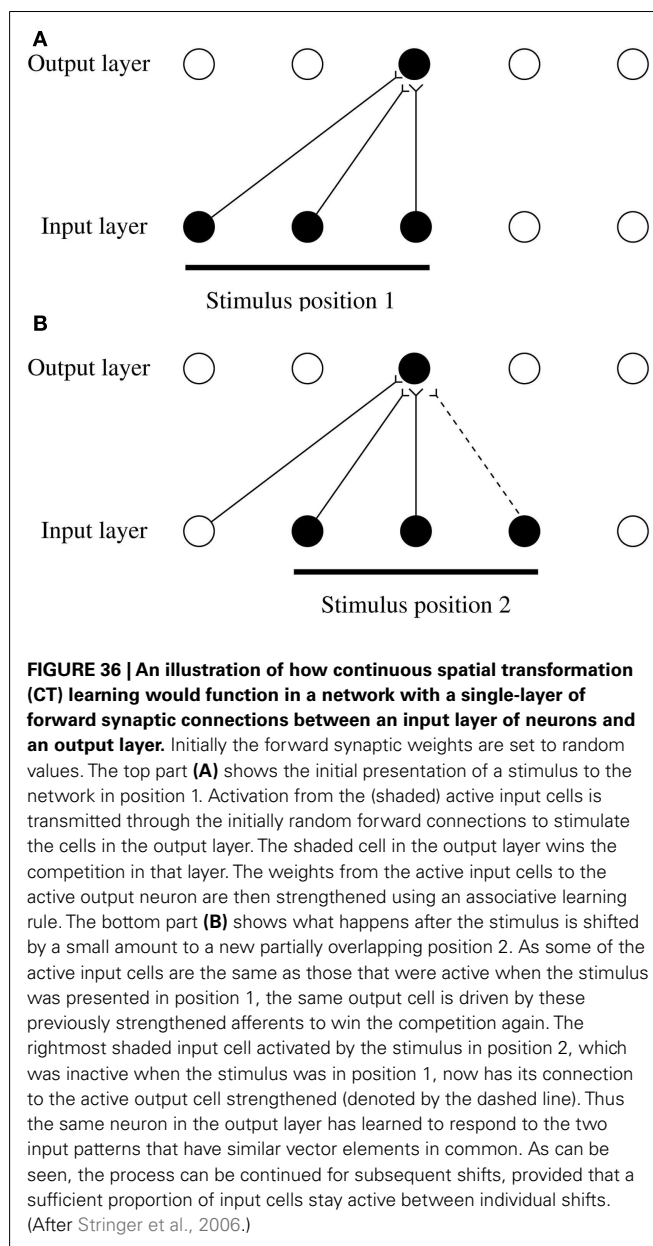
The concept is illustrated in **Figure 36**. During the presentation of a visual image at one position on the retina that activates neurons in layer 1, a small winning set of neurons in layer 2 will modify (through associative learning) their afferent connections from layer 1 to respond well to that image in that location. When the same image appears later at nearby locations, so that there is spatial continuity, the same neurons in layer 2 will be activated because some of the active afferents are the same as when the image was in the first position. The key point is that if these afferent connections have been strengthened sufficiently while the image is in the first location, then these connections will be able to continue to activate the same neurons in layer 2 when the image appears in overlapping nearby locations. Thus the same neurons in the output layer have learned to respond to inputs that have similar vector elements in common.

As can be seen in **Figure 36**, the process can be continued for subsequent shifts, provided that a sufficient proportion of input cells stay active between individual shifts. This whole process is repeated throughout the network, both horizontally as the image moves on the retina, and hierarchically up through the network. Over a series of stages, transform invariant (e.g., location invariant) representations of images are successfully learned, allowing the network to perform invariant object recognition. A similar CT learning process may operate for other kinds of transformation, such as change in view or size.

Stringer et al. (2006) demonstrated that VisNet can be trained with continuous spatial transformation learning to form view-invariant representations. They showed that CT learning requires the training transforms to be relatively close together spatially so that spatial continuity is present in the training set; and that the order of stimulus presentation is not crucial, with even interleaving with other objects possible during training, because it is spatial continuity rather than the temporal continuity that drives the self-organizing learning with the purely associative synaptic modification rule.

Perry et al. (2006) extended these simulations with VisNet of view-invariant learning using CT to more complex 3D objects, and using the same training images in human psychophysical investigations, showed that view-invariant object learning can occur when spatial but not temporal continuity applies in a training condition in which the images of different objects were interleaved. However, they also found that the human view-invariance learning was better if sequential presentation of the images of an object was used, indicating that temporal continuity is an important factor in human invariance learning.

Perry et al. (2010) extended the use of continuous spatial transformation learning to translation invariance. They showed that translation-invariant representations can be learned by continuous spatial transformation learning; that the transforms must be close for this to occur; that the temporal order of presentation of each transformed image during training is not crucial for learning to occur; that relatively large numbers of transforms can be learned; and that such continuous spatial transformation learning can be usefully combined with temporal trace training.



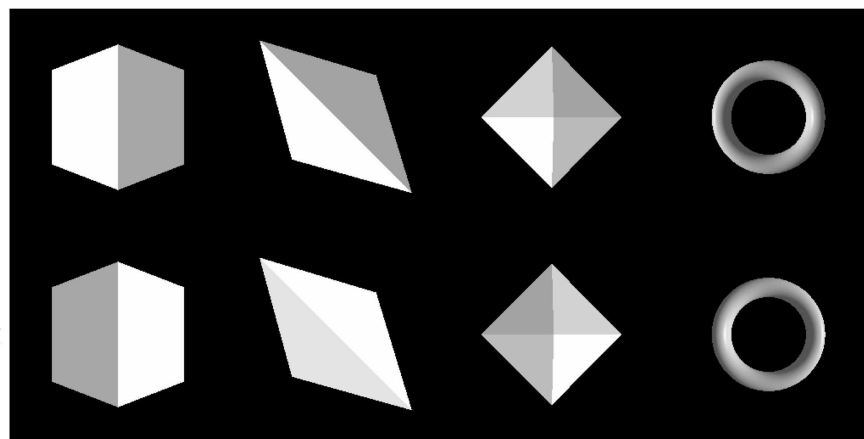
**FIGURE 36 | An illustration of how continuous spatial transformation (CT) learning would function in a network with a single-layer of forward synaptic connections between an input layer of neurons and an output layer.** Initially the forward synaptic weights are set to random values. The top part (**A**) shows the initial presentation of a stimulus to the network in position 1. Activation from the (shaded) active input cells is transmitted through the initially random forward connections to stimulate the cells in the output layer. The shaded cell in the output layer wins the competition in that layer. The weights from the active input cells to the active output neuron are then strengthened using an associative learning rule. The bottom part (**B**) shows what happens after the stimulus is shifted by a small amount to a new partially overlapping position 2. As some of the active input cells are the same as those that were active when the stimulus was presented in position 1, the same output cell is driven by these previously strengthened afferents to win the competition again. The rightmost shaded input cell activated by the stimulus in position 2, which was inactive when the stimulus was in position 1, now has its connection to the active output cell strengthened (denoted by the dashed line). Thus the same neuron in the output layer has learned to respond to the two input patterns that have similar vector elements in common. As can be seen, the process can be continued for subsequent shifts, provided that a sufficient proportion of input cells stay active between individual shifts. (After Stringer et al., 2006.)

## 5.11. LIGHTING INVARIANCE

Object recognition should occur correctly even despite variations of lighting. In an investigation of this, Rolls and Stringer (2006) trained VisNet on a set of 3D objects generated with OpenGL in which the viewing angle and lighting source could be independently varied (see **Figure 37**). After training with the trace rule on all the 180 views (separated by 1°, and rotated about the vertical axis in **Figure 37**) of each of the four objects under the left lighting condition, we tested whether the network would recognize the objects correctly when they were shown again, but with the source of the lighting moved to the right so that the objects appeared different (see **Figure 37**). With this protocol, lighting invariant object recognition by VisNet was demonstrated (Rolls and Stringer, 2006).

Left Lighting

Right Lighting



**FIGURE 37 | Lighting invariance.** VisNet was trained on a set of 3D objects (cube, tetrahedron, octahedron, and torus) generated with OpenGL in which for training the objects had left lighting, and for testing the objects had right

lighting. Just one view of each object is shown in the Figure, but for training and testing 180 views of each object separated by  $1^\circ$  were used. (After Rolls and Stringer, 2006.)

Some insight into the good performance with a change of lighting is that some neurons in the inferior temporal visual cortex respond to the outlines of 3D objects (Vogels and Biederman, 2002), and these outlines will be relatively consistent across lighting variations. Although the features about the object represented in VisNet will include more than the representations of the outlines, the network may because it uses distributed representations of each object generalize correctly provided that some of the features are similar to those present during training. Under very difficult lighting conditions, it is likely that the performance of the network could be improved by including variations in the lighting during training, so that the trace rule could help to build representations that are explicitly invariant with respect to lighting.

## 5.12. INVARIANT GLOBAL MOTION IN THE DORSAL VISUAL SYSTEM

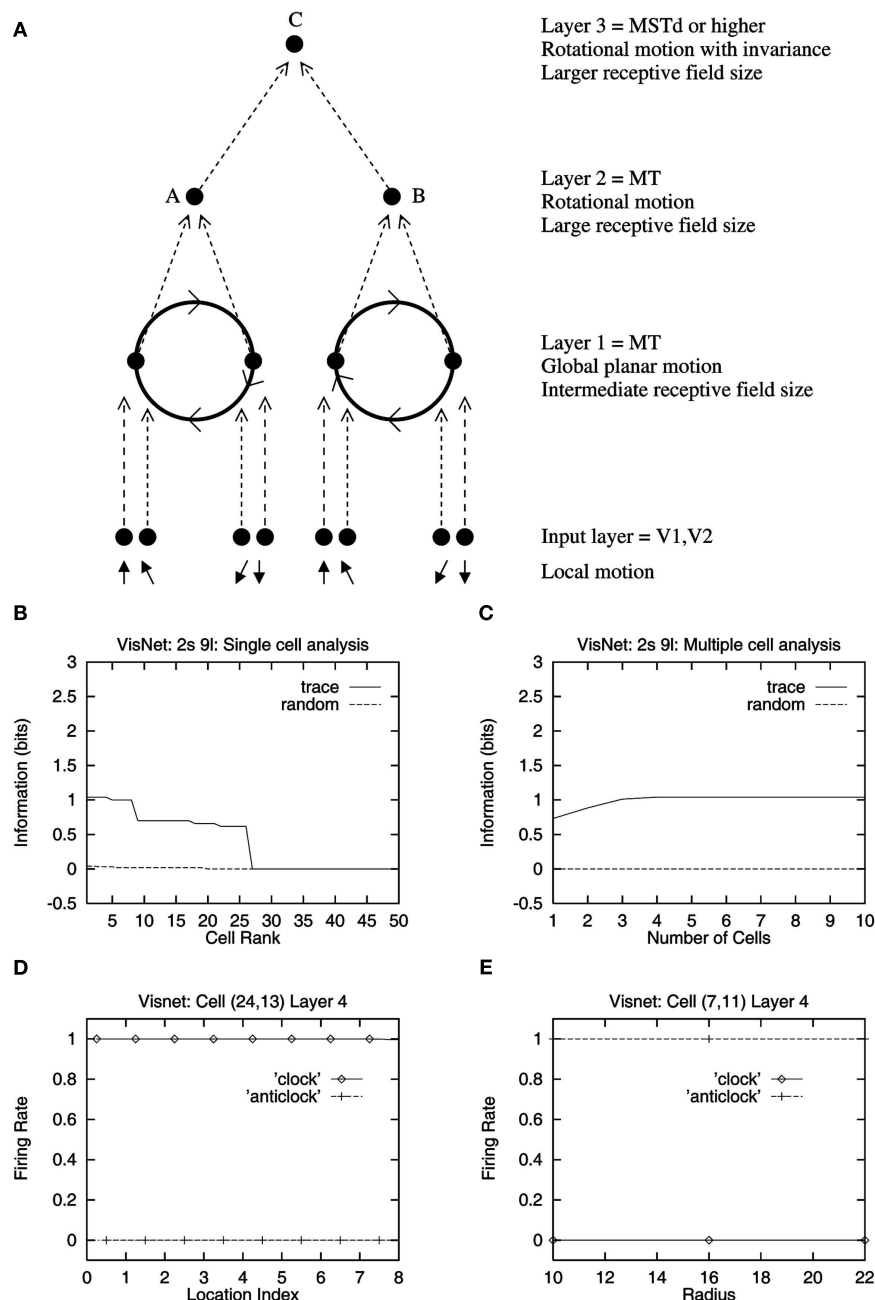
A key issue in understanding the cortical mechanisms that underlie motion perception is how we perceive the motion of objects such as a rotating wheel invariantly with respect to position on the retina, and size. For example, we perceive the wheel shown in **Figure 38A** rotating clockwise independently of its position on the retina. This occurs even though the local motion for the wheels in the different positions may be opposite. How could this invariance of the visual motion perception of objects arise in the visual system? Invariant motion representations are known to be developed in the cortical dorsal visual system. Motion-sensitive neurons in V1 have small receptive fields (in the range  $1\text{--}2^\circ$  at the fovea), and can therefore not detect global motion, and this is part of the aperture problem (Wurtz and Kandel, 2000b). Neurons in MT, which receives inputs from V1 and V2, have larger receptive fields (e.g.,  $5^\circ$  at the fovea), and are able to respond to planar global motion, such as a field of small dots in which the majority (in practice as few as 55%) move in one direction, or to the overall direction of a moving plaid, the orthogonal grating components of which have motion at  $45^\circ$  to the overall motion (Movshon et al., 1985; Newsome et al., 1989). Further on in the dorsal visual system, some neurons in macaque visual area MST (but not MT) respond

to rotating flow fields or looming with considerable translation invariance (Graziano et al., 1994; Geesaman and Andersen, 1996). In the cortex in the anterior part of the superior temporal sulcus, which is a convergence zone for inputs from the ventral and dorsal visual systems, some neurons respond to object-based motion, for example, to a head rotating clockwise but not anticlockwise, independently of whether the head is upright or inverted which reverses the optic flow across the retina (Hasselmo et al., 1989b).

In a unifying hypothesis with the design of the ventral cortical visual system Rolls and Stringer (2007) proposed that the dorsal visual system uses a hierarchical feed-forward network architecture (V1, V2, MT, MSTd, parietal cortex) with training of the connections with a short-term memory trace associative synaptic modification rule to capture what is invariant at each stage. The principle is illustrated in **Figure 38A**. Simulations showed that the proposal is computationally feasible, in that invariant representations of the motion flow fields produced by objects self-organize in the later layers of the architecture (see examples in **Figures 38B–E**). The model produces invariant representations of the motion flow fields produced by global in-plane motion of an object, in-plane rotational motion, looming vs receding of the object. The model also produces invariant representations of object-based rotation about a principal axis. Thus it is proposed that the dorsal and ventral visual systems may share some unifying computational principles Rolls and Stringer (2007). Indeed, the simulations of Rolls and Stringer (2007) used a standard version of VisNet, with the exception that instead of using oriented bar receptive fields as the input to the first layer, local motion flow fields provided the inputs.

## 6. LEARNING INVARIANT REPRESENTATIONS OF SCENES AND PLACES

The primate hippocampal system has neurons that respond to a view of a spatial scene, or when that location in a scene is being looked at in the dark or when it is obscured (Rolls et al., 1997a, 1998; Robertson et al., 1998; Georges-François et al., 1999; Rolls and Xiang, 2006; Rolls, 2008b). The representation is relatively



**FIGURE 38 | (A)** Two rotating wheels at different locations rotating in opposite directions. The local flow field is ambiguous. Clockwise or counterclockwise rotation can only be diagnosed by a global flow computation, and it is shown how the network is expected to solve the problem to produce position-invariant global motion-sensitive neurons. One rotating wheel is presented at any one time, but the need is to develop a representation of the fact that in the case shown the rotating flow field is always clockwise, independently of the location of the flow field. **(B–D)** Translation invariance, with training on 9 locations. **(B)** Single cell information measures showing that some layer 4 neurons have

perfect performance of 1 bit (clockwise vs anticlockwise) after training with the trace rule, but not with random initial synaptic weights in the untrained control condition. **(C)** The multiple cell information measure shows that small groups of neurons have perfect performance. **(D)** Position invariance illustrated for a single cell from layer 4, which responded only to the clockwise rotation, and for every one of the 9 positions. **(E)** Size invariance illustrated for a single cell from layer 4, which after training with three different radii of rotating wheel, responded only to anticlockwise rotation, independently of the size of the rotating wheels. (After Rolls and Stringer, 2007.)

invariant with respect to the position of the macaque in the environment, and of head direction, and eye position. The requirement for these spatial view neurons is that a position in the spatial scene

is being looked at. (There is an analogous set of place neurons in the rat hippocampus that respond in this case when the rat is in a given position in space, relatively invariantly with respect to

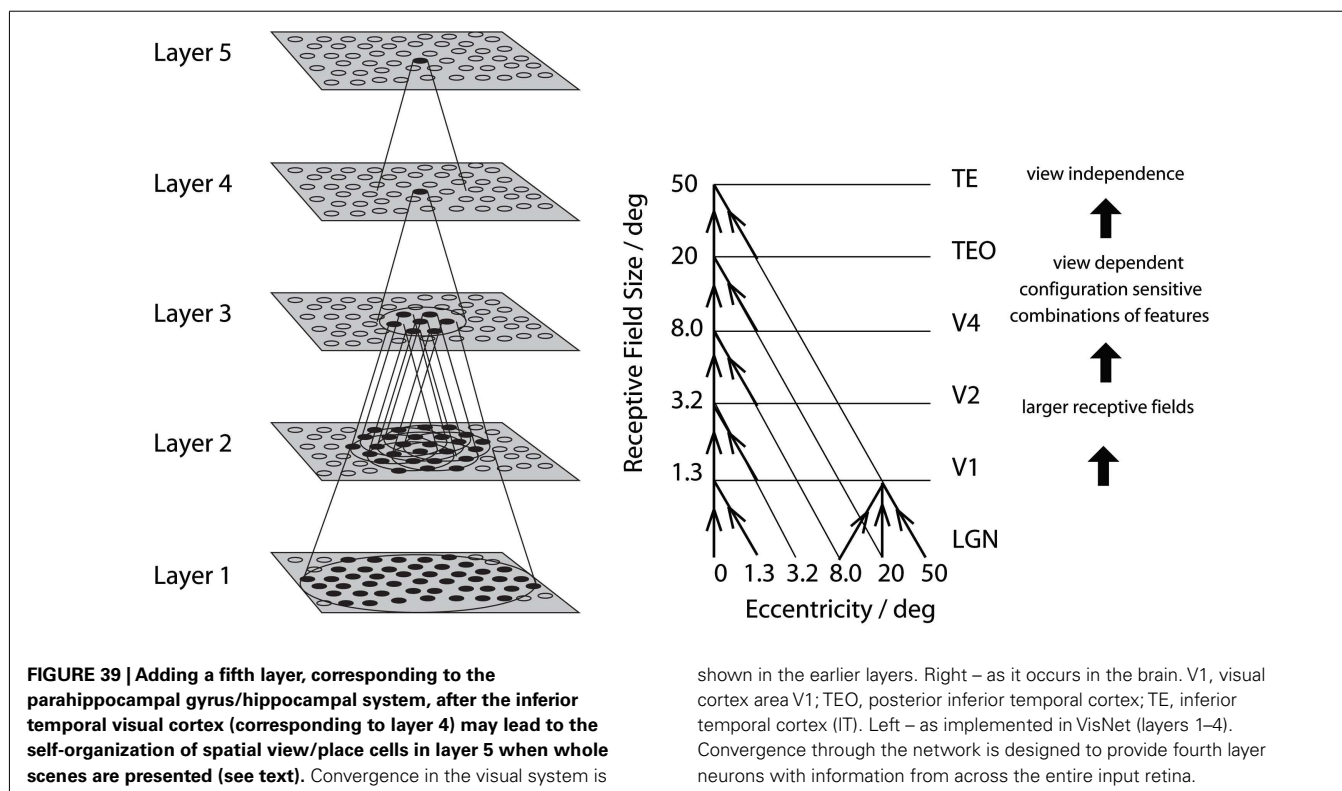
head direction (McNaughton et al., 1983; O'Keefe, 1984; Muller et al., 1991).) How might these spatial view neurons be set up in primates?

Before addressing this, it is useful to consider the difference between a spatial view or scene representation, and an object representation. An object can be moved to different places in space or in a spatial scene. An example is a motor car that can be moved to different places in space. The object is defined by a combination of features or parts in the correct relative spatial position, but its representation is independent of where it is in space. In contrast, a representation of space has objects in defined relative spatial positions, which cannot be moved relative to one another in space. An example might be Trafalgar Square, in which Nelson's column is in the middle, and the National Gallery and St Martin's in the Fields church are at set relative locations in space, and cannot be moved relative to one another. This draws out the point that there may be some computational similarities between the construction of an object and of a scene or a representation of space, but there are also important differences in how they are used. In the present context we are interested in how the brain may set up a spatial view representation in which the relative position of the objects in the scene defines the spatial view. That spatial view representation may be relatively invariant with respect to the exact position from which the scene is viewed (though extensions are needed if there are central objects in a space through which one moves).

It is now possible to propose a unifying hypothesis of the relation between the ventral visual system, and primate hippocampal spatial view representations (Rolls, 2008b; Rolls et al., 2008). Let us consider a computational architecture in which a fifth layer is added to the VisNet architecture, as illustrated in **Figure 39**. In

the anterior inferior temporal visual cortex, which corresponds to the fourth layer of VisNet, neurons respond to objects, but several objects close to the fovea (within approximately  $10^\circ$ ) can be represented because many object-tuned neurons have asymmetric receptive fields with respect to the fovea (Aggelopoulos and Rolls, 2005; see Section 5.9). If the fifth layer of VisNet performs the same operation as previous layers, it will form neurons that respond to combinations of objects in the scene with the positions of the objects relative spatially to each other incorporated into the representation (as described in Section 5.4). The result will be spatial view neurons in the case of primates when the visual field of the primate has a narrow focus (due to the high-resolution fovea), and place cells when as in the rat the visual field is very wide (De Araujo et al., 2001; Rolls, 2008b). The trace-learning rule in layer 5 should help the spatial view or place fields that develop to be large and single, because of the temporal continuity that is inherent when the agent moves from one part of the view or place space to another, in the same way as has been shown for the entorhinal grid cell to hippocampal place cell mapping (Rolls et al., 2006b; Rolls, 2008b).

The hippocampal dentate granule cells form a network expected to be important in this competitive learning of spatial view or place representations based on visual inputs. As the animal navigates through the environment, different spatial view cells would be formed. Because of the overlapping fields of adjacent spatial view neurons, and hence their coactivity as the animal navigates, recurrent collateral associative connections at the next stage of the system, CA3, could form a continuous attractor representation of the environment (Rolls, 2008b). We thus have a hypothesis for how the spatial representations are formed as a



natural extension of the hierarchically organized competitive networks in the ventral visual system. The expression of such spatial representations in CA3 may be particularly useful for associating those spatial representations with other inputs, such as objects or rewards (Rolls, 2008b).

We have performed simulations to test this hypothesis with VisNet simulations with conceptually a fifth layer added (Rolls et al., 2008). Training now with whole scenes that consist of a set of objects in a given fixed spatial relation to each other results in neurons in the added layer that respond to one of the trained whole scenes, but do not respond if the objects in the scene are rearranged to make a new scene from the same objects. The formation of these scene-specific representations in the added layer is related to the fact that in the inferior temporal cortex (Aggelopoulos and Rolls, 2005), and in the VisNet model (Rolls et al., 2008), the receptive fields of inferior temporal cortex neurons shrink and become asymmetric when multiple objects are present simultaneously in a natural scene. This also provides a solution to the issue of the representation of multiple objects, and their relative spatial positions, in complex natural scenes (Rolls, 2008b).

Consistently, in a more artificial network trained by gradient ascent with a goal function that included forming relatively time invariant representations and decorrelating the responses of neurons within each layer of the 5-layer network, place-like cells were formed at the end of the network when the system was trained with a real or simulated robot moving through spatial environments (Wyss et al., 2006), and slowness as an asset in learning spatial representations has also been investigated by others (Wiskott and Sejnowski, 2002; Wiskott, 2003; Franzius et al., 2007). It will be interesting to test whether spatial view cells develop in a VisNet fifth layer if trained with foveate views of the environment, or place cells if trained with wide angle views of the environment (cf. De Araujo et al., 2001), and the utility of testing this with a VisNet-like architecture is that it embodies a biologically plausible implementation based on neuronally plausible competitive learning and a short-term memory trace-learning rule.

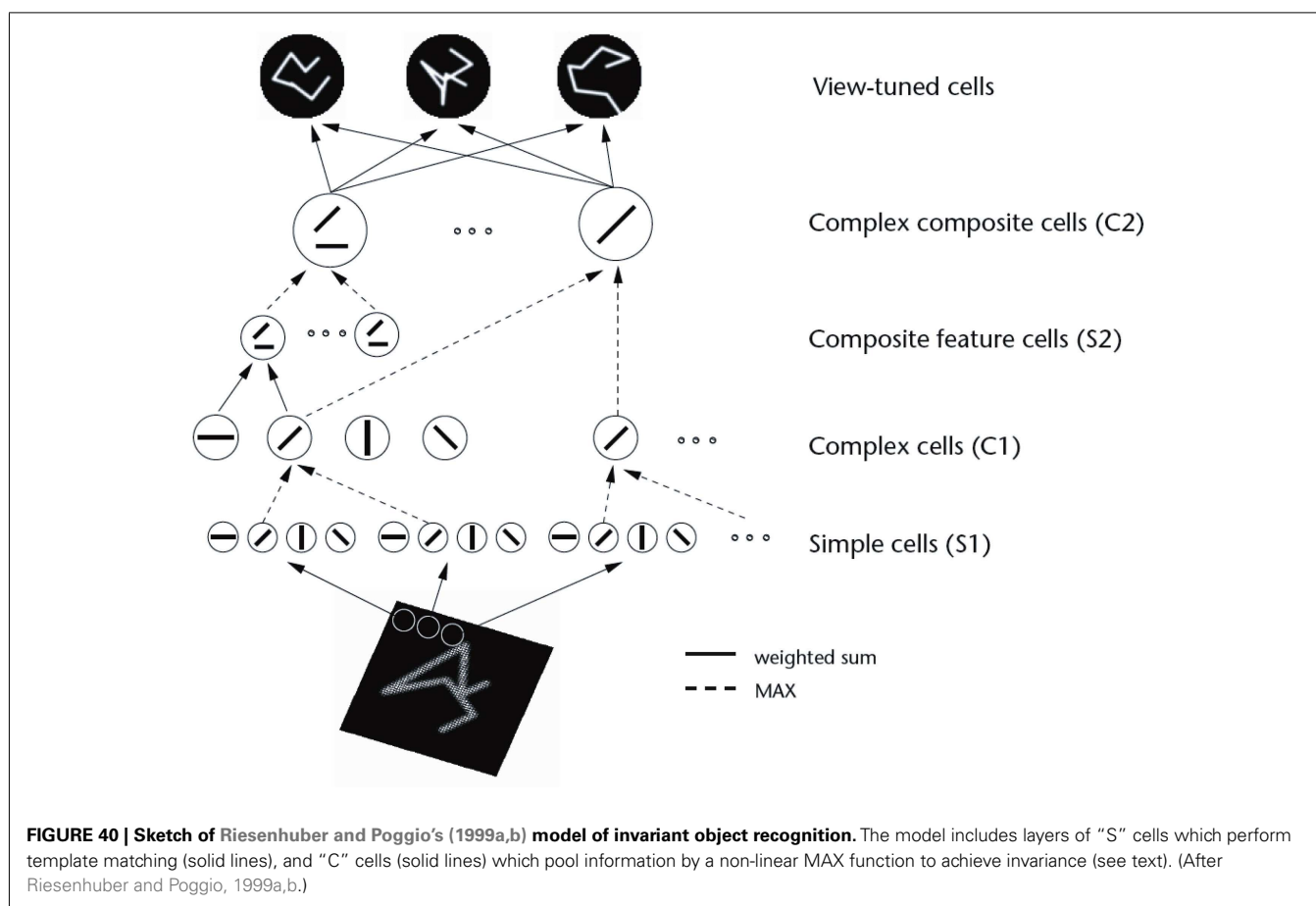
It is an interesting part of the hypothesis just described that because spatial views and places are defined by the relative spatial positions of fixed landmarks (such as buildings), slow learning of such representations over a number of trials might be useful, so that the neurons come to represent spatial views or places, and do not learn to represent a random collection of moveable objects seen once in conjunction. In this context, an alternative brain region to the dentate gyrus for this next layer of VisNet-like processing might be the parahippocampal areas that receive from the inferior temporal visual cortex. Spatial view cells are present in the parahippocampal areas (Rolls et al., 1997a, 1998, 2005b; Robertson et al., 1998; Georges-François et al., 1999), and neurons with place-like fields (though in some cases as a grid, Hafting et al., 2005) are found in the rat medial entorhinal cortex (Moser and Moser, 1998; Brun et al., 2002; Fyhn et al., 2004; Moser, 2004). These spatial view and place-like representations could be formed in these regions as, effectively, an added layer to VisNet. Moreover, these cortical regions have recurrent collateral connections that could implement a continuous attractor representation. Alternatively, it is possible that these parahippocampal spatial representations reflect the effects of backprojections from the hippocampus to the

entorhinal cortex and thus to parahippocampal areas. In either case, it is an interesting and unifying hypothesis that an effect of adding an additional layer to VisNet-like ventral stream visual processing might with training in a natural environment lead to the self-organization, using the same principles as in the ventral visual stream, of spatial view or place representations in parahippocampal or hippocampal areas (Rolls, 2008b; Rolls et al., 2008). Such spatial view representations are relatively invariant with respect to the position from which the scene is viewed (Georges-François et al., 1999), but are selective to the relative spatial position of the objects that define the spatial view (Rolls, 2008b; Rolls et al., 2008).

## 7. FURTHER APPROACHES TO INVARIANT OBJECT RECOGNITION

A related approach to invariant object recognition is described by Riesenhuber and Poggio (1999b), and builds on the hypothesis that not just shift invariance (as implemented in the Neocognitron of Fukushima (1980)), but also other invariances such as scale, rotation, and even view, could be built into a feature hierarchy system, as suggested by Rolls (1992) and incorporated into VisNet (Wallis et al., 1993; Wallis and Rolls, 1997; Rolls and Milward, 2000; Rolls and Stringer, 2007; Rolls, 2008b; see also Perrett and Oram, 1993). The approach of Riesenhuber and Poggio (1999b) and its developments (Riesenhuber and Poggio, 1999a, 2000; Serre et al., 2007a,b,c) is a feature hierarchy approach that uses alternate “simple cell” and “complex cell” layers in a way analogous to (Fukushima, 1980; see Figure 40).

The function of each S cell layer is to build more complicated features from the inputs, and works by template matching. The function of each “C” cell layer is to provide some translation invariance over the features discovered in the preceding simple cell layer (as in Fukushima, 1980), and operates by performing a MAX function on the inputs. The non-linear MAX function makes a complex cell respond only to whatever is the highest activity input being received, and is part of the process by which invariance is achieved according to this proposal. This C layer process involves “implicitly scanning over afferents of the same type differing in the parameter of the transformation to which responses should be invariant (for instance, feature size for scale invariance), and then selecting the best-matching afferent” (Riesenhuber and Poggio, 1999b). Brain mechanisms by which this computation could be set up are not part of the scheme, and the model does not incorporate learning in its architecture, so does not yet provide a biologically plausible model of invariant object recognition. The model receives as its inputs a set of symmetric spatial-frequency filters that are closely spaced in spatial-frequency, and maps these through pairs of convergence followed by MAX function layers, without learning. Whatever output appears in the final layer is then tested with a support vector machine to measure how well the output can be used by this very powerful subsequent learning stage to categorize different types of image. Whether that is a good test of invariance learning is a matter for discussion (Pinto et al., 2008; see Section 8). The approach taken in VisNet is that instead of using a benchmark test of image exemplars from which to learn categories (Serre et al., 2007a,b,c), instead VisNet is trained to generalize across transforms of objects that provide the training set. However, the fact that the model of Poggio, Riesenhuber, Serre and



colleagues does use a hierarchical approach to object recognition does represent useful convergent thinking toward how invariant object recognition may be implemented in the brain. Similarly, the approach of training a five-layer network with a more artificial gradient ascent approach with a goal function that does however include forming relatively time invariant representations and decorrelating the responses of neurons within each layer (Wyss et al., 2006; both processes that have their counterpart in VisNet), also reflects convergent thinking.

Further evidence consistent with the approach developed in the investigations of VisNet described in this paper comes from psychophysical studies. Wallis and Bülthoff (1999) and Perry et al. (2006) describe psychophysical evidence for learning of view-invariant representations by experience, in that the learning can be shown in special circumstances to be affected by the temporal sequence in which different views of objects are seen.

Another related approach, from the machine learning area, is that of convolutional networks. Convolutional Networks are a biologically inspired trainable architecture that can learn invariant features. Each stage in a ConvNet is composed of a filter bank, some non-linearities, and feature pooling layers. With multiple stages, a ConvNet can learn multi-level hierarchies of features (LeCun et al., 2010). Non-linearities that include rectification and local contrast normalization are important in such systems (Jarrett et al., 2009; and are of course properties of VisNet). Applications have been

developed to visual object recognition and vision navigation for off-road mobile robots. Ullman has considered the use of features in a hierarchy to help with processes such as segmentation and object recognition (Ullman, 2007).

Another approach to the implementation of invariant representations in the brain is the use of neurons with Sigma-Pi synapses. Sigma-Pi synapses effectively allow one input to a synapse to be multiplied or gated by a second input to the synapse (Rolls, 2008b). The multiplying input might gate the appropriate set of the other inputs to a synapse to produce the shift or scale change required. For example, the multiplying input could be a signal that varies with the shift required to compute translation invariance, effectively mapping the appropriate set of  $x_i$  inputs through to the output neurons depending on the shift required (Olshausen et al., 1993, 1995; Mel et al., 1998; Mel and Fiser, 2000). Local operations on a dendrite could be involved in such a process (Mel et al., 1998). The explicit neural implementation of the gating mechanism seems implausible, given the need to multiply and thus remap large parts of the retinal input depending on shift and scale modifying connections to a particular set of output neurons. Moreover, the explicit control signal to set the multiplication required in V1 has not been identified. Moreover, if this was the solution used by the brain, the whole problem of shift and scale invariance could in principle be solved in one-layer of the system, rather than with the multiple hierarchically organized set of layers actually used



in the brain, as shown schematically in **Figure 1**. The multiple-layers actually used in the brain are much more consistent with the type of scheme incorporated in VisNet. Moreover, if a multiplying system of the type hypothesized by Olshausen et al. (1993), Mel et al. (1998), and Olshausen et al. (1995) was implemented in a multilayer hierarchy with the shift and scale change emerging gradually, then the multiplying control signal would need to be supplied to every stage of the hierarchy. A further problem with such approaches is how the system is trained in the first place.

## 8. MEASURING THE CAPACITY OF VisNet

For a theory of the brain mechanisms of invariant object recognition, it is important that the system should scale up, so that if a model such as VisNet was the size of the human visual system, it would have comparable performance. Most of the research with VisNet to date has focused on the principles of operation of the system, and what aspects of invariant object recognition the model can solve (Rolls, 2008b). In this section I consider how the system performs in its scaled up version (VisNetL, with  $128 \times 128$  neurons in each of 4 layers). I compare the capacity of VisNetL with that of another model, HMAX, as that has been described as competing with state of the art systems (Serre et al., 2007a,b,c; Mutch and Lowe, 2008), and I raise interesting issues about how to measure the capacity of systems for invariant object recognition in natural scenes.

The tests (performed by L. Robinson of the Department of Computer Science, University of Warwick, UK and E. T. Rolls) utilized a benchmark approach incorporated in the work of Serre, Mutch, Poggio and colleagues (Serre et al., 2007b,c; Mutch and Lowe, 2008) and indeed typical of many standard approaches in computer vision. This uses standard datasets such as the Caltech-256 (Griffin et al., 2007) in which sets of images from different categories are to be classified.

### 8.1. OBJECT BENCHMARK DATABASES

The Caltech-256 dataset (Griffin et al., 2007) is comprised of 256 object classes made up of images that have many aspect ratios, sizes and differ quite significantly in quality (having being manually collated from web searches). The objects within the images show significant intra-class variation and have a variety of poses, illumination, scale, and occlusion as expected from natural images (see examples in **Figure 41**). In this sense, the Caltech-256 database is considered to be a difficult challenge to object recognition systems. I come to the conclusion below that the benchmarking approach with this type of dataset is not useful for training a system that must learn invariant object representations. The reason for this is that the exemplars of each category in the Caltech-256 dataset are too discontinuous to provide a basis for learning invariant object representations. For example, the exemplars within a category in these datasets may be very different indeed.

Partly because of the limitations of the Caltech-256 database for training in invariant object recognition, we also investigated training with the Amsterdam Library of Images (ALOI; Geusebroek et al., 2005) database<sup>1</sup>. The ALOI database takes a different

approach to the Caltech-256, and instead of focusing on a set of natural images within a category, provides images with a systematic variation of pose and illumination for 1,000 small objects. Each object is placed onto a turntable and photographed in consistent conditions at 5° increments, resulting in a set of images that not only show the whole object (with regard to out of plane rotations), but does so with some continuity from one image to the next (see examples in **Figure 42**).

### 8.2. THE HMAX MODELS USED FOR COMPARISON WITH VISNETL

The performance of VisNetL was compared against a standard HMAX model (Serre et al., 2007b,c; Mutch and Lowe, 2008), and a HMAX model scaled down to have a comparable complexity (in terms, for example, of the number of neurons) to that of VisNetL. The scaled down HMAX model is referred to as HMAX\_min. The current HMAX family models have in the order of 10 million computational units (Serre et al., 2007b), which is at least 100 times the number contained within the current implementation of VisNetL (which uses  $128 \times 128$  neurons in each of 4 layers, i.e., 65,536 neurons). In producing HMAX\_min, we aimed to maintain the architectural features of HMAX, and primarily to scale it down. HMAX\_min is based upon the “base” implementation of Mutch and Lowe (2008)<sup>2</sup>. The minimal version used in the comparisons differs from this base HMAX implementation in two significant ways. First, HMAX\_min has only 4 scales compared to the 10 scales of HMAX. (Care was taken to ensure that HMAX\_min still covered the same image size range – 256, 152, 90, and 53 pixels.) Second, the number of distinct units in the S2 “template matching” layer was limited to only 25 in HMAX\_min, compared to 2,000 in HMAX. This results in a scaled down model HMAX\_min, with approximately 12,000 units in the C1 layer, 75,000 units in the S2 layer, and 25 in the upper C2 layer, which is much closer to the 65,536 neurons of VisNetL. (The 75,000 units in S2 allow for every C2 neuron to be connected by its own weight to a C1 neuron.; When counting the number of neurons in the models, the number of neurons in S1 is not included, as they just provide the inputs to the models.)

### 8.3. PERFORMANCE ON A CALTECH-256 TEST

VisNetL and the two HMAX models were trained to discriminate between two object classes from the Caltech-256 database, the *teddy-bear* and *cowboy-hat* (see examples in **Figure 41**). Sixty image examples of each class were rescaled to  $256 \times 256$  and converted to gray-scale, so that shape recognition was being investigated. The 60 images from each class were randomly partitioned into training and testing sets, with the training set size ranging over 1, 5, 15 and 30 images, and the corresponding testing set being the remainder of the 60 images in the cross-validation design. A linear support vector machine (libSVM, Chang and Lin, 2011) approach operating on the output of layer 4 of VisNetL was used to compare the categorization of the trained images with that of the test images, as that is the approach used by HMAX (Serre et al., 2007b,c; Mutch and Lowe, 2008). The standard default parameters of the support vector machine were used in identical form for the VisNetL and HMAX tests.

<sup>1</sup><http://staff.science.uva.nl/aloi/>

<sup>2</sup><http://cbcl.mit.edu/jmutch/cns/index.html>



**FIGURE 41 |** Example images from the Caltech-256 database for two object classes, teddy-bears and cowboy-hats.



**FIGURE 42 |** Example images from the two object classes within the ALOI database, (A) 90 (rubber duck) and (B) 93 (black shoe). Only the 45° increments are shown.

**Figure 43** shows the performance of all three models when performing the task with the Caltech-256 dataset. It is clear that VisNetL performed better than HMAX\_min as soon as there were reasonable numbers of training images, and this was confirmed statistically using the Chi-square test. It is also shown that the full HMAX model (as expected given its very large number of neurons) exhibits higher performance than that of VisNetL and HMAX\_min.

#### 8.4. PERFORMANCE WITH THE AMSTERDAM LIBRARY OF IMAGES

Eight classes of object (with designations 36, 90, 93, 103, 138, 156, 203, 161) from the dataset were chosen (see **Figure 42**, for example). Each class comprises of 72 images taken at 5° increments through the full 360° out of plane rotation. Three sets of training images were used. (1) Three training images per class were taken at 315, 0, and 45°. (2) Eight training images encompassing the entire rotation of the object were taken in 45° increments. (3) Eighteen training images also encompassing the entire rotation of the object were taken in 20° increments. The testing set consisted for each object of the remaining orientations from the set of 72 that were not present in the particular training set. The aim of using the different training sets was to investigate how

close in viewing angle the training images need to be; and also to investigate the effects of using different numbers of training images.

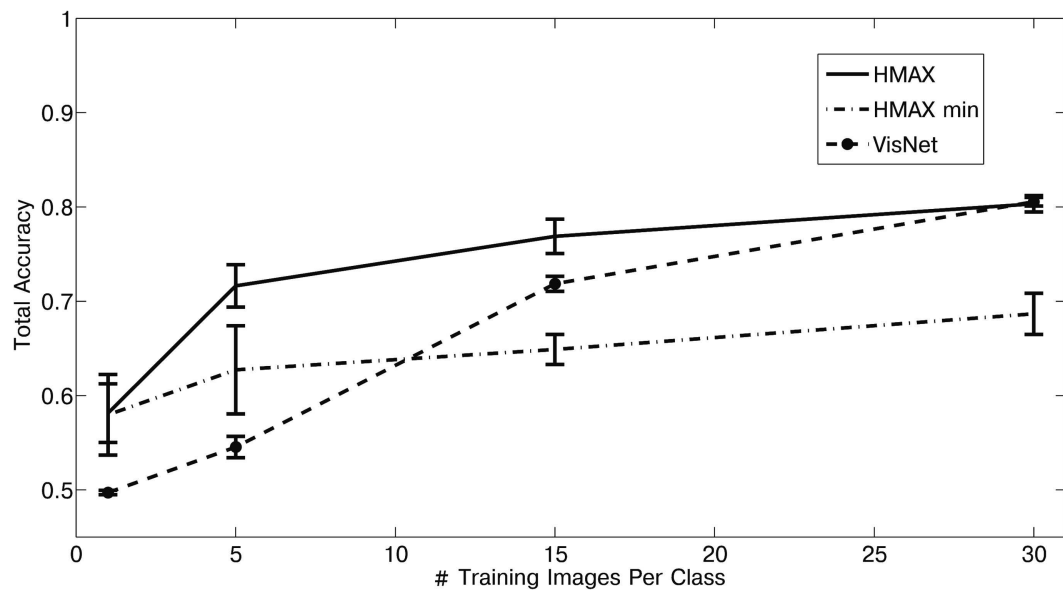
**Figure 44** shows that VisNetL performed better than HMAX\_min as soon as there were even a few training images, with HMAX as expected performing better. VisNetL performed almost as well as the very much larger HMAX as soon as there were reasonable numbers of training images.

What VisNetL can do here is to learn view-invariant representations using its trace-learning rule to build feature analyzers that reflect the similarity across at least adjacent views of the training set. Very interestingly, with 8 training images, the view spacing of the training images was 45°, and the test images in the cross-validation design were the intermediate views, 22.5° away from the nearest trained view. This is promising, for it shows that enormous numbers of training images with many different closely spaced views are not necessary for VisNetL. Even 8 training views spaced 45° apart produced reasonable training.

#### 8.5. INDIVIDUAL LAYER PERFORMANCE

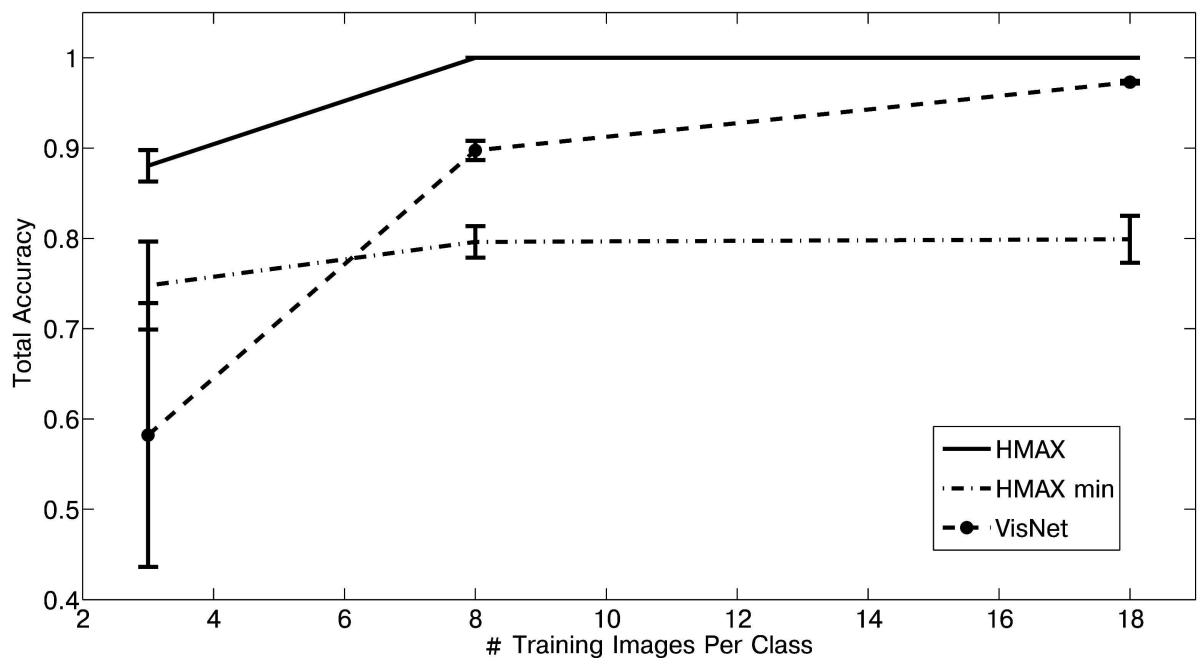
To test whether the VisNet hierarchy is actually performing useful computations with these datasets the simulations were re-run,





**FIGURE 43 | Performance of VisNetL, HMAX, and HMAX\_min on the classification task using the Caltech-256 dataset.** The error bars show the standard error of the means over 5 cross-validation trials with different images chosen at random for the training set on each trial. It is clear that

VisNetL performs better than HMAX\_min, and this was confirmed statistically using the Chi-square test performed with 30 training images and 30 cross-validation test images in each of two categories (Chi-square = 8.09,  $df = 1$ ,  $p = 0.0025$ ).



**FIGURE 44 | Performance of VisNetL, HMAX\_min, and HMAX on the classification task with 8 classes using the Amsterdam Library of Images dataset.** It is clear that VisNetL performs better than HMAX\_min, and this

was confirmed statistically using the Chi-square test performed with 18 training images 20° apart in view and 54 cross-validation testing images 5° apart in each of eight categories (Chi-square = 110.58,  $df = 1$ ,  $p = 10^{-3}$ ).

though this time instead of only training the SVM on the activity generated in the final layer, four identical SVM's were trained independently on the activities of each of the four layers. If the VisNet

hierarchy is actually forming useful representations with these datasets then we should see the discriminatory power of SVMs trained on each layer increase as we traverse the hierarchy.

When the Caltech-256 dataset was used to train VisNetL there was very little difference in the measured performance of classifiers trained on each layer. This is revealing, for it shows that the Caltech-256 dataset does not have sufficient similarity between the exemplars within a given class for the trace-learning rule utilized in VisNet to perform useful learning. Thus, at least with a convergent feature hierarchy network trained in this way, there is insufficient similarity and information in the exemplars of each category of the Caltech-256 to learn to generalize in a view-invariant way to further exemplars of that category.

In contrast, when the ALOI dataset was used to train VisNetL the later layers performed better (layer 2–72% correct; layer 3–84% correct; layer 4–86% correct;  $p < 0.001$ ). Thus there is sufficient continuity in the images in the ALOI dataset to support view-invariance learning in this feature hierarchy network.

## 8.6. EVALUATION

One conclusion is that VisNetL performs comparably to a scaled down version of HMAX on benchmark tests. This is reassuring, for HMAX has been described as competing with state of the art systems (Serre et al., 2007a,b,c; Mutch and Lowe, 2008).

A second conclusion is that image databases such as the Caltech-256 that are used to test the performance of object recognition systems (Serre et al., 2007a,b,c; Mutch and Lowe, 2008; and in many computer vision approaches) are inappropriate as training sets for systems that perform invariant visual object recognition. Instead, for such systems, it will be much more relevant to train on image sets in which the image exemplars within a class show much more continuous variation. This provides the system with the opportunity to learn invariant representations, instead of just doing its best to categorize images into classes from relatively limited numbers of images that do not allow the system to learn the rules of the transforms that objects undergo in the real-world, and that can be used to help object recognition when objects may be seen from different views. This is an important conclusion for research in the area. Consistently, others are realizing that invariant visual object recognition is a hard problem (Pinto et al., 2008; DiCarlo et al., 2012). In this context, the hypotheses presented in this paper are my theory of how invariant visual object recognition is performed by the brain (Rolls, 1992, 2008b), and the model VisNet tests those hypotheses and provides a model for how invariant visual object representations can be learned (Rolls, 2008b).

Third, the findings described here are encouraging with respect to training view-invariant representations, in that the training images with the ALOI dataset could be separated by as much as 45° to still provide for view-invariant object recognition with cross-validation images that were never closer than 22.5° to a training image. This is helpful, for it is an indication that large numbers of different views will not need to be trained with the VisNet architecture in order to achieve good view-invariant object recognition.

## 9. DIFFERENT PROCESSES INVOLVED IN DIFFERENT TYPES OF OBJECT IDENTIFICATION

To conclude this paper, it is proposed that there are (at least) three different types of process that could be involved in object identification. The first is the simple situation where different objects can

be distinguished by different non-overlapping sets of features (see Section 3.1). An example might be a banana and an orange, where the list of features of the banana might include yellow, elongated, and smooth surface; and of the orange its orange color, round shape, and dimpled surface. Such objects could be distinguished just on the basis of a list of the properties, which could be processed appropriately by a competitive network, pattern associator, etc. No special mechanism is needed for view-invariance, because the list of properties is very similar from most viewing angles. Object recognition of this type may be common in animals, especially those with visual systems less developed than those of primates. However, this approach does not describe the shape and form of objects, and is insufficient to account for primate vision. Nevertheless, the features present in objects are valuable cues to object identity, and are naturally incorporated into the feature hierarchy approach.

A second type of process might involve the ability to generalize across a small range of views of an object, that is within a generic view, where cues of the first type cannot be used to solve the problem. An example might be generalization across a range of views of a cup when looking into the cup, from just above the near lip until the bottom inside of the cup comes into view. This type of process includes the learning of the transforms of the surface markings on 3D objects which occur when the object is rotated, as described in Section 5.6. Such generalization would work because the neurons are tuned as filters to accept a range of variation of the input within parameters such as relative size and orientation of the components of the features. Generalization of this type would not be expected to work when there is a catastrophic change in the features visible, as, for example, occurs when the cup is rotated so that one can suddenly no longer see inside it, and the outside bottom of the cup comes into view.

The third type of process is one that can deal with the sudden catastrophic change in the features visible when an object is rotated to a completely different view, as in the cup example just given (cf. Koenderink, 1990). Another example, quite extreme to illustrate the point, might be when a card with different images on its two sides is rotated so that one face and then the other is in view. This makes the point that this third type of process may involve arbitrary pairwise association learning, to learn which features and views are different aspects of the same object. Another example occurs when only some parts of an object are visible. For example, a red-handled screwdriver may be recognized either from its round red handle, or from its elongated silver-colored blade.

The full view-invariant recognition of objects that occurs even when the objects share the same features, such as color, texture, etc. is an especially computationally demanding task which the primate visual system is able to perform with its highly developed temporal lobe cortical visual areas. The neurophysiological evidence and the neuronal network analyses described here and elsewhere (Rolls, 2008b) provide clear hypotheses about how the primate visual system may perform this task.

## 10. CONCLUSION

We have seen that the feature hierarchy approach has a number of advantages in performing object recognition over other approaches (see Section 3), and that some of the key computational

issues that arise in these architectures have solutions (see Sections 4 and 5). The neurophysiological and computational approach taken here focuses on a feature hierarchy model in which invariant representations can be built by self-organizing learning based on the statistics of the visual input.

The model can use temporal continuity in an associative synaptic learning rule with a short-term memory trace, and/or it can use spatial continuity in continuous spatial transformation learning.

The model of visual processing in the ventral cortical stream can build representations of objects that are invariant with respect to translation, view, size, and lighting.

The model uses a feature combination neuron approach with the relative spatial positions of the objects specified in the feature combination neurons, and this provides a solution to the binding problem.

The model has been extended to provide an account of invariant representations in the dorsal visual system of the global motion produced by objects such as looming, rotation, and object-based movement.

The model has been extended to incorporate top-down feedback connections to model the control of attention by biased competition in, for example, spatial and object search tasks (Deco and Rolls, 2004; Rolls, 2008b).

The model has also been extended to account for how the visual system can select single objects in complex visual scenes, how multiple objects can be represented in a scene, and how invariant representations of single objects can be learned even when multiple objects are present in the scene.

It has also been suggested in a unifying proposal that adding a fifth layer to the model and training the system in spatial environments will enable hippocampus-like spatial view neurons or place cells to develop, depending on the size of the field of view (Section 6).

We have thus seen how many of the major computational issues that arise when formulating a theory of object recognition in the ventral visual system (such as feature binding, invariance learning, the recognition of objects when they are in cluttered natural scenes, the representation of multiple objects in a scene, and learning invariant representations of single objects when there are multiple objects in the scene), could be solved in the brain, with tests of the hypotheses performed by simulations that are consistent with complementary neurophysiological results.

The approach described here is unifying in a number of ways. First, a set of simple organizational principles involving a hierarchy of cortical areas with convergence from stage to stage, and

competitive learning using a modified associative learning rule with a short-term memory trace of preceding neuronal activity, provide a basis for understanding much processing in the ventral visual stream, from V1 to the inferior temporal visual cortex. Second, the same principles help to understand some of the processing in the dorsal visual stream by which invariant representations of the global motion of objects may be formed. Third, the same principles continued from the ventral visual stream onward to the hippocampus help to show how spatial view and place representations may be built from the visual input. Fourth, in all these cases, the learning is possible because the system is able to extract invariant representations because it can utilize the spatio-temporal continuities and statistics in the world that help to define objects, moving objects, and spatial scenes. Fifth, a great simplification and economy in terms of brain design is that the computational principles need not be different in each of the cortical areas in these hierarchical systems, for some of the important properties of the processing in these systems to be performed.

In conclusion, we have seen how the invariant recognition of objects involves not only the storage and retrieval of information, but also major computations to produce invariant representations. Once these invariant representations have been formed, they are used for many processes including not only recognition memory (Rolls, 2008b), but also associative learning of the rewarding and punishing properties of objects for emotion and motivation (Rolls, 2005, 2008b, 2013), the memory for the spatial locations of objects and rewards, the building of spatial representations based on visual input, and as an input to short-term memory, attention, decision, and action selection systems (Rolls, 2008b).

## ACKNOWLEDGMENTS

Edmund T. Rolls is grateful to Larry Abbott, Nicholas Aggelopoulos, Roland Baddeley, Francesco Battaglia, Michael Booth, Gordon Baylis, Hugo Critchley, Gustavo Deco, Martin Ekliffe, Leonardo Franco, Michael Hasselmo, Nestor Parga, David Perrett, Gavin Perry, Leigh Robinson, Simon Stringer, Martin Tovee, Alessandro Treves, James Tromans, and Tristan Webb for contributing to many of the collaborative studies described here. Professor R. Watt, of Stirling University, is thanked for assistance with the implementation of the difference of Gaussian filters used in many experiments with VisNet and VisNet2. Support from the Medical Research Council, the Wellcome Trust, the Oxford McDonnell Centre in Cognitive Neuroscience, and the Oxford Centre for Computational Neuroscience ([www.oxcns.org](http://www.oxcns.org), where .pdfs of papers are available) is acknowledged.

## REFERENCES

- Abbott, L. F., Rolls, E. T., and Tovee, M. J. (1996). Representational capacity of face coding in monkeys. *Cereb. Cortex* 6, 498–505.
- Abeles, M. (1991). *Corticonics: Neural Circuits of the Cerebral Cortex*. Cambridge: Cambridge University Press.
- Ackley, D. H., Hinton, G. E., and Sejnowski, T. J. (1985). A learning algorithm for Boltzmann machines. *Cogn. Sci.* 9, 147–169.
- Aggelopoulos, N. C., Franco, L., and Rolls, E. T. (2005). Object perception in natural scenes: encoding by inferior temporal cortex simultaneously recorded neurons. *J. Neurophysiol.* 93, 1342–1357.
- Aggelopoulos, N. C., and Rolls, E. T. (2005). Natural scene perception: inferior temporal cortex neurons encode the positions of different objects in the scene. *Eur. J. Neurosci.* 22, 2903–2916.
- Amit, D. J. (1989). *Modelling Brain Function*. New York: Cambridge University Press.
- Anzai, A., Peng, X., and Van Essen, D. C. (2007). Neurons in monkey visual area V2 encode combinations of orientations. *Nat. Neurosci.* 10, 1313–1321.
- Arathorn, D. (2002). *Map-Seeking Circuits in Visual Cognition: A Computational Mechanism for Biological and Machine Vision*. Stanford, CA: Stanford University Press.
- Arathorn, D. (2005). “Computation in the higher visual cortices: map-seeking circuit theory and

- application to machine vision," in *Proceedings of the AIPR 2004: 33rd Applied Imagery Pattern Recognition Workshop*, 73–78.
- Ballard, D. H. (1990). "Animate vision uses object-centred reference frames," in *Advanced Neural Computers*, ed. R. Eckmiller (Elsevier: Amsterdam), 229–236.
- Barlow, H. B. (1972). Single units and sensation: a neuron doctrine for perceptual psychology. *Perception* 1, 371–394.
- Barlow, H. B. (1985). "Cerebral cortex as model builder," in *Models of the Visual Cortex*, eds D. Rose and V. G. Dobson (Chichester: Wiley), 37–46.
- Barlow, H. B., Kaushal, T. P., and Mitchison, G. J. (1989). Finding minimum entropy codes. *Neural Comput.* 1, 412–423.
- Bartlett, M. S., and Sejnowski, T. J. (1997). "Viewpoint invariant face recognition using independent component analysis and attractor networks," in *Advances in Neural Information Processing Systems*, Vol. 9, eds M. Mozer, M. Jordan, and T. Petsche (Cambridge, MA: MIT Press), 817–823.
- Baylis, G. C., Rolls, E. T., and Leonard, C. M. (1985). Selectivity between faces in the responses of a population of neurons in the cortex in the superior temporal sulcus of the monkey. *Brain Res.* 342, 91–102.
- Baylis, G. C., Rolls, E. T., and Leonard, C. M. (1987). Functional subdivisions of temporal lobe neocortex. *J. Neurosci.* 7, 330–342.
- Bennett, A. (1990). Large competitive networks. *Network* 1, 449–462.
- Biederman, I. (1972). Perceiving real-world scenes. *Science* 177, 77–80.
- Biederman, I. (1987). Recognition-by-components: a theory of human image understanding. *Psychol. Rev.* 94, 115–147.
- Binford, T. O. (1981). Inferring surfaces from images. *Artif. Intell.* 17, 205–244.
- Blumberg, J., and Kreiman, G. (2010). How cortical neurons help us see: visual recognition in the human brain. *J. Clin. Invest.* 120, 3054–3063.
- Bolles, R. C., and Cain, R. A. (1982). Recognizing and locating partially visible objects: the local-feature-focus method. *Int. J. Robot. Res.* 1, 57–82.
- Booth, M. C. A., and Rolls, E. T. (1998). View-invariant representations of familiar objects by neurons in the inferior temporal visual cortex. *Cereb. Cortex* 8, 510–523.
- Boussaoud, D., Desimone, R., and Ungerleider, L. G. (1991). Visual topography of area TEO in the macaque. *J. Comp. Neurol.* 306, 554–575.
- Brady, M., Ponce, J., Yuille, A., and Asada, H. (1985). Describing surfaces, A. I. Memo 882. *Artif. Intell.* 17, 285–349.
- Brincat, S. L., and Connor, C. E. (2006). Dynamic shape synthesis in posterior inferotemporal cortex. *Neuron* 49, 17–24.
- Bruce, V. (1988). *Recognising Faces*. Hillsdale, NJ: Erlbaum.
- Brun, V. H., Otnass, M. K., Molden, S., Steffenach, H. A., Witter, M. P., Moser, M. B., and Moser, E. I. (2002). Place cells and place recognition maintained by direct entorhinal-hippocampal circuitry. *Science* 296, 2243–2246.
- Buckley, M. J., Booth, M. C. A., Rolls, E. T., and Gaffan, D. (2001). Selective perceptual impairments following perirhinal cortex ablation. *J. Neurosci.* 21, 9824–9836.
- Buhmann, J., Lange, J., von der Malsburg, C., Vorbrüggen, J. C., and Würtz, R. P. (1991). "Object recognition in the dynamic link architecture: parallel implementation of a transputer network," in *Neural Networks for Signal Processing*, ed. B. Kosko (Englewood Cliffs, NJ: Prentice-Hall), 121–159.
- Carlson, E. T., Rasquinha, R. J., Zhang, K., and Connor, C. E. (2011). A sparse object coding scheme in area v4. *Curr. Biol.* 21, 288–293.
- Cerella, J. (1986). Pigeons and perceptors. *Pattern Recognit.* 19, 431–438.
- Chakravarty, I. (1979). A generalized line and junction labeling scheme with applications to scene analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* 1, 202–205.
- Chang, C.-C., and Lin, C.-J. (2011). LIB-SVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol.* 2, 27.
- Dane, C., and Bajcsy, R. (1982). "An object-centred three-dimensional model builder," in *Proceedings of the 6th International Conference on Pattern Recognition*, Munich, 348–350.
- Daugman, J. (1988). Complete discrete 2D-Gabor transforms by neural networks for image analysis and compression. *IEEE Trans. Acoust.* 36, 1169–1179.
- De Araujo, I. E. T., Rolls, E. T., and Stringer, S. M. (2001). A view model which accounts for the response properties of hippocampal primate spatial view cells and rat place cells. *Hippocampus* 11, 699–706.
- De Valois, R. L., and De Valois, K. K. (1988). *Spatial Vision*. New York: Oxford University Press.
- Deco, G., and Rolls, E. T. (2004). A neurodynamical cortical model of visual attention and invariant object recognition. *Vision Res.* 44, 621–644.
- Deco, G., and Rolls, E. T. (2005a). Attention, short term memory, and action selection: a unifying theory. *Prog. Neurobiol.* 76, 236–256.
- Deco, G., and Rolls, E. T. (2005b). Neurodynamics of biased competition and cooperation for attention: a model with spiking neurons. *J. Neurophysiol.* 94, 295–313.
- Desimone, R., and Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annu. Rev. Neurosci.* 18, 193–222.
- DeWeese, M. R., and Meister, M. (1999). How to measure the information gained from one symbol. *Network* 10, 325–340.
- DiCarlo, J. J., and Maunsell, J. H. R. (2003). Anterior inferotemporal neurons of monkeys engaged in object recognition can be highly sensitive to object retinal position. *J. Neurophysiol.* 89, 3264–3278.
- DiCarlo, J. J., Zoccolan, D., and Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron* 73, 415–434.
- Dolan, R. J., Fink, G. R., Rolls, E. T., Booth, M., Holmes, A., Frackowiak, R. S. J., and Friston, K. J. (1997). How the brain learns to see objects and faces in an impoverished context. *Nature* 389, 596–599.
- Dow, B. W., Snyder, A. Z., Vautin, R. G., and Bauer, R. (1981). Magnification factor and receptive field size in foveal striate cortex of the monkey. *Exp. Brain Res.* 44, 213–218.
- Edelman, S. (1999). *Representation and Recognition in Vision*. Cambridge, MA: MIT Press.
- Elliffe, M. C. M., Rolls, E. T., Parga, N., and Renart, A. (2000). A recurrent model of transformation invariance by association. *Neural Netw.* 13, 225–237.
- Elliffe, M. C. M., Rolls, E. T., and Stringer, S. M. (2002). Invariant recognition of feature combinations in the visual system. *Biol. Cybern.* 86, 59–71.
- Engel, A. K., Konig, P., Kreiter, A. K., Schillen, T. B., and Singer, W. (1992). Temporal coding in the visual system: new vistas on integration in the nervous system. *Trends Neurosci.* 15, 218–226.
- Farah, M. J. (2000). *The Cognitive Neuroscience of Vision*. Oxford: Blackwell.
- Farah, M. J., Meyer, M. M., and McMullen, P. A. (1996). The living/nonliving dissociation is not an artifact: giving an a priori implausible hypothesis a strong test. *Cogn. Neuropsychol.* 13, 137–154.
- Faugeras, O. D. (1993). *The Representation, Recognition and Location of 3-D Objects*. Cambridge, MA: MIT Press.
- Faugeras, O. D., and Hebert, M. (1986). The representation, recognition and location of 3-D objects. *Int. J. Robot. Res.* 5, 27–52.
- Feldman, J. A. (1985). Four frames suffice: a provisional model of vision and space. *Behav. Brain Sci.* 8, 265–289.
- Fenske, M. J., Aminoff, E., Gronau, N., and Bar, M. (2006). Top-down facilitation of visual object recognition: object-based and context-based contributions. *Prog. Brain Res.* 155, 3–21.
- Field, D. J. (1987). Relations between the statistics of natural images and the response properties of cortical cells. *J. Opt. Soc. Am. A* 4, 2379–2394.
- Field, D. J. (1994). What is the goal of sensory coding? *Neural Comput.* 6, 559–601.
- Finkel, L. H., and Edelman, G. M. (1987). "Population rules for synapses in networks," in *Synaptic Function*, eds G. M. Edelman, W. E. Gall, and W. M. Cowan (New York: John Wiley & Sons), 711–757.
- Földiák, P. (1991). Learning invariance from transformation sequences. *Neural Comput.* 3, 193–199.
- Földiák, P. (1992). *Models of Sensory Coding*. Technical Report CUED/F-INFENG/TR 91. Department of Engineering, University of Cambridge, Cambridge.
- Folstein, J. R., Gauthier, I., and Palmeri, T. J. (2010). Mere exposure alters category learning of novel objects. *Front. Psychol.* 1:40. doi:10.3389/fpsyg.2010.00040
- Franco, L., Rolls, E. T., Aggelopoulos, N. C., and Jerez, J. M. (2007). Neuronal selectivity, population sparseness, and ergodicity in the inferior temporal visual cortex. *Biol. Cybern.* 96, 547–560.
- Franco, L., Rolls, E. T., Aggelopoulos, N. C., and Treves, A. (2004). The use of decoding to analyze the contribution to the information of the correlations between the firing of simultaneously recorded neurons. *Exp. Brain Res.* 155, 370–384.

- Franzius, M., Sprekeler, H., and Wiskott, L. (2007). Slowness and sparseness lead to place, head-direction, and spatial-view cells. *PLoS Comput. Biol.* 3, e166. doi:10.1371/journal.pcbi.0030166
- Freedman, D. J., and Miller, E. K. (2008). Neural mechanisms of visual categorization: insights from neurophysiology. *Neurosci. Biobehav. Rev.* 32, 311–329.
- Freiwald, W. A., Tsao, D. Y., and Livingstone, M. S. (2009). A face feature space in the macaque temporal lobe. *Nat. Neurosci.* 12, 1187–1196.
- Frey, B. J., and Jojic, N. (2003). Transformation-invariant clustering using the EM algorithm. *IEEE Trans. Pattern Anal. Mach. Intell.* 25, 1–17.
- Fries, P. (2005). A mechanism for cognitive dynamics: neuronal communication through neuronal coherence. *Trends Cogn. Sci. (Regul. Ed.)* 9, 474–480.
- Fries, P. (2009). Neuronal gamma-band synchronization as a fundamental process in cortical computation. *Annu. Rev. Neurosci.* 32, 209–224.
- Fukushima, K. (1975). Cognitron: a self-organizing neural network. *Biol. Cybern.* 20, 121–136.
- Fukushima, K. (1980). Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol. Cybern.* 36, 193–202.
- Fukushima, K. (1988). Neocognitron: a hierarchical neural network model capable of visual pattern recognition unaffected by shift in position. *Neural Netw.* 1, 119–130.
- Fukushima, K. (1989). Analysis of the process of visual pattern recognition by the neocognitron. *Neural Netw.* 2, 413–420.
- Fukushima, K. (1991). Neural networks for visual pattern recognition. *IEEE Trans. E* 74, 179–190.
- Fukushima, K., and Miyake, S. (1982). Neocognitron: a new algorithm for pattern recognition tolerant of deformations and shifts in position. *Pattern Recognit.* 15, 455–469.
- Fyhn, M., Molden, S., Witter, M. P., Moser, E. I., and Moser, M.-B. (2004). Spatial representation in the entorhinal cortex. *Science* 204, 1258–1264.
- Gardner, E. (1988). The space of interactions in neural network models. *J. Phys. A Math. Gen.* 21, 257–270.
- Garthwaite, J. (2008). Concepts of neural nitric oxide-mediated transmission. *Eur. J. Neurosci.* 27, 2783–3802.
- Geesaman, B. J., and Andersen, R. A. (1996). The analysis of complex motion patterns by form/cue invariant MSTd neurons. *J. Neurosci.* 16, 4716–4732.
- Georges-François, P., Rolls, E. T., and Robertson, R. G. (1999). Spatial view cells in the primate hippocampus: allocentric view not head direction or eye position or place. *Cereb. Cortex* 9, 197–212.
- Geusebroek, J.-M., Burghouts, G. J., and Smeulders, A. W. M. (2005). The Amsterdam library of object images. *Int. J. Comput. Vis.* 61, 103–112.
- Gibson, J. J. (1950). *The Perception of the Visual World*. Boston: Houghton Mifflin.
- Gibson, J. J. (1979). *The Ecological Approach to Visual Perception*. Boston: Houghton Mifflin.
- Grabenhorst, F., and Rolls, E. T. (2011). Value, pleasure, and choice systems in the ventral prefrontal cortex. *Trends Cogn. Sci. (Regul. Ed.)* 15, 56–67.
- Graziano, M. S. A., Andersen, R. A., and Snowden, R. J. (1994). Tuning of MST neurons to spiral motions. *J. Neurosci.* 14, 54–67.
- Griffin, G., Holub, A., and Perona, P. (2007). *The Caltech-256*. Caltech Technical Report, Los Angeles, 1–20.
- Grimson, W. E. L. (1990). *Object Recognition by Computer*. Cambridge, MA: MIT Press.
- Griniasty, M., Tsodyks, M. V., and Amit, D. J. (1993). Conversion of temporal correlations between stimuli to spatial correlations between attractors. *Neural Comput.* 35, 1–17.
- Gross, C. G., Desimone, R., Albright, T. D., and Schwartz, E. L. (1985). Inferior temporal cortex and pattern recognition. *Exp. Brain Res.* 11(Suppl.), 179–201.
- Hafting, T., Fyhn, M., Molden, S., Moser, M. B., and Moser, E. I. (2005). Microstructure of a spatial map in the entorhinal cortex. *Nature* 436, 801–806.
- Hasselmo, M. E., Rolls, E. T., and Baylis, G. C. (1989a). The role of expression and identity in the face-selective responses of neurons in the temporal visual cortex of the monkey. *Behav. Brain Res.* 32, 203–218.
- Hasselmo, M. E., Rolls, E. T., Baylis, G. C., and Nalwa, V. (1989b). Object-centered encoding by face-selective neurons in the cortex in the superior temporal sulcus of the monkey. *Exp. Brain Res.* 75, 417–429.
- Hawken, M. J., and Parker, A. J. (1987). Spatial properties of the monkey striate cortex. *Proc. R. Soc. Lond. B Biol. Sci.* 231, 251–288.
- Hegde, J., and Van Essen, D. C. (2000). Selectivity for complex shapes in primate visual area V2. *J. Neurosci.* 20, RC61.
- Hegde, J., and Van Essen, D. C. (2003). Strategies of shape representation in macaque visual area V2. *Vis. Neurosci.* 20, 313–328.
- Hegde, J., and Van Essen, D. C. (2007). A comparative study of shape representation in macaque visual areas V2 and V4. *Cereb. Cortex* 17, 1100–1116.
- Herrnstein, R. J. (1984). “Objects, categories, and discriminative stimuli,” in *Animal Cognition*, Chap. 14, eds H. L. Roitblat, T. G. Bever, and H. S. Terrace (Hillsdale, NJ: Lawrence Erlbaum and Associates), 233–261.
- Hertz, J. A., Krogh, A., and Palmer, R. G. (1991). *Introduction to the Theory of Neural Computation*. Wokingham: Addison-Wesley.
- Hestrin, S., Sah, P., and Nicoll, R. (1990). Mechanisms generating the time course of dual component excitatory synaptic currents recorded in hippocampal slices. *Neuron* 5, 247–253.
- Hinton, G. E. (2010). Learning to represent visual input. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 365, 177–184.
- Hinton, G. E., Dayan, P., Frey, B. J., and Neal, R. M. (1995). The “wake-sleep” algorithm for unsupervised neural networks. *Science* 268, 1158–1161.
- Hinton, G. E., and Ghahramani, Z. (1997). Generative models for discovering sparse distributed representations. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 352, 1177–1190.
- Hinton, G. E., and Sejnowski, T. J. (1986). “Learning and relearning in Boltzmann machines,” in *Parallel Distributed Processing*, Vol. 1, Chap. 7, eds D. Rumelhart and J. L. McClelland (Cambridge, MA: MIT Press), 282–317.
- Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci. U. S. A.* 79, 2554–2558.
- Hubel, D. H., and Wiesel, T. N. (1962). Receptive fields, binocular interaction, and functional architecture in the cat’s visual cortex. *J. Physiol.* 160, 106–154.
- Hubel, D. H., and Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *J. Physiol.* 195, 215–243.
- Hummel, J. E., and Biederman, I. (1992). Dynamic binding in a neural network for shape recognition. *Psychol. Rev.* 99, 480–517.
- Puttenlocher, D. P., and Ullman, S. (1990). Recognizing solid objects by alignment with an image. *Int. J. Comput. Vis.* 5, 195–212.
- Ito, M. (1984). *The Cerebellum and Neural Control*. New York: Raven Press.
- Ito, M. (1989). Long-term depression. *Annu. Rev. Neurosci.* 12, 85–102.
- Ito, M., and Komatsu, H. (2004). Representation of angles embedded within contour stimuli in area V2 of macaque monkeys. *J. Neurosci.* 24, 3313–3324.
- Itti, L., and Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Res.* 40, 1489–1506.
- Jarrett, K., Kavukcuoglu, K., Ranzato, M., and Lecun, Y. (2009). “What is the best multi-stage architecture for object recognition?” in *2009 IEEE 12th International Conference on Computer Vision (ICCV)*, 2146–2153.
- Jiang, F., Dricot, L., Weber, J., Righi, G., Tarr, M. J., Goebel, R., and Rossion, B. (2011). Face categorization in visual scenes may start in a higher order area of the right fusiform gyrus: evidence from dynamic visual stimulation in neuroimaging. *J. Neurophysiol.* 106, 2720–2736.
- Koch, C. (1999). *Biophysics of Computation*. Oxford: Oxford University Press.
- Koenderink, J. J. (1990). *Solid Shape*. Cambridge, MA: MIT Press.
- Koenderink, J. J., and Van Doorn, A. J. (1979). The internal representation of solid shape with respect to vision. *Biol. Cybern.* 32, 211–217.
- Koenderink, J. J., and van Doorn, A. J. (1991). Affine structure from motion. *J. Opt. Soc. Am. A* 8, 377–385.
- Kourtzi, Z., and Connor, C. E. (2011). Neural representations for object perception: structure, category, and adaptive coding. *Annu. Rev. Neurosci.* 34, 45–67.
- Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., Tanaka, K., and Bandettini, P. A. (2008). Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron* 60, 1126–1141.
- Krieman, G., Koch, C., and Fried, I. (2000). Category-specific visual responses of single neurons in the human medial temporal lobe. *Nat. Neurosci.* 3, 946–953.
- Land, M. F. (1999). Motion and vision: why animals move their eyes. *J. Comp. Physiol. A* 185, 341–352.
- Land, M. F., and Collett, T. S. (1997). “A survey of active vision in invertebrates,” in *From Living Eyes to Seeing Machines*, eds M. V. Srinivasan

- and S. Venkatesh (Oxford: Oxford University Press), 16–36.
- LeCun, Y., Kavukcuoglu, K., and Faret, C. (2010). “Convolutional networks and applications in vision,” in *2010 IEEE International Symposium on Circuits and Systems*, 253–256.
- Lee, T. S. (1996). Image representation using 2D Gabor wavelets. *IEEE Trans. Pattern Anal. Mach. Intell.* 18, 959–971.
- Leen, T. K. (1995). From data distributions to regularization in invariant learning. *Neural Comput.* 7, 974–981.
- Leibo, J. Z., Mutch, J., Rosasco, L., Ullman, S., and Poggio, T. (2010). *Learning Generic Invariances in Object Recognition: Translation and Scale*. MIT-CSAIL-TR-2010-061, Cambridge.
- Li, N., and DiCarlo, J. J. (2008). Unsupervised natural experience rapidly alters invariant object representation in visual cortex. *Science* 321, 1502–1507.
- Li, S., Mayhew, S. D., and Kourtzi, Z. (2011). Learning shapes spatiotemporal brain patterns for flexible categorical decisions. *Cereb. Cortex*. doi: 10.1093/cercor/bhr309. [Epub ahead of print].
- Liu, J., Harris, A., and Kanwisher, N. (2010). Perception of face parts and face configurations: an fMRI study. *J. Cogn. Neurosci.* 22, 203–211.
- Logothetis, N. K., Pauls, J., Bulthoff, H. H., and Poggio, T. (1994). View-dependent object recognition by monkeys. *Curr. Biol.* 4, 401–414.
- Logothetis, N. K., Pauls, J., and Poggio, T. (1995). Shape representation in the inferior temporal cortex of monkeys. *Curr. Biol.* 5, 552–563.
- Logothetis, N. K., and Sheinberg, D. L. (1996). Visual object recognition. *Annu. Rev. Neurosci.* 19, 577–621.
- Lowe, D. (1985). *Perceptual Organization and Visual Recognition*. Boston: Kluwer.
- Marr, D. (1982). *Vision*. San Francisco: Freeman.
- Marr, D., and Nishihara, H. K. (1978). Representation and recognition of the spatial organization of three dimensional structure. *Proc. R. Soc. Lond. B Biol. Sci.* 200, 269–294.
- McNaughton, B. L., Barnes, C. A., and O’Keefe, J. (1983). The contributions of position, direction, and velocity to single unit activity in the hippocampus of freely-moving rats. *Exp. Brain Res.* 52, 41–49.
- Mel, B. W. (1997). SEEMORE: combining color, shape, and texture histogramming in a neurally-inspired approach to visual object recognition. *Neural Comput.* 9, 777–804.
- Mel, B. W., and Fiser, J. (2000). Minimizing binding errors using learned conjunctive features. *Neural Comput.* 12, 731–762.
- Mel, B. W., Ruderman, D. L., and Archie, K. A. (1998). Translation-invariant orientation tuning in visual “complex” cells could derive from intradendritic computations. *J. Neurosci.* 18, 4325–4334.
- Mikami, A., Nakamura, K., and Kubota, K. (1994). Neuronal responses to photographs in the superior temporal sulcus of the rhesus monkey. *Behav. Brain Res.* 60, 1–13.
- Milner, P. (1974). A model for visual shape recognition. *Psychol. Rev.* 81, 521–535.
- Miyashita, Y. (1988). Neuronal correlate of visual associative long-term memory in the primate temporal cortex. *Nature* 335, 817–820.
- Miyashita, Y., and Chang, H. S. (1988). Neuronal correlate of pictorial short-term memory in the primate temporal cortex. *Nature* 331, 68–70.
- Montague, P. R., Gally, J. A., and Edelman, G. M. (1991). Spatial signalling in the development and function of neural connections. *Cereb. Cortex* 1, 199–220.
- Moser, E. I. (2004). Hippocampal place cells demand attention. *Neuron* 42, 183–185.
- Moser, M. B., and Moser, E. I. (1998). Functional differentiation in the hippocampus. *Hippocampus* 8, 608–619.
- Movshon, J. A., Adelson, E. H., Gizzi, M. S., and Newsome, W. T. (1985). “The analysis of moving visual patterns,” in *Pattern Recognition Mechanisms*, eds C. Chagas, R. Gattass, and C. G. Gross (New York: Springer-Verlag), 117–151.
- Mozar, M. C. (1991). *The Perception of Multiple Objects: A Connectionist Approach*. Cambridge, MA: MIT Press.
- Muller, R. U., Kubie, J. L., Bostock, E. M., Taube, J. S., and Quirk, G. J. (1991). “Spatial firing correlates of neurons in the hippocampal formation of freely moving rats,” in *Brain and Space*, ed. J. Paillard (Oxford: Oxford University Press), 296–333.
- Mundy, J., and Zisserman, A. (1992). “Introduction – towards a new framework for vision,” in *Geometric Invariance in Computer Vision*, eds J. Mundy and A. Zisserman (Cambridge, MA: MIT Press), 1–39.
- Mutch, J., and Lowe, D. G. (2008). Object class recognition and localization using sparse features with limited receptive fields. *Int. J. Comput. Vis.* 80, 45–57.
- Newsome, W. T., Britten, K. H., and Movshon, J. A. (1989). Neuronal correlates of a perceptual decision. *Nature* 341, 52–54.
- Oja, E. (1982). A simplified neuron model as a principal component analyzer. *J. Math. Biol.* 15, 267–273.
- O’Keefe, J. (1984). “Spatial memory within and without the hippocampal system,” in *Neurobiology of the Hippocampus*, ed. W. Seifert (London: Academic Press), 375–403.
- Olshausen, B. A., Anderson, C. H., and Van Essen, D. C. (1993). A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *J. Neurosci.* 13, 4700–4719.
- Olshausen, B. A., Anderson, C. H., and Van Essen, D. C. (1995). A multiscale dynamic routing circuit for forming size- and position-invariant object representations. *J. Comput. Neurosci.* 2, 45–62.
- Orban, G. A. (2011). The extraction of 3D shape in the visual system of human and nonhuman primates. *Annu. Rev. Neurosci.* 34, 361–388.
- O’Reilly, J., and Munakata, Y. (2000). *Computational Explorations in Cognitive Neuroscience*. Cambridge, MA: MIT Press.
- Parga, N., and Rolls, E. T. (1998). Transform invariant recognition by association in a recurrent network. *Neural Comput.* 10, 1507–1525.
- Peng, H. C., Sha, L. F., Gan, Q., and Wei, Y. (1998). Energy function for learning invariance in multi-layer perceptron. *Electron. Lett.* 34, 292–294.
- Perrett, D. I., and Oram, M. W. (1993). Neurophysiology of shape processing. *Image Vis. Comput.* 11, 317–333.
- Perrett, D. I., Rolls, E. T., and Caan, W. (1982). Visual neurons responsive to faces in the monkey temporal cortex. *Exp. Brain Res.* 47, 329–342.
- Perrett, D. I., Smith, P. A. J., Potter, D. D., Mistlin, A. J., Head, A. S., Milner, D., and Jeeves, M. A. (1985). Visual cells in temporal cortex sensitive to face view and gaze direction. *Proc. R. Soc. Lond. B Biol. Sci.* 223, 293–317.
- Perry, G., Rolls, E. T., and Stringer, S. M. (2006). Spatial vs temporal continuity in view invariant visual object recognition learning. *Vision Res.* 46, 3994–4006.
- Perry, G., Rolls, E. T., and Stringer, S. M. (2010). Continuous transformation learning of translation invariant representations. *Exp. Brain Res.* 204, 255–270.
- Pinto, N., Cox, D. D., and DiCarlo, J. J. (2008). Why is real-world visual object recognition hard? *PLoS Comput. Biol.* 4, e27. doi:10.1371/journal.pcbi.0040027
- Poggio, T., and Edelman, S. (1990). A network that learns to recognize three-dimensional objects. *Nature* 343, 263–266.
- Pollen, D., and Ronner, S. (1981). Phase relationship between adjacent simple cells in the visual cortex. *Science* 212, 1409–1411.
- Rao, R. P. N., and Ruderman, D. L. (1999). “Learning lie groups for invariant visual perception,” in *Advances in Neural Information Processing Systems*, Vol. 11, eds M. S. Kearns, S. A. Solla, and D. A. Cohn (Cambridge: MIT Press), 810–816.
- Renart, A., Parga, N., and Rolls, E. T. (2000). “A recurrent model of the interaction between the prefrontal cortex and inferior temporal cortex in delay memory tasks,” in *Advances in Neural Information Processing Systems*, Vol. 12, eds S. Solla, T. Leen, and K.-R. Mueller (Cambridge, MA: MIT Press), 171–177.
- Rhodes, P. (1992). The open time of the NMDA channel facilitates the self-organization of invariant object responses in cortex. *Soc. Neurosci. Abstr.* 18, 740.
- Riesenhuber, M., and Poggio, T. (1998). “Just one view: invariances in inferotemporal cell tuning,” in *Advances in Neural Information Processing Systems*, Vol. 10, eds M. I. Jordan, M. J. Kearns, and S. A. Solla (Cambridge, MA: MIT Press), 215–221.
- Riesenhuber, M., and Poggio, T. (1999a). Are cortical models really bound by the “binding problem”? *Neuron* 24, 87–93.
- Riesenhuber, M., and Poggio, T. (1999b). Hierarchical models of object recognition in cortex. *Nat. Neurosci.* 2, 1019–1025.
- Riesenhuber, M., and Poggio, T. (2000). Models of object recognition. *Nat. Neurosci.* 3(Suppl.), 1199–1204.
- Ringach, D. L. (2002). Spatial structure and symmetry of simple-cell receptive fields in macaque primary visual cortex. *J. Neurophysiol.* 88, 455–463.
- Robertson, R. G., Rolls, E. T., and Georges-François, P. (1998). Spatial view cells in the primate hippocampus: effects of removal of view details. *J. Neurophysiol.* 79, 1145–1156.
- Rolls, E. T. (1989a). “Functions of neuronal networks in the hippocampus and neocortex in memory,” in *Neural Models of Plasticity: Experimental and Theoretical Approaches*,

- Chap. 13, eds J. H. Byrne and W. O. Berry (San Diego, CA: Academic Press), 240–265.
- Rolls, E. T. (1989b). “The representation and storage of information in neuronal networks in the primate cerebral cortex and hippocampus,” in *The Computing Neuron*, Chap. 8, eds R. Durbin, C. Miall, and G. Mitchison (Wokingham: Addison-Wesley), 125–159.
- Rolls, E. T. (1992). Neurophysiological mechanisms underlying face processing within and beyond the temporal cortical visual areas. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 335, 11–21.
- Rolls, E. T. (1994). Brain mechanisms for invariant visual recognition and learning. *Behav. Processes* 33, 113–138.
- Rolls, E. T. (1995). Learning mechanisms in the temporal lobe visual cortex. *Behav. Brain Res.* 66, 177–185.
- Rolls, E. T. (1999). *The Brain and Emotion*. Oxford: Oxford University Press.
- Rolls, E. T. (2000). Functions of the primate temporal lobe cortical visual areas in invariant visual object and face recognition. *Neuron* 27, 205–218.
- Rolls, E. T. (2003). Consciousness absent and present: a neurophysiological exploration. *Prog. Brain Res.* 144, 95–106.
- Rolls, E. T. (2005). *Emotion Explained*. Oxford: Oxford University Press.
- Rolls, E. T. (2006). “Consciousness absent and present: a neurophysiological exploration of masking,” in *The First Half Second*, Chap. 6, eds H. Ogmen and B. G. Breitmeyer (Cambridge, MA: MIT Press), 89–108.
- Rolls, E. T. (2007a). “Invariant representations of objects in natural scenes in the temporal cortex visual areas,” in *Representation and Brain*, Chap. 3, ed. S. Funahashi (Tokyo: Springer), 47–102.
- Rolls, E. T. (2007b). The representation of information about faces in the temporal and frontal lobes of primates including humans. *Neuropsychologia* 45, 124–143.
- Rolls, E. T. (2007c). Sensory processing in the brain related to the control of food intake. *Proc. Nutr. Soc.* 66, 96–112.
- Rolls, E. T. (2008a). Face representations in different brain areas, and critical band masking. *J. Neuropsychol.* 2, 325–360.
- Rolls, E. T. (2008b). *Memory, Attention, and Decision-Making. A Unifying Computational Neuroscience Approach*. Oxford: Oxford University Press.
- Rolls, E. T. (2008c). Top-down control of visual perception: attention in natural vision. *Perception* 37, 333–354.
- Rolls, E. T. (2011a). David Marr’s vision: floreat computational neuroscience. *Brain* 134, 913–916.
- Rolls, E. T. (2011b). “Face neurons,” in *The Oxford Handbook of Face Perception*, Chap. 4, eds A. J. Calder, G. Rhodes, M. H. Johnson, and J. V. Haxby (Oxford: Oxford University Press), 51–75.
- Rolls, E. T. (2012). *Neuroculture: On the Implications of Brain Science*. Oxford: Oxford University Press.
- Rolls, E. T. (2013). *Emotion and Decision-Making Explained*. Oxford: Oxford University Press.
- Rolls, E. T., Aggelopoulos, N. C., Franco, L., and Treves, A. (2004). Information encoding in the inferior temporal visual cortex: contributions of the firing rates and the correlations between the firing of neurons. *Biol. Cybern.* 90, 19–32.
- Rolls, E. T., Aggelopoulos, N. C., and Zheng, F. (2003). The receptive fields of inferior temporal cortex neurons in natural scenes. *J. Neurosci.* 23, 339–348.
- Rolls, E. T., and Baylis, G. C. (1986). Size and contrast have only small effects on the responses to faces of neurons in the cortex of the superior temporal sulcus of the monkey. *Exp. Brain Res.* 65, 38–48.
- Rolls, E. T., Baylis, G. C., Hasselmo, M., and Nalwa, V. (1989). “The representation of information in the temporal lobe visual cortical areas of macaque monkeys,” in *Seeing Contour and Colour*, eds J. Kulikowski, C. Dickinson, and I. Murray (Oxford: Pergamon).
- Rolls, E. T., Baylis, G. C., and Hasselmo, M. E. (1987). The responses of neurons in the cortex in the superior temporal sulcus of the monkey to band-pass spatial frequency filtered faces. *Vision Res.* 27, 311–326.
- Rolls, E. T., Baylis, G. C., and Leonard, C. M. (1985). Role of low and high spatial frequencies in the face-selective responses of neurons in the cortex in the superior temporal sulcus. *Vision Res.* 25, 1021–1035.
- Rolls, E. T., and Deco, G. (2002). *Computational Neuroscience of Vision*. Oxford: Oxford University Press.
- Rolls, E. T., and Deco, G. (2006). Attention in natural scenes: neurophysiological and computational bases. *Neural Netw.* 19, 1383–1394.
- Rolls, E. T., Franco, L., Aggelopoulos, N. C., and Jerez, J. M. (2006a). Information in the first spike, the order of spikes, and the number of spikes provided by neurons in the inferior temporal visual cortex. *Vision Res.* 46, 4193–4205.
- Rolls, E. T., Stringer, S. M., and Elliot, T. (2006b). Entorhinal cortex grid cells can map to hippocampal place cells by competitive learning. *Network* 17, 447–465.
- Rolls, E. T., Franco, L., and Stringer, S. M. (2005a). The perirhinal cortex and long-term familiarity memory. *Q. J. Exp. Psychol. B.* 58, 234–245.
- Rolls, E. T., Xiang, J.-Z., and Franco, L. (2005b). Object, space and object-space representations in the primate hippocampus. *J. Neurophysiol.* 94, 833–844.
- Rolls, E. T., and Grabenhorst, F. (2008). The orbitofrontal cortex and beyond: from affect to decision-making. *Prog. Neurobiol.* 86, 216–244.
- Rolls, E. T., and Milward, T. (2000). A model of invariant object recognition in the visual system: learning rules, activation functions, lateral inhibition, and information-based performance measures. *Neural Comput.* 12, 2547–2572.
- Rolls, E. T., Robertson, R. G., and Georges-François, P. (1997a). Spatial view cells in the primate hippocampus. *Eur. J. Neurosci.* 9, 1789–1794.
- Rolls, E. T., Treves, A., and Tovee, M. J. (1997b). The representational capacity of the distributed encoding of information provided by populations of neurons in the primate temporal visual cortex. *Exp. Brain Res.* 114, 149–162.
- Rolls, E. T., Treves, A., Tovee, M., and Panzeri, S. (1997c). Information in the neuronal representation of individual stimuli in the primate temporal visual cortex. *J. Comput. Neurosci.* 4, 309–333.
- Rolls, E. T., and Stringer, S. M. (2000). On the design of neural networks in the brain by genetic evolution. *Prog. Neurobiol.* 61, 557–579.
- Rolls, E. T., and Stringer, S. M. (2001). Invariant object recognition in the visual system with error correction and temporal difference learning. *Network* 12, 111–129.
- Rolls, E. T., and Stringer, S. M. (2006). Invariant visual object recognition: a model, with lighting invariance. *J. Physiol. Paris* 100, 43–62.
- Rolls, E. T., and Stringer, S. M. (2007). Invariant global motion recognition in the dorsal visual system: a unifying theory. *Neural Comput.* 19, 139–169.
- Rolls, E. T., and Tovee, M. J. (1994). Processing speed in the cerebral cortex and the neurophysiology of visual masking. *Proc. R. Soc. Lond. B Biol. Sci.* 257, 9–15.
- Rolls, E. T., and Tovee, M. J. (1995a). The responses of single neurons in the temporal visual cortical areas of the macaque when more than one stimulus is present in the visual field. *Exp. Brain Res.* 103, 409–420.
- Rolls, E. T., and Tovee, M. J. (1995b). Sparseness of the neuronal representation of stimuli in the primate temporal visual cortex. *J. Neurophysiol.* 73, 713–726.
- Rolls, E. T., Tovee, M. J., and Panzeri, S. (1999). The neurophysiology of backward visual masking: information analysis. *J. Cogn. Neurosci.* 11, 335–346.
- Rolls, E. T., Tovee, M. J., Purcell, D. G., Stewart, A. L., and Azzopardi, P. (1994). The responses of neurons in the temporal cortex of primates, and face identification and detection. *Exp. Brain Res.* 101, 474–484.
- Rolls, E. T., and Treves, A. (1998). *Neural Networks and Brain Function*. Oxford: Oxford University Press.
- Rolls, E. T., and Treves, A. (2011). The neuronal encoding of information in the brain. *Prog. Neurobiol.* 95, 448–490.
- Rolls, E. T., Treves, A., Robertson, R. G., Georges-François, P., and Panzeri, S. (1998). Information about spatial view in an ensemble of primate hippocampal cells. *J. Neurophysiol.* 79, 1797–1813.
- Rolls, E. T., Tromans, J. M., and Stringer, S. M. (2008). Spatial scene representations formed by self-organizing learning in a hippocampal extension of the ventral visual system. *Eur. J. Neurosci.* 28, 2116–2127.
- Rolls, E. T., Webb, T. J., and Deco, G. (2012). Communication before coherence. *Eur. J. Neurosci.* (in press).
- Rolls, E. T., and Xiang, J.-Z. (2006). Spatial view cells in the primate hippocampus, and memory recall. *Rev. Neurosci.* 17, 175–200.
- Rosenblatt, F. (1961). *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*. Washington, DC: Spartan.
- Sakai, K., and Miyashita, Y. (1991). Neural organisation for the long-term memory of paired associates. *Nature* 354, 152–155.



- Sato, T. (1989). Interactions of visual stimuli in the receptive fields of inferior temporal neurons in macaque. *Exp. Brain Res.* 77, 23–30.
- Selfridge, O. G. (1959). "Pandemonium: a paradigm for learning," in *The Mechanization of Thought Processes*, eds D. Blake and A. Uttley (London: H. M. Stationery Office), 511–529.
- Serre, T., Kreiman, G., Kouh, M., Cadieu, C., Knoblich, U., and Poggio, T. (2007a). A quantitative theory of immediate visual recognition. *Prog. Brain Res.* 165, 33–56.
- Serre, T., Oliva, A., and Poggio, T. (2007b). A feedforward architecture accounts for rapid categorization. *Proc. Natl. Acad. Sci. U.S.A.* 104, 6424–6429.
- Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., and Poggio, T. (2007c). Robust object recognition with cortex-like mechanisms. *IEEE Trans. Pattern Anal. Mach. Intell.* 29, 411–426.
- Shadlen, M. N., and Movshon, J. A. (1999). Synchrony unbound: a critical evaluation of the temporal binding hypothesis. *Neuron* 24, 67–77.
- Shashua, A. (1995). Algebraic functions for recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 17, 779–789.
- Shevelev, I. A., Novikova, R. V., Lazareva, N. A., Tikhomirov, A. S., and Sharaev, G. A. (1995). Sensitivity to cross-like figures in cat striate neurons. *Neuroscience* 69, 51–57.
- Sillito, A. M., Grieve, K. L., Jones, H. E., Cudeiro, J., and Davis, J. (1995). Visual cortical mechanisms detecting focal orientation discontinuities. *Nature* 378, 492–496.
- Singer, W. (1999). Neuronal synchrony: a versatile code for the definition of relations? *Neuron* 24, 49–65.
- Singer, W., Gray, C., Engel, A., König, P., Artaola, A., and Brocher, S. (1990). Formation of cortical cell assemblies. *Cold Spring Harb. Symp. Quant. Biol.* 55, 939–952.
- Singer, W., and Gray, C. M. (1995). Visual feature integration and the temporal correlation hypothesis. *Annu. Rev. Neurosci.* 18, 555–586.
- Spiridon, M., Fischl, B., and Kanwisher, N. (2006). Location and spatial profile of category-specific regions in human extrastriate cortex. *Hum. Brain Mapp.* 27, 77–89.
- Spruston, N., Jonas, P., and Sakmann, B. (1995). Dendritic glutamate receptor channel in rat hippocampal CA3 and CA1 pyramidal neurons. *J. Physiol.* 482, 325–352.
- Stan-Kiewicz, B., and Hummel, J. (1994). "Metricat: a representation for basic and subordinate-level classification," in *Proceedings of the 18th Annual Conference of the Cognitive Science Society*, ed. G. W. Cottrell (San Diego: Erlbaum), 254–259.
- Stringer, S. M., Perry, G., Rolls, E. T., and Proske, J. H. (2006). Learning invariant object recognition in the visual system with continuous transformations. *Biol. Cybern.* 94, 128–142.
- Stringer, S. M., and Rolls, E. T. (2000). Position invariant recognition in the visual system with cluttered environments. *Neural Netw.* 13, 305–315.
- Stringer, S. M., and Rolls, E. T. (2002). Invariant object recognition in the visual system with novel views of 3D objects. *Neural Comput.* 14, 2585–2596.
- Stringer, S. M., and Rolls, E. T. (2008). Learning transform invariant object recognition in the visual system with multiple stimuli present during training. *Neural Netw.* 21, 888–903.
- Stringer, S. M., Rolls, E. T., and Tromans, J. M. (2007). Invariant object recognition with trace learning and multiple stimuli present during training. *Network* 18, 161–187.
- Sutherland, N. S. (1968). Outline of a theory of visual pattern recognition in animal and man. *Proc. R. Soc. Lond., B, Biol. Sci.* 171, 297–317.
- Sutton, R. S. (1988). Learning to predict by the methods of temporal differences. *Mach. Learn.* 3, 9–44.
- Sutton, R. S., and Barto, A. G. (1981). Towards a modern theory of adaptive networks: expectation and prediction. *Psychol. Rev.* 88, 135–170.
- Sutton, R. S., and Barto, A. G. (1998). *Reinforcement Learning*. Cambridge, MA: MIT Press.
- Tanaka, K. (1993). Neuronal mechanisms of object recognition. *Science* 262, 685–688.
- Tanaka, K. (1996). Inferotemporal cortex and object vision. *Annu. Rev. Neurosci.* 19, 109–139.
- Tanaka, K., Saito, C., Fukada, Y., and Moriya, M. (1990). "Integration of form, texture, and color information in the inferotemporal cortex of the macaque," in *Vision, Memory and the Temporal Lobe*, Chap. 10, eds E. Iwai and M. Mishkin (New York: Elsevier), 101–109.
- Tanaka, K., Saito, H., Fukada, Y., and Moriya, M. (1991). Coding visual images of objects in the inferotemporal cortex of the macaque monkey. *J. Neurophysiol.* 66, 170–189.
- Tou, J. T., and Gonzalez, A. G. (1974). *Pattern Recognition Principles*. Reading, MA: Addison-Wesley.
- Tovee, M. J., and Rolls, E. T. (1995). Information encoding in short firing rate epochs by single neurons in the primate temporal visual cortex. *Vis. Cogn.* 2, 35–58.
- Tovee, M. J., Rolls, E. T., and Azzopardi, P. (1994). Translation invariance and the responses of neurons in the temporal visual cortical areas of primates. *J. Neurophysiol.* 72, 1049–1060.
- Tovee, M. J., Rolls, E. T., and Ramachandran, V. S. (1996). Rapid visual learning in neurons of the primate temporal visual cortex. *Neuroreport* 7, 2757–2760.
- Tovee, M. J., Rolls, E. T., Treves, A., and Bellis, R. P. (1993). Information encoding and the responses of single neurons in the primate temporal visual cortex. *J. Neurophysiol.* 70, 640–654.
- Trappenberg, T. P., Rolls, E. T., and Stringer, S. M. (2002). "Effective size of receptive fields of inferior temporal visual cortex neurons in natural scenes," in *Advances in Neural Information Processing Systems*, Vol. 14, eds T. G. Dietterich, S. Becker, and Z. Ghahramani (Cambridge, MA: MIT Press), 293–300.
- Treves, A., and Rolls, E. T. (1991). What determines the capacity of autoassociative memories in the brain? *Network* 2, 371–397.
- Tromans, J. M., Harris, M., and Stringer, S. M. (2011). A computational model of the development of separate representations of facial identity and expression in the primate visual system. *PLoS ONE* 6, e25616. doi:10.1371/journal.pone.0025616
- Tromans, J. M., Page, J. I., and Stringer, S. M. (2012). Learning separate visual representations of independently rotating objects. *Network*. PMID: 22364581. [Epub ahead of print].
- Tsao, D. Y., Freiwald, W. A., Tootell, R. B., and Livingstone, M. S. (2006). A cortical region consisting entirely of face-selective cells. *Science* 311, 617–618.
- Tsao, D. Y., and Livingstone, M. S. (2008). Mechanisms of face perception. *Annu. Rev. Neurosci.* 31, 411–437.
- Tsodyks, M. V., and Feigel'man, M. V. (1988). The enhanced storage capacity in neural networks with low activity level. *Europhys. Lett.* 6, 101–105.
- Ullman, S. (1996). *High-Level Vision, Object Recognition and Visual Cognition*. Cambridge, MA: Bradford/MIT Press.
- Ullman, S. (2007). Object recognition and segmentation by a fragment-based hierarchy. *Trends Cogn. Sci. (Regul. Ed.)* 11, 58–64.
- Van Essen, D., Anderson, C. H., and Felleman, D. J. (1992). Information processing in the primate visual system: an integrated systems perspective. *Science* 255, 419–423.
- Vogels, R., and Biederman, I. (2002). Effects of illumination intensity and direction on object coding in macaque inferior temporal cortex. *Cereb. Cortex* 12, 756–766.
- von der Malsburg, C. (1973). Self-organization of orientation-sensitive columns in the striate cortex. *Kybernetik* 14, 85–100.
- von der Malsburg, C. (1990). "A neural architecture for the representation of scenes," in *Brain Organization and Memory: Cells, Systems and Circuits*, Chap. 18, eds J. L. McGaugh, N. M. Weinburger, and G. Lynch (Oxford: Oxford University Press), 356–372.
- Wallis, G., and Baddeley, R. (1997). Optimal unsupervised learning in invariant object recognition. *Neural Comput.* 9, 883–894.
- Wallis, G., and Bülthoff, H. (1999). Learning to recognize objects. *Trends Cogn. Sci. (Regul. Ed.)* 3, 22–31.
- Wallis, G., and Rolls, E. T. (1997). Invariant face and object recognition in the visual system. *Prog. Neurobiol.* 51, 167–194.
- Wallis, G., Rolls, E. T., and Foldiak, P. (1993). Learning invariant responses to the natural transformations of objects. *Proc. Int. Jt. Conf. Neural Netw.* 2, 1087–1090.
- Wasserman, E., Kirkpatrick-Steger, A., and Biederman, I. (1998). Effects of geon deletion, scrambling, and movement on picture identification in pigeons. *J. Exp. Psychol. Anim. Behav. Process.* 24, 34–46.
- Watanabe, S., Lea, S. E. G., and Dittrich, W. H. (1993). "What can we learn from experiments on pigeon discrimination?" in *Vision, Brain, and Behavior in Birds*, eds H. P. Zeigler and H.-J. Bischof (Cambridge, MA: MIT Press), 351–376.
- Weiner, K. S., and Grill-Spector, K. (2011). Neural representations of faces and limbs neighbor in human high-level visual cortex: evidence for a new organization principle. *Psychol. Res.* PMID: 22139022. [Epub ahead of print].
- Widrow, B., and Hoff, M. E. (1960). "Adaptive switching circuits," in *1960 IRE WESCON Convention Record, Part 4* (New York: IRE), 96–104. [Reprinted in Anderson and Rosenfeld, 1988].
- Widrow, B., and Stearns, S. D. (1985). *Adaptive Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall.
- Winston, P. H. (1975). "Learning structural descriptions from examples," in *The Psychology of Computer Vision*, ed. P. H. Winston (New York: McGraw-Hill), 157–210.



- Wiskott, L. (2003). Slow feature analysis: a theoretical analysis of optimal free responses. *Neural Comput.* 15, 2147–2177.
- Wiskott, L., and Sejnowski, T. J. (2002). Slow feature analysis: unsupervised learning of invariances. *Neural Comput.* 14, 715–770.
- Womelsdorf, T., Schoffelen, J. M., Oostenveld, R., Singer, W., Desimone, R., Engel, A. K., and Fries, P. (2007). Modulation of neuronal interactions through neuronal synchronization. *Science* 316, 1609–1612.
- Wurtz, R. H., and Kandel, E. R. (2000a). “Central visual pathways,” in *Principles of Neural Science*, 4th Edn, Chap. 27, eds E. R. Kandel, J. H. Schwartz, and T. M. Jessell (New York: McGraw-Hill), 543–547.
- Wurtz, R. H., and Kandel, E. R. (2000b). “Perception of motion depth and form,” in *Principles of Neural Science*, 4th Edn, Chap. 28, eds E. R. Kandel, J. H. Schwartz, and T. M. Jessell (New York: McGraw-Hill), 548–571.
- Wyss, R., Konig, P., and Verschure, P. F. (2006). A model of the ventral visual system based on temporal stability and local memory. *PLoS Biol.* 4, e120. doi:10.1371/journal.pbio.0040120
- Yamane, S., Kaji, S., and Kawano, K. (1988). What facial features activate face neurons in the inferotemporal cortex of the monkey? *Exp. Brain Res.* 73, 209–214.
- Yamane, Y., Carlson, E. T., Bowman, K. C., Wang, Z., and Connor, C. E. (2008). A neural code for three-dimensional object shape in macaque inferotemporal cortex. *Nat. Neurosci.* 11, 1352–1360.
- Yi, D. J., Turk-Browne, N. B., Flombaum, J. I., Kim, M. S., Scholl, B. J., and Chun, M. M. (2008). Spatiotemporal object continuity in human ventral visual cortex. *Proc. Natl. Acad. Sci. U.S.A.* 105, 8840–8845.
- Zhao, Q., and Koch, C. (2011). Learning a saliency map using fixated locations in natural scenes. *J. Vis.* 11, 9.
- Zucker, S. W., Dobbins, A., and Iverson, L. (1989). Two stages of curve detection suggest two styles of visual computation. *Neural Comput.* 1, 68–81.
- Conflict of Interest Statement:** The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 01 November 2011; accepted: 23 May 2012; published online: 19 June 2012.

Citation: Rolls ET (2012) Invariant visual object and face recognition: neural and computational bases, and a model, VisNet. *Front. Comput. Neurosci.* 6:35. doi: 10.3389/fncom.2012.00035

Copyright © 2012 Rolls. This is an open-access article distributed under the terms of the Creative Commons Attribution Non Commercial License, which permits non-commercial use, distribution, and reproduction in other forums, provided the original authors and source are credited.



# Object recognition can be viewpoint dependent or invariant – it's just a matter of time and task

Branka Milivojevic<sup>1,2\*</sup>

<sup>1</sup> Donders Institute for Brain, Cognition and Behaviour, Radboud University Nijmegen, Nijmegen, Netherlands

<sup>2</sup> Experimental and Developmental Psychology, Utrecht University, Utrecht, Netherlands

\*Correspondence: branka.mili@gmail.com

As we move through our environment, we encounter familiar objects from various viewpoints. Despite the ensuing variability of the images projected onto the retina, we have seemingly little difficulty when it comes to recognizing objects we encounter. We can, however, see how the objects are oriented, suggesting that object recognition is to a certain degree dissociable from perception of other object “features” such as orientation. Changes in orientation of objects, particularly inversion, can also affect how we perceive the objects. A particularly illustrative example (shown in **Figure 1**) is that of the Thatcher illusion (Thompson, 1980), where the grotesque appearance of a face with its inverted eyes and mouth is “hidden” when the whole face is also inverted. The percept itself, therefore, is affected by the change in orientation. In addition, there are also subtle effects of viewpoint changes on object recognition itself. For example, identifying rotated objects is more difficult when they are briefly presented than when viewing time is unlimited (Lawson and Jolicoeur, 2003), and identifying a face is considerably more difficult the face has been inverted (Yin, 1969), as is discrimination between characters “b” and “d,” or “p” and “q” which requires (physical or mental) rotation of the characters to upright, before we can be certain which letter we are looking at (Corballis and McLaren, 1984).

These subtle, yet persistent, effects of viewpoint changes on perception and recognition arise as a consequence of how visual object processing is handled by the brain. Here, I discuss how neural mechanisms underlying visual processing give rise to perception and recognition which can be both viewpoint dependent and viewpoint invariant depending on the timing of those processes, as well as specific task demands or current “perceptual goals” of an individual. To do so, I will firstly explain how temporal dynamics of low-level visual processing may give rise to impaired

recognition at short viewing latencies and suggest that this may also relate to effects of viewpoint changes on perceptual experience. I will then discuss how the perceptual goals of an individual determines whether recognition is accomplished in viewpoint invariant or dependent manner with a particular focus on cognitive operations thought to be subserved by ventral and dorsal visual streams, namely object recognition and mental rotation, respectively.

## PERCEPTION IS AFFECTED BY POINT OF VIEW

Change in orientation must affect processing of visual information. For example, as our viewpoint changes, so does the shape of the image that falls on the retina. In the case of picture-plane rotations, the orientation of the edges of that shape will also change and thus stimulate different populations of orientation-tuned visually responsive neurons in primary visual cortex. However, these initial effects of orientation-changes on neural processing probably do not give rise to altered perceptual experience such as those associated with inversion of a Thatcherized face.

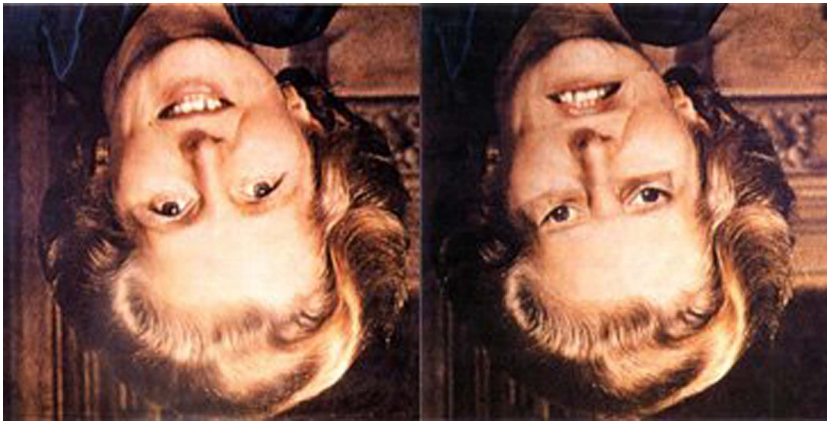
Inversion affects how we perceive the spatial relations between objects’ features and may, as James (1890) suggested, depend on perceptual experience with an object at a given orientation. This could explain why recognition of faces is particularly impaired by inversion: faces are most frequently seen the right way up, and are thought to be recognized using information about the configuration of the constituent features. As mirror reversal is also a special case of a configural change where the relative configuration of object’s features remains the same but reverses in its left–right orientation, this could also explain why mirror-images are difficult to tell apart when they are rotated away from a canonical viewpoint, and which is why we must rotate objects into alignment with our egocentric reference frames before

we can distinguish between parity-defined characters such as “b” and “d” (Corballis and McLaren, 1984). Interestingly, neural responses to unaltered and thatcherized images also follow the perceptual illusion and disappear as the face is rotated away from upright (Milivojevic et al., 2003a).

On neural level, large changes in the viewpoint of an object, such as inversion of faces (Rossion et al., 2000) and alphanumeric characters (Milivojevic et al., 2008), result in delays of the N170 component. The N170 is thought to reflect object classification, and inversion-related delays of N170 possibly reflect increases in time required to accumulate sufficient neural activity to reach a threshold at which recognition can occur (Perrett et al., 1998; Heekeren et al., 2008). If changes in viewpoint delay visual object encoding, this could explain why accurate recognition of rotated objects requires longer viewing times than recognition of canonically oriented objects (Jolicoeur and Landau, 1984; Lawson and Jolicoeur, 2003; Mack and Palmeri, 2011).

## VIEWPOINT MATTERS ONLY FOR SOME PERCEPTUAL GOALS

Task-dependent effect of viewpoint changes on neural processing are only observed around 250 ms after stimulus onset and coincide with the P2 component of the ERP. For example, if the observers need to determine whether a rotated alphanumeric character is normal or mirror-reversed, they will mentally rotate it to upright before making the decision. Although the beginning of mental rotation is later than the P2, parity decisions are associated with linear increases of P2 amplitudes while this is not the case for P2 preceding categorization of alphanumeric characters which does not require mental rotation (Milivojevic et al., 2011). Interestingly, similar increases in P2 amplitudes can be observed as a consequence of stimulus degradation, either by addition of noise (Banko et al., 2011) or by occlusion



**FIGURE 1 | Unaltered and “thatcherized” version of Margaret Thatcher’s face.** The grotesque appearance of the face when its eyes and mouth are inverted is hidden by the inversion of the whole image. Rotating the pictures to upright makes discrimination between the two versions of the face easier.

(Doniger et al., 2000), but not size transformation (Muthukumaraswamy et al., 2003), suggesting that changes in orientation degrade certain types of perceptual information which may be required for task-specific decision making, and may be, thus, associated with some form of perceptual decision making (Heekeren et al., 2008; Schendan and Lucia, 2009, 2010), such as whether sufficient information is available for the perceptual goal to be achieved. This decision would then trigger other visuospatial cognitive operations, such as mental rotation or more detailed inspection of individual features of an object. Those cognitive operations would lead to acquisition of additional information about the object which would, in turn, enable a more accurate completion of the perceptual task at hand. For the purpose of illustration, two types of “perceptual goals” that depend on object orientation will be described: object identification and parity-based recognition.

### IDENTIFICATION IS VIEWPOINT DEPENDENT BUT CATEGORISATION IS NOT

As already mentioned, face recognition is worse when faces are inverted (Yin, 1969), both in terms of reduced recognition accuracy and increased reaction times. This seems to be the case both for familiar and unfamiliar faces, and may be a consequence of disrupted neural processing underlying object classification although a causal relationship has not been firmly established. It should be noted here that faces are nevertheless recog-

nized as faces, what seems to be disrupted is the *identification* of the face as belonging to a particular person or *identification* of an emotional expression, while differentiation between categories of “face” and “non-face” objects is largely unimpaired by inversion.

The difference in viewpoint-sensitivity of identification and categorization has also been established for other classes of objects. For example, identifying letters of the alphabet is affected by character orientation while the same is not the case for between-category decisions such as letter–digit categorization (Corballis et al., 1978). In a sense, categorization may relate to recognition at a basic or entry level described by Roch (Rosch et al., 1976), while identification may be more closely related subordinate-level recognition. Object recognition at basic level (e.g., deciding a shape is a dog) are not affected by changes in viewpoint, while subordinate-level decisions (e.g., identifying a dog as a poodle) are affected by viewpoint changes in terms of reaction times and accuracy (Hamm and McMullen, 1998).

Studies which have directly compared identification and categorization of objects using neuroimaging methods are scarce. Nevertheless, studies investigating neural correlates of rotated-object categorization show little evidence of orientation-dependence at visual processing stages beyond the initial encoding of the objects (see above). In contrast, studies investigating rotated-object recognition either as identity-matching or in terms of explicit identification show that there is

an increase in activity in areas involved in object recognition within the inferior temporal cortex for various object classes such as faces (Haxby et al., 1999), bodies (Brandman and Yovel, 2010), landscapes (Epstein et al., 2006). Some authors have suggested that this increase in activity may reflect a shift in recognition strategy from one that is based on the whole shape to one that is based on the analysis of individual object features (i.e., details Jolicoeur, 1990).

### RECOGNIZING PARITY-DEFINED SHAPES REQUIRES MENTAL ROTATION

Decisions regarding the direction of the left–right axis of an object, or its *handedness*, require alignment between the object and our own egocentric frame of reference. For example, deciding whether a shoe is the left or the right one requires either physical or *mental* rotation of the shoe into alignment with our feet, or the feet with the shoe. The same holds for any object class that has a well-defined left–right orientation, such as alphanumeric characters, which can be readily recognized as “backward” if they have been mirror-reversed (Cooper and Shepard, 1973) – but only if they are presented at upright. Rotated characters require rotation to their canonical upright before we can notice if they are normal or backward, particularly if they are rotated by a large degree (Kung and Hamm, 2010). When the identity of an object depends on its left–right parity, as is the case with lower-case letters “b” and “d” or “p” and “q,” then the discrimination of such characters also requires rotation to upright before it can be successfully recognized (Corballis and McLaren, 1984).

This suggests that information regarding the identity of the object must be extracted before information about the handedness of an object can be determined. Although generally we need to recognize an object before mental rotation begins (Heil et al., 1996; Schendan and Lucia, 2009), this cannot be the case for objects whose identity depends on their handedness, such as “b” and “d” or “p” and “q.” With the exception of alphanumeric characters, there are not many commonly encountered objects whose identity is defined by parity (i.e., a hand is a hand irrespective of whether it is a left one or a right one) and those objects can be seen as special case whose identity cannot be determined at all orientations. For these objects, identification from a feature-based

descriptor such as “a semi-circle attached at an end of a long stem” could lead to selection of possible four candidates, and the remaining possibilities would need to be resolved with mental rotation.

Mental rotation has been associated with linear increases in centro-parietal negativity between ~400 and 800 ms after stimulus onset (e.g., Milivojevic et al., 2009b) which last somewhat longer for larger angular departures from upright (Milivojevic et al., 2003b; Hamm et al., 2004). The ERP correlates of mental rotation are probably generated by a distributed network of sources localized (Milivojevic et al., 2009b) within a network of prefrontal and posterior parietal areas which has been identified using fMRI (e.g., Milivojevic et al., 2009a). Whether these areas also subserve recognition of rotated parity-defined objects is still unclear as this particular question has not been investigated using neuroimaging.

## SUMMARY AND CONCLUSION

Although changes in viewpoint rarely interfere with common perceptual goals, such as categorizing objects into basic categories, this type of viewpoint invariant recognition can only be achieved after initial viewpoint-dependent neural processing has been accomplished. Depending on current perceptual goals, changes in viewpoint may impose certain recognition costs, observable in terms of increased response latencies or reduced accuracy. These costs are likely to reflect increased cognitive demands associated with recognition of misoriented shapes such as detailed analysis of object features or mental rotation of the shape to its canonical upright. In this sense, recognition of objects will always be affected by changes in viewpoint early on in the visual processing stream, but these effects will taper off with time. At later visual processing stages, some types of perceptual goals such as object identification or parity discrimination, will require additional processing operations which will give rise to viewpoint dependent behavioral performance.

## ACKNOWLEDGMENTS

I would like to thank Michael Corballis, Jeff Hamm, and Maarten Boksem for their helpful comments regarding earlier versions of the manuscript.

## REFERENCES

- Banko, E. M., Gal, V., Kortvelyes, J., Kovacs, G., and Vidnyanszky, Z. (2011). Dissociating the effect of noise on sensory processing and overall decision difficulty. *J. Neurosci.* 31, 2663–2674.
- Brandman, T., and Yovel, G. (2010). The body-inversion effect is mediated by face-selective, not body-selective, mechanisms. *J. Neurosci.* 30, 10534–10540.
- Cooper, L. A., and Shepard, R. N. (1973). “Chronometric studies of the rotation of mental images,” in *Visual Information Processing*, ed. W. G. Chase (New York: Academic Press), 75–176.
- Corballis, M. C., and McLaren, R. (1984). Winding one's ps and qs: mental rotation and mirror-image discrimination. *J. Exp. Psychol. Hum. Percept. Perform.* 10, 318–327.
- Corballis, M. C., Zbrodoff, N. J., Shetzer, L. I., and Butler, P. B. (1978). Decisions about identity and orientation of rotated letters and digits. *Mem. Cognit.* 6, 98–107.
- Doniger, G. M., Foxe, J. J., Murray, M. M., Higgins, B. A., Snodgrass, J. G., Schroeder, C. E., and Javitt, D. C. (2000). Activation timecourse of ventral visual stream object-recognition areas: high density electrical mapping of perceptual closure processes. *J. Cogn. Neurosci.* 12, 615–621.
- Epstein, R. A., Higgins, J. S., Parker, W., Aguirre, G. K., and Cooperman, S. (2006). Cortical correlates of face and scene inversion: a comparison. *Neuropsychologia* 44, 1145–1158.
- Hamm, J. P., Johnson, B. W., and Corballis, M. C. (2004). One good turn deserves another: an event-related brain potential study of rotated mirror-normal letter discriminations. *Neuropsychologia* 42, 810–820.
- Hamm, J. P., and McMullen, P. A. (1998). Effects of orientation on the identification of rotated objects depend on the level of identity. *J. Exp. Psychol. Hum. Percept. Perform.* 24, 413–426.
- Haxby, J. V., Ungerleider, L. G., Clark, V. P., Schouten, J. L., Hoffman, E. A., and Martin, A. (1999). The effect of face inversion on activity in human neural systems for face and object perception. *Neuron* 22, 189–199.
- Heekeren, H. R., Marrett, S., and Ungerleider, L. G. (2008). The neural systems that mediate human perceptual decision making. *Nat. Rev. Neurosci.* 9, 467–479.
- Heil, M., Bajric, J., Rösler, F., and Hennighausen, E. (1996). Event-related potentials during mental rotation: disentangling the contributions of character classification and image transformation. *J. Psychophysiol.* 10, 326–335.
- James, W. (1890). *Principles of Psychology*. London: Macmillan.
- Jolicoeur, P. (1990). Identification and disoriented objects: a dual systems theory. *Mind Lang.* 5, 387–410.
- Jolicoeur, P., and Landau, M. J. (1984). Effects of orientation on the identification of simple visual patterns. *Can. J. Psychol.* 38, 80–93.
- Kung, E., and Hamm, J. P. (2010). A model of rotated mirror/normal letter discriminations. *Mem. Cognit.* 38, 206–220.
- Lawson, R., and Jolicoeur, P. (2003). Recognition thresholds for plane-rotated pictures of familiar objects. *Acta Psychol. (Amst.)* 112, 17–41.
- Mack, M. L., and Palmeri, T. J. (2011). The timing of visual object categorization. *Front. Psychol.* 2:165. doi: 10.3389/fpsyg.2011.00165
- Milivojevic, B., Corballis, M. C., and Hamm, J. P. (2008). Orientation sensitivity of the N1 evoked by letters and digits. *J. Vis.* 8(10), 1–14. doi: 10.1167/1168.1110.1111
- Milivojevic, B., Hamm, J. P., and Corballis, M. C. (2009a). Functional neuroanatomy of mental rotation. *J. Cogn. Neurosci.* 21, 945–959.
- Milivojevic, B., Hamm, J. P., and Corballis, M. C. (2009b). Hemispheric dominance for mental rotation: it is a matter of time. *Neuroreport* 20, 1507–1512.
- Milivojevic, B., Hamm, J. P., and Corballis, M. C. (2011). About turn: how object orientation affects categorisation and mental rotation. *Neuropsychologia* 49, 3758–3767.
- Milivojevic, B., Clapp, W. C., Johnson, B. W., and Corballis, M. C. (2003a). Turn that frown upside down: ERP effects of thatcherization of misoriented faces. *Psychophysiology* 40, 967–978.
- Milivojevic, B., Johnson, B. W., Hamm, J. P., and Corballis, M. C. (2003b). Non-identical neural mechanisms for two types of mental transformation: event-related potentials during mental rotation and mental paper folding. *Neuropsychologia* 41, 1345–1356.
- Muthukumaraswamy, S. D., Johnson, B. W., and Hamm, J. P. (2003). A high-density ERP comparison of mental rotation and mental size transformation. *Brain Cogn.* 52, 271–280.
- Perrett, D. I., Oram, M. W., and Ashbridge, E. (1998). Evidence accumulation in cell populations responsive to faces: an account of generalisation of recognition without mental transformations. *Cognition* 67, 111–145.
- Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., and Boyes-Braem, E. (1976). Basic objects in natural categories. *Cogn. Psychol.* 8, 382–439.
- Rossion, B., Gauthier, I., Tarr, M. J., Despland, P., Bruyer, R., Linotte, S., and Crommelinck, M. (2000). The N170 occipito-temporal component is delayed and enhanced to inverted faces but not to inverted objects: an electrophysiological account of face-specific processes in the human brain. *Neuroreport* 11, 69–74.
- Schendan, H. E., and Lucia, L. C. (2009). Visual object cognition precedes but also temporally overlaps mental rotation. *Brain Res.* 1294, 91–105.
- Schendan, H. E., and Lucia, L. C. (2010). Object-sensitive activity reflects earlier perceptual and later cognitive processing of visual objects between 95 and 500ms. *Brain Res.* 1329, 124–141.
- Thompson, P. (1980). Margaret Thatcher: a new illusion. *Perception* 9, 483–484.
- Yin, R. K. (1969). Looking at upside down faces. *J. Exp. Psychol.* 81, 141–145.

Received: 05 October 2011; accepted: 23 April 2012; published online: 11 May 2012.

Citation: Milivojevic B (2012) Object recognition can be viewpoint dependent or invariant – it's just a matter of time and task. *Front. Comput. Neurosci.* 6:27. doi: 10.3389/fncom.2012.00027

Copyright © 2012 Milivojevic. This is an open-access article distributed under the terms of the Creative Commons Attribution Non Commercial License, which permits non-commercial use, distribution, and reproduction in other forums, provided the original authors and source are credited.





# Renewing the respect for similarity

Shimon Edelman\* and Reza Shahbazi

Department of Psychology, Cornell University, Ithaca, NY, USA

**Edited by:**

Evgeniy Bart, Palo Alto Research Center, USA

**Reviewed by:**

Florentin Wörgötter, University Goettingen, Germany  
Evgeniy Bart, Palo Alto Research Center, USA

**\*Correspondence:**

Shimon Edelman, Department of Psychology, Cornell University, Ithaca, NY 14853-7601, USA.  
e-mail: se37@cornell.edu

In psychology, the concept of similarity has traditionally evoked a mixture of respect, stemming from its ubiquity and intuitive appeal, and concern, due to its dependence on the framing of the problem at hand and on its context. We argue for a renewed focus on similarity as an explanatory concept, by surveying established results and new developments in the theory and methods of similarity-preserving associative lookup and dimensionality reduction—critical components of many cognitive functions, as well as of intelligent data management in computer vision. We focus in particular on the growing family of algorithms that support associative memory by performing hashing that respects local similarity, and on the uses of similarity in representing structured objects and scenes. Insofar as these similarity-based ideas and methods are useful in cognitive modeling and in AI applications, they should be included in the core conceptual toolkit of computational neuroscience. In support of this stance, the present paper (1) offers a discussion of conceptual, mathematical, computational, and empirical aspects of similarity, as applied to the problems of visual object and scene representation, recognition, and interpretation, (2) mentions some key computational problems arising in attempts to put similarity to use, along with their possible solutions, (3) briefly states a previously developed similarity-based framework for visual object representation, the Chorus of Prototypes, along with the empirical support it enjoys, (4) presents new mathematical insights into the effectiveness of this framework, derived from its relationship to locality-sensitive hashing (LSH) and to concomitant statistics, (5) introduces a new model, the Chorus of Relational Descriptors (ChoRD), that extends this framework to scene representation and interpretation, (6) describes its implementation and testing, and finally (7) suggests possible directions in which the present research program can be extended in the future.

**Keywords:** object recognition, scene interpretation, scene space, shape space, similarity, view space, visual structure

## 1. THE UBIQUITY OF SIMILARITY

The effectiveness of an embodied cognitive system in fending for itself depends on its ability to gain insights into its situation that may not be immediately obvious, either because the properties of interest are not explicit in its sensory assessment of the outside world, or, more interestingly, because they are projections into a potential future. Species that share an ecological niche cannot entirely avoid the need for forethought, or reasoning about the future (Dewey, 1910; Craik, 1943; Dennett, 2003; Edelman, 2008; Bar, 2011). Indeed, evolutionary experiments in which a species seemingly drops out of the smarts race by opting for thicker armor or bigger teeth are merely bets that these bodily attributes will continue to be effective in the future. Such bets that are likely to go horribly wrong when a competitor invents the next brainy countermeasure to brawn.

Forethought works because the world is “well-behaved” in the sense that the future *resembles* the remembered past and can be often enough estimated from it, in relevant respects, and up to a point. In particular, similar consequences are likely to follow from similar observable causes—an observation that has influenced philosophical thought since Aristotle and that has been expressed forcefully by Hume (1748, ch. IX):

ALL our reasonings concerning matter of fact are founded on a species of Analogy, which leads us to expect from any cause the

same events, which we have observed to result from similar causes. Where the causes are entirely similar, the analogy is perfect, and the inference, drawn from it, is regarded as certain and conclusive. [...] But where the objects have not so exact a similarity, the analogy is less perfect, and the inference is less conclusive; though still it has some force, in proportion to the degree of similarity and resemblance.

While Hume’s observation applies to visual objects and scenes just as it does to all of cognition, bringing out similarity in vision and putting it to work requires some extra ingenuity on the part of any visual system, natural or artificial. In particular, to obtain information regarding the *shapes* of the objects that are present in the scene, the visual system must overcome the effects of the orientation of objects, of their juxtaposition, and of illumination. As it turns out that these computational challenges are subsumed under the general rubric of similarity-based processing, we shall begin by considering the most general issues first<sup>1</sup>.

<sup>1</sup>We discuss a similarity-based approach to dealing with the effects of orientation and juxtaposition of objects in scenes later in this paper. For related approaches to countering the effect of illumination, which rely on similarity to previously encountered exemplars, see for instance (Shashua, 1992; Sali and Ullman, 1998). Evidence that the human visual system relies on prior experience in its treatment of illumination in face recognition is offered by Moses et al. (1996).

The past several decades saw a concerted effort to put the explanatory role of similarity in psychology on a mathematical foundation. One well-known approach has employed set-theoretical tools (Tversky, 1977; Tversky and Gati, 1978); another one resulted in the development, from first principles, of a theory of similarity based on metric representation spaces (Shepard, 1980, 1984, 1987). In the present brief overview, we initially focus on the metric-space approach (although, as we shall see, the differences between the two turn out to be immaterial).

The basic premise of the metric theory of similarity posits that a perceiver encodes entities that are of interest to it, such as visual objects, scenes, or events, as points in a representation space in which perceived similarity between two items is monotonically related to their proximity. Shepard (1987) showed that a few fundamental assumptions, such as the Bayes theorem and the maximum entropy principle, lead to a representation space endowed with the Minkowski  $l_p$  metric (with  $p = 1$  if its dimensions are separable (Attneave, 1950; Garner and Felfoldy, 1970) and  $p = 2$  if they are not), and that the dependence of generalization from one item to another on their similarity—that is, on the representation-space distance—is negative exponential.

This dependence of generalization on representation-space distance had been found to hold for a range of taxa and tasks, from hue discrimination in goldfish to vowel categorization in humans. Shepard (1987) interpreted the ubiquity of this pattern as evidence for a universal law of generalization. This idea has been revisited in a special issue of the *Behavioral and Brain Sciences* (Shepard, 2001), where it has also been given a Bayesian formulation (Tenenbaum and Griffiths, 2001). Its empirical support has also been broadened. In a typical study, a confusion table for a set of stimuli is first formed by measuring same/different error rates for each pair of stimuli (this can be accomplished by various means; cf. Cutzu and Edelman, 1998). The table is then submitted to multidimensional scaling (MDS; Beals et al., 1968; Shepard, 1980), which yields a spatial configuration of the stimuli in a metric space of prescribed dimensionality (usually two or three) that best fits the confusion table data. Finally, the probability of generalization is plotted against distance in this “psychological space,” invariably resulting in a negative exponential dependence.

Chater and Vitányi (2003) have recently shown that this dependence of generalization on similarity must hold in principle even without the assumption that items are represented by points in a Minkowski metric space. Resorting instead to the notion of algorithmic information distance, defined as the length of the shortest program that transforms the representations of the two items that are being compared into one another, Chater and Vitányi derived the same negative exponential dependence as in Shepard’s formulation. They also noted that their “generalized law of generalization” holds even for “complex visual or linguistic material that seems unlikely to embed naturally into a multidimensional psychological space.”

Combined with the assumption that the world is well-behaved in the sense that similar situations occur often enough and have similar consequences, Shepard’s Universal Law of generalization suggests that cognitive processes that guide behavior all conform to the same functional template. A cognitive system faced with a

potentially novel situation needs (1) to determine where the new representation lands in the space of prior experience, (2) to look up records of the consequences of responses to similar situations, (3) to use those in thinking ahead to likely outcomes of possible responses, and (4) to generate an actual response while taking into account these data. Notably, this functional template applies all across cognition, from perception (as when conceptual knowledge is distilled from similar pieces of episodic information) to thinking (as in case-based reasoning) and action (where behavioral plans and motor programs are synthesized from whatever worked in the past).

In the remainder of this paper, we offer a series of discussions highlighting a series of conceptual, mathematical, computational, and empirical aspects of similarity, as applied to the problems of visual object and scene representation, recognition, and interpretation. Section 2 discusses certain issues with similarity and argues that these need not prevent it from being a useful explanatory concept in cognition. Sections 3 and 4 offer, respectively, a very brief introduction to a similarity-based framework for visual object representation, the Chorus of Prototypes, and an equally brief overview of the empirical support it enjoys (with multiple references to a detailed treatment elsewhere). In section 5, we present some new mathematical insights into the effectiveness of this framework, derived from its relationship to locality-sensitive hashing (LSH) and to concomitant statistics. Section 6 introduces a new model, the Chorus of Relational Descriptors (ChoRD), that extends this framework to scene representation and interpretation. An implementation and testing of the ChoRD model is described in section 7. Finally, section 8 offers some conclusions and suggests possible directions in which the present research program can be extended in the future.

## 2. THE PROBLEMATICITY OF SIMILARITY

Although first-principles considerations of the kind invoked by Shepard (1987), Tenenbaum and Griffiths (2001), and Chater and Vitányi (2003) clearly suggest that similarity should serve as an indispensable and broad foundation for cognition, its status as an explanatory concept in psychology and in neuroscience has been subject to much doubt (Goodman, 1972; Tversky, 1977; Tversky and Gati, 1978; Rips, 1989; Medin et al., 1993; Townsend and Thomas, 1993; Hahn and Chater, 1998). The prime reason for this is the ambiguity of similarity with regard to items that vary along independent or potentially conflicting dimensions.

Any two objects or situations that are not identical to each other are bound to be similar in some respects and dissimilar in others. As Eisler (1960, p. 77) put it, “An observer instructed to estimate the similarity of e.g., two differently colored weights, is supposed to ask: in what respect?” Because the *respects* in which objects are to be compared do generally depend on the task and on the mindset that the subject brings to it, similarity appears to be too ill-defined to have explanatory value for the psychologist or, indeed, practical value for the perceiver.

This conceptual difficulty is, however, not insurmountable. Rather than seeking an ironclad, universally valid set of similarity relations that are prior to any experience, cognitive systems use their experience in interacting with the world to learn the respects in which various situations should be considered as similar, by

tracking the *consequences* of their actions. The similarity question thus turns out to be an instance of the well-known computational problem of credit assignment (Minsky, 1961). Here, it takes the form of the need to differentiate between those features (dimensions) of similarity of two items that are, in the context of the task, predictive of the consequences of generalizing between them, and those that are not<sup>2</sup>.

In general, the credit assignment problem has both temporal (diachronic) and structural aspects. The former has to do with apportioning credit to each of a potentially long sequence of actions, and the latter—to the various dimensions of the situation/action representation. With regard to similarity-based processing, it is the dimensionality of the representation space that is of prime concern. The three related computational problems discussed below all arise from the typically *high dimensionality* of measurement and representation spaces.

The need for high-dimensional representation spaces in cognition stems in turn from the foundational role of experience in the planning of future behavior. To increase the chances that at least some of the stored data would bring out the similarity patterns on which generalization can be based, an advanced cognitive system must measure up as many episodes of its interaction with the world as possible, while making each measurement as detailed as possible. It is no wonder, then, that the amount of information that the brains of long-lived animals in complex ecosystems must capture, process, and store is vast (Merker, 2004). To understand how the brains of such animals, including ourselves, manage this deluge of data, we must first identify the computational principles that are in the play.

## 2.1. THE TUG OF WAR BETWEEN CONTENT-BASED RETRIEVAL AND GENERALIZATION

Seeing that storage as such appears to be cheap (e.g., Brady et al., 2008), the main problem here is retrieval. In other words, if a vast amount of data is stored against a possible future need, the efficiency of retrieval becomes all the more important. Clearly, retrieval must be selective: only those records that are similar to the present experience must be brought to the fore. Moreover, retrieval must be fast: a sequential scan of the full contents of the multitude of stored items will not do. A computational scheme that fulfills these requirements is *hashing* (Aho et al., 1974). By storing each item under a key that is computed from its content and that uniquely specifies a memory address, hashing allows fast associative recall: a test item can be looked up in constant time, independent of the number of stored items. In that respect, hashing is like a massively parallel, content-addressable biological memory system, in which a cue can be compared simultaneously to multiple stored items (see Willshaw et al., 1969 for an early computational model and Lamdan and Wolfson, 1988 for an early application in a computer vision system for object recognition).

To minimize recall mistakes stemming from memory collisions, hashing functions in data management applications were traditionally engineered to map any two items, even similar ones,

to very different addresses. This way, the probability of confusing distinct items could be kept low—but only at the expense of destroying any similarity relationships that may hold over the items. Because under a classical hashing scheme two similar and therefore possibly related cues may wind up very far apart in the representation space, simply “looking around” the address of the best-matching item for anything that may be worth retrieving along with it would not work. Thus, while enabling content-based retrieval, classical hashing hinders similarity-based generalization.

## 2.2. THE CHALLENGE OF DIMENSIONALITY REDUCTION

Earlier in this section we noted that the measurement space in which objects external to the system are first represented is likely to be high-dimensional. Indeed, in the human visual system, the nominal dimensionality of the input signal from each eye is equal to the number of axons that comprise the optic nerve, or about  $10^6$ . Any perceivable similarities over visual objects or scenes must, therefore, exist as patterns in that multidimensional signal<sup>3</sup>. The task of finding such patterns is, however, extremely hard.

What kind of measurement-space pattern could be useful for similarity-based generalization? Two generic types of patterns are those that afford categorization and those that support regression (Edelman and Intrator, 2002; Bishop, 2006). In the first case, a number of previously encountered exemplars fall into a small number of distinct categories according to some characteristics, making it possible to categorize a new item by its similarity to each of those. In the second case, exemplars cluster in a subspace of dimensionality that is lower than that of the original measurement space. In each of the two cases, subsequent generalization becomes possible because the description of the data in terms of the patterns is simpler than the original representation (as per the Minimum Description Length (MDL) principle; cf. Adriaans and Vitányi, 2007).

The problem is that the characteristics that define the “small number” of clusters or the “lower-dimensional” subspace in the above formulation need not correspond to any of the original measurement dimensions by themselves. The similarity of two spatially sampled visual objects, for instance, is always distributed over a multitude of pixels (that is, dimensions) rather than being confined to a single pixel. The visual system must find the right function of pixel values (e.g., a rotation of the original space followed by a projection onto a subspace, if the function is constrained to be linear) under which the sought-after similarity pattern—in the two-category case, a bimodal distribution—is made explicit (in the sense of Marr, 1982).

The linear version of the problem of finding such a function is known as projection pursuit (Huber, 1985). By the central limit theorem, most low-dimensional projections of a high-dimensional “cloud” of points will be approximately normal,

<sup>2</sup>Cf. Shepard's (1987) notion of consequential regions, and the need for differential valuation of stimulus dimensions implied by the Ugly Duckling Theorem (Watanabe, 1969, pp. 376–377).

<sup>3</sup>This observation applies to natural or analog similarities, not symbolic or conventional ones. Thus, a heap of 19 marbles is naturally similar to a heap of 20 marbles under any of a wide range of visual measurement schemes, whereas under most schemes the number 19 on this page is only conventionally similar to the number 20. A natural similarity space for shapes is discussed in (Edelman, 1999, 3.2–3.3).

that is, they will look like noise. Consequently, an “interesting” projection is one that yields a distribution that deviates from normality, e.g., because it is bimodal, or perhaps heavy-tailed (Intrator and Cooper, 1992). Algorithms based on this approach can be extremely effective in cases where the pattern of interest is indeed linear (e.g., two linearly separable clusters of data points side by side). They are, however, of no avail in the general case, where no linear projection can do the job (e.g., if the pattern consists of two concentric spherical shells of data points).

### 2.3. THE COMPLEXITY OF LEARNING FROM EXAMPLES

A complementary problem to the separation of a pattern into a few clusters or a subspace of a few dimensions is that of pattern build-up. How many data points suffice to define a pattern that can support reliable generalization? This question is of central concern in machine learning (along with the related issue of the number of degrees of freedom of the learning mechanism; e.g., Haussler, 1992). Intuitively, learning from examples can be seen as an instance of function approximation (Poggio, 1990), which suggests that the set of examples must cover the domain of the sought-after function in a representative manner<sup>4</sup>.

The need to cover the representation space with examples implies that the number of required data points depends exponentially on the number of dimensions of the representation space—a problem known as the curse of dimensionality (Bellman, 1961). While it can be circumvented in supervised learning on a task-by-task basis<sup>5</sup>, the problem of dimensionality in an exploratory (unsupervised) setting or in a situation where transfer of performance is expected between tasks (Intrator and Edelman, 1996) must be addressed by undertaking dimensionality reduction prior to learning.

### 2.4. THE TRUTH IS OUT THERE

The last computational consideration that we would like to bring to bear on the problem of learning and use of similarity is that perceptual similarity (as opposed to arbitrary associations that the cognitive system may form following experience) is “out there” in the world, waiting to be transduced into the measurement space and preserved and discovered in the reduced-dimensionality representation. In the domain of visual object shapes, for instance, natural similarity relations arise from the mathematics of shape parametrization, where certain uniqueness results have been proved (see Edelman, 1999, App.C for references). As noted in the introduction, these relations are in principle discoverable by agents situated in the world, insofar as similar causes tend to lead to similar consequences.

This observation suggests that perceptual representations should be evaluated on the basis of their *veridicality*—the degree to which they preserve the qualities of the objects “out there.” In particular, a veridical representation scheme that preserves

relational qualities such as similarity amounts to what Shepard (1987, 2001; cf. Shepard and Chipman, 1970) termed a second-order isomorphism between the representations and their targets (this must be distinguished from first-order isomorphism, which posits representations that individually resemble their respective objects and which, it should be noted, merely postpones the problem of making sense of the world rather than solving it; Edelman, 1999)<sup>6</sup>.

We may therefore conclude that the twofold computational challenge that any perceptual system must address is (1) to achieve veridical representation of similarities among objects, so as to forge a link between sensory data and consequentially responsible behavior, and (2) to do so in a low-dimensional representation space, so as to allow effective pattern discovery and learning from experience. The rest of this article offers a brief overview of a comprehensive computational theory that explains how the primate system for visual object recognition solves these two problems. This theory has been implemented and tested both as a computer vision system and as a model of biological vision and is backed by behavioral and neurobiological findings, as detailed in the references.

## 3. A SIMILARITY-BASED FRAMEWORK FOR VISUAL OBJECT PROCESSING: THE CHORUS OF PROTOTYPES

In problems that arise in visual object processing (see Table 1), the nature of the stimulus universe and certain generic properties of visual systems ensure that veridical representation of distal object similarities in a low-dimensional space is easy to achieve (for a detailed argument, based on properties of smooth mappings, see Edelman, 1999). In this section, we outline a computational framework that offers a solution to these problems, which is based on the idea of putting similarity itself to work in constructing a representation space for distal objects. Because it represents each

**Table 1 | A hierarchy of tasks arising in visual object and scene processing.**

Task	What needs to be done	What it takes
Recognition	Dealing with novel views of shapes	Tolerance to extraneous factors (pose, illumination, etc.)
Categorization	Dealing with novel instances of known categories	Tolerance to within-category differences
Open-ended representation	Dealing with shapes that differ from familiar categories	Representing a novel shape without necessarily categorizing it
Structural analysis	Reasoning about (i) the arrangement of parts in an object; (ii) the arrangement of objects in a scene	Explicit coding of parts and relationships of objects and scenes

<sup>4</sup>Note that this formulation is related to the more general view of the problem of learning from examples as the estimation of the joint probability density over input and output variables.

<sup>5</sup>The support vector approach to supervised learning can solve classification and regression tasks directly in a high-dimensional space; see Cortes and Vapnik (1995) for an early formulation and Malisiewicz et al. (2011) for a recent application.

<sup>6</sup>Despite its intuitive appeal and deep roots that go back to Plato, the first-order isomorphism approach is also infeasible in practice (given the computational difficulties associated with the task of reconstructing the world from sensory data) and is a poor model of human performance (given that subjects are in fact very bad at such reconstruction).



stimulus by a vector of its similarities to a small set of reference objects, this framework is called the “Chorus of Prototypes” (Edelman, 1995, 1999).

The Chorus framework is founded on the observation that, no matter how high-dimensional the measurement space of a visual system is, certain events and relationships of interest “out there” in the world give rise to representational signatures whose structure ensures tractability. One behaviorally important type of such event is the rotation of a rigid object in front of the observer around a fixed axis (or, equivalently, the circumambulation of the object by the observer). Provided that the imaging function that maps the object’s geometry into the representation space is smooth, the footprint of the rotation event in the representation space will be a one-dimensional manifold—a smooth curve (which, moreover, will loop back upon itself, due to the cyclic nature of the rotation event)<sup>7</sup>. For rotation around three mutually orthogonal axes, the manifold will be three-dimensional<sup>8</sup>.

### 3.1. OBJECT VIEW SPACES

Because the representation of the set of views of a rotating object—its *view space*—has the manifold property, the views can be related to one another by computationally tractable procedures. In particular, given that the view space is smooth, a small number of exemplars (representation-space points that encode particular views of the object) typically suffice to interpolate it, using any of the many existing methods for function approximation. One such method, which, as we shall see in the next section, is especially interesting from the neurobiological standpoint, is approximation by a linear superposition of radial basis functions (Poggio and Edelman, 1990; Poggio and Girosi, 1990).

This corresponds to representing any view of the object by its similarities to a handful of exemplar views that can be learned from experience (Poggio and Edelman, 1990; this, in turn, implies that the view space for the object, as well as a decision function for object identity, can take the form of a weighted sum of the outputs of a set of neurons each of which is broadly tuned to one of the exemplar views). While recognition performance of this mechanism can be highly tolerant to viewpoint changes (if the exemplars are chosen so as to jointly cover the view space well), it is not fully viewpoint-invariant—but neither is the performance of human subjects (Bülthoff and Edelman, 1992; Edelman and Bülthoff, 1992; Edelman, 1999; DiCarlo and Cox, 2007; more about this in section 4).

### 3.2. OBJECT SHAPE SPACES

Edelman (1995) noted that the principles that facilitate this kind of low-dimensional representation of relationships between different views of the same object apply also to the relationships between different object shapes. Specifically, object shapes that are not too dissimilar from each other—say, a duck, a goose, and a chicken—can be meaningfully morphed into one another by simple linear interpolation of some fiducial features such as

edge configurations, so that intermediate shapes do make sense. Indeed, they form a smooth, low-dimensional manifold.

This implies that under a smooth representation mapping, the set of view spaces of the objects in such a “tight” shape category—its collective *shape space*—can be interpolated by the same means that support the interpolation of individual view spaces (Edelman, 1998). Moreover, because the view spaces of the shapes in question will be roughly parallel to each other, learning a view-related task for one shape would readily transfer to another (Intrator and Edelman, 1996, 1997; Edelman and Duvdevani-Bar, 1997). For instance, learning to predict the appearance of a three quarters view of one face from its frontal view would work also for other faces (Lando and Edelman, 1995; Duvdevani-Bar et al., 1998).

With regards to implementation, the shape space can be approximated by the same means as the view space, as a weighted sum of tuned unit responses, which serve as basis functions. If each of the units is tuned to an entire view space of some object (which may itself appear at a range of orientations), together they will span the shape space for the family of objects in question. Given a potentially novel stimulus, each such tuned unit effectively signals how distant (that is, dissimilar) it is from its preferred shape, or “prototype.” The joint ensemble activity (which inspired the name *Chorus of Prototypes*; Edelman, 1995) pinpoints the location of the stimulus in shape space, just as in a land survey the distances to a handful of landmarks jointly fix the location of a test point in the terrain.

### 3.3. THE CHORUS TRANSFORM

Formally, representing a new view by its similarities to familiar views or a new shape by its similarities to familiar shapes are both instances of an application of the Chorus Transform (Edelman, 1999). Let  $\mathbf{p}_1, \dots, \mathbf{p}_n$  be  $n$  prototypes and let  $\mathbf{x}$  be an input vector,  $\mathbf{p}_k, \mathbf{x} \in \mathbb{R}^d$ . The Chorus Transform (CT) is defined as follows:

$$CT(\mathbf{x}) = \frac{1}{\sqrt{n}} \begin{pmatrix} \|\mathbf{x} - \mathbf{p}_1\| \\ \vdots \\ \|\mathbf{x} - \mathbf{p}_n\| \end{pmatrix} \quad (1)$$

The application of this transform  $CT: \mathbb{R}^d \rightarrow \mathbb{R}^n$  results in dimensionality reduction, if the number of prototypical objects,  $n$ , is smaller than the dimensionality of the measurement space  $d$ .

Edelman (1999, App.B) showed that the Chorus Transform can support a logarithmic dimensionality reduction, while approximately preserving the inter-point distances in the original space (the proof of this claim is based on a theorem due to Bourgain, 1985). In other words, even with a very small number of prototypes— $O(\log d)$ , where  $d$  is the dimensionality of the original space—the relative positions of the data points in the new, low-dimensional space approximate their original layout, implying that the original similarity relations, and with them category boundaries, etc., are largely preserved<sup>9</sup>.

<sup>7</sup>For definitions of formal concepts such as smoothness and manifolds, and for other mathematical details, see Edelman (1999).

<sup>8</sup>If the object is opaque, the manifold will be piecewise smooth.

<sup>9</sup>Recent developments in neighborhood-preserving embedding and immersion (Bartal et al., 2011) improve on the Johnson and Lindenstrauss (1984) result that had been cited by Edelman (1999). The original J-L lemma states

A statistically robust version of CT can be derived by observing that a representation based on distances to a set of points (prototypes) is related to vector quantization (Linde et al., 1980; the following exposition is borrowed from Edelman, 1999, App.B). A vector quantizer  $Q$  is a mapping from a  $d$ -dimensional Euclidean space,  $\mathcal{S}$ , into a finite set  $\mathcal{C}$  of *code vectors*,  $Q: \mathcal{S} \rightarrow \mathcal{C}$ ,  $\mathcal{C} = \{p_1, p_2, \dots, p_n\}$ ,  $p_i \in \mathcal{S}$ ,  $i = 1, 2, \dots, n$ . Every  $n$ -point vector quantizer partitions  $\mathcal{S}$  into  $n$  regions,  $R_i = \{x \in \mathcal{S} : Q(x) = p_i\}$ ; the Voronoi diagram is an example of such a partition. Whereas vector quantization encodes each input pattern in terms of *one* of the code vectors chosen by the nearest-neighbor principle (Cover and Hart, 1967), Chorus does so in terms of similarities to several prototypes. This parallel suggests that a discretized representation of the input space, related to the Voronoi diagram, can be obtained by considering ranks of distances to prototypes, instead of the distances themselves.

Let  $\mathbf{p}_1, \dots, \mathbf{p}_n$  be  $n$  prototypes, and consider a representation that associates with each input stimulus the *Rank Order* of its *Distances* to the prototypes (*ROD*). That is, an input  $\mathbf{x}$  is represented by an ordered list of indices  $ROD(\mathbf{x}) = (i_1, i_2, \dots, i_n)$ , meaning that among all prototypes  $\mathbf{p}_i$ ,  $\mathbf{x}$  is the most similar to  $\mathbf{p}_{i_1}$ , then to  $\mathbf{p}_{i_2}$ , and so on. Note that the index  $i$  always heads the list  $ROD(\mathbf{p}_i)$  corresponding to the prototype  $\mathbf{p}_i$  (a prototype is most similar to itself). The total number of distinct representations under the *ROD* scheme is  $n!$  (the number of permutations of the  $n$  indices). To compare two representations, one may use Spearman rank order correlation of the index lists.

#### 4. EXPERIMENTAL SUPPORT FOR THE CHORUS FRAMEWORK

The Chorus framework has been implemented and evaluated as a computer vision system for recognition and categorization of isolated objects (Duvdevani-Bar and Edelman, 1999) and for class-based generalization (Lando and Edelman, 1995; Edelman and Duvdevani-Bar, 1997). It had also generated predictions for behavioral, electrophysiological, and imaging experiments, all of which were subsequently corroborated. The relevant studies, which are mentioned briefly in this section, have been discussed at great length elsewhere (Edelman, 1998, 1999).

The basic tenet of the Chorus model—that object vision is fundamentally viewpoint-dependent because its functional building block is a unit broadly tuned to a specific view of a specific object—received early support from psychophysical (Bülthoff and Edelman, 1992; Edelman and Bülthoff, 1992) and neurophysiological (Logothetis et al., 1994; Logothetis and Pauls, 1995; Wachsmuth et al., 1994; Perrett and Oram, 1998) experiments. Subsequent studies consolidated the notion that object recognition is characterized not by invariance but by tolerance to extraneous factors such as orientation and retinal position, which,

that any  $n$ -point subset of Euclidean space can be embedded in  $O(\epsilon^{-2} \log n)$  dimensions with at most  $(1 + \epsilon)$  distortion of the inter-point distances. In contrast, the new *local dimension reduction* lemma (Bartal et al., 2011) offers a likewise bounded-distortion embedding into a space whose dimensionality does not depend on  $n$ , as long as it is the local and not the global structure of the data set that is to be preserved. It remains to be seen whether this embedding method can be carried out by mechanisms whose biological implementation is as straightforward as that of the Chorus scheme.

furthermore, depends on the task and on the prior experience with the objects in question (Dill and Edelman, 2001; DiCarlo and Maunsell, 2003; Cox et al., 2005; Rust and DiCarlo, 2010).

A particularly interesting feature of the Chorus framework is that object representations that it posits are *generically veridical* with regard to inter-object similarities. As noted above, the dimensionality reduction method employed by the Chorus model—representing each stimulus by its distances to shape-space landmarks—is guaranteed to approximately preserve original similarities among stimulus shapes, insofar as it implements the random subspace projection method of near-isomorphic embedding (Johnson and Lindenstrauss, 1984; Bourgain, 1985). The predicted metrically veridical perception of object similarities has indeed been demonstrated in behavioral and physiological studies with humans (Cutzu and Edelman, 1996, 1998; Edelman et al., 1998, 1999; Giese et al., 2008; Panis et al., 2008) and monkeys (Sugihara et al., 1998; Op de Beeck et al., 2001).

In summary, results from human and monkey psychophysics and physiology suggest, as predicted by the Chorus framework, (1) that the visual system seeks tolerance rather than invariance to object transformations (Rust and DiCarlo, 2010), as predicted by the view- and shape-space idea (Edelman et al., 1998; DiCarlo and Cox, 2007), (2) that object translation can be disruptive, especially for structure representation (Dill and Edelman, 2001; Cox et al., 2005; Kravitz et al., 2008), as predicted by the retinotopy of the classical receptive fields that are the functional building blocks of the Chorus model, (3) that this trait is compatible with extrastriate neural response properties (Vogels, 1999; Gallant et al., 2000; DiCarlo and Maunsell, 2003), and (4) that the peculiarities in the manner in which primate vision deals with object structure (Tsunoda et al., 2001; Newell et al., 2005; van Dam and Hommel, 2010) can be accounted for by a fragment-based scheme that relies on binding by retinotopy (Edelman and Intrator (2003)).

#### 5. A RENEWED INTEREST IN THE MATHEMATICS OF SIMILARITY AND THE CHORUS TRANSFORM

The past decade saw a variety of new and exciting developments in the theory of similarity-preserving associative recall, which are proving to be widely useful in computer vision, notably LSH (Andoni and Indyk, 2008). Furthermore, some old ideas for embedding structured data in vector spaces, such as holographic reduced representations (Plate, 1991), are being rediscovered and applied (Jones and Mewhort, 2007), albeit not in the visual domain. We see both these sets of development as important to visual scene representation and processing: the former contribute to the struggle against the curse of dimensionality, while the latter suggest computationally convenient and neurally plausible ways of dealing with structure. In this section and in section 6, we briefly describe representative methods from these two domains and show that they are either related to the Chorus Transform or can benefit from its application.

##### 5.1. THE CHORUS TRANSFORM IMPLEMENTS LOCALITY-SENSITIVE HASHING (LSH)

Significant progress in similarity-based high-dimensional data management has been recently brought about by the development of new algorithms that perform hashing while respecting local

similarity (Andoni and Indyk, 2008; Paulevé et al., 2010). The growing family of LSH algorithms “effectively enables the reduction of the approximate nearest neighbor problem for worst-case data to the exact nearest neighbor problem over random (or pseudorandom) point configuration in low-dimensional spaces” (Andoni and Indyk, 2008). Both steps in this process—forming the random projections and quantizing the resulting low-dimensional space into address bins—rely on the same computational principles that underlies the Chorus Transform and can be carried out by the same mechanism, namely, a set of tuned units.

As outlined in **Figure 1**, the process begins by choosing a number of hash functions from a family of functions  $\mathcal{H} = \{h : \mathbb{R}^d \rightarrow U\}$  that satisfies the LSH condition: the probability  $P_1$  of mapping two data points  $\mathbf{p}, \mathbf{q} \in \mathbb{R}^d$  to the same bin must be larger than the probability  $P_2$  of mapping them to different bins if the points are close together —

$$\text{if } \|\mathbf{p} - \mathbf{q}\| \leq R \text{ then } \Pr_{\mathcal{H}}[h(\mathbf{p}) = h(\mathbf{q})] \geq P_1 \quad (2)$$

$$\text{if } \|\mathbf{p} - \mathbf{q}\| \geq cR \text{ then } \Pr_{\mathcal{H}}[h(\mathbf{p}) = h(\mathbf{q})] \leq P_2 \quad (3)$$

where  $R$  is the radius of the neighborhood that defines proximity and  $c > 1$  is a constant (which defines an “exclusion zone” around the  $R$ -neighborhood). Each of the hash functions is then used to construct a hash table, which are populated by points from the given data-set. The lookup procedure for a query point  $\mathbf{q}$  iterates over the hash tables and returns retrieved points that fall within an  $R$ -neighborhood of  $\mathbf{q}$ .

Now, consider the “multidimensional line partitioning” LSH family described by Andoni and Indyk (2008, p. 121). A hash function from this family first performs a random projection of

the data point  $\mathbf{p}$  into  $\mathbb{R}^t$ , where  $t$  is super-constant [i.e., grows slowly with  $n$ , as in  $t = o(\log n)$ ]. The space  $\mathbb{R}^t$  is then partitioned into cells, and the hash function is made to return the index of the cell that contains the projected point  $\mathbf{p}$ .

This last part suggests a ready parallel to the Chorus Transform. Specifically, the receptive fields of the tuned units representing the prototypes effectively function as the cells in the second step of the above procedure (the first step being the projection of the probe point on the manifold defined implicitly by the choice of prototypes). To complete the analogy, the outputs of the tuned units can be thresholded (as in the *ROD* version of the transform), so that the resulting code consists of the identities (that is, indices) of units whose activation by the probe point exceeds the threshold.

The original Chorus Transform, without thresholding, can be seen to carry out *kernelized* LSH (a variant introduced by Kulis and Grauman (2009), which, as those authors note, is applicable to both vector and non-vector data). In a recent development of this approach, He et al. (2010, p.1133) defined the space  $V_j$  onto which the data are projected by the  $j^{\text{th}}$  hashing function by a linear combination of “landmarks”  $\{\mathbf{z}_n\}$  in the kernel space. This idea leads to the hash function.

$$h(\mathbf{p}) = \text{sign}(\mathbf{a}^T \mathbf{k}_p - \mathbf{b}) \quad (4)$$

where  $\mathbf{a}$  are the linear combination weights and

$$\mathbf{k}_x = [K(\mathbf{x}, \mathbf{z}_1), \dots, K(\mathbf{x}, \mathbf{z}_n)]^T \quad (5)$$

are the kernel values between  $\mathbf{x}$  and each of the landmark points  $\mathbf{z}_n$ . With the distance function  $\|\cdot\|$  serving as the kernel and  $\mathbf{z}_n$

#### Preprocessing:

1. Choose  $L$  functions  $g_j$ ,  $j = 1, \dots, L$ , by setting  $g_j = (h_{1,j}, h_{2,j}, \dots, h_{k,j})$ , where the  $h$  functions are chosen at random from an LSH family  $\mathcal{H}$ .
2. Construct  $L$  hash tables containing the dataset points hashed by the functions  $g_j$ .

#### Query algorithm for a test point $\mathbf{q}$ :

1. For each  $j = 1, 2, \dots, L$  do
  - (a) Retrieve the points from the bucket  $g_j(\mathbf{q})$  in the  $j^{\text{th}}$  hash table.
  - (b) For each retrieved point, compute the distance to  $\mathbf{q}$  and report the point if it is correct (i.e., an  $R$ -near neighbor of  $\mathbf{q}$ ).
  - (c) Stop as soon as the number of reported points reaches a preset threshold.

**FIGURE 1 | The locality-sensitive hashing (LSH) scheme (after Andoni and Indyk, 2008, Figure 2).** For an explanation of how the Chorus Transform implements LSH, see section 5.1.

as the prototypes, this corresponds precisely to an application of the Chorus Transform to the data point  $\mathbf{x}$ .

## 5.2. THE CHORUS TRANSFORM COMPUTES CONCOMITANT STATISTICS

In their discussion of LSH families, Andoni and Indyk (2008, p. 120) note that if the Jaccard similarity, defined for two sets  $A$  and  $B$  as  $s(A, B) = |A \cap B| / |A \cup B|$ , is used as a basis for hashing, the LSH framework is thereby extended to include the so-called *minwise hashing* methods. Minwise hashing (Broder, 1997; Li and König, 2011) is a special case of pairwise characterization of ordered sets through their concomitant statistics (Eshghi and Rajaram, 2008, Section 4), and is best explained as such.

Consider  $n$  independent sample pairs,  $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$  obtained from a bivariate distribution  $f(x, y)$ . In the theory of rank order statistics,  $y_k$  is called the *concomitant* of  $x_k$ . Formally, concomitant theory captures the relation between the order statistics of  $x$  and  $y$  in the form of a rank distribution given by  $\Pr[\text{Rank}(y_i) = j \mid \text{Rank}(x_i) = k]$ .

Let  $\prod_{1,1}^n$  be the probability that the smallest of  $x_i$  is the concomitant of the smallest of  $y_i$ . The link to the LSH theory now becomes apparent: if the smallest element among  $x_i$  is identical to that of  $y_i$ , it must lie in the intersection of the two sets, which implies that the probability  $\prod_{1,1}^n$  is equal to the Jaccard similarity between them (this is the defining insight behind minwise hashing, due to Broder, 1997).

Eshghi and Rajaram (2008) observe that the same reasoning holds not just for the smallest (lowest-ranking) pair but also for any range of smallest concomitant ranking pairs of the two sets. They proceed to define a “min  $k$ -multi-hash” LSH family based on this observation. For us, it is of interest because the smallest  $k$  values in a Chorus Transform—a representation that supports LSH—are effectively computed by retaining the smallest  $k$  out of the  $n$  distances to the prototypes that define it<sup>10</sup>.

In a related vein, Yagnik et al. (2011) introduce the Winner Take All (WTA) hash, “a sparse embedding method that transforms the input feature space into binary codes such that Hamming distance in the resulting space closely correlates with rank similarity measures.” Their hash functions define the similarity between two points by the degree to which their feature dimension rankings agree. Yagnik et al. (2011) point out that the simplest of such measures is the pairwise order function  $PO(x, y) = \sum_i \sum_{j < i} T((x_i - x_j)(y_i - y_j))$ , where  $x_i$  and  $y_i$  are the  $i^{\text{th}}$  dimension values of  $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$  and  $T$  is a threshold function,  $T(x) = 1$  if  $x > 0$  and  $T(x) = 0$  otherwise.

Whereas Yagnik et al. (2011) proceed to define their WTA hash family using random permutations of feature dimensions, it can also be formulated in terms of the Chorus Transform. To that end, in lieu of permuting the dimensions, all we have to do is administer a vector of random biases (drawn from a

predetermined set of random vectors) to the landmark units; each such bias vector effectively permutes the rank order of the unit responses. Given that under the Chorus Transform, the output representation by distances to prototypes preserves the rank order of data point similarities in the original space (Edelman, 1999, App.B), the above procedure is exactly equivalent to the one proposed by Yagnik et al. (2011), with the added advantage of being carried out in a more convenient low-dimensional space.

## 6. EXTENDING THE CHORUS FRAMEWORK TO COVER STRUCTURAL SIMILARITY

The kinds of visual stimuli discussed up to now in this paper did not include objects composed of parts or scenes containing multiple objects, such as those depicted in **Figure 2**, or that which you will see if you raise your eyes from this paragraph and look around you. In this section we first list some of the functional requirements posed by structured scenes and the challenges presented by those requirements. We then briefly mention a previously published biologically motivated model of scene processing (Edelman and Intrator, 2003). Finally, we outline a new computational approach to scene interpretation, the Chorus of Relational Descriptors (ChoRD), which uses CT on all the representational levels: for representing shapes, their relationships, and entire scenes.

### 6.1. FUNCTIONAL REQUIREMENTS AND CHALLENGES IN COMPOSITE SCENE INTERPRETATION: SYSTEMATICITY AND STRUCTURAL ALIGNMENT

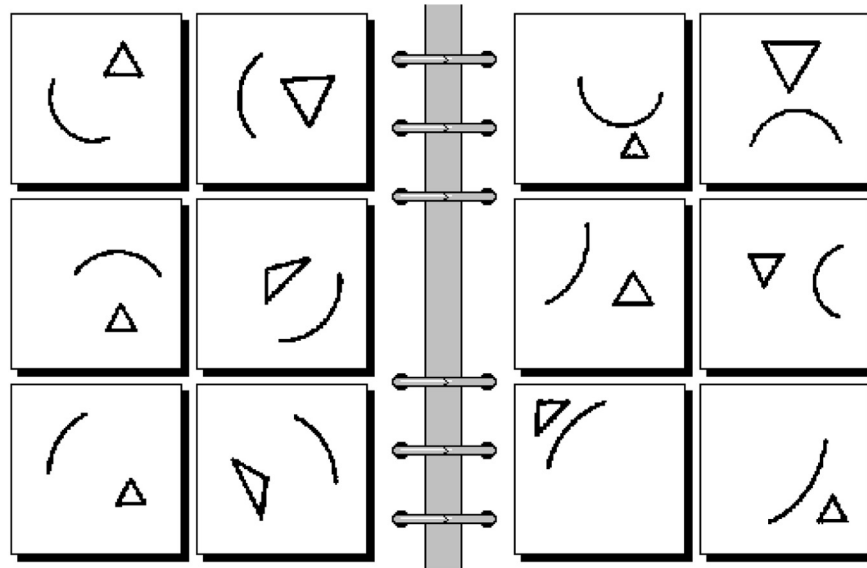
Operational parsimony, which in animal vision translates into evolutionary pressure, dictates that a visual system should represent a structured scene hierarchically, in terms of intermediate-size parts and their spatial relations, if such a representation is warranted for the family of scenes at hand by the MDL principle (Rissanen, 1987; Adriaans and Vitányi, 2007). Ideally, therefore, the representation of scene structure would be fully compositional in the classical sense of Frege (1891)<sup>11</sup>.

A compositional representation would allow the visual system to be *systematic* in its interpretation of parts and relations—a desideratum that is traditionally invoked in support of compositional models based on MDL (Bienenstock et al., 1997). Formally, an agent employing symbolic representations is systematic if its ability to entertain the proposition  $R(a, b)$  implies a concomitant ability to entertain the proposition  $R(b, a)$ . In vision, this would mean that a system that can make sense of a scene in which a man rides a donkey should also be able to make sense of a scene in which a donkey rides a man (Edelman and Intrator, 2003, **Figure 1**). In practice, however, human cognition is often far from systematic in its dealing with structure, and so is unlikely to rely on fully compositional representations (see Johnson, 2004 for informal arguments and Edelman and Intrator, 2003 for empirical evidence).

<sup>10</sup>These are the  $k$  landmarks that are the closest to the probe data point; cf. the discussion of the relationship between CT and vector quantization in section 3.3. We also note that this idea is related to the coding scheme of Thorpe et al. (1996) and the MAX model of Rousselet et al. (2003).

<sup>11</sup>For a thorough introduction to the principle of compositionality, see (Szabó, 2008); for a discussion in the context of vision, see (Edelman and Intrator, 2003).





**FIGURE 2 | Problem #75 of the 100-long sequence of challenges to pattern recognition posed by Bongard (1970).** The task is to determine what distinguishes the scenes on the left from the scenes on the right. To answer this question, it is not enough to list the shapes that appear in

the scenes: their spatial attitudes and relations must be made explicit too. This representational requirement is often referred to as (a spatial counterpart to) *structural systematicity* (Edelman and Intrator, 2003). See text for discussion.

If a modicum of systematicity is to be preserved, a certain amount of spatial analysis must be carried out (Edelman and Intrator, 2003), so as to enable *structural alignment* (Markman and Gentner, 1993)—a procedure in which parts and relations found in one scene are matched to parts and relations found in the other<sup>12</sup>. Consider, for instance, the two scenes at the top of **Figure 3**. Disparate as these scenes are, certain parallels can be drawn between some fragments of one and fragments of the other. In particular, the vertical ridge at the center of the sandstone depression in the scene on the left resembles the narrow vertical lean-to attached to the wall of the building depicted in the scene on the right. Furthermore, each of the two circular windows on both sides of this vertical feature can be matched, respectively, to two rounded (but not very circular) holes in the scene on the left. In each of the two scenes, the spatial arrangement of the matched fragments forms a stylized face (two eyes and a nose between them)—a realization that in turn suggests structural similarity to the spatial composition of the head of the owl in the scene on the bottom left and, stretching the imagination a bit, to the Chinese character on the bottom right of **Figure 3**.

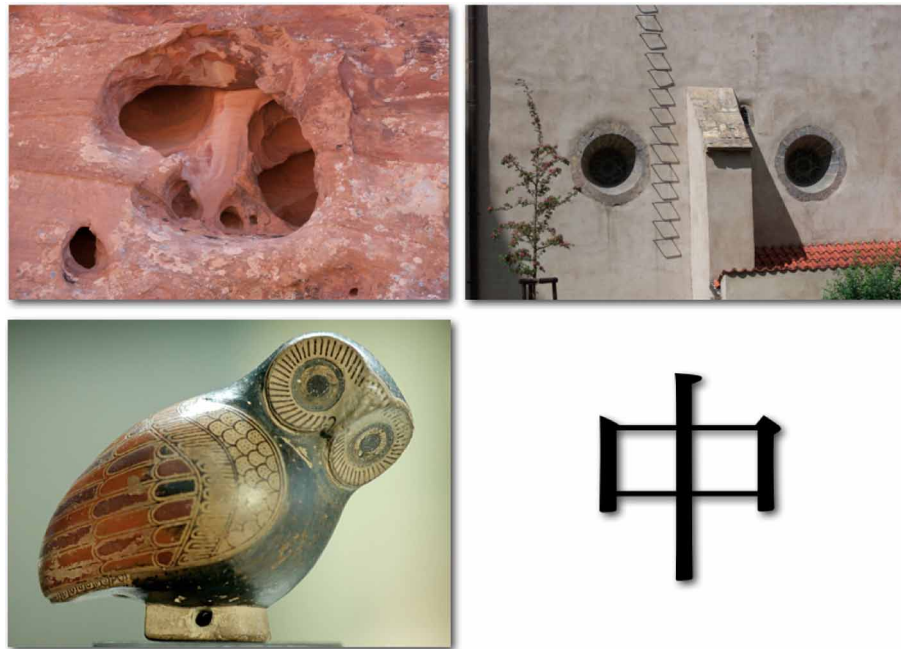
Structural alignment thus turns the question of scene interpretation (and with it also the question of scene similarity) into a nested set of questions about similarities of scene parts and their relations. The four scenes resemble each other (up to a point) *because* each one consists of individually alignable fragments (the

“eyes” and the “nose”) that, moreover, form the same spatial pattern on a larger scale. Given a proper interpretation of each of those scenes, we can answer questions such as “what shape appears to the left of the vertical feature?”, “what feature appears between the rounded ones?” or “what is the structural counterpart of *this* vertical feature in the other scene?”

What kind of representation can meet these functional needs without running afoul of constraints imposed by neural implementation? Let us suppose for the moment that the representations of structured objects or scenes are themselves made to possess an analogous symbolic structure. Following this logic, the representation of a scene composed of two shapes, one above the other, could take the form of an ordered pair of the two feature vectors corresponding to the two constituent shapes. This approach, however, creates a dilemma. On the one hand, it relies on abstract relational binding (which is how the ordered pairing of constituents is implemented in symbolic models; see, e.g., Hummel and Holyoak, 1998; Hummel, 2001). Although such an implementation, being fully compositional, would result in ideal systematicity, it is not, we believe, entirely biologically or behaviorally plausible, as noted above<sup>13</sup>. On the other hand, eschewing symbolic binding in favor of a more biologically relevant approach, such as representing composite scenes by bags of features each of which carries both shape and location information

<sup>12</sup>Structural alignment differs from shape alignment for recognition, introduced by Huttenlocher and Ullman (1987) and Ullman (1989), in that it operates on the objects’ parts (which, further, could be defined in terms of their function rather than shape) and relations, instead of on the global shapes of the objects.

<sup>13</sup>Concerns about biological plausibility arise also with regard to the otherwise fascinating idea of representing structured objects in the same metric space as simple ones, as in the Holographic Reduced Representations of Plate (1991) and other approaches based on similar mathematical principles (e.g., Jones and Mewhort, 2007; Sahlgren et al., 2008; Basile et al., 2011).



**FIGURE 3 | Four scenes for which possibilities for structural alignment can be profitably explored.** Image sources: *top left*, a pattern in weathered sandstone, Lower Muley Twist Canyon, Capitol Reef National Park, Utah; *top right*, the eastern wall

of the Old Synagogue, Jewish Quarter, Prague; *bottom left*, a proto-Corinthian figurine of an owl, ca. 640 B.C. (from the antiquities collection at the Louvre); *bottom right*, the Chinese character for “middle” (*zhōng*).

(cf. the “what + where” features of Rao et al., 1997; see also Op de Beeck and Vogels, 2000) has problems of its own in supporting structural alignment, insofar as scene constituents are not easy to address selectively in such a representation.

## 6.2. AN EARLY APPROACH: THE CHORUS OF FRAGMENTS

Edelman and Intrator (2000; 2003) attempted to avoid both horns of the above dilemma by developing the Chorus of Prototypes into a non-compositional model of structure representation that exhibits appropriately limited systematicity. Instead of positing generic parts and abstract relations, their *Chorus of Fragments* model relied on the scene layout and on binding by retinotopy to represent structure and on multiple location-bound shape spaces to represent its constituents. The resulting model exhibited a degree of systematicity, in that it interpreted correctly spatial rearrangements of shapes familiar to it through training (namely, digit shapes). It also showed productivity, in that it performed nearly equally well for novel shapes, which had had no “what” units dedicated to them (letter shapes).

The model, described in detail by Edelman and Intrator (2003), consisted of “what + where” units, which by definition respond selectively in a graded manner both to stimulus shape and to its location (Rao et al., 1997; Op de Beeck and Vogels, 2000). During learning, it relied on multiple fixations to train the functional equivalent of a shape-tuned (“what”) unit parameterized by location (“where”). This functionality, which can be thought of as gain modulation through covert attention shifts (Connor et al., 1997; Salinas and Abbott, 1997; Salinas and Thier, 2000), offers a solution of sorts to the problem of constituent

addressing, which, as we just mentioned, arises in structural alignment. During testing, a single fixation of the composite stimulus by the model sufficed for interpreting it—that is, for making explicit, through the pattern of the units’ responses, of what shape was present at what location in the stimulus.

## 6.3. A NEW IDEA: CHORUS OF RELATIONAL DESCRIPTORS (ChoRD)

While the CoF model did the right thing in predicating a full representation of a scene on multiple fixations of its constituents, it implemented the “what + where” functionality using a black-box learning mechanism (a bottleneck autoencoder; DeMers and Cottrell, 1993) that performed the task while leaving its inner workings opaque. In this section, we describe a new approach to implementing limited systematicity and thereby supporting various structure-related tasks, which is characterized by two main features. First, similar to the CoF model, it is constrained by the architectural and functional considerations that call for distributed, graded, low-dimensional representations. Second, it improves on the CoF model by dealing explicitly with the many related versions of the same scene arising from multiple fixations, and by doing so through recourse to the same computational mechanism that is at the core of CT: representation by similarities to multiple prototypes. Because of that, the new approach has also the advantage of being related to the similarity-preserving hashing methods that are being currently used in computer vision (as we pointed out in preceding sections).

The new approach, Chorus of Relational Descriptors, or ChoRD, represents a given scene by multiple entries in an

associative memory. The memory system is implemented by a hash table of the LSH type, in which (1) each of the possibly many entries for a given scene uses one of the scene's regions of interest (ROIs) as the key, and (2) key values falling within a certain range of similarity to a given ROI are all mapped to the same record. The record associated with a key ROI is the scene minus that ROI; it is represented by a list of the remaining ROIs along with the spatial displacement of each of them relative to the key ROI.

To give a concrete example, consider a scene consisting of an object, **A**, which appears *above* another object, **B** (in general, of course, a scene can consist of more than two objects). Representations of this scene will be stored in the hash table under two keys,  $ROI(A)$  and  $ROI(B)$ —and so will scenes that contain objects sufficiently similar to **A** and **B**. In particular, the representation stored under  $ROI(A)$  will consist of the list  $\{ROI(B), \text{dir}(A, B)\}$ , where the last element encodes the direction from **A** to **B**.

The ChoRD model that we just outlined uses *CT* on two levels. First, and most fundamentally, both the ROIs comprising the scene and their relative spatial displacements with regard to each other are represented by vectors of distances to select sets of shape and layout prototypes, respectively. Second, given that an LSH-based representation is itself equivalent to *CT* (as we showed in section 5.1), the entire scene is de facto represented in a distributed, redundant, graded fashion by the ensemble of records associated with its constituent ROIs, in a manner that neither discards the spatial structure of the scene, nor attempts to capture it categorically, as the symbolic models aim to do.

## 7. TESTING A SIMPLE IMPLEMENTATION OF ChoRD

We now describe a series of tests of the ChoRD model, carried out in the simple domain of scenes composed of two ROIs each (a detailed examination of the model's performance and its scaling to more complex scenes will be reported elsewhere; Shahbazi and Edelman, in preparation). Each scene was constructed by

embedding two object images, drawn from six most populous object categories in the LabelMe database (Russell et al., 2008), in a black background. The objects were converted to grayscale and scaled to a size of  $50 \times 50$  pixels; the entire scene was  $150 \times 150$  pixels (see **Figure 7** for some scene examples). While this type of test image will probably fail to impress computer vision practitioners, it has the advantage of allowing a very tight control over the scene parameters, which is why such scenes are at present widely used in behavioral and imaging studies (e.g., Newell et al., 2005; Hayworth et al., 2011; MacEvoy and Epstein, 2011; Zhang et al., 2011), some of whose results we replicate below.

### 7.1. ENCODING THE ROIs AND THEIR LAYOUT

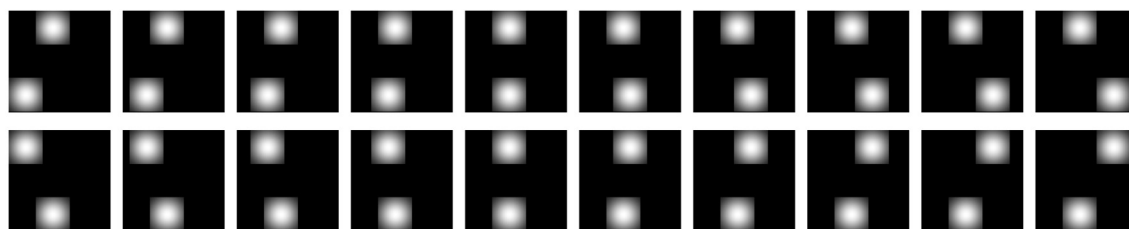
Regions of interest (ROIs) were detected in the scene by sliding a Gaussian patch along the image and locating the ROI at the place that resulted in a maximum sum of the pixel values of the convolved image. The size of the Gaussian patch was made to match the size of the objects. Ten objects were chosen at random from the list of LabelMe objects to serve as the prototypes for *CT* (see **Figure 4**). Each of those was represented by a list of outputs of Gabor filters at two different scales, 5 and 10 pixels, and two orientations,  $0^\circ$  and  $90^\circ$ <sup>14</sup>. Every detected ROI patch was represented by the list of filter values, then encoded by the 10-prototype *CT*.

To encode the spatial structure or layout of the scene, we represented it by similarities to a set of 10 layout prototypes. Fixation-dependent encoding was simulated by using one such set of 10 layouts for cases in which the top ROI was fixated and another one for cases in which the bottom ROI was fixated (see **Figure 5**). Each layout prototype consisted of two Gaussian

<sup>14</sup>The original implementation of *CT*-based object recognition (Duvdevani-Bar and Edelman, 1999) used an even simpler ROI representation with great effect. In a modern computer vision setting, a SIFT-based representation (Lowe, 1999) would be used.



**FIGURE 4 | The 10 shape prototypes used in conjunction with *CT* to encode the ROIs comprising the scenes (see section 7.1).** Each ROI detected in a scene was represented by a 10-dimensional vector of its respective similarities to these 10 images.



**FIGURE 5 | The layout prototypes used in conjunction with *CT* to encode the spatial structure of scenes (see section 7.1).** There are two different sets of such prototypes. One set of 10 prototypes is used for encoding the scene when the top ROI is fixated; the other set of 10

prototypes is used when the bottom ROI is fixated. For each situation (scene + fixation), the scene structure was thus represented by a 10-dimensional vector of similarities between the layout of the scene's ROIs and the 10 layout prototypes.



image patches. The image location of one of these, corresponding to the would-be scene placement of the reference or key ROI for the given fixation, was fixed, and the location of the other differed systematically among the 10 prototypes, spanning collectively a range of displacements as illustrated in **Figure 5**. The entire scene's layout was therefore encoded relative to the fixation point (the location of the key ROI) by listing its image-based similarities to the 10 displacement prototypes.

The entire procedure whereby the representation of a scene was computed is illustrated in **Figure 6**. Altogether, the complete representation of a scene for a given fixation ("entry" or key) point consisted of the concatenation of (1) a 10-dimensional representation of the fixation ROI, (2) a 10-dimensional representation of the other ROI, and (3) a 10-dimensional representation of the spatial layout relative to fixation. Scene representations constructed in this manner were entered into an LSH table, implemented using Shakhnarovich's Matlab code with ten 64-bit hash tables (Shakhnarovich, 2008).

The LSH functionality (which, as we showed in section 6, is equivalent to that of *CT*) subsequently allowed content-based lookup—a key ingredient in testing the resulting ChoRD model on additional scenes, which could be familiar or novel in some respects. In the experiments described in the remainder of this section, we tested the ability of the ChoRD model to support certain systematicity-related queries and to replicate several behavioral and imaging studies involving human subjects.

Following training (that is, populating the LSH with scene representations), each familiar scene is represented redundantly,

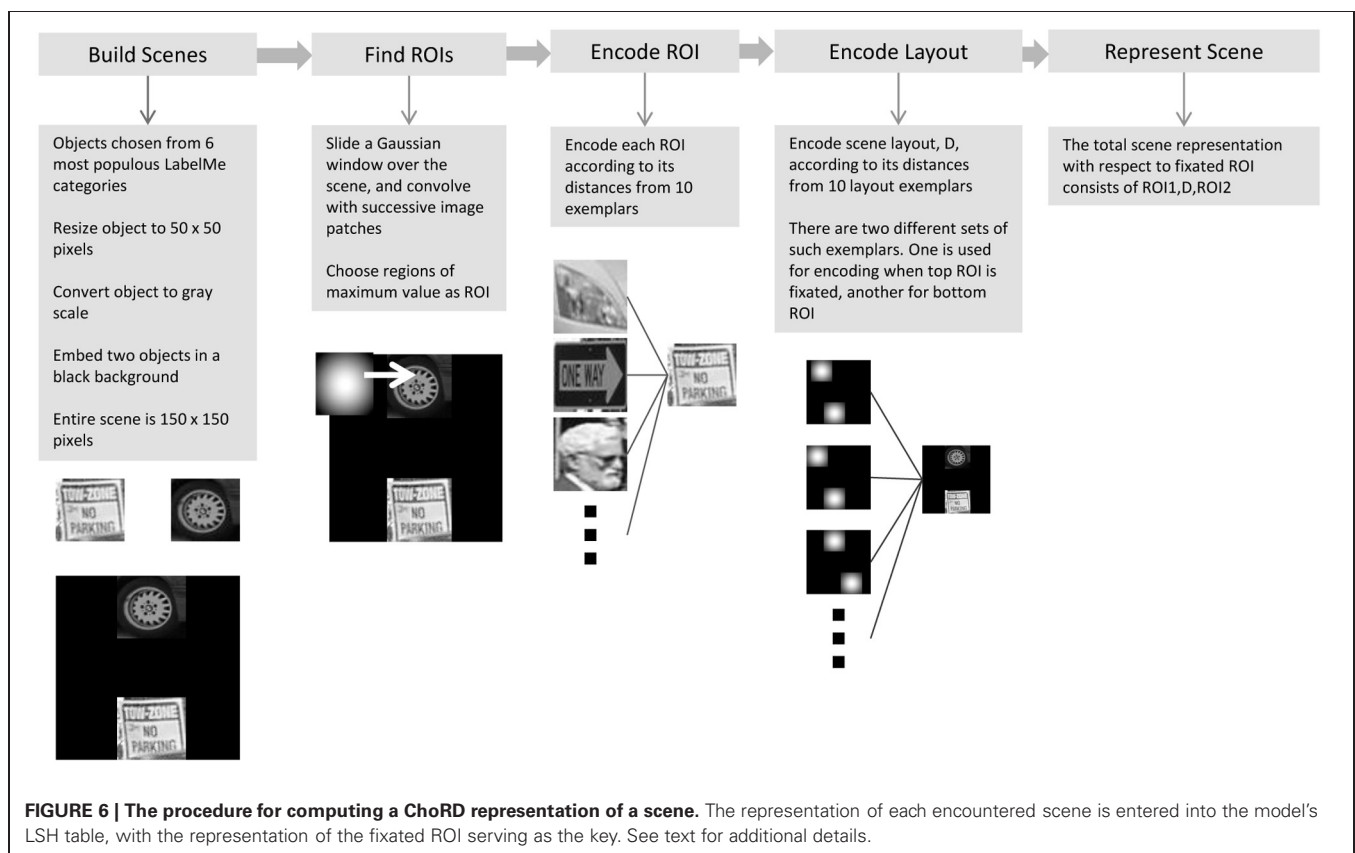
by as many records as it has ROIs. Given a test scene, the model's LSH table returns all the representations that match the ROIs contained in it. Importantly, because of the locality-sensitive property of the hashing scheme that we used, a novel scene—that is, a scene that differs somewhat from the familiar examples either in its ROIs or in their locations, or both—results in the retrieval of familiar scenes that are sufficiently similar to it. Thus, we expected the model's performance to degrade gracefully when tested on progressively more novel stimuli, rather than crash.

## 7.2. EXPERIMENT 1: PRODUCTIVITY

Our first experiment tested the model's productivity: its ability to deal with moderate novelty as just defined. Each of the test stimuli in this experiment had one novel and one familiar object in a familiar configuration, two novel objects in a familiar configuration, or two familiar objects in a novel configuration. The dissimilarity between the test scene and the representation retrieved in response to it was defined as

$$\Delta_k = \|ROI_{11} - ROI_{12}\| + \|D_{11} - D_{12}\| + \|ROI_{21} - ROI_{22}\| \quad (6)$$

where  $ROI_{ij}$  is the  $i^{th}$  ROI of scene  $j$ , and  $D_{ij}$  is the layout representation of scene  $j$  relative to  $ROI_{ij}$ . Identical computations were performed by fixating each of the two objects in



the test scene, yielding  $\Delta_1$  and  $\Delta_2$ , which were then averaged together to form the composite dissimilarity between the two scenes.

We remark that the form of Eq. 6 glosses over the conceptual difficulty inherent in trying to deal simultaneously with multiple shape and location differences. This difficulty is universal in that it arises in any attempt to compare composite entities (say, estimating the similarity of two sets of fruit containing one apple and one orange each), including certain structural alignment tasks (section 6.1). In psychology, this corresponds to the classical problem of scaling (Shepard, 1987), which is beyond the scope of the present discussion. Thankfully, in the present context of *testing* a given model (rather than defining the representation that serves as its foundation), this difficulty amounts merely to a matter of preference that may or may not be given to some components of the composite dissimilarity, depending on the task. This can be done simply by weighting those components as needed. Our choice in Equation 6 corresponds to using equal weights for all.

The experiment was performed on 6000 test scenes in three different conditions: condition N, 2000 test scenes with one novel object; condition NN, 2000 test scenes with two novel objects; and condition L with 2000 test scenes with two familiar objects in a new spatial layout. For each condition, the test scene was encoded according to both possible fixations, and the query was performed for both encodings. For each query, the five nearest neighbors were retrieved and their (dis)similarity to the test scene was computed. The reported results are for the best match obtained (i.e., the most similar scene retrieved from the hash table). **Figure 7** shows examples of test scenes (on the left) and their corresponding five most similar scenes retrieved from the table.

To investigate the contribution of *CT* to the model's performance, we carried out another experiment, this time using the raw filter-based encoding of the scenes. **Figure 8** shows side by side the results for the raw and *CT*-encoded scenes. Note that there is no significant difference in the similarity of the test and retrieved scenes for different conditions in the non-*CT* version.

### 7.3. EXPERIMENT 2: SENSITIVITY TO GRADUAL CHANGE

In the second experiment, we measured the similarity of two scenes represented by the ChoRD model, in one of which the two objects were progressively displaced relative to each other (see **Figure 9**). Newell et al. (2005) found that the performance of human subjects in this situation indicated their reliance on representations that yielded graded similarity, rather than breaking down categorically as the layout of the manipulated scene changed. To simulate their study, we generated a series of test scenes with the same two objects. By keeping one object's position constant and displacing the other one, the relative positions of the objects were changed, either horizontally or vertically, in increments of 10 pixels. **Figure 10** shows the resulting dissimilarities between reference and test scenes. The experiment was performed on 2000 different scenes, with five levels of displacement tested for each scene, and resulted in a gradual increase of dissimilarity with displacement. A linear regression fit the results well:  $R^2 = 0.72$ ,  $F_{(9998)} = 2.06 \times 10^4$  ( $p < 2.2 \times 10^{-16}$ ).

### 7.4. EXPERIMENT 3: SENSITIVITY TO DIFFERENT TYPES OF QUALITATIVE CHANGE

Our third experiment examined the ChoRD model's representation of relative similarities of scenes that were subjected to certain structural transformations. It has been patterned on the imaging study of Hayworth et al. (2011), who showed that for human subjects the BOLD response of brain areas implicated in scene representation is more sensitive to some structural transformations than to others. In particular, for scenes composed of two objects, switching the two objects around resulted in a larger release of adaptation, compared to simply translating both objects within the scene while keeping their relative positions unchanged.

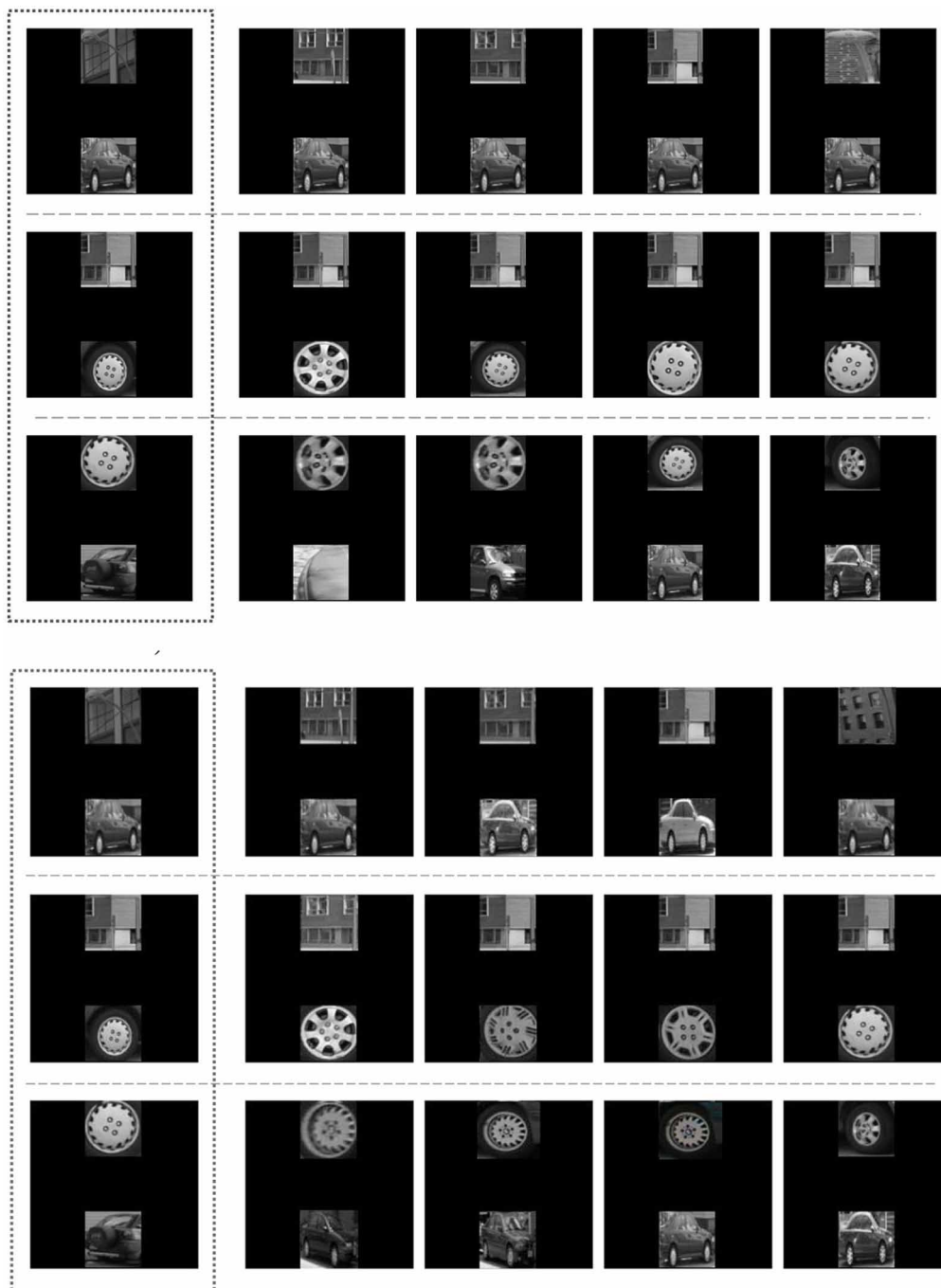
To replicate this finding, we constructed test scenes related to reference ones in three ways: through a joint translation of both objects (condition T), or reversal of the objects' locations (condition R), or both (condition TR). Two thousand scenes were generated for each of these conditions. The results, plotted in **Figure 11**, conform to those of Hayworth et al. (2011).

### 7.5. THE ChoRD MODEL: A DISCUSSION

We have tested the ChoRD model on simple scenes composed of two objects, in three experiments. In the first experiment, the model exhibited a degree of productivity, that is, an ability to deal, systematically, with scenes that differed in various ways from those to which it had been exposed during "training" (cf. Edelman and Intrator, 2003). In the second experiment, we found that the model's estimate of similarity between a reference scene and a series of test scenes differing from it progressively was it self graded—a finding that echoed that of Newell et al. (2005) in a similar setup. In the third experiment, we used the model to replicate one of the findings of an fMRI adaptation study (Hayworth et al., 2011), which found differential effects on brain activation of two types of scene transformation: joint translation vs. switching around of the scene's constituents. All these results were obtained by a model that used *CT* on every relevant representational level to reduce dimensionality and enact tolerance to moderate novelty, supporting our assertion of the importance of similarity-based representations in scene processing.

In addition to being rooted in our own earlier work on similarity-based object and scene representation (Edelman, 1999; Edelman et al., 2002; Edelman and Intrator, 2003), the ChoRD model can be seen as related to several contemporary lines of thinking in computer vision, as mentioned very briefly below (a detailed comparison will be offered in Shahbazi and Edelman, in preparation). In particular, the location-specific *CT*-based representations used here resemble the locality-constrained linear coding of Wang et al. (2010). The relationship between *CT* and vector quantization (VQ), from which Wang et al. (2010) derive their approach, has been noted and analyzed in (Edelman, 1999; cf. section 3.3). Continuing this parallel, the graded manner in which *CT* codes the similarities between the target object and prototype shapes may be compared to the variant of VQ that uses soft assignment (van Gemert et al., 2010).

Whereas many computer vision methods for image representation and retrieval rely on the bag of (visual) words idea (which goes back to the first histogram-based approaches developed two decades ago), there is an increasing number of attempts to extend

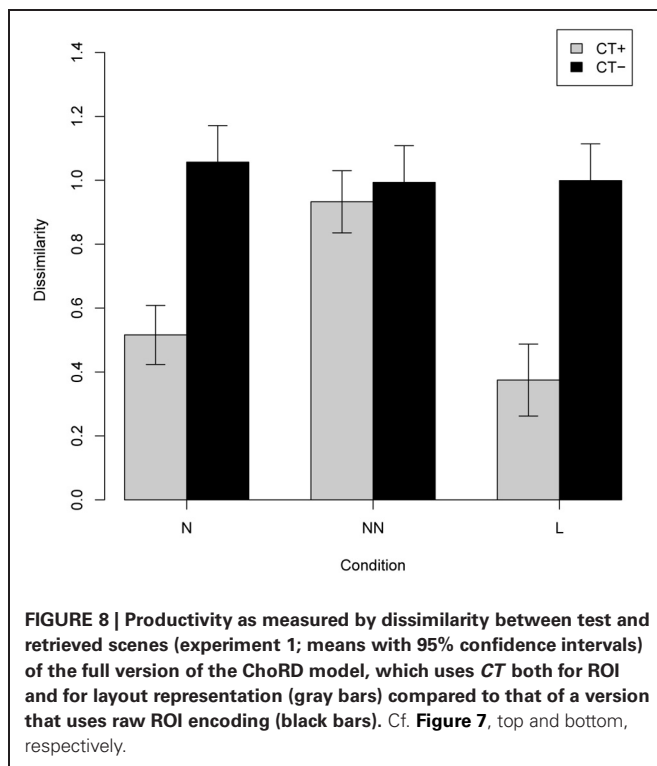


**FIGURE 7 | Experiment 1, testing productivity.** See section 7.2 for a description of the procedure and **Figure 8** for quantitative results. *Above:* the performance of the ChoRD model, which uses *CT* to represent ROIs. The leftmost column shows test scenes; the other columns show the best five matches retrieved from the model's LSH table, in the decreasing order of

similarity to the test scene. **Top row:** One novel object at position  $ROI_1$ . **Middle row:** One novel object at position  $ROI_2$ . **Bottom row:** Two novel objects. *Below:* the performance of a version of the model that uses raw ROI encoding rather than one based on *CT* (the layout was still encoded with *CT*).

this simple and powerful principle to capture some of the scene structure (and not just the mere presence in it of certain objects). One step in this direction is expressed by the “context challenge” of Torralba (2003), which led to the development of such

successful systems for context-based recognition as that of Divvala et al. (2009). Our model can be seen to engage with this challenge by coding scenes relative to certain “entry points” or key objects, for which the rest of the scene then constitutes a context



(of course, it still needs to be tested in an actual context-based recognition task).

We single out the work of Zhang et al. (2011) on image retrieval using geometry-preserving visual phrases (GVP) as the closest to ChoRD among the present computer vision approaches. Rather than trying to make scene structure matter by subjecting a set of images, preselected on the basis of bag of visual words similarity, to a spatial voting test (RANSAC; Fischler and Bolles, 1981), Zhang et al. (2011) incorporate information about relative spatial locations of the features forming a visual phrase into its representation (hence “geometry-preserving”). Compared to GVP, the ChoRD model appears to be more flexible and open-ended, insofar as it relies on CT in representing both the features and their layout.

Insofar as the ChoRD model represents a scene by a set of records keyed to its constituents and stored in an LSH table, it can be said to treat a scene merely as a big object. Imaging evidence

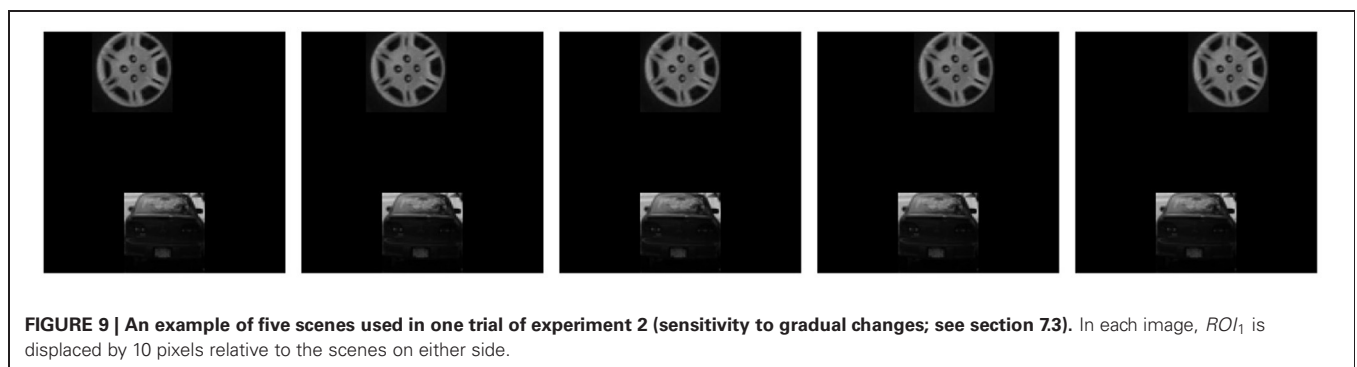
for this kind of scene representation in the lateral occipital complex in the human brain has been reported recently by MacEvoy and Epstein (2011), who write that “patterns of activity evoked in LO by scenes are well predicted by linear combinations of the patterns evoked by their constituent objects.” Notably, there was no evidence of such summation in the parahippocampal place area (PPA), implicated by previous studies in the representation of scene structure (Epstein and Kanwisher, 1998; Bar, 2004). In comparison, in the ChoRD model, the spatial structure of the scene is not lost in summation, as it would be under a bag of features approach. This pattern of results suggests to us the following tentative double analogy: (1) between the (distributed, CT-based) ChoRD representation of constituent shape and the LO complex, and (2) between the (also CT-based) ChoRD representation of scene layout and the PPA.

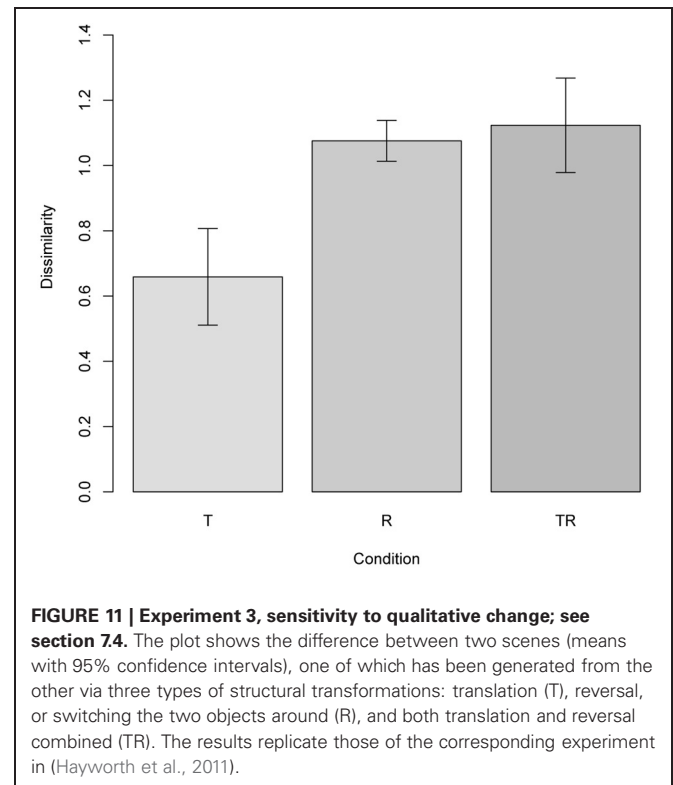
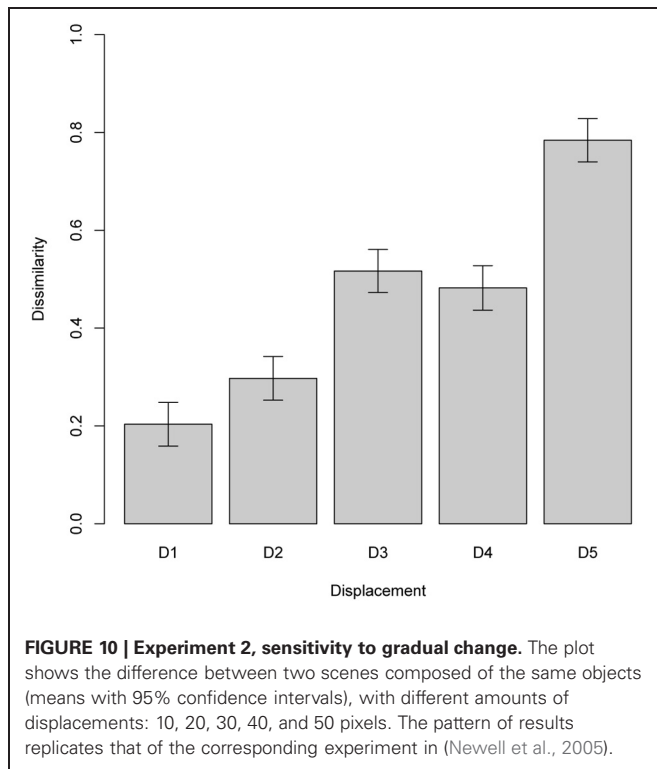
## 8. CONCLUSIONS

In the first part of this paper, we surveyed the role of similarity in theories and models of object recognition and described some newly discovered computational parallels between the Chorus Transform, or CT (an idea that received a book-length treatment in Edelman, 1999) and the widely popular computer vision methods of similarity-preserving hashing and dimensionality reduction. In the second part, we described the outcome of some (rather preliminary) tests of the ChoRD model, which extends CT so as to support a joint representation of scene content and layout. In this concluding section, we outline some of the directions in which the similarity project can be extended.

Taken together, our findings suggest that similarity to prototypes may constitute a viable general approach to representing structured objects and scenes. In particular, the same CT-based method can be used to span view spaces of individual shapes and shape spaces of object categories (Edelman, 1999), as well as “scene spaces” defined by objects and their spatial relations (the present work). From the computational standpoint, this is an exciting development, given that scene-related work in computer vision tended until recently to focus on scene categorization rather than interpretation (Oliva and Torralba, 2001; Lazebnik et al., 2006; Loeff and Farhadi, 2008).

The approach proposed here can support scene interpretation (over and above categorization), insofar as a list of objects, contexts, and relations to which a given scene is similar constitutes a rather complete representation of its content and structure (just





like in a text local adjacency relations within character n-grams jointly enforce global structure of phrases; cf. Wickelgren, 1969; Mel and Fiser, 2000). In computer vision, similar ideas underlie the work on “visual phrases” (Sadeghi and Farhadi, 2011; Zhang et al., 2011) and Conditional Random Fields (Kulkarni et al., 2011, **Figure 3**). To ensure flexibility, this representation should be parameterized by task, so that the similarity patterns revealed by it could focus on shape similarity (say) in some cases and on spatial relation similarity in others; a related idea has been proposed by Edelman and Intrator (2003, **Figures 6 and 7**).

We believe that further development of the similarity-based representational framework outlined in this paper should focus on the following three issues.

**Neural implementation.** Edelman and Intrator (2003) discussed the biological plausibility of their similarity-based scheme that coded scene fragments and their spatial relations (which they called the Chorus of Fragments). Indeed, this approach seems quite amenable to a neural implementation: a set of laterally interacting receptive fields, each tuned to an object category and embedded in a retinotopic map, would seem to do the job. More thought needs, however, to be given to the implementation of tuning. In particular, units that employ radial basis functions are not good at rejecting false positives. This calls for alternatives such as Exemplar-SVM (Malisiewicz et al., 2011), which may, perhaps, be amenable to implementation by augmenting RBF units with massive inhibition (Wang et al., 2000).

**Scalability.** Much progress has been achieved in computer vision by methods that utilize huge databases of images (e.g., Malisiewicz and Efros, 2009). Given the close relationship between the Chorus framework and similarity-tolerant hashing,

which we detailed in section 5, those methods may be on a convergence course with our approach. This may in turn result in a biologically inspired emulation of the vast human memory for visual objects and scenes (e.g., Brady et al., 2008).

**A probabilistic turn.** The Chorus framework is deterministic in its operation, its only stochastic aspect being the choice of prototypes during learning; it is also purely feedforward. While such models may be adequate for categorization tasks (Serre et al., 2008), they do not allow for the kind of flexibility that is afforded by the generative Bayesian approach (Tenenbaum and Griffiths, 2001; Chater et al., 2006). It is often the case, however, that successful models of learning and inference can be recast in Bayesian terms with very little modification (Edelman and Shahbazi, 2011). Developing the Chorus framework into a hierarchical generative model<sup>15</sup> is, therefore, a worthwhile future pursuit, which may take as its starting points the use of maximum-entropy reasoning and the Bayes theorem by Shepard (1987) and the generative theory of similarity proposed by Kemp et al. (2005).

In summary, we remark that the idea that similarity could play a key explanatory role in vision (as well as in other cognitive sciences) has experienced ups and downs in the centuries since its introduction by Hume. The Chorus project has previously shown that coding objects by their similarities to select prototypes can support a veridical representation of distal similarities

<sup>15</sup>The importance of hierarchy in this context is underscored by the recent finding that human observers learn to interpret hierarchically structured scenes more readily than others (Shahbazi et al., 2011).



among objects “out there” in the world, and to do so in a low-dimensional space that affords effective learning from experience. The ChoRD approach to representing structure enables the extension of the Chorus framework to composite objects and scenes. Moreover, the deep parallels between the Chorus idea and

similarity-preserving hashing techniques indicate that the resulting methods could be made to scale up to deal with massive amounts of visual data. These developments suggest that vision researchers would do well to renew their respect for similarity and assign it a key role in their conceptual toolkit.

## REFERENCES

- Adriaans, P., and Vitányi, P. M. B. (2007). “The power and perils of MDL,” in *Proceedings of the IEEE International Symposium on Information Theory (ISIT)* (Nice, France), 2216–2220.
- Aho, A. V., Hopcroft, J. E., and Ullman, J. D. (1974). *The Design and Analysis of Computer Algorithms*. Reading, MA: Addison-Wesley.
- Andoni, A., and Indyk, P. (2008). Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. *Commun. ACM* 51, 117–122.
- Attneave, F. (1950). Dimensions of similarity. *Am. J. Psychol.* 63, 516–556.
- Bar, M. (2004). Visual objects in context. *Nat. Rev. Neurosci.* 5, 617–629.
- Bar, M. (ed.). (2011). *Prediction in the Brain*. New York, NY: Oxford University Press.
- Bartal, Y., Recht, B., and Schulman, L. J. (2011). “Dimensionality reduction: beyond the Johnson-Lindenstrauss bound,” in *Proceedings of the 22nd SODA (ACM-SIAM Symposium on Discrete Algorithms)*, 86.
- Basile, P., Caputo, A., and Semeraro, G. (2011). “Encoding syntactic dependencies by vector permutation,” in *Proceedings of the GEMS 2011 Workshop on Geometrical Models of Natural Language Semantics, EMNLP 2011*, eds S. Padó and Y. Peirsman (Edinburgh, Scotland), 4351.
- Beals, R., Krantz, D. H., and Tversky, A. (1968). The foundations of multidimensional scaling. *Psychol. Rev.* 75, 127–142.
- Bellman, R. E. (1961). *Adaptive Control Processes*. Princeton, NJ: Princeton University Press.
- Bienenstock, E., Geman, S., and Poter, D. (1997). “Compositionality, MDL priors, and object recognition,” in *Neural Information Processing Systems*, Vol. 9, eds M. C. Mozer, M. I. Jordan, and T. Petsche (Cambridge, MA: MIT Press), 838–844.
- Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Berlin: Springer.
- Bongard, M. M. (1970). *Pattern Recognition*. Rochelle Park, NJ: Spartan Books.
- Bourgain, J. (1985). On Lipschitz embedding of finite metric spaces in Hilbert space. *Israel J. Math.* 52, 46–52.
- Brady, T. F., Konkle, T., Alvarez, G. A., and Oliva, A. (2008). Visual long-term memory has a massive storage capacity for object details. *Proc. Natl. Acad. Sci. U.S.A.* 105, 14325–14329.
- Broder, A. Z. (1997). “On the resemblance and containment of documents,” in *Proceedings of the Compression and Complexity of Sequences* (Salerno, Italy), 21–27.
- Bülthoff, H. H., and Edelman, S. (1992). Psychophysical support for a 2-D view interpolation theory of object recognition. *Proc. Natl. Acad. Sci. U.S.A.* 89, 60–64.
- Chater, N., and Vitányi, P. (2003). The generalized universal law of generalization. *J. Math. Psychol.* 47, 346–369.
- Chater, N., Tenenbaum, J. B., and Yuille, A. (2006). Probabilistic models of cognition: conceptual foundations. *Trends Cogn. Sci.* 10, 287–291.
- Connor, C. E., Preddie, D. C., Gallant, J. L., and Van Essen, D. C. (1997). Spatial attention effects in macaque area V4. *J. Neurosci.* 17, 3201–3214.
- Cortes, C., and Vapnik, V. (1995). Support-vector networks. *Mach. Learn.* 20, 273–297.
- Cover, T., and Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory* 13, 21–27.
- Cox, D. D., Meier, P., Oertelt, N., and DiCarlo, J. J. (2005). ‘Breaking’ position-invariant object recognition. *Nat. Neurosci.* 8, 1145–1147.
- Craik, K. J. W. (1943). *The Nature of Explanation*. Cambridge, England: Cambridge University Press.
- Cutzu, F., and Edelman, S. (1996). Faithful representation of similarities among three-dimensional shapes in human vision. *Proc. Natl. Acad. Sci. U.S.A.* 93, 12046–12050.
- Cutzu, F., and Edelman, S. (1998). Representation of object similarity in human vision: psychophysics and a computational model. *Vision Res.* 38, 2227–2257.
- DeMers, D., and Cottrell, G. (1993). “Nonlinear dimensionality reduction,” in *Advances in Neural Information Processing Systems* 5, eds S. J. Hanson, J. D. Cowan, and C. L. Giles (Washington, DC: Morgan Kaufmann), 580–587.
- Dennett, D. C. (2003). *Freedom Evolves*. New York, NY: Viking.
- Dewey, J. (1910). *How We Think*. Lexington, MA: D. C. Heath.
- DiCarlo, J. J., and Cox, D. D. (2007). Untangling invariant object recognition. *Trends Cogn. Sci.* 11, 333–341.
- DiCarlo, J. J., and Maunsell, J. H. R. (2003). Anterior inferotemporal neurons of monkeys engaged in object recognition can be highly sensitive to object retinal position. *J. Neurophysiol.* 89, 3264–3278.
- Dill, M., and Edelman, S. (2001). Imperfect invariance to object translation in the discrimination of complex shapes. *Perception* 30, 707–724.
- Divvala, S. K., Hoiem, D., Hays, J., Efros, A. A., and Hebert, M. (2009). “An empirical study of context in object detection,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (Cambridge, MA).
- Duvdevani-Bar, S., and Edelman, S. (1999). Visual recognition and categorization on the basis of similarities to multiple class prototypes. *Int. J. Comput. Vis.* 33, 201–228.
- Duvdevani-Bar, S., Edelman, S., Howell, A. J., and Buxton, H. (1998). “A similarity-based method for the generalization of face recognition over pose and expression,” in *Proceedings of the 3rd International Symposium on Face and Gesture Recognition (FG98)*, eds S. Akamatsu and K. Mase (Washington, DC), 118–123.
- Edelman, S. (1995). Representation, similarity, and the chorus of prototypes. *Minds Mach.* 5, 45–68.
- Edelman, S. (1998). Representation is representation of similarity. *Behav. Brain Sci.* 21, 449–498.
- Edelman, S. (1999). *Representation and Recognition in Vision*. Cambridge, MA: MIT Press.
- Edelman, S. (2008). *Computing the Mind: How the Mind Really Works*. New York, NY: Oxford University Press.
- Edelman, S., and Bülthoff, H. H. (1992). Orientation dependence in the recognition of familiar and novel views of 3D objects. *Vision Res.* 32, 2385–2400.
- Edelman, S., and Intrator, N. (2000). (Coarse coding of shape fragments) + (retinotopy)  $\approx$  representation of structure. *Spat. Vis.* 13, 255–264.
- Edelman, S., and Intrator, N. (2002). “Models of perceptual learning,” in *Perceptual Learning*, eds M. Fahle and T. Poggio (Berlin: MIT Press), 337–353.
- Edelman, S., and Intrator, N. (2003). Towards structural systematicity in distributed, statically bound visual representations. *Cogn. Sci.* 27, 73–109.
- Edelman, S., and Shahbazi, R. (2011). Survival in a world of probable objects. *Behav. Brain Sci.* 34, 197–198. A commentary on *Bayesian Fundamentalism or Enlightenment? On the Explanatory Status and Theoretical Contributions of Bayesian Models of Cognition* by Jones and Love.
- Edelman, S., and Duvdevani-Bar, S. (1997). “Similarity-based viewspace interpolation and the categorization of 3D objects,” in *Proceedings of the Similarity and Categorization Workshop*, (Department of AI, University of Edinburgh), 75–81.
- Edelman, S., Bülthoff, H. H., and Bülthoff, I. (1999). Effects of parametric manipulation of inter-stimulus similarity on 3D object recognition. *Spat. Vis.* 12, 107–123.
- Edelman, S., Grill-Spector, K., Kushnir, T., and Malach, R. (1998). Towards direct visualization of the internal shape representation space by fMRI. *Psychobiology* 26, 309–321.
- Edelman, S., Intrator, N., and Jacobson, J. S. (2002). “Unsupervised learning of visual structure,” in *Proceedings of the 2nd International Workshop on Biologically Motivated Computer Vision, Volume 2525 of Lecture Notes in Computer Science*, eds H. H. Bülthoff, C. Wallraven, S.-W. Lee, and T. Poggio (New York, NY: Springer), 629–643.
- Eisler, H. (1960). Similarity in the continuum of heaviness with some methodological and theoretical considerations. *Scand. J. Psychol.* 1, 69–81.
- Epstein, R., and Kanwisher, N. (1998). A cortical representation of the local

- visual environment. *Nature* 392, 598–601.
- Eshghi, K., and Rajaram, S. (2008). “Locality sensitive hash functions based on concomitant rank order statistics,” in *The 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD08)*, (New York, NY).
- Fischler, M. A., and Bolles, R. C. (1981). Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM* 24, 381–395.
- Frege, G. (1891). “On sense and reference,” in *Translations from the Philosophical Writings of G. Frege*, eds P. Geach and M. Black (Oxford: Blackwell). Translated as “On Sense and Meaning” (1993), 56–78.
- Gallant, J. L., Shoup, R. E., and Mazer, J. A. (2000). A human extrastriate area functionally homologous to Macaque V4. *Neuron* 27, 227–235.
- Garner, W. R., and Felfoldy, G. L. (1970). Integrality of stimulus dimensions in various types of information processing. *Cogn. Psychol.* 1, 225–241.
- Giese, M. A., Thornton, I., and Edelman, S. (2008). Metrics of the perception of body movement. *J. Vis.* 8, 1–18.
- Goodman, N. (1972). *Seven Strictures on Similarity*. Indianapolis, IN: Bobbs Merill.
- Hahn, U., and Chater, N. (1998). “Similarity and rules: distinct? exhaustive? empirically distinguishable?” in *Similarity and symbols in human thinking*, eds S. A. Sloman and L. J. Rips (Cambridge, MA: MIT Press), 111–144.
- Hausser, D. (1992). Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Inf. Comput.* 100, 78–150.
- Hayworth, K. J., Lescroart, M. D., and Biederman, I. (2011). Neural encoding of relative position. *J. Exp. Psychol. Hum. Percept. Perform.* 37, 1032–1050.
- He, J., Liu, W., and Chang, S.-F. (2010). “Scalable similarity search with optimized kernel hashing,” in *Proceedings of the Knowledge Discovery and Data Mining (KDD’10)*, (Boston, MA). 1129–1138.
- Huber, P. J. (1985). Projection pursuit (with discussion). *Ann. Stat.* 13, 435–475.
- Hume, D. (1748). *An Enquiry Concerning Human Understanding*. Available online at <http://eserver.org/18th/hume-enquiry.html>.
- Hummel, J. E. (2001). Complementary solutions to the binding problem in vision: implications for shape perception and object recognition. *Vis. Cogn.* 8, 489–517.
- Hummel, J. E., and Holyoak, K. J. (1998). Distributed representations of structure: a theory of analogical access and mapping. *Psychol. Rev.* 104, 427–466.
- Huttenlocher, D. P., and Ullman, S. (1987). “Object recognition using alignment,” in *Proceedings of the 1st International Conference on Computer Vision*, (London, England; Washington, DC: IEEE), 102–111.
- Intrator, N., and Cooper, L. N. (1992). Objective function formulation of the BCM theory of visual cortical plasticity: statistical connections, stability conditions. *Neural Netw.* 5, 3–17.
- Intrator, N., and Edelman, S. (1996). How to make a low-dimensional representation suitable for diverse tasks. *Connect. Sci.* 8, 205–224.
- Intrator, N., and Edelman, S. (1997). Learning low dimensional representations of visual objects with extensive use of prior knowledge. *Network* 8, 259–281.
- Johnson, K. E. (2004). On the systematicity of language and thought. *J. Philos.* 111–139.
- Johnson, W. B., and Lindenstrauss, J. (1984). Extensions of Lipschitz mappings into a Hilbert space. *Contemp. Math.* 26, 189–206.
- Jones, M. N., and Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychol. Rev.* 114, 137.
- Kemp, C., Bernstein, A., and Tenenbaum, J. B. (2005). “A generative theory of similarity,” in *Proceedings of the Twenty-Seventh Annual Conference of the Cognitive Science Society*, (Washington, DC).
- Kravitz, D. J., Vinson, L. D., and Baker, C. I. (2008). How position dependent is visual object recognition? *Trends Cogn. Sci.* 12, 114–122.
- Kulis, B., and Grauman, K. (2009). “Kernelized locality-sensitive hashing for scalable image search,” in *Proceedings of the 12th International Conference on Computer Vision (ICCV)*, 2130–2137.
- Kulkarni, G., Premraj, V., Dhar, S., Li, S., Choi, Y., Berg, A. C., and Berg, T. L. (2011). “Baby talk: understanding and generating simple image descriptions,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, eds T. Boult, T. Kanade, and S. Peleg, (Colorado Springs, CO).
- Lamdan, Y., and Wolfson, H. (1988). “Geometric hashing: a general and efficient recognition scheme,” in *Proceedings of the 2nd International Conference on Computer Vision*, (Tarpon Springs, FL; Washington, DC: IEEE), 238–251.
- Lando, M., and Edelman, S. (1995). Receptive field spaces and class-based generalization from a single view in face recognition. *Network* 6, 551–576.
- Lazebnik, S., Schmid, C., and Ponce, J. (2006). “Beyond bags of features: spatial pyramid matching for recognizing natural scene categories,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (Washington, DC).
- Li, P., and König, A. C. (2011). Theory and applications of b-bit min-wise hashing. *Commun. ACM* 54, 101–109.
- Linde, Y., Buzo, A., and Gray, R. (1980). An algorithm for vector quantizer design. *IEEE Trans. Commun.* 28, 84–95.
- Loeff, N., and Farhadi, A. (2008). “Scene discovery by matrix factorization,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, (New York, NY).
- Logothetis, N. K., Pauls, J., Poggio, T., and Bülthoff, H. H. (1994). View dependent object recognition by monkeys. *Curr. Biol.* 4, 404–441.
- Logothetis, N., and Pauls, J. (1995). Psychophysical and physiological evidence for viewer-centered object representations in the primate. *Cereb. Cortex* 3, 270–288.
- Lowe, D. G. (1999). “Object recognition from local scale-invariant features,” in *Proceedings of the International Conference on Computer Vision*, Vol. 2. (Washington, DC), 1150–1157.
- MacEvoy, S. P., and Epstein, R. A. (2011). Constructing scenes from objects in human occipitotemporal cortex. *Nat. Neurosci.* 14, 1323–1331.
- Malisiewicz, T., and Efros, A. A. (2009). “Beyond categories: the visual Memex model for reasoning about object relationships,” in *Proceedings of the 22nd Neural Information Processing Systems Conference (NIPS)*, eds Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, 1222–1230.
- Malisiewicz, T., Gupta, A., and Efros, A. A. (2011). “Ensemble of exemplar-SVMs for object detection and beyond,” in *Proceedings of the International Conference on Computer Vision (ICCV) 2011*, eds D. Metaxas, L. Quan, A. Sanfeliu, and L. Van Cool, (Barcelona, Spain).
- Markman, A., and Gentner, D. (1993). Structural alignment during similarity comparisons. *Cogn. Psychol.* 25, 431–467.
- Marr, D. (1982). *Vision*. San Francisco, CA: W. H. Freeman.
- Medin, D. L., Goldstone, R. L., and Gentner, D. (1993). Respects for similarity. *Psychol. Rev.* 100, 254–278.
- Mel, B. W., and Fiser, J. (2000). Minimizing binding errors using learned conjunctive features. *Neural Comput.* 12, 247–278.
- Merker, B. (2004). Cortex, countercurrent context, and dimensional integration of lifetime memory. *Cortex* 40, 559–576.
- Minsky, M. (1961). Steps toward artificial intelligence. *Proc. Inst. Radio Eng.* 49, 8–30.
- Moses, Y., Ullman, S., and Edelman, S. (1996). Generalization to novel images in upright and inverted faces. *Perception* 25, 443–462.
- Newell, F. N., Sheppard, D., Edelman, S., and Shapiro, K. (2005). The interaction of shape- and location-based priming in object categorisation: evidence for a hybrid what+where representation stage. *Vision Res.* 45, 2065–2080.
- Oliva, A., and Torralba, A. (2001). Modeling the shape of the scene: a holistic representation of the spatial envelope. *Int. J. Comput. Vis.* 42, 145–175.
- Op de Beeck, H., and Vogels, R. (2000). Spatial sensitivity of Macaque inferior temporal neurons. *J. Comp. Neurol.* 426, 505–518.
- Op de Beeck, H., Wagemans, J., and Vogels, R. (2001). Inferotemporal neurons represent low-dimensional configurations of parameterized shapes. *Nat. Neurosci.* 4, 1244–1252.
- Panis, S., Vangeneugden, J., Op de Beeck, H. P., and Wagemans, J. (2008). The representation of subordinate shape similarity in human occipitotemporal cortex. *J. Vis.* 8, 1–15.
- Paulevé, L., Jégou, H., and Amsaleg, L. (2010). Locality sensitive hashing: A comparison of hash function types and querying mechanisms. *Pattern Recognit. Lett.* 31, 1348–1358.
- Perrett, D. I., and Oram, M. W. (1998). Visual recognition based on temporal cortex cells: viewer-centred processing of pattern configuration. *Z. Naturforsch.* 53, 518–541.



- Plate, T. A. (1991). "Holographic reduced representations: convolution algebra for compositional distributed representations," in *Proceedings of the 12th International Joint Conference on Artificial Intelligence (IJCAI)*, eds J. Mylopoulos and R. Reiter (San Mateo, CA: Morgan Kaufmann), 30–35.
- Poggio, T. (1990). A theory of how the brain might work. *Cold Spring Harbor Symposia on Quantitative Biology*, LV, 899–910.
- Poggio, T., and Girosi, F. (1990). Regularization algorithms for learning that are equivalent to multilayer networks. *Science* 247, 978–982.
- Poggio, T., and Edelman, S. (1990). A network that learns to recognize three-dimensional objects. *Nature* 343, 263–266.
- Rao, S. C., Rainer, G., and Miller, E. K. (1997). Integration of what and where in the primate prefrontal cortex. *Science* 276, 821–824.
- Rips, L. J. (1989). "Similarity, typicality, and categorization," in *Similarity, Analogy, and Thought*, eds S. Vosniadu and A. Ortony (Cambridge: Cambridge University Press), 21–59.
- Rissanen, J. (1987). "Minimum description length principle," in *Encyclopedia of Statistical Sciences*, Vol. 5, eds S. Kotz and N. L. Johnson (Washington, DC: J. Wiley and Sons), 523–527.
- Rousset, G. A., Thorpe, S. J., and Fabre-Thorpe, M. (2003). Taking the MAX from neuronal responses. *Trends Cogn. Sci.* 7, 99–102.
- Russell, B., Torralba, A., Murphy, K., and Freeman, W. T. (2008). LabelMe: a database and web-based tool for image annotation. *Int. J. Comput. Vis.* 77, 157–173.
- Rust, N., and DiCarlo, J. (2010). Selectivity and tolerance ('invariance') both increase as visual information propagates from cortical area V4 to IT. *J. Neurosci.* 30, 12978–12995.
- Sadeghi, M. A., and Farhadi, A. (2011). "Recognition using visual phrases," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, eds T. Boult, T. Kanade, and S. Peleg (Colorado Springs, CO).
- Sahlgren, M., Holst, A., and Kanerva, P. (2008). "Permutations as a means to encode order in word space," in *Proceedings of the 30th Annual Meeting of the Cognitive Science Society*, eds B. C., Love, K. McRae, and V. M. Sloutsky (Austin, TX: Cognitive Science Society), 1300–1305.
- Sali, E., and Ullman, S. (1998). "Recognizing novel 3-D objects under new illumination and viewing position using a small number of example views or even a single view," in *Proceedings of the International Conference on Computer Vision (ICCV)*, IEEE (Washington, DC), 153–164.
- Salinas, E., and Abbott, L. F. (1997). Invariant visual responses from attentional gain fields. *J. Neurophysiol.* 77, 3267–3272.
- Salinas, E., and Thier, P. (2000). Gain modulation: a major computational principle of the central nervous system. *Neuron* 27, 15–21.
- Serre, T., Oliva, A., and Poggio, T. (2008). A feedforward architecture accounts for rapid categorization. *Proc. Natl. Acad. Sci. U.S.A.* 104, 6424–6429.
- Shahbazi, R., Field, D. J., and Edelman, S. (2011). "The role of hierarchy in learning to categorize images," in *Proceedings of the 33rd Cognitive Science Society Conference*, eds L. Carlson, C. Holscher, and T. Shipley (Boston, MA).
- Shakhnarovich, G. (2008). Matlab LSH Toolbox. Retrieved on 2/1/2012 from <http://ttic.uchicago.edu/~gregory/code/lsh/lshcode.tar.gz>.
- Shashua, A. (1992). "Illumination and view position in 3D visual recognition," in *Neural Information Processing Systems*, Vol. 4, eds J. Moody, S. J. Hanson, and R. L. Lippman (San Mateo, CA: Morgan Kaufmann), 404–411.
- Shepard, R. N. (1980). Multidimensional scaling, tree-fitting, and clustering. *Science* 210, 390–397.
- Shepard, R. N. (1984). Ecological constraints on internal representation: resonant kinematics of perceiving, imagining, thinking, and dreaming. *Psychol. Rev.* 91, 417–447.
- Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science* 237, 1317–1323.
- Shepard, R. N. (2001). Perceptual-cognitive universals as reflections of the world. *Behav. Brain Sci.* 24, 581–601.
- Shepard, R. N., and Chipman, S. (1970). Second-order isomorphism of internal representations: shapes of states. *Cogn. Psychol.* 1, 1–17.
- Sugihara, T., Edelman, S., and Tanaka, K. (1998). Representation of objective similarity among three-dimensional shapes in the monkey. *Biol. Cyber.* 78, 1–7.
- Szabó, Z. (2008). "Compositionality," in *The Stanford Encyclopedia of Philosophy*, ed E. N. Zalta (online).
- Tenenbaum, J. B., and Griffiths, T. L. (2001). Generalization, similarity, and bayesian inference. *Behav. Brain Sci.* 24, 629–641.
- Thorpe, S., Fize, D., and Marlot, C. (1996). Speed of processing in the human visual system. *Nature* 381, 520–522.
- Torralba, A. (2003). Context challenge: contextual priming for object detection. *Int. J. Comput. Vis.* 53, 169–191.
- Townsend, J. T., and Thomas, R. D. (1993). "On the need for a general quantitative theory of pattern similarity," in *Foundations of Perceptual Theory*, ed S. C. Masin, (Amsterdam: Elsevier), 297–368.
- Tsunoda, K., Yamane, Y., Nishizaki, M., and Tanifuji, M. (2001). Complex objects are represented in macaque inferotemporal cortex by the combination of feature columns. *Nat. Neurosci.* 4, 832–838.
- Tversky, A. (1977). Features of similarity. *Psychol. Rev.* 84, 327–352.
- Tversky, A., and Gati, I. (1978). "Studies of similarity," in *Cognition and Categorization*, eds E. Rosch and B. Lloyd (Washington, DC: Erlbaum), 79–98.
- Ullman, S. (1989). Aligning pictorial descriptions: an approach to object recognition. *Cognition* 32, 193–254.
- van Dam, W. O., and Hommel, B. (2010). How object-specific are object files? evidence for integration by location. *J. Exp. Psychol. Hum. Percept. Perform.* 36, 1184–1192.
- van Gemert, J. C., Veenman, C. J., Smeulders, A. W. M., and Geusebroek, J.-M. (2010). Visual word ambiguity. *IEEE Trans. Pattern Anal. Mach. Intell.* 32, 1271–1283.
- Vogels, R. (1999). Effect of image scrambling on inferior temporal cortical responses. *Neuroreport* 10, 1811–1816.
- Wachsmuth, E., Oram, M. W., and Perrett, D. I. (1994). Recognition of objects and their component parts: responses of single units in the temporal cortex of the macaque. *Cereb. Cortex* 5, 509–522.
- Wang, J., Yang, J., Yu, K., Lv, K., Huang, T., and Gong, Y. (2010). "Locality-constrained linear coding for image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (San Francisco, CA), 3360–3367.
- Wang, Y., Fujita, I., and Murayama, Y. (2000). Neuronal mechanisms of selectivity for object features revealed by blocking inhibition in inferotemporal cortex. *Nat. Neurosci.* 3, 807–813.
- Watanabe, S. (1969). *Knowing and Guessing: A Quantitative Study of Inference and Information*. New York, NY: Wiley.
- Wickelgren, W. (1969). Context-sensitive coding, associative memory, and serial order in (speech) behavior. *Psychol. Rev.* 76, 115.
- Willshaw, D. J., Buneman, O. P., and Longuet-Higgins, H. C. (1969). Non-holographic associative memory. *Nature* 222, 960–962.
- Yagnik, J., Strelow, D., Ross, D. A., and Lin, R. (2011). "The power of comparative reasoning," in *Proceedings of the International Conference on Computer Vision (ICCV'11)*, eds D. Metaxas, L. Quan, A. Sanfeliu, and L. Van Gool (Barcelona, Spain).
- Zhang, Y., Meyers, E. M., Bichot, N. P., Serre, T., Poggio, T. A., and Desimone, R. (2011). Object decoding with attention in inferior temporal cortex. *Proc. Natl. Acad. Sci. U.S.A.* 108, 8850–8855.
- Zhang, Y., Jia, Z., and Chen, T. (2011). "Image retrieval with geometry-preserving visual phrases," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Providence, RI).

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 15 November 2011; accepted: 24 June 2012; published online: 13 July 2012.

Citation: Edelman S and Shahbazi R (2012) Renewing the respect for similarity. *Front. Comput. Neurosci.* 6:45. doi: 10.3389/fncom.2012.00045

Copyright © 2012 Edelman and Shahbazi. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in other forums, provided the original authors and source are credited and subject to any copyright notices concerning any third-party graphics etc.



# Learned non-rigid object motion is a view-invariant cue to recognizing novel objects

Lewis L. Chuang<sup>1</sup>, Quoc C. Vuong<sup>1</sup> and Heinrich H. Bülthoff<sup>1,2\*</sup>

<sup>1</sup> Department of Perception, Cognition and Action, Max Planck Institute for Biological Cybernetics, Tübingen, Germany

<sup>2</sup> Department of Brain and Cognitive Engineering, Korea University, Seoul, Korea

## Edited by:

Jay Hegdé, Georgia Health Sciences University, USA

## Reviewed by:

Fang Fang, Peking University, China  
David D. Cox, Harvard University, USA

## \*Correspondence:

Heinrich H. Bülthoff, Department of Perception, Cognition and Action, Max Planck Institute for Biological Cybernetics, Spemannstrasse 41, 72076 Tübingen, Germany.  
e-mail: hhb@tuebingen.mpg.de

There is evidence that observers use learned object motion to recognize objects. For instance, studies have shown that reversing the learned direction in which a rigid object rotated in depth impaired recognition accuracy. This motion reversal can be achieved by playing animation sequences of moving objects in reverse frame order. In the current study, we used this sequence-reversal manipulation to investigate whether observers encode the motion of dynamic objects in visual memory, and whether such dynamic representations are encoded in a way that is dependent on the viewing conditions. Participants first learned dynamic novel objects, presented as animation sequences. Following learning, they were then tested on their ability to recognize these learned objects when their animation sequence was shown in the same sequence order as during learning or in the reverse sequence order. In Experiment 1, we found that non-rigid motion contributed to recognition performance; that is, sequence-reversal decreased sensitivity across different tasks. In subsequent experiments, we tested the recognition of non-rigidly deforming (Experiment 2) and rigidly rotating (Experiment 3) objects across novel viewpoints. Recognition performance was affected by viewpoint changes for both experiments. Learned non-rigid motion continued to contribute to recognition performance and this benefit was the same across all viewpoint changes. By comparison, learned rigid motion did not contribute to recognition performance. These results suggest that non-rigid motion provides a source of information for recognizing dynamic objects, which is not affected by changes to viewpoint.

**Keywords:** visual object recognition, motion, spatio-temporal signature, non-rigid motion, reversal effect, view-dependency, rigid motion, depth rotation

## INTRODUCTION

Object motion can play an important role in the detection and perception of three-dimensional (3D) objects. For example, the perceptual system can use translational motion to group image fragments of the same object and segregate it from a cluttered background (Fahle, 1993; Nygård et al., 2009). In addition, an object's 3D structure and shape can be recovered from a sequence of two-dimensional (2D) images that depict its rotations in depth using structure-from-motion computations (Ullman, 1979; Grzywacz and Hildreth, 1987).

The role of object motion is not limited to shape recovery. There is evidence that object motion *per se* can be directly used to recognize objects (e.g., Stone, 1998, 1999; Lander and Bruce, 2000; Knappmeyer et al., 2003; Liu and Cooper, 2003; Newell et al., 2004; Vuong and Tarr, 2006; Vuong et al., 2009; Setti and Newell, 2010). For example, Johansson's (1973) classic point-light display demonstrates that an observer can use only the motion of dots attached to the joints of an otherwise invisible human actor to recognize the actor's action (e.g., walking or dancing), sex, or even identity if the observer is highly familiar with the actor (Cutting and Kozlowski, 1977). Other studies have shown that manipulating an object's learned motion can affect observers' performance on different recognition tasks (e.g., Stone, 1998, 1999; Liu and Cooper, 2003).

However, it is not clear how object motion is encoded in visual memory. To address this issue, we tested observers' ability to recognize dynamic objects from different perspective viewpoints. When an object is seen from different viewpoints, it projects different 2D retinal images (e.g., imagine viewing a car from the side or from above). Importantly, the larger the difference between two viewpoints is, the more visually dissimilar the projected images will be. For static objects, measuring how viewpoint changes affect recognition performance has helped to reveal how static object features (e.g., edges and parts) are encoded in visual memory (e.g., Biederman, 1987; Tarr et al., 1998; Foster and Gilson, 2002). There is evidence from different recognition tasks that static features can be encoded in a view-invariant or view-dependent manner (see Peissig and Tarr, 2007, for a review). Using a similar strategy, we systematically manipulated the viewpoint to determine whether object motion is encoded in a view-invariant or view-dependent manner.

Features that are encoded in a view-invariant manner in visual memory are robust to changes in viewing conditions (e.g., viewpoint change or illumination change). In comparison, features that are encoded in a view-dependent manner are stored in visual memory as they appear to an observer under specific viewing conditions (e.g., like a template). They are thus less robust to changes to viewing conditions. One way to distinguish between these two types of

features is to test recognition performance across changes in viewpoints (Peissig and Tarr, 2007). That is, one can test how observers' recognition performance (e.g., accuracy and/or response times) varies with changes in viewpoint. Typically, recognition performance decreases with increasing differences between a familiar and a novel viewpoint (e.g., Bülthoff and Edelman, 1992; Tarr et al., 1998). This robust viewpoint effect across many stimuli and recognition tasks has motivated many computational models to adopt a view-dependent approach to understanding visual object recognition (e.g., Serre et al., 2007; Ullman, 2007; for a view-invariant approach, see Hummel and Biederman, 1992).

To date, only a few studies have investigated how object motion affected recognition performance across changes in viewpoint. For example, the recognition of non-rigid facial motion (e.g., expressions) has been shown to be less affected by viewpoint changes than the recognition of rigid (e.g., head nodding) and non-rigid facial motion combined (Watson et al., 2005). The recognition of point-light walkers has also been shown to be influenced by view-dependent information and insensitive to distortions of the human body's 3D structure (Bülthoff et al., 1998). More recently, Vuong et al. (2009) found that observers could use the articulatory motion of novel objects to help them recognize objects across larger viewpoint changes. These articulatory motions are similar to the movements of the human body.

Stone (1998) referred to the learned motion of a dynamic object as its spatio-temporal signature. He demonstrated that observers directly used these signatures for object recognition (Stone, 1998, 1999). In his studies, observers first learned a small set of novel amoeboid objects that rotated rigidly in depth with a tumbling motion. During the learning phase, the objects always rotated in depth in the same manner (and particularly in the same direction). These objects were presented as an animation consisting of an ordered sequence of views (i.e., a video). When observers' reached a learning criterion, Stone reversed the rotation direction of these now familiar objects, by presenting the learned animation sequence in reverse frame order (i.e., presenting videos of the learned objects backward). This *sequence-reversal* manipulation reduced recognition accuracy by as much as 22%. Importantly, this manipulation does not disrupt the spatial properties of the 2D images in the animation sequence nor does it disrupt structure-from-motion processes (Ullman, 1979). Therefore, sequence-reversal effects supported the claim that a moving object provides dynamic information *per se* for recognition, in addition to static shape information (Stone, 1998, 1999).

Sequence-reversal has been used extensively to study the role of object motion in recognition across different tasks, stimuli, and even species. The sequence-reversal effect has been demonstrated with a large set of 32 rigidly rotating objects, which were implicitly learned (Liu and Cooper, 2003). In addition, the effect has been shown to be more prominent when observers identified objects with highly similar shapes compared to those with highly distinctive 3D structures (Vuong and Tarr, 2006). In addition, Wang and Zhang (2010) showed that observers were also sensitive to local frame sequences. In their study, they took an animation sequence and divided it into shorter sub-sequences. They then reversed the frame order within these "local" sub-sequences, while preserving the "global" order of the sub-sequences themselves. They found

that observers' recognition performance was impaired in this case. The sequence-reversal effect has also been demonstrated with non-rigidly moving faces (Lander and Bruce, 2000). Finally, this effect has even been shown with pigeons, indicating that sequence-reversal disrupts a source of visual information that is not unique to human cognition (Spetch et al., 2006).

The current experiments were conducted to investigate the effect of sequence-reversal on the recognition of dynamic amoeboid objects across changes in viewpoint. These objects were chosen because they lack a distinctive geometric structure and because they do not constitute a highly familiar object class (e.g., faces). If observers rely on an object's motion, sequence-reversal would impair recognition performance, compared to preserving the learned sequence order. On the other hand, there would be no influence of sequence-reversal if recognition depends strictly on static view-dependent information (e.g., 2D shape features) because these features are not disrupted by this manipulation. In addition, we investigated how the effect of sequence-reversal interacted with viewpoint changes for non-rigid and rigid object motion.

## MATERIALS AND METHODS

Three experiments were conducted to assess how participants encoded object motion learned from a specific viewpoint. In particular, the experiments were designed to determine whether object motion was encoded for recognition in a view-invariant or view-dependent manner (Watson et al., 2005; Perry et al., 2006; Vuong et al., 2009). Each experiment consisted of a familiarization phase, followed by a testing phase. In the familiarization phase, participants learned two objects that deformed non-rigidly (Experiments 1–2) or rotated rigidly in depth over time (Experiment 3). Each object's motion was the same on every trial during this phase. In the testing phase, observers were required to discriminate the learned target objects from two new distracter objects.

To replicate previous findings (e.g., Stone, 1998, 1999; Lander and Bruce, 2000; Liu and Cooper, 2003; Vuong and Tarr, 2006), we first investigated if sequence-reversal affected the recognition of novel non-rigidly deforming objects on an old-new recognition task (Experiment 1a) and a two-interval forced-choice (2IFC) task (Experiment 1b). Following this, we investigated the effect of sequence-reversal on recognizing non-rigidly deforming (Experiment 2) or rigidly rotating (Experiment 3) objects across a range of novel viewpoints.

## PARTICIPANTS

Seventy volunteers (age range: 18–35 years) were recruited from the Institute's participant database – E1a: 16; E1b: 14; E2: 24; E3: 21. They were paid 8€/h for their time and provided informed consent, approved by the local ethics committee. All participants had normal or corrected-to-normal vision and did not participate in more than one experiment.

## APPARATUS

The experiments were conducted on a Macintosh G4 computer, which was controlled by customized MATLAB software that used the PsychToolBox extension (Brainard, 1997; Pelli, 1997). The

stimuli were presented on a 21" CRT monitor with a resolution of  $1152 \times 864$  pixels and a refresh rate of 75 Hz. Participants were seated 60 cm from the screen. All responses were collected from a standard keyboard.

## MATERIAL

**Figure 1** shows an example of the 3D amoeboid object used in the current study. All the visual stimuli were bounded by a square that was centered on the screen (diagonal  $\approx 15.6^\circ$ ). The experimental stimuli were derived from animation sequences of 100 numerically labeled images ( $320 \times 320$  pixels) that depicted the objects moving smoothly over time (22 frames/s), either deforming non-rigidly (Experiments 1 and 2) or rotating rigidly in depth (Experiment 3). Each sequence was rendered from seven camera viewpoints (see **Figure 1B**).

The 3D objects and their animation sequences were produced using 3D Studio Max (v. 7; Autodesk, Montreal). For each object, the 3D coordinates of a sphere's vertices were smoothly modulated by the application of a series of random sinusoidal deformation fields in Studio Max (see **Figure 1A**). By randomly shifting the phase of the sinusoidal deformation fields applied to the base sphere, we could synthesize amoeboids with different 3D shapes (Norman et al., 1995).

Non-rigid deformations could be introduced by shifting the phases of these sinusoidal deformation fields simultaneously at a rate of  $\sim 0.16$  cycles every 20th frame. This induced a smooth deformation of each object's 3D structure over time. Alternatively, each object could be rigidly rotated about its center to a new pose every 20th frame. This produced a smooth rigid tumbling motion that did not deform the object's 3D structure. A randomly determined sequence of poses ensured that each object had a unique rigid rotational path in depth.

Altogether, 4 non-rigidly deforming objects were created for Experiment 1, 16 non-rigidly deforming objects for Experiment 2, and 16 rigidly rotating objects for Experiment 3. For each participant, four objects were randomly selected from the set of possible objects of the relevant experiment. Two of the objects were randomly assigned to be targets and two as distracters.

A virtual camera was positioned in front of each object and focused on its center of mass. This was designated as the

$0^\circ$  viewpoint (the white camera in **Figure 1B**). This camera was rotated clockwise or counter-clockwise along the azimuth. Ordered sequences of 100 images were then rendered for each object from seven viewpoints ( $0^\circ, \pm 20^\circ, \pm 40^\circ, \pm 60^\circ$ ; see **Figure 1B**). Video examples are provided as Supplementary Material. All participants learned the objects from the  $0^\circ$  viewpoint during the familiarization phase. In addition, a grayscale luminance noise pattern served as a mask.

## EXPERIMENTAL PROCEDURE

**Figure 2** illustrates the trial sequence on the familiarization and testing phases for the old-new recognition task (Experiment 1) and the 2IFC task (Experiments 1b, 2, and 3).

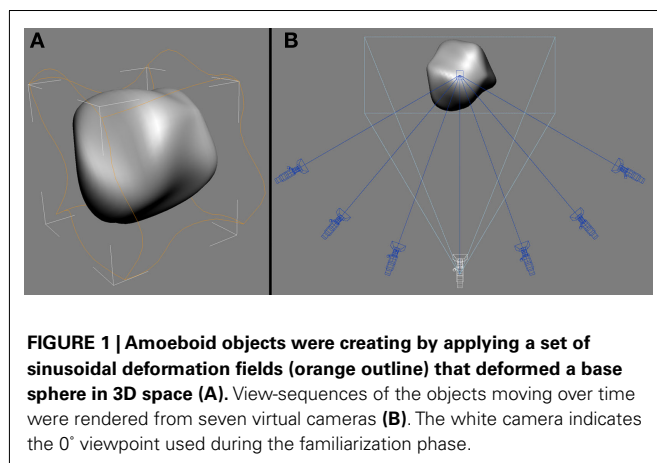
### Familiarization phase

The familiarization phase was the same for all three experiments. During this phase, one of the two target objects was presented on each trial. The stimulus was a 75-image sequence that was sampled from the object's full 100-image sequence. These sequences were always presented in numerically ascending-order. After the presentation of each stimulus ( $\sim 3.4$  s), a noise mask appeared until participants responded with one of two keys (i.e., *y* or *b*) to indicate the object's identity. Each target was randomly assigned a key. Participants were provided with an auditory feedback for incorrect responses. Every participant performed 104 familiarization trials.

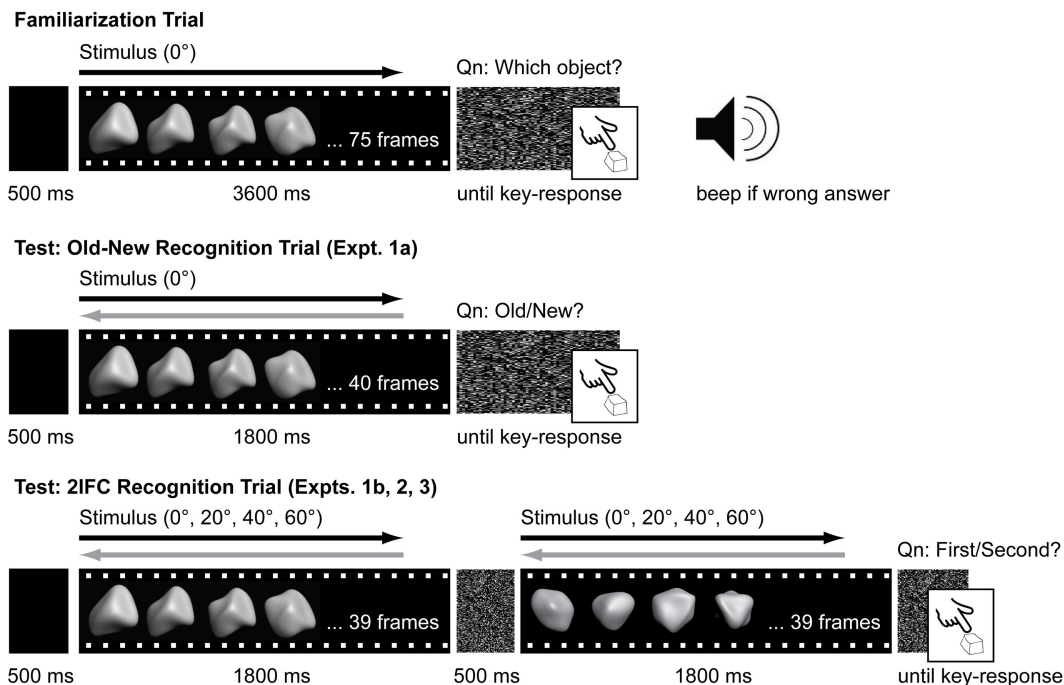
### Testing phase

During the testing phase, participants had to discriminate targets learned during the familiarization phase from distracters. In this phase, the stimuli were shorter animation sequences (i.e., 40 sequential images for Experiment 1a; and 39 sequential images for Experiments 1b, 2, and 3) of the two targets learned during the familiarization phase or two distracters. These sequences lasted  $\sim 1.8$  s each. For the old-new recognition task (Experiment 1a), participants were presented with one stimulus on each trial and had to decide whether that stimulus was *old* (i.e., one of the targets) or *new* (i.e., one of the distracters) by responding with one of two keys after the stimulus presentation ended. For the 2IFC task (Experiments 1b–3), two stimuli were presented sequentially on each trial, one of which was a target and one of which was a distracter. The target and distracter were separated by a 500 ms noise mask. There was also a noise mask presented at the end of the second interval, which stayed on the screen until participants responded. Each target object appeared equally often in the first and second interval. Participants had to decide which interval contained the target object. They were only allowed to respond after both stimuli had been presented. In all experiments, participants were encouraged to respond as quickly and as accurately as possible.

The dynamic objects could be shown in either ascending or descending-order frame sequences. For the target objects, the ascending-order sequence was the *same* (learned) object motion and the descending-order sequence was the *reverse* object motion. For Experiments 2 and 3, target objects could be presented from all seven viewpoints (i.e.,  $0^\circ, \pm 20^\circ, \pm 40^\circ, \pm 60^\circ$ ). The distracter objects in these two experiments were presented at one of these viewpoints, which were randomly chosen. Participants were informed







**FIGURE 2 | Timeline of familiarization and test trials.** From top to bottom: Familiarization trial for all experiments, Old-new recognition test trial for Experiment 1a, 2IFC recognition test trial for Experiments 1b, 2, and 3. On familiarization trials, animations for the stimuli were always presented in ascending sequence order (black arrow) and objects were always presented

from the 0° viewpoint. On test trials, animations for the stimuli could be animated in either the same sequence order as during familiarization (black arrow) or in reverse sequence order (gray arrow). With the exception of Experiment 1, objects could also be presented from a range of perspective viewpoints (0°, ±20°, ±40°, ±60°).

that the target objects' motion could be reversed relative to their motion in the familiarization phase. They were instructed to continue to respond to these as targets.

The test stimuli were sampled only from the central range of the full 100-image sequences (i.e., images 26–75); images that comprised this range were presented equally often during the familiarization phase. The four objects (two targets and two distracters) were presented equally often. There were an equal number of trials in all test conditions (sequence order in Experiment 1; sequence order and viewpoint difference in Experiments 2 and 3). There were a total of 352 test trials for Experiment 1a, 192 trials for Experiment 1b, and 224 trials for Experiments 2 and 3.

## RESULTS

Recognition performance in the test conditions was measured by sensitivity ( $d'$ ; MacMillan and Creelman, 1991). **Figure 3** summarizes sensitivity scores for Experiments 1, 2, and 3, which were collapsed for the direction of the viewpoint difference in Experiments 2 and 3. In the present study, we focused on observers' sensitivity data because we were interested in how object motion was encoded in visual memory. Nonetheless, it should be noted that response-time results were consistent with sensitivity scores and there was no evidence of any speed-accuracy trade-offs. The sensitivity data were submitted to paired-sampled  $t$ -tests or repeated-measures analyses of variance (ANOVAs). Confidence intervals were computed using the within-subjects error term from the sequence order condition (Experiment 1) or its interaction with viewpoint

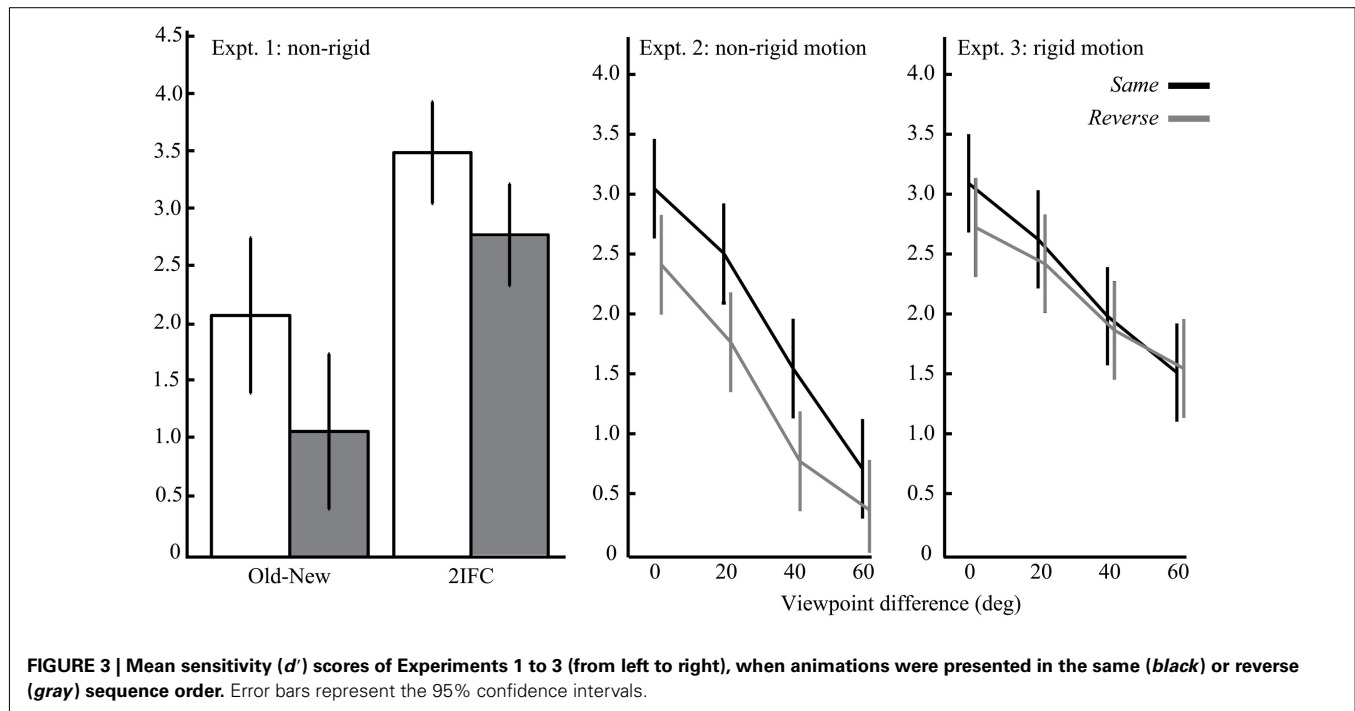
difference (Experiments 2 and 3), where appropriate (Loftus and Masson, 1994). An  $\alpha$ -level of 0.05 indicated statistical significance. Greenhouse–Geisser corrections were applied when the assumption of sphericity was violated. In addition, effect sizes were computed as Cohen's  $d$  and partial  $\eta^2$  for the  $t$ -tests and ANOVAs respectively (Morris and DeShon, 2002).

## EXPERIMENT 1

Experiment 1 tested the effect of sequence-reversal of non-rigidly deforming amoeboids on an old-new recognition (Experiment 1a) and 2IFC task (Experiment 1b). A significant main effect of sequence order was found on  $d'$  scores (E1a:  $t_{15} = 3.19$ , Cohen's  $d = 0.81$ ; E1b:  $t_{13} = 3.49$ , Cohen's  $d = 1.02$ ). Participants were more sensitive in recognizing learned objects when they were animated in the same sequence order as during the familiarization phase than when they were animated with the reverse order. Like previous studies on rigid object motion (Stone, 1998, 1999; Liu and Cooper, 2003; Vuong and Tarr, 2006; Wang and Zhang, 2010), these results show that recognition performance is similarly sensitive to learned non-rigid motion.

## EXPERIMENT 2

In Experiment 2, we tested the effect of sequence-reversal of non-rigidly deforming objects across different viewpoints using the 2IFC task. The participants'  $d'$  scores were submitted to a repeated-measures ANOVA for the test conditions of sequence order (same, reverse) and viewpoint difference (0°, ±20°, ±40°, ±60°).



Main effects were found for both sequence order ( $F_{1,23} = 13.0$ , partial  $\eta^2 = 0.36$ ) and viewpoint difference ( $F_{2,44.9} = 42.8$ , partial  $\eta^2 = 0.65$ ). Sequence-reversal and novel objects viewpoints produced lower  $d'$  scores. In addition,  $d'$  decreased linearly as a function of viewpoint difference, as revealed by a significant linear trend ( $F_{1,23} = 67.2$ , partial  $\eta^2 = 0.75$ ). There was no significant interaction between sequence order and viewpoint difference ( $F_{1,69} = 0.66$ , partial  $\eta^2 = 0.03$ ). That is, the sequence-reversal effect was constant across the different viewpoints. Taken together, these findings show that the recognition of non-rigidly deforming objects was sensitive to changes to the learned viewpoint as well as learned object motion.

### EXPERIMENT 3

Experiment 3 was identical to Experiment 2 except that we tested the effect of sequence reversal with rigidly rotating objects. The  $d'$  data from Experiment 3 were submitted to the same ANOVA as in Experiment 2. In contrast to Experiment 2, there was no significant effect of sequence order ( $F_{1,20} = 2.18$ , partial  $\eta^2 = 0.10$ ). However like the previous experiment, there was a significant effect of viewpoint difference ( $F_{3,60} = 13.3$ , partial  $\eta^2 = 0.40$ ). More specifically,  $d'$  decreased linearly as a function of viewpoint difference ( $F_{1,20} = 22.9$ , partial  $\eta^2 = 0.53$ ). There was no significant interaction between sequence order and viewpoint difference in Experiment 3 ( $F_{1,60} = 0.56$ , partial  $\eta^2 = 0.03$ ). Thus, the recognition of rigidly rotating objects in this experiment was sensitive to changes to the learned viewpoint but not to learned object motion.

### DISCUSSION

In the current study, we used a sequence-reversal manipulation to test the extent to which observers encoded object motion *per se*

during learning, and how robust such dynamic representations are to viewpoint changes (Stone, 1998, 1999; Liu and Cooper, 2003; Vuong and Tarr, 2006; Wang and Zhang, 2010). We found a sequence-reversal effect for non-rigidly deforming objects across a variety of tasks (Experiments 1 and 2): Observers performed more accurately (as measured by sensitivity) when target objects were shown in the same sequence order than when they were shown in the reverse sequence order, even though sequence reversal did not disrupt the objects' 3D structure or set of available 2D images. We also found a large viewpoint effect when observers were tested with these objects (Experiment 2): Observers' sensitivity decreased with increasing viewpoint changes from the learned viewpoint. Importantly, however, the benefit of preserving the learned object motion was constant across all magnitudes of viewpoint change. In contrast to non-rigid motion, we found a viewpoint effect but no sequence-reversal effect when the objects rotated rigidly in depth (Experiment 3). Taken together, these results provide insights into how object motion is encoded in visual memory, and provide important constraints for different models of object recognition.

### LEARNED NON-RIGID OBJECT MOTION PROVIDES A VIEW-INVARIANT BENEFIT TO DYNAMIC OBJECT RECOGNITION

In combination with previous studies, our results suggest that the process of visual object recognition relies on both view-dependent shape information as well as motion information (Stone, 1998, 1999; Liu and Cooper, 2003; Vuong and Tarr, 2006; Wang and Zhang, 2010). This conclusion has several important implications. First, by using visually similar amoeboid objects that did not have distinctive static shape features, our results directly show that non-rigid object motion can be encoded in visual object memory. Second, learned non-rigid object motion contributes directly

to the recognition process in a view-invariant manner, although dynamic objects seem to be encoded in view-dependent manner. That is, the pattern of recognition performance suggests that the contribution of learned non-rigid object motion does not deteriorate with increasing disparity between learned and novel viewpoints. Lastly, our findings extend the results from previous studies showing that non-rigid object motion can facilitate view generalization (Watson et al., 2005; Vuong et al., 2009). Importantly, our results show that this facilitation is not restricted to a highly familiar object class (i.e., faces) or restricted to only articulatory motion.

The pattern of recognition performance in Experiment 2 – namely, a consistent contribution of object motion across viewpoint differences – mirrors one that has been reported before (Foster and Gilson, 2002). Foster and Gilson observed that certain object properties, such as the number of discernible parts, led to a uniform benefit to the recognition of novel bent-wire objects, regardless of the viewpoint of the test objects. Objects that were discriminable on the basis of the number of their parts were better recognized than those that did not differ with respect to this property. Nonetheless, observers' recognition performance with these objects also decreased with increasing differences in viewpoint.

Foster and Gilson (2002) proposed that the successful recognition of an object can depend on multiple sources of information, those that are accessible across views and those that are dependent on view-familiarity. Visual object recognition can rely on either or both contributions. Like the number of object parts, learned non-rigid motion could constitute an object property that can be accessed across a range of viewpoints and, thus, provides a view-invariant benefit to recognition. However, recognition can also continue to rely on view-dependent information such as image-based features of an object's shape.

Interestingly, we did not find a significant benefit of learned motion for rigidly rotating objects (Experiment 3). Previous studies which demonstrated a reversal effect with rigid rotation used the same tumbling motion across all objects (Stone, 1998, 1999; Liu and Cooper, 2003; Vuong and Tarr, 2006). In our current study, each object had a unique tumbling motion. Future work will be necessary to determine if this stimulus difference could account for the contrasting results. However, it should be noted that the reversal effect is not automatic; it can be mediated by factors such as shape similarity and task difficulty (Liu and Cooper, 2003; Vuong and Tarr, 2006). For example, it has been shown to be more prominent in the recognition of blobby objects similar to the ones used here and less so with objects which have highly distinctive parts (Vuong and Tarr, 2006). In addition, it is more apparent in the recognition of objects that were learned moving fast compared to those that were learned moving slow (Balas and Sinha, 2009).

Future experiments will be needed to determine the particular *spatio-temporal* aspects of motion that are encoded to give rise to the view-invariant benefit we observed here. For example, optic-flow patterns could be directly represented as a dynamic object property for subsequent recognition (Casile and Giese, 2005). In the next two sections, we outline some possible mechanisms that could explain the contribution of object motion to recognition.

## TEMPORAL ASSOCIATIONS FOR LEARNING OBJECT MOTION

In a dynamic environment, subsequent views of the same object tend to occur in close temporal proximity, even if these views are drastically different from each other. Several researchers have suggested that this temporal contingency can induce time-dependent Hebbian learning between neuronal units – possibly in the anterior inferotemporal (IT) brain regions (Miyashita, 1988) – that is sensitive to the order of view-dependent shape features present in successive images of an animation sequence (Wallis and Bülthoff, 2001; Wallis, 2002). Learning these spatio-temporal associations of a dynamic object can be reinforced with repeated exposure to that object undergoing the same motion. Thus, a learned animation sequence will lead to a larger neural response than a reversed animation sequence (Wallis, 1998).

Our results are consistent with this form of temporal-associative learning. While a temporal-associative account of dynamic object learning remains plausible, it is unlikely to fully explain the contribution of learned object motion to recognition performance. For example, a purely temporal-associative account suggests that the contribution of learned motion to object recognition is automatic, regardless of whether the motion is rigid or non-rigid. However, we did not find any benefits of rigid motion for object recognition in our study.

## HIERARCHICAL MODELS FOR THE RECOGNITION OF LEARNED OBJECT MOTION

In addition to temporal-associative mechanisms, other researchers have proposed hierarchical-processing mechanisms that could provide insights into how object motion can be encoded in visual memory and contribute to object recognition in a view-invariant manner. Generally, these hierarchical models assume that visual features are progressively processed from simple features (e.g., edges) to more complex features that are conjunctions of simpler ones (Riesenhuber and Poggio, 1999; Serre et al., 2007).

Although these models were originally proposed for static features, they can be extended to include dynamic features. For example, Giese and Poggio (2003) introduced a motion pathway that operates in parallel with a form pathway. This motion pathway contributes to visual recognition by processing visual motion in a feed-forward and hierarchical fashion, employing principles similar to those proposed for the form pathway (Riesenhuber and Poggio, 1999). Giese and Poggio's model proposes that visual motion is first processed in early visual cortex (V1, V2) by direction-selective neurons. The motion signals are subsequently pooled by detectors for local optic-flow patterns such as translation and expansion in the temporal lobe (e.g., hMT+). Eventually, these relatively simple optic-flow patterns are pooled by detectors that respond selectively to complex optic-flow patterns that define the individual moments of familiar movement sequences (e.g., STS). Thus, complex static and dynamic features at the end of both pathways can, in principle, encode the unique spatio-temporal patterns of an object's learned motion.

Giese and Poggio's (2003) model was originally intended for the recognition of biological motion. Nonetheless, it should also generalize to the recognition of novel object classes with unique



spatio-temporal patterns. Indeed, our results in combination with previous studies suggest that different types of motion (rigid versus non-rigid) can lead to more accurate recognition across different viewpoint changes (see also, Watson et al., 2005; Perry et al., 2006; Vuong et al., 2009; Wallis et al., 2009). Within Giese and Poggio's model, this would suggest that recognition performance is influenced by optic-flow patterns, in the mid- and especially the later processing stages of visual motion. Speculatively, these features could capture the motion information that our participants relied upon for object recognition (Watson et al., 2005; Perry et al., 2006; Vuong et al., 2009; Wallis et al., 2009).

## CONCLUSION

The contribution of learned object motion to the recognition of dynamic objects is view-invariant. However, our results suggest that any such contributions of object motion are not automatic but may depend on the requirements of the recognition task instead. Computational models of object recognition

should consider the contribution of motion-based information, independently from image-based information about an object's shape. Future studies should also investigate the conditions that lead to a stronger reliance on certain types of information over others.

## ACKNOWLEDGMENTS

This research was supported by the Max Planck Society and the WCU (World Class University) program through the National Research Foundation of Korea funded by the Ministry of Education, Science and Technology (R31-10008). We would like to thank Profs. Ian Thornton, Roland Fleming, Christian Wallraven, and Dr. Isabelle Bühlhoff for their helpful comments.

## SUPPLEMENTARY MATERIAL

The Movies S1–S8 for this article can be found online at [http://www.frontiersin.org/Computational\\_Neuroscience/10.3389/fncom.2012.00026/abstract](http://www.frontiersin.org/Computational_Neuroscience/10.3389/fncom.2012.00026/abstract)

## REFERENCES

- Balas, B., and Sinha, P. (2009). A speed-dependent inversion effect in dynamic object matching. *J. Vis.* 9, 1–13.
- Biederman, I. (1987). Recognition-by-components: a theory of human image understanding. *Psychol. Rev.* 94, 115–147.
- Brainard, D. H. (1997). The psychophysics toolbox. *Spat. Vis.* 10, 433–436.
- Bülthoff, H. H., and Edelman, S. (1992). Psychophysical support for a two-dimensional view interpolation theory of object recognition. *Proc. Natl. Acad. Sci. U.S.A.* 89, 60–64.
- Bülthoff, I., Bülthoff, H. H., and Sinha, P. (1998). Top-down influences on stereoscopic depth-perception. *Nat. Neurosci.* 1, 254–257.
- Casile, A., and Giese, M. A. (2005). Critical features for the recognition of biological motion. *J. Vis.* 5, 348–360.
- Cutting, J. E., and Kozlowski, L. T. (1977). Recognizing friends by their walk – gait perception without familiarity cues. *Bull. Psychon. Soc.* 9, 353–356.
- Fahle, M. (1993). Figure-ground discrimination from temporal information. *Proc. R. Soc. Lond. B Biol. Sci.* 254, 199–203.
- Foster, D. H., and Gilson, S. J. (2002). Recognizing novel three-dimensional objects by summing signals from parts and views. *Proc. R. Soc. Lond. B Biol. Sci.* 269, 1939–1947.
- Giese, M. A., and Poggio, T. (2003). Neural mechanisms for the recognition of biological movements. *Nat. Rev. Neurosci.* 4, 179–192.
- Grzywacz, N. M., and Hildreth, E. C. (1987). Incremental rigidity scheme for recovering structure from motion – position-based versus velocity-based formulations. *J. Opt. Soc. Am. A* 4, 503–518.
- Hummel, J. E., and Biederman, I. (1992). Dynamic binding in a neural network for shape recognition. *Psychol. Rev.* 99, 480–517.
- Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Percept. Psychophys.* 14, 201–211.
- Knappmeyer, B., Thornton, I. M., and Bülthoff, H. H. (2003). The use of facial motion and facial form during the processing of identity. *Vision Res.* 43, 1921–1936.
- Lander, K., and Bruce, V. (2000). Recognizing famous faces: exploring the benefits of facial motion. *Ecol. Psychol.* 12, 259–272.
- Liu, T., and Cooper, L. A. (2003). Explicit and implicit memory for rotating objects. *J. Exp. Psychol. Learn. Mem. Cogn.* 29, 554–562.
- Loftus, G. R., and Masson, M. E. J. (1994). Using confidence intervals in within-subject designs. *Psychon. Bull. Rev.* 1, 476–490.
- MacMillan, N. A., and Creelman, C. D. (1991). *Detection Theory: A User's Guide*. Cambridge: Cambridge University Press.
- Miyashita, Y. (1988). Neuronal correlate of visual associative long-term memory in the primate temporal cortex. *Nature* 335, 817–820.
- Morris, S. B., and DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychol. Methods* 7, 105–125.
- Newell, F. N., Wallraven, C., and Huber, S. (2004). The role of characteristic motion in object categorization. *J. Vis.* 4, 118–129.
- Norman, J. F., Todd, J. T., and Phillips, F. (1995). The perception of surface orientation from multiple sources of optical information. *Percept. Psychophys.* 57, 629–636.
- Nygård, G. E., Looy, T. V., and Wage-mans, J. (2009). The influence of orientation jitter and motion on contour saliency and object identification. *Vision Res.* 49, 2475–2484.
- Peissig, J. J., and Tarr, M. J. (2007). Visual object recognition: do we know more now than we did 20 years ago? *Annu. Rev. Psychol.* 58, 75–96.
- Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: transforming numbers into movies. *Spat. Vis.* 10, 437–442.
- Perry, G., Rolls, E. T., and Stringer, S. M. (2006). Spatial vs temporal continuity in view invariant visual object recognition learning. *Vision Res.* 46, 3994–4006.
- Riesenhuber, M., and Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature* 2, 1019–1025.
- Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., and Poggio, T. (2007). Robust object recognition with cortex-like mechanisms. *IEEE Trans. Pattern Anal. Mach. Intell.* 29, 411–426.
- Setti, A., and Newell, F. N. (2010). The effect of body and part-based motion on the recognition of unfamiliar objects. *Vis. Cogn.* 18, 456–480.
- Spetch, M. L., Friedman, A., and Vuong, Q. C. (2006). Dynamic object recognition in pigeons and humans. *Learn. Behav.* 34, 215–228.
- Stone, J. V. (1998). Object recognition using spatiotemporal signatures. *Vision Res.* 38, 947–951.
- Stone, J. V. (1999). Object recognition: view-specificity and motion-specificity. *Vision Res.* 39, 4032–4044.
- Tarr, M. J., Williams, P., Hayward, W. G., and Gauthier, I. (1998). Three-dimensional object recognition is viewpoint dependent. *Nat. Neurosci.* 1, 275–277.
- Ullman, S. (1979). The interpretation of structure from motion. *Proc. R. Soc. Lond. B Biol. Sci.* 203, 405–426.
- Ullman, S. (2007). Object recognition and segmentation by a fragment-based hierarchy. *Trends Cogn. Sci. (Regul. Ed.)* 11, 58–64.
- Vuong, Q. C., Friedman, A., and Plante, C. (2009). Modulation of viewpoint effects in object recognition by shape and two kinds of motion cues. *Perception* 38, 1628–1648.
- Vuong, Q. C., and Tarr, M. J. (2006). Structural similarity and spatiotemporal noise effects on learning dynamic novel objects. *Perception* 35, 497–510.
- Wallis, G. (1998). Spatio-temporal influences at the neural level of object recognition. *Network* 9, 265–278.
- Wallis, G. (2002). The role of object motion in forging long-term representations of objects. *Vis. Cogn.* 9, 233–247.
- Wallis, G., Backus, B. T., Langer, M., Huebner, G., and Bülthoff, H. H. (2009). Learning illumination- and orientation-invariant representations of objects through temporal association. *J. Vis.* 9, 1–8.
- Wallis, G., and Bülthoff, H. H. (2001). Effects of temporal association on

- recognition memory. *Proc. Natl. Acad. Sci. U.S.A.* 98, 4800–4804.
- Wang, Y., and Zhang, K. (2010). Decomposing the spatiotemporal signature in dynamic 3D object recognition. *J. Vis.* 10, 1–16.
- Watson, T., Johnston, A., Hill, H. C. H., and Troje, N. F. (2005). Motion as a cue for viewpoint invariance. *Vis. Cogn.* 12, 1291–1308.
- Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.
- Received: 17 January 2012; accepted: 22 April 2012; published online: 22 May 2012.
- Citation: Chuang LL, Vuong QC and Bülthoff HH (2012) Learned non-rigid object motion is a view-invariant cue to recognizing novel objects. *Front. Comput. Neurosci.* 6:26. doi: 10.3389/fncom.2012.00026
- Copyright © 2012 Chuang, Vuong and Bülthoff. This is an open-access article distributed under the terms of the Creative Commons Attribution Non Commercial License, which permits non-commercial use, distribution, and reproduction in other forums, provided the original authors and source are credited.



# Low-level contrast statistics are diagnostic of invariance of natural textures

Iris I. A. Groen<sup>1\*</sup>, Sennay Ghebrea<sup>1,2</sup>, Victor A. F. Lamme<sup>1</sup> and H. Steven Scholte<sup>1</sup>

<sup>1</sup> Department of Psychology, Cognitive Neuroscience Group, University of Amsterdam, Amsterdam, Netherlands

<sup>2</sup> Intelligent Systems Lab Amsterdam, Institute of Informatics, University of Amsterdam, Amsterdam, Netherlands

## Edited by:

Jay Hegd , Georgia Health Sciences University, USA

## Reviewed by:

Udo Ernst, University of Bremen, Germany

Guillaume A. Rousselet, University of Glasgow, UK

Jeremy Freeman, New York University, USA

## \*Correspondence:

Iris I. A. Groen, Department of Psychology, Cognitive Neuroscience Group, University of Amsterdam, Weesperplein 4, 1018 XA, Amsterdam, Netherlands.  
e-mail: i.i.a.groen@uva.nl

Texture may provide important clues for real world object and scene perception. To be reliable, these clues should ideally be invariant to common viewing variations such as changes in illumination and orientation. In a large image database of natural materials, we found textures with low-level contrast statistics that varied substantially under viewing variations, as well as textures that remained relatively constant. This led us to ask whether textures with constant contrast statistics give rise to more invariant representations compared to other textures. To test this, we selected natural texture images with either high (HV) or low (LV) variance in contrast statistics and presented these to human observers. In two distinct behavioral categorization paradigms, participants more often judged HV textures as “different” compared to LV textures, showing that textures with constant contrast statistics are perceived as being more invariant. In a separate electroencephalogram (EEG) experiment, evoked responses to single texture images (single-image ERPs) were collected. The results show that differences in contrast statistics correlated with both early and late differences in occipital ERP amplitude between individual images. Importantly, ERP differences between images of HV textures were mainly driven by illumination angle, which was not the case for LV images: there, differences were completely driven by texture membership. These converging neural and behavioral results imply that some natural textures are surprisingly invariant to illumination changes and that low-level contrast statistics are diagnostic of the extent of this invariance.

**Keywords:** textures, image statistics, EEG, contrast, natural images, invariance, dissimilarity analysis, illumination

## INTRODUCTION

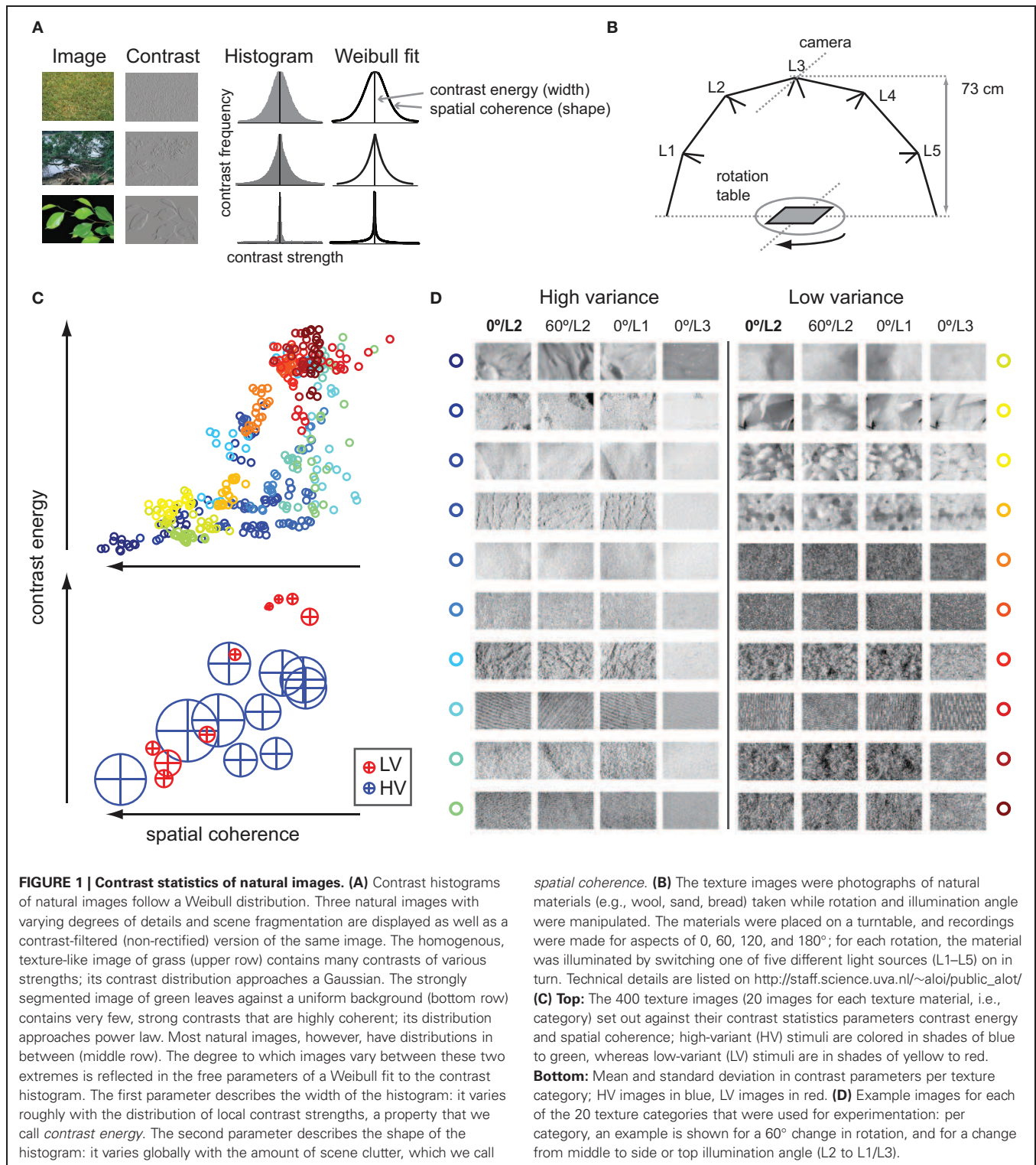
Despite the complexity and variability of everyday visual input, the human brain rapidly translates light falling onto the retina into coherent percepts. One of the relevant features to accomplish this feat is texture information (Bergen and Julesz, 1983; Malik and Perona, 1990; Elder and Velisavljević, 2009). Texture—“the stuff in the image” (Adelson and Bergen, 1991)—is a property of an image region that can be used by early visual mechanisms for initial segmentation of the visual scene into regions (Landy and Graham, 2004), to separate figure from ground (Nothdurft, 1991) or to judge 3D shape from 2D input (Malik and Rosenholtz, 1997; Li and Zaidi, 2000). The relevance of texture for perception of natural images is demonstrated by the finding that a computational model based on texture statistics accurately predicted human natural scene categorization performance (Renninger and Malik, 2004).

In general, a desirable property for any visual feature is perceptual invariance to common viewing variations such as illumination and viewing angle. Whereas invariance is often defined at the level of cognitive templates (e.g., Biederman, 1987) or as a “goal” of visual coding that needs to be achieved by multiple consecutive transformation along the visual pathway (Riesenhuber and Poggio, 1999; DiCarlo and Cox, 2007), there is another possible interpretation: invariance may also be present to a certain degree

in the natural world. Specifically, it can be hypothesized that textures that are more invariant will provide more reliable cues for object and scene perception.

The effects of viewing conditions on textures have been previously studied by Geusebroek and Smeulders (2005), who showed that changes in image recording conditions of natural materials are well characterized as changes in underlying contrast statistics. Specifically, two parameters fitted to the contrast histogram of natural images described the spatial structure of several different materials completely. These parameters express the width and outline of the histogram (Figure 1A) and carry information about perceptual characteristics of natural textures such as regularity and roughness (Geusebroek and Smeulders, 2005).

Recently, we found that for a set of natural images, the same statistics explain up to 80% of the variance of event-related potentials (ERPs) recorded from visual cortex (Ghebrea et al., 2009). We proposed that the two contrast parameters reflect relevant perceptual dimensions of natural images, namely the amount of contrast energy and spatial coherence in a scene. Importantly, we found that these parameters can be reliably approximated by linear summation of the output of localized contrast filters modeled after LGN cells (Scholte et al., 2009), suggesting that these statistics may be available to visual cortex directly from its pre-cortical contrast responses.



In the present work, we evaluated contrast statistics of a large set of natural textures that were recorded under different viewing conditions (Geusebroek and Smeulders, 2005). The contrast energy and spatial coherence of a substantial amount of textures covaried with viewing conditions. However, the statistics of some

textures remained remarkably constant under these variations. If the visual system is indeed highly sensitive to variability in low-level image statistics, differences between textures in terms of this variability should have a consequence for their perceptual processing. Specifically, textures with constant contrast statistics



across differences in viewing conditions may form more invariant representations compared to other textures.

To test this hypothesis, we asked whether perceptual invariance (experiment 1 and 2) and invariance in evoked neural responses (experiment 3) to natural textures under changes in viewing conditions was associated with variance in contrast statistics. We selected images of multiple textures that differed in two recording conditions: illumination angle and rotation (**Figure 1B**). Based on variance in contrast statistics, textures were labeled as either “HV” or “LV,” **Figure 1C**; example images of each texture category are shown in **Figure 1D**. In experiment 1, human observers performed a same-different categorization task on pairs of images that were either from the same or a different texture category. We tested whether variance in contrast statistics influenced categorization accuracy: we predicted that compared to HV textures, images from the same LV texture would appear more similar (i.e., higher accuracy of same-texture trials) and would also be less often confused with other textures (higher accuracy on different-texture trials), indicating higher “perceived invariance.” In experiment 2, we addressed the same question using another behavioral paradigm—an oddity task—in which participants selected one of three images belonging to a different texture category. We predicted that when presented with two texture images from the same HV category, participants would more often erroneously pick one of these images as the odd-one-out, indicating less “perceived invariance” on these trials. In experiment 3, event-related EEG responses (ERPs) to individually presented texture images were collected and used to examine differences in neural processing between HV and LV textures and to evaluate the contribution of each of the two image parameters (contrast energy and spatial coherence) over the course of the ERP. Specifically, we related differences in image statistics to differences in single-image responses; an avenue that more researchers are beginning to explore (Philiastides and Sajda, 2006; Scholte et al., 2009; van Rijsbergen and Schyns, 2009; Gaspar et al., 2011; Rousselet et al., 2011). The advantage of this approach relative to traditional ERP analysis (which is based on averaging many trials within a condition or an a priori-determined set of stimuli) is that it provides a richer and more detailed impression of the data and that it allows us to examine how differences between individual images can give rise to categorical differences in a bottom-up way.

The results show that variance in contrast statistics correlates with perceived texture similarity under changes in rotation and illumination, as well as differences in neural responses due to illumination changes. They suggest that low-level contrast statistics are informative about the degree of perceptual invariance of natural textures.

## MATERIALS AND METHODS

### COMPUTATION OF IMAGE STATISTICS

#### Contrast filtering

We computed image contrast information according to the standard linear-nonlinear model. For the initial linear filtering step we used contrast filters modeled after well-known receptive fields of LGN-neurons (Bonin et al., 2005). As described in detail in Ghebreab et al. (2009) each location in the image was filtered

with Gaussian second-order derivative filters spanning multiple octaves in spatial scale, following Croner and Kaplan (1995). Two separate spatial scale octave ranges were applied to derive two image parameters. For the contrast energy parameter, each image location was processed by filters with standard deviations 0.16, 0.32, 0.64, 1.28, 2.56 in degrees; for the spatial coherence parameter, the filter bank consisted of octave scales of 0.2, 0.4, 0.8, 1.6, and 3.2°. The output of each filter was normalized with a Naka-Rushton function with five semi-saturation constants between 0.15 and 1.6 to cover the spectrum from linear to non-linear contrast gain control in the LGN (Croner and Kaplan, 1995).

#### Response selection

From the population of gain- and scale-specific filters, one filter response was selected for each location in the image using minimum reliable scale selection (Elder and Zucker, 1998): a spatial scale control mechanism in which the smallest filter with output higher than what is expected to be noise for that specific filter is selected. In this approach (similar steps are implemented in standard feed-forward filtering models, e.g., Riesenhuber and Poggio, 1999) a scale-invariant contrast representation is achieved by minimizing receptive field size while simultaneously maximizing response reliability (Elder and Zucker, 1998). As previously (Ghebreab et al., 2009), noise thresholds for each filter were determined in a separate set of images (a selection of 1800 images from the Corel database) and set to half a standard deviation of the average contrast present in that dataset for a given scale and gain.

#### Approximation of Weibull statistics

Applying the selected filter at each to location to the image results in a contrast magnitude map. Based on the different octave filter banks, one contrast magnitude map was derived for the contrast energy parameter and one for the spatial coherence parameter. These contrast maps were then converted into two 256-bin histograms. It has been demonstrated that contrast distributions of most natural images adhere to a Weibull distribution (Geusebroek and Smeulders, 2002). The Weibull function is given by:

$$p(r) = ce^{\left(\frac{r-\mu}{\beta}\right)^{\gamma}} \quad (1)$$

where  $c$  is a normalization constant and  $\mu$ ,  $\beta$ , and  $\gamma$  are the free parameters that represent the origin, scale and shape of the distribution, respectively. The value of the origin parameter  $\mu$  is generally close to zero for natural images. The *contrast energy* parameter ( $\beta$ ) varies with the range of contrast strengths present in the image. The *spatial coherence* parameter ( $\gamma$ ) describes the outline of the distribution and varies with the degree of correlation between local contrast values.

As mentioned, these two parameters can also be approximated in a more biologically plausible way: we demonstrated that simple summation of X- and Y-type LGN output corresponded strikingly well with the fitted Weibull parameters (Scholte et al., 2009). Similarly, if the outputs of the multi-scale, octave filter banks (Ghebreab et al., 2009) used here—reflecting the entire range of receptive field sizes of the LGN—are linearly summed, we obtain values that correlate even stronger with the Weibull parameters

obtained from the contrast histogram at minimal reliable scale (Ghebreab et al., under review). In the present stimulus set, the approximation based on summation of the two filter banks correlated  $r = 0.99$  and  $r = 0.95$  with respectively the beta and gamma parameter of a Weibull function fitted to the contrast histogram. For all analyses presented here, these biologically realistic approximations based on linear summation were used instead of the fitted parameters.

## EXPERIMENT 1: BEHAVIORAL CATEGORIZATION WITH A SAME-DIFFERENT TASK

### Subjects

In total, 28 subjects participated in the first behavioral categorization experiment. The experiment was approved by the ethical committee of the University of Amsterdam and all participants gave written informed consent prior to participation. They were rewarded for participation with either study credits or financial compensation (7€ for one hour of experimentation). The data from two participants was excluded because mean behavioral performance was at chance level (50%).

### Stimuli

Texture images were selected from a large database of natural materials (throughout the document, we will refer to these as “texture categories,” <http://staff.science.uva.nl/~aloi/publicatlot/>) that were photographed under various systematic manipulations (illumination angle, rotation, viewing angle, and illumination color). For the subset used in the present study, images (grayscale,  $512 \times 342$  pixels) of each texture category varied only in illumination angle (five different light sources) and rotation (0, 60, 120, or  $180^\circ$ ), while viewing angle ( $0^\circ$  azimuth) and illumination color (white balanced) were held constant. This selection yielded 20 unique images per texture category. For all 250 categories in the database, contrast statistics were computed for this subset of images. Based on the resulting contrast energy and spatial coherence parameters, textures were designated as either HV or LV if the variance in both parameter values was more than 0.5 standard deviation above (HV) or below (LV) the median variance for all textures. From those two selections, 10 texture categories were randomly chosen; however, care was taken that the mean parameter values of the selected categories were representative of the range of the entire database. The final selection thus yielded 20 texture categories, 10 of which formed the “HV condition” and 10 that formed the “LV condition,” with each category consisting of 20 images that were systematically manipulated in illumination angle and rotation. Thus, in total, 400 images were used for experimentation.

### Procedure

On each trial, two images were presented which were from the same or a different texture category. Stimuli were presented on a 19 inch Dell monitor with a resolution of  $1280 \times 1024$  pixels and a frame rate of 60 Hz. Participants were seated approximately 90 cm from the monitor and completed four blocks of 380 trials each. A block contained four breaks, after which subject could continue the task by means of a button press. On each trial, a fixation cross appeared on the center of the screen; after

an interval of 500 ms, a pair of stimuli was presented simultaneously for 50 ms, separated by a gap of 86 pixels (**Figure 2A**). A mask (see below) followed after 100 ms, and stayed on screen for 200 ms. Subjects were instructed to indicate if the stimuli were from the same or a different texture category by pressing one of two designated buttons on a keyboard (“z” and “m”) that were mapped to the left or the right hand. Within one block, one stimulus from one texture category was once paired with a stimulus from another texture category (190 trials). Stimuli were drawn without replacement, such that each image occurred once in each block, but were randomly paired with the images from the other texture category on each block. For the other 190 trials, the two stimuli were from the same texture category: for each texture category, 10 pairs were randomly chosen, resulting in 200 trials (20 from each texture category), from which 10 were then randomly removed (but never more than one from each category) such that 190 trials remained. The ratio of different-category vs. same-category comparisons was thus 1, which was explicitly communicated to the subjects prior to the test phase. Subjects were shown a few example textures, which contained examples of both illumination and rotation changes, and they also performed 20 practice trials before starting the actual experiment (none of these examples occurred in the experiment; the practice trials contained comparisons of both illumination and rotation changes between the two presented texture images). Masks were created by dividing each of the 400 texture stimuli up in mini-blocks of  $9 \times 16$  pixels: a mask was created by drawing equal amounts of these mini-blocks from each stimulus and placing those at random positions in a frame of  $512 \times 342$  pixels. Unique masks were randomly assigned to each of the 400 trials within a block, and were repeated over blocks. Per trial, the same mask was presented at both stimulus locations. Stimuli were presented using Matlab Psychophysics Toolbox (Brainard, 1997; Pelli, 1997).

### Data analysis

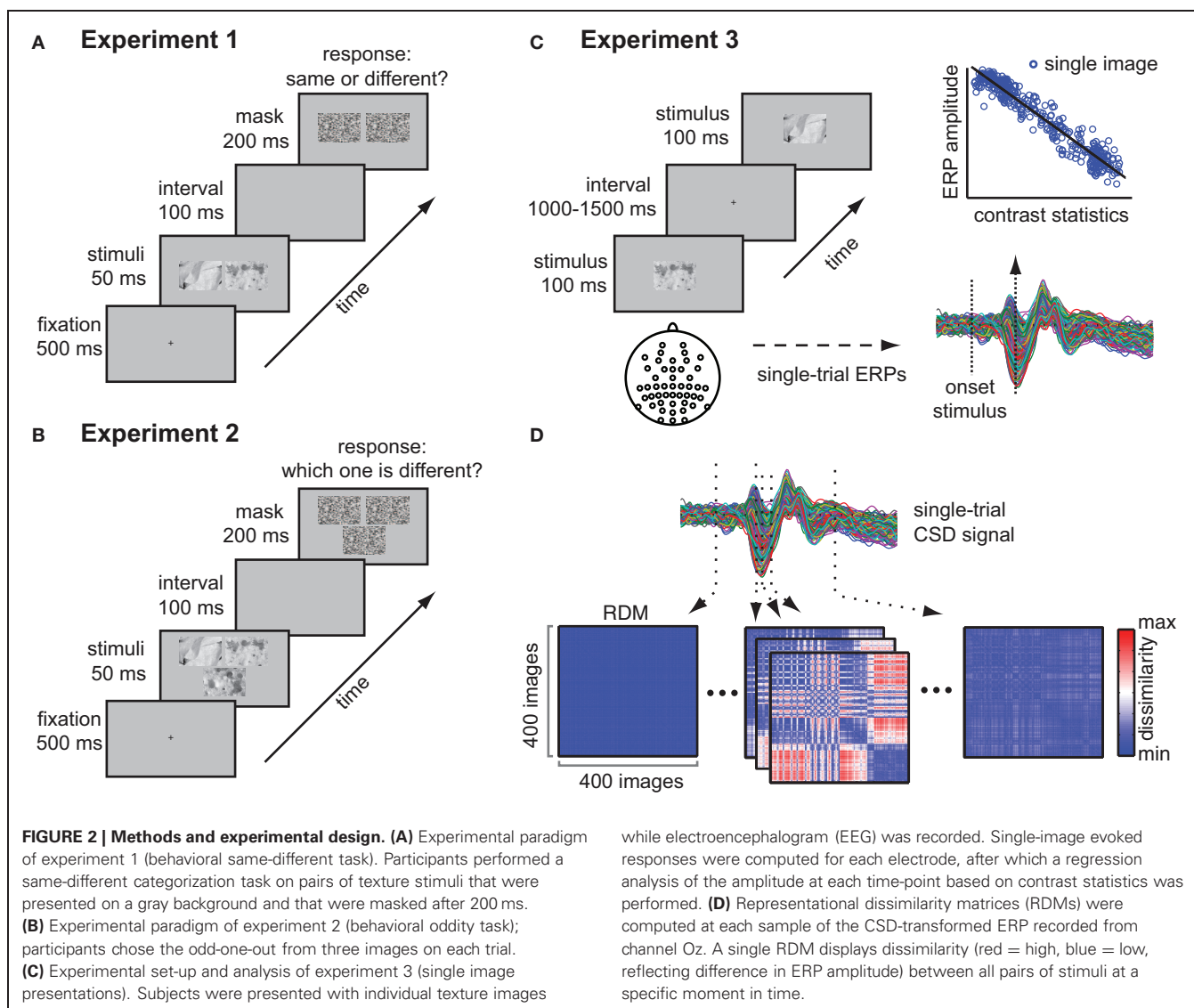
Mean accuracy for each subject was determined by calculating percent correct over four repeated blocks and was done separately for different-category vs. same-category comparisons and for trials on which two HV categories were compared vs. trials on which two LV categories were compared ( $2 \times 2$  design). Different-category trials on which HV categories were compared to LV categories were excluded from analysis. Only the responses, and not the reaction times (RTs) were recorded: as a consequence a number of trials in which subjects may have responded too fast (for instance before 200 ms) were included in the analysis. This results in a potential underestimation of the error rate.

## EXPERIMENT 2: BEHAVIORAL CATEGORIZATION WITH AN ODDITY TASK

### Subjects

In total, 18 subjects participated in the second behavioral categorization experiment, which was approved by the ethical committee of the University of Amsterdam. All participants gave written informed consent prior to participation. They were rewarded for participation with either study





credits or financial compensation (7€/hour), as previously. The data from two participants was excluded because mean performance was at chance level (33%, one participant) or because RTs demonstrated an outlier ( $>2$  standard deviations away from the mean across all participants, one participant).

### Stimuli and procedure

The same set of 400 texture images was used as in the first behavioral experiment. However, for this task, on each trial three images were presented: two images from the same texture category (the “same pair”), and one from a different category (the “odd-one-out”), see **Figure 2B**. Stimuli were presented on a 19-inch ASUS monitor with a resolution of  $1920 \times 1080$  pixels and a frame rate of 60 Hz. The procedure was identical to the first behavioral experiment, i.e., the stimuli were presented simultaneously: in this case, two images were positioned adjacent to each other and the third image was located either below

or above the other two (this was counterbalanced across participants). The three images were separated by equal gaps of 120 pixels. The position of the odd stimulus was randomized over trials. Subjects were instructed to indicate which image was from the different texture category by pressing one of three designated buttons on a keyboard (“1”, “2”, and “3” on the NUM pad of the keyboard) using their right hand. Within one block, each texture category was twice paired with every other texture category by randomly drawing a stimulus from both categories (380 trials). For one half of the trials, the image from the first category was designated as the “odd-one-out,” whereas from the second category, another stimulus was drawn to form the second half of the “same pair.” For the other half of the trials, the procedure was reversed, such that each texture category once formed the odd stimulus, and once formed the paired stimulus. Compared to the first experiment, the trials were thus always “different-category trials,” but on a given trial, each texture category could be the odd stimulus or the same-pair stimulus,

allowing us to test whether variance of the same-pair texture images influenced performance: we predicted that increased variance in contrast statistics of the same-pair stimuli would lead to more errors (i.e., selecting one of the same-pair as the odd-one-out).

### Data analysis

As in the first experiment, except that trials in which the participant responded before 200 ms after stimulus-onset were excluded. To allow comparison with the same-different accuracy data from the previous experiment, we first selected only the trials on which either two HV or two LV texture categories were compared (ignoring trials on which one HV and one LV category were compared). The same comparison was done for RTs. In a subsequent analysis, we did include all trials but split them into two groups in two different ways: namely (1) based on whether the odd stimulus was LV or HV or (2) based on whether the same-pair were HV or LV. This allowed us to test whether the variance of the odd stimulus vs. the variance of the same-pair was associated with increased error rates in selection of the odd stimulus.

## EXPERIMENT 3: EEG EXPERIMENT

### Subjects

Seventeen volunteers participated and were rewarded with study credits or financial compensation (7€/hour for 2,5 h of experimentation). The data from two subjects was excluded because the participant blinked consistently shortly after trial onset in more than 50% of the trials (one subject) or because their vision deviated from normal (one subject) which became clear in another experiment conducted in the same session. This study was approved by the ethical committee of the University of Amsterdam and all participants gave written informed consent prior to participation.

### EEG data acquisition

The same set of stimuli was used as in the behavioral experiment. In addition, for each image a phase-scrambled version was created, which were presented randomly intermixed with the actual textures, with equal proportions of the two types of images. Stimuli were presented on an ASUS LCD-screen with a resolution of  $1024 \times 768$  pixels and a frame rate of 60 Hz. Subjects were seated 90 cm from the monitor such that stimuli subtended  $11 \times 7.5^\circ$  of visual angle. During EEG acquisition, a stimulus was presented one at a time in the center of the screen on a gray background for 100 ms, on average every 1500 ms (range 1000–2000 ms; **Figure 2C**). Each stimulus was presented twice, in two separate runs. Subjects were instructed to indicate on each trial whether the image was an actual texture or a phase-scrambled image: a few examples of the two types of images were displayed prior to the experiment. Response mappings were counterbalanced between the two separate runs for each subject. Stimuli were presented using the Presentation software ([www.neurobs.com](http://www.neurobs.com)). EEG Recordings were made with a Biosemi 64-channel Active Two EEG system (Biosemi Instrumentation BV, Amsterdam, NL, <http://www.biosemi.com/>) using the standard 10–10 systems with additional occipital electrodes (I1 and I2), which replaced two

frontal electrodes (F5 and F6). Eye movements were monitored with a horizontal and vertical electro-oculogram (EOG) and were aligned with the pupil location when the participants looked straight ahead. Data was sampled at 256 Hz. The Biosemi hardware is completely DC-coupled, so no high-pass filter is applied during recording of the raw data. A Bessel low-pass filter was applied starting at 1/5th of the sample rate.

### EEG data preprocessing

The raw data was pre-processed using Brain Vision Analyzer (BVA) by taking the following steps: (1) offline referencing to earlobe electrodes, (2) applying a high-pass filter at 0.1 Hz (12 dB/octave), a low-pass filter at 30 Hz (24 dB/octave); because low-pass filters in BVA have a graded descent, additionally two notch filters at 50 (for line noise) and 60 Hz (for monitor noise) were applied, (3) automatic removal of deflections larger than  $250 \mu\text{V}$ . Trials were segmented into epochs starting 100 ms before stimulus onset and ending 500 ms after stimulus onset. These epochs were corrected for eye movements by removing the influence of ocular-generated EEG using a regression analysis based on the EOG channels (Gratton et al., 1983). Baseline correction was performed based on the data between  $-100$  and  $0$  ms relative to stimulus onset; artifacts were rejected using maximal allowed voltage steps of  $50 \mu\text{V}$ , minimal and maximal allowed amplitudes of  $-75$  and  $75 \mu\text{V}$  and a lowest allowed activity of  $0.50 \mu\text{V}$ . The resulting ERPs were converted to Current Source Density (CSD) responses (Perrin, 1989). This conversion results in a signal that is more localized in space, which has the advantage of more reliably reflecting activity of neural tissue underlying the recording electrode (Nunez and Srinivasan, 2006). Trials in which the same individual image was presented were averaged over the two runs, resulting in an image-specific ERP (single-image ERP).

### Regression analyses on single-image ERPs

To test whether differences between neural responses correlated with differences in contrast statistics between images, we conducted regression analyses on the single-image ERPs (**Figure 2C**). We first performed this analysis on ERPs averaged across subjects to test whether contrast energy and spatial coherence could explain consistent differences between images. For each channel and time-point, the image parameters (contrast energy and spatial coherence) were entered together as linear regressors on ERP amplitude, resulting in a measure of model fit ( $r^2$ ) over time (each sample of the ERP) and space (each electrode). To statistically evaluate the specific contribution of each parameter to the explained variance for the two different image conditions (HV en LV), we ran regressions at the single subject level (these analyses were restricted to electrode Oz). For this, we constructed a model with four predictors of interest (constant term + LV contrast energy, HV contrast energy, LV spatial coherence, HV spatial coherence). The obtained  $\beta$ -coefficients for each predictor were subsequently tested against zero by means of  $t$ -tests, which were Bonferroni-corrected for multiple comparisons based on the number of time-points for which the comparison was performed (154 samples). Finally, to test whether each predictor

contributed unique variance, we conducted a stepwise version of the two-parameter (contrast energy and spatial coherence for both LV and HV images) regression analysis for each single subject. In this analysis, a predictor was entered to the model if it was significant at  $\alpha < 0.05$ , and was removed again if  $\alpha > 0.10$ ; as an initial model, none of the parameters were included. We then counted, at every time-point, for how many subjects the full model was chosen, or only one of the predictors was included.

### Representational similarity analysis

To better examine how variance between individual visual stimuli arises over time, and how differences between individual images relate to image variance (HV/LV) and image manipulations (rotation/illumination), we computed representational dissimilarity matrices (RDMs; Kriegeskorte et al., 2008) based on single-image ERPs recorded at channel Oz. We computed, for each subject separately, at each time-point, for all pairs of images the difference between their evoked ERP amplitude (Figure 2D). As a result we obtained a single RDM containing  $400 \times 400$  “dissimilarity” values between all pairs of images at each time-point. Within one such matrix, the pixel value of each cell reflects the difference in ERP amplitude of the corresponding two images indicated by the row- and column number.

### Comparison between dissimilarity matrices

To compare the dissimilarities between evoked ERPs by individual images with corresponding differences in image statistics between those images, we computed a pair-wise dissimilarity matrix based on both image parameter values combined. For each pair of images, we computed the sum of the absolute differences between the (normalized) contrast energy (CE) and spatial coherence (SC) values of those two images [e.g.,  $(CE_{\text{image1}} + SC_{\text{image1}}) - (CE_{\text{image2}} + SC_{\text{image2}})$ , etc.], resulting in one difference value reflecting the combined difference in image parameters between the two images. For each subject, this matrix was compared with the RDMs based on the ERP data using a Mantel test for two-dimensional correlations (Daniels, 1944).

### Computation of luminance and AIC-values

To obtain a simple description of luminance for each image, we computed the mean luminance value per image (LUM) by averaging the pixel values (0–255) of each individual image. For the EEG analysis, to compare the regression results based on LUM with those obtained with contrast statistics, we used Akaike’s information criterion (AIC; Akaike, 1973). The AIC-values were computed by transforming the residual sum of squares (RSSs) of each regression analysis using

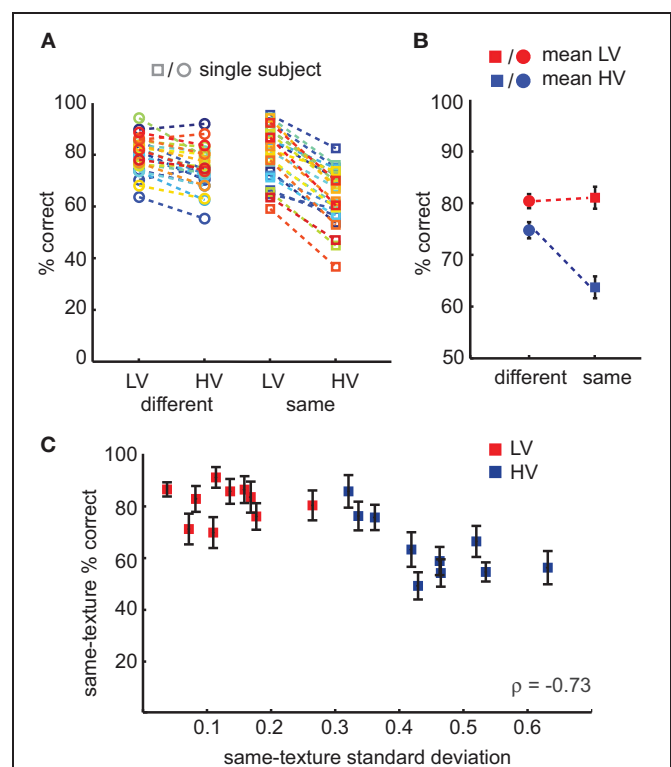
$$\text{AIC} = n \cdot \log(\text{RSS} / n) + 2k \quad (2)$$

where  $n$  = number of images and  $k$  is the number of predictor variables ( $k = 2$  for contrast statistics, and  $k = 1$  for LUM). AIC can be used for model selection given a set of candidate models of the same data; the preferred model has minimum AIC-value.

## RESULTS

### EXPERIMENT 1: SAME-DIFFERENT CATEGORIZATION

Categorization accuracy was determined separately for HV and LV trials and for same-category and different-category comparisons (Figure 3A). A repeated-measures, two-way ANOVA indicated a significant main effect of variance [ $F_{(1, 25)} = 298.9, p < 0.0001$ ], but not of type of comparison [ $F_{(1, 25)} = 3.6, p = 0.07$ ]; however, there was a significant interaction between variance and comparison [ $F_{(1, 25)} = 61.8, p < 0.0001$ ; Figure 3B]. Subsequent paired  $t$ -tests revealed that participants performed better for LV than HV textures at both different-category [ $t_{(25)} = 6.1, p < 0.0001$ , mean difference = 6%,  $ci = 4\text{--}8\%$ ] and same-category comparisons [ $t_{(25)} = 16.3$ , mean difference = 17%,  $ci = 15\text{--}19\%$ ,  $p < 0.0001$ ], but also for different-category HV comparisons relative to same-category HV comparisons [ $t_{(25)} = 3.4$ , mean difference = 11%,  $ci = 4\text{--}17\%$ ,  $p = 0.002$ ]. These results show that participants generally made more errors on trials in which they compared two different HV texture categories; in addition, they more often incorrectly judged two



**FIGURE 3 | Results of the behavioral same-different experiment.**

(A) Accuracy scores for individual subjects according to task conditions: subjects compared a pair of images that were either from different (circles) or same (squares) texture categories, which could either be low-varient or high-varient. Trials in which HV images were compared with LV images were excluded from the analysis. (B) Mean accuracy per condition, demonstrating an interaction effect between texture variance (HV, blue vs. LV, red) and type of comparison (same vs. different trial). (C) Accuracy on same-texture trials correlates with category specific variance in contrast statistics. Error bars indicate s.e.m.

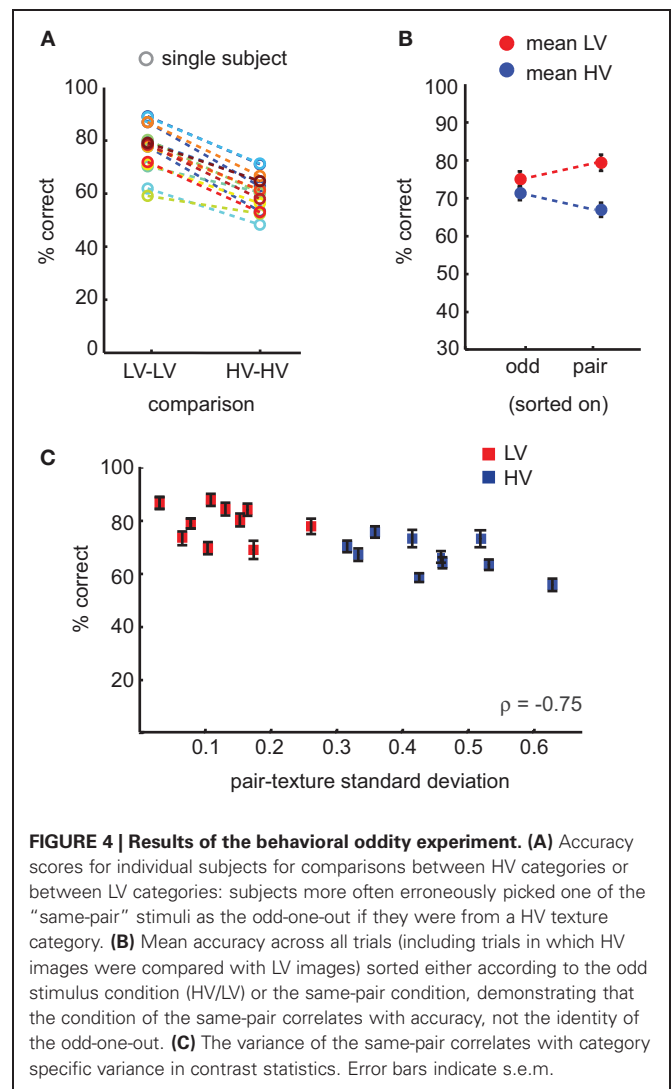
images from the same HV texture category as different than vice versa (two different HV images as the same category). This finding suggests that LV texture categories are easier to categorize than HV texture categories and that images from the same HV texture category are perceived as less similar. This latter conclusion is supported by an additional analysis performed on the accuracy scores, in which we correlated the specific amount of variance in contrast statistics with the average number of same-texture errors. We found that variance in contrast statistics correlated with same-texture accuracy across all texture categories (Spearman's  $\rho = -0.73$ ,  $p < 0.0001$ , **Figure 3C**). This result suggests that the specific amount of variance in contrast statistics influences perceived similarity of same-texture images: more variance implies less similarity.

As subjects always compared only two images on each trial, we cannot be certain to what degree they based their judgment on the between-stimulus differences vs. the difference of these two images compared to all other images in the stimulus set. To investigate this more explicitly, we conducted another behavioral experiment using an oddity task, in which each trial consisted of three images that were drawn from two different texture categories. In this task, subjects always make a difference judgment: they have to pick the most distinct stimulus (the “odd-one-out”) and thus actively compare differences *between* texture categories with differences *within* texture categories. If variance in contrast statistics of a texture category indeed determines its perceived invariance, we would expect that for comparisons between images with high variance, it is more difficult to accurately decide which stimulus is different.

## EXPERIMENT 2: ODDITY CATEGORIZATION

Categorization accuracy on comparisons of HV texture categories was significantly lower compared to comparisons of LV texture categories [ $t_{(15)} = 14.4$ , mean difference = 17%,  $ci = 14\text{--}20\%$ ,  $p < 0.0001$ ]; **Figure 4A**. Participants were also significantly faster on LV trials compared to HV trials [ $t_{(15)} = -3.5$ , mean difference = 27 ms,  $ci = 10\text{--}43$  ms,  $p < 0.004$ ]. If we compute accuracy across all possible comparisons of texture categories (also including HV-LV comparisons), and split the data either according to the variance of the odd stimulus, or to the variance of the same-pair stimulus on each trial, we see that specifically the variance of the same-pair images is correlated with differences in accuracy (**Figure 4B**): on trials at which the same-pair was from a HV texture category, subjects more often erroneously chose one of that pair as the odd-one-out. As in the previous experiment, we correlated the amount of variance in contrast statistics of the same-pair with accuracy, and we again find a significant correlation ( $\rho = -0.75$ ,  $p < 0.0001$ ; **Figure 4C**), indicating that with increasing variance in contrast statistics, images from the same texture category are more often perceived as different.

Overall, the results of the two behavioral experiments indicate that low variance in contrast statistics allows observers to more accurately categorize images of natural textures. Importantly, images of a texture category with constant statistics under different viewing conditions are more accurately recognized as the same category compared to images from categories with variable



**FIGURE 4 | Results of the behavioral oddity experiment. (A)** Accuracy scores for individual subjects for comparisons between HV categories or between LV categories: subjects more often erroneously picked one of the “same-pair” stimuli as the odd-one-out if they were from a HV texture category. **(B)** Mean accuracy across all trials (including trials in which HV images were compared with LV images) sorted either according to the odd stimulus condition (HV/LV) or the same-pair condition, demonstrating that the condition of the same-pair correlates with accuracy, not the identity of the odd-one-out. **(C)** The variance of the same-pair correlates with category specific variance in contrast statistics. Error bars indicate s.e.m.

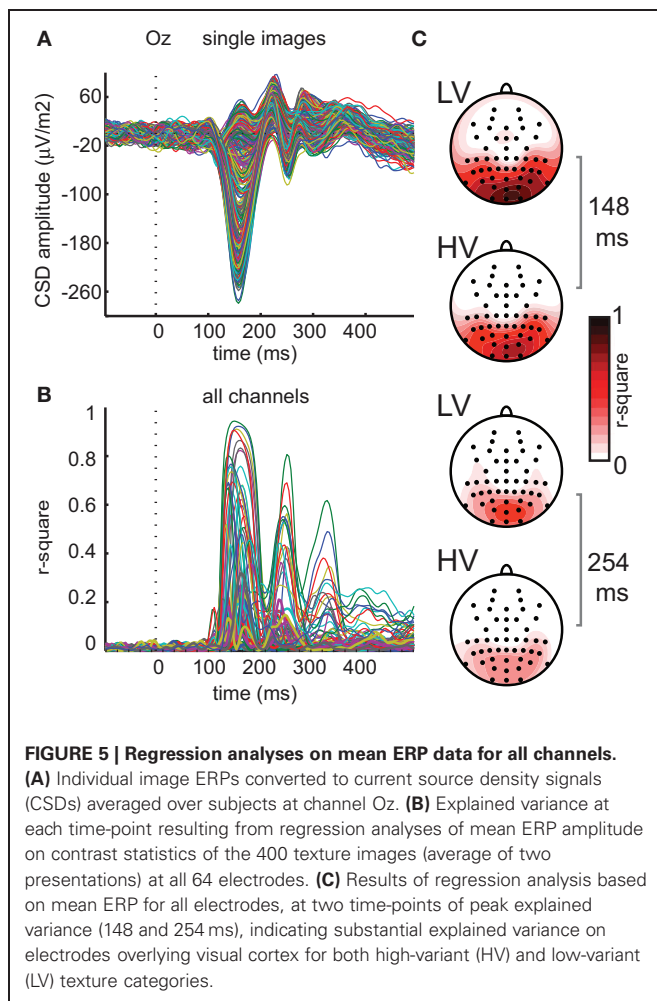
statistics, suggesting that textures categories with little variance in contrast statistics are perceived as more invariant.

## EXPERIMENT 3: EEG

### Contrast statistics explain variance in occipital ERP signals

As a first-pass analysis, we first averaged single-image ERPs over subjects, after which a simple regression model with two predictors (contrast energy and spatial coherence) was fitted based on these “subject-averaged” ERPs at every channel and time-point. Despite individual differences between subjects in EEG responses (e.g., in mean evoked response amplitude, likely due to individual differences in cortical folding), this analysis revealed a highly reliable ERP waveform time-locked to the presentation of the stimulus (**Figure 5A**). This time-locked ERP nonetheless varied substantially between individual images, mostly between 100 and 300 ms after stimulus-onset. The results show that early in time, nearly all ERP variance is explained by the image parameters (maximal  $r^2 = 0.94$  at 148 ms,  $p < 0.0001$  on channel Oz, **Figure 5B**). Also at later time-points and at other electrodes, there is substantial (e.g., more than 50%) explained variance. If we





examine the results for all channels simultaneously (**Figure 5C**), we see that explained variance is highest at occipital channels, and subsequently wears off toward more parietal and lateral electrodes. This localization is similar for both early and late time-points (i.e., mostly central-occipital).

This result shows that low-level image statistics can explain a high amount of variance, both early and late in time, of image-specific differences across participants. To test more precisely (1) whether these effects were present in all participants, (2) which of the two image parameters contributed most to the explained variance, and (3) whether these contributions differed between the two conditions (LV/HV), we selected the electrode with the highest  $r^2$ -value (Oz) and conducted regression analyses at the single-subject level using a model containing four parameters (see Materials and Methods): LV contrast energy, HV contrast energy, LV spatial coherence, HV spatial coherence. The results showed that contrast statistics explained a substantial amount of variance between individual images in each participant. Mean explained variance across subjects peaked 156 ms after stimulus onset ( $r^2 = 0.65$ , mean  $p < 0.0001$ , Bonferroni-corrected; **Figure 6A**); peak values for individual subjects ranged between  $r^2 = 0.49$ – $0.85$  at

144–168 ms after stimulus-onset and were all highly significant (all  $p < 0.0001$ ).

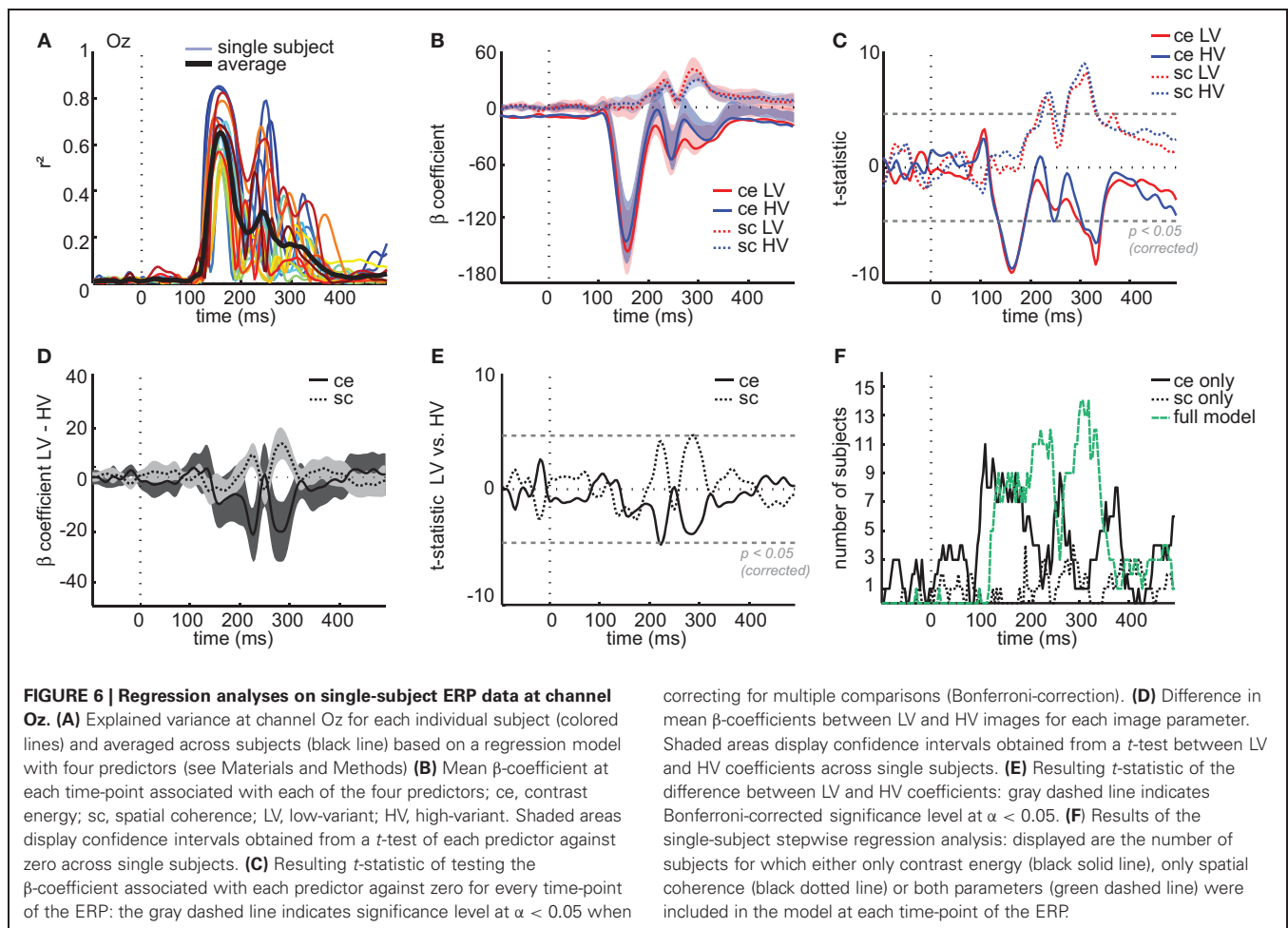
If we compare the time courses of the  $\beta$ -coefficients associated with each predictor (**Figure 6B**), we observe that contrast energy and spatial coherence have distinct time courses. Statistical comparisons of each coefficient against zero across participants (**Figure 6C**) show that ERP amplitude at an early time interval is mostly correlated with contrast energy [between 136 and 183 ms, all  $t_{(15)} < -5.1$ , max  $t_{(15)} = -9.0$ , all  $p < 0.0003$ ], which correlates again much later in time [between 305 and 340 ms, all  $t_{(15)} < -5.1$ , max  $t_{(15)} = -6.5$ , all  $p < 0.0001$ ]. Spatial coherence only contributes significantly to the explained variance between 220 and 240 ms, [all  $t_{(15)} > 4.7$ , max  $t_{(15)} = 6.1$ , all  $p < 0.003$ ; again between 274 and 330 ms, all  $t_{(15)} > 5.4$ , max  $t_{(15)} = 9.0$ , all  $p < 0.0003$ ]. Importantly, at most time-points the temporal profile of each predictor is comparatively similar for HV and LV images; differences between the beta coefficients of these two conditions are relatively small (**Figure 6D**). For both image parameters, the difference between HV and LV images appears to be substantial only at two time-intervals between 150 and 300 ms, but statistical tests of these differences were right at the threshold of Bonferroni-corrected significance [contrast energy at 223 ms,  $t_{(15)} = -4.8$ ,  $p = 0.0002$ ; spatial coherence, at 285 and 289 ms,  $t_{(15)} = 4.6$ ,  $p = 0.0003$ ; **Figure 6E**]. Given the small effects and the borderline significance, this issue cannot be resolved with the current dataset.

Finally, to test whether the two image parameters explain unique variance, we conducted stepwise regression analyses (see Materials and Methods) based on evoked responses of single subjects on ERPs recorded at channel Oz. At each time-point of the ERP, we counted for how many participants (a) either the full model was chosen or (b) only one predictor was included in the model (**Figure 6F**). The results show that early in time, the contrast energy parameter alone is preferred over the full model, but that later in time (from  $\sim 200$  ms onwards), for most subjects the spatial coherence parameter is also included. This suggests that especially later in time, spatial coherence adds additional explanatory power to the regression model.

These results show that across subjects, differences in early ERP amplitude between individual images correlate with variance in contrast statistics of those images for both HV and LV textures. Whereas contrast energy explains most variance early in time, both parameters become significantly correlated with ERP amplitude at later time intervals. These regression results do not reveal, however, whether these differences are related to texture category (categorical differences), or if they occur as a result of variations in recording conditions. We investigated this in the next section.

### Dissimilarities between images map onto contrast statistics

To examine the origin of the variance between individual trials, we computed (for each subject separately) RDMs based on differences in evoked responses between individual images (see Materials and Methods and **Figure 2D**). In brief, to build an RDM, we compute for each possible combination of individual images the difference in evoked ERP amplitude, and convert the result into a color value. The advantage of this approach is that RDMs allow us to see at once how images are (dis)similar



to all other images, and how this relates to texture category “membership.”

To demonstrate the result of this analysis, we selected the RDM at the time-point of maximal explained variance for the subject-averaged regression analysis (148 ms after stimulus-onset) for each subject and simply averaged the resulting matrix over subjects. In this RDM (Figure 7A), every consecutive 20 rows/columns index all images from one specific texture category; these categories are sorted according to their mean contrast energy and spatial coherence values (i.e., distance from zero in the contrast statistics space in Figure 1C). If we visually examine the RDM, we observe that differences between HV images (lower right quadrant) occur at different positions than for LV images (upper left quadrant). Specifically, for HV stimuli, there are larger differences *within* textures, whereas for LV stimuli, the differences are largest *between* textures: i.e., within a  $20 \times 20$  “square,” images are “similarly dissimilar” from other textures. This result suggests that LV images cluster more by texture category than HV images, which are highly different even within a given texture.

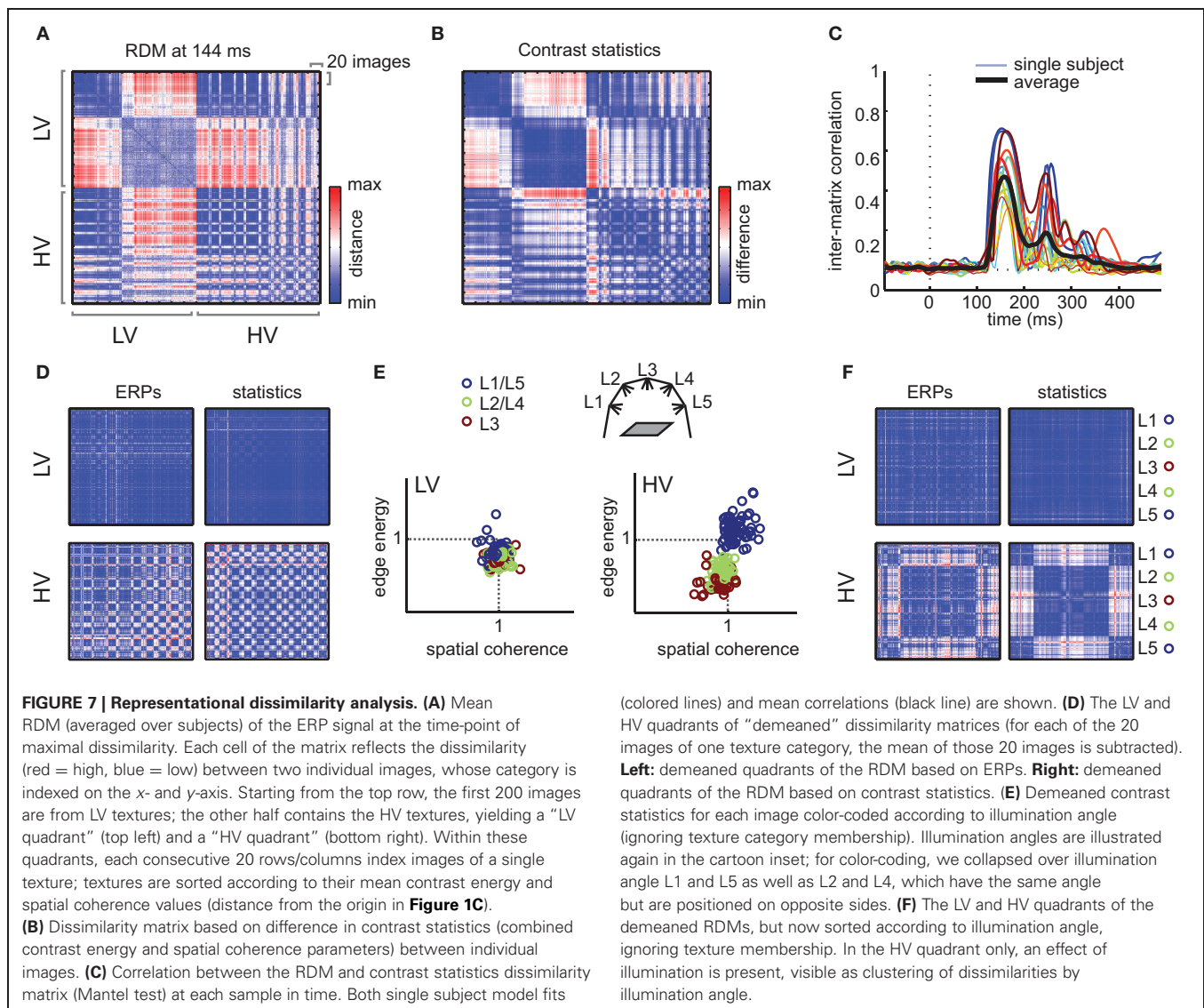
Next, we tested to what extent these image-specific differences in ERP amplitude were similar to differences in contrast statistics. We calculated another  $400 \times 400$  difference matrix, in

which we simply subtracted the parameter values of each image from the values of each other image (Figure 7B, see Materials and Methods). Based on visual inspection, it is clear that the relative dissimilarities between individual images in contrast statistics are very similar to the ERP differences. A test of the inter-matrix correlation at each time-point (Figure 7C) indicated that the RDM of the ERP signal correlated significantly with the difference matrix based on contrast statistics; between 137 and 227 ms after stimulus onset, the correlation was significant for all 17 subjects (range peak  $r = 0.31$ – $0.72$ , all  $p < 0.01$ , Bonferroni-corrected).

#### Dissimilarities between HV stimuli reflect illumination changes

Presumably, the higher dissimilarities *within* HV textures result from variability in responses driven by changes in recording conditions. To isolate these effects, we computed “demeaned” versions of the RDMs, by dividing the evoked response to each image by the mean response to all 20 images of its texture category, before computing the differences between individual images. As a result, we obtain RDMs that only reflect differences in variance from the mean response to that texture category, ignoring differences between the means of different categories. Analogously, for the contrast statistics matrix, we divided the parameter values between images of a given texture by the mean contrast energy





and spatial coherence value of that texture and subsequently computed the image-specific differences.

As one would expect, dissimilarities between images in LV stimuli have completely disappeared in these demeaned RDMs (**Figure 7D**, displaying only the HV-HC and LV-LV quadrants). This demonstrates that all differences between LV stimuli indeed reflect differences *between* texture categories. For HV stimuli however, dissimilarities *within* texture categories remain after demeaning; moreover, we observe a “plaid-like” pattern in the RDM, which suggests that dissimilarities of individual HV images do not fluctuate randomly, but are present in a regular manner. What manipulation is driving these dissimilarities? If we investigate the clustering of images based on demeaned contrast statistics (**Figure 7E**), we see that for HV stimuli, the variance from the mean is caused by changes in illumination direction: the illumination change “moves” the stimulus to another location the contrast statistics space in a consistent manner. As a final demonstration, we resorted all images in the RDMs based on illumination direction instead of texture category: in the resulting

RDM, the differences between ERPs now cluster with illumination changes (**Figure 7F**), confirming that dissimilarities within HV categories result mostly from illumination differences.

These results again show that differences between individual images in ERP responses are correlated with differences in contrast statistics for those images. Importantly, they reveal that differences between HV textures occur for other reasons than differences between LV textures. For HV textures, we observe that manipulations of illumination angle are reflected in the RDM: instead of clustering by category (which would be evidenced by within-texture *similarity* and between-texture *dissimilarity*), images are selectively dissimilar for one illumination angle compared to another. For LV stimuli, the pattern of results is different: stimuli do cluster by category, meaning that all images of a given texture are “similarly dissimilar” from other textures (or similar, if the mean of the other images is very nearby in “contrast statistics space,” **Figure 1C**). Overall, this suggests that the amount of variance in contrast statistics correlates with variance between neural responses resulting from variations in recording

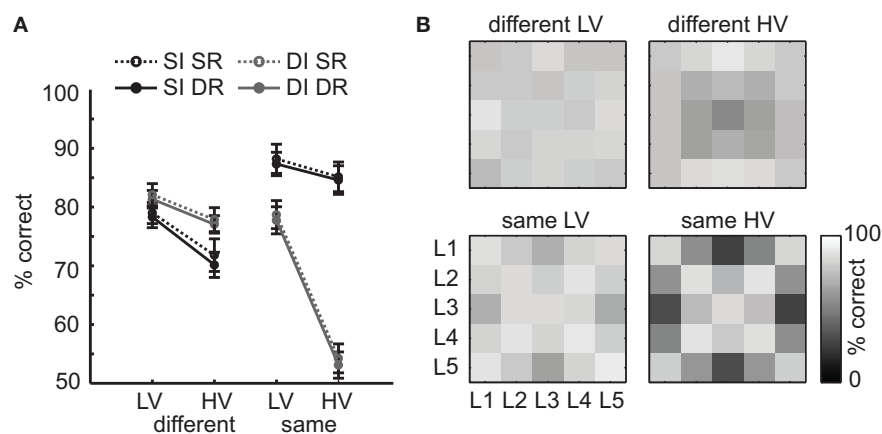
conditions, specifically illumination differences. Textures that vary little in contrast statistics appear to form a more “invariant” representation in terms of evoked responses.

### Image manipulations: rotation versus illumination

The results of the EEG experiment suggest that the high variance of HV texture images is related to a higher sensitivity of these textures to changes in illumination direction: the large differences that remain *within* texture categories after subtracting differences *between* texture categories appear to be driven by differences in illumination angle. Based on this finding, we can expect that the main result we observed in behavioral categorization (increased error rates on HV texture categories) is also driven by effects of illumination direction, rather than image rotations. To address this question, we *post-hoc* sorted the data from the same-different experiment based on whether the two presented images differed in (a) rotation only, (b) illumination only, or (c) both rotation and illumination, and separately computed the accuracies for each of the different conditions (same LV, same HV, different LV, different HV). Because the pairing of individual images was randomized over trials (see Materials and Methods), there were unequal amounts of manipulation differences for each subject and condition. To increase the number of trials per condition and to be able to compare across conditions, we collapsed over same angles from different sides, as in the EEG RDM analysis (see **Figure 7E**, i.e., we counted a pair of images of the same texture category that were illuminated from angle 2/4 or 1/5 as “same illumination”). As a result, we obtained four different “trial-types”: same illumination, same rotation (SI, SR), same illumination, different rotation (SI, DR), different illumination, same rotation (DI, SR), or different illumination, different rotation (DI, DR). The results show that across

all trial-types, accuracy is lower for HV than for LV stimuli [**Figure 8A**; main effect of variance on both same-category and different-category comparisons, all  $F_{(1, 25)} > 27.9$ ,  $p < 0.0001$ ]. However, as predicted, on same-category trials most errors are made when illumination is changed compared to when rotation is changed and illumination is kept constant [main effect of illumination,  $F_{(1, 25)} > 262.9$ ,  $p < 0.0001$ ]. Importantly, this effect is much larger for the HV texture categories than for the LV texture categories [significant interaction between variance and illumination,  $F_{(1, 25)} = 162.1$ ,  $p < 0.0001$ ]. This analysis thus shows that the influence of illumination angles differs for HV vs. LV texture categories: it again demonstrates that the categories from the latter condition are more invariant to these manipulations than other categories. Interestingly, the effect is reversed for different pairs [more errors for *same* illumination trials; main effect of illumination,  $F_{(1, 25)} = 20.6$ ,  $p < 0.0001$ ] suggesting that in this case, illumination changes “help” to distinguish different texture categories more easily. Most importantly, these results support the conclusion that the extent to which a given texture category is sensitive to illumination changes can be derived from contrast statistics.

To demonstrate the effects of illumination changes more clearly, we sorted the DI accuracies based on the exact illumination angle (L1–L5, see **Figures 1B** and **7E**) that was used for each of the two presented stimuli on a given trial (given the small effect of rotation, we now collapsed over same rotation and different rotation trial-types, i.e., over SR and DR trials). The results of this analysis are displayed as confusion matrixes in **Figure 8B** (diagonals represent the SI trials). Here, it can be observed that on same-category HV trials (lower right matrix), most errors are made when the change in illumination angle was large (e.g., a pairing of L3 and L1/L5). For same-category LV trials (lower



**FIGURE 8 | Post-hoc analysis of the effect of image manipulations on accuracy of the same-different categorization experiment.**

**(A)** Accuracies for each condition (HV/LV) and type of comparison (same/different) were computed separately for trials in which the two images were either photographed under same illumination and same rotation (SI, SR), same illumination and different rotation (SI, DR), different illumination and same rotation (DI, SR), or different illumination and different rotation (DI, DR). The results show that the participants always made more errors on HV texture categories than LV texture categories: however, this

effect is strongest for same-comparison HV trials where there was a difference in illumination angle between images. Error bars indicate s.e.m. **(B)** Effect of illumination angle (L1–L5, see **Figure 1B**) on same-illumination (SI, diagonal values) and different-illumination (DI, off-diagonal values) trials (collapsed over SR and DR trials), for each condition and type of comparison. Most errors on same-category HV trials are made for images that have the largest difference in illumination angle (i.e., L3 vs. L1/L5), whereas most errors on different-category HV trials are made for illumination angle L3.

left matrix), however, this effect is much weaker, indicating that LV texture categories are less responsive to illumination changes. Interestingly, from the confusion matrix of the different category HV trials (upper right panel), it appears that most errors were made when the two different texture images were photographed under angle L3, suggesting that in general, images from different texture categories become more similar when the light is shone right from above; likely, this is due to higher saturation of the image (overexposure) under this illumination angle. This effect is however again absent for the different-category LV trials (upper left panel), suggesting that these saturation effects are less likely to occur for images that are LV in contrast statistics.

### Luminance statistics

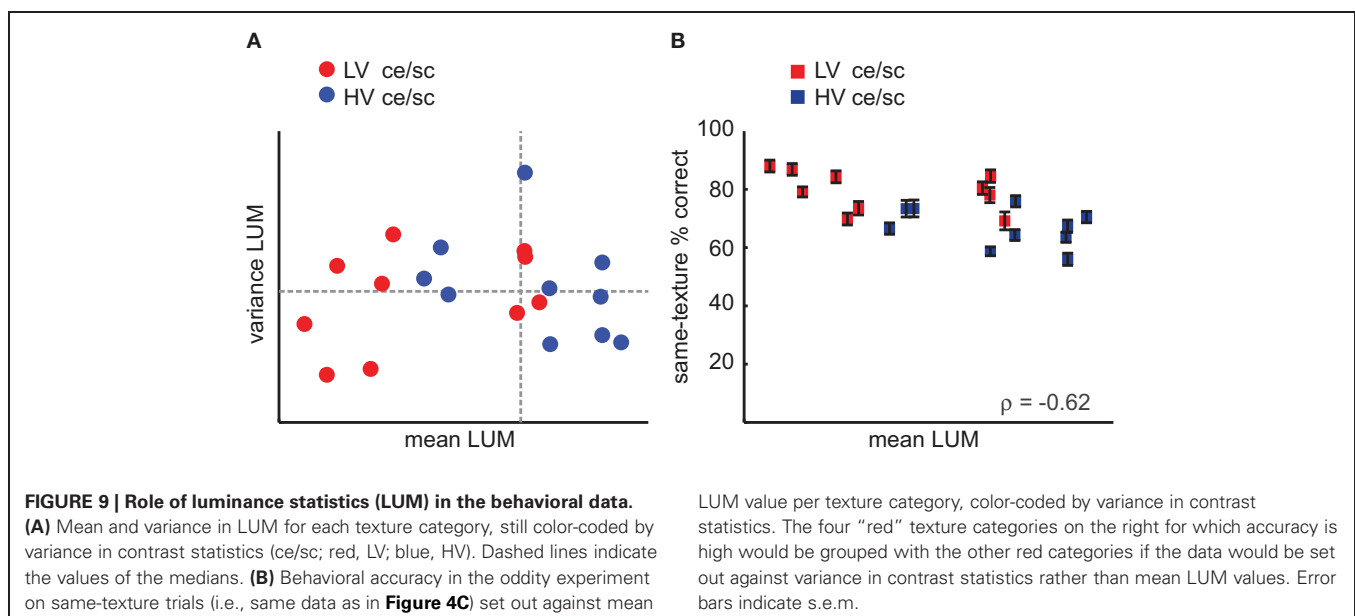
These behavioral and EEG results suggest that contrast statistics of the texture categories are diagnostic of perceived variance under changes in illumination. The behavioral results further suggest that the *amount* of illumination change is directly related to the perceived similarity on same-texture HV trials, and that the specific illumination angle used on different-texture HV trials may also influence the observed similarity (Figure 8B). Does this mean that a simple description of differences in luminance between images (i.e., brightness), rather than contrast statistics, would describe the same pattern of results? To test this, we computed the mean luminance (LUM, see Materials and Methods) of each image and tested to what extent differences in luminance were correlated with differences in behavioral categorization and EEG responses.

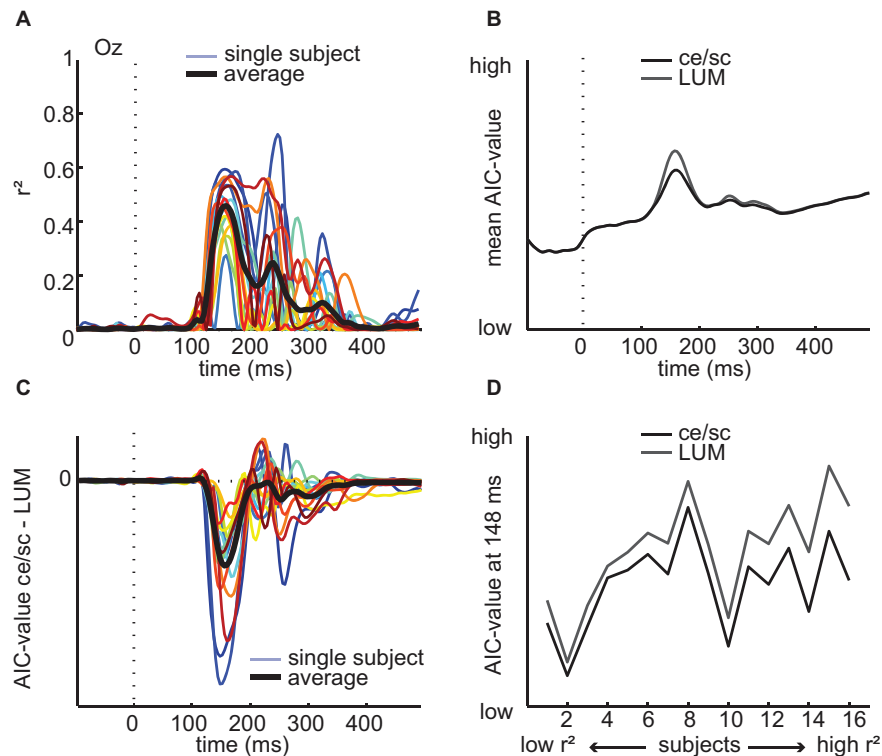
First of all, LUM values of individual images were indeed highly correlated with contrast energy ( $\rho = -0.69$ ,  $p < 0.001$ ), and somewhat lower but significantly correlated with spatial coherence ( $\rho = -0.38$ ,  $p < 0.001$ ). However, if we split the texture categories into high and low variance conditions based on LUM values per category, we do not find the same texture categories in each condition as in the original division based on

contrast statistics; see Figure 9A. In fact, about half of the LV categories are “HV” in LUM if we separate the categories using a median split. The correlation of variance in LUM and accuracy on same-texture trials was either not significant ( $\rho = -0.32$ ,  $p = 0.15$ , for the oddity experiment) or significant but lower compared to the correlation with contrast statistics ( $\rho = -0.51$ ,  $p = 0.02$  for the same-different experiment), suggesting that variance in brightness rather than contrast is not an alternative explanation for the finding that observers perceive images from HV texture categories more often as “different.”

A majority of LV texture categories have low mean LUM values, but this is not the case for all categories, suggesting that HV texture categories are not systematically brighter than LV categories. The correlations of same-texture accuracy and mean LUM per texture category are also inconsistent: significant in the oddity experiment ( $\rho = -0.62$ ,  $p < 0.005$ ), but not in the same-different experiment ( $\rho = -0.35$ ,  $p = 0.12$ ). As can be seen in Figure 9B, the correlation with behavioral accuracy can be explained by the partial overlap of the LV/HV categories and the low/high LUM values.

In the EEG data, differences in LUM explained less variance (peak mean  $r^2$  across subjects = 0.44, between 0.27 and 0.72 for individual subjects, Figure 10A) than differences in contrast statistics. To compare the model fits directly, we used Akaike’s information criterion (Akaike, 1973) to compute AIC-values (see Materials and Methods) based on the residuals of the regression analyses. In this analysis, a lower AIC-value indicates a better fit to the data, or “more information,” whereby models with more parameters are penalized. The mean AIC-values obtained at each time-point of the ERP are shown in Figure 10B, where it can be seen that in the early time interval where contrast statistics correlate with the ERP (~140–180 ms), the regression model based on LUM has a higher AIC-value and thus worse predictive power than the regression model based on contrast statistics (note that this is despite the fact that the contrast





**FIGURE 10 | Regression analyses based on LUM values on single-subject ERP data and comparison with contrast statistics. (A)** Explained variance at channel Oz for each individual subject (colored lines) and averaged across subjects (black line) from a regression model with LUM values **(B)** AIC-values of the regression results based on LUM compared to contrast statistics (ce/sc); low AIC-value indicates better model fit. **(C)** Single-subject differences in AIC-value between LUM and contrast statistics over time.

**(D)** Subject-specific differences in AIC-values at the time-point where the difference between the two models in mean AIC-values is largest (148 ms), sorted based on their maximal  $r^2$  values of the contrast statistics model. For subjects with higher  $r^2$  values, the difference in AIC-values becomes somewhat larger, suggesting that high explained variance on contrast statistics is not coupled with increased fit of both models simultaneously, but rather with a better fit of contrast statistics compared to LUM values.

statistics contain two parameters and LUM only one, which benefits the latter's AIC-value). For most subjects, the AIC-value based on contrast statistics is consistently lower than the AIC-value from regression on LUM, also later in time (**Figure 10C**). Between the two models, individual subjects' peak  $r^2$  values and corresponding time-points were highly correlated ( $\rho = 0.87$ ,  $p < 0.0001$  and  $\rho = 0.67$ ,  $p < 0.005$ , respectively), suggesting that subjects with high explained variance for contrast statistics also had high explained variance for LUM and that these peaks occurred around the same time. Interestingly however, the difference in AIC-value is largest for subjects with high maximal  $r^2$ -values (**Figure 10D**), suggesting that increased explained variance is associated with a larger difference in goodness of fit or "information" between the alternative models (LUM vs. contrast statistics).

These analyses suggest that despite the high correlations between contrast statistics and simple luminance values, contrast statistics provide a better predictor of perceived invariance as well as differences in evoked activity for this set of texture images. This is not unexpected: from physiology, it is known that neurons in LGN effectively band-pass filter contrast values from the visual input (De Valois and De Valois, 1990). Indeed, repeated band-pass filtering of visual information seems a fundamental

property of visual cortex, resulting in increasingly invariant representations (Bouvier et al., 2009). From this perspective, contrast information is itself more invariant than luminance. Our results suggest that this hierarchical increase in invariance, obtained by filtering, is not equal for all types of textures: after contrast filtering, each image becomes more invariant in information content, but some texture images become more invariant than others, possibly forming a more reliable building block for further processing.

## DISCUSSION

In a large database of natural textures, we selected images with low-level contrast statistics that were either constant or variable under changes in illumination angle and orientation. In both EEG and behavior, we showed that textures with little variation in low-level contrast statistics were perceived as more invariant (experiment 1 and 2) and led to more invariant representations at the neural level (experiment 3). The higher the variance in contrast statistics within a given texture, the higher the probability of subjects judging two images of that texture as different categories, specifically if the images differ in illumination direction. Accordingly, high-variant textures give rise to neural evoked responses that are clearly modulated by illumination



direction, which is not the case for low-variant textures, as predicted.

Interestingly, as indicated by higher accuracy on same-texture comparisons in the behavioral experiment, textures with low variance in contrast statistics remained more perceptually similar under different illumination (and rotation) conditions. This was explained by the finding that for LV textures, we observed “clustering by texture” of dissimilarities in single-image ERPs between images, whereas there was clustering by illumination direction for HV textures. These results suggest that distance *between* different textures in terms of contrast statistics—which an observer may use to estimate whether two stimuli are from the same or from a different texture category—are more reliable for LV textures than for HV textures. This is not surprising if one examines the clustering of LV vs. HV stimuli in “contrast statistics space” (Figure 1C): as a natural consequence of the lower variance *within* LV textures, the differences *between* texture categories become more similar for images of LV textures.

This work extends recent findings that statistical variations in low-level information are important for understanding generalization over single images (Karklin and Lewicki, 2009). In addition, it has been demonstrated that behavioral categorization accuracy can be predicted using a computational model of visual processing: a neural network consisting of local filters that were first allowed to adapt to the statistics of the natural environment could accurately predict behavioral performance on an object categorization task (Serre et al., 2007). Compared to the latter study, however, in our case there was no training or tuning of a network on a separate set of stimuli such that statistical regularities were implicitly encoded: here, perceived texture similarity was inferred directly from explicitly modeled contrast statistics.

In addition to behavioral categorization, we were able to test the contribution of our two contrast parameters to evoked neural responses using EEG. It is well known that early ERP components can be modulated by low-level properties of (simple) visual input (Luck, 2005). Our finding that contrast energy of single-image responses to natural stimuli is correlated with ERP amplitude around 140–180 ms is also consistent with previous reports of an early time-frame where stimulus-related differences drive evoked responses, e.g., between face stimuli (Philiastides and Sajda, 2006; van Rijsbergen and Schyns, 2009). These authors used classification techniques on single-trial ERPs to show that at later time intervals, differences between individual images correspond to either a more refined representation of the information relevant for the task (van Rijsbergen and Schyns, 2009) or the actual decision made by the subject (Philiastides and Sajda, 2006), suggesting that over the course of the ERP, the visual representation is transformed “away” from simple low-level properties to information that is task-relevant. In this light, it is remarkable that our second image parameter, spatial coherence, is specifically correlated with late ERP activity—around 200 and 300 ms—and that it explains additional variance compared to contrast energy alone specifically in this time interval.

One possible explanation of this apparent discrepancy is that the spatial coherence parameter is itself correlated with more refined or relevant features of natural images: essentially constituting a “summary statistic” of visual input that can be used

for rapid decision-making (Oliva and Torralba, 2006). Another interesting hypothesis is that this low-level image parameter is predictive of the availability of diagnostic information, reflecting higher “quality in stimulus information” (Gaspar et al., 2011) or less noise in the stimulus (Bankó et al., 2011; Rousselet et al., 2011), which may influence the accumulation of information for decision-making (Philiastides et al., 2006). Since our two stimulus conditions (HV/LV) were defined based on variance in both contrast energy and spatial coherence, we cannot test which of the two parameters is more strongly correlated with behavioral accuracy. Also, our work is substantially different from these previous reports in that our experiments did not require formation of a high-level representation (e.g., recognition of a face/car), but merely a same-different judgment, essentially constituting a low-level task.

Another difference between our results and those reported in the face processing literature (see e.g., Rousselet et al., 2008) is the localization of our effects. Maximal sensitivity of evoked activity to faces and objects is found at lateral-occipital and parietal electrodes (PO), whereas our correlations, obtained with texture images, are clustered around occipital electrode Oz. This is not unexpected since textural information is thought to be processed in early visual areas such as V2 (Kastner et al., 2000; Scholte et al., 2008; Freeman and Simoncelli, 2011).

In this paper, we specifically aimed to test whether invariance, in addition to a “goal” of visual encoding, could be defined as a property of real-world visual features (in this case, textures). In the first scenario, one would expect the representation of the visual input to change over time to (gradually) become more invariant. Our behavioral results however indicate that variance in low-level properties of natural textures (contrast statistics, presumably derived from very early visual information) can already predict the perceived invariance by human observers under specific viewing manipulations. Moreover, it demonstrates that there are interesting differences between natural textures in terms of this invariance: some textures appear to be surprisingly invariant. It has been argued that, in evolution, mechanisms have evolved for detecting “stable features” in visual input because they are important for object recognition (Chen et al., 2003). In light of the present results, a biologically realistic instantiation of such a stable feature could be “a texture patch whose contrast statistics do not change under viewing variations.” This natural invariance is rooted in physical properties of natural images, but is present at the level of image statistics (stochastic invariance). Such invariance may play an important role in stochastic approaches to computer vision, such as the successful bag-of-words approach (Feifei et al., 2007; Jégou et al., 2011). For example, a patch of a visual scene with more invariant contrast statistics may provide a more reliable “word” for categorization in a bag of words model for scene recognition (Gavves et al., 2011).

Our results suggest that these stochastic invariances are not only reflected in occipital ERPs recorded at the scalp, but that the human visual system may actively exploit them: in the present data, LV textures did not only give rise to more reliable differences between texture categories in evoked responses, but were also associated with more reliable judgments about similarity between different textures—i.e., with behavioral outcome.

The link between contrast statistics and categorization accuracy leads to the interesting hypothesis that in more naturalistic tasks such as object detection or natural scene processing, image elements that are stochastically invariant, i.e., reliable, may weigh more heavily in perceptual decision-making than variable, “unreliable” elements.

In sum, the present results show that low-level contrast statistics correlate with variance of natural texture images in terms of evoked responses, as well as perceived perceptual similarity; they suggest that textures with little variance in contrast statistics may give rise to more invariant neural representations. Simply put, invariance in simple, physical contrast information may lead to a more invariant perceptual representation. This makes us wonder about visual invariance as a general real-world property: how much of it can be derived from image statistics? Are there other

low-level visual features that differ in their degree of invariance? Next to studying top-down, cognitive invariance, or transformations performed by the visual system to achieve invariance of visual input, exploring to what extent “natural invariances” exist and whether they play a role in visual processing may provide an exciting new avenue in the study of natural scene perception.

## ACKNOWLEDGMENTS

We thank Marissa Bainathsah, Vincent Barneveld, Timmo Heijmans, and Kelly Wols for behavioral data collection of the same-different experiment. This work is part of the Research Priority Program “Brain and Cognition” at the University of Amsterdam and was supported by an Advanced Investigator grant from the European Research Council.

## REFERENCES

- Adelson, E. H., and Bergen, J. R. (1991). “The plenoptic function and the elements of early vision,” in *Computational Models of Visual Processing*, eds M. S. Landy and J. A. Movshon (Cambridge, MA: MIT Press), 3–20.
- Akaike, H. (1973). “Information theory and an extension of the maximum likelihood principle,” in *Second International Symposium on Information Theory*, eds B. N. Petrov and F. Csaki (Budapest: Akademiai Kiado), 267–281.
- Bankó, E. M., Gál, V., Körtvélyes, J., Kovács, G., and Vidnyánszky, Z. (2011). Dissociating the effect of noise on sensory processing and overall decision difficulty. *J. Neurosci.* 31, 2663–2674.
- Bergen, J. R., and Julesz, B. (1983). Parallel versus serial processing in rapid pattern discrimination. *Nature* 303, 696–698.
- Biederman, I. (1987). Recognition-by-components: a theory of human image understanding. *Psychol. Rev.* 94, 115–147.
- Bonin, V., Mante, V., and Carandini, M. (2005). The suppressive field of neurons in lateral geniculate nucleus. *J. Neurosci.* 25, 10844–10856.
- Bouvier, J., Rosasco, L., and Poggio, T. (2009). On invariance in hierarchical models. *Adv. Neural Inf. Process. Syst.* 22, 162–170.
- Brainard, D. H. (1997). The psychophysics toolbox. *Spat. Vis.* 10, 433–436.
- Chen, L., Zhang, S., and Srinivasan, M. V. (2003). Global perception in small brains: topological pattern recognition in honey bees. *Proc. Natl. Acad. Sci. U.S.A.* 100, 6884–6889.
- Croner, L. J., and Kaplan, E. (1995). Receptive fields of P and M ganglion cells across the primate retina. *Vision Res.* 35, 7–24.
- Daniels, H. E. (1944). The relation between measures of correlation in the universe of sample permutations. *Biometrika* 33, 129–135.
- De Valois, R. L., and De Valois, K. K. (1990). *Spatial Vision*. New York, NY: Oxford University Press.
- DiCarlo, J. J., and Cox, D. D. (2007). Untangling invariant object recognition. *Trends Cogn. Sci.* 11, 333–341.
- Elder, J. H., and Velisavljević, L. (2009). Cue dynamics underlying rapid detection of animals in natural scenes. *J. Vis.* 9, 1–20.
- Elder, J. H., and Zucker, S. W. (1998). Local scale control for edge detection and blur estimation. *IEEE Trans. Pattern Anal. Mach. Intell.* 20, 699–716.
- Feifei, L., Fergus, R., and Perona, P. (2007). Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories. *Comput. Vis. Image Underst.* 106, 59–70.
- Freeman, J., and Simoncelli, E. P. (2011). Metamers of the ventral stream. *Nat. Neurosci.* 14, 1195–1200.
- Gaspar, C. M., Rousset, G. A., and Pernet, C. R. (2011). Reliability of ERP and single-trial analyses. *Neuroimage* 58, 620–629.
- Gavves, E., Snoek, C. G. M., and Smeulders, A. W. M. (2011). Visual synonyms for landmark image retrieval. *Comput. Vis. Image Underst.* 116, 238–249.
- Geusebroek, J., and Smeulders, A. W. M. (2002). “A physical explanation for natural image statistics,” in *Proceedings of the 2nd International Workshop on Texture Analysis and Synthesis*, ed M. Chantler (Edinburgh, Scotland: Heriot-Watt University), 47–52.
- Geusebroek, J., and Smeulders, A. W. M. (2005). A six-stimulus theory for stochastic texture. *Int. J. Comp. Vis.* 62, 7–16.
- Ghebreab, S., Smeulders, A. W. M., Scholte, H. S., and Lamme, V. A. F. (2009). A biologically plausible model for rapid natural image identification. *Adv. Neural Inf. Process. Syst.* 22, 629–637.
- Gratton, G., Coles, M. G. H., and Donchin, E. (1983). A new method for off-line removal of ocular artifact. *Electroencephalogr. Clin. Neurophysiol.* 55, 468–484.
- Jégou, H., Perronnin, F., Douze, M., Jorge, S., Patrick, P., and Schmid, C. (2011). Aggregating local image descriptors into compact codes. *IEEE Trans. Pattern Anal. Mach. Intell.* 99. doi: 10.1109/TPAMI.2011.235
- Karklin, Y., and Lewicki, M. S. (2009). Emergence of complex cell properties by learning to generalize in natural scenes. *Nature* 457, 83–86.
- Kastner, S., Weerd, P. D., and Ungerleider, L. G. (2000). Texture segregation in the human visual cortex: a functional MRI study. *J. Neurophysiol.* 2453–2457.
- Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., Tanaka, K., and Bandettini, P. A. (2008). Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron* 60, 1126–1141.
- Landy, M. S., and Graham, N. (2004). “Visual perception of texture,” in *The Visual Neurosciences*, eds L. M. Chalupa and J. S. Werner (Cambridge, MA: MIT Press), 1106–1118.
- Li, A., and Zaidi, Q. (2000). Perception of three-dimensional shape from texture is based on patterns of oriented energy. *Vision Res.* 40, 217–242.
- Luck, S. J. (2005). *An Introduction to the Event-Related Potential Technique*. Cambridge, MA: MIT Press.
- Malik, J., and Perona, P. (1990). Preattentive texture discrimination with early vision mechanisms. *J. Opt. Soc. Am. A* 7, 923–932.
- Malik, J., and Rosenholtz, R. (1997). Computing local surface orientation and shape from texture for curved surfaces. *Int. J. Comp. Vis.* 23, 149–168.
- Nothdurft, H. C. (1991). Texture segmentation and pop-out from orientation contrast. *Vision Res.* 31, 1073–1078.
- Nunez, P. L., and Srinivasan, R. (2006). *The Neurophysics of EEG*, 2nd Edn. Oxford, UK: Oxford University Press.
- Oliva, A., and Torralba, A. (2006). Building the gist of a scene: the role of global image features in recognition. *Prog. Brain Res.* 155, 23–36.
- Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: transforming numbers into movies. *Spat. Vis.* 437–442.
- Perrin, F. (1989). Spherical splines for scalp potential and current density mapping. *Electroencephalogr. Clin. Neurophysiol.* 72, 184–187.
- Philastides, M. G., Ratcliff, R., and Sajda, P. (2006). Neural representation of task difficulty and decision making during perceptual categorization: a timing diagram. *J. Neurosci.* 26, 8965–8975.
- Philastides, M. G., and Sajda, P. (2006). Temporal characterization of the neural correlates of perceptual decision making in the human brain. *Cereb. Cortex* 16, 509–518.



- Renninger, L. W., and Malik, J. (2004). When is scene identification just texture recognition? *Vision Res.* 44, 2301–2311.
- Riesenhuber, M., and Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nat. Neurosci.* 2, 1019–1025.
- Rousselet, G. A., Gaspar, C. M., Kacper, P., and Pernet, C. R. (2011). Modeling single-trial ERP reveals modulation of bottom-up face visual processing by top-down task constraints (in some subjects). *Front. Psychol.* 2:137. doi: 10.3389/fpsyg.2011.00137
- Rousselet, G. A., Husk, J. S., Bennett, P. J., and Sekuler, A. B. (2008). Time course and robustness of ERP object and face differences. *J. Vis.* 8, 1–18.
- Scholte, H. S., Ghebreab, S., Waldorp, L., Smeulders, A. W. M., and Lamme, V. A. F. (2009). Brain responses strongly correlate with Weibull image statistics when processing natural images. *J. Vis.* 9, 1–15.
- Scholte, H. S., Jolij, J., Fahrenfort, J. J., and Lamme, V. A. (2008). Feedforward and recurrent processing in scene segmentation: electroencephalography and functional magnetic resonance imaging. *J. Cogn. Neurosci.* 20, 2097–2109.
- Serre, T., Oliva, A., and Poggio, T. (2007). A feedforward architecture accounts for rapid categorization. *Proc. Natl. Acad. Sci. U.S.A.* 104, 6424–6429.
- van Rijsbergen, N. J., and Schyns, P. G. (2009). Dynamics of trimming the content of face representations for categorization in the brain. *PLoS Comp. Biol.* 5:e1000561. doi: 10.1371/journal.pcbi.1000561
- Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 15 December 2011; accepted: 23 May 2012; published online: 08 June 2012.

Citation: Groen IIA, Ghebreab S, Lamme VAF and Scholte HS (2012) Low-level contrast statistics are diagnostic of invariance of natural textures. *Front. Comput. Neurosci.* 6:34. doi: 10.3389/fncom.2012.00034

Copyright © 2012 Groen, Ghebreab, Lamme and Scholte. This is an open-access article distributed under the terms of the Creative Commons Attribution Non Commercial License, which permits non-commercial use, distribution, and reproduction in other forums, provided the original authors and source are credited.



# Invariant object recognition based on extended fragments

Evgeniy Bart<sup>1\*</sup> and Jay Hegdé<sup>2</sup>

<sup>1</sup> Palo Alto Research Center, Intelligent Systems Laboratory, Palo Alto, CA, USA

<sup>2</sup> Department of Ophthalmology, Vision Discovery Institute and Brain and Behavior Discovery Institute, Georgia Health Sciences University, Augusta, GA, USA

## Edited by:

Klaus R. Pawelzik, Universität Bremen, Germany

## Reviewed by:

Germán Mato, Centro Atómico Bariloche, Argentina

Christian Leibold, Ludwig Maximilians University, Germany

## \*Correspondence:

Evgeniy Bart, Palo Alto Research Center, 3333 Coyote Hill Road, Palo Alto, CA 94304, USA.  
e-mail: bart@parc.com

Visual appearance of natural objects is profoundly affected by viewing conditions such as viewpoint and illumination. Human subjects can nevertheless compensate well for variations in these viewing conditions. The strategies that the visual system uses to accomplish this are largely unclear. Previous computational studies have suggested that in principle, certain types of object fragments (rather than whole objects) can be used for invariant recognition. However, whether the human visual system is actually capable of using this strategy remains unknown. Here, we show that human observers can achieve illumination invariance by using object fragments that carry the relevant information. To determine this, we have used novel, but naturalistic, 3-D visual objects called “digital embryos.” Using novel instances of whole embryos, not fragments, we trained subjects to recognize individual embryos across illuminations. We then tested the illumination-invariant object recognition performance of subjects using fragments. We found that the performance was strongly correlated with the mutual information (MI) of the fragments, provided that MI value took variations in illumination into consideration. This correlation was not attributable to any systematic differences in task difficulty between different fragments. These results reveal two important principles of invariant object recognition. First, the subjects can achieve invariance at least in part by compensating for the changes in the appearance of small local features, rather than of whole objects. Second, the subjects do not always rely on generic or pre-existing invariance of features (i.e., features whose appearance remains largely unchanged by variations in illumination), and are capable of using learning to compensate for appearance changes when necessary. These psychophysical results closely fit the predictions of earlier computational studies of fragment-based invariant object recognition.

**Keywords:** form vision, illumination constancy, informative fragments, invariant recognition, mutual information

## INTRODUCTION

We rarely encounter a given object under the same viewing conditions twice: the viewpoint, illumination, retinal size, and background all tend to differ from one encounter to the next. Yet, we have little difficulty in recognizing an object for what it is while ignoring the irrelevant image variations. How the visual system accomplishes this invariant recognition of objects (also referred to as perceptual constancy) has remained largely unclear (for reviews, see Walsh and Kulikowski, 1998; Wallis and Bulthoff, 1999; Christou and Bulthoff, 2000; Rolls, 2008; Biederman and Cooper, 2009). This is both because the underlying computational problems are profoundly difficult, and because experimental and computational studies have so far largely focused on understanding object recognition without these variations.

Previous studies have shown that the visual system can use local, informative image fragments of a given object, rather than the whole object, in order to recognize the object under constant viewing conditions (Ullman et al., 2002; Harel et al., 2007; Ullman, 2007; Hegdé et al., 2008; Lerner et al., 2008; Kromrey et al., 2010). Such image fragments are referred to as “informative fragments.” Computational studies indicate that this fragment-based approach is also beneficial specifically for invariant object

recognition (Bart et al., 2004; Ullman and Bart, 2004), including for pose and illumination invariance.

These studies have identified two broad functional sub-categories of informative fragments useful for invariant recognition. One sub-category of fragments, referred to as “Invariant fragments,” are those local features whose appearance is largely resistant to variations in viewing conditions. For instance, the appearance of the hairline changes relatively little under variations of illumination, which therefore makes it useful for illumination-invariant face recognition. On the other hand, the appearance of many features changes significantly with viewing conditions, which makes them unsuitable as invariant fragments.

“Extended fragments” are a second sub-category of fragments useful for invariant object recognition. In contrast to invariant fragments, extended fragments do not require feature appearance to be stable under changes in viewing conditions. Instead, an extended fragment records the appearance of the given feature under all viewing conditions of interest. In principle, this may involve simply memorizing the appearance of a given feature under each set of viewing conditions. An extended fragment can then be used for recognizing the feature regardless of the viewing

conditions. For instance, even though the appearance of a nose changes under variations of illumination, it can still be useful for recognition if one learns how a nose looks under various illuminations. Since extended fragments do not depend on feature appearance being resistant to viewing conditions, any feature can be used as an extended fragment. Therefore, extended fragments may provide more information to the visual system and thus achieve better performance. However, extended fragments may be more difficult to learn than invariant fragments. This is because, in order to use a feature as an extended fragment, one must somehow learn its appearance under the various viewing conditions.

The extent to which the human visual system actually uses either extended or invariant fragments in object recognition is largely unclear. The mechanisms by which we learn either type of fragments, and conditions under which they can be learned, are also unknown. While previous studies have addressed the question of feature learning in general [e.g., (Kobatake and Tanaka, 1994; Schyns et al., 1998; Wallis and Bulthoff, 1999; Wallis et al., 2009)], it is unclear whether and to what extent the mechanisms suggested by these studies can generalize to learning extended or invariant fragments, given that the nature of object fragments is fundamentally different from the features addressed by these studies (for details, see Ullman, 2007; Hegdé et al., 2008).

The present study focused on testing a specific hypothesis, namely that the visual system is capable of using extended and/or invariant fragments to help achieve a particular type of perceptual constancy, namely illumination-invariant object recognition. In particular, we varied the direction of illumination while holding all other viewing parameters, including other illumination parameters such as brightness or color of illumination, constant. Note that the general framework of extended and invariant fragments is not limited to illumination; in particular, it has been used for pose-invariant recognition as well (Bart et al., 2004; Ullman and Bart, 2004).

We have previously shown, in the context of the aforementioned informative fragments, that both humans and monkeys automatically learn the fragments when they learn new object categories, and can use the learned fragments to recognize whole objects (Hegd  et al., 2008; Kromrey et al., 2010). We therefore use a similar experimental design in the present study to characterize how the human visual system learns and uses extended and/or invariant fragments for illumination-invariant object recognition. We find that human subjects can automatically learn extended fragments when they learn new objects, and can use the learned extended fragments to recognize whole objects.

## MATERIALS AND METHODS

### PARTICIPANTS

Five adult volunteer human subjects (three females) with normal or corrected-to-normal vision participated in this study. All protocols used in this study conformed to the relevant regulatory standards, and were approved in advance by the Human Assurance Committee of the Georgia Health Sciences University, where the psychophysical experiments were carried out. All subjects gave informed consent prior to participating in the study.

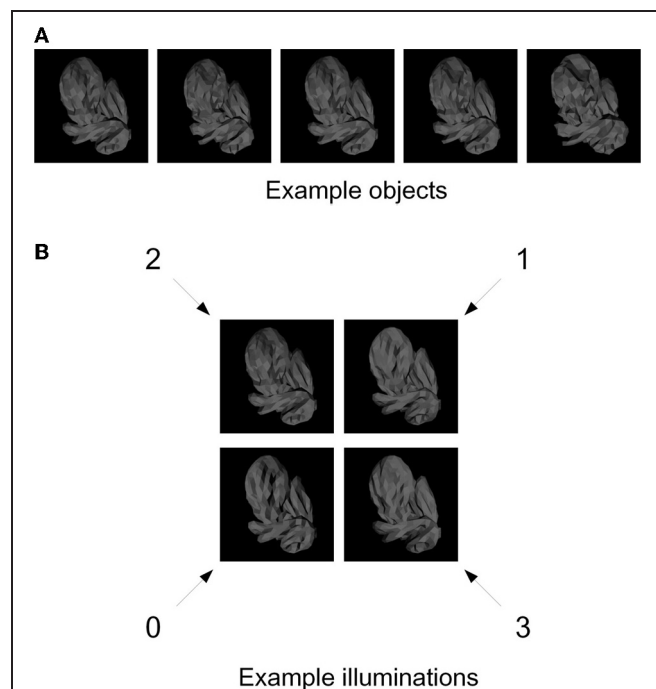
### STIMULI

We generated 50 novel, naturalistic virtual 3-D objects called “digital embryos” using a custom implementation of the Virtual Phylogenesis (VP) algorithm (Brady and Kersten, 2003; Heg  et al., 2008; Hauffen et al., in press). All 50 embryos were descendants of the same parent object, and thus constituted a single naturalistic “category.” The overall appearance of all objects was similar, with relatively small variations distinguishing individual objects from each other, so that distinguishing one embryo from another was nontrivial (see, e.g., **Figure 1A**). It is important to note that these shape variations were not imposed externally, but rather arose randomly during VP. To the extent that VP simulates the natural processes of morphogenesis and phylogenesis, these variations can be considered naturalistic (Hegd  et al., 2008; Hauffen et al., in press).

For each embryo, we generated four different images, corresponding to four different directions of illumination (illuminated from top left, top right, bottom left, and bottom right; see **Figure 1B**) using the 3DS Max graphics toolkit (Autodesk, Inc., San Rafael, CA).

### FRAGMENT SELECTION

Difference-of-Gaussians (DoG) interest points were located in each embryo image as described by (Lowe, 2004). A  $20 \times 20$ -pixel



**FIGURE 1 | Training stimuli. (A)** Five example digital embryos from our training set. All five embryos are shown under the same illumination. Note that the embryos are perceptually similar enough that distinguishing among them is not trivial. **(B)** The four directions of illumination used in our experiments. The directions are denoted by arbitrary numbers: 0 (illuminated from bottom left), 1 (from top right), 2 (from top left), and 3 (from bottom right). The same digital embryo is shown under the four illumination directions to illustrate the appearance changes induced by changes in illumination.

window around each interest point was extracted to form a candidate fragment.

For each fragment, the mutual information (MI) was computed. The MI  $I(F; L)$  between the fragment  $F$  and the object identity label  $L$  is defined as  $I(F; L) = H(L) - H(L | F)$ , where  $H(X)$  is the entropy of the random variable  $X$  and measures the uncertainty in the value of  $X$ . Thus,  $H(L)$  is the uncertainty in the identity label of the given image in the absence of any information, and  $H(L | F)$  is the uncertainty in the identity given the information in the fragment  $F$ . Therefore, MI of the fragment  $F$  measures how much the uncertainty about object identity decreases by using the given fragment.

In practice, the MI can be computed by using the expression

$$I(F; L) = \sum_{f,l} p(f, l) \log \frac{p(f, l)}{p(f)p(l)}. \quad (1)$$

The quantities of interest  $p(f, l)$ ,  $p(f)$ , and  $p(l)$  can be evaluated from the training images, i.e., the set of images used as input to the fragment selection process. For example, the quantity  $p(F = 1)$  is the probability that a given fragment is present in an image. Similarly, the quantity  $p(F = 1, l)$  is the probability that the fragment is present in an image of object  $l$ .

The presence of fragments in images was determined by using the absolute value of normalized cross-correlation (ANCC), as previously described in (Bart et al., 2004; Ullman and Bart, 2004). Briefly, to determine whether a given  $20 \times 20$ -pixel fragment  $V$  was present in a given image  $X$ , ANCC was first computed between the fragment and all  $20 \times 20$ -pixel windows in the image. The highest ANCC value was taken; this highest value is denoted  $A(V, X)$ . If  $A(V, X)$  was above a pre-determined threshold, the fragment was considered present in the image ( $F = 1$ ); otherwise, it was considered absent ( $F = 0$ ). Thus, a  $20 \times 20$ -pixel fragment and a threshold determine the variable  $F$  and can be used to compute MI. The appropriate value of the threshold itself was determined by considering multiple threshold values for each fragment and selecting the threshold that maximized MI, as in (Bart et al., 2004; Ullman and Bart, 2004). ANCC values themselves were computed as follows. For two  $20 \times 20$ -pixel windows  $V, W$ , normalized cross-correlation is defined as

$$\begin{aligned} NCC(V, W) \\ = \frac{\frac{1}{20 \times 20} \sum_{x=1}^{20} \sum_{y=1}^{20} (V[x, y] - \bar{V})(W[x, y] - \bar{W})}{\sigma_V \sigma_W}, \end{aligned} \quad (2)$$

where  $V[x, y]$  is the pixel value at position  $(x, y)$  in the window  $V$ ,  $\bar{V}$  is the average of all pixel values in  $V$ ,  $\sigma_V$  is the standard deviation of the pixel values, and similarly for  $W$ . Normalized cross-correlation has values between  $+1$  and  $-1$ ; the value  $+1$  indicates perfect correlation,  $-1$  indicates perfect anti-correlation, and  $0$  indicates no correlation. The ANCC therefore has values between  $0$  and  $1$ , with lower values indicating weaker correlation and higher values indicating stronger positive or anti-correlation (in practice, anti-correlation rarely occurs in our images; data not shown). An example of MI computation and threshold selection is given in Figure 2.

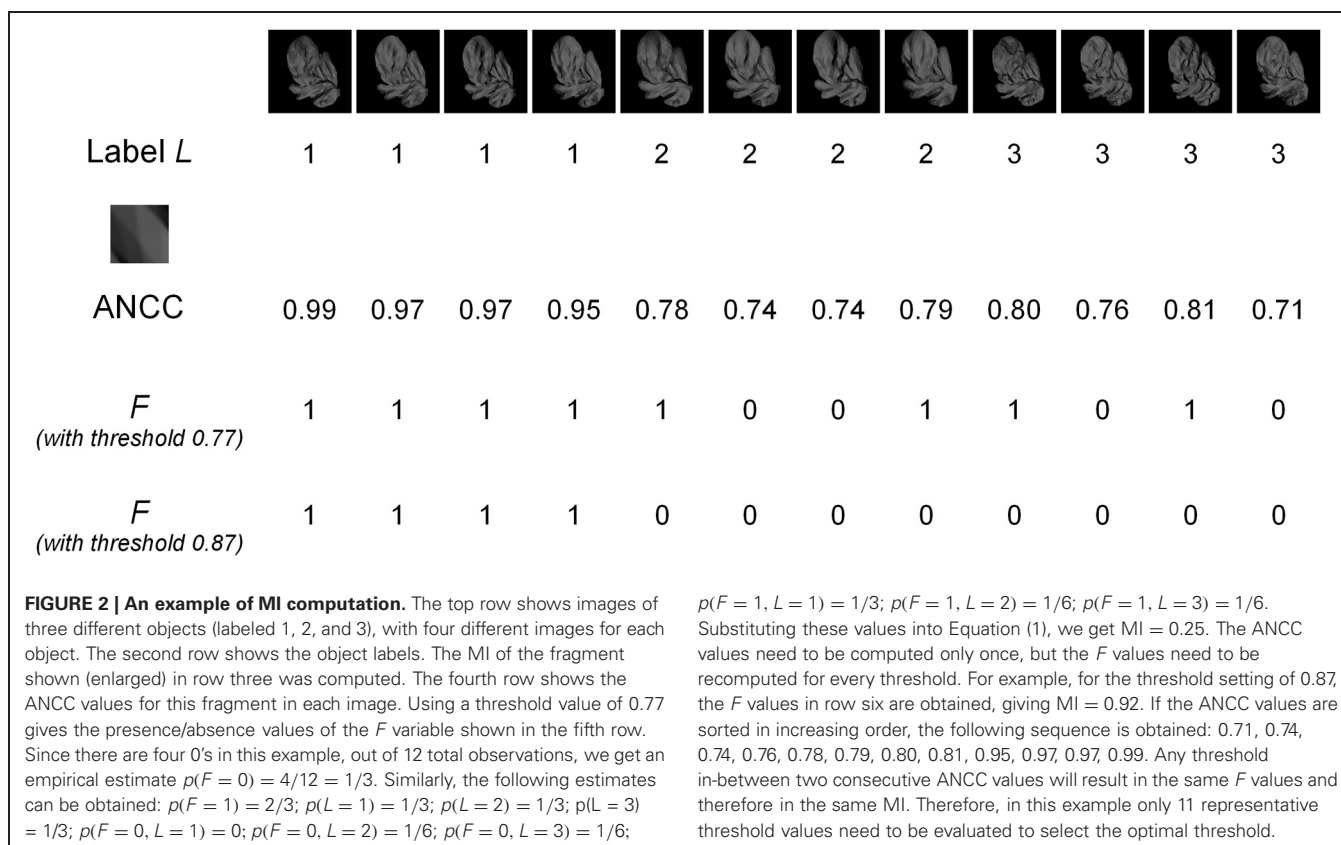
There are two ways to use ANCC to determine a fragment's presence in a given image. One is to render the fragment under a fixed illumination (say, illumination 0) and use the rendering at this illumination to compute ANCC, regardless of which illumination the given image is in. Mathematically, we set  $F = 1$  if the single template's ANCC value is above the threshold and compute MI using Equation (1). Of course, if the fragment appearance changes across illuminations, the results will be poor when the fragment illumination is different from the image illumination. This method of computation therefore implicitly assumes that the fragment's appearance is invariant to viewing conditions. When fragments are used in this manner, they are called "invariant fragments," and MI computed in this manner is called "Invariant MI" and denoted by  $I_{\text{inv}}$ . See Bart et al. (2004), Ullman and Bart (2004) for details.

A second method of using ANCC to determine the presence of a fragment in a given image is to learn the appearance of each fragment under all illuminations in question. Computationally, this requires rendering and storing for each fragment the four templates, one for each illumination, as illustrated in Figure 3. In a biological system, this could be achieved by learning the appearance of a given feature in a given set of training examples. Given an image in a particular illumination, all four templates are matched to it using ANCC, and the best-matching template is selected in order to calculate the similarity. Mathematically, we set  $F = 1$  if the maximal ANCC value over all four templates is above the threshold. The advantage of this method is that matching across illuminations is no longer necessary. In most cases, the template with the best ANCC value will automatically be the one that matches the image illumination (Bart et al., 2004; Ullman and Bart, 2004), thus eliminating comparison across illuminations. This generally results in much better similarity estimates and improved recognition performance (Bart et al., 2004; Ullman and Bart, 2004). The disadvantage is that training examples are needed, and the learning process may be difficult. When fragments are used in this manner, they are called "extended fragments," and MI computed in this manner is called "Extended MI" and denoted by  $I_{\text{ext}}$ . See Bart et al. (2004); Ullman and Bart (2004) for details.

For each candidate fragment, both  $I_{\text{ext}}$  and  $I_{\text{inv}}$  were calculated. Four "goodness" measures were defined as follows:

- $G_1 = I_{\text{ext}} + I_{\text{inv}}$  favors fragments that have high Extended MI and high Invariant MI.
- $G_2 = I_{\text{ext}} - I_{\text{inv}}$  favors fragments that have high Extended MI and low Invariant MI.
- $G_3 = -I_{\text{ext}} + I_{\text{inv}}$  favors fragments that have low Extended MI and high Invariant MI.
- $G_4 = -I_{\text{ext}} - I_{\text{inv}}$  favors fragments that have low Extended MI and low Invariant MI.

For each of these measures, the fragments were sorted according to the decreasing value of the measure. Note that  $G_3 = -G_2$ ; the reason to use both is that we wanted to have fragments with high  $I_{\text{ext}}$  and low  $I_{\text{inv}}$ , as well as fragments with low  $I_{\text{ext}}$  and high  $I_{\text{inv}}$ . This allowed us to disassociate between  $I_{\text{ext}}$  and  $I_{\text{inv}}$  and determine how each separately affects the performance. Similarly,



$G_4 = -G_1$ ; the reason to have  $G_1$  was to assess any additive effects of  $I_{\text{ext}}$  and  $I_{\text{inv}}$ , while the reason to have  $G_4$  was to assess the performance of uninformative fragments. The top 20 fragments were selected for each measure, subject to the constraint that a fragment's visual similarity (as measured by ANCC) to any previously selected fragment could not exceed 0.9. This resulted in a total of 80 fragments.

The 20 fragments selected by measure  $G_1$  are shown in **Figure 7**. Note that there are still many fragments in this set that are visually similar to each other and thus redundant. Therefore, five non-redundant fragments were selected from this set manually by the authors (fragments 1–5 in **Figure 3** and fragments 3, 4, 14, 17, and 20 in **Figure 7**). Similarly, five non-redundant fragments out of each subset of 20 were selected for the other goodness measures. This resulted in the final set of 20 non-redundant fragments shown in **Figure 3**. Note that this final set contains five fragments selected by each of the four goodness measures.

### TRAINING IN ILLUMINATION-INVARIANT OBJECT RECOGNITION

















































































Except where noted otherwise, the procedures used in the psychophysical training phase (this section) and testing phase (see next section) of the experiment were identical to those described by us previously (Hegde et al., 2008). Briefly, during the training phase, we trained the subjects to recognize individual digital embryos across illuminations using a simultaneous match-to-sample task. In this task, the subjects had to match a single sample embryo at one illumination at the center of the screen to an array

of ten test embryos at another illumination arranged along the periphery of the screen (**Figure 4**). The subjects were allowed unlimited time to examine the images and arrive at a decision. Once the subjects reported their decision using a key press, visual feedback was provided (including the correct response, if the subject's response was wrong). The subjects had unlimited time to re-examine the display in light of the feedback. During initial training, the subjects were not required in any way to learn the fragments, nor were they even told of their existence. The performance was monitored across the training blocks (**Figure 5**). After a subject's performance remained asymptotic at above-chance levels for at least three sequential training blocks of 50 trials each (binomial tests,  $p < 0.05$ ), the subjects moved to the testing phase (see below). All the subjects achieved asymptotic learning within 10 blocks (not shown). To minimize day-to-day forgetting of the learned objects, each subject carried out up to 50 "refresher" training trials at the start of each testing day. Note that during these refresher trials the subjects were aware of the existence of fragments, although they still weren't explicitly asked to learn them.

### TESTING ILLUMINATION-INVARIANT OBJECT RECOGNITION USING FRAGMENTS

During the testing phase, the subjects performed an object identification task on the sole basis of a given fragment. In each trial, a composite object showing a sample fragment at illumination 0 was displayed at the center of the screen. Two test embryos at illumination 3 abutted the composite object. All stimuli



Index	Illumination				$MI_{ext}$	$MI_{inv}$
	0	1	2	3		
1					0.89	0.86
2					0.95	0.70
3					0.80	0.66
4					0.77	0.67
5					0.77	0.65
6					0.83	0.05
7					0.81	0.11
8					0.84	0.14
9					0.84	0.16
10					0.80	0.14
11					0.26	0.74
12					0.25	0.65
13					0.27	0.65
14					0.31	0.67
15					0.27	0.62
16					0.09	0.08
17					0.13	0.08
18					0.12	0.11
19					0.12	0.10
20					0.05	0.19

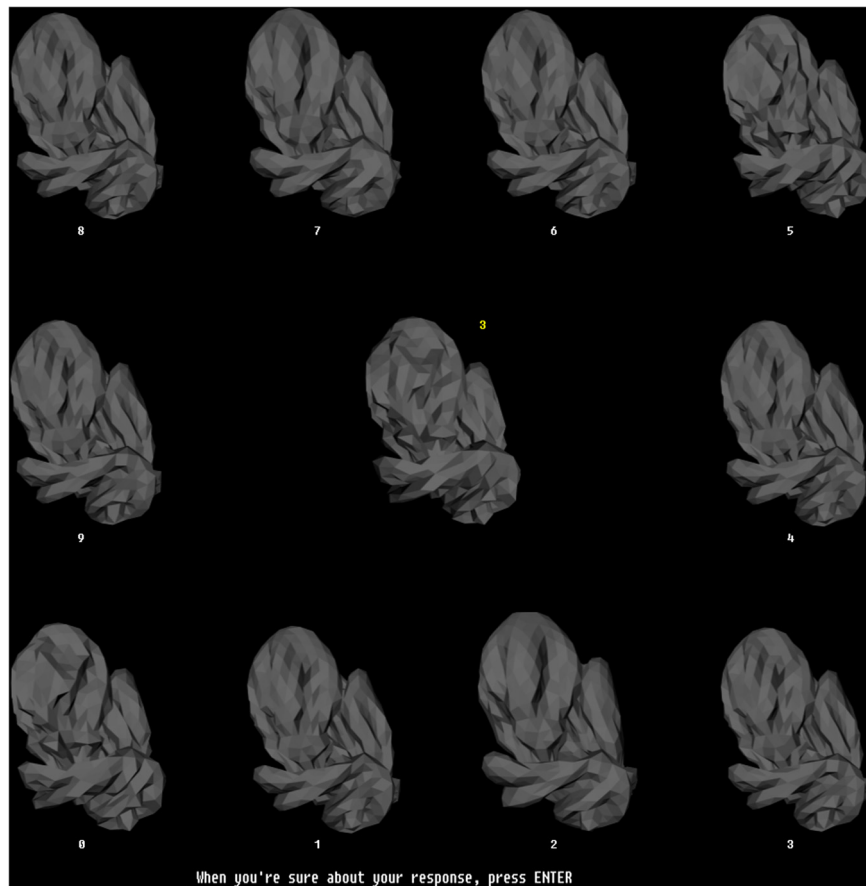
**FIGURE 3 | The 20 fragments used in our experiments.** The appearance of each fragment under each of the four illuminations is shown, as well as the corresponding Extended and Invariant MI values ( $MI_{ext}$  and  $MI_{inv}$ , respectively).

were presented simultaneously for 3000 ms (**Figure 6**). Only one fragment in the composite object was clearly visible (see below). This fragment (called the “sample fragment”) was also present in one of the test embryos (“positive embryo,” presented on a randomly chosen side during a given trial) and absent from the other test embryo (“negative embryo”). Following a 200 ms random noise mask, subjects had unlimited time to indicate, based on the

sample fragment in the composite object, whether the composite object was the same as the left test embryo or the right test embryo.

The composite object was generated by graphically overlaying the sample fragment over a randomly drawn “background” embryo. The composite object was shown to the subject behind a rectangular translucent occluder with a hole, so that only the





**FIGURE 4 | The training paradigm.** This figure illustrates the configuration of stimuli during a typical trial during the training phase. During each trial, a randomly selected sample embryo was shown in the center in a randomly selected illumination. The 10 test stimuli were shown simultaneously arrayed along the periphery of the screen. The illumination was the same across all sample stimuli, but was different from the illumination of the sample embryo. The test embryos were assigned randomly to numbered locations (white numbers). One of the test embryos was the same object as the sample embryo, but at a different illumination. The subjects had to identify the test

embryo that matched the sample embryo, and enter the number of this test embryo using the computer's keyboard, which then appeared as a yellow number next to the sample embryo. Note that this task required the subjects to generalize across the illuminations. The subjects pressed another key to finalize their response. After the subjects finalized their response, they received visual feedback (not shown), along with the correct response, if the subject's response was incorrect. Subjects had unlimited time both to perform the task and to examine the subsequent feedback. The stimulus configuration shown subtended  $26^\circ \times 26^\circ$  during the actual experiments.

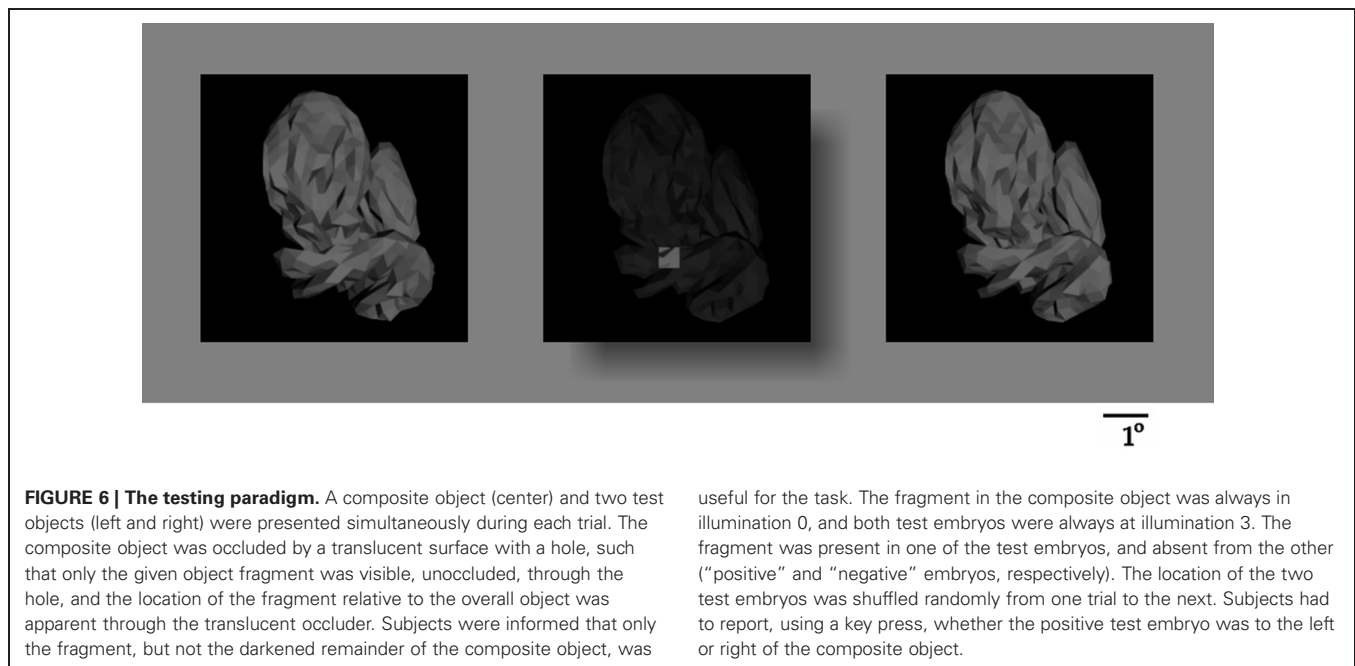
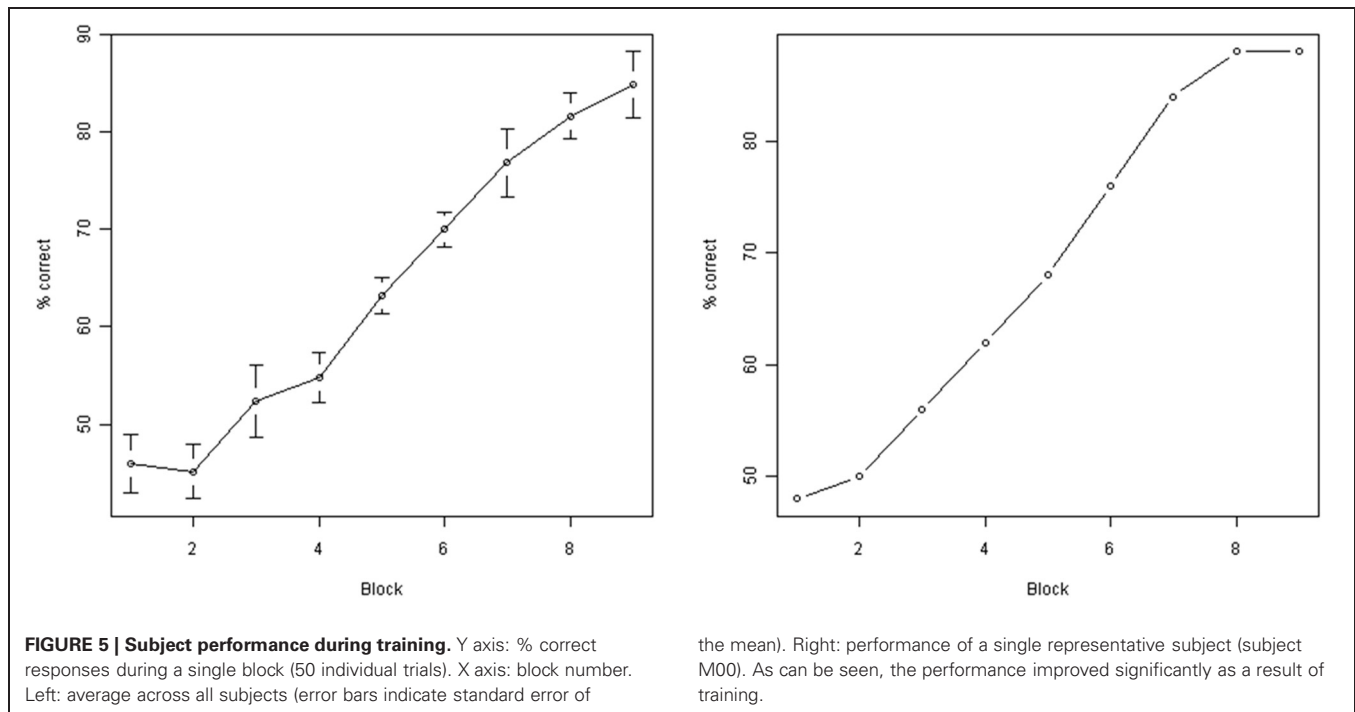
sample fragment ( $0.53^\circ \times 0.53^\circ$ ) was visible unhindered through the hole in its proper position on the object, whereas the rest of the object appeared as a faded “background” (see **Figure 6**). This design helped ensure that the subjects saw the sample fragment in its proper spatial context. This design is better than presenting the sample fragment by itself without the spatial context, because it minimizes the possibility that the subject may have to use semantic and spatial cues (e.g., configural cues, such “the corner of the left eye”) to help perform the task.

Subjects were informed that only the unoccluded fragment of the composite object was useful for the task, and that the faded background portion of the composite object (i.e., the portion visible behind the translucent occluder) was randomly selected, so that they would not be able to perform the task above chance levels using the background object.

Two different test objects (called “test embryos”) were shown on either side of the composite object. Whole objects, rather

than just fragments, were used as test objects to help ensure that (1) the task involved object identification, as opposed to simple visual matching of individual fragments, and (2) task required only implicit perceptual learning and not declarative (or explicit) association between a fragment and an object.

A sample fragment and two test embryos (one positive and one negative) constitute a “testing configuration.” For each fragment in **Figure 3**, five embryos in which the fragment was most active, and five embryos in which it was least active, were selected. This activation level was measured by finding the highest ANCC value among all illuminations of a given embryo. All 25 possible testing configurations for each of the 20 fragments were created, resulting in 500 total testing configurations. This choice of testing configurations was motivated by the following considerations. The test embryos need to be visually distinguishable on the basis of the sample fragment; otherwise, the trial will be



















































































meaningless as the fragment will provide no information as to the correct answer. Embryos with highest fragment activation were compared to embryos with the lowest fragment activation to maximize this visual distinguishability. We used five embryos of each type, because fewer than 25 configurations per fragment might be insufficient to ensure thorough testing, while more than 25 configurations would make the testing too long and laborious for the subjects.

No feedback was provided during testing. Each fragment was presented over six randomly interleaved repetitions for each subject, so each subject performed 3000 trials during the testing phase.

#### DATA ANALYSIS

The results were analyzed using scripts custom-written in R (r-project.org) and Matlab (Mathworks, Natick, MA, USA).

Index	Illumination				MI <sub>ext</sub>	MI <sub>inv</sub>
	0	1	2	3		
1					0.95	0.83
2					0.92	0.83
3					0.89	0.86
4					0.95	0.70
5					0.92	0.72
6					0.86	0.78
7					0.83	0.72
8					0.90	0.64
9					0.88	0.64
10					0.91	0.61
11					0.78	0.73
12					0.83	0.66
13					0.84	0.64
14					0.80	0.66
15					0.85	0.60
16					0.75	0.69
17					0.77	0.67
18					0.90	0.53
19					0.83	0.60
20					0.77	0.65

**FIGURE 7 | The 20 fragments selected by the  $G_1$  measure.** The appearance of each fragment under each of the four illuminations is shown, as well as the fragment's Extended and Invariant MI.

Additional details of the analyses are provided in the “Results” section, where underlying rationale will be clearer.

## RESULTS

Our study was aimed at testing the hypothesis that the human visual system *can* use invariant and/or extended fragments to achieve invariant object recognition. During the testing

phase of the experiment, the subjects had to determine which of the two test embryos contained the sample fragment (i.e., which one was the “positive” embryo). This task was difficult, because the sample fragment was presented in illumination 0, while both test embryos were presented in illumination 3. This difference in illumination induced a significant change in appearance (see, e.g., **Figure 3**) that the

subjects had to compensate for in order to perform the task properly.

Several possible strategies for performing this task were evaluated. One possible strategy is that during each trial, the subject matches the fragment's visual appearance directly to both test embryos and selects the embryo that resembles the fragment more closely. Another possibility is that the subject discounts the illumination (for example, by somehow transforming the fragment into the embryo's illumination or vice versa) and then performs the visual comparison. A third possibility is that, as suggested by a computational model of invariance (Bart et al., 2004; Ullman and Bart, 2004) and our previous experiments (Hegdé et al., 2008; Kromrey et al., 2010), the subjects preferentially learn fragments that are useful for the object recognition task they were previously trained on. This usefulness can be measured objectively by using MI.

MI can be calculated under one of two hypotheses. One possibility is that the subjects assume illumination invariance, or preferentially seek out and exploit invariant fragments. The other possibility is that the subjects make no invariance assumptions and instead use learning to compensate for appearance changes across illumination by using extended fragments.

Five predictor variables corresponding to the strategies outlined above were computed for each testing configuration (i.e., the set of the sample fragment and two test embryos presented during a given trial):

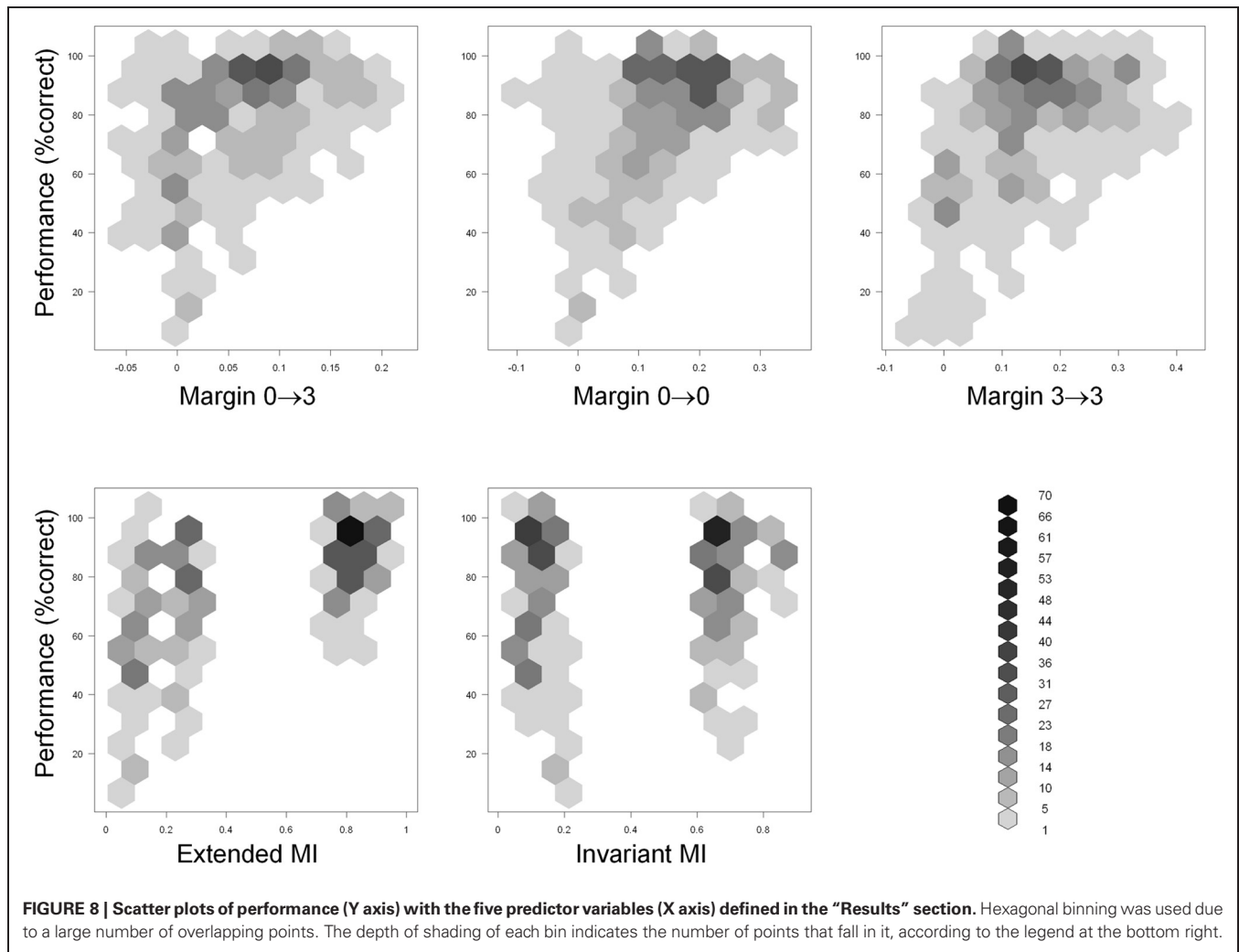
- (1) M03 was the difference in visual similarity of the fragment to the “positive” and “negative” test images shown to the subject. Visual similarity was measured by ANCC, as described above. Denoting the fragment rendered in illumination 0 by  $V_0$ , the positive test embryo rendered in illumination 3 by  $X_3^+$ , and the negative test embryo rendered in illumination 3 by  $X_3^-$ , M03 was defined as  $A(V_0, X_3^+) - A(V_0, X_3^-)$ , where  $A$  was the ANCC value, as defined above. If the subjects used the naive strategy of direct matching by visual appearance, this M03 would be expected to correlate strongly with performance. Note that in practice, this strategy is likely to result in poor performance, since the fragment and the embryo images had different illuminations. This variable is also called Margin  $0 \rightarrow 3$ , which refers to the fact that a fragment in illumination 0 is matched to images in illumination 3.
- (2) M00 was the difference in visual similarity of the fragment to the “positive” and “negative” embryos rendered in illumination 0 (same illumination as the fragment). M00 was defined as  $A(V_0, X_0^+) - A(V_0, X_0^-)$ , where  $X_0^+$  was the positive test embryo rendered in illumination 0, and  $X_0^-$  was the negative test embryo rendered in illumination 0. This variable is also called Margin  $0 \rightarrow 0$ . If the subjects mentally transformed the embryo images to illumination 0 and then used matching by visual appearance, this value would be expected to correlate strongly with performance.
- (3) M33 was the difference in visual similarity of the fragment, rendered in illumination 3 (same illumination as the embryo images) to the “positive” and “negative” images displayed. M33 was defined as  $A(V_3, X_3^+) - A(V_3, X_3^-)$ , where  $X_3^+$  was the positive test embryo rendered in illumination 3,  $X_3^-$  was the negative test embryo rendered in illumination 3, and  $V_3$  was the fragment rendered in illumination 3. This variable is also called Margin  $3 \rightarrow 3$ . If the subjects mentally transformed the fragment image to illumination 3 and then used matching by visual appearance, this value would be expected to correlate strongly with performance.
- (4) Extended MI ( $MI_{\text{ext}}$ ) measured how useful the given fragment is for object recognition for subjects who use extended fragments.
- (5) Invariant MI ( $MI_{\text{inv}}$ ) measured how useful the given fragment is for object recognition for subjects who rely on the invariance of features across illumination.

Note that the first three variables, in general, change from one stimulus configuration to the next, while the last two variables have the same value for all 25 configurations involving a single fragment.

Scatter plots of performance with the five predictor variables are shown in **Figure 8**. Examination of performance averaged across all subjects revealed that the subjects systematically underperformed in many configurations despite abundant visual cues. This suggests (although does not, by itself, prove) that visual appearance alone was insufficient to explain the subjects' performance. To help discern whether this is indeed the case, we defined a configuration to be “visually recognizable” if the margin M03 was above 0.05 (note that the absolute values of normalized correlation range from 0 to 1). This threshold is shown as a red vertical line in **Figure 9**. The underlying intuition was that this amount of visual difference is easily detectable by human observers and can therefore be interpreted reliably. This intuition is confirmed by the fact that 67 configurations with M03 less than 0.05 were recognized correctly in over 80% of the trials (blue rectangle in **Figure 9**). In other words, even smaller margins were sufficient to allow reliable recognition. However, there were 49 configurations with a margin above 0.05 whose recognition rate was between 50 and 70% (green rectangle in **Figure 9**). Note that 50% recognition is expected by chance. In other words, even though these configurations contained sufficient visual cues to perform the task, the subjects systematically failed to do so. Similar results can be obtained using M00 or M33 to define visual recognizability instead of M03 (see **Figure 8**). These informal considerations support, although do not by themselves prove, the notion that factors other than visual recognizability significantly affect subjects' performance.

To rigorously analyze the intuition presented above, we fitted a linear regression to the data that accounted for the average performance in terms of the aforementioned five independent variables. An examination of the fitted model revealed that  $MI_{\text{ext}}$  was the only independent variable that contributed significantly to the fit (**Table 1**). This contribution was highly significant ( $p = 1.5 \times 10^{-14}$ ,  $F$ -test). The contributions of the three visual variables (M03, M00, and M33), as well as the contribution of  $MI_{\text{inv}}$ , were each statistically insignificant ( $p > 0.05$ ).

We also compared the regression with the three purely visual variables (M03, M00, and M33) to regression with all five variables. Adding the MI-based variables had a highly significant effect ( $p = 7 \times 10^{-15}$ ,  $F$ -test). In other words, even after



**FIGURE 8 |** Scatter plots of performance (Y axis) with the five predictor variables (X axis) defined in the “Results” section. Hexagonal binning was used due to a large number of overlapping points. The depth of shading of each bin indicates the number of points that fall in it, according to the legend at the bottom right.

accounting for the purely appearance-based factors given by the variables M03, M00, and M33, the MI-based variables explained a significant additional fraction of variance. In contrast, the performance of the two MI-based variables by themselves did not improve further after adding the three purely visual variables ( $p = 0.09$ ,  $F$ -test). That is, the visual variables add no information beyond that already contained in the MI-based variables.

These analyses further support the conclusion that subjects do not rely on visual appearance alone, and can preferentially use extended fragments that are useful for the recognition task.

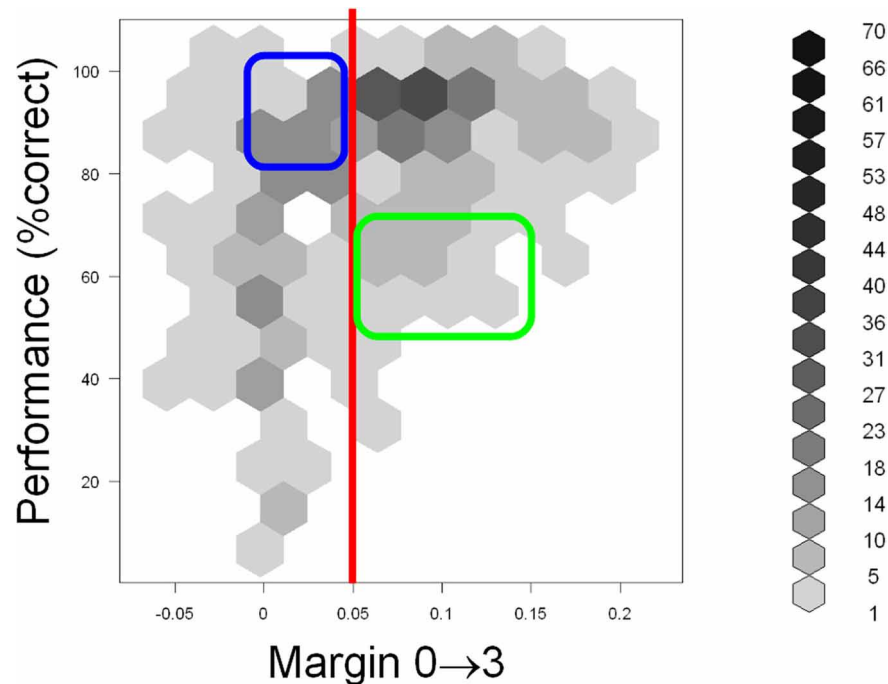
## DISCUSSION

### INVARIANT OBJECT RECOGNITION BASED ON FRAGMENTS

Our results empirically confirm, for the first time, the hypothesis that the human visual system can use extended fragments to achieve invariant object recognition. The results found no support for the use of invariant fragments by the visual system. Note that this does not necessarily mean that the visual system cannot use invariant fragments for invariant object recognition under

any circumstances; rather, it only shows that invariant fragments were not used in the current experiment. Nonetheless, it is worth noting that the statistical power of the sample was adequate enough to find affirmative evidence that the human visual system is capable of using extended fragments for invariant recognition. It is also worth noting that the fact that the visual system can use extended fragments under our experimental conditions does not necessarily mean that extended fragments are the universal, much less the sole, means by which the visual system achieves invariant object recognition in general, or illuminant-invariant recognition in particular (also see below).

The demonstration that the visual system is capable of using extended fragments is significant, for two main reasons. First, it provides the empirical “existence proof” for a hitherto theoretical idea. Second, as extensively noted by previous studies, fragment-based object approach is a substantially different approach to object recognition in general, and invariant object recognition in particular, than the conventional approaches based on whole objects [cf. (Wallis and Bulthoff, 1999; Christou and Bulthoff, 2000; Rolls, 2008; Biederman and Cooper, 2009; Wallis et al., 2009)]. Therefore, the empirical demonstration that the visual



**FIGURE 9 | Scatter plot of performance (Y axis) with the M03 variable (X axis) defined in the “Results” section.** Hexagonal binning was used due to a large number of overlapping points. The depth of shading of each bin indicates the number of points that fall in it, according to the legend on the right. The red line indicates the “visual recognizability” threshold, defined in

the “Results” section. Note that testing configurations remained discernible even below this threshold (blue rectangle). However, subjects systematically underperformed in some highly recognizable configurations (green rectangle), indicating that factors other than visual recognizability affected performance. See text for details.

**Table 1 | Coefficients of linear regression for the five independent variables and the intercept term.**

Variable	Estimate	Std. Error	Partial $r^2$	$p$ value (F-test)
Intercept	21.0	0.2	(Not Applicable)	$2.0 \times 10^{-16}$
M03	0.5	0.3	0.004	0.06
M00	0.5	0.4	0.003	0.15
M33	0.1	0.4	0.0001	0.71
MI <sub>ext</sub>	2.4	0.3	0.08	$1.5 \times 10^{-14}$
MI <sub>inv</sub>	0.4	0.2	0.004	0.06

With all terms included, the value of  $r^2$ , the coefficient of determination, was 0.39.

system can use fragments for this purpose opens important new avenues of future research for invariant object recognition in general, and illumination-invariant object recognition in particular (also see below).

### ILLUMINATION-INVARIANT OBJECT RECOGNITION

Several key implications of our results for illumination-invariant object recognition are worth noting. First, if the subjects only used visual cues to perform the testing task, then the performance would be explained by the margin variables and would not be affected by MI. Since adding MI in fact improves

the fit highly significantly ( $p = 7 \times 10^{-15}$ , F-test), we conclude that the subjects preferentially use informative fragments that are useful for the recognition task they were trained with. In contrast, uninformative fragments are neglected, even when sufficient visual information is available for accurate recognition.

Second, if the subjects compensated for illumination effects at the level of whole objects, then illumination of all features of a given object would be compensated for in a similar manner. The performance would thus depend only on how visually recognizable a given feature is after accounting for illumination. In practice, however, fragments with similar visual recognizability have dramatically different recognition rates (see “Results” for details). These considerations indicate that illumination compensation occurs on a feature level, rather than on a whole object level.

Finally, if the subjects assumed (implicitly or explicitly) that individual features were invariant to illumination, then the usefulness of individual features for recognition would be given by the Invariant MI. However, Invariant MI did not contribute significantly to explaining performance ( $p > 0.05$ , F-test). In contrast, the contribution of Extended MI, computed under the assumption that illumination is compensated for by extended fragments, rather than by assuming invariance, was highly significant ( $p = 1.5 \times 10^{-14}$ , F-test). Thus, subjects are highly unlikely to have assumed invariance, but rather must have



compensated for viewing conditions by using extended fragments.

Note that for the training task we have used, computational simulations predict invariant features to perform much poorer than extended fragments. We cannot therefore conclude that subjects always use extended fragments. It is possible that when invariant features are sufficient to perform a task, those would be used instead of, or in addition to, extended fragments. However, as noted above, our results do provide an “existence proof” that subjects are capable of using extended fragments, and do use them when needed. Further work is necessary to determine under what conditions extended fragments can be learned and used. However, our “existence proof” is by itself an important conclusion, because using extended fragments is a nontrivial task.

Together, the above arguments support two main conclusions. First, illumination invariance is not achieved on a whole-object level. Rather, the illumination is compensated for feature-by-feature, with some features being preferred over others. The preferred features are those which support the recognition task, and their appearance variations are compensated for more carefully. Second, the subjects do not rely on invariance of individual features. Rather, they are capable of using extended fragments to compensate for appearance changes when necessary. Both conclusions fit closely with the computational model for invariant object recognition developed in Bart et al. (2004) and Ullman and Bart (2004).

#### LEARNING DURING TRAINING VS. PRIOR LEARNING

Using extended fragments to compensate for illumination requires familiarity with the visual appearance of a given object feature under various illuminations. This familiarity may be achieved by learning during the training process. Alternatively, this familiarity may be achieved by generalizing from previous visual experience, or may even be innate. The demonstration that subjects can use extended fragments at all is novel and interesting by itself, regardless of the exact learning mechanism used. We therefore did not attempt to establish the learning mechanism conclusively in this experiment.

In principle, some generalization from prior experience might have occurred in our experiment. For example, a corner may be recognizable as a corner under many different illuminations without dedicated training. However, it seems unlikely that such generalization would affect informative and uninformative fragments differentially, as in our experiment. There were no systematic visual differences between different fragments (see **Figure 3**). Moreover, the notion of MI itself is highly task-specific. For example, by computing MI for a different task where only two (rather than four) illuminations are used, the informativeness of fragments in **Figure 3** changes dramatically. In particular, some of the uninformative fragments in **Figure 3** become highly informative for this modified task (data not shown). The fact that subjects preferentially compensate for illumination changes of fragments informative for the given specific task, rather than for a number of possible alternative tasks, indicates that generalization from prior experience, if it exists, is modulated substantially by learning.

#### IMPLICATIONS FOR THE NEURAL MECHANISMS FOR ILLUMINATION-INVARIANT OBJECT RECOGNITION

The neural mechanisms by which the visual system learns extended fragments, or uses them to achieve illumination-invariant object recognition, remain to be characterized. However, previous neuroimaging studies in human subjects have shown, using informative fragments, that the lateral occipital complex and the posterior fusiform gyrus are preferentially responsive to fragments with high MI values (Lerner et al., 2008), also see Harel et al. (2007). Both of these brain regions are known to play a central role in visual object recognition (Grill-Spector et al., 1999, 2000, 2001, 2004; Grill-Spector and Malach, 2004). Both of these regions have also been previously shown to play important roles in perceptual learning, albeit of whole visual objects (Gauthier and Tarr, 1997; Gauthier et al., 1998, 1999; Bukach et al., 2006; Wong et al., 2009). Taken together, these considerations suggest the possibility that these two brain regions play a key role in learning and/or using extended fragments for illumination-invariant object recognition.

It has been observed that object representations become both more selective and more invariant as they propagate upstream in the visual system (see, e.g., Rust and Dicarlo, 2010). This is thought to be a consequence of the hierarchical architecture of the visual system, where cells at higher levels pool input from several lower-level cells and thus become more tolerant of changes than each individual lower-level cell (Riesenhuber and Poggio, 1999). Our results are consistent with this view, because the features we have used are quite high-level, and are expected to be processed in high-level visual areas, and can therefore be expected to be quite tolerant of viewing conditions.

#### FUTURE DIRECTIONS

It is worth noting that our results, although highly statistically significant, only account for about 40% of the variability in the subjects' performance (**Table 1**). In the future, it would be interesting to determine what factors account for the remaining variability. One potential source of this residual variability is that our sample sizes, even with 3000 trials per subject (see “Materials and Methods”), were nonetheless relatively small from the statistical viewpoint. Using more testing configurations per fragment and repeating the experiments with more subjects and more trials per subject would help reduce the intrinsic randomness in the performance. Another potential source of variability is a scenario where the subjects learn features at a smaller scale than those extracted computationally, but learn different subsets of these smaller features. Although extracting such small features is easy computationally, it may present practical problems for our current experimental setup. This is because even the current features are small enough to cause visibility concerns. However, designing an experiment where smaller features are not useful for recognition, or using a different testing paradigm, may alleviate this problem.

Although the experiments in the current work addressed illumination invariance, it should be noted that our experimental setup can readily be used to test other types of invariant recognition, such as viewpoint (or pose) invariance, size

(or scale) invariance, etc. This could be a particularly interesting direction for future work, especially since the underlying computations are fundamentally the same (Bart et al., 2004; Ullman and Bart, 2004). This is not necessarily to say, however, that the underlying neural mechanisms are the same. Indeed, given that the relevant visual features tend to be processed differently by the visual system (Felleman and Van Essen, 1991; DeYoe et al., 1994; Vuilleumier et al., 2002; Grill-Spector and Malach, 2004), the underlying neural mechanisms are likely to be substantially

different. However, the fragment-based approach provides a common, rigorous conceptual framework for the experimental study of many different types of perceptual invariance.

## ACKNOWLEDGMENTS

Matthew Maestri provided excellent technical assistance with the psychophysical experiments. This work was supported in part by the U.S. Army Research Laboratory and the U.S. Army Research Office grant W911NF-11-1-0105 to Jay Hegdé.

## REFERENCES

- Bart, E., Byvatov, E., and Ullman, S. (2004). "View-invariant recognition using corresponding object fragments," in *Proceedings ECCV, Part II*, (New York, NY: Springer-Verlag), 152–165.
- Biederman, I., and Cooper, E. E. (2009). Translational and reflectional priming invariance: a retrospective. *Perception* 38, 809–817.
- Brady, M. J., and Kersten, D. (2003). Bootstrapped learning of novel objects. *J. Vis.* 3, 413–422.
- Bukach, C. M., Gauthier, I., and Tarr, M. J. (2006). Beyond faces and modularity: the power of an expertise framework. *Trends Cogn. Sci.* 10, 159–166.
- Christou, C., and Bulthoff, H. H. (2000). Perception, representation and recognition: a holistic view of recognition. *Spat. Vis.* 13, 265–275.
- DeYoe, E. A., Felleman, D. J., Van Essen, D. C., and McClendon, E. (1994). Multiple processing streams in occipitotemporal visual cortex. *Nature* 371, 151–154.
- Felleman, D. J., and Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cereb. Cortex* 1, 1–47.
- Gauthier, I., and Tarr, M. J. (1997). Becoming a "Greeble" expert: exploring mechanisms for face recognition. *Vision Res.* 37, 1673–1682.
- Gauthier, I., Tarr, M. J., Anderson, A. W., Skudlarski, P., and Gore, J. C. (1999). Activation of the middle fusiform 'face area' increases with expertise in recognizing novel objects. *Nat. Neurosci.* 2, 568–573.
- Gauthier, I., Williams, P., Tarr, M. J., and Tanaka, J. (1998). Training 'greeble' experts: a framework for studying expert object recognition processes. *Vision Res.* 38, 2401–2428.
- Grill-Spector, K., Knouf, N., and Kanwisher, N. (2004). The fusiform face area subserves face perception, not generic within-category identification. *Nat. Neurosci.* 7, 555–562.
- Grill-Spector, K., Kourtzi, Z., and Kanwisher, N. (2001). The lateral occipital complex and its role in object recognition. *Vision Res.* 41, 1409–1422.
- Grill-Spector, K., Kushnir, T., Edelman, S., Avidan, G., Itzhak, Y., and Malach, R. (1999). Differential processing of objects under various viewing conditions in the human lateral occipital complex. *Neuron* 24, 187–203.
- Grill-Spector, K., Kushnir, T., Hendler, T., and Malach, R. (2000). The dynamics of object-selective activation correlate with recognition performance in humans. *Nat. Neurosci.* 3, 837–843.
- Grill-Spector, K., and Malach, R. (2004). The human visual cortex. *Annu. Rev. Neurosci.* 27, 649–677.
- Harel, A., Ullman, S., Epshtein, B., and Bentin, S. (2007). Mutual information of image fragments predicts categorization in humans: electrophysiological and behavioral evidence. *Vision Res.* 47, 2010–2020.
- Hauffen, K., Bart, E., Brady, M., Kersten, D., and Hegdé, J. (in press). Creating objects and object categories for studying perception and perceptual learning. *J. Vis. Exp.*
- Hegdé, J., Bart, E., and Kersten, D. (2008). Fragment-based learning of visual object categories. *Curr. Biol.* 18, 597–601.
- Kobatake, E., and Tanaka, K. (1994). Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. *J. Neurophysiol.* 71, 856–867.
- Kromrey, S., Maestri, M., Hauffen, K., Bart, E., and Hegdé, J. (2010). Fragment-based learning of visual object categories in non-human primates. *PLoS ONE* 5:e15444. doi: 10.1371/journal.pone.0015444
- Lerner, Y., Epshtein, B., Ullman, S., and Malach, R. (2008). Class information predicts activation by object fragments in human object areas. *J. Cogn. Neurosci.* 20, 1189–1206.
- Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* 60, 91–110.
- Riesenhuber, M., and Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nat. Neurosci.* 2, 1019–1025.
- Rolls, E. T. (2008). *Memory, Attention, and Decision-Making: A Unifying Computational Neuroscience Approach*. Oxford; New York, NY: Oxford University Press.
- Rust, N. C., and Dicarlo, J. J. (2010). Selectivity and tolerance ("invariance") both increase as visual information propagates from cortical area V4 to IT. *J. Neurosci.* 30, 12978–12995.
- Schyns, P. G., Goldstone, R. L., and Thibaut, J. P. (1998). The development of features in object concepts. *Behav. Brain Sci.* 21, 1–17. discussion: 17–54.
- Ullman, S. (2007). Object recognition and segmentation by a fragment-based hierarchy. *Trends Cogn. Sci.* 11, 58–64.
- Ullman, S., and Bart, E. (2004). Recognition invariance obtained by extended and invariant features. *Neural Netw.* 17, 833–848.
- Ullman, S., Vidal-Naquet, M., and Sali, E. (2002). Visual features of intermediate complexity and their use in classification. *Nat. Neurosci.* 5, 682–687.
- Vuilleumier, P., Henson, R. N., Driver, J., and Dolan, R. J. (2002). Multiple levels of visual object constancy revealed by event-related fMRI of repetition priming. *Nat. Neurosci.* 5, 491–499.
- Wallis, G., Backus, B. T., Langer, M., Huebner, G., and Bulthoff, H. (2009). Learning illumination- and orientation-invariant representations of objects through temporal association. *J. Vis.* 9, 6.
- Wallis, G., and Bulthoff, H. (1999). Learning to recognize objects. *Trends Cogn. Sci.* 3, 22–31.
- Walsh, V., and Kulikowski, J. (eds.). (1998). *Perceptual Constancy: Why Things Look As They Do*. New York, NY: Cambridge University Press.
- Wong, A. C., Palmeri, T. J., Rogers, B. P., Gore, J. C., and Gauthier, I. (2009). Beyond shape: how you learn about objects affects how they are represented in visual cortex. *PLoS ONE* 4:e8405. doi: 10.1371/journal.pone.0008405

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 25 January 2012; accepted: 17 July 2012; published online: 24 August 2012.

Citation: Bart E and Hegdé J (2012) Invariant object recognition based on extended fragments. *Front. Comput. Neurosci.* 6:56. doi: 10.3389/fncom.2012.00056

Copyright © 2012 Bart and Hegdé. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in other forums, provided the original authors and source are credited and subject to any copyright notices concerning any third-party graphics etc.



# Learning and disrupting invariance in visual recognition with a temporal association rule

Leyla Isik<sup>\*†</sup>, Joel Z. Leibo<sup>†</sup> and Tomaso Poggio

Center for Biological and Computational Learning, McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge, MA, USA

## Edited by:

Evgeniy Bart, Palo Alto Research Center, USA

## Reviewed by:

Peter König, University of

Osnabrück, Germany

Jay Hegdè, Georgia Health Sciences University, USA

## \*Correspondence:

Leyla Isik, Center for Biological and Computational Learning, McGovern Institute for Brain Research, Massachusetts Institute of Technology, Bldg. 46-5155, 77 Massachusetts Avenue, Cambridge, MA, USA.  
e-mail: lisik@mit.edu

<sup>†</sup> These authors contributed equally to this work.

Learning by temporal association rules such as Foldiak's trace rule is an attractive hypothesis that explains the development of invariance in visual recognition. Consistent with these rules, several recent experiments have shown that invariance can be broken at both the psychophysical and single cell levels. We show (1) that temporal association learning provides appropriate invariance in models of object recognition inspired by the visual cortex, (2) that we can replicate the "invariance disruption" experiments using these models with a temporal association learning rule to develop and maintain invariance, and (3) that despite dramatic single cell effects, a population of cells is very robust to these disruptions. We argue that these models account for the stability of perceptual invariance despite the underlying plasticity of the system, the variability of the visual world and expected noise in the biological mechanisms.

**Keywords:** object recognition, invariance, vision, trace rule, cortical models, inferotemporal cortex, visual development

## 1. INTRODUCTION

A single object can give rise to a wide variety of images. The pixels (or photoreceptor activations) that make up an image of an object change dramatically when the object is moved relative to its observer. Despite these large changes in sensory input, the brain's ability to recognize objects is relatively unimpeded. Temporal association methods are promising solutions to the problem of how to build computer vision systems that achieve similar feats of invariant recognition (Foldiak, 1991; Wallis and Rolls, 1997; Wiskott and Sejnowski, 2002; Einhauser et al., 2005; Spratling, 2005; Wyss et al., 2006; Franzius et al., 2007; Masquelier and Thorpe, 2007; Masquelier et al., 2007). These methods associate temporally adjacent views under the assumption that temporal adjacency is usually a good cue that two images are of the same object. For example, an eye movement from left to right causes an object to translate on the visual field from right to left; under such a rule, the cells activated by the presence of the object on the right will be linked with the cells activated by the presence of the object on the left. This linkage can be used to signal that the two views represent the same object—despite its change in retinal position.

Recent experimental evidence suggests that the brain may also build invariance with this method. Furthermore, the natural temporal association-based learning rule remains active even after visual development is complete (Wallis and Bulthoff, 2001; Cox et al., 2005; Li and DiCarlo, 2008, 2010; Wallis et al., 2009). This paper addresses the wiring errors that must occur with such a continually active learning rule due to regular disruptions of temporal contiguity (from lighting changes, sudden occlusions, or biological imperfections, for example).

Experimental studies of temporal association involve putting observers in an altered visual environment where objects change identity across saccades. Cox et al. (2005) showed that after about an hour of exposure to an altered environment, where objects changed identity at a specific retinal position, the subjects mistook one object for another at the swapped position while preserving their ability to discriminate the same objects at other positions. A subsequent physiology experiment by Li and DiCarlo using a similar paradigm showed that individual neurons in primate anterior inferotemporal cortex (AIT) change their selectivity in a position-dependent manner after less than an hour of exposure to the altered visual environment (Li and DiCarlo, 2008).

The Li and DiCarlo experiment did not include a behavioral readout, so the effects of the manipulation on the monkey's perception are not currently known, however, the apparent robustness of our visual system suggests it is highly unlikely that the monkey would really be confused between such different looking objects (e.g., a teacup and a sailboat) after such a short exposure to the altered visual environment. In contrast, the Cox et al. psychophysics experiment had a similar timecourse (a significant effect was present after 1 h of exposure) but used much more difficult to discriminate objects ("Greebles" Gauthier and Tarr, 1997).

In this paper, we describe a computational model of invariance learning that shows how strong effects at the single cell level, like those observed in the experiments by Li and DiCarlo do not necessarily cause confusion on the neural population level, and hence do not imply perceptual effects. Our

simulations show that a population of cells is surprisingly robust to large numbers of mis-wirings due to errors of temporal association.

## 2. MATERIALS AND METHODS

### 2.1. HIERARCHICAL MODELS OF OBJECT RECOGNITION

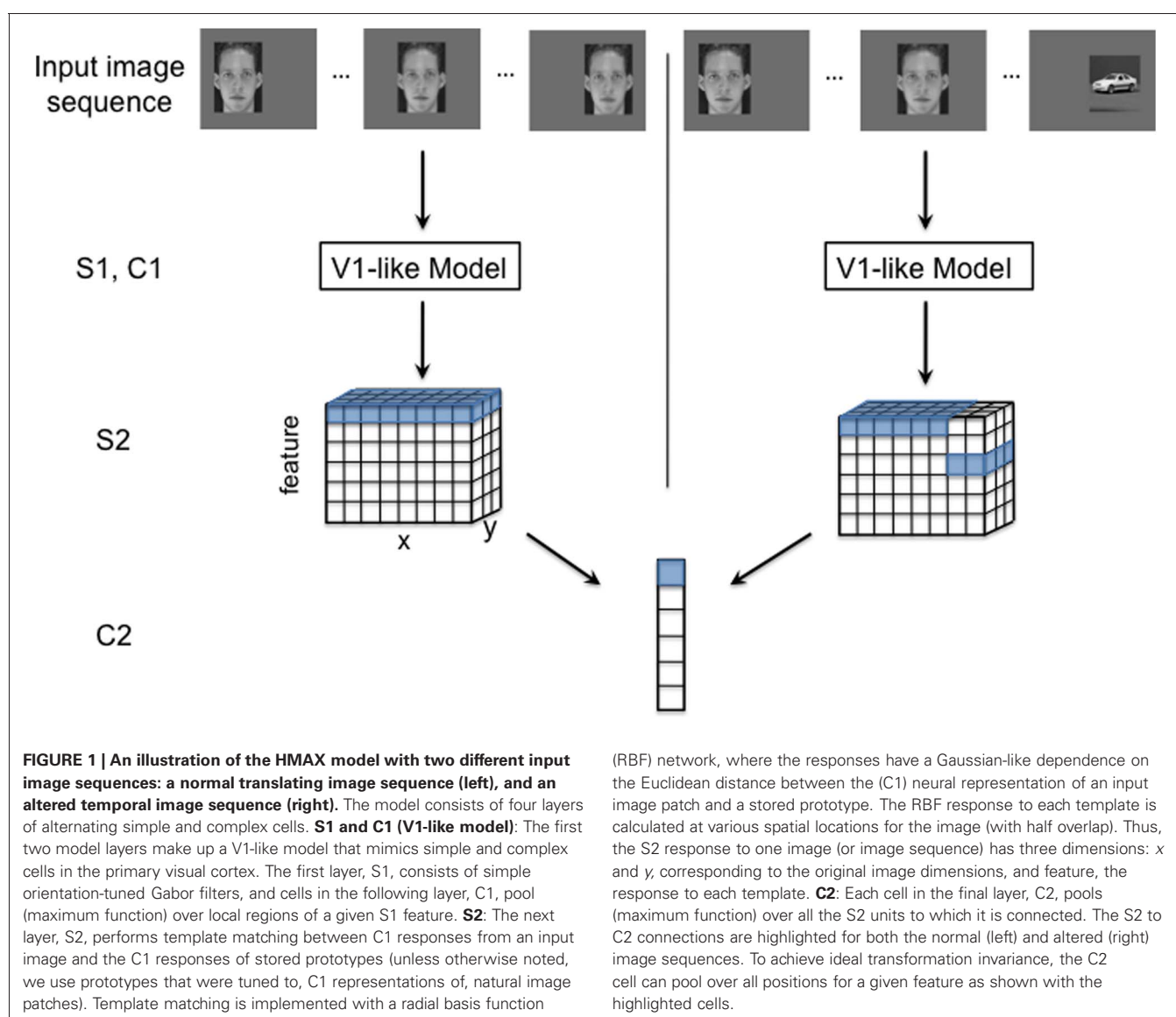
We examine temporal association learning with a class of cortical models inspired by Hubel and Wiesel's famous studies of visual cortex (Hubel and Wiesel, 1962). These models contain alternating layers of simple S cells or feature detectors to build specificity, and complex C cells that pool over simple cells to build invariance (Fukushima, 1980; Riesenhuber and Poggio, 1999; Serre et al., 2007). We will focus on one particular such model, HMAX (Serre et al., 2007). The differences between these models are likely irrelevant to the issue we are studying, and thus our results will generalize to other models in this class.

### 2.2. THE HMAX MODEL

In this model, simple (S) cells compute a measure of their input's similarity to a stored optimal feature via a gaussian radial basis function (RBF) or a normalized dot product. Complex (C) cells pool over S cells by computing the *max* response of all the S cells with which they are connected. These operations are typically repeated in a hierarchical manner, with the output of one C layer feeding into the next S layer and so on. The model used in this report had four layers:  $S1 \rightarrow C1 \rightarrow S2 \rightarrow C2$ . The caption of **Figure 1** gives additional details of the model's structure.

In our implementation of the HMAX model, the response of a C2 cell—associating templates  $w$  at each position  $t$ —is given by:

$$r_w(x) = \max_t \left( \exp \left( -\frac{1}{2\sigma} \sum_{j=1}^n (w_{t,j} - x_j)^2 \right) \right) \quad (1)$$





In the hardwired model, each template  $w_t$  is replicated at all positions, thus the C2 response models the outcome of a previous temporal association learning process that associated the patterns evoked by a template at each position. The C2 responses of the hardwired model are invariant to translation (Serre et al., 2007; Leibo et al., 2010). The remainder of this report is focused on the model with learned pooling domains. Section 2.3 describes the learning procedure and **Figure 2** compares the performance of the hardwired model to an HMAX model with learned C2 pooling domains.

As in Serre et al. (2007), we typically obtain S2 templates from patches of natural images (except where noted in **Figure 3**). The focus of this report is on learning the pooling domains. The choice of templates, i.e., the learning of selectivity (as opposed to invariance) is a separate issue with a large literature of its own<sup>1</sup>.

### 2.3. TEMPORAL ASSOCIATION LEARNING

Temporal association learning rules provide a plausible way to learn transformation invariance through natural visual experience (Foldiak, 1991; Wallis and Rolls, 1997; Wiskott and Sejnowski, 2002; Einhauser et al., 2005; Spratling, 2005; Wyss et al., 2006; Franzius et al., 2007; Masquelier and Thorpe, 2007; Masquelier et al., 2007). Objects typically move in and out of our visual field much slower than they transform due to changes

in pose and position. Based on this difference in timescale we can group together cells that are tuned to the same object under different transformations.

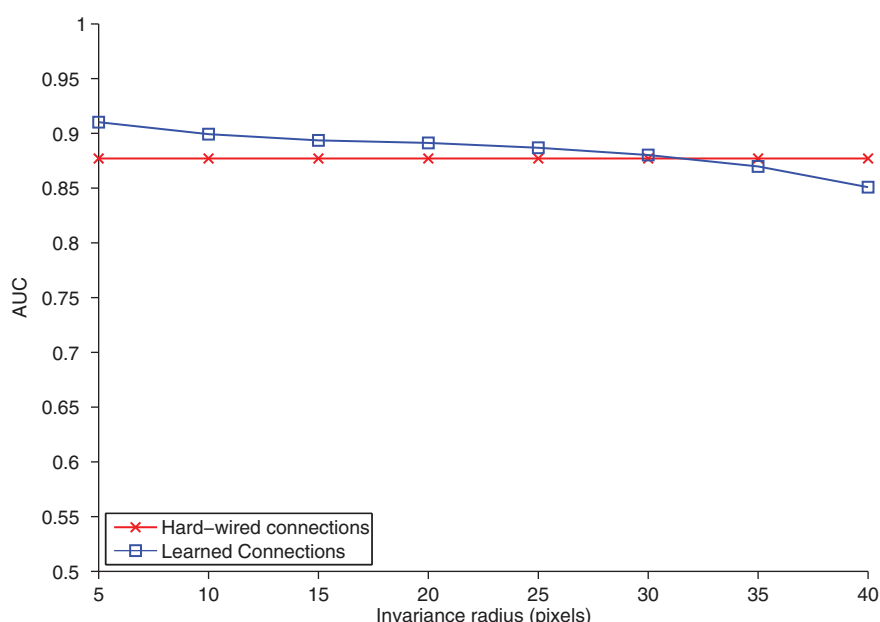
Our model learns translation invariance from a sequence of images of continuously translating objects. During a training phase prior to each simulation, the model's S2 to C2 connections are learned by associating the patterns evoked by adjacent images in the training sequence as shown in **Figure 1**, left.

The training phase is divided into temporal association periods. During each temporal association period the highly active S2 cells become connected to the same C2 cell. One C2 cell is learned during each association period. When modeling "standard" (undisrupted) visual experience, as in **Figure 2**, each association period contains all views of a single object at each retinal position. If temporally adjacent images really depict the same object at different positions, then this procedure will group all the S2 cells that were activated by viewing the object, no matter what spatial location elicited the response. The outcome of this learning procedure in one association period is illustrated in **Figure 1**, left. The C2 cell produced by this process pools over its connected S2 cells. The potential effect of a temporally altered image sequence is illustrated in **Figure 1**, right. This altered training will likely result in mis-wirings between the S2 and C2 neurons, which could ultimately alter the system's performance.

<sup>1</sup>See Leibo et al. (2010) for a discussion of the impact of template-choice on HMAX results with a similar translation-invariant recognition task to the one used here.

#### 2.3.1. Learning rule

In Foldiak's original trace rule, shown in **Equation 2**, the weight of a synapse  $w_{ij}$  between an input cell  $x_j$  and output cell  $y_i$  is



**FIGURE 2 | The area under the ROC curve (AUC) (ordinate) plotted for the task of classifying (nearest neighbors) objects appearing on an interval of increasing distance from the reference position (abscissa).** The model was trained and tested on separate training and testing sets, each with 20 car and 20 face images. For temporal association learning, one C2 unit is learned for each association period or training image, yielding 40 learned C2 units.

One hard-wired C2 unit was learned from each natural image patch that S2 cells were tuned to, yielding 10 hard-wired C2 units. Increasing the number of hard-wired features has only a marginal effect on classification accuracy. For temporal association learning, the association period  $\tau$  was set to the length of each image sequence (12 frames), and the activation threshold  $\theta$  was empirically set to 3.9 standard deviations above the mean activation.

strengthened proportionally to the input activity and the trace or average of recent output activity at time  $t$ . The dependence of the trace on previous activity decays over time with the  $\delta$  term (Foldiak, 1991).

Foldiak trace rule:

$$\begin{aligned}\Delta w_{ij}^{(t)} &\propto x_j \bar{y}_i^{(t)} \\ \bar{y}_i^{(t)} &= (1 - \delta) \bar{y}_i^{(t-1)} + \delta y_i^{(t)}\end{aligned}\quad (2)$$

In the HMAX model, connections between S and C cells are binary. Additionally, in our training case we want to learn connections based on image sequences of a known length, and thus for simplicity should include a hard time window rather than a decaying time dependence. Thus we employed a modified trace rule that is appropriate for learning S2 to C2 connections in the HMAX model.

Modified trace rule for the HMAX model:

$$\begin{aligned}\text{for } t \text{ in } \tau : \\ \text{if } x_j > \theta, \quad w_{ij} = 1 \\ \text{else,} \quad w_{ij} = 0\end{aligned}\quad (3)$$

With this learning rule, one C2 cell with index  $i$  is produced for each association period. The length of the association period is  $\tau$ .

### 3. RESULTS

#### 3.1. TRAINING FOR TRANSLATION INVARIANCE

We model natural invariance learning with a training phase where the model learns to group different representations of a given object based on the learning rule in **Equation 3**. Through the learning rule, the model groups continuously translating images that move across the field of view over each association period  $\tau$ . An example of a translating image sequence is shown at the top, left of **Figure 1**. During this training phase, the model learns the domain of pooling for each C2 cell.

#### 3.2. ACCURACY OF TEMPORAL ASSOCIATION LEARNING

To test the performance of the HMAX model with the learning rule in **Equation 3**, we train the model with a sequence of training images. Next, we compare the learned model's performance to that of the hard-wired HMAX (Serre et al., 2007) on a translation-invariant recognition task. In standard implementations of the HMAX model, the S2 to C2 connections are hard-wired, each C2 cell pools all the S2 responses for a given template globally over all spatial locations. This pooling gives the model translation invariance and mimics the outcome of an idealized temporal association process.

The task is a 20 face and 20 car identification task, where the target images are similar (but not identical) for different translated views<sup>2</sup>. We collect hard-wired C2 units and C2 units

learned from temporal sequences of the faces and cars. We then used a nearest neighbor classifier to compare the correlation of C2 responses for translated objects to those in a given reference position. The accuracy of the two methods (hard-wired and learned from test images) versus translation is shown in **Figure 2**. The two methods performed equally well. This confirms that the temporal associations learned from this training yield correct invariance.

#### 3.3. MANIPULATING THE TRANSLATION INVARIANCE OF A SINGLE CELL

In their physiology experiments Li and DiCarlo identified AIT cells that responded preferentially to one object over another, they then performed altered temporal association training where the two objects were swapped at a given position (Li and DiCarlo, 2008). To model these experiments we perform temporal association learning (described by **Equation 3**) with a translating image of one face and one car. For this simulation, the S2 units are tuned to the same face and car images (see **Figure 1** caption) to mimic object-selective cells that are found in AIT. Next we select a "swap position" and perform completely new, altered training with the face and car images swapped only at that position (see **Figure 1**, top right). After the altered training, we observe the response (of one C2 cell) to the two objects at the swap position and another non-swap position in the visual field that was unaltered during training.

As shown in **Figure 3**, the C2 response for the preferred object at the swap position (but not the non-swap position) is lower after training, and the C2 response to the non-preferred object is higher at the swap position. As in the physiology experiments performed by Li and DiCarlo, these results are object and position specific. Though unsurprising, this result draws a parallel between the response of a single C2 unit and the physiological response of a single cell.

#### 3.4. INDIVIDUAL CELL VERSUS POPULATION RESPONSE

In the previous section we modeled the single cell results of Li and DiCarlo, namely that translation-invariant representations of objects can be disrupted by a relatively small amount of exposure to altered temporal associations. However, single cell changes do not necessarily reflect whole population or perceptual behavior and no behavioral tests were performed on the animals in this study.

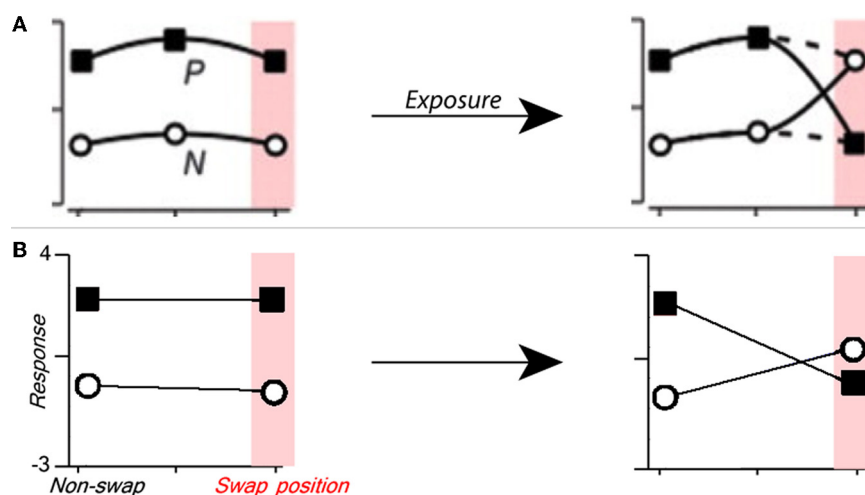
A cortical model with a temporal association learning rule provides a way to model population behavior with swap exposures similar to the ones used by Li and DiCarlo (2008, 2010). A C2 cell in the HMAX model can be treated as analogous to an AIT cell (as tested by Li and DiCarlo), and a C2 vector as a population of these cells. We can thus apply a classifier to this cell population to obtain a model of behavior or perception.

#### 3.5. ROBUSTNESS OF TEMPORAL ASSOCIATION LEARNING WITH A POPULATION OF CELLS

We next model the response of a population of cells to different amounts of swap exposure, as illustrated in **Figure 1**, right. The translating image sequence with which we train the model replicates visual experience, and thus jumbling varying amounts of these training images is analogous to presenting different

<sup>2</sup>The invariance-training and testing datasets come from a concatenation of two datasets from: ETH80 (<http://www.d2.mpi-inf.mpg.de/Datasets/ETH80>) and ORL (<http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html>). Except when noted, the image patches used to obtain the S2 templates were obtained from a different, unrelated, collection of natural images; see Serre et al. (2007) for details.





**FIGURE 3 | Manipulating single cell translation invariance through altered visual experience. (A)** Figure from Li and DiCarlo (2008)

summarizing the expected results of swap exposure on a single cell. *P* is the response to preferred stimulus, and *N* is that to non-preferred stimulus.

**(B)** The response of a C2 cell tuned to a preferred object before (left) and after (right) altered visual training where the preferred and non-preferred

objects were swapped at a given position. To model the experimental paradigm used in Wallis and Bulthoff (2001), Cox et al. (2005), and Li and DiCarlo (2008, 2010), altered training and final testing were performed on the same altered image sequence. The C2 cell's relative response (Z-score) to the preferred and non-preferred objects is shown on the ordinate, and the position (swap or non-swap) is shown on the abscissa.

amounts of altered exposure to a test subject as in (Li and DiCarlo, 2008, 2010). These disruptions also model the mis-associations that may occur with temporal association learning due to sudden changes in the visual field (such as light, occlusions, etc.), or other imperfections of the biological learning mechanism. During each training phase we randomly swap different face and car images in the image sequences with a certain probability, and observe the effect on the response of a classifier to a population of C2 cells. The performance, as measured by area under the ROC curve (AUC), versus different neural population sizes (number of C2 cells) is shown in **Figure 4** for several amounts of altered exposure. We measured altered exposure by the probability of flipping a face and car image in the training sequence.

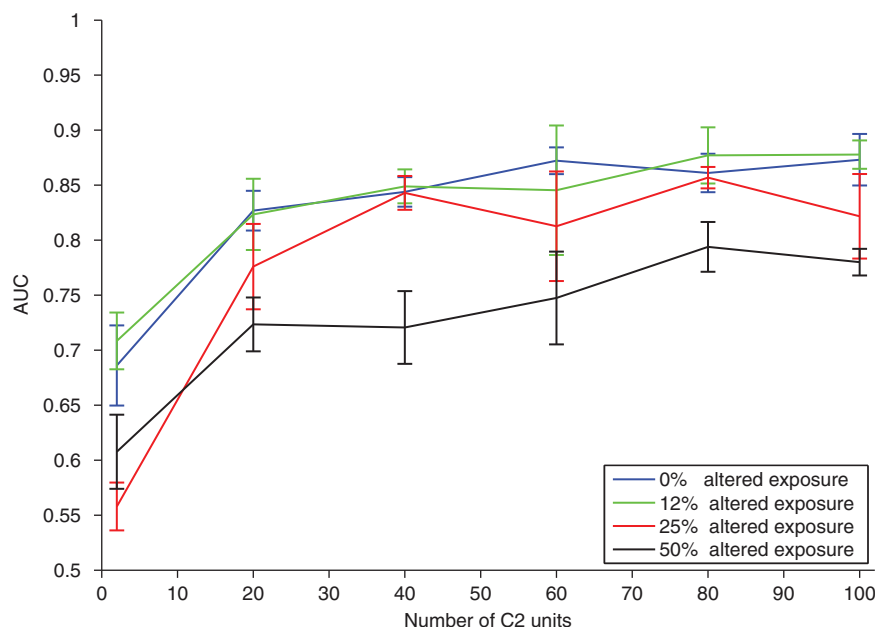
A small amount of exposure to altered temporal training (0.125 probability of flipping each face and car) has negligible effects, and the model under this altered training performs as well as with normal temporal training. A larger amount of exposure to altered temporal training (0.25 image flip probability) is not significantly different than perfect temporal training, especially if the neural population is large enough. With enough C2 cells, each of which is learned from a temporal training sequence, the effects of small amounts of jumbling in training images are insignificant. Even with half altered exposure (0.5 image flip probability), if there are enough C2 cells, then classification performance is still reasonable. This is likely because with similar training (multiple translating faces or cars) redundant C2 cells are formed, creating robustness to association errors that occurred during altered training. Similar redundancies are likely to occur in natural vision. This indicates that in natural learning mis-wirings do not have a strong effect on learning translation invariance, particularly with familiar objects or tasks.

#### 4. DISCUSSION

We use a cortical model inspired by Hubel and Wiesel (1962), where translation invariance is learned through a variation of Foldiak's trace rule (Foldiak, 1991) to model the visual response to altered temporal exposure. We first show that this temporal association learning rule is accurate by comparing its performance to that of a similar model with hard-wired translation invariance (Serre et al., 2007). This extends previous modeling results by Masquelier et al. (2007) for models of V1 to higher levels in the visual recognition architecture. Next, we test the robustness of translation invariance learning on single cell and whole population responses. We show that even if single cell translation invariance is disrupted, the whole population is robust enough to maintain invariance despite a large number of mis-wirings.

The results of this study provide insight into the evolution and development of transformation invariance mechanisms in the brain. It is unclear why a translation invariance learning rule, like the one we modeled, and those confirmed by Cox et al. (2005) and Li and DiCarlo (2008, 2010), would remain active after development. We have shown that the errors associated with a continuously active learning rule are negligible, and thus it may be simpler to leave these processes active than to develop a mechanism to turn them off.

Extending this logic to other transformations is interesting. Translation is a *generic* transformation; all objects translate in the same manner, so translation invariance, in principle, can be learned during development for all types of objects. This is not true of "non-generic" or *class-specific* transformations, such as rotation in depth, which depends on the 3-D structure of an individual object or class of objects (Vetter et al., 1995; Leibo et al., 2010, 2011). For example, knowledge of how 2-D images of faces rotate in depth can be used to predict how a new face will



**FIGURE 4 | Results of a translation invariance task ( $\pm 40$  pixels) with varying amounts of altered visual experience.** To model the experimental paradigm used in (Wallis and Bulthoff, 2001; Cox et al., 2005; Li and DiCarlo, 2008, 2010; Wallis et al., 2009), training and testing were performed on the same altered image sequence. The performance (AUC) on the same translation-invariant recognition task

as in **Figure 2**, with a nearest neighbor classifier, versus the number of C2 units. Different curves have a different amount of exposure to altered visual training as measured by the probability of swapping a car and face image during training. The error bars show  $\pm$  one standard deviation over runs using different natural image patches as S2 templates.

appear after a rotation. However, knowledge of how faces rotate is not useful for predicting the appearance of non-face objects after the same 3-D transformation. Many transformations are class-specific in this sense<sup>3</sup>. One hypothesis as to why invariance-learning mechanisms remain active in the mature visual system could be a continuing need to learn and refine invariant representations for more objects under non-generic transformations.

Disrupting rotation in depth has been studied in psychophysics experiments. Wallis and Bulthoff showed that training subjects with slowly morphing faces, disrupts viewpoint invariance after only a few instances of altered training (Wallis and Bulthoff, 2001; Wallis et al., 2009). This effect occurs with a faster time course than observed in the translation invariance experiments (Cox et al., 2005). One possible explanation for this time discrepancy is that face processing mechanisms are higher-level than those for the “greeble objects” and thus easier to disrupt. However, we conjecture that the strong, fast effect has to do with the type of transformation rather than the specific class of stimuli.

Unlike generic transformations, class-specific transformations cannot be generalized between objects with different properties. It is even possible that we learn non-generic transformations of novel objects through a memory-based architecture that requires the visual system to store each viewpoint of a novel

object. Therefore, it is logical that learning rules for non-generic transformations should remain active as we are exposed to new objects throughout life.

In daily visual experience we are exposed more to translations than rotations in depth, so through visual development or evolutionary mechanisms there may be more cells dedicated to translation-invariance than rotation-invariance. We showed that the size of a population of cells has a significant effect on its robustness to altered training, see **Figure 4**. Thus rotation invariance may also be easier to disrupt, because there could be fewer cells involved in this process.

Two plausible hypotheses both point to rotation (class-specific) versus translation (generic) being the key difference between the Wallis and Bulthoff and Cox et al. experiments. We conjecture that if an experiment controlled for variables such as the type and size of the stimulus, class-specific invariances would be easier to disrupt than generic invariances.

This study shows that despite unavoidable disruptions, models based on temporal association learning are quite robust and therefore provide a promising solution for learning invariance from natural vision. These models will also be critical in understanding the interplay between the mechanisms for developing different types of transformation invariance.

## ACKNOWLEDGMENTS

This work was supported by the following grants: NSF-0640097, NSF-0827427, NSF-0645960, DARPA-DSO, AFSOR FA8650-50-C-7262, AFSOR FA9550-09-1-0606.

<sup>3</sup>Changes in illumination are another example of a class-specific transformation. These depend on both 3-D structure and material properties of objects (Leibo et al., 2011).

## REFERENCES

- Cox, D., Meier, P., Oertelt, N., and DiCarlo, J. J. (2005). 'Breaking' position-invariant object recognition. *Nat. Neurosci.* 8, 1145–1147.
- Einhauser, W., Hipp, J., Eggert, J., Korner, E., and Konig, P. (2005). Learning viewpoint invariant object representations using a temporal coherence principle. *Biol. Cybern.* 93, 79–90.
- Foldiak, P. (1991). Learning invariance from transformation sequences. *Neural Comput.* 3, 194–200.
- Franzius, M., Sprekeler, H., and Wiskott, L. (2007). Slowness and sparseness lead to place, head-direction, and spatial-view cells. *PLoS Comput. Biol.* 3:e166. doi: 10.1371/journal.pcbi.0030166
- Fukushima, K. (1980). Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol. Cybern.* 36, 193–201.
- Gauthier, I., and Tarr, M. (1997). Becoming a "greeble" expert: exploring mechanisms for face recognition. *Vision Res.* 37, 1673–1682.
- Hubel, D. H., and Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cats visual cortex. *J. Physiol.* 160, 106–154.
- Leibo, J. Z., Mutch, J., and Poggio, T. (2011). "Why the brain separates face recognition from object recognition," in *Advances in Neural Information Processing Systems (NIPS)*, (Cambridge, MA).
- Leibo, J. Z., Mutch, J., Rosasco, L., Ullman, S., and Poggio, T. (2010). Learning generic invariances in object recognition: translation and scale. MIT-CSAIL-TR-2010–2061.
- Li, N., and DiCarlo, J. J. (2008). Unsupervised natural experience rapidly alters invariant object representation in visual cortex. *Science* 321, 1502–1507.
- Li, N., and DiCarlo, J. J. (2010). Unsupervised natural visual experience rapidly reshapes size-invariant object representation in inferior temporal cortex. *Neuron* 67, 1062–1075.
- Masquelier, T., Serre, T., Thorpe, S., and Poggio, T. (2007). Learning complex cell invariance from natural videos: a plausible proof. MIT-CSAIL-TR-2007–2060.
- Masquelier, T., and Thorpe, S. J. (2007). Unsupervised learning of visual features through spike timing dependent plasticity. *PLoS Comput. Biol.* 3:e31. doi: 10.1371/journal.pcbi.0030031
- Riesenhuber, M., and Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nat. Neurosci.* 2, 1019–1025.
- Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., and Poggio, T. (2007). Robust object recognition with cortex-like mechanisms. *IEEE Trans. Pattern Anal. Mach. Intell.* 29, 411–426.
- Spratling, M. (2005). Learning viewpoint invariant perceptual representations from cluttered images. *IEEE Trans. Pattern Anal. Mach. Intell.* 27, 753–761.
- Vetter, T., Hurlbert, A., and Poggio, T. (1995). View-based models of 3D object recognition: invariance to imaging transformations. *Cereb. Cortex* 3, 261–269.
- Wallis, G., Backus, B. T., Langer, M., Huebner, G., and Bulthoff, H. (2009). Learning illumination- and orientation-invariant representations of objects through temporal association. *J. Vis.* 9, 1–8.
- Wallis, G., and Bulthoff, H. (2001). Effects of temporal association on recognition memory. *Proc. Natl. Acad. Sci. U.S.A.* 98, 4800–4804.
- Wallis, G., and Rolls, E. T. (1997). Invariant face and object recognition in the visual system. *Prog. Neurobiol.* 51, 167–194.
- Wiskott, L., and Sejnowski, T. J. (2002). Slow feature analysis: unsupervised learning of invariances. *Neural Comput.* 14, 715–770.
- Wyss, R., Konig, P., and Verschure, P. (2006). A model of the ventral visual system based on temporal stability and local memory. *PLoS Biol.* 4:e120. doi: 10.1371/journal.pbio.0040120

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 01 November 2011; accepted: 27 May 2012; published online: 25 June 2012.

Citation: Isik L, Leibo JZ and Poggio T (2012) Learning and disrupting invariance in visual recognition with a temporal association rule. *Front. Comput. Neurosci.* 6:37. doi: 10.3389/fncom.2012.00037

Copyright © 2012 Isik, Leibo and Poggio. This is an open-access article distributed under the terms of the Creative Commons Attribution Non Commercial License, which permits non-commercial use, distribution, and reproduction in other forums, provided the original authors and source are credited.



# Transformation-invariant visual representations in self-organizing spiking neural networks

Benjamin D. Evans\* and Simon M. Stringer

Department of Experimental Psychology, Centre for Theoretical Neuroscience and Artificial Intelligence, University of Oxford, Oxford, UK

## Edited by:

Evgeniy Bart, Palo Alto Research Center, USA

## Reviewed by:

Evgeniy Bart, Palo Alto Research Center, USA  
Jay Hegd , Georgia Health Sciences University, USA

## \*Correspondence:

Benjamin D. Evans, Department of Experimental Psychology, Centre for Theoretical Neuroscience and Artificial Intelligence, University of Oxford, South Parks Road, Oxford, OX1 3UD, UK.  
e-mail: benjamin.evans@psy.ox.ac.uk

The ventral visual pathway achieves object and face recognition by building transformation-invariant representations from elementary visual features. In previous computer simulation studies with rate-coded neural networks, the development of transformation-invariant representations has been demonstrated using either of two biologically plausible learning mechanisms, Trace learning and Continuous Transformation (CT) learning. However, it has not previously been investigated how transformation-invariant representations may be learned in a more biologically accurate spiking neural network. A key issue is how the synaptic connection strengths in such a spiking network might self-organize through Spike-Time Dependent Plasticity (STDP) where the change in synaptic strength is dependent on the relative times of the spikes emitted by the presynaptic and postsynaptic neurons rather than simply correlated activity driving changes in synaptic efficacy. Here we present simulations with conductance-based integrate-and-fire (IF) neurons using a STDP learning rule to address these gaps in our understanding. It is demonstrated that with the appropriate selection of model parameters and training regime, the spiking network model can utilize either Trace-like or CT-like learning mechanisms to achieve transform-invariant representations.

**Keywords:** transformation-invariant visual object recognition, integrate and fire, spiking neural net, continuous transformation learning, trace learning, inferior temporal cortex

## 1. INTRODUCTION

The increasingly complex cell response properties of the primate ventral visual stream strongly suggest the functional organization of this pathway is that of a feature hierarchy. Cells in the early stages (V1) are found to be sensitive to oriented bars and edges appearing in particular locations on the retina (Hubel and Wiesel, 1968). Information analysis of natural scenes reveals these features to be the most statistically independent components of such images (Bell and Sejnowski, 1997; van Hateren and van der Schaaf, 1998) and hence the most natural “building-blocks” for such a system. Through successive layers, there follows a convergence of receptive fields allowing neurons at the end of the pathway in anterior Inferotemporal cortex (aIT) to view the entire retina and respond to increasingly complex stimuli (Tanaka, 1996). Here, and more recently in the medial temporal lobe (Quiroga et al., 2005), neurons have been found which respond with translation (Op de Beeck and Vogels, 2000), size (Ito et al., 1995) and view invariance (Booth and Rolls, 1998) to objects (Tanaka et al., 1991) and faces (Desimone, 1991).

Several groups have attempted to understand how elementary features may be combined into more complex view-invariant representations of whole objects with hierarchical feed-forward neural network models such as the *Neocognitron* (Fukushima, 1988), the *SEEMORE* system (Mel, 1997), the *HMAX* model (Riesenhuber and Poggio, 1999) and *VisNet* (Wallis and Rolls, 1997). These models are all composed of “rate-coded” neurons (McCulloch and Pitts, 1943) which consist of applying a non-linear function (e.g., threshold or sigmoid) to a weighted sum

of inputs (Boolean, or real values) which they receive at each computational step<sup>1</sup>.

Within this paradigm, two main biologically plausible learning mechanisms have been discovered which explain how different views of the same object may be bound together and recognized as the same entity. The first of these—*Trace learning* (Földi k, 1991)—relies upon temporal continuity, while the second—*Continuous Transformation (CT) learning* (Stringer et al., 2006)—relies upon spatial continuity to associate together successive transforms and build view-invariant representations in later layers. While the properties of these mechanisms have been explored extensively in rate-coded models, it remains an open question as to how they might map onto a more biologically realistic spiking-neuron paradigm.

Spiking Neural Networks (SNN) can solve problems at least as complex as those that rate-coded models can solve (  ma and Orponen, 2003), which in turn have greater computational power than Turing machines, and as such have been applied to a wide variety of problems, including modeling object recognition (Michler et al., 2009). By more faithfully modeling the electrical properties of neurons, spiking neural network model parameters may be more meaningfully mapped onto the biophysical properties of their real counterparts. This motivates the use

<sup>1</sup>These early neuron models were designed to show that the elementary components of the brain could compute elementary logic functions. The belief commonly held at the time being that intelligence is based upon symbolic reasoning, which in turn rests upon the foundations of logic.

of the conductance-based “leaky” integrate-and-fire (LIF) model (described in section 2) over models which are computationally cheaper or have a less apparent correspondence to measurable biological parameters such as the Spike Response Model (Gerstner and Kistler, 2006) or Izhikevich’s null-cline derived model (2003).

Since time is explicitly and accurately modeled in SNNs, they allow quantitative investigation of the time-course of processing on such tasks (Thorpe et al., 2000) providing further arguments against rate-coding on the basis that Poisson rate-codes are too inefficient to account for the rapidity of information processing in the human visual system<sup>2</sup> (Thorpe et al., 1996; Rullen and Thorpe, 2001). Furthermore, SNNs allow the investigation of qualitative effects such as the selective representation of one stimulus over another by the synchronization of its population of feature-neurons as found in neurophysiological studies (Kreiter and Singer, 1996; Fries et al., 2002). Similarly, the phenomenon of Spike-Time Dependent Plasticity (STDP) and its effect upon learning transformation-invariant representations may only be investigated by modeling individual spikes which is of great importance to the present research.

Hebb originally conjectured that synapses effective at evoking a response should grow stronger (Hebb, 1949), capturing a causal relationship between the two neurons. This was eventually simplified (partly for the purposes of rate-coded models) to become interpreted as any long-lasting synapse-specific form of modification dependent upon correlations between presynaptic and postsynaptic firing. This is usually expressed in the form  $\delta w_{ij} = k y_i x_j$ , where  $\delta w_{ij}$  is the change in synaptic strength,  $k$  is a learning rate constant, and  $x_j$  and  $y_i$  are the firing rates of the presynaptic and postsynaptic neurons (see e.g., Rolls and Treves, 1998).

Progress in neurophysiology has shown, however, that the all-or-nothing nature of an action potential means that the information may be conveyed by the number *and* the timing of action potentials (Ferster and Spruston, 1995; Maass and Bishop, 1999), typically neglecting their size and shape in modeling. In other words neurons communicate by a *pulse* code (a time series of discrete binary events) rather than simply a *rate* code (a moving average level of activity) which has been convincingly demonstrated in the sensory systems of several organisms, such as echolocating bats (Kuwabara and Suga, 1993) and the visual systems of flies (Bialek et al., 1991).

It is also now well-established that *synaptic plasticity* is sensitive to the relative timing of the presynaptic and postsynaptic spikes (Markram et al., 1997; Dan and Poo, 2006), typically becoming approximately exponentially less sensitive as the time difference increases (Bi and Poo, 1998). This has been found to take several forms in different brain regions (Abbott and Nelson, 2000) but here we focus on the form observed in retinotectal connections and neocortical and hippocampal pyramidal cells where *pre* → *post* spike pairs lead to synaptic potentiation (with greater effect over shorter intervals) and the opposite ordering of spikes leads to synaptic depression.

The challenge now is to investigate how the timing of spikes affects the self-organization of the system applied to the

problem of developing transformation-invariant representations and understanding how the CT and Trace learning mechanisms, which have been developed in the context of rate-coded models, might fit into a model of STDP.

## 2. METHODS

### 2.1. NETWORK ARCHITECTURE

While the ventral visual stream is typically modeled as four or more layers of neurons with excitatory modifiable feed-forward synapses and a mechanism of lateral inhibition, here we seek to understand the mechanisms operating at each layer which ultimately may lead to transformation-invariant representations, hence a simpler architecture is used.

The model consists of two layers of excitatory pyramidal neurons with one layer of modifiable feed-forward synapses between them (as shown in **Figure 1**). Within each layer there are also inhibitory interneurons with non-plastic lateral synaptic connections to and from the excitatory neurons to produce a degree of competition between the excitatory neurons.

For all presented simulations we have used 400 excitatory neurons and 100 inhibitory neurons in each layer, with full connectivity. Each neuron is based upon the standard conductance-based leaky integrate and fire (LIF) model (see for example Rolls and Treves, 1998) while the equations for STDP at the Excitatory-Excitatory ( $E \rightarrow E$ ) synapses are adapted from Perrinet et al. (2001).

### 2.2. DIFFERENTIAL EQUATIONS

#### 2.2.1. Cell equations

Depolarization of the neuron’s membrane potential is described by Equation 1 and the cell (and synapse) constants were chosen to be as biologically accurate as possible based upon the available neurophysiological literature (see **Table 1** for a full list).

The cell membrane potential for a given neuron (indexed by  $i$ ) is driven up by presynaptic excitatory conductances (or direct current injection) and towards the inhibitory reversal potential (typically down) by presynaptic inhibitory conductances, decaying back to its resting state over a time course determined by the properties of its membrane.

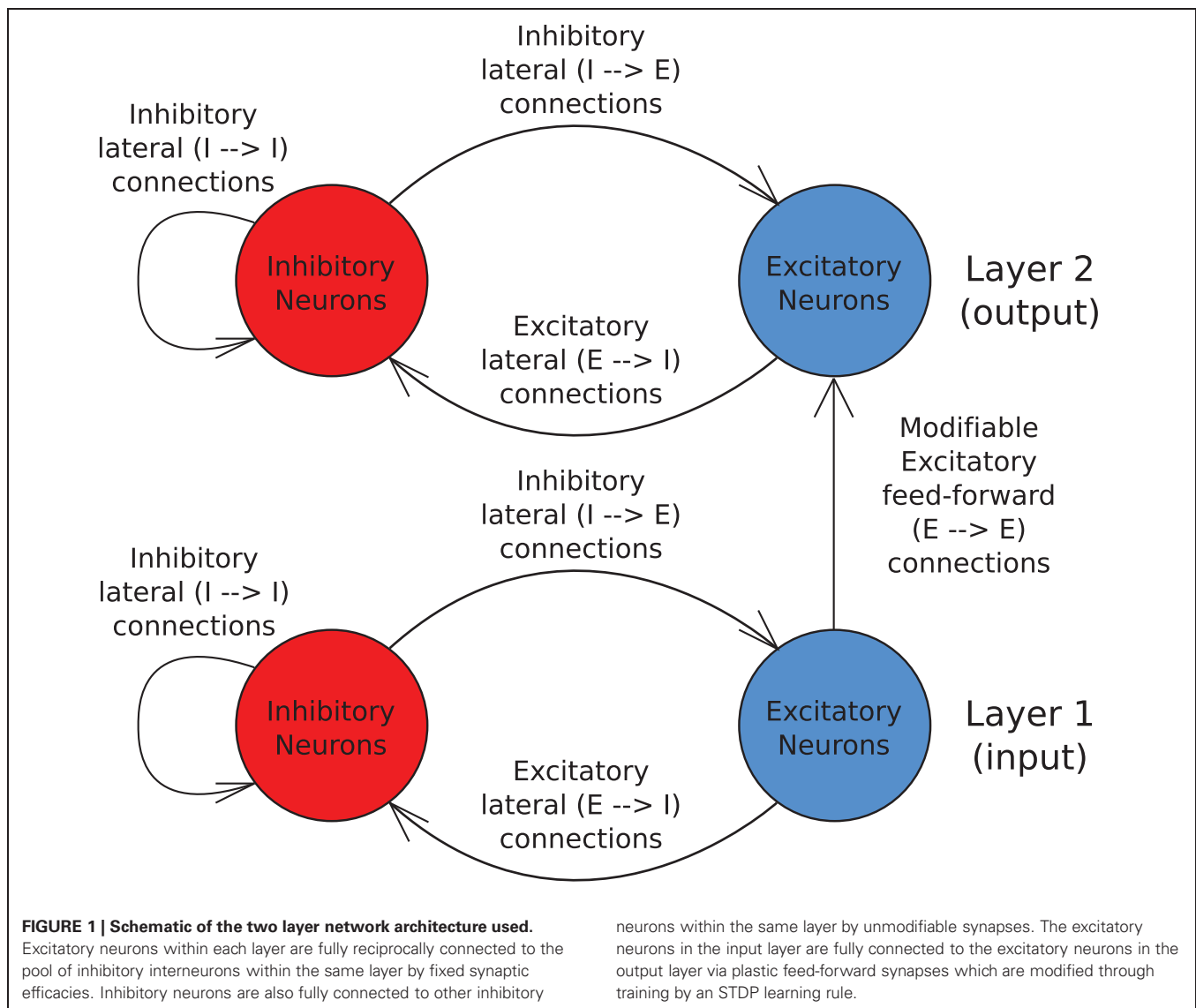
$$\tau_m^y \frac{dV_i(t)}{dt} = V_0^\gamma - V_i(t) + R^\gamma I_i(t) + R^\gamma I_i^{ext}(t) + \sigma \cdot \xi(t) \cdot \sqrt{\tau_m^y} \quad (1)$$

Here  $\tau_m$  represents the membrane time constant, defined as  $\tau_m = C_m/g_0$ , where  $C_m$  is the membrane capacitance,  $g_0$  is the membrane leakage conductance and  $R$  is the membrane resistance, ( $R = 1/g_0$ ).  $V_0$  denotes the resting potential of the cell (indexed by  $\gamma$  along with these other class-specific parameters),  $I_i(t)$  represents the total synaptic current (described in Equation 2) and  $I_i^{ext}(t)$  models the injected current.

In addition, Gaussian white noise was added to the cell membrane potential with zero mean and standard deviation  $\sigma = 0.015 \cdot (\Theta - V_H)$  as used by Masquelier et al. (2009). Here,  $\xi(t)$  is a Wiener (Gaussian) variable (where  $\xi(t)$  represents  $\frac{dW}{dt}$ ) satisfying the definition of the Wiener process such that  $\langle \xi \rangle = 0$  and  $\langle \xi(t)\xi(s) \rangle = \delta(t - s)$ , where  $\delta(\cdot)$  is the Dirac delta function and  $\sigma$  tunes the amplitude of the noise (the standard deviation of  $\tau_m^y$

<sup>2</sup>Typically, only 100–150 ms is required to respond to complex stimuli.





noise in units of Volts) since  $\xi$  has unit variance. The noise term,  $\xi$ , is importantly scaled by (the square root) of the time constant,  $\tau_m$ , which means that the amplitude of the noise is scaled up or down as the system speeds up (short  $\tau_m$ ) or slows down (long  $\tau_m$ ), respectively. The dimension of the  $\xi$  term is  $\text{time}^{-1}$  and so  $\xi$  is scaled by  $\sqrt{\tau_m}$  to make the equation dimensionally consistent.

The total synaptic conductance is the sum of conductances of all presynaptic neurons of each type (excitatory and inhibitory) with inhibitory conductances being negative.

$$I_i(t) = \sum_{\gamma} \sum_j g_{ij}(t) (\hat{V}^{\gamma} - V_i(t)) \quad (2)$$

Here  $\hat{V}$  represents the reversal potential of a particular class of synapse (denoted again by  $\gamma$ ) which consists of Excitatory and Inhibitory neurons  $\{E, I\}$  and  $j$  indexes the presynaptic neurons of each class.

## 2.2.2. Synaptic conductance equations

The synaptic conductance of a particular synapse,  $g(t)$ , (indexed by  $ij$ ) is governed by a decay term  $\tau_g$  and a Dirac delta function for when spikes occur, which correspond to the first and second terms of Equation 3. The Dirac delta function is defined as follows:

$$\delta(x) = \begin{cases} \infty & \text{if } x = 0 \\ 0 & \text{otherwise} \end{cases} \quad \text{where, } \int_{-\infty}^{+\infty} \delta(x) dx = 1.$$

The conduction delay for a particular synapse is denoted by  $\Delta t_{ij}$  and each spike is indexed by  $l$  as a separate train for each presynaptic neuron. A biological scaling constant,  $\lambda$  (set in all simulations to be 5 ns) has been introduced to scale the synaptic efficacy  $\Delta g_{ij}$  which lies between unity and zero.

$$\frac{dg_{ij}(t)}{dt} = -\frac{g_{ij}(t)}{\tau_g} + \lambda \Delta g_{ij}(t) \sum_l \delta(t - \Delta t_{ij} - t_j^l) \quad (3)$$



**Table 1 | Parameters used in the simulations.**

Parameter	Symbol	Value	Reference
Cue current	$I^{ext}$	1.0 nA	*
Cue period {training, testing}	$t_{cue}$	{100, 250} ms	
Time step	$\Delta t$	0.02 ms	
<b>NETWORK PARAMETERS</b>			
No. of layers	$N_L$	2	
No. of excitatory cells per layer	$N_E$	400	
No. of inhibitory cells per layer	$N_I$	100	
No. of afferent excit. connections per excit. neuron	$S_{EE}$	400	
No. of afferent excit. connections per inhib. neuron	$S_{EI}$	400	
No. of afferent inhib. connections per excit. neuron	$S_{IE}$	100	
No. of afferent inhib. connections per inhib. neuron	$S_{II}$	100	
<b>CELLULAR PARAMETERS</b>			
Excitatory cell somatic capacitance	$C_m^E$	500 pF	§
Inhibitory cell somatic capacitance	$C_m^I$	214 pF	§
Excitatory cell somatic leakage conductance	$g_0^E$	25 nS	§
Inhibitory cell somatic leakage conductance	$g_0^I$	18 nS	§
Excitatory cell membrane time constant	$\tau_m^E$	20 ms	§
Inhibitory cell membrane time constant	$\tau_m^I$	12 ms	§
Excitatory cell resting potential	$V_0^E$	−74 mV	§
Inhibitory cell resting potential	$V_0^I$	−82 mV	§
Excitatory firing threshold potential	$\Theta^E$	−53 mV	§
Inhibitory firing threshold potential	$\Theta^I$	−53 mV	§
Excitatory after-spike hyperpolarization potential	$V_H^E$	−57 mV	§
Inhibitory after-spike hyperpolarization potential	$V_H^I$	−58 mV	§
Excitatory reversal potential	$\hat{V}^E$	0 mV	§
Inhibitory reversal potential	$\hat{V}^I$	−70 mV	§
Absolute refractory period	$\tau_R$	2 ms	§
<b>SYNAPTIC PARAMETERS</b>			
Synaptic neurotransmitter concentration	$\alpha_C$	0.5	†
Proportion of unblocked NMDA receptors	$\alpha_D$	0.5	†
Presynaptic STDP time constant	$\tau_C$	{3, 75} ms	†
Postsynaptic STDP time constant	$\tau_D$	{5, 125} ms	†
Synaptic learning rate	$\rho$	0.1	†
Plastic ( $E \rightarrow E$ ) synaptic conductance range, CT	$\lambda \cdot \Delta g^{EE}$	{0, 4} nS	*
Plastic ( $E \rightarrow E$ ) synaptic conductance range, Trace	$\lambda \cdot \Delta g^{EE}$	{0, 1.25} nS	*
Change in synaptic conductance ( $I \rightarrow E$ )	$\lambda \cdot \Delta g^{IE}$	{0.5, 2.5} nS	*
Change in synaptic conductance ( $E \rightarrow I$ )	$\lambda \cdot \Delta g^{EI}$	5.0 nS	*
Change in synaptic conductance ( $I \rightarrow I$ )	$\lambda \cdot \Delta g^{II}$	5.0 nS	*
Excitatory-Excitatory synaptic time constant	$\tau_{EE}$	{2, 150} ms	*
Inhibitory-Excitatory synaptic time constant	$\tau_{IE}$	5 ms	§
Excitatory-Inhibitory synaptic time constant	$\tau_{EI}$	2 ms	§
Inhibitory-Inhibitory synaptic time constant	$\tau_{II}$	5 ms	§

Most integrate and fire parameters were taken from Troyer et al. (1998) (derived originally from McCormick et al. 1985) as indicated by §. Plasticity parameters (denoted by †) are taken from Perrinet et al. (2001). Parameters marked with \* were tuned for the reported simulations.

### 2.2.3. Synaptic learning equations

The following differential equations describe the STDP occurring at each modifiable *Excitatory – Excitatory* ( $E \rightarrow E$ ) synapse. Here  $i$  labels the postsynaptic neuron. The recent presynaptic activity,  $C_{ij}(t)$ , is modeled by Equation 4 which may be interpreted as the concentration of neurotransmitter (glutamate) released into the synaptic cleft (Perrinet et al., 2001) and is bounded

by  $[0, 1]$  for  $0 \leq \alpha_C < 1$ .

$$\frac{dC_{ij}(t)}{dt} = -\frac{C_{ij}(t)}{\tau_C} + \alpha_C (1 - C_{ij}(t)) \sum_l \delta(t - \Delta t_{ij} - t_j^l) \quad (4)$$

The presynaptic spikes drive  $C_{ij}(t)$  up at a synapse according to the model parameter  $\alpha_C$ , which then the current value of

$C_{ij}(t)$ , which then decays back to 0 over a time course governed by  $\tau_C$ .

The recent postsynaptic activity,  $D_i(t)$ , is modeled by Equation 5 which may be interpreted as the proportion of unblocked NMDA receptors as a result of recent depolarization through back-propagated action potentials (Perrinet et al., 2001).

$$\frac{dD_i(t)}{dt} = -\frac{D_i(t)}{\tau_D} + \alpha_D (1 - D_i(t)) \sum_k \delta(t - t_i^k) \quad (5)$$

Unlike with the conduction of action potentials to postsynaptic neurons, there is no conduction delay associated with  $D_i$  since the cell body is assumed to be arbitrarily close to the receiving synapses, and it is the same for a given (postsynaptic) neuron rather than each of its synapses since the effects of a postsynaptic spike are assumed to have an equal impact on all receiving synapses.

The strength of the synaptic weight,  $\Delta g_{ij}(t)$ , is then modified according to Equation 6, which is governed by the time course variable  $\tau_{\Delta g}$ .

$$\tau_{\Delta g} \frac{d\Delta g_{ij}(t)}{dt} = (1 - \Delta g_{ij}(t)) C_{ij}(t) \sum_k \delta(t - t_i^k) - \Delta g_{ij}(t) D_i(t) \sum_l \delta(t - \Delta t_{ij} - t_j^l) \quad (6)$$

Note that the postsynaptic spike train (indexed by  $k$ ) is now associated with the presynaptic state variable ( $C$ ) and vice versa. If  $C$  is high (due to recent presynaptic spikes) at the time of a postsynaptic spike, then the synaptic weight is increased (LTP) whereas if  $D$  is high (from recent postsynaptic spikes) at the time of a presynaptic spike then the weight is decreased (LTD).

The weight updates are also multiplicative, meaning that the amount of potentiation decreases as the synapse strengthens, as has been found experimentally (Bi and Poo, 1998). Theoretically,

this weight-dependent potentiation yields a normal distribution of synaptic efficacies rather than pushing each weight to one extreme or the other (van Rossum et al., 2000) as would be the case with an additive form of STDP.

### 2.3. NUMERICAL SCHEME

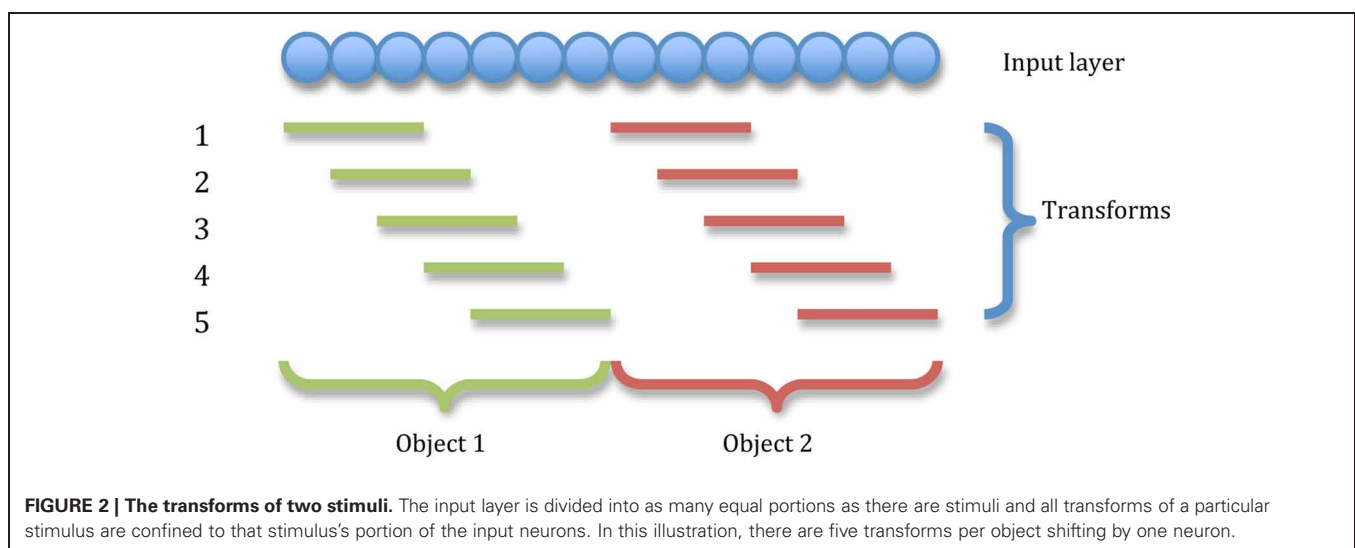
The differential equations described above are converted to finite difference equations and simulated using the Forward-Euler numerical scheme with a time step  $\Delta t = 0.02$  ms. In the finite difference equations, the Dirac delta function has been replaced by the discrete approximation,  $S(x)$  as defined in Amit and Brunel (1997). Finally, in the original description, the change in synaptic weight (Equation 6) was instantaneous and so  $\Delta t/\tau_{\Delta g}$  is defined to be a learning rate constant,  $\rho$ , in the corresponding finite difference equation.

### 2.4. TRAINING AND STIMULI

Stimuli are represented by injecting a small amount of current directly into the cell bodies of a particular set of excitatory input neurons continuously throughout the cue period. This pattern of stimulated neurons is gradually shifted across the input layer representing successive transforms of the stimulus (see **Figure 2**).

The size of a stimulus and the amount of neurons each of its transforms is shifted by allows us to precisely control the degree of overlap between transforms of each stimulus. Spatial continuity is crucial to the functioning of the CT mechanism, whereas the trace mechanism requires temporal continuity to associate successive transforms together, (which can be controlled independently through model time constants). In this way, we may eliminate the operation of one mechanism to study the other in isolation and hence disentangle their contributions to the network's capacity for invariance learning.

During training, the set of stimuli are presented in a random order with all transforms for a given stimulus being presented in succession before presenting the next stimulus's transforms. Presentation of all stimuli in this manner constitutes one training epoch, and the total training period comprised of five such epochs.



After completion of training, learning is switched off (prohibiting further synaptic modification) and the network is presented with all transforms of all stimuli in order (resetting the neurons to their resting state between transforms) and the resultant firing in both input and output layers is saved for analysis.

## 2.5. PERFORMANCE MEASURES

Two information-theoretic<sup>3</sup> measures are used to assess the network's performance which reflect the extent to which cells respond invariantly to a particular stimulus over several transforms but differently to other stimuli [for more details see Rolls and Milward (2000); Elliffe et al. (2002)]. The work presented has used spiking neural networks because we believe that their richer dynamics they model are critical for learning to solve the problem of object recognition (transformation-invariant cell responses). However, analysis of macaque visual cortical neuron responses has found that after learning, the majority of the information about stimulus identity is contained within the *firing rates* rather than the detailed timing of spikes (Tovee et al., 1994). As such, we adopt a dual approach whereby the network self-organizes through spiking dynamics but the information content with respect to stimulus identity is assessed through the output cell's firing rates.

During testing each transform of each stimulus was presented to the input layer of the network. Each neuron was reset (allowed to settle) after presentation of each transform such that the activity due to one transform did not affect the responses to later transforms. After testing, the spikes of each output neuron were placed into a different bin for each transform of each stimulus and the corresponding firing rate for each cell was calculated. Based upon these firing rates, the stimulus-specific single-cell information  $I(s, R)$  was calculated according to Equation 7, which gives the amount of information in a set of responses  $R$  of a single cell about a specific stimulus  $s$ . The set of responses,  $R$  consisted of the firing rate of a cell to every stimulus presented in every location.

$$I(s, R) = \sum_{r \in R} P(r|s) \log_2 \frac{P(r|s)}{P(r)} \quad (7)$$

Good performance for a cell would entail stimulus specificity (with generality across most or all transforms of that stimulus), meaning a large response to one or a few stimuli regardless of their position (transform) and small responses to other stimuli. We therefore compute the maximum amount of information a neuron conveys about *any* of the stimuli rather than the average amount it conveys about the whole set  $S$  of stimuli (which would be the mutual information).

If all the output cells learnt to respond to the same stimulus then there would be no discriminability and the information about the set of stimuli  $S$  would be poor. To test this, the multiple cell information measure is used which calculates the information about the set of stimuli from a population of up to 10 output neurons. This population consisted of the subset of up to five cells which had, according to the single cell measure, the most information about each of the two stimuli. Ideally, we would

calculate the mutual information (the average amount of information about which stimulus was shown from the responses of all cells after a single presentation of a stimulus, averaged across all stimuli), however, the high dimensionality of the neural response space and the limited sampling of these distributions is prohibitive.

Instead, a decoding procedure is used to estimate the stimulus  $s'$  that gave rise to the particular firing rate response vector on each trial. From this a probability table is then constructed of the real stimuli  $s$  and the decoded stimuli  $s'$ , from which the mutual information is calculated (Equation 8).

$$I(s, s') = \sum_{s, s'} P(s, s') \log_2 \frac{P(s, s')}{P(s)P(s')} \quad (8)$$

A Bayesian decoding procedure is used for this purpose, whereby the firing rates of each cell in the ensemble vector to each transform of each stimulus in turn is fitted to a Gaussian distribution parameterized by these means and standard deviations of each cell's responses to all other transforms of each stimulus separately to yield an estimate of  $P(r_c|s')$ . Taking the product of these probabilities over all cells in the response vector with  $P(s')$  and then normalizing the resultant joint probability distribution gives an estimate of  $P(s'|r)$ , (Földiák, 1993). These probability distributions are factored into a confusion matrix of  $P(s, s')$  over many iterations to smooth the effects of randomly sampling the output cells. From this decoding and cross-validation procedure, the probability tables are constructed for calculating the multiple cell information measure, further details of which may be found in Rolls et al. (1997). This measure should increase up to the theoretical maximum  $\log_2 N_S$  bits, (where  $N_S$  is the number of stimuli), as a larger population of cells is used, only if those cells have become tuned to different stimuli.

## 3. SIMULATIONS

In the simulations described below we investigated invariance learning in a spiking neural network with STDP utilizing two different learning mechanisms. For details of the methods and parameters used for the following simulations, please refer to section 2 and **Table 1**, respectively.

### 3.1. CONTINUOUS TRANSFORMATION LEARNING

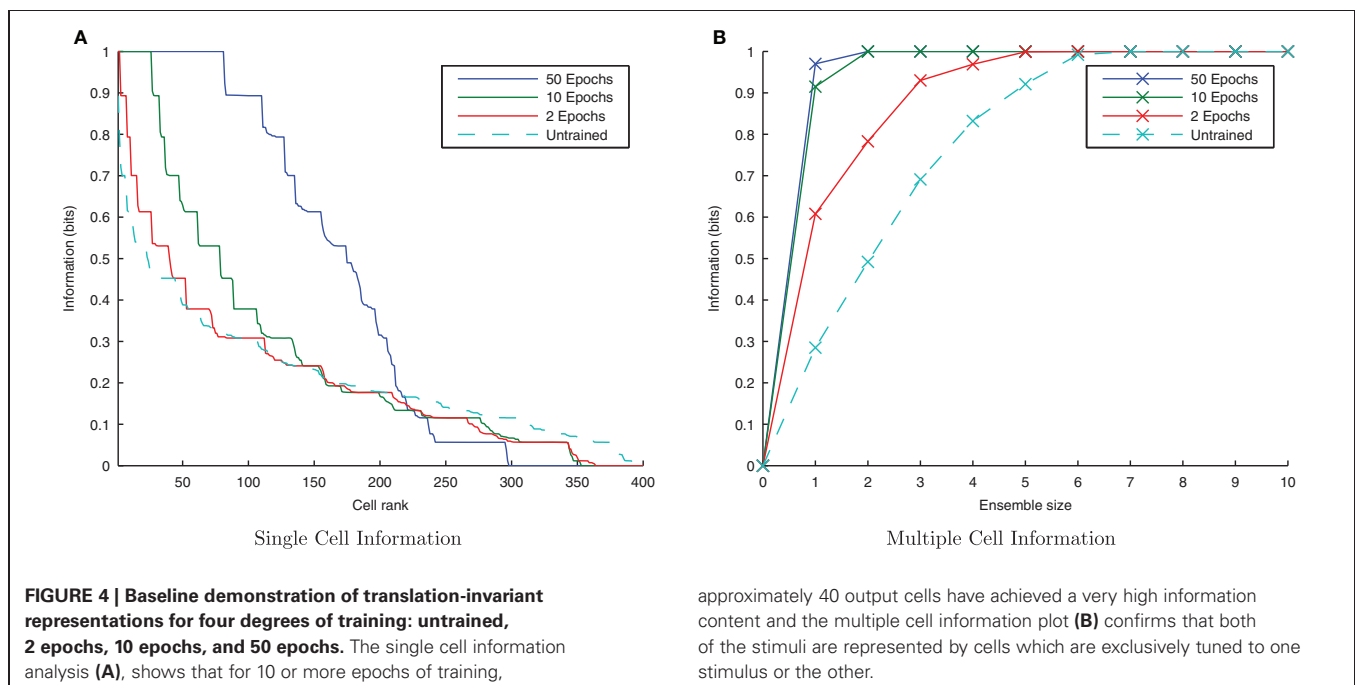
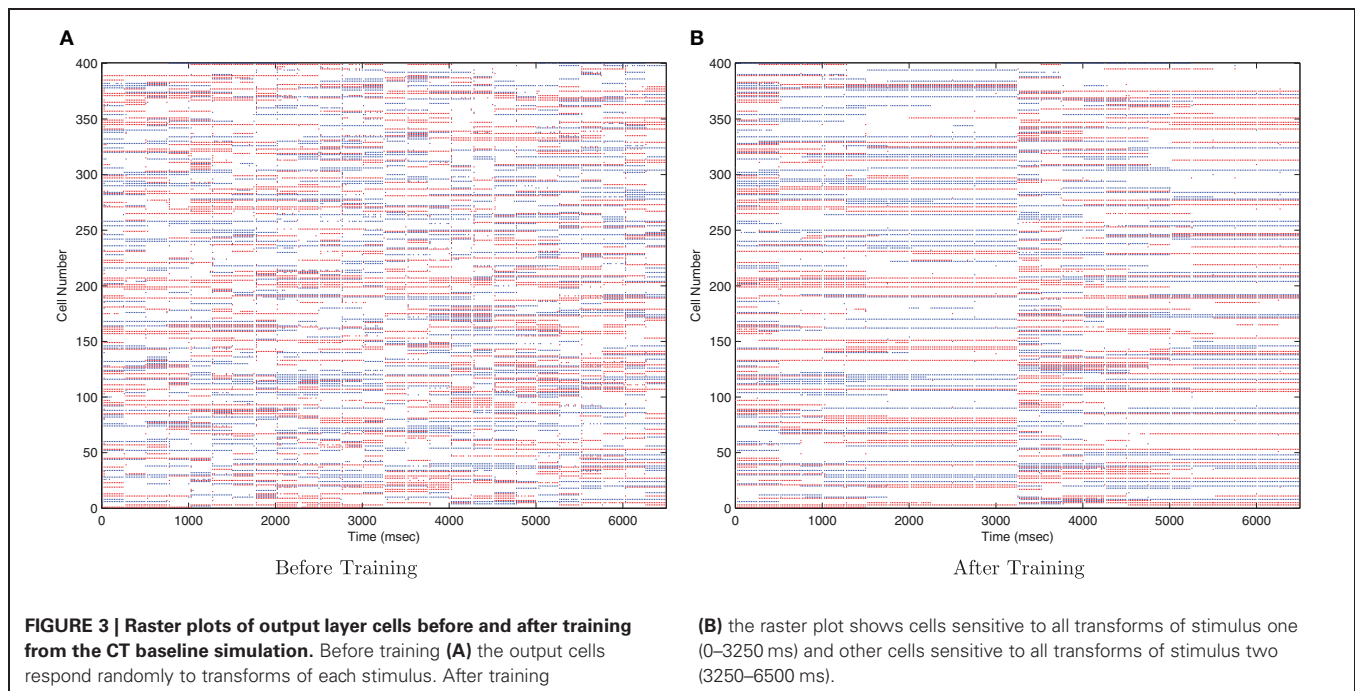
Continuous Transformation (CT) learning relies upon the spatial continuity of continuously transforming stimuli and a purely associative (Hebbian) learning rule with lateral competition to associate together successive transforms of a stimulus (Stringer et al., 2006). Presentation of an initial transform will excite one or more postsynaptic neurons and through the Hebbian learning rule, will strengthen the synapses between those cells. If there is enough overlap (similarity) between the original and a new transform, the same postsynaptic neuron(s) will be excited and so increase their synaptic strengths to the neurons of the current transform. This process can continue across a series of overlapping transforms until they are all mapped onto the same output cells. Since similar images are more likely to be transforms of the same object than different stimuli, the CT mechanism provides an explanation for how transformation-invariant representations may develop in the ventral visual system.

<sup>3</sup>For a general introduction to Information Theory see MacKay, 2003.

In this set of simulations, the parameters were chosen to encourage the operation of the CT learning mechanism (Stringer et al., 2006) while excluding any trace-like effects (Földiák, 1991). To this end, spatial overlap between successive transforms was generally kept high (13 transforms per stimulus each covering 56 neurons and shifting by 12 neurons per transform by default). Also a short time constant of 2 ms was used for the Excitatory-Excitatory (feed-forward) synaptic conductances,  $\tau_{EE}$ . These conditions were hypothesised to support a CT-like learning mechanism in a spiking neural network.

### 3.1.1. Invariance learning with CT

This simulation demonstrates the formation of transformation-invariant representations in the output cells through STDP as illustrated by the raster plots in **Figure 3** (which contrast the untrained with the trained network) and the information plots of **Figures 4A** and **B**. The level of inhibition had to be tuned so that the spikes from additional neurons from successive transforms (in the input layer) could be brought into phase with those already firing from the previous transforms. While the feed-forward excitatory weights were plastic and hence modified through learning,



their maximum level was set to 4 nS to achieve a reasonable level of output layer activity for the network size and connectivity. It can be seen from the pre-training raster plot (Figure 3A) that before learning, output neurons respond to a random set of transforms of each stimulus. However, post-training (Figure 3B), there are several cells which are responsive across the whole set of transforms for the first stimulus which are presented contiguously over the first 3250 ms, and several other cells which respond to all transforms of the second stimulus presented contiguously over the second 3250 ms.

In accordance with the raster plot, the  $I(s, R)$  (single-cell information measure) plots show many more cells in the network have attained the maximum information content (1 bit) than in the untrained case, demonstrating both transformation-invariance and stimulus specificity. Examining the  $I(s, s')$  (multiple cell information measure) plots shows that the maximum information about the stimulus set  $S$  is reached with fewer than the 10 available cells of the output ensemble of the highest scoring cells (in terms of their  $I(s, R)$  values), thus confirming that both stimuli are represented invariantly.

### 3.1.2. Temporal specificity

By default, the learning time constants,  $\tau_C$  and  $\tau_D$ , used in these simulations are 15 and 25 ms in accordance with Perrinet (2003). Here we reran the same simulations but shortened or lengthened these time constants by a factor of five (maintaining the same 3/5 ratio) to give 3/5 ms and 75/125 ms for  $\tau_C/\tau_D$ , respectively. Figure 5A shows a trend of a much greater information content in the network with the shorter (more temporally specific) time constants (3/5 ms) with the accompanying  $I(s, s')$  plot confirming that both stimuli are being represented (see Figure 5B). Network performance drops, however, with the longer (less temporally specific) STDP time constants 75/125 as the learning rule is less

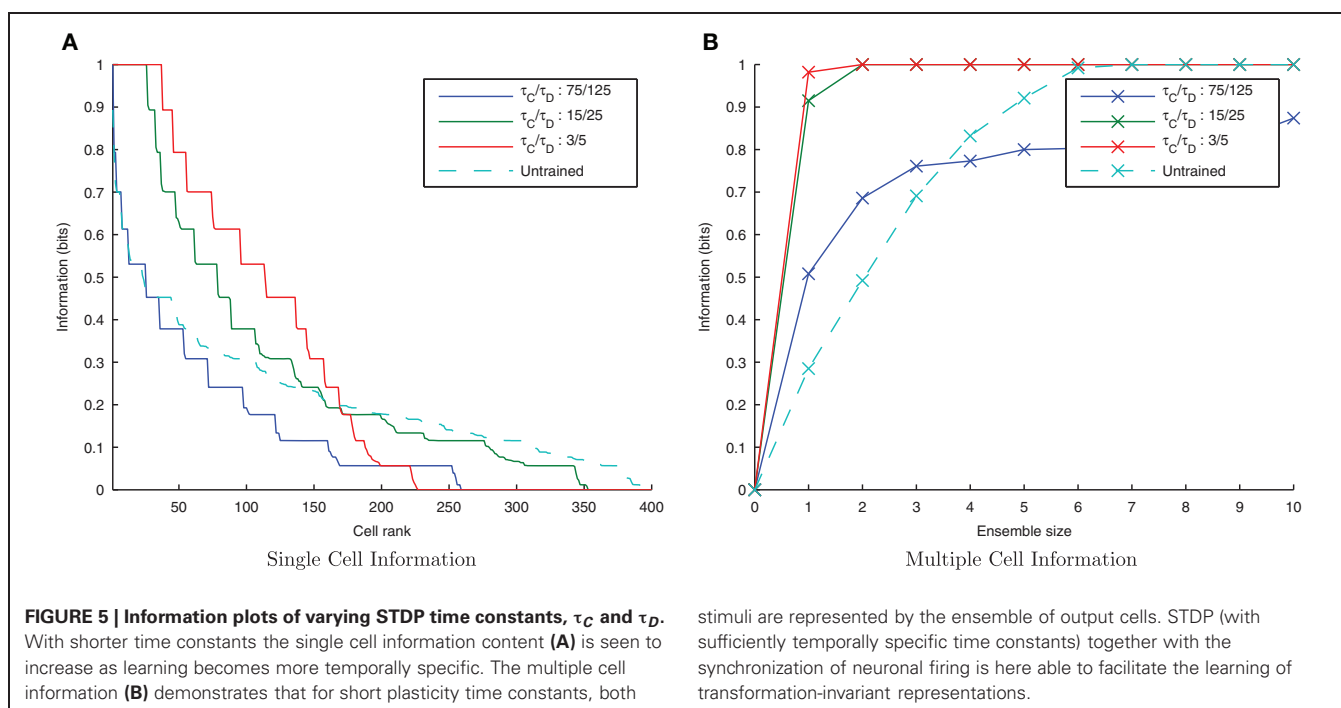
capable of capturing the temporally specific causal relationship of the input/output spike volleys.

The effect of shortening the STDP time constants is that after a pre-post spike pairing results in LTP, the following presynaptic spike from the next wave comes a relatively long time after the initial pair, such that the effect of its post-pre LTD is significantly lessened. The synaptic weight distributions in Figure 6 support this, exhibiting a peaked distribution of synaptic efficacies arising from the initially flat uniform distribution (as expected from a multiplicative model of STDP in the standard case,  $\tau_C = 15$  ms,  $\tau_D = 25$  ms, Figure 6B) and more peaked distributions with shorter STDP time constants (Figure 6C) indicating more specific learning. The higher proportions of large weights with the shorter learning time constants are what might be expected from an unbalanced learning rule dominated by LTP when waves of input spikes are widely spaced relative to the time delay until the postsynaptic spikes which they cause. In contrast, the weight distribution with the longer STDP time constants is smoother, indicating a less trained layer of synaptic weights (Figure 6A).

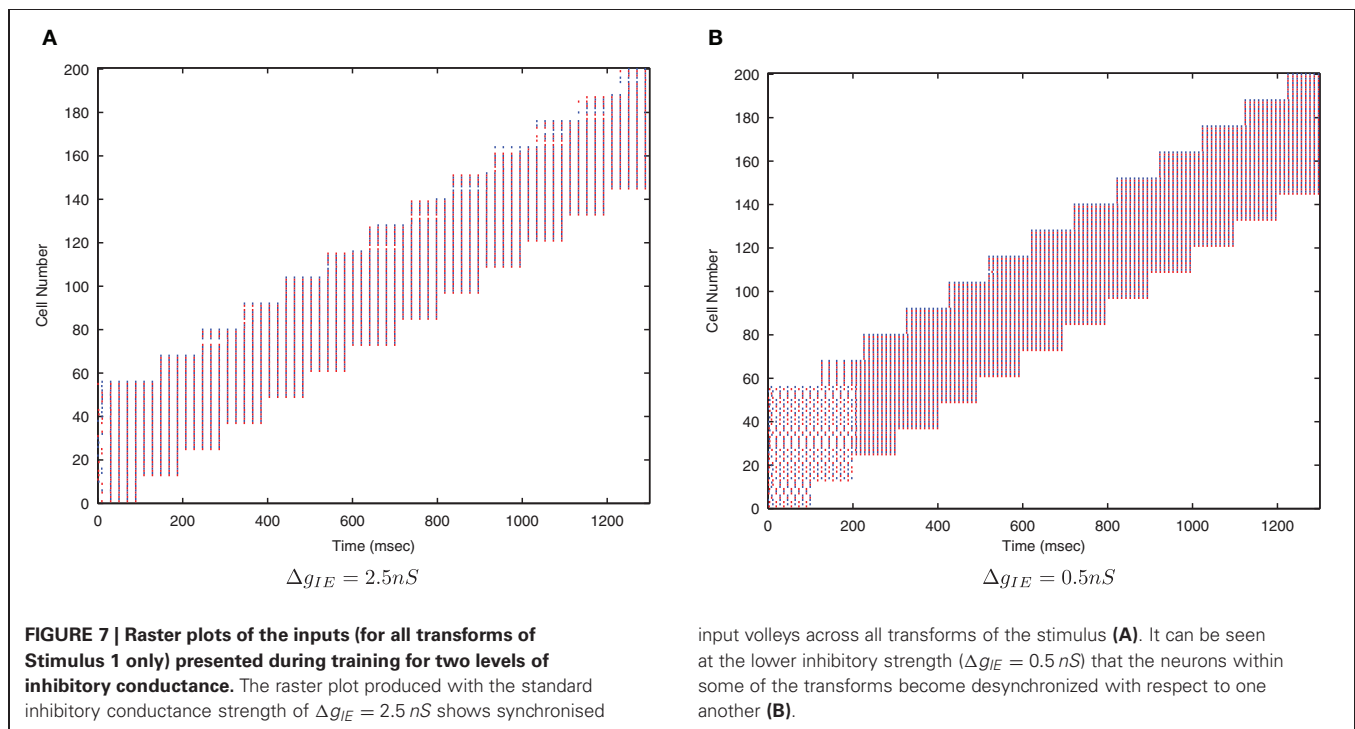
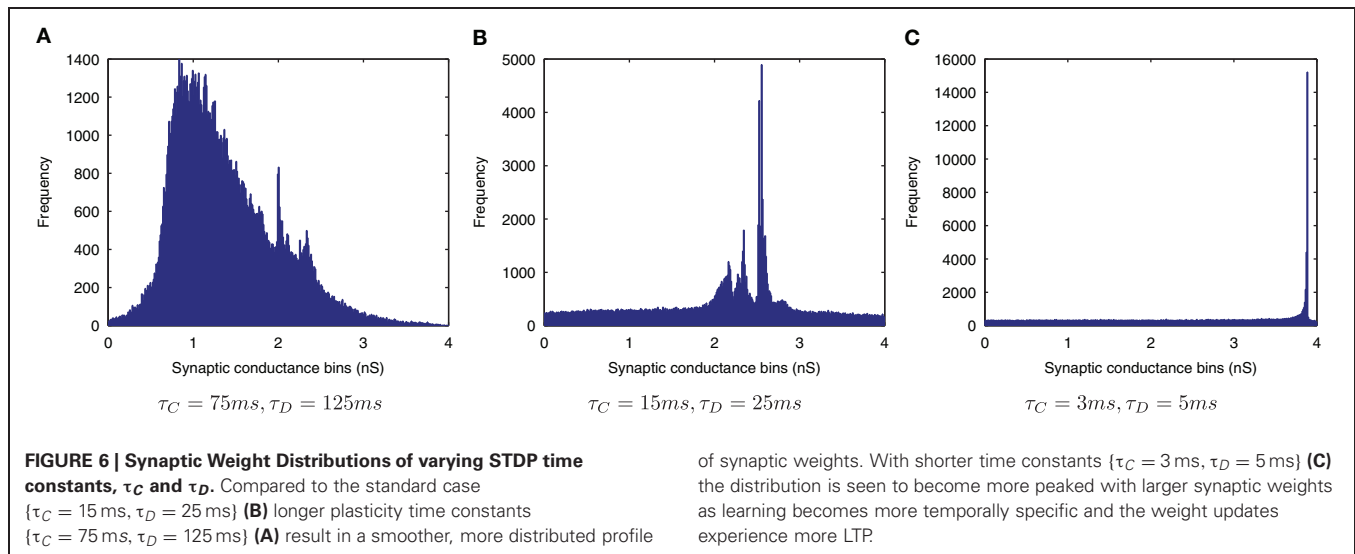
### 3.1.3. Lateral inhibition and synchrony

From earlier simulations, it is apparent that this training paradigm and the STDP model are very sensitive to the effects of the strength of inhibition on the synchronization of input spikes. We therefore systematically varied the strength of  $\Delta g_{IE}$ , the *Inhibitory*  $\rightarrow$  *Excitatory* conductances (which were non-plastic) to understand these effects in more detail.

Figure 7 shows that as the level of inhibition is reduced and the cell membrane potential noise begins to cause jitter in the spike timings, the new input layer neurons from successive transforms no longer fire in phase with those neurons from previous transforms. This reduces invariance learning in the output layer, where the information content can also be seen to be reduced (Figure 8).







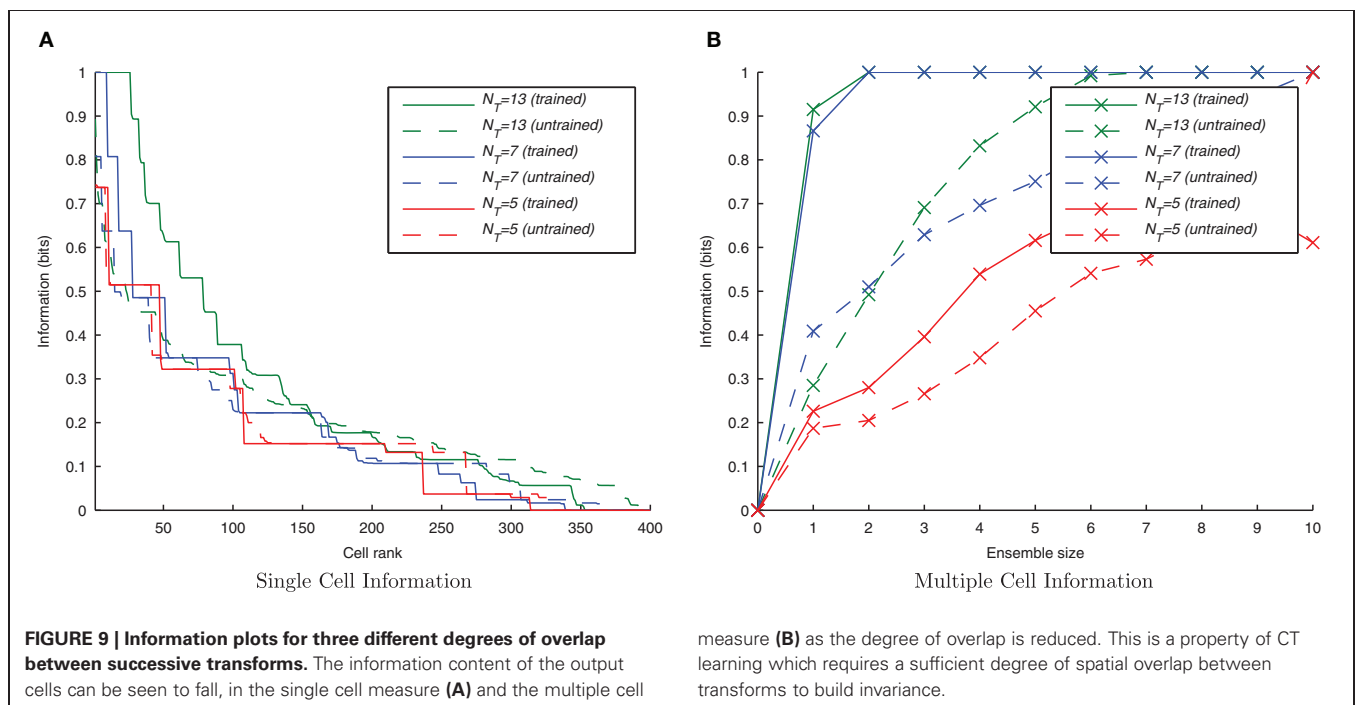
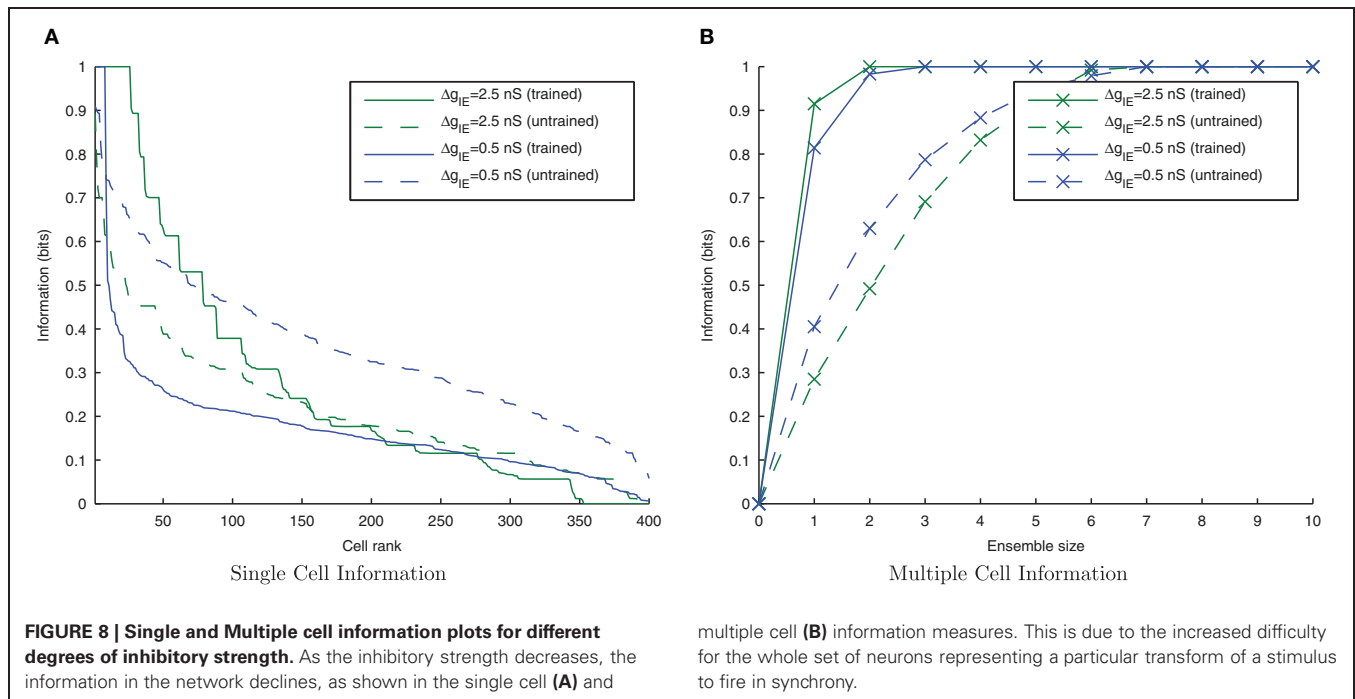
### 3.1.4. Degree of overlap

From previous rate-coded simulations it is clear that CT learning requires a high degree of resemblance among adjacent members of a set of transforms in order to associate them together. If this mechanism is being employed in the present spiking model, its performance should suffer by reducing this transform similarity. This was tested by removing intermediate transforms leaving only every 2nd or 3rd transform from the original sets of 13 transforms per stimulus (with a consecutive transform overlap of 44 neurons) such that there were only 7 or 5 transforms per stimulus, respectively. Since they still occupied the same proportion of

the input layer, the degree of overlap between any two consecutive transforms was correspondingly lower, being 32 or 20 neurons, respectively.

It can be seen from **Figure 9** that by reducing the spatial overlap between successive transforms of each object, the information content of the output layer declines (despite there being fewer transforms to associate together) since there are fewer cells that respond invariantly across all transforms of a given stimulus. This confirms that the network is learning invariance by a spiking equivalent of the CT learning mechanism.



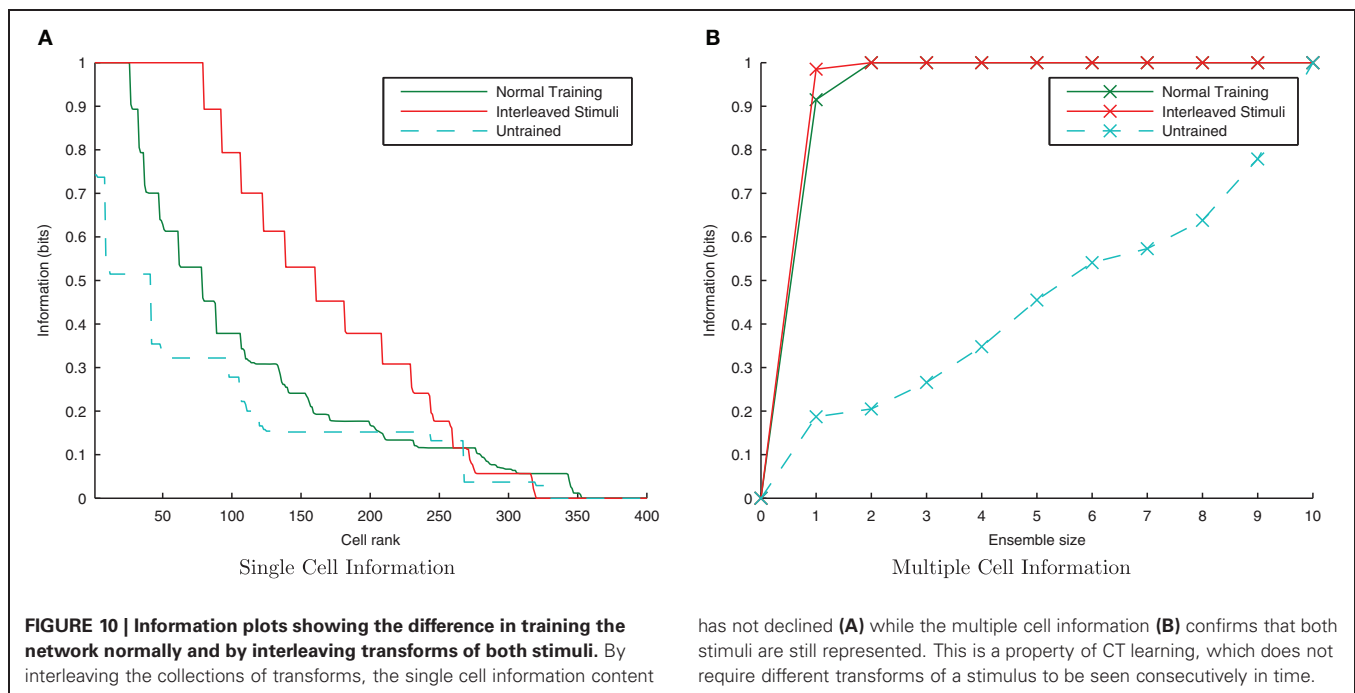


### 3.1.5. Interleaved transforms

Since the degree of similarity between any two transforms of a stimulus is the same regardless of when they are presented to the network, under a CT learning regime it should not matter whether the transforms are seen close together in time or not. One of the key properties of CT learning is therefore its ability to enable a network to learn about stimuli, even when their transforms are interleaved with those of another stimulus

(analogous to learning to recognize two faces or objects as the viewer saccades back and forth between them). To test this hypothesis we presented transforms of each stimulus alternately i.e.,  $S_1^{t_1}, S_2^{t_1}, S_1^{t_2}, S_2^{t_2}, \dots, S_1^{t_n}, S_2^{t_n}$ . If neurons are able to develop transformation-invariant responses with this training paradigm, it proves the learning mechanism is not utilizing a temporal trace.

Here it is evident from the information analysis (Figure 10) that the network has managed to learn about the individual



stimuli with this additional constraint. Both the single and multiple information measures show not just comparable results to the consecutive presentation of each stimulus's transforms during training but surprisingly, an improvement over the standard case. Examining the input layer rasters, this enhancement to learning from interleaving the stimuli seems to be due to the fact that under normal training, the first one or two spike volleys of a new transform have not yet recruited the additional neurons (which were not part of the previous transform) due to the lateral inhibition suppressing them. In contrast, neurons in the overlapping region are under constant stimulation from the injected current and so continue to fire, whereas those neurons exclusive to the previous transform stop firing when no longer stimulated with direct current.

When the stimuli are interleaved, however, all of the input neurons representing the new transform are stimulated by current injection at the same time (rather than their cell membrane potentials starting at different points in the stimulation cycle) and so fire simultaneously from the very first volley. Using a stimulating direct current of 1 nA with the cell body parameters and network connectivity given in **Table 1**, the neurons will fire approximately five complete volleys of spikes in the 100 ms presentation period (50 Hz). The ultimate effect of this training difference is that in the interleaved case, each transform will be represented by five complete spike volleys (as opposed to only three or four in the standard case) and hence will be trained more fully (with more useful weight updates) over the same training duration.

### 3.1.6. Randomized transform order

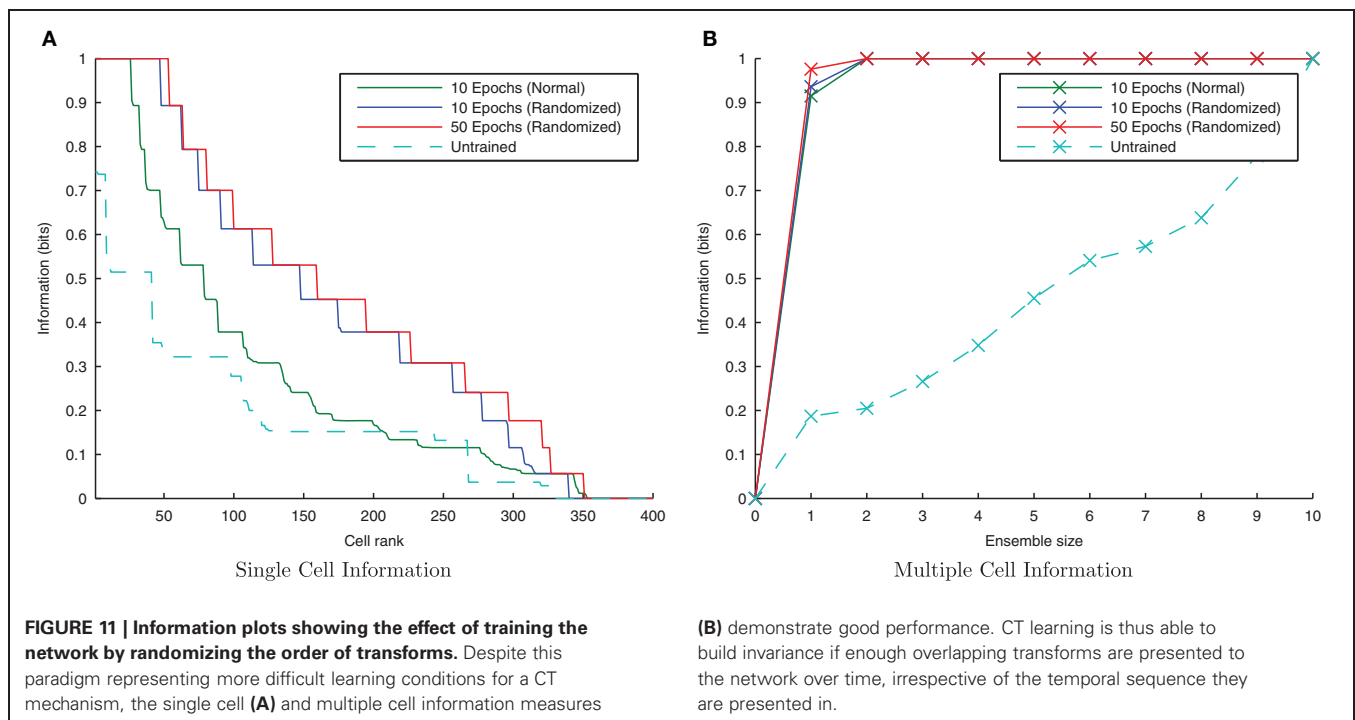
CT learning is also able to form transformation-invariant representations when the individual transforms of an object are presented in a random order during training. This is analogous

to learning to recognize a face or object from a number of random "snapshot" views rather than seeing it move smoothly. The consequence of such a training regime is that there is not necessarily any overlap between two consecutive transforms in time. At the beginning of training when the feed-forward weights are randomly initialized, this training regime may mean that different output neurons learn to respond to different subsets of each stimulus' transforms, thus making it harder for the similarity-based CT mechanism to associate all related (overlapping) transforms together onto the same output neurons. If, however, there is a sufficient number of such training epochs and degree of competition in the output layer, eventually each transform will be randomly followed by a similar enough transform such that the same postsynaptic cell is fired which eventually learns invariance across the whole set of transforms.

Initially randomizing the order of transforms degraded the network performance as expected. However, building upon the learning enhancement found in the previous simulations with interleaving the stimuli, simulations were repeated with simultaneously randomized transform order and interleaved stimuli. **Figure 11** demonstrates that the network is able to cope with randomizing the order of the transforms.

### 3.2. TRACE LEARNING

Trace learning utilizes the temporal continuity of objects in the world to learn transformation-invariant representations (Földiák, 1991). The mechanism relies upon the proposal that over short time scales, successive images are more likely to be transforms of the same object rather than different objects. The trace learning rule (Földiák, 1991; Wallis and Rolls, 1997) uses these temporal statistics of visual input by incorporating a temporal trace of the previous (typically postsynaptic) neural activity



into a simple Hebbian learning rule, which helps to maintain firing in the same output cell(s) when successive transforms are presented. Through further Hebbian synaptic modifications, successive transforms may become associated together onto the same output cells leading to transformation-invariant neurons.

In contrast to the CT simulations (section 3.1), here we lengthen the synaptic time constant  $\tau_{EE}$  to 150 ms to explore the hypothesis that by continuing to bleed current into a post-synaptic neuron, the activity generated by one transform may be associated with the next. In this way, a temporal trace effect may be achieved, allowing a spiking neural network to learn through temporal rather than spatial continuity.

### 3.2.1. Invariance learning with a temporal trace

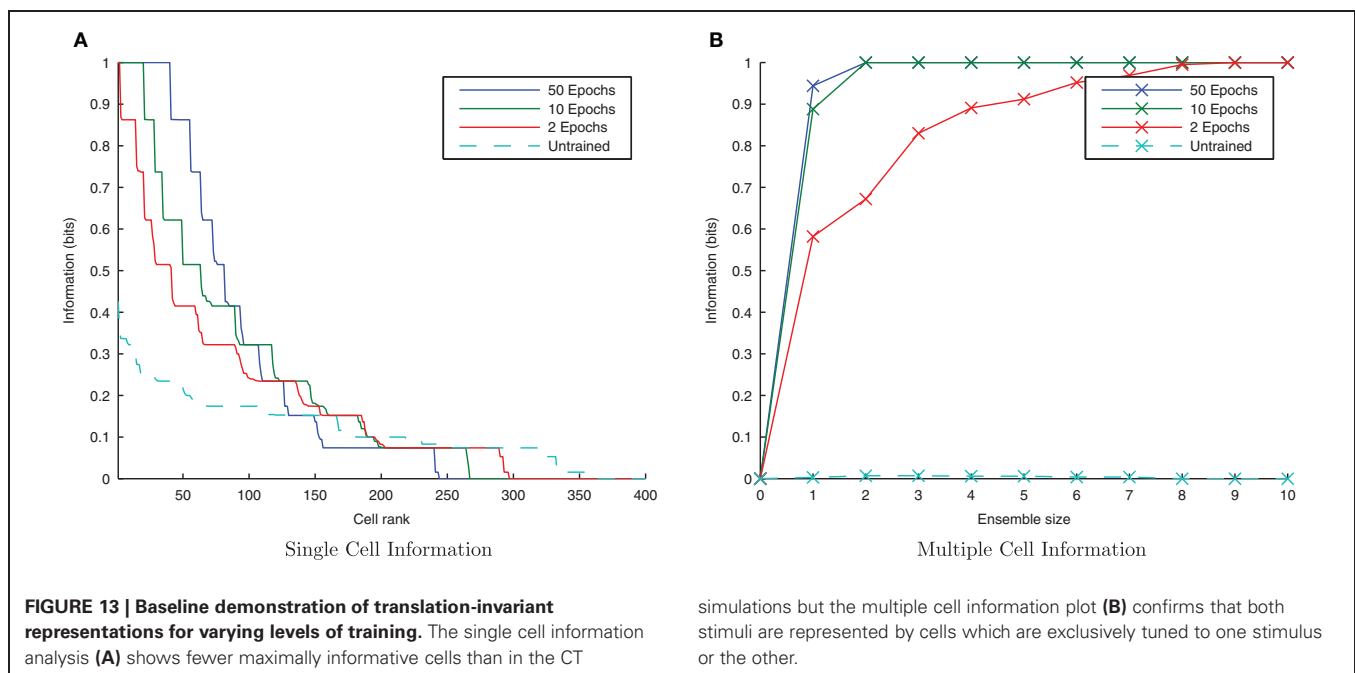
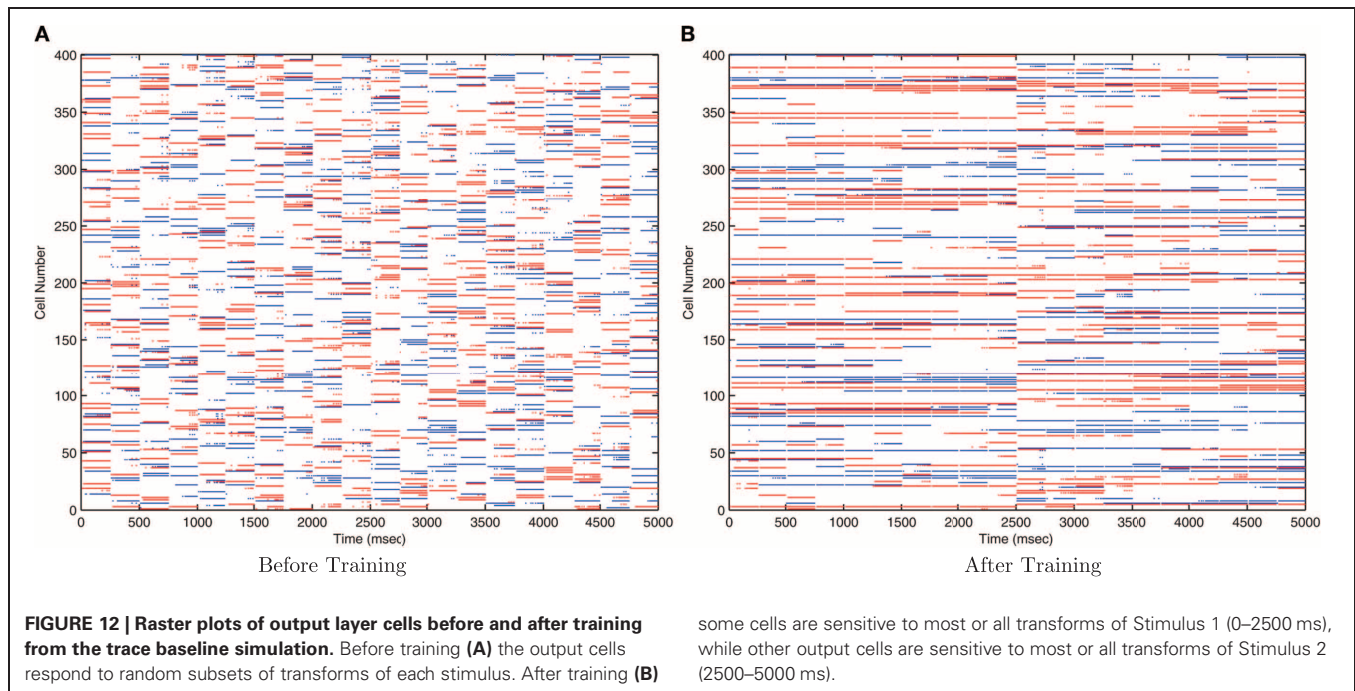
This simulation demonstrates the formation of transformation-invariant representations in the output cells through STDP and a trace-like effect from longer  $E \rightarrow E$  synaptic time constants. The other parameters remained the same as in the CT simulations except that the maximum strength of the plastic feed-forward excitatory synapses was reduced to 1.25 nS (from 4 nS) to compensate for the greater degree of excitation arising from the longer feed-forward synaptic time constant. Also the stimuli were changed such that in the following trace simulations, there are 10 transforms per stimulus (consisting of 20 neurons each) which are shifted by 20 neurons for each transform such that there is no spatial overlap between transforms. Since these transforms are orthogonal, any CT effects from spatial overlap are eliminated. Additionally, since the spatio-temporal statistics of natural stimuli tend to have different transforms of the same stimulus closer together in time more frequently than transforms of different stimuli, the neurons were allowed to settle between

presentation of the two sets of transforms (stimuli) to effectively reduce the temporal continuity between different stimuli so as to avoid introducing an artificial trace effect between them. These changes allow for a controlled investigation of whether orthogonal transforms may be linked together by a trace-like learning mechanism by lengthening the excitatory synaptic conductance.

Due to the random initialization of the feed-forward weights, output neurons before training respond to a random set of transforms of each stimulus (Figure 12A), whereas after training Figure 12B shows both stimuli are represented by cells which are invariant to most transforms of their respective stimuli, while the information plots (Figure 13) confirm that both stimuli may be identified with a small ensemble of output neurons. In earlier simulations without allowing the neurons to settle between each set of transforms (not shown here), the multiple cell information measure was found to drop with further training. This was caused by the association of the two stimuli together since they are presented consecutively in time during training with long synaptic time constants, so the last transform of the first stimulus was still active as the first transform of the second stimulus was presented, thereby leading to their association.

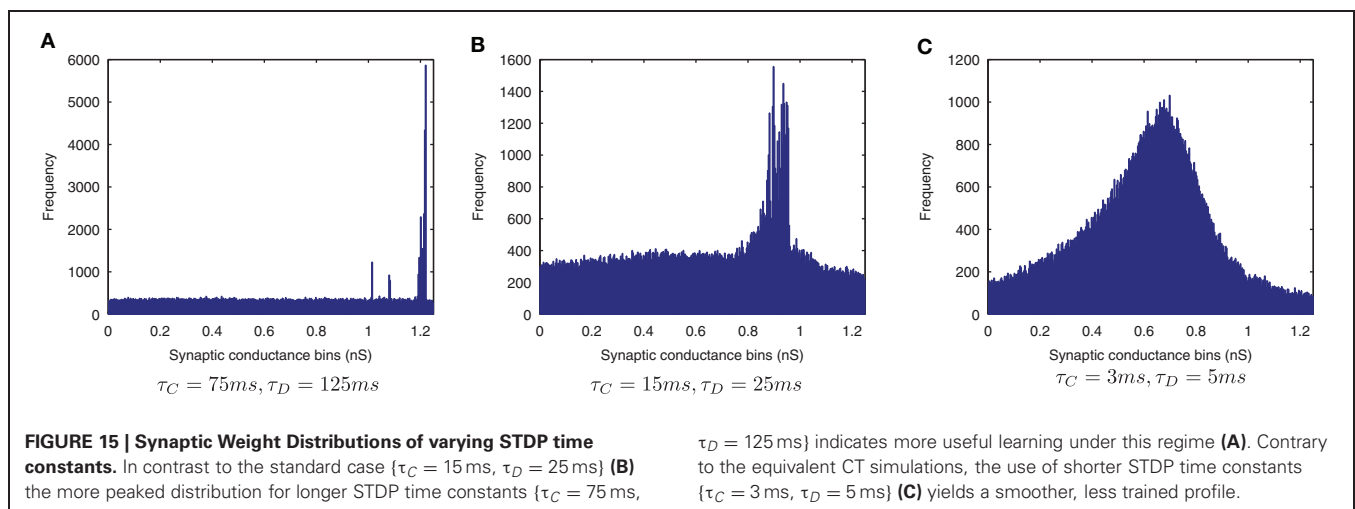
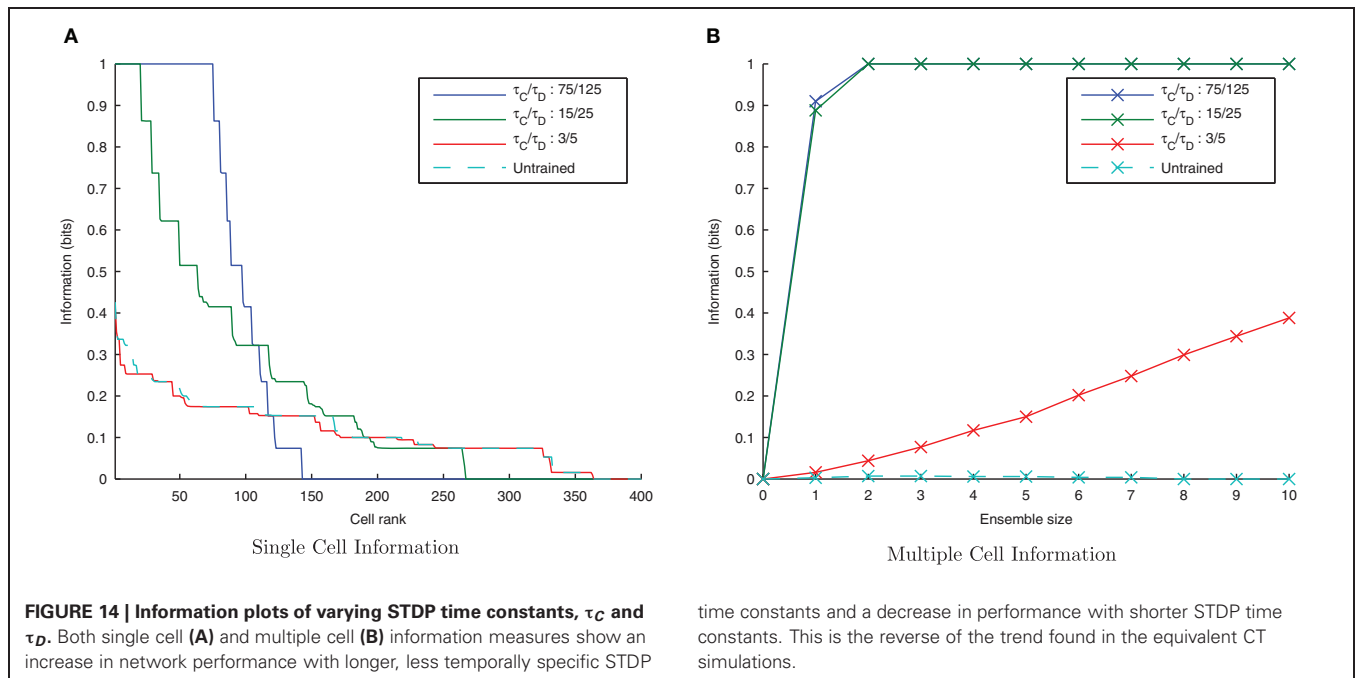
### 3.2.2. Temporal specificity

Lengthening the synaptic conductance time constant,  $\tau_{EE}$ , may affect the dynamics of synaptic plasticity in unforeseen ways, so it was important to explore a range of values for the plasticity time constants as for the first set of CT simulations. As before,  $\tau_C = 15$  ms and  $\tau_D = 25$  ms were used as standard for the learning time constants but here they are shortened and lengthened by a factor of five (keeping the same 3/5 ratio) for comparison (while  $\tau_{EE}$  remains fixed at 150 ms).



The results are shown in the information plots of **Figure 14** and the synaptic weight distributions of **Figure 15**. In contrast to the previous CT simulations, the network performance degrades with more temporally specific (shorter) STDP time constants (**Figure 14**) but improves with longer, less specific STDP time constants (the reverse trend). Similarly, this opposite trend is borne out by the synaptic weight distributions (**Figure 15**) exhibiting a smoother profile (indicating less useful training) for short STDP time constants and a more peaked profile for longer, less temporally specific STDP time constants.

This reverse effect may be understood in the context of the two learning mechanisms whereby CT learning performs best with tightly synchronized, temporally-specific causal spike volleys, hence a temporally specific form of STDP is most appropriate. In contrast, trace learning requires activity to continue over an extended period of time between different transforms in order to associate them together, and as such the relationship it needs to capture is less temporally specific and thus a less specific form of STDP is better suited for this purpose.



### 3.2.3. Lateral inhibition and synchrony

As the level of inhibition is reduced, and the effects of timing jitter from the cellular membrane potential noise become more prominent, the new input layer neurons from successive transforms no longer fire in phase with those neurons from previous transforms and the information content of the output layer (Figure 16) can be seen to be reduced.

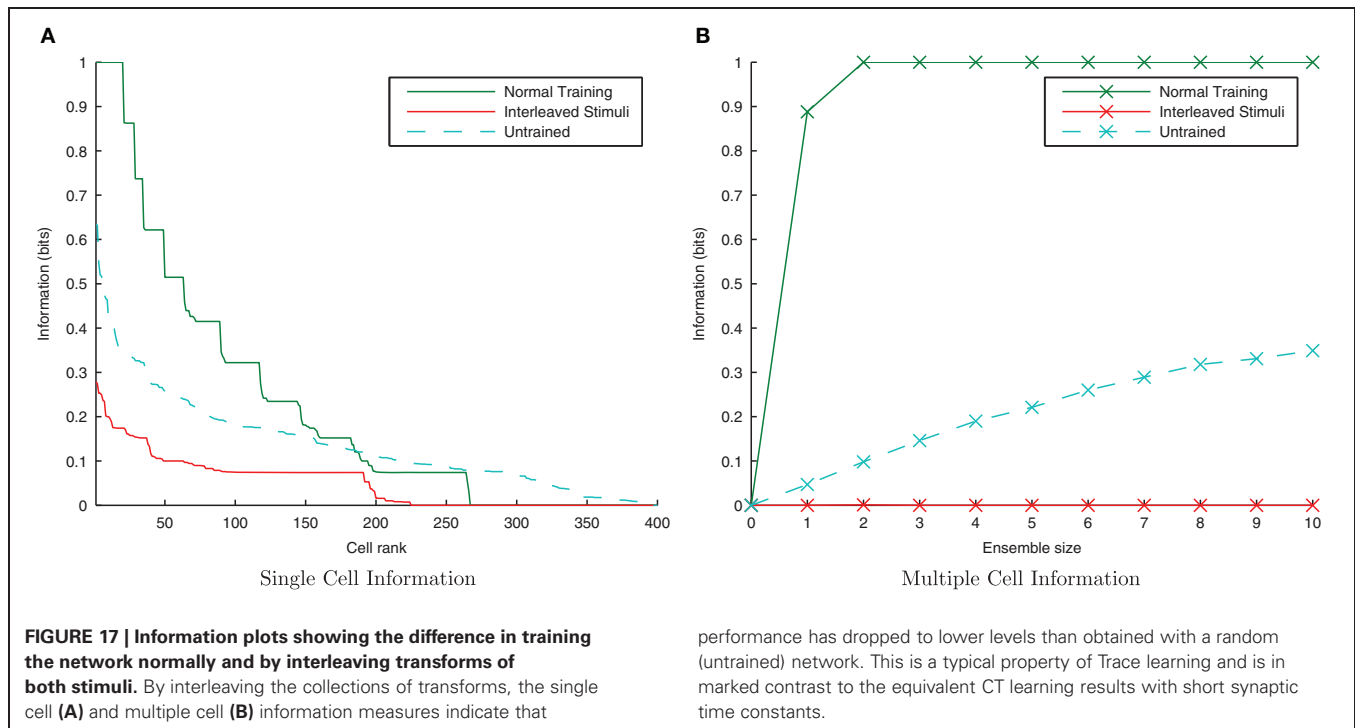
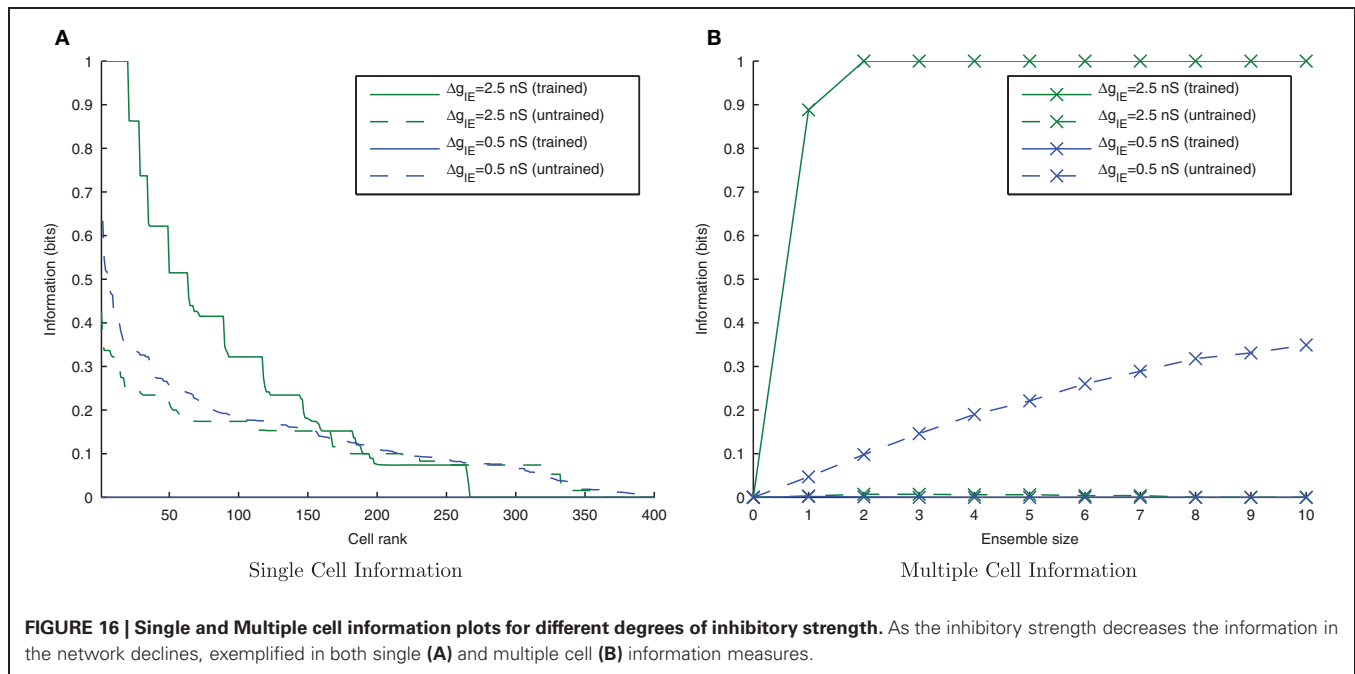
### 3.2.4. Interleaved transforms

By interleaving transforms of the two stimuli alternately through time, transforms from different stimuli should be associated together by their temporal continuity with a trace-like mechanism. Unlike in the previous CT simulations (where the association is not time-dependent, only similarity dependent), this inter-stimulus association should lead to a large drop in

information since the network will be unable to distinguish between the two stimuli. The neurons were not allowed to settle between presentations of different stimuli (as with previous trace simulations) as this would negate the effect of interleaving the stimuli and undermine the purpose of this section of simulations.

From Figure 17 it is evident that interleaving the transforms of the two stimuli has significantly reduced the information content of the network as expected. In the interleaved case, the single-cell information content (Figure 17A) has dropped to a poorer level than the untrained case (tested with a random uniform distribution of synaptic weights) as transforms from each stimuli have been associated together, meaning the output cells are less able to discriminate between stimuli than in their initial untrained, random state. From the

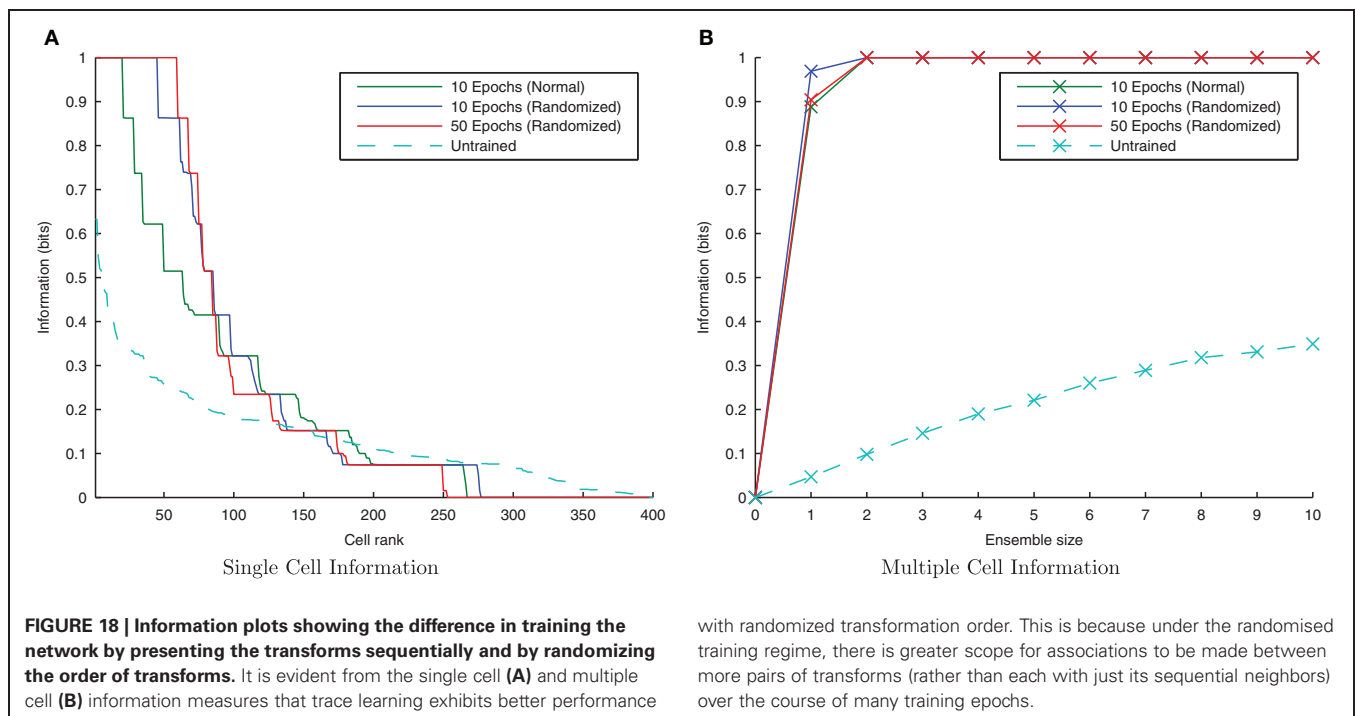




multiple-cell information plot (Figure 17B) it is clear that virtually all transforms of all stimuli have become associated together since even using the ten best single-cell information neurons barely raises the multiple cell information measure above 0-bits since the cells are unable to discriminate one stimulus from the other.

### 3.2.5. Randomized transform order

If the network is using a temporal trace to associate orthogonal transforms together, randomizing the order of those transforms within a stimulus block, but still presenting all transforms of one stimulus followed by all transforms of the other, should not significantly degrade its performance. Moreover,



there is good reason to expect that the performance should be improved slightly, as this training paradigm will help to associate each transform,  $S_1^n$ , with each other from the same stimulus rather than just its neighboring transforms,  $S_1^{n-1}$  and  $S_1^{n+1}$ .

It is clear from **Figure 18** that with the longer synaptic conductance time course ( $\tau_{EE} = 150$  ms) and the same degree of training, the randomized transform case has performed better than the standard non-randomized paradigm as expected, improving further with more epochs of training. In previous simulations this training paradigm proved difficult for the CT mechanism with an initially random set of feed-forward weights, since different pools of output neurons were stimulated by randomly ordered transforms (due to having less spatial overlap between consecutive transforms on average). In the case of randomly ordered transforms with the trace learning mechanism, however, the lower degree of spatial overlap is irrelevant as the same pool of output neurons is kept active for all the transforms of a particular stimulus by virtue of the longer synaptic time constants and the consecutive presentation of all transforms of a particular stimulus (albeit not necessarily in order).

#### 4. DISCUSSION

In the above simulations we have shown that a biologically realistic spiking neural network with STDP can operate in two very different ways to achieve transformation-invariant representations. These simulations lend more biological plausibility to the Trace and CT learning mechanisms, which may be utilized by the same model with slight differences in the training environment or the physical parameters of the neurons.

With short synaptic conductance time constants between the pyramidal neurons ( $\tau_{EE}$ ), the model works similarly to the CT learning mechanism. In this case, the network requires the transforms to be spatially overlapping (as a direct consequence of the learning mechanism) but can cope with interleaving the transforms of different stimuli and thus bears the characteristics of the equivalent rate-coded mechanism (Stringer et al., 2006). Importantly, this mechanism is sensitive to the strength of lateral inhibition, which under optimal conditions serves to maintain the synchronous firing of neurons representing the novel part of an unseen transform with those already potentiated from previous learning of another transform. Without this effect of lateral inhibition, these novel neurons will most likely fire outside the time window for significant LTP, and may possibly come after the postsynaptic neuron has fired leading to LTD.

Lengthening the very same synaptic conductance time constants ( $\tau_{EE}$ ), enables the model to work with a Trace learning mechanism. In this case the network uses temporal continuity to associate together orthogonal (completely non-overlapping) transforms and consequently fails to develop invariance and stimulus specificity if the transforms of different stimuli are interleaved. While these properties are the same as for the classic McCulloch-Pitts neuron, it is interesting to note that in such a rate-coded model, the trace term is associated with the presynaptic or (more commonly) postsynaptic neuron (Rolls and Milward, 2000). In contrast, in a conductance-based spiking neural network, the trace can instead be associated with the individual synapses between two connected neurons. This is a measurable property of biological neurons and suggests where to focus neurophysiological investigation aiming to understand invariance learning mechanisms.

While the Trace and CT learning mechanisms have been studied here in isolation, it seems likely that a combination of both would be employed to varying degrees depending upon the statistics of the inputs to each layer of the brain. In early layers (e.g., V1), the patterns of stimulation are likely to change more from one transform to another since the neurons here are highly specific in their sensitivity to a location and orientation. In later layers, however, such as Inferotemporal cortex (IT), the invariance built in the earlier layers will mean that inputs to these cells are less changeable from one transform to another. Having passed through several layers of pyramidal cells with lateral inhibition acting at each stage, the spike volleys representing a stimulus may also become more synchronized (Diesmann et al., 1999). Under these conditions, we therefore expect that as the similarity between transforms increases through the layers, the CT mechanism will become more prominent and trace effects will become less important, which would be evidenced by progressively quicker synaptic conductance decays (shorter time constants).

If it is the case that the ventral visual system uses an effective synaptic time constant between the two extremes presented in the simulations here, we would therefore predict that the type of learning occurring for any given stimuli would be highly dependent on how those stimuli are presented, for example with rapidly transforming (and hence spatially dissimilar successive views) leading to more of a Trace learning regime, whereas temporally separate exposures would require a high degree of similarity between the views for the CT mechanism to work.

The work presented here is a first step toward understanding how the Trace and CT learning rules may be utilised in a spiking neural network, and as such will naturally have limitations. So far, the model has been presented with orthogonal, non-overlapping “toy” stimuli rather than the more distributed, spatially overlapping stimuli found in the natural world. Whilst we acknowledge that these highly idealized representations are somewhat lacking in ecological validity, they were employed in order to isolate each learning mechanism in a precise and identifiable way. Further work would benefit however from exploring these learning mechanisms with more natural, spatially overlapping stimuli.

A further limitation concerns the Trace learning mechanism. By lengthening the time constant of the feed-forward synaptic conductances,  $\tau_{EE}$ , the excitatory activity reaching the output neurons decays more slowly and results in much higher firing rates in the output neurons (approximately 200 spikes/s) than in the CT simulations (approximately 50 spikes/s). While these rates are still within the realms of biological plausibility, they are towards the edge of it and so the conclusions would be on firmer ground through exploring additional mechanisms to reduce these high firing rates.

#### 4.1. FUTURE DIRECTIONS

In understanding the dynamics of learning transformation-invariant representations in spiking neural networks, we have only demonstrated *translation* invariance so far. A natural

extension to this body of work would therefore be to investigate this learning process with other kinds of transforms commonly found in natural visual scenes and investigated in rate-coded models including, for example, rotations (Stringer et al., 2006), occlusions (Stringer and Rolls, 2000) and changes in scale (Wallis and Rolls, 1997). This would provide a more general understanding of the variations in the problems of visual object recognition that the visual system must overcome.

Furthermore, the use of realistic 3D shapes and faces will also allow the model to be more directly compared to psychophysical data, both in terms of the effects on representations formed from exposure to realistic images (Simoncelli, 2003; David et al., 2004; Felsen and Dan, 2005; Felsen et al., 2005) and testing if invariance learning may be achieved at natural speeds of transformation (e.g., rotation). Neuronal parameters such as the synaptic time constants (e.g.,  $\tau_{EE}$ ) and the learning time constants ( $\tau_C$  and  $\tau_D$ ) may be crucial to invariance learning with realistic stimuli. Exploring the interaction between the speed of transformation of objects and the parameters of the model should lead to concrete predictions which may be tested against neurophysiological data. For example this may reveal an upper-threshold of stimulus movement speed which still allows transformation-invariant representations to form, or even that our visual systems typically use a number of static views to learn invariance.

Natural stimuli will also test the model’s ability to learn transformation-invariant representations with effectively distributed, overlapping representations rather than the orthogonal non-overlapping representations employed so far. This would mean that the network could no longer appear to solve the problem through learning about retinal location.

In addition to enhancing the ecological validity of the stimuli and their presentation paradigm, the model itself could be modified to incorporate additional features found in its biological counterpart including lateral excitatory connectivity, cell firing-rate adaptation and multiple layers of feed-forward weights, some or all of which may prove to be necessary for solving the more complex invariance learning problems, for instance, with natural scenes composed of multiple objects.

## 5. CONCLUSION

In the work presented here, we have demonstrated how a spiking neural network may exhibit two very different modes of invariance learning, which share the characteristic properties of their rate-coded counterparts. This was achieved in a single model by changing, (most notably), the time constant of the feed-forward synaptic conductances and the properties of the stimulus sets. Through developing more biologically accurate spiking models in this way, we may build upon incites from previous work to more fully understand the detailed mechanisms of visual invariance learning in the brain.

## ACKNOWLEDGMENTS

This research was supported by the ESRC and the Wellcome Trust.

## REFERENCES

- Abbott, L. F., and Nelson, S. B. (2000). Synaptic plasticity: taming the beast. *Nat. Neurosci.* 3, 1178–1183.
- Amit, D. J., and Brunel, N. (1997). Dynamics of a recurrent network of spiking neurons before and following learning. *Network* 8, 373–404.
- Bell, A. J., and Sejnowski, T. J. (1997). The “independent components” of natural scenes are edge filters. *Vision Res.* 37, 3327–3338.
- Bi, G.-Q., and Poo, M.-m. (1998). Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type. *J. Neurosci.* 18, 10464–10472.
- Bialek, W., Rieke, F., de Ruyter van Steveninck, R. R., and Warland, D. (1991). Reading a neural code. *Science* 252, 1854–1857.
- Booth, M. C. A., and Rolls, E. T. (1998). View-invariant representations of familiar objects by neurons in the inferior temporal visual cortex. *Cereb. Cortex* 8, 510–523.
- Dan, Y., and Poo, M.-m. (2006). Spike timing-dependent plasticity: from synapse to perception. *Physiol. Rev.* 86, 1033–1048.
- David, S. V., Vinje, W. E., and Gallant, J. L. (2004). Natural stimulus statistics alter the receptive field structure of V1 neurons. *J. Neurosci.* 24, 6991–7006.
- Desimone, R. (1991). Face-selective cells in the temporal cortex of monkeys. *J. Cogn. Neurosci.* 3, 1–8.
- Diesmann, M., Gewaltig, M. O., and Aertsen, A. (1999). Stable propagation of synchronous spiking in cortical neural networks. *Nature* 402, 529–533.
- Elliffe, M. C. M., Rolls, E. T., and Stringer, S. M. (2002). Invariant recognition of feature combinations in the visual system. *Biol. Cybern.* 86, 59–71.
- Felsen, G., and Dan, Y. (2005). A natural approach to studying vision. *Nat. Neurosci.* 8, 1643–1646.
- Felsen, G., Touryan, J., Han, F., and Dan, Y. (2005). Cortical sensitivity to visual features in natural scenes. *PLoS Biol.* 3:e342. doi: 10.1371/journal.pbio.0030342
- Ferster, D., and Spruston, N. (1995). Cracking the neuronal code. *Science* 270, 756–757.
- Földiák, P. (1991). Learning invariance from transformation sequences. *Neural Comput.* 3, 194–200.
- Földiák, P. (1993). “The ‘ideal homunculus’: statistical inference from neuronal population responses,” in *Computation and Neural Systems*, eds F. H. Eeckman and J. M. Bower (Norwell, MA: Kluwer Academic Publishers), 55–60.
- Fries, P., Schröder, J. H., Roelfsema, P. R., Singer, W., and Engel, A. K. (2002). Oscillatory neuronal synchronization in primary visual cortex as a correlate of stimulus selection. *J. Neurosci.* 22, 3739–3754.
- Fukushima, K. (1988). Neocognitron: a hierarchical neural network capable of visual pattern recognition. *Neural Netw.* 1, 119–130.
- Gerstner, W., and Kistler, W. (2006). *Spiking Neuron Models: Single Neurons, Populations, Plasticity*, 3rd Edn. Cambridge, UK: Cambridge University Press.
- Hebb, D. O. (1949). *The Organization of Behaviour: A Neuropsychological Theory*. New York, NY: Wiley.
- Hubel, D. H., and Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *J. Physiol.* 195, 215–243.
- Ito, M., Tamura, H., Fujita, I., and Tanaka, K. (1995). Size and position invariance of neuronal response in monkey inferotemporal cortex. *J. Neurophysiol.* 73, 218–226.
- Izhikevich, E. M. (2003). Simple model of spiking neurons. *IEEE Trans. Neural Netw.* 14, 1569–1572. Available online at: <http://vesicle.nsi.edu/users/izhikevich/publications/whichmod.htm>
- Kreiter, A. K., and Singer, W. (1996). Stimulus-dependent synchronization of neuronal responses in the visual cortex of the awake macaque monkey. *J. Neurosci.* 16, 2381–2396.
- Kuwabara, N., and Suga, N. (1993). Delay lines and amplitude selectivity are created in subthalamic auditory nuclei: the brachium of the inferior colliculus of the mustached bat. *J. Neurophysiol.* 69, 1713–1724.
- Maass, W., and Bishop, C. M. (eds.) (1999). *Pulsed Neural Networks*. Cambridge, MA: MIT Press.
- MacKay, D. J. C. (2003). *Information Theory, Inference and Learning Algorithms*. Cambridge, CA: Cambridge University Press.
- Markram, H., Lübke, J., Frotscher, M., and Sakmann, B. (1997). Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs. *Science* 275, 213–215.
- Masquelier, T., Hugues, E., Deco, G., and Thorpe, S. J. (2009). Oscillations, phase-of-firing coding, and spike timing-dependent plasticity: an efficient learning scheme. *J. Neurosci.* 29, 13484–13493.
- McCormick, D. A., Connors, B. W., Lighthall, J. W., and Prince, D. A. (1985). Comparative electrophysiology of pyramidal and sparsely spiny stellate neurons of the neocortex. *J. Neurophysiol.* 54, 782–806.
- McCulloch, W., and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biol.* 5, 115–133.
- Mel, B. (1997). Seemore: combining color, shape, and texture histogramming in a neurally inspired approach to visual object recognition. *Neural Comput.* 9, 777–804.
- Michler, F., Eckhorn, R., and Wachtler, T. (2009). Using spatiotemporal correlations to learn topographic maps for invariant object recognition. *J. Neurophysiol.* 102, 953–964.
- Op de Beeck, H., and Vogels, R. (2000). Spatial sensitivity of macaque inferior temporal neurons. *J. Comp. Neurol.* 426, 505–518.
- Perrinet, L. (2003). *Comment Déchiffrer le Code Impulsionnel de la Vision? Étude du Flux Parallèle, Asynchrone et éparé dans le Traitement Visuel Ultra-Rapide*. Ph.D. thesis, Université Paul Sabatier, Toulouse, France.
- Perrinet, L., Delorme, A., Samuelides, M., and Thorpe, S. J. (2001). Networks of integrate-and-fire neuron using rank order coding A: how to implement spike time dependent hebbian plasticity. *Neurocomputing* 38–40, 817–822.
- Quiroga, R. Q., Reddy, L., Kreiman, G., Koch, C., and Fried, I. (2005). Invariant visual representation by single neurons in the human brain. *Nature* 435, 1102–1107.
- Riesenhuber, M., and Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nat. Neurosci.* 2, 1019–1025.
- Rolls, E. T., and Milward, T. (2000). A model of invariant object recognition in the visual system: learning rules, activation functions, lateral inhibition, and information-based performance measures. *Neural Comput.* 12, 2547–2572.
- Rolls, E. T., and Treves, A. (1998). *Neural Networks and Brain Function*. Oxford University Press, Oxford.
- Rolls, E. T., Treves, A., and Tovee, M. J. (1997). The representational capacity of the distributed encoding of information provided by populations of neurons in primate temporal visual cortex. *Exp. Brain Res.* 114, 149–162.
- Rullen, R. V., and Thorpe, S. J. (2001). Rate coding versus temporal order coding: what the retinal ganglion cells tell the visual cortex. *Neural Comput.* 13, 1255–1283.
- Šima, J., and Orponen, P. (2003). General-purpose computation with neural networks: a survey of complexity theoretic results. *Neural Comput.* 15, 2727–2778.
- Simioncelli, E. P. (2003). Vision and the statistics of the visual environment. *Curr. Opin. Neurobiol.* 13, 144–149.
- Stringer, S. M., Perry, G., Rolls, E. T., and Proskew, J. H. (2006). Learning invariant object recognition in the visual system with continuous transformations. *Biol. Cybern.* 94, 128–142.
- Stringer, S. M., and Rolls, E. T. (2000). Position invariant recognition in the visual system with cluttered environments. *Neural Netw.* 13, 305–315.
- Tanaka, K. (1996). Representation of visual features of objects in the inferotemporal cortex. *Neural Netw.* 9, 1459–1475.
- Tanaka, K., Saito, H., Fukada, Y., and Morioka, M. (1991). Coding visual images of objects in the inferotemporal cortex of the macaque monkey. *J. Neurophysiol.* 66, 170–189.
- Thorpe, S., Fize, D., and Marlot, C. (1996). Speed of processing in the human visual system. *Nature* 381, 520–522.
- Thorpe, S. J., Delorme, A., Van Rullen, R., and Paquier, W. (2000). “Reverse engineering of the visual system using networks of spiking neurons,” in *Proceedings. ISCAS 2000 Geneva. The 2000 IEEE International Symposium on Circuits and Systems (2000)*, Vol. 4, (Geneva), 405–408.
- Tovee, M. J., Rolls, E. T., and Azzopardi, P. (1994). Translation invariance in the responses to faces of single neurons in the temporal visual cortical areas of the alert macaque. *J. Neurophysiol.* 72, 1049–1060.
- Troyer, T. W., Krukowski, A. E., Priebe, N. J., and Miller, K. D. (1998). Contrast-invariant orientation tuning in cat visual cortex:

- thalamocortical input tuning and correlation-based intracortical connectivity. *J. Neurosci.* 18, 5908–5927.
- van Hateren, J. H., and van der Schaaf, A. (1998). Independent component filters of natural images compared with simple cells in primary visual cortex. *Proc. Biol. Sci.* 265, 359–366.
- van Rossum, M. C., Bi, G.-Q., and Turrigiano, G. G. (2000). Stable hebbian learning from spike timing-dependent plasticity. *J. Neurosci.* 20, 8812–8821.
- Wallis, G., and Rolls, E. T. (1997). Invariant face and object recognition in the visual system. *Prog. Neurobiol.* 51, 167–194.
- Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.
- Received: 10 February 2012; accepted: 25 June 2012; published online: 25 July 2012.
- Citation: Evans BD and Stringer SM (2012) Transformation-invariant visual representations in self-organizing spiking neural networks. *Front. Comput. Neurosci.* 6:46. doi: 10.3389/fncom.2012.00046
- Copyright © 2012 Evans and Stringer. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in other forums, provided the original authors and source are credited and subject to any copyright notices concerning any third-party graphics etc.





# Learning view invariant recognition with partially occluded objects

James M. Tromans\*, Irina Higgins and Simon M. Stringer

Experimental Psychology, Oxford Foundation for Theoretical Neuroscience and Artificial Intelligence, University of Oxford, Oxford, UK

## Edited by:

Evgeniy Bart, Palo Alto Research Center, USA

## Reviewed by:

Evgeniy Bart, Palo Alto Research Center, USA  
Jay Hegd , Georgia Health Sciences University, USA

## \*Correspondence:

James M. Tromans, Experimental Psychology, Oxford Foundation for Theoretical Neuroscience and Artificial Intelligence, University of Oxford, Oxford, UK.  
e-mail: james.tromans@psy.ox.ac.uk

This paper investigates how a neural network model of the ventral visual pathway, VisNet, can form separate view invariant representations of a number of objects seen rotating together. In particular, in the current work one of the rotating objects is always partially occluded by the other objects present during training. A key challenge for the model is to link together the separate partial views of the occluded object into a single view invariant representation of that object. We show how this can be achieved by Continuous Transformation (CT) learning, which relies on spatial similarity between successive views of each object. After training, the network had developed cells in the output layer which had learned to respond invariantly to particular objects over most or all views, with each cell responding to only one object. All objects, including the partially occluded object, were individually represented by a unique subset of output cells.

**Keywords:** object recognition, continuous transformation, occlusion, inferior temporal cortex

## 1. INTRODUCTION

It is important to understand how invariant representations of individual objects are built in the primate visual system even when multiple objects are present in natural scenes. Neurophysiological research has provided substantial evidence showing that over successive stages, the visual system develops neurons that respond with view, size, and position (translation) invariance to objects or faces (Desimone, 1991; Tanaka et al., 1991; Rolls, 1992, 2000; Perrett and Oram, 1993; Rolls and Deco, 2002). For example, it has been shown that the inferior temporal visual cortex has neurons that respond to faces and objects with translation (Kobatake and Tanaka, 1994; Tovee et al., 1994; Ito et al., 1998; Op De Beeck and Vogels, 2000), and view (Hasselmo et al., 1989; Booth and Rolls, 1998) invariance.

The “biased competition hypothesis” of attention suggested that feedback connections are necessary to build separate representations of individual objects in a complex scene by providing the mechanism for attentional selection (Rolls and Deco, 2002). However, it has been shown that this separation can be achieved without the need for an attentional mechanism using purely feed-forward connectivity in a hierarchical neural network model of the ventral visual pathway, VisNet (Stringer et al., 2007). The statistical properties of the input stimuli play a crucial role, whereby the features within individual objects occur more frequently together than the features between different objects. As such, although the role of feedback connections is an important area for future research, they will not be implemented in the present study.

Stringer and Rolls (2000) showed that a hierarchical neural network model of the ventral visual pathway, VisNet, could recognize objects presented against natural cluttered scenes, providing the model had been previously trained with each object presented individually transforming against a blank background. However, the network failed to learn to recognize individual objects if the

objects were presented against a natural cluttered background during training.

Recent studies by Stringer and Rolls (2008) and Stringer et al. (2007) have shown how VisNet may cope with complex scenes during training, and learn invariant representations of individual objects even when no single object is seen in isolation. These modeling studies used the statistics of the natural environment where features within an object occur together more frequently than features between different objects. Specifically, VisNet could learn invariant representations of individual objects if different combinations of transforming objects were seen at different times.

However, a further challenge is to explain how invariant representations can be learned when the objects are partially occluded by one another during learning. Stringer et al. (2007) proposed that Continuous Transformation (CT) learning (Stringer et al., 2006) combined with the statistical independence of objects presented in different combinations might allow the network to solve this problem. Specifically, consider presenting a number of objects to the network in different subset combinations, but where one of the objects is always partially occluded by whichever objects it is currently shown with. The hypothesis is that the network will simultaneously form separate representations of all of the different objects, where an invariant representation of the partially occluded object is formed by linking together the different partial views through CT learning. However, Stringer et al. (2007) provided no simulation evidence that this could work. In this paper we demonstrate for the first time this process operating with simulated three dimensional rotating objects. It is important to investigate this issue because objects in the natural environment will often overlap. This task is more difficult than simply forming separate representations of different objects because, in order for the network to build a complete invariant representation of the partially occluded object, the network has to link together

the different partial views of the object as well as separate these partial views from the other objects present.

In the simulations described below, we show how VisNet can form separate view invariant representations of individual objects seen rotating together, where one of the rotating objects is always partially occluded by the other objects present during training. The network develops cells in the output layer which have learned to respond invariantly to particular objects over most or all views, with each cell responding to only one object. All objects, including the partially occluded object, are individually represented in this way by a unique subset of output cells. This learning process relies on the statistical independence of the objects that are shown in different combinations, as well as an invariance learning mechanism known as CT learning that is described next.

## 2. MATERIALS AND METHODS

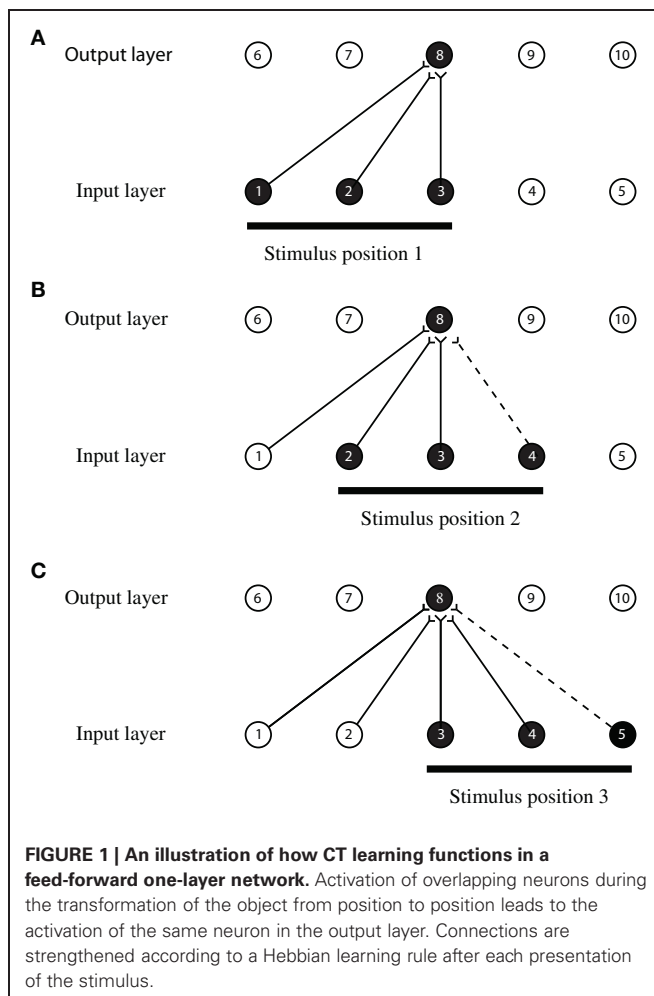
### 2.1. CONTINUOUS TRANSFORMATION LEARNING

A leading computational theory of how the ventral visual pathway in the brain may develop neurons that respond to objects with transform (e.g., view or location) invariance is CT learning. CT learning uses an associative (Hebbian) synaptic modification rule (Stringer et al., 2006) that can exploit the image similarity across successive transforms (e.g., views) of a continuously transforming object in order to develop output neurons which respond to the object over all transforms. Because CT learning is based on the standard Hebbian learning rule, it is biologically plausible.

An idealized version of the CT learning process outlining the theoretical principle is illustrated in **Figure 1** and operates as follows. The network shown has an input layer where stimuli are presented, and an output layer where transform invariant representations develop through learning. The output layer operates as a competitive network, where individual cells send inhibitory projections to the other cells in this layer (not shown in **Figure 1**), and thereby compete with each other. Initially, the weights of the feedforward synaptic connections are set to random values. Then, during learning, a stimulus is initially presented in position 1 (shown in **Figure 1A**) and is represented by three active neurons in the input layer (neurons 1, 2, and 3). Activity propagates through the random feedforward connections to the output layer, where one of the neurons, say neuron 8, wins the competition. The simultaneous activation of neurons in the input and output layers causes the synaptic connections between them to become strengthened according to a Hebbian learning rule

$$\delta w_{ij} = \alpha y_i x_j \quad (1)$$

where  $\delta w_{ij}$  is the increment in the synaptic weight  $w_{ij}$ ,  $y_i$  is the firing rate of the post-synaptic neuron  $i$ ,  $x_j$  is the firing rate of the pre-synaptic neuron  $j$ , and  $\alpha$  is the learning rate. To restrict and limit the growth of each neuron's synaptic weight vector,  $w_i$  for the  $i$ th neuron, its length is normalized at the end of each timestep during training as is usual in competitive learning (Hertz et al., 1991). This is necessary to ensure that one or a few neurons do not always win the competition. If there was no normalization of synaptic weights during a simple Hebbian learning procedure, just a few neurons may eventually learn to respond strongly to nearly all of the input patterns. Neurophysiological evidence for



synaptic weight normalization is provided by Royer and Pare (2003).

As the stimulus moves from position 1 to position 2 (shown in **Figure 1B**), it causes activation in the input layer to also move along one neuron at a time. Therefore, when the stimulus is in position 2, it causes neurons 2, 3, and 4 to become active. The overlap in the input space allows two neurons in the input layer to remain active (neurons 2 and 3) during both transformations. The activation of the same neurons in the input layer causes the same neuron in the output layer (neuron 8) to become active again because the connections have already been strengthened when the stimulus was in position 1. The simultaneous activation of the output neuron, with input neurons 2, 3, and the additional input neuron 4 causes their synaptic connections to become strengthened according to the Hebbian learning rule. Therefore, the activation of neuron 8 will now become associated with the activation of neurons 2, 3, and 4. As the stimulus continues to move from one position to the next, the process repeats itself and the same neuron in the output layer remains activated. This output neuron becomes a position invariant neuron. A more comprehensive description of CT learning and simulation results in the context of invariant object recognition is provided by Stringer et al. (2006) and Perry et al. (2006).

## 2.2. LEARNED OBJECT SELECTIVITY

CT learning develops transform invariant representations that are object-specific. That is, as long as each object is not always presented together with another particular object transforming in lock-step during training, individual neurons typically learn to respond to one object only (Stringer et al., 2006). Consider an object rotating at a particular retinal location during training. Successive views of the object are represented by the outputs of the oriented input filters representing V1 simple cells as described in Equation 2. CT learning utilizes Hebbian competitive learning. At each presentation of a view of an object during training, activity is propagated up through successive neuronal layers in the network. Within each layer, a small subset of neurons wins the competition.

The feedforward synaptic connections from the input filters to the first layer of neurons are modified according to a Hebbian learning rule (Equation 1). This learning rule strengthens only the synaptic connections from those V1 filters that are activated by the particular visual form of the object view currently presented. The weight vector of each of the first layer neurons gradually shifts during learning to point in the same direction as the V1 input pattern(s) that it is learning to respond to. Since each first layer neuron computes its activation (Equation 3) according to the dot product of its weight vector and the current input pattern from the V1 input layer, after training each neuron will respond in proportion to the similarity between the current input pattern and the input pattern(s) the neuron learned to respond to during training. That is, each neuron will respond maximally to the input pattern that it has learned to respond to during training, and generalize to other input patterns depending on their similarity (Hertz et al., 1991).

The competitive Hebbian learning rule operates in a similar manner for the feedforward connections between all of the later layers of the network. This ensures that a subset of neurons in layer 1 and all successive layers learn to respond to the pattern of visual features present in the current view of the trained object. This means that the subset of output neurons in the higher layers of VisNet learn to respond to the visual form of the current view of the trained object and not its retinal location.

If output neurons simply learned to respond to retinal location, then the feedforward connections would need to be strengthened from *all* of the V1 filter inputs in a particular location regardless of the visual form of the objects. But this cannot occur because the Hebbian learning rule ensures that only the synaptic connections coming from those V1 input filters actually activated by the particular visual form of the object can be strengthened.

As described above in the Materials and Methods section on CT learning, the Hebbian learning rule is able to learn to associate different views of the object onto the same active output neurons as long as the different object images presented during training cover a space of smoothly changing views. Again, only those V1 input filters that were activated by the different object views can become associated with the active subset of neurons in the higher layers. So, even after many stimulus views have been presented, the neurons in the later layers of the network cannot learn to respond to all of the V1 filters in a particular retinal location. Thus, after training, the output neurons become object-specific.

The output neurons will respond maximally to different views of the particular object that has been learned.

If another different untrained object is presented in the same retinal location as the first trained object, then there will be a rather different pattern of V1 input filters activated. However, as discussed above, the output of each of the neurons in the network reflects the similarity between the input pattern in the previous layer that it learned to respond to during training and the currently tested input pattern. This means that the neurons in the higher layers that have been previously trained to respond to the first object will respond to the second untrained object in proportion to the degree of visual similarity between the two objects. Therefore, due to the properties of Hebbian competitive learning, the neurons through the higher layers of the network must learn to respond the visual forms of objects rather than locations.

However, there is a potential conflict between the need to develop representations that are object-specific and the need to develop transform invariant representations within each object. In principle, if two different objects have similar transforms, then the CT learning mechanism may encourage output neurons to learn to respond invariantly across both objects. This is a fundamental issue with CT learning, which we are continuing to investigate. In simulation studies, we have found that increasing the size of the VisNet architecture improves the ability of the model to learn separate representations of similar faces for example. It is also possible that combining CT learning with a trace learning rule (Foldiak, 1991) could improve the ability of the network to form separate invariant representations of different objects. Although, this has not been implemented in the simulations reported here, which use only a standard Hebbian learning rule.

## 2.3. MULTIPLE OBJECTS

How the brain can build invariant representations of individual objects even when multiple objects are present in a scene is a very important question in natural vision. How the visual system learns about individual objects rather than the combination of objects that make up the scene has only recently been investigated successfully in a biologically realistic model (Stringer et al., 2007; Stringer and Rolls, 2008). The features that make up a given object occur together more frequently when presented during training compared to the features that make up different objects. Depending on how often a given object is presented during training with another object, the frequency of how often features between these two different objects occur together will vary. However, the features that make up any individual object are always presented with one another and are therefore completely correlated.

It has been shown that a competitive network will operate usefully in this situation. The network will learn primarily to form representations that reflect the high probability of co-occurrence of features from one object and do not reflect the features of other objects presented simultaneously during training if the object being trained is seen much more frequently than it is presented with any other object.

In order for a competitive network to build representations of individual objects, there must be a statistical decoupling between

the features that comprise each of the objects presented during training. Providing that there are a sufficient number of objects present during training, each object will be presented with many other objects and therefore the features within the object will appear together significantly more often than they are coupled with features of any other object. It has been previously demonstrated that this allows a competitive network to form transform invariant representations of individual objects, rather than the combinations of objects seen during training, by a mechanism such as CT learning (Stringer and Rolls, 2008).

#### 2.4. LEARNING TO RECOGNISE PARTIALLY OCCLUDED TRANSFORMING OBJECTS

The major new problem addressed in this paper is how VisNet can form separate view invariant representations of individual objects seen rotating together, where one of the rotating objects is always partially occluded by the other objects present during training. To create view invariant representations of the occluded object, the network will have to separate it from the occluding objects and link together different partial views to create a representation of the whole object. The potential solution described by Stringer and Rolls (2008) for separating out individual objects in a scene with multiple objects present will be used to separate the occluded and the occluding objects. CT learning will be used to link together the different transforms of each individual object, including associating together the occluded and unoccluded views. By training VisNet with multiple objects that partially occlude one another, we show that our model of the ventral visual stream is able to learn reliably in increasingly realistic visual environments.

#### 2.5. OBJECTS

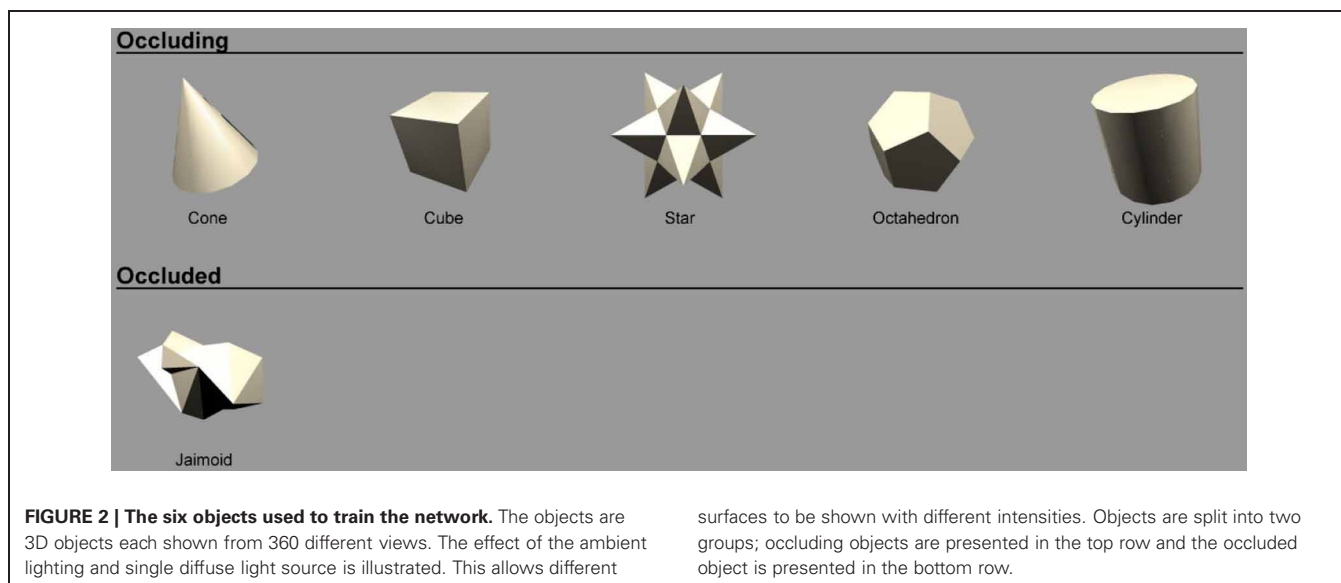
**Figure 2** shows the objects used to train the network. There were  $N = 6$  continuously rotating 3D objects on a gray background. Previous research (Stringer and Rolls, 2008) has shown that  $N = 6$  objects is sufficient to allow VisNet to develop

representations of individual objects when the network was trained on object pairs. The objects were designed and created using the 3D modeling tool Swift 3D 5.4. Ambient lighting with a diffuse light source was added to allow different surfaces to be shown with different intensities. Each object rotated in depth around the vertical axis in  $1^\circ$  steps over  $360^\circ$ . This step size was chosen because past research (Stringer et al., 2006) has revealed that it was sufficiently small for CT learning to operate. The 360 views of each object were then exported as 2D JPG images and encapsulated as Adobe Shock Wave Files. The objects were then aligned and organized using Adobe Flash CS4.

The stimulus set was comprised of five occluding objects and one occluded object. During each training sequence, the occluded object was shown rotating with one of the occluding objects. In all cases, each object would rotate about its own vertical axis and, therefore, all axes were in parallel with one another. The spatial arrangement of the objects is shown in **Figure 3**. The occluding objects were presented in a pentagon formation. The occluded object, the Jaimoid (irregular multifaceted three dimensional object, **Figure 4**), was always presented in the center of the pentagon.

Each of the occluding objects was placed at one of the five points of the pentagon, partially overlapping the Jaimoid at the center. The occluding objects were equidistant from the center of the occluded object, therefore occluding it to the same extent. The occluded object was always behind the occluding objects and in the middle of the pentagon formation. This spatial formation was chosen because it was necessary to ensure that different parts of the occluded object were covered by the five occluding objects.

In our simulations the objects were rotating at the same speeds. However, the correlations that would arise between corresponding view points of the objects are broken due to the fact that objects are paired with different objects on different occasions. This allows the network to form separate representations of different objects in the output layer.





## 2.6. THE VISNET MODEL

The model architecture (VisNet) implemented by Wallis et al. (1993) and Wallis and Rolls (1997) that is used to investigate the properties of CT learning in this paper is based on the following: (1) A series of hierarchical competitive networks with local graded inhibition. (2) Convergent connections to each neuron from a topologically corresponding region of the preceding layer, leading to an increase in the receptive field size of neurons through the visual processing areas. (3) Synaptic plasticity based on a Hebb-like learning rule. Model simulations which incorporated these hypotheses with a modified associative learning rule to incorporate a short term memory trace of previous neuronal activity (Foldiak, 1991) were shown to be capable of producing object-selective but translation and view invariant representations (Wallis and Rolls, 1997; Rolls and Milward, 2000; Rolls and Stringer, 2001).

CT learning and trace learning are two biologically plausible learning mechanisms that have been used to model invariance learning in the visual system. Each may explain how neurons at the end of the ventral visual pathway learn to respond to visual stimuli with transform (e.g., position or view) invariance. With CT learning (Stringer et al., 2006), a standard Hebb learning rule is able to encourage output neurons to learn to respond invariantly across different transforms of an object. CT learning

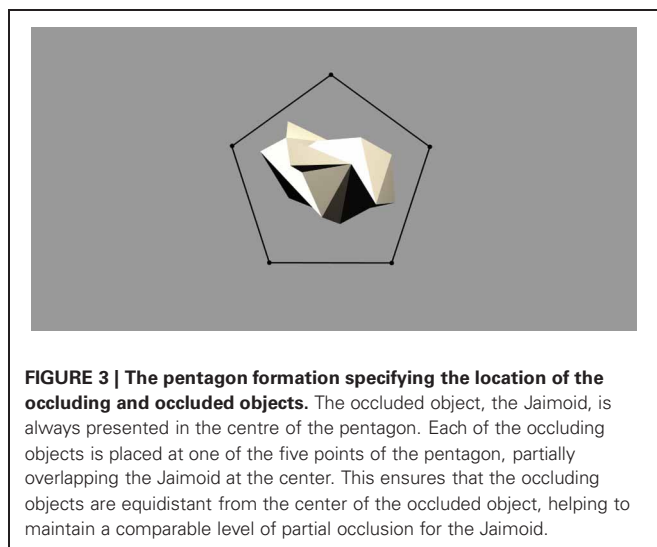
utilizes the spatial overlap or similarity between different transforms of an object in order to produce invariant responses. In contrast, the trace learning rule (Foldiak, 1991) incorporates a memory trace of recent neuronal activity, which is able to exploit the temporal continuity of the different transforms of an object in order to produce invariant responses. The trace learning rule assumes that in the natural visual world different transforms of an object tend to occur close together in time. In this paper, we will explore only the performance of the CT learning mechanism, which relies on the simpler Hebb learning rule.

The CT learning principle in the model architecture (VisNet) uses only spatial continuity in the input objects to drive the Hebbian associative learning with no temporal trace. In principle, the CT learning mechanism we describe could operate in various forms of feedforward neural network, with different forms of associative learning rule or different ways of implementing competition between neurons within each layer.

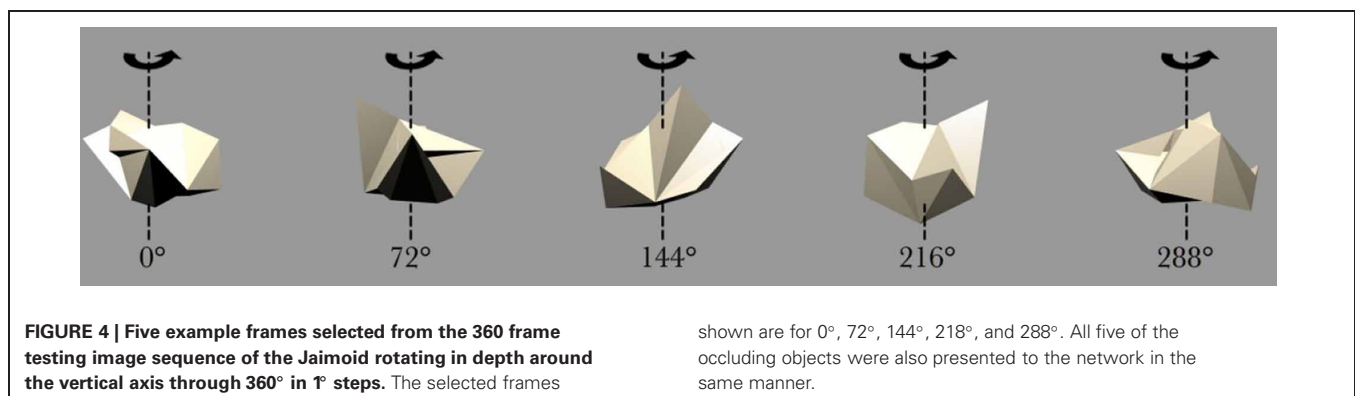
The model consists of a hierarchical series of four layers of competitive networks that are intended to model the hierarchy of processing areas in the ventral visual stream, which include V2, V4, the posterior inferior temporal cortex, and the anterior inferior temporal cortex, as shown in **Figure 5**. The forward connections to individual cells are derived from a topologically corresponding region of the preceding layer, using a Gaussian distribution of connection probabilities. These distributions are defined by a radius which will contain approximately 67% of the connections from the preceding layer. The values used are given in **Table 1**.

Before the objects are presented to the network's input layer they are pre-processed by a set of input filters which accord with the general tuning profiles of simple cells in V1. The filters provide a unique pattern of filter outputs for each transform of each visual object, which is passed through to the first layer of VisNet. The input filters used are computed by weighting the difference of two Gaussians by a third orthogonal Gaussian according to the following:

$$\Gamma_{xy}(\rho, \theta, f) = \rho \left[ e^{-\left(\frac{x \cos \theta + y \sin \theta}{\sqrt{2}/f}\right)^2} - \frac{1}{1.6} e^{-\left(\frac{x \cos \theta + y \sin \theta}{1.6\sqrt{2}/f}\right)^2} \right] e^{-\left(\frac{x \sin \theta - y \cos \theta}{3\sqrt{2}/f}\right)^2} \quad (2)$$



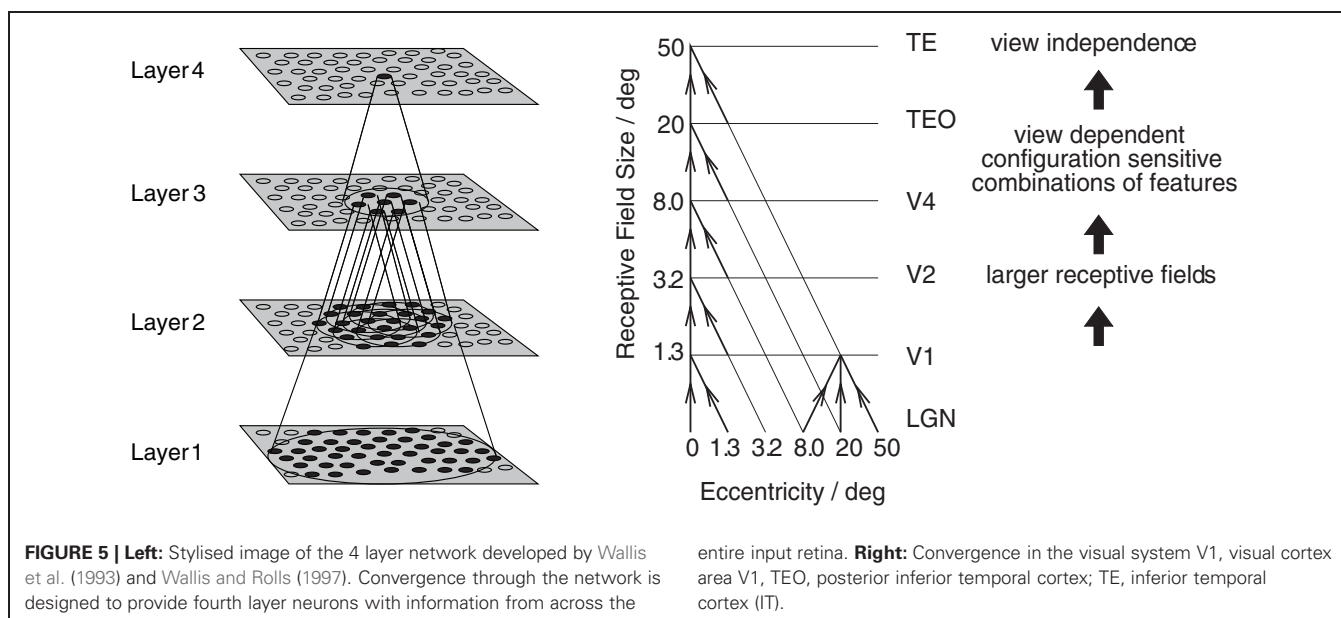
**FIGURE 3 | The pentagon formation specifying the location of the occluding and occluded objects.** The occluded object, the Jaimoid, is always presented in the centre of the pentagon. Each of the occluding objects is placed at one of the five points of the pentagon, partially overlapping the Jaimoid at the center. This ensures that the occluding objects are equidistant from the center of the occluded object, helping to maintain a comparable level of partial occlusion for the Jaimoid.



**FIGURE 4 | Five example frames selected from the 360 frame testing image sequence of the Jaimoid rotating in depth around the vertical axis through 360° in 1° steps.** The selected frames

shown are for 0°, 72°, 144°, 218°, and 288°. All five of the occluding objects were also presented to the network in the same manner.





**Table 1 | Network dimensions showing the number of connections per neuron and the radius in the preceding layer from which 67% are received.**

	Dimensions	Number of connections	Radius
Layer 4	32 × 32	100	12
Layer 3	32 × 32	100	9
Layer 2	32 × 32	100	6
Layer 1	32 × 32	272	6
Retina	128 × 128 × 32	–	–

where  $f$  is the filter spatial frequency,  $\theta$  is the filter orientation, and  $\rho$  is the sign of the filter, i.e.,  $\pm 1$ . Individual filters are tuned to spatial frequency (0.0625–0.5 cycles/pixel); orientation ( $0^\circ$ – $135^\circ$  in steps of  $45^\circ$ ); and sign ( $\pm 1$ ). Example input filters are shown in **Figure 6**. In previous studies, we have found that four filter orientations  $\theta$  is the minimal number needed to distinguish effectively between different visual objects presented to the retina. The number of layer 1 connections to each spatial frequency filter group is given in **Table 2**. Our model incorporates four octaves of filter frequencies. There are more connections from high frequency filters than low frequency filters. This enables the high frequency filters to cover a similar region of the input as the low frequency filters. Past neurophysiological research has shown that models based on difference-of-Gaussians functions are superior to those based on the Gabor function or the second differential of a Gaussian. Although the DOG-based models have more free parameters, they can account better for the variety of shapes of spatial contrast sensitivity functions observed in cortical cells and, unlike other models, they provide a detailed description of the organization of subregions of the receptive field that is consistent with the physiological constraints imposed by earlier stages in the visual pathway. (Hawken and Parker, 1987).

The activation  $h_i$  of each neuron  $i$  in the network is set equal to a linear sum of the inputs  $y_j$  from afferent neurons  $j$  weighted by the synaptic weights  $w_{ij}$ . That is,

$$h_i = \sum_j w_{ij} y_j \quad (3)$$

where  $y_j$  is the firing rate of neuron  $j$ , and  $w_{ij}$  is the strength of the synapse from neuron  $j$  to neuron  $i$ .

Within each layer, competition is graded rather than winner-take-all, and is implemented in two stages. First, to implement lateral inhibition, the activation  $h$  of neurons within a layer are convolved with a spatial filter,  $I$ , where  $\delta$  controls the contrast and  $\sigma$  controls the width, and  $a$  and  $b$  index the distance away from the center of the filter

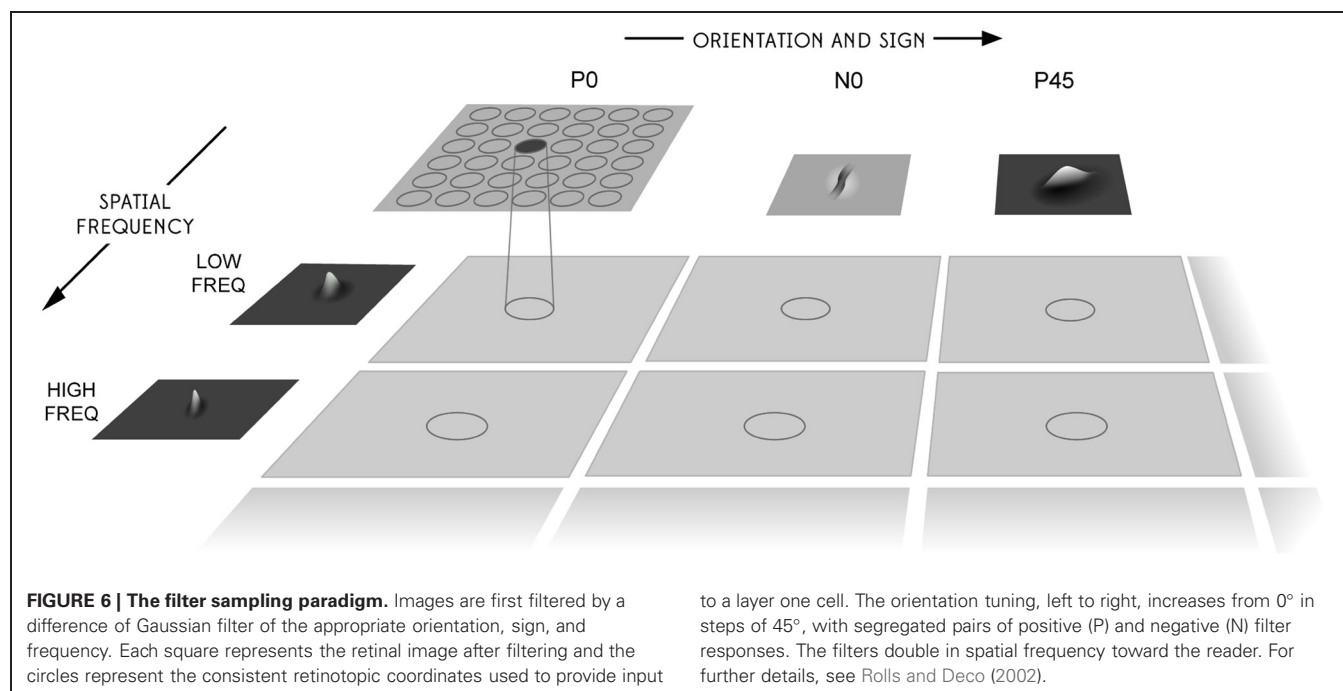
$$I_{a,b} = \begin{cases} -\delta e^{-\frac{a^2+b^2}{\sigma^2}} & \text{if } a \neq 0 \text{ or } b \neq 0, \\ 1 - \sum_{\substack{a \neq 0 \\ b \neq 0}} I_{a,b} & \text{if } a = 0 \text{ and } b = 0. \end{cases} \quad (4)$$

The lateral inhibition parameters are given in **Table 3**.

Next, contrast enhancement is applied by means of a sigmoid activation function

$$y = f^{\text{sigmoid}}(r) = \frac{1}{1 + e^{-2\beta(r-\alpha)}} \quad (5)$$

where  $r$  is the activation (or firing rate) after lateral inhibition,  $y$  is the firing rate after contrast enhancement, and  $\alpha$  and  $\beta$  are the sigmoid threshold and slope respectively. The parameters  $\alpha$  and  $\beta$  are constant within each layer, although  $\alpha$  is adjusted to control the sparseness  $a$  of the firing rates. The sparseness  $a$  of the firing within a layer can be defined, by extending the binary notion of



to a layer one cell. The orientation tuning, left to right, increases from 0° in steps of 45°, with segregated pairs of positive (P) and negative (N) filter responses. The filters double in spatial frequency toward the reader. For further details, see Rolls and Deco (2002).

**Table 2 | Layer 1 connectivity.**

Frequency	0.5	0.25	0.125	0.0625
Number of connections	201	50	13	8

The numbers of connections from each spatial frequency set of filters are shown. The spatial frequency is in cycles per pixel.

**Table 3 | Lateral inhibition parameters.**

Layer	1	2	3	4
Radius, $\sigma$	1.38	2.7	4.0	6.0
Contrast, $\delta$	1.5	1.5	1.6	1.4

the proportion of neurons that are firing, as

$$a = \frac{\left( \sum_{i=1}^N y_i / N \right)^2}{\sum_{i=1}^N y_i^2 / N} \quad (6)$$

where  $y_i$  is the firing rate of the  $i$ th neuron in the set of  $N$  neurons (Rolls and Treves, 1990, 1998). For the simplified case of neurons with binarized firing rates = 0/1, the sparseness is the proportion  $\in [0, 1]$  of neurons that are active. To set the sparseness to, say, 5% in VisNet simulations, the threshold  $\alpha$  is set to the value of the 95th percentile point of the activations within the layer.

The parameters for the sigmoid activation function are shown in **Table 4**. These fall squarely within the standard VisNet sigmoid parameter values which have been previously optimised to provide reliable and robust performance (Stringer et al., 2006, 2007; Stringer and Rolls, 2008).

**Table 4 | Sigmoid parameters.**

Layer	1	2	3	4
Percentile	95	95	88	91
Slope $\beta$	190	40	75	26

## 2.7. TRAINING PROCEDURE

The lateral inhibition and contrast enhancement stages of the VisNet model aim to simulate the function of inhibitory interneurons. In the brain, inhibitory interneurons effect direct competition between nearby excitatory cells within each layer of the ventral visual pathway. The way in which contrast enhancement is currently implemented in VisNet allows us to control the sparseness of firing rates within each layer. This is a useful aspect of the model, which allows us to explore the effects of sparseness on network performance. Although, it should be noted that the current contrast enhancement mechanism is not as realistic as implementing local inhibitory neurons explicitly because it is a global operation across each entire layer.

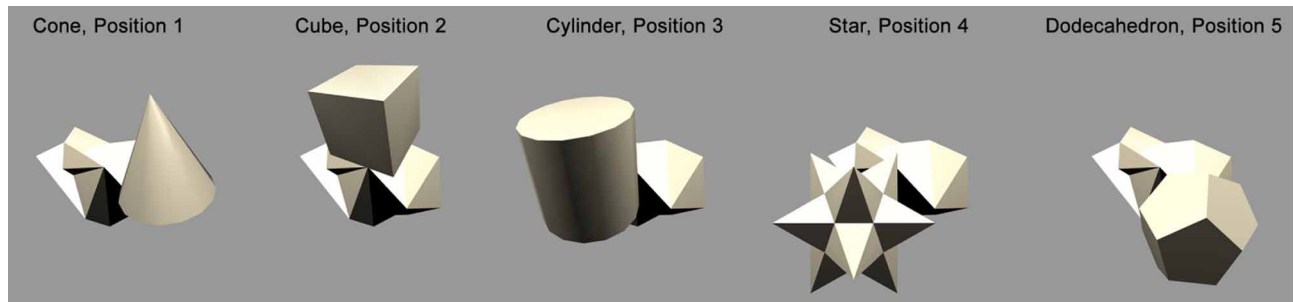
The occluded object, the Jaimoid, paired with each of the five surrounding occluding objects, is presented to VisNet with both objects rotating over 360° (**Figure 7**). Each full revolution over 360° of the pair is followed by the occluded object paired with a different occluding object in a different location around the pentagon formation. This process is repeated until the occluded object is paired with all five occluding objects. The rotating objects are presented as follows: cone, position 1; cube, position 2; cylinder, position 3; star, position 4; dodecahedron, position 5; Jaimoid, centrally, at position 6 (**Figures 8 and 9**).

In addition, all possible pairings of the five occluding objects are then presented in a similar fashion rotating over 360°. This helped VisNet to learn separate representations of the objects by



**FIGURE 7 | Five example frames selected from the 360 frame training image sequence of the Jaimoid and the Cone rotating through 360° in 1° steps.** The selected frames shown are for 0°, 72°, 144°, 216°, and 288°.

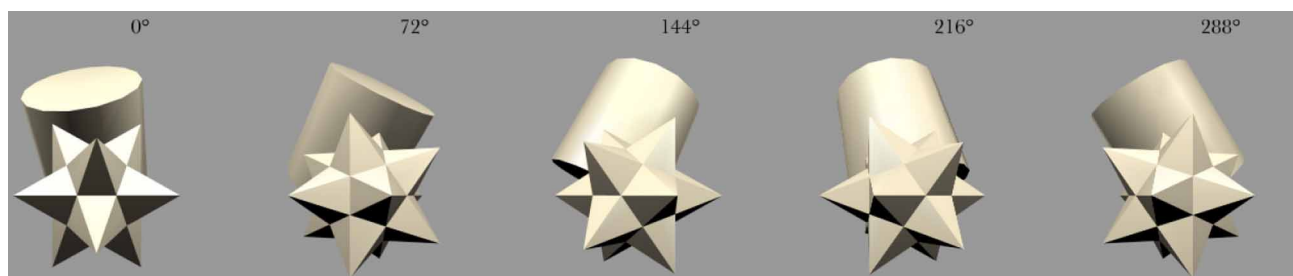
The Jaimoid is also occluded by the four other occluding objects in separate image sequences. Therefore, in total, there are five image sequences used during training, each containing 360 frames.



**FIGURE 8 | Five example frames of the Jaimoid occluded by all five occluding objects in their five corresponding positions.**

The occluding objects are arranged around the pentagon formation so that they are equidistant from the center of the Jaimoid. The rotating

objects used during training are presented in the same locations: cone, position 1; cube, position 2; cylinder, position 3; star, position 4; dodecahedron, position 5; Jaimoid, position 6.



**FIGURE 9 | Five example frames of two occluding stimuli (cylinder and star) rotating together, demonstrating the typical overlap between the occluding objects.**

using the statistics of the natural environment where the features within an object occur together more frequently than features between different objects (Stringer and Rolls, 2008). It should be noted that adjacent pairs of occluding objects would also sometimes overlap during training, leading to one occluding object being partially occluded by another occluding object.

At each image presentation, the activation of individual neurons within a layer is calculated, then their firing rates are calculated, and the feedforward synaptic weights between layers  $w_{ij}$  are updated according to Equation 1. This process is repeated for each layer in turn for all 4 layers of the VisNet model. One training epoch consists of the occluded object paired with all

five occluding objects across all 360 transforms followed by all possible pairings of the occluding objects rotating over all 360°.

In this manner, the network is trained one layer at a time starting with layer 1 and finishing with layer 4. Fifty training epochs were used for layers 1–4. The learning rate for layers 1–4 were 0.109, 0.1, 0.1, and 0.1, respectively.

Due to high computational expense, the Jaimoid was the only object that was partially occluded by its neighbours in the simulations described below. However, the underlying theory described above predicts that similar effects would be found if the simulations were repeated with more objects partially occluded by each other during training.

## 2.8. TESTING PROCEDURE

During testing, the synaptic weights within the model are fixed and cannot be altered. Firstly, in order to test whether VisNet had built an invariant representation of the central partially occluded object, the occluded object was presented individually, rotating around the vertical axis over  $360^\circ$  (**Figure 4**). The surrounding five occluding objects were also presented in isolation in a similar fashion to verify that VisNet had built invariant representations of these objects too. The neuronal outputs of the network were then recorded during the testing presentations of each view of each object.

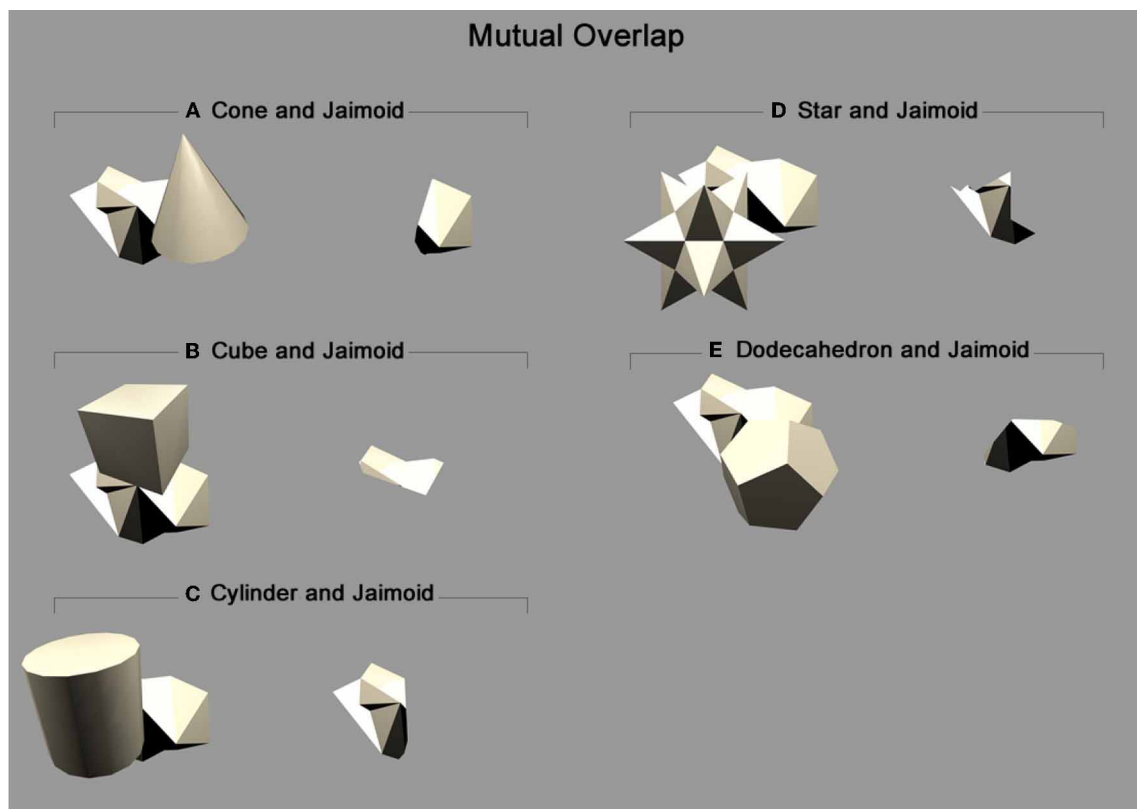
Secondly, VisNet was also tested with six novel objects rotating over  $360^\circ$  in  $1^\circ$  steps. This test demonstrates whether VisNet has learned to respond to the specific objects presented during training or whether VisNet has learned to respond selectively to only the location where these objects were presented.

Finally, VisNet was also tested with the different partial views of the occluded object as presented in **Figure 10** and exemplified in **Figure 11**. As the different object pairs rotate together over  $360^\circ$ , parts of the partially occluded object are not visible. By testing VisNet with the partial views that were not visible during training it is possible to establish whether VisNet is able to bind together the partially occluded views of the occluded object into one holistic invariant representation. This

test is important because it shows that VisNet does not need to rely on a key component of the training stimuli in order to recognise it.

The network's ability to recognise which object is shown during testing is assessed using two information theoretic measures: single and multiple cell information. Full details on the application of these measures to VisNet are given by Stringer et al. (2006). These measures reflect the extent to which cells respond invariantly to an object over a number of different views (transforms), but respond differently to different objects. The single cell information measure is applied to individual cells in layer 4 of the VisNet model, and measures how much information is available from the response of a single cell about the stimulus that was presented. The single cell information measure for each cell shows the maximum amount of information that the cell conveys about any one object. This is computed using the following formula with details provided by Rolls et al. (1997) and Rolls and Milward (2000). The object-specific information  $I(s, R)$  is the amount of information the set of responses  $R$  has about a specific object  $s$ , and is given by

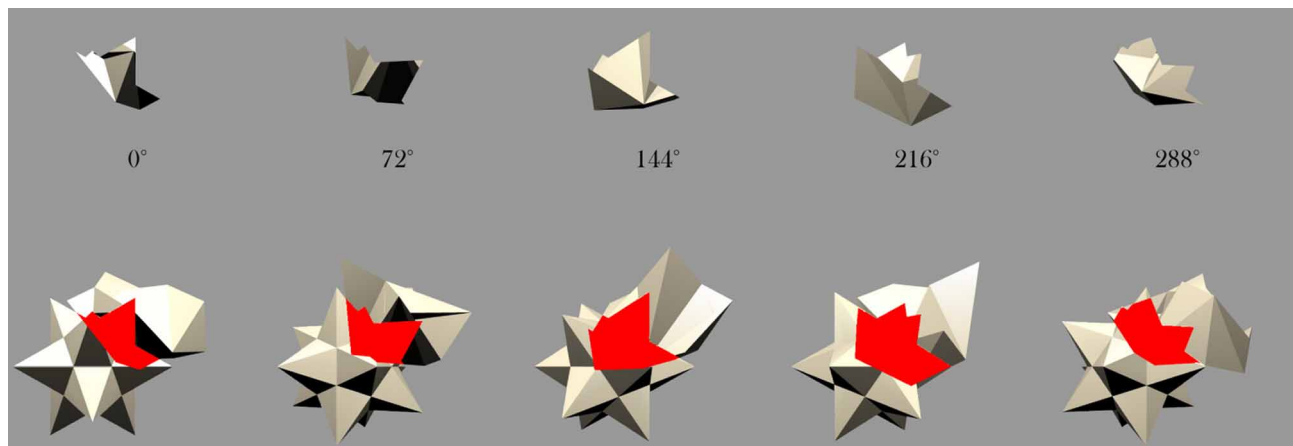
$$I(s, R) = \sum_{r \in R} P(r|s) \log_2 \frac{P(r|s)}{P(r)}, \quad (7)$$



**FIGURE 10 | Mutual overlap: Areas of mutual overlap between the occluding and occluded objects during one example frame of rotation.**

As the two objects rotate together in lock-step, the area of mutual overlap creates a partial view of the occluded object: **(A)** Shows the cone and Jaimoid;

**(B)** Cube and Jaimoid; **(C)** Cylinder and Jaimoid; **(D)** Star and Jaimoid; **(E)** Dodecahedron and Jaimoid. Each pair is presented alongside the partial view it creates. VisNet must learn to associate together all of the different partial views of the occluded object to build an exclusively invariant representation.



**FIGURE 11 | Star and Jaimoid mutual overlap: Five example frames of the star and Jaimoid as they rotate together in lock-step.** As in **Figure 10**, the area of mutual overlap creates a partial view

of the Jaimoid. This is shown in this specific example where the star and Jaimoid are presented over five equally spaced viewing angles.

where  $r$  is an individual response from the set of responses  $R$ . However, the single cell information measure cannot give a complete assessment of VisNet's performance with respect to invariant object recognition. If the amount of information provided by a single cell is not sufficient to differentiate between which objects are present during testing, the network may have failed to learn, or a distributed representation may have formed that needs information from a population of neurons to encode which object is present. Furthermore, if all output cells learned to respond to the same object then there would in fact be relatively little information available about the set of objects  $S$ , and single cell information measures alone would not reveal this. To address these issues, we also calculate a multiple cell information measure, which assesses the amount of information that is available about the whole set of objects from a population of neurons.

Procedures for calculating the multiple cell information measure are described in detail by Rolls et al. (1997) and Rolls and Milward (2000). From a single presentation of an object, we calculate the average amount of information obtained from the responses of all the cells regarding which object is shown. This is achieved through a decoding procedure that estimates which object  $s'$  gives rise to the particular firing rate response vector on each trial. A probability table of the real objects  $s$  and the decoded objects  $s'$  is then constructed. From this probability table, the mutual information is calculated as

$$I(S, S') = \sum_{s, s'} P(s, s') \log_2 \frac{P(s, s')}{P(s)P(s')}. \quad (8)$$

Multiple cell information values are calculated for the subset of cells which, according to the single cell analysis, have the most information about which object is shown. In particular, the multiple cell information is calculated from the first five cells for each object that had the most single cell information about that object. This results in a population of 30 cells given that there were six objects. Previous research (Stringer and Rolls, 2000) found this to

be a sufficiently large subset to demonstrate that invariant representations of each object presented during testing were formed, and that each object could be uniquely identified.

### 3. RESULTS

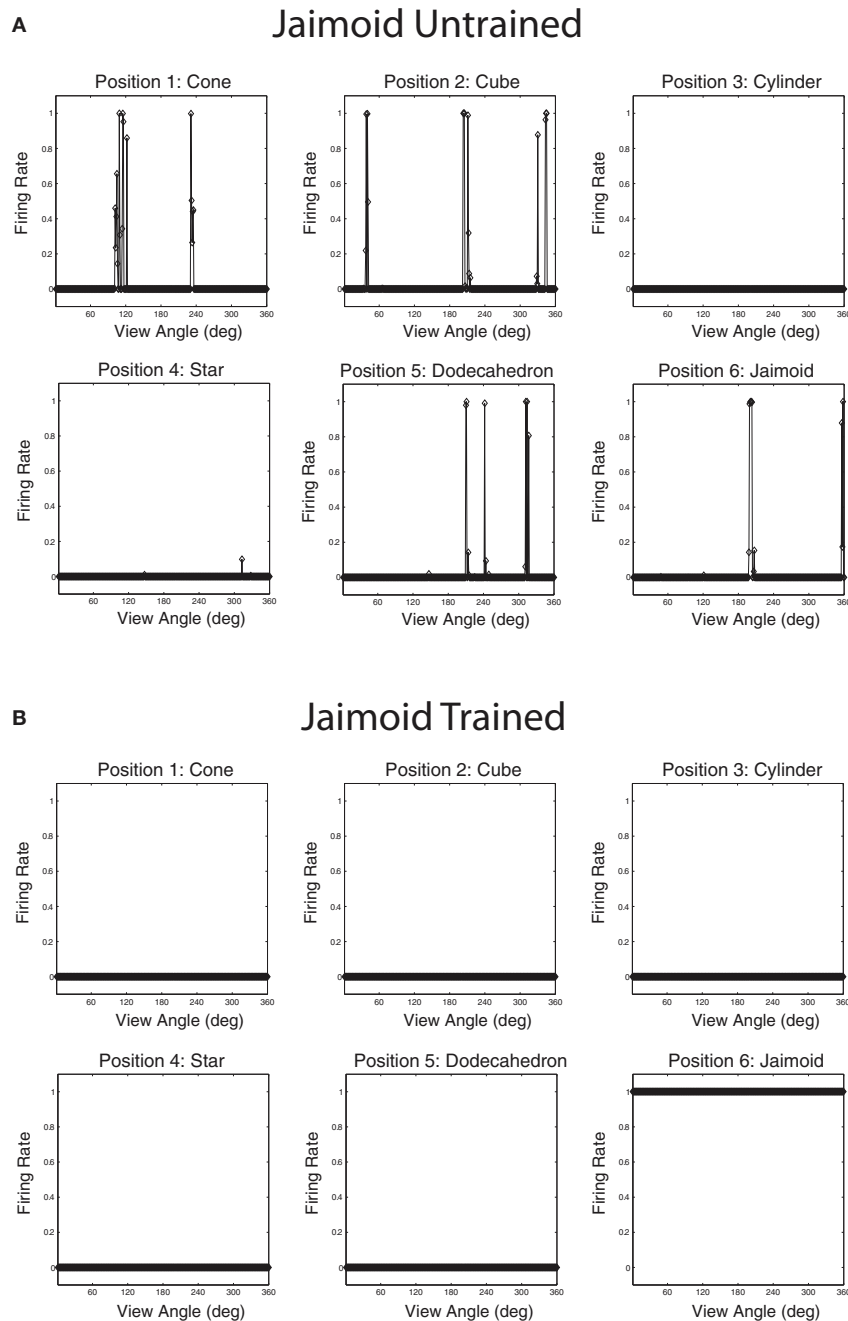
#### 3.1. ANALYSIS OF INDIVIDUALLY ROTATING OBJECTS

After the network had been trained on pairings of the occluded and five occluding objects, we tested whether the network had built transform invariant representations of the objects through a CT learning effect. By presenting the rotating objects individually (**Figure 4**) to the network we were able to record the cell response properties of the neurons in the fourth layer of VisNet for each of the objects. A large number of individual experiments were performed across different parameters and random seeds to ensure the consistency and validity of the results. However, the results presented are all collected as part of the same individual experiment.

Populations of cells that responded invariantly to the individual objects were found. These cells responded to only one object and to no views of any of the other objects. **Figures 12A,B** show the cell response plots for cell (4, 17), selected at random, as each object is rotated through 360° in 1° steps. **Figure 12A** shows the responses of the cell before training and **Figure 12B** shows the cell responses after training. The six response plots of cell (4, 17) before training show that the cell responds at random to the six objects. After training, the cell has learned to respond to the central occluded object, the Jaimoid, invariantly and does not respond to any view of any of the other objects.

**Figures 13A,B** show the cell response plots for cell (19, 1) before and after training, respectively. Before training, the cell responds to the objects randomly. After training this cell has learned to respond invariantly to all 360 views of the Dodecahedron, which was one of the occluding objects, and to no views of any other objects. Furthermore, although not shown here, other output cells learned to respond in a selective and invariant manner to each of the other occluding objects.





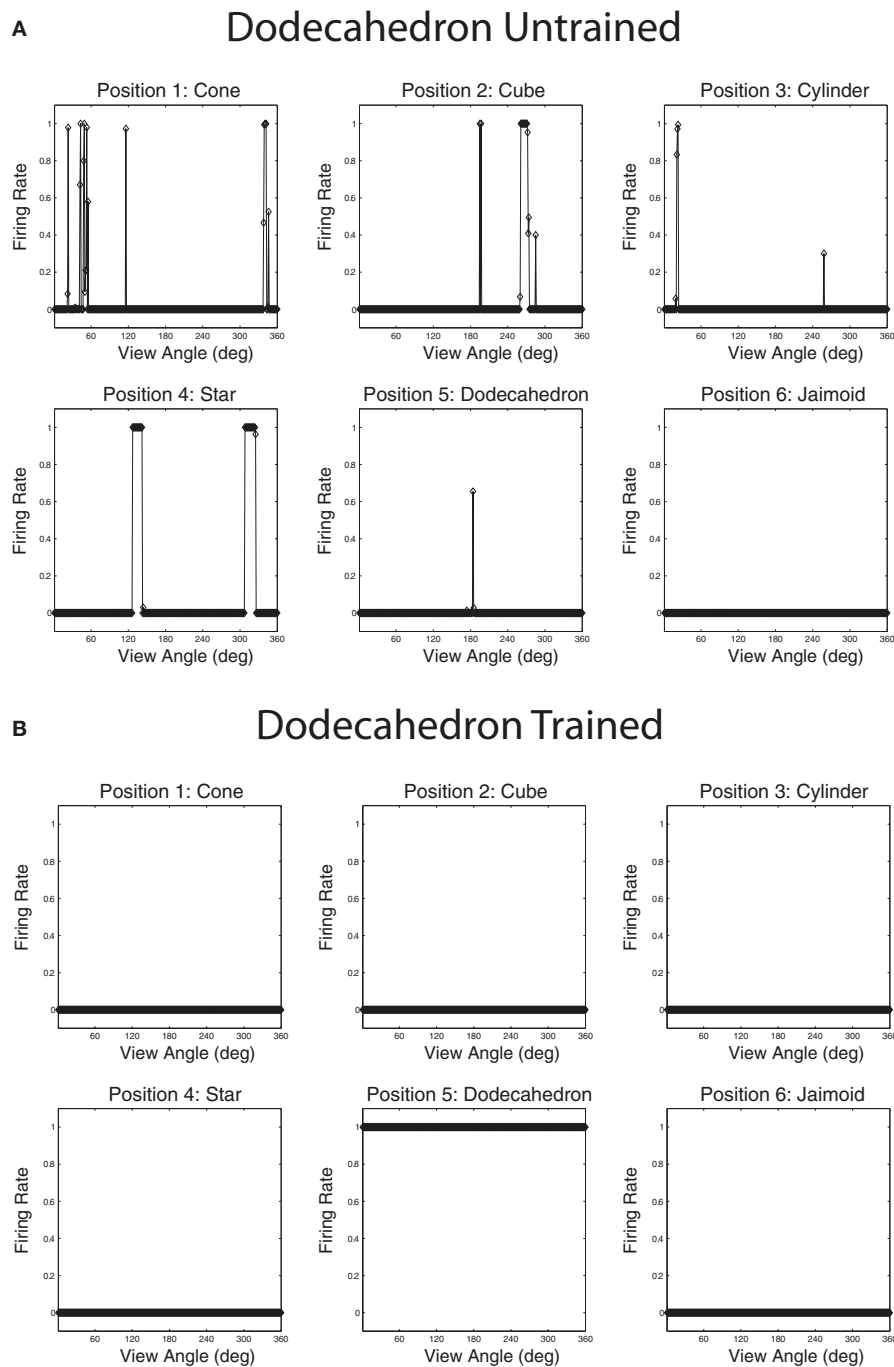
**FIGURE 12 | The firing rate responses of cell (4, 17) in the 4th (output) layer of VisNet to the central occluded object (Jaimoid) and the five surrounding occluding objects as they rotated through 360° in 1° steps before and after training.** Before training, it can be seen that the cell

responds randomly to different views of different objects. After training, it can be seen that the cell's response pattern has changed. This cell responds to all the views of Jaimoid, and to none of the views of the other objects.

Thus, all of the objects were represented individually. When different sparseness values throughout the layers were investigated, results were found to be robust. As the sparseness was gradually increased, a more distributed representation began to form with fewer exclusive cells whereby each cell began to respond to more than one object. In this situation, object identity is still encoded but over a population of cells.

### 3.2. ANALYSIS OF CELL FIRING PROPERTIES IN EARLIER LAYERS

The analyses described above were applied to the output (fourth) layer of the network. Cells in the output layer receive information through the feedforward synaptic connections from across the entire input retina. However, cells in the earlier layers receive more localized input from the retina due to the topographical feedforward connectivity present within the model. Therefore, we



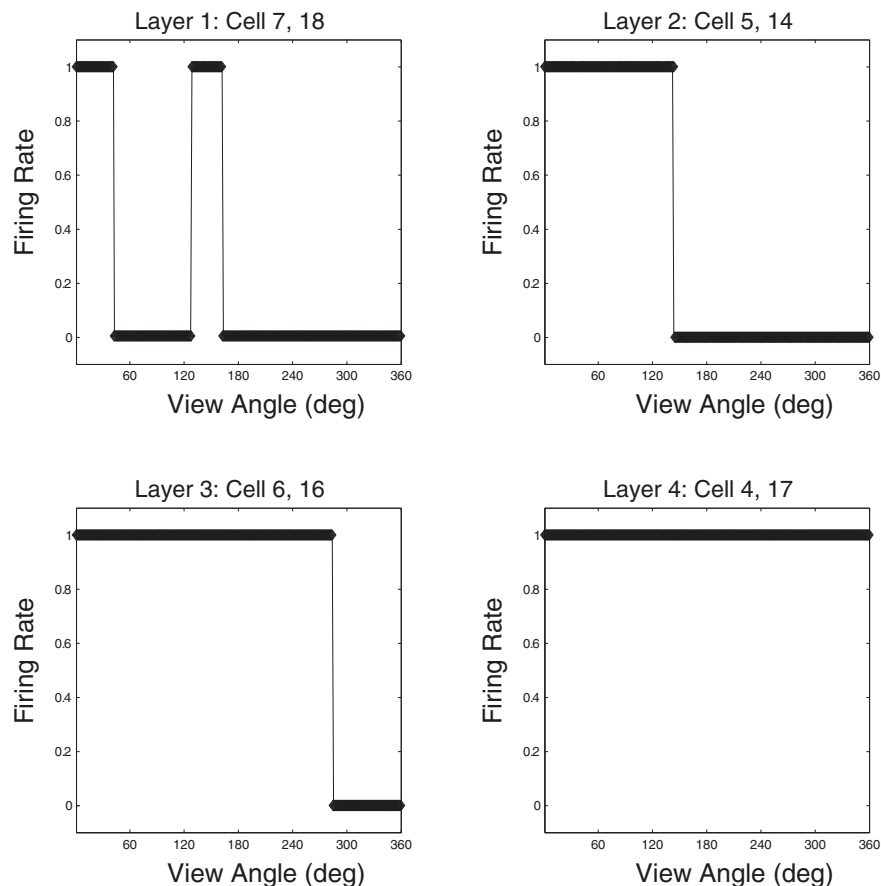
**FIGURE 13 | The firing rate responses of cell (19, 1) in the fourth (output) layer of VisNet to the central occluded object (Jaimoid) and the five surrounding occluding objects as they rotated through 360° in 1° steps before and after training.** Before training it can be seen that the cell

responds randomly to different views of different objects. After training it can be seen that the cell's response pattern has changed. This cell has become an exclusive invariant cell for the Dodecahedron. It responds invariantly to all 360 views of the Dodecahedron and to no views of any other object.

carried out additional analyses of the cell response properties in the earlier layers after training.

Response plots for cells that have learned to respond to the Jaimoid are presented for each of the four layers in **Figure 14**. It was found that the responses of cells in layer 3 of the network

were similar to those in the output (fourth) layer. That is, cells in layer 3 were both object-selective (responding exclusively to their preferred object) and highly transform (view) invariant. In layer 2 of the network, cells were object-selective but showed more modest levels of transform invariance due to the limited convergence



**FIGURE 14 | Example Jaimoid response plots for all four layers.**

A prototypical example cell is presented for each of the 4 layers within the VisNet model. For each cell, its response plot is presented with respect to the example object, the Jaimoid. It can be seen that a typical layer 3 cell is highly transform invariant while a layer 2 cell shows more modest levels of

view invariance. Layer 1 cells demonstrate very little view invariance and responded to a very narrow set of views. In all cases, these cells responded exclusively to their preferred object, in this case the Jaimoid, and did not learn to respond to any views of any of the other objects. Comparable responses exist for all six of the objects presented during training.

of feedforward connections from the retina. Individual layer 1 cells receive projections from a very limited region of the retina. These cells were object-selective but provided very low levels of transform invariance. Stringer and Rolls (2008) have shown that a one-layer network with full feedforward connectivity can learn output representations that are object-selective and completely transform invariant, even when trained on pairs of objects simultaneously. Although, these authors did not look at the case of realistic visual objects that are partially occluding during training.

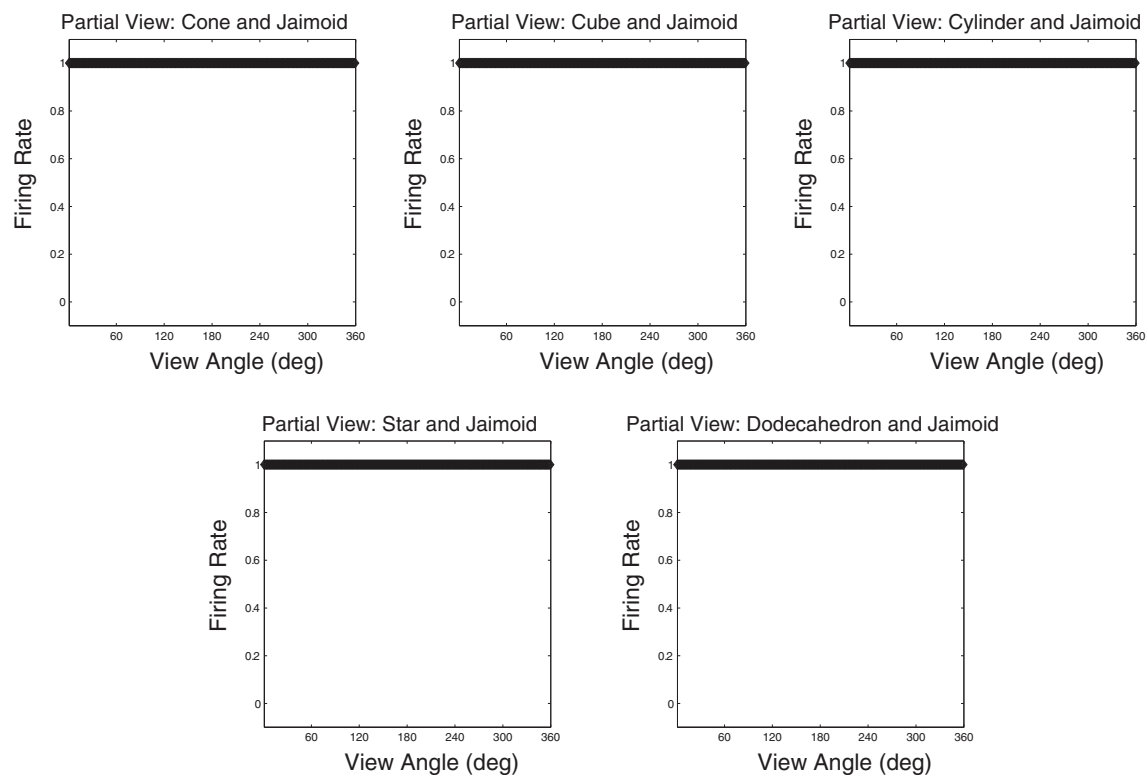
### 3.3. ANALYSIS OF PARTIAL VIEW RESPONSE

To better understand how the network has learned to represent the partially occluded object, the different fragmented partial views of the occluded object that were obscured at different times during training were presented separately to the model during testing for all 360 views. This is a fundamental test to ensure that output neurones have learned to respond to the fragmented parts of the partially occluded object. A similar but easier test would have been to only present the two different halves of the Jaimoid to the network for testing. By presented the smaller fragmented views, it may shown that VisNet has successfully learned to bind

these partial views together form a complete invariant representation of the Jaimoid. This test is necessary to show that output neurons do not just relying on a key-marker or partial view of the Jaimoid in order to recognise it.

Figure 15 shows that neurons in the output layer of the VisNet model were able to successfully bind together all of the different partial views into a holistic invariant representation. Specifically, the exact same output neurons (e.g., cell 4, 17) that responded invariantly when presented with the complete Jaimoid (e.g., Figure 12) were also activated in an identical manner when presented with the various partial views. Each and every partial view caused the same output neurons to respond invariantly as if the whole object had been presented in its entirety.

This important novel result shows that a feedforward hierarchical model of the ventral visual system such as VisNet does not need to rely on particular parts, or key-markers, of an object in order to recognize it. Furthermore, it shows that such a biologically inspired network is able to not only build invariant representation of individual objects despite the fact that pairs of objects were presented during training, but it also shows that such a network can solve a far more complex problem, that is, building



**FIGURE 15 | Partial view response plots for cell 4, 17.** Cell response plots are presented after testing the network with the fragmented partial views of the Jaimoid, as exemplified in **Figure 10**. It can be seen that the example cell 4, 17 that responded invariantly to the complete view of the Jaimoid also responds in an identical manner to the different

fragmented partial views of the Jaimoid. Cell 4, 17 has learned to bind together these different partial views into a holistic representation and responds equally well to all of them, thus proving that this cell does not rely on a specific partial view or key-marker in order to recognize the Jaimoid.

invariant representations in a multi-object environment even when the different objects are partially occluding one another, as is often the case in the real world.

### 3.4. LOCATION VERSUS OBJECT SELECTIVITY

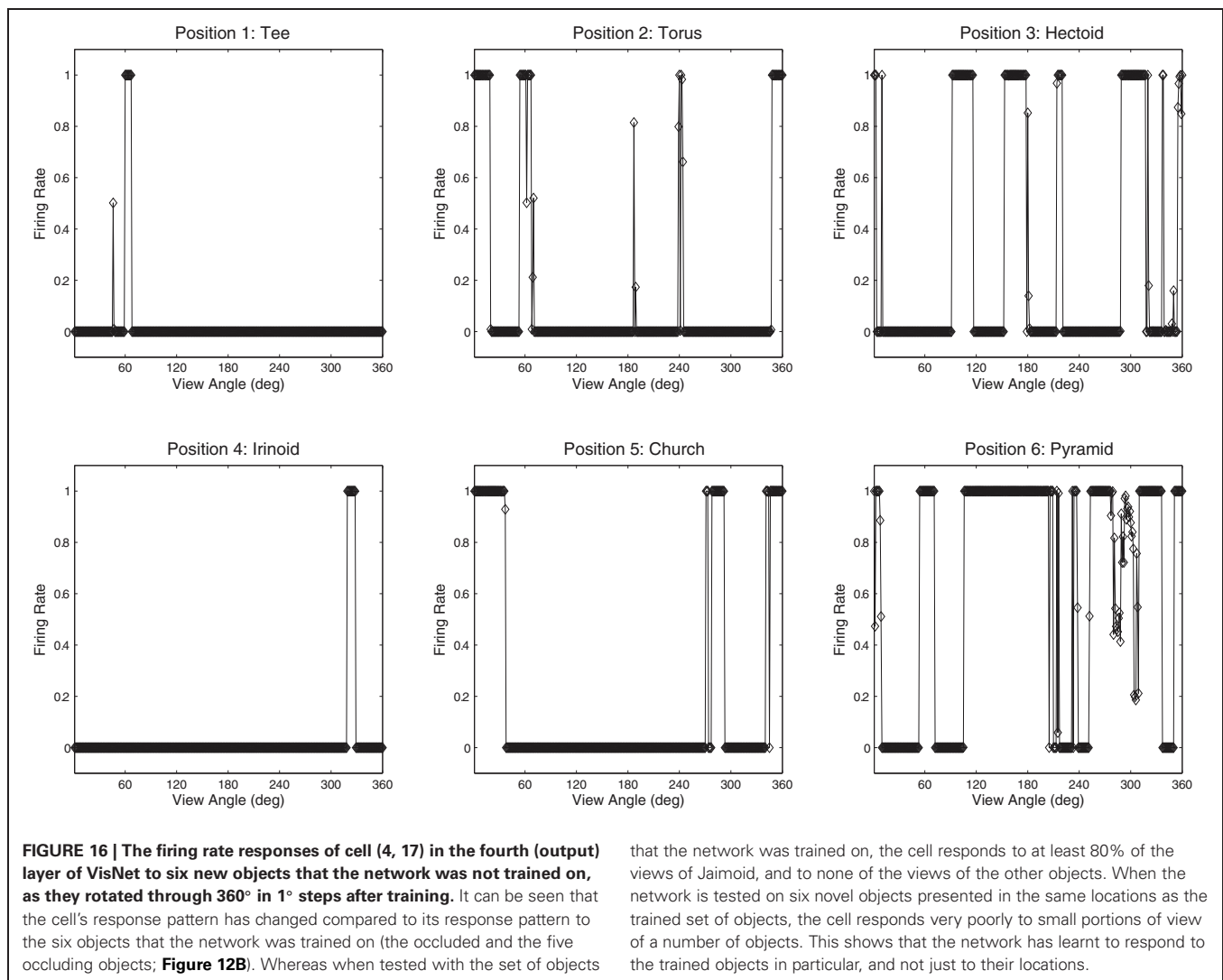
An important question is whether the output cells learned to respond to the visual form of the objects or merely to the retinal locations. In order to minimize the possibility that the output cells had learned to respond to the locations, the objects were presented to the network in overlapping locations as shown in **Figure 8**. Even with the objects presented in highly overlapping locations, the output cells learned to respond to the objects themselves, and to no views of any of the other partially overlapping objects.

The network was also tested with six novel objects, such as a pyramid, rotating over 360° in 1° steps. These objects are novel in the sense that the network was not trained with them and was only exposed to them for testing. The novel objects were presented in the same locations as the original occluding and occluded objects. If the network had learnt to respond to the individual trained objects rather than the locations, then the responses of the output cells to the novel objects should be less clearly tuned than to the trained objects. That is, the cells should not respond so uniformly (invariantly) over the different views of

any particular novel object, and the cell responses should not be selective to individual novel objects. **Figures 16** and **17** show cell response plots for cell (4, 17) and (19, 1) after testing the network with the novel objects. To reiterate, when tested with the original set of objects, cell (4, 17) responds invariantly to the Jaimoid, and to none of the views of the other objects (**Figure 12B**). However, when the network is tested on six novel objects presented in the same locations as the trained set of objects, the cell responds very poorly to small portions of view of a number of objects. Similarly, when tested with the original set of objects, cell (19, 1) responds invariantly to all 360 views of the Dodecahedron and to no views of any other object (**Figure 13B**). When tested on the six novel objects presented in the same locations as the trained set of objects, the cell responds very poorly to small portions of view of a number of objects. These results help demonstrate that the network has learnt to respond to the trained objects in particular, and not just to their locations.

### 3.5. INFORMATION ANALYSIS

Single cell information analysis was conducted to confirm whether the network had developed cells that responded invariantly to their preferred object (**Figure 18**). The unbroken line represents the results obtained after presenting the six original trained objects to a network after training on the 360 views of



that the network was trained on, the cell responds to at least 80% of the views of Jaimoid, and to none of the views of the other objects. When the network is tested on six novel objects presented in the same locations as the trained set of objects, the cell responds very poorly to small portions of view of a number of objects. This shows that the network has learnt to respond to the trained objects in particular, and not just to their locations.

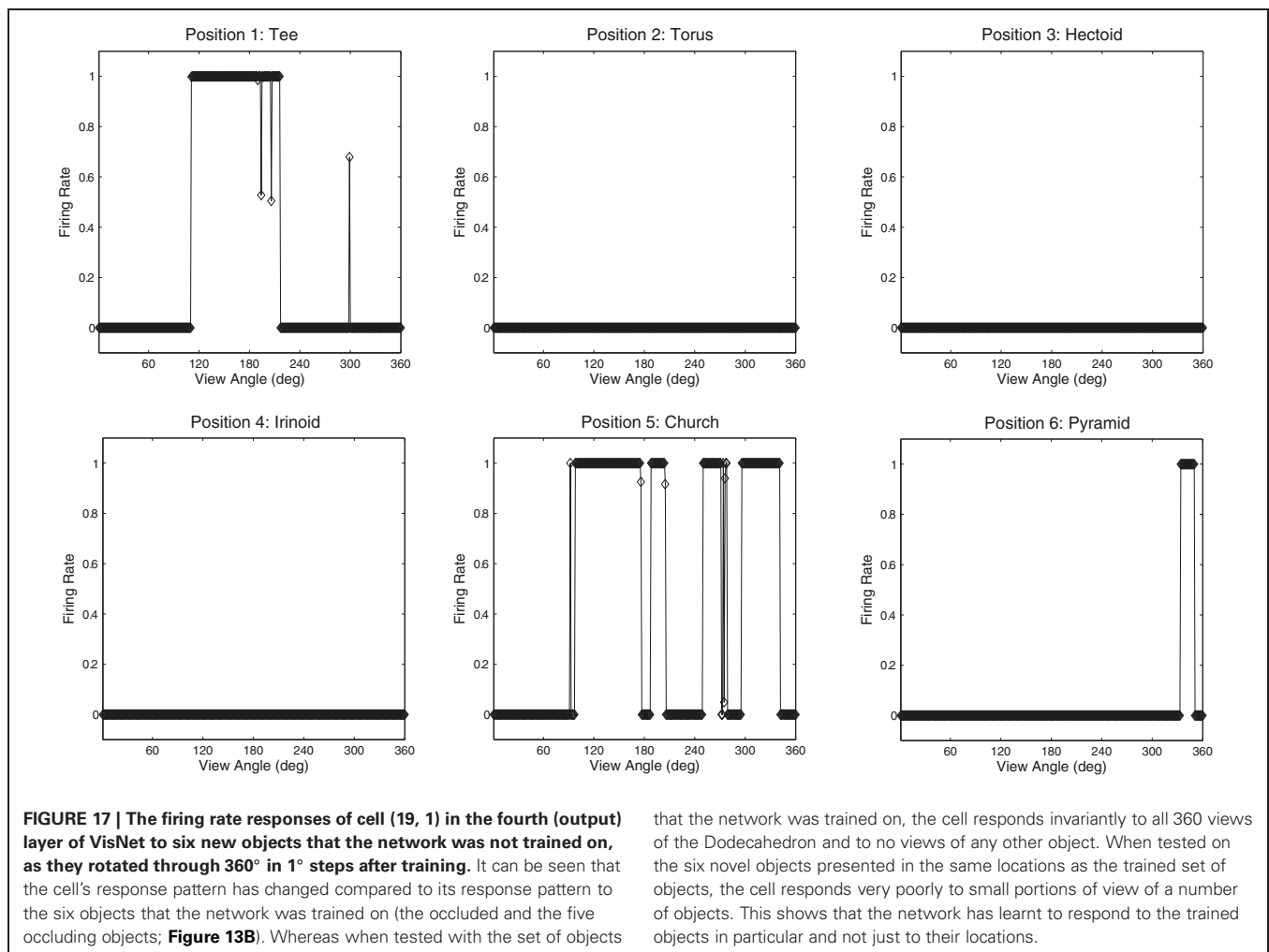
all possible object pairs. The dashed line represents the results obtained after presenting six novel objects rotating in the same positions as the six original trained objects. The dotted line represents the results after presenting the six original objects to a random untrained network. Single cell information measures for the fourth layer neurons ranked in order of their invariance to the objects are shown. It can be seen that training the network on the object pairs has lead to many of the fourth layer neurons attaining the maximal level of single cell information of 2.58 bits for the trained objects. These neurons have learned to respond to all of the views of their preferred object. However, when the network, which had been trained on the six original objects, was tested with novel objects, no cells reached the maximum level of information. This reflected the fact that the output cells of the trained network were not able to respond to the novel objects in a view-invariant or object-selective manner, as shown in **Figures 16** and **17**. These results thus further demonstrate that when the network was trained on the six original objects, the output cells had learned to respond selectively to the trained objects and not the untrained objects. This in turn confirms that the network learned

to respond to the visual forms of the trained objects rather than their retinal locations.

However, it is unclear whether all of the six objects are individually represented by a unique subset of invariant output cells. Indeed, it is possible that these cells are responding to the same object and are, therefore, unable to provide information regarding which object is present. To ensure that there are cells that respond preferentially to each of the six objects multiple cell information analysis was performed.

**Figure 19** shows the multiple cell information analysis obtained when VisNet was tested with the six individual objects rotating through 360° in 1° steps. Multiple cell information analysis results are also plotted for six novel objects that the network was not previously trained on. These novel objects were rotating in the same positions as the six objects on which the network was originally trained. Results are presented having tested the trained network with the original set of objects (unbroken line), after testing the trained network with the novel set of objects (dashed line) and with a random untrained network (dotted line). After the network was trained and tested with the original set of objects,





that the network was trained on, the cell responds invariantly to all 360 views of the Dodecahedron and to no views of any other object. When tested on the six novel objects presented in the same locations as the trained set of objects, the cell responds very poorly to small portions of view of a number of objects. This shows that the network has learnt to respond to the trained objects in particular and not just to their locations.

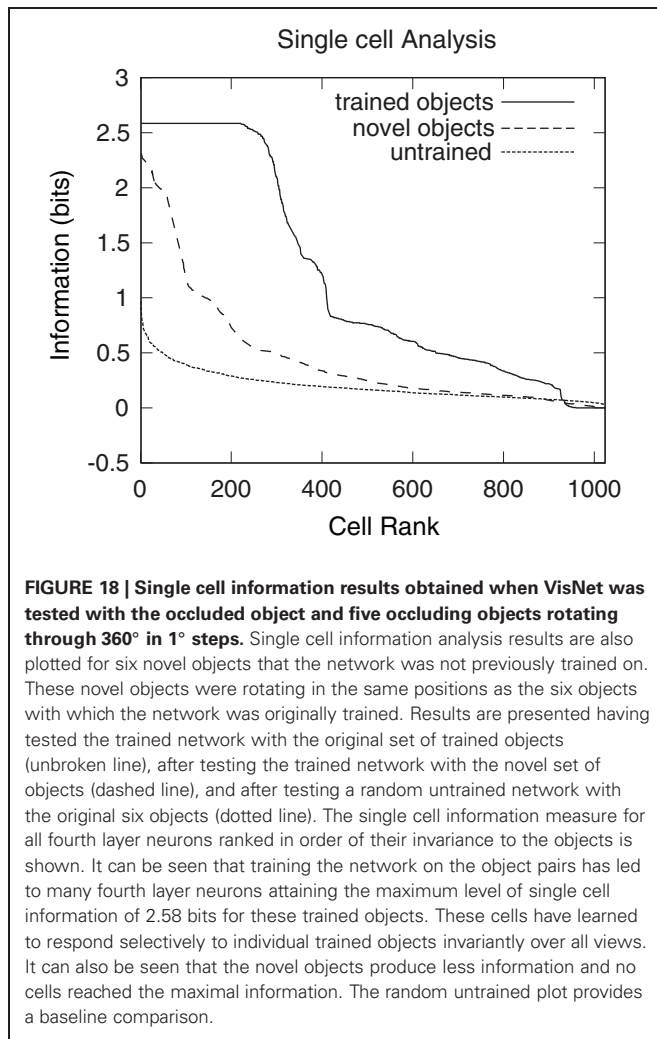
over 2.5 bits of information was reached (substantially higher than 1.1 bits reached by the untrained network or 1.7 bits reached by the trained network that was tested on the novel set of objects) suggesting that the single cell information results included cells that preferentially responded to all six objects. These plots show that the network did not learn to respond to the locations of the objects, and instead bound together different views of the occluding and occluded object to form object specific representations. This was also confirmed by inspection of the cell response plots as shown in **Figures 12B** and **16**, as well as **Figures 13B** and **17**.

#### 4. DISCUSSION

An important question in natural vision is how the brain forms invariant representations of objects that are always partially occluded by other objects during learning. In a real world visual environment, this will often be the case. *Stringer and Rolls (2008)* have shown that a biologically plausible competitive neural network (VisNet) can develop invariant representations of individual objects when no single object is seen in isolation. In this paper we demonstrate for the first time how such a network might form an invariant representation of an object that is always partially occluded by other objects. The mechanism employed for

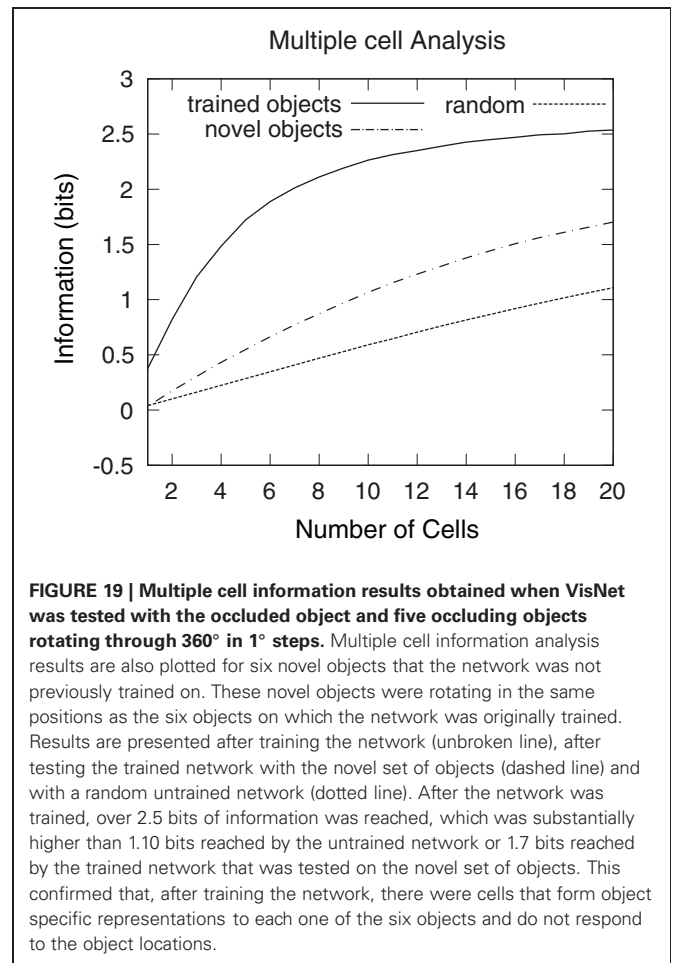
invariance learning is CT learning. CT learning uses the spatial continuity between the views of individual objects as they transform in the real world, combined with associative learning of feedforward connection weights.

It was found that, after training the network with a rotating object that is always partially occluded, the network is able to form view invariant representation of the partially occluded object. In addition, by testing the network with the fragmented partial views of the occluded object in isolation, it was also shown that the same output neurons that learned to respond to the Jaimoid when presented in its entirety also responded in an identical manner to each of the fragmented partial view sequences. This shows that the network has learned to bind together the different partial views of the occluded object presented during training into a holistic invariant representation despite always seeing the Jaimoid partially occluded and, therefore, never in isolation. It was also found that view invariant representations are also formed for all five occluding objects. This is a challenging task since the occluding objects were always overlapping the occluded object and therefore VisNet had to learn to separate the objects. This is the first time such learning has been shown to happen in a biologically inspired model of the ventral visual system.



Despite the fact that the objects were presented in the same location during training and testing, the network was able to form representations of the objects' identities instead of just learning to respond to particular locations. During training there was significant overlap between the objects (Figures 8 and 9), which would have precluded VisNet from learning about each object just because it was presented in the same location. Instead, VisNet built separate representations of each of the individual objects. This is confirmed by invariant cells that responded maximally to all views of only one of the stimuli and not to any views of any other stimuli. After training, all of the stimuli were represented in this way. Given that the objects were highly overlapping during training, if the network had learned to respond to location instead of stimulus identity, then the network would not have developed cells which responded specifically and invariantly to individual objects, with all of the objects represented uniquely in this way.

This conclusion is also confirmed by additional results obtained with a novel set of six objects presented during testing. These novel objects were presented in the exact same locations as the six original objects that were presented during training. The output neurons that learned to respond preferentially and



invariantly to the original trained objects were then inspected and an example response plot was presented. It was shown that these output neurons did not respond in an exclusive or invariant manner to any of these novel objects thus confirming that the network had learned object selectivity. The residual firing that was present within the response plots can be explained by the fact that after the network has been trained, neurons will respond in proportion to how similar the novel input pattern is to the previous learned patterns. By virtue of the fact that the input spaces are not orthogonal with respect to the input filters that they activate when objects occupy the same input space on the retina, they will cause an degree of activity in the network based on previous learning. This is a fundamental property of competitive networks, which will try to generalize to novel input patterns depending on their similarity to the previous learned input patterns.

Most artificial computer vision systems designed by software engineers do not seek to mimic processing exactly as it is carried out in the brain. Also, the challenges addressed by artificial computer vision systems are often more focused, for example, on the problem of object or face recognition after training the system with individual segmented objects or faces. For such a task, non-biologically inspired artificial visual systems often rely on either template matching or searching for the presence of a subset of key features in order to recognise a partially occluded

object (Ullmann, 1992; Ying and Castañon, 2002; Do et al., 2005). However, as computational neuroscientists, we are interested in the more ecological problem of understanding how the primate visual system learns in an unsupervised manner to make sense of complex natural visual scenes containing multiple objects. This is a valuable long term goal, which may ultimately offer engineers powerful new approaches to intelligent visual scene analysis and object recognition. Understanding how the primate brain learns to process visual input from scenes will involve the step-by-step uncovering of many key neurodynamical mechanisms, which ultimately blend and work together in the brain. In this current paper, we have examined the problem of how the primate visual system might develop separate transform-invariant representations of individual objects even if these objects are always seen partially occluding each other during unsupervised learning. The simulations reported here have shown for the first time how a biologically plausible model, VisNet, of the ventral visual pathway, with a Hebbian associative synaptic learning rule, is able to solve this particular problem.

#### 4.1. FUTURE WORK

The results described above have shown how a biologically plausible neural network model of the ventral visual pathway, VisNet, is able to develop object-selective and rotation-invariant representations of objects that were partially occluding each other during training. However, more work needs to be done to explore the limits of this mechanism.

For example, what proportion of an object needs to be visible? Future research could investigate the exact degree to which objects can be occluded before learning of the partially occluded object breaks down. This avenue of research could address the effect of occlusion by two or more objects at a time, or where there is more than one occluded object. The results of these experiments presented within this study suggest that the extent to which the partially occluded object is covered could increase quite considerably so long as the different parts of the occluded object have all become visible at some point during training.

The use of simple geometric shapes is a limitation of the current study that should be addressed as part of future research. The choice to use simple geometric shapes was not to help VisNet solve the task at hand. These shapes allowed for a level of control necessary to answer the question “how” has VisNet solved this problem. This level of control is harder to achieve with more complex objects, but their use is a sensible next step to explore their effect on the self-organization of the network. The authors believe that the types of objects used will not have any qualitative impact on the results presented within this paper, but this should be confirmed. So long as the resolution of the retina is high enough

to convey the necessary detail of the more natural objects, then the VisNet model will make use of the same principles discussed within this study to solve the problem.

In the simulations described above, individual objects rotated on the same part of the retina. Perhaps a more challenging problem is the *translation* of objects across the retina. This would happen naturally as an observer shifts their gaze around a visual scene. In this case, all of the objects would be seen moving over the entire retina. The input representations of the objects would then fully overlap over all possible locations on the retina. Yet the network must still form separate output representations of the objects, which are also translation invariant. We hypothesize that the network described in this paper should still be able to solve this problem using similar learning principles.

Another limitation of the current study is that it explores only one type of invariance learning mechanism, CT learning (Stringer et al., 2006). This binds different transforms of a particular object together by exploiting the spatial similarity that exists between the different transforms of that object. As discussed, CT learning relies on a simple Hebbian learning rule. It would be very interesting to investigate if similar results can be achieved with a different type of biologically plausible learning rule such as Trace learning (Foldiak, 1991; Wallis and Rolls, 1997). This alternative learning rule exploits temporal continuity of successive transforms of an object in order to build a transform-invariant representation of that object. Trace learning utilizes a memory trace of the recent firing of the post-synaptic cell.

In natural vision, objects are not always moving with respect to one another, nor with respect to the viewer. Sometimes objects are simply static, and one object will occlude another and yet in many situations we are still able to learn to recognize the partially occluded object. A typical situation of this kind might occur when we view some faces in a photograph, for example. The current VisNet model would not be able to solve such a training paradigm because it relies on the statistical decoupling of features between different objects that can occur through independent movement in order to tell them apart. However, our laboratory has recently shown that this more difficult problem can be solved using spiking neural network dynamics. In such a model, the times of individual action potentials are simulated, and the synaptic plasticity can be dependent on the times of the pre- and post-synaptic spikes (Bi and Poo, 1998).

#### ACKNOWLEDGMENTS

This research was supported by the Wellcome Trust, and by the Economic and Social Research Council. James M. Tromans is an ESRC-supported graduate student.

#### REFERENCES

- Bi, G. G., and Poo, M. M. (1998). Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type. *J. Neurosci.* 18, 10464–10472.
- Booth, M. C., and Rolls, E. T. (1998). View-invariant representations of familiar objects by neurons in the inferior temporal visual cortex. *Cereb. Cortex* 8, 510–523.
- Desimone, R. (1991). Face-selective cells in the temporal cortex of monkeys. *J. Cogn. Neurosci.* 3, 1–8.
- Do, Q. V., Lozo, P., and Jain, L. C. (2005). A Vision System for Partially Occluded Landmark Recognition, Vol. 3809/2005. Berlin/Heidelberg: Springer.
- Foldiak, P. (1991). Learning invariance from transformation sequences. *Neural Comput.* 3, 194–200.
- Hasselmo, M. E., Rolls, E. T., Baylis, G. C., and Nalwa, V. (1989). Object-centered encoding by face-selective neurons in the cortex in the superior temporal sulcus of the monkey. *Exp. Brain Res.* 75, 417–429.
- Hawken, M. J., and Parker, A. J. (1987). Spatial properties of neurons in the monkey striate cortex. *Proc. R. Soc. Lond. B. Biol. Sci.* 231, 251–288.
- Hertz, J., Krogh, A., and Palmer, R. G. (1991). *Introduction to the Theory of*

- Neural Computation*. Workingham, UK: Addison Wesley
- Ito, M., Tamura, H., Fujita, I., and Tanaka, K. (1998). Size and position invariance of neuronal response in monkey inferotemporal cortex. *J. Neurophysiol.* 73, 218–226.
- Kobatake, E., and Tanaka, K. (1994). Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. *J. Neurophysiol.* 71, 856–867.
- Op De Beeck, H., and Vogels, R. (2000). Spatial sensitivity of macaque inferior temporal neurons. *J. Comp. Neurol.* 426, 505–518.
- Perrett, D. I., and Oram, M. W. (1993). Neurophysiology of shape processing. *Image Vis. Comput.* 11, 317–333.
- Perry, G., Rolls, E. T., and Stringer, S. M. (2006). Spatial vs temporal continuity in view invariant visual object recognition learning. *Vision Res.* 46, 3994–4006.
- Rolls, E. T. (1992). Neurophysiological mechanisms underlying face processing within and beyond the temporal cortical visual areas. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 335, 11–20.
- Rolls, E. T. (2000). Functions of the primate temporal lobe cortical visual areas in invariant visual object and face recognition. *Neuron* 27, 205–218.
- Rolls, E. T., and Deco, G. (2002). *Computational Neuroscience of Vision*. Oxford, UK: Oxford University Press.
- Rolls, E. T., and Milward, T. (2000). A model of invariant object recognition in the visual system: learning rules, activation functions, lateral inhibition, and information-based performance measures. *Neural Comput.* 12, 2547–2572.
- Rolls, E. T., and Stringer, S. M. (2001). Invariant object recognition in the visual system with error correction and temporal difference learning. *Network* 12, 111–129.
- Rolls, E., and Treves, A. (1990). The relative advantages of sparse versus distributed encoding for associative neuronal networks in the brain. *Network* 1, 407–421.
- Rolls, E. T., and Treves, A. (1998). *Neural Networks and Brain Function*, 1st Edn. Oxford: Oxford University Press.
- Rolls, E. T., Treves, A., Tovee, M. J., and Panzeri, S. (1997). Information in the neuronal representation of individual stimuli in the primate temporal visual cortex. *J. Comput. Neurosci.* 4, 309–333.
- Royer, S., and Pare, D. (2003). Conservation of total synaptic weight through balanced synaptic depression and potentiation. *Nature* 422, 518–522.
- Stringer, S. M., and Rolls, E. T. (2000). Position invariant recognition in the visual system with cluttered environments. *Neural Netw.* 13, 305–315.
- Stringer, S. M., Perry, G., Rolls, E. T., and Proske, J. H. (2006). Learning invariant object recognition in the visual system with continuous transformations. *Biol. Cybern.* 94, 128–142.
- Stringer, S. M., and Rolls, E. T. (2008). Learning transform invariant object recognition in the visual system with multiple stimuli present during training. *Neural Netw.* 21, 888–903.
- Stringer, S. M., Rolls, E. T., and Tromans, J. M. (2007). Invariant object recognition with trace learning and multiple stimuli present during training. *Network* 18, 161–187.
- Tanaka, K., Saito, H., Fukada, Y., and Morioka, M. (1991). Coding visual images of objects in the inferotemporal cortex of the macaque monkey. *J. Neurophysiol.* 66, 170–189.
- Tovee, M. J., Rolls, E. T., and Azzopardi, P. (1994). Translation invariance in the responses to faces of single neurons in the temporal visual cortical areas of the alert macaque. *J. Neurophysiol.* 72, 1049–1060.
- Ullmann, J. R. (1992). Analysis of 2-D occlusion by subtracting out. *IEEE Trans. Pattern Anal. Mach. Intell.* 14, 485–489.
- Wallis, G., and Rolls, E. T. (1997). Invariant face and object recognition in the visual system. *Prog. Neurobiol.* 51, 167–194.
- Wallis, G., Rolls, E. T., and Foldiak, P. (1993). Learning invariant responses to the natural transformations of objects. *Int. Jt. Conf. Neural Netw.* 2, 1087–1090.
- Ying, Z., and Castañón, D. (2002). Partially occluded object recognition using statistical models. *Int. J. Comput. Vis.* 49, 57–78.

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 10 February 2012; accepted: 27 June 2012; published online: 25 July 2012.

Citation: Tromans JM, Higgins I and Stringer SM (2012) Learning view invariant recognition with partially occluded objects. *Front. Comput. Neurosci.* 6:48. doi: 10.3389/fncom.2012.00048

Copyright © 2012 Tromans, Higgins and Stringer. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in other forums, provided the original authors and source are credited and subject to any copyright notices concerning any third-party graphics etc.



# Spatially invariant computations in stereoscopic vision

Michel Vidal-Naquet<sup>1\*</sup> and Sergei Gepshtein<sup>2</sup>

<sup>1</sup> Brain Science Institute, RIKEN, Wako-shi, Saitama, Japan

<sup>2</sup> The Salk Institute for Biological Studies, La Jolla, CA, USA

## Edited by:

Evgeniy Bart, Palo Alto Research Center, USA

## Reviewed by:

Brent Doiron, University of Pittsburgh, USA

Evgeniy Bart, Palo Alto Research Center, USA

## \*Correspondence:

Michel Vidal-Naquet, Brain Science Institute, RIKEN, Wako-shi, Saitama, Wako-shi, Japan.  
e-mail: michel.vidalnaquet@gmail.com

Perception of stereoscopic depth requires that visual systems solve a correspondence problem: find parts of the left-eye view of the visual scene that correspond to parts of the right-eye view. The standard model of binocular matching implies that similarity of left and right images is computed by inter-ocular correlation. But the left and right images of the same object are normally distorted relative to one another by the binocular projection, in particular when slanted surfaces are viewed from close distance. Correlation often fails to detect correct correspondences between such image parts. We investigate a measure of inter-ocular similarity that takes advantage of spatially invariant computations similar to the computations performed by complex cells in biological visual systems. This measure tolerates distortions of corresponding image parts and yields excellent performance over a much larger range of surface slants than the standard model. The results suggest that, rather than serving as disparity detectors, multiple binocular complex cells take part in the computation of inter-ocular similarity, and that visual systems are likely to postpone commitment to particular binocular disparities until later stages in the visual process.

**Keywords:** adaptive, binocular matching, complex cell, correlation, flexible matching, perception of slant, stereopsis

## INTRODUCTION

Stereoscopic vision depends on *binocular matching*: a process that finds which parts of the left and right eye's images correspond to the same source in the visual scene (**Figure 1**). The difference between positions of the corresponding image parts is called *binocular disparity*, a key source of information for perception of stereoscopic depth.

In the standard view of binocular matching, the corresponding parts of left and right images are found using *inter-ocular correlation* as the measure of image similarity. This view is supported by neurophysiological evidence. The *disparity energy model* (Ohzawa et al., 1990; Qiang, 1994; Cumming and Parker, 1997; Ohzawa, 1998; Cumming and DeAngelis, 2001; Haefner and Cumming, 2008) describes function of binocular complex cells which are thought to play a key role in the computation of binocular disparity (and which are sometimes described as “disparity detectors”). Responses of modeled binocular complex cells to some stimuli are well approximated by a computation similar to inter-ocular correlation (Fleet et al., 1996; Qian and Zhu, 1997; Anzai et al., 1999), and so a simplifying assumption is often made that inter-ocular correlation can be used to predict outcomes of the computation of similarity in biological vision. In psychophysical studies of stereopsis, for example, inter-ocular correlation is commonly used to explain limitations of stereoscopic vision (Tyler, 1973; Cormack et al., 1991; Banks et al., 2004, 2005; Filippini and Banks, 2009), in particular the decline in the ability for stereopsis at large slants of stimulus surfaces.

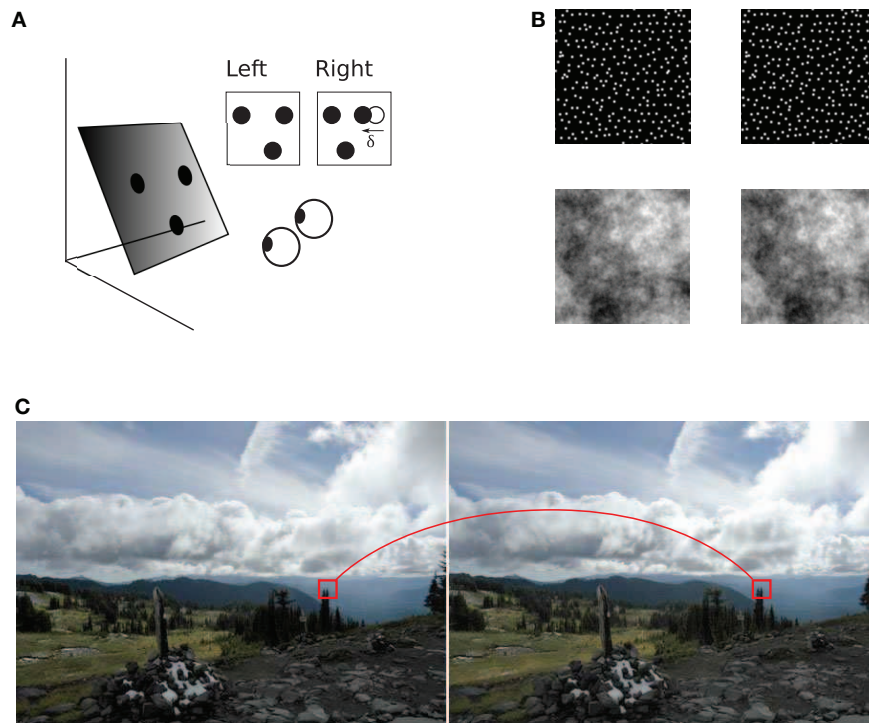
The computation of similarity of left and right images using inter-ocular correlation has two shortcomings. First, correlation of image regions fails to capture an important characteristic of

complex cells: spatial invariance of their responses (even though the disparity energy model does capture this invariance). The disregard for spatial invariance misses an important aspect of the biological computation. Studies of other visual functions showed that spatial invariance endows visual systems with important computational abilities, e.g., in object perception (Riesenhuber and Poggio, 1999; Ullman et al., 2002; Yu et al., 2002; Serre et al., 2007a,b) and in efficient encoding of natural scenes (Hyvarinen and Hoyer, 2000; Karklin and Lewicki, 2009).

Second, inter-ocular correlation is biased in favor of stimuli that are uncommon in the natural viewing conditions. Inter-ocular correlation is “rigid” in the sense it does not tolerate large distortions of corresponding image parts: correlation peaks where image parts are identical and it rapidly declines where image parts are dissimilar. But only rarely do identical left and right images occur in the natural environment. Because of the geometry of binocular projection, parts of the left and right images are generally dissimilar (**Figure 1A**), especially when stimulus surfaces are slanted and viewed from a short distance (Pollard et al., 1986; Filippini and Banks, 2009). It is therefore not surprising that a recent study of human perception found that the correlation operation fails to explain human perception in stimuli that involve slanted surfaces (Allenmark and Read, 2010). We refer to this implicit bias of matching by correlation as the *assumption of uniform disparity*.

In the following we propose that the computation of binocular similarity in biological vision should be modeled using an operation which, first, takes advantage of the spatial invariance found in binocular complex cells and, second, avoids the inapt assumption of uniform disparity. We investigate a “flexible” measure of





**FIGURE 1 | Binocular geometry.** (A) A slanted plane with three dots painted on it is viewed from two slightly different vantage points. Left and right projections of the dots are shown in the insets. Coordinates of dot projections in the two images are generally different, illustrated for one of the dots using vector  $\delta$ . The horizontal extent of this vector is called horizontal binocular disparity. The triangle formed by the three dots in the right image is distorted with respect to the triangle in the left image (B) Examples of stereograms (image pairs) used in the present study: a random-dot

stereogram on the top and a stereogram with 1/f luminance distribution on the bottom. Both stereograms depict a slanted plane which the reader may experience by cross fusion. (C) Binocular correspondence. The visual system must establish which parts of the two images correspond to the same source in the scene. A pair of such corresponding image parts is shown in a stereogram of a natural scene ("Accidental stereo pair." Online image. Flickr. <http://www.flickr.com/photos/abulafia/829612/>, Creative Commons).

similarity that tolerates distortions of the corresponding parts of left and right images. We implement this measure using a *MAX-pooling operation*, which has been successfully used for modeling spatially invariant computations by complex cells in service of other functions of biological vision (Riesenhuber and Poggio, 1999; Serre et al., 2007a,b).

In a series of computational experiments, we simulate a tilt discrimination task using stimuli that portray a wide range of surface slants. The stimuli are composed of two types of texture: random dots (common in psychophysical studies of stereopsis, e.g., Banks et al., 2004; Filippini and Banks, 2009; Allenmark and Read, 2010) and patterns that imitate statistics of luminance in natural images (Ruderman and Bialek, 1994).

We find that the spatially invariant computation of inter-ocular similarity supports excellent performance over a significantly larger range of stimulus slants than the rigid computation. This is because the flexible measure of similarity can adapt to different amounts of inter-ocular distortion in different parts of the stimulus.

We also find that in stimuli with naturalistic image statistics, the flexible measure is more effective than methods previously advanced to overcome inter-ocular distortions, such as image blurring, supporting the view that spatially invariant

computation of inter-ocular similarity is particularly suitable for stereoscopic vision in the natural visual environment.

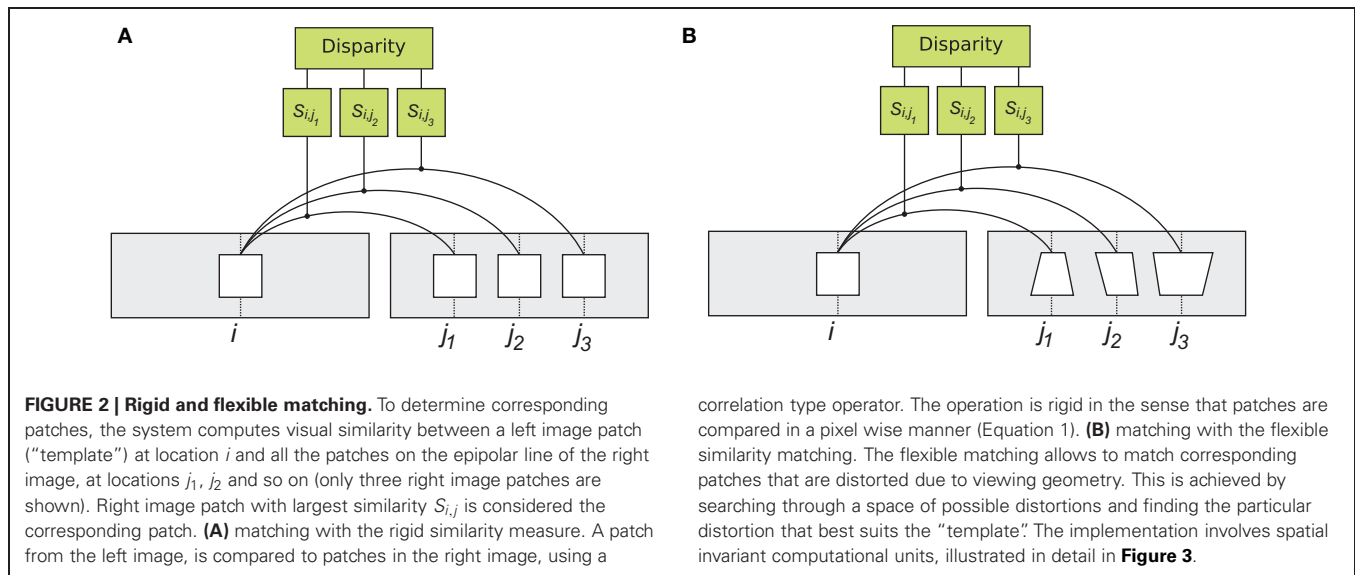
## MODELS AND METHODS

We first describe the two methods for measurement of inter-ocular similarity compared in our experiments: rigid matching and flexible matching (Figure 2). We then describe the computations we used to evaluate performance of these matching methods. (We chose to do so using a tilt discrimination task because it allowed us to compare matching methods comprehensively: across many directions of disparity change, which is particularly important in the complex stimulus of Experiment 2.)

### RIGID MATCHING

Normalized Cross-correlation is commonly used for modeling of binocular matching in biological vision (Tyler and Julesz, 1978; Cormack et al., 1991; Banks et al., 2004, 2005; Filippini and Banks, 2009). For image parts ("patches" or "templates")  $\mathbf{L}$  from the left image and  $\mathbf{R}$  from the right image, this measure is

$$C(\mathbf{L}, \mathbf{R}) = \frac{1}{\sigma_L \sigma_R} \sum_{x,y=1}^N (L_{(x,y)} - \bar{L})(R_{(x,y)} - \bar{R}), \quad (1)$$



where  $L_{(x,y)}$  and  $R_{(x,y)}$  are the luminances at coordinates  $(x, y)$ ,  $\bar{L}$  and  $\bar{R}$  are the average luminances,  $\sigma_L$  and  $\sigma_R$  are the standard deviations of luminance distributions, and  $N$  is the number of image elements within each patch used in the computation.

This measure is “rigid” in the sense the inter-ocular similarity is computed using unaltered image patches, i.e., as they are in the left and right projections of the visual scene (**Figures 2A** and **3A**). The rigid computation of similarity favors matching of image parts that are identical (up to a luminance multiplication and shift), which is why estimates of similarity of corresponding patches rapidly decline when luminance patterns in the left and right images are misaligned (**Figure 3C**, top). Thus, rigid matching is likely to miss binocular correspondences when local image distortions are large, which happens when surface slant is high.

To contrast the rigid measure of inter-ocular similarity with the measure we review next (Equation 5), we write it as

$$S_{i,j}^{\text{rig}} = C(\mathbf{L}_i, \mathbf{R}_j), \quad (2)$$

where  $C$  is as in Equation 1, and  $\mathbf{L}_i$  and  $\mathbf{R}_j$  stand for the left and right image patches of the same size.

### FLEXIBLE MATCHING

We compared the rigid measure of inter-ocular similarity with another measure, introduced here, which we called “flexible” because it tolerates small distortions of corresponding image parts. Now the computation of Equation 1 is applied independently to parts (“sub-patches” or “sub-templates”) of  $\mathbf{L}$  and  $\mathbf{R}$ . The parts may undergo small independent displacements with respect to their original locations, emulating properties of multiple complex cells tuned to adjacent spatial locations (Riesenhuber and Poggio, 1999; Ullman et al., 2002; Serre et al., 2007a,b; Ullman, 2007).

Flexible matching is illustrated in **Figures 3B–D**. Patch  $\mathbf{L}_i$  is divided to  $T$  parts: sub-templates  $\mathbf{L}_i^k$ , where  $k \in [1, \dots, T]$  is the

sub-template index. (In the experiments we tested divisions of the templates into different numbers of sub-templates of equal size: four, nine, and 16.) Patch similarity  $S_{i,j}^{\text{flex}}$  is computed in two steps:

1. Correlation is determined as in Equation 1 separately for each sub-template  $\mathbf{L}_i^k$ , over a set of contiguous horizontal coordinates  $\mathbf{M}_k^j$  (**Figures 3C–D**). The maximal similarity is

$$S_{i,j}^k = \max_{u \in \mathbf{M}_k^j} (C(\mathbf{L}_i^k, \mathbf{R}_u^j)), \quad (3)$$

where  $u$  is the horizontal position of sub-template in the right image. Equation 3 is the MAX-pooling operation. Length  $\mu$  of set  $\mathbf{M}_k^j$  is called *template flexibility*. It is a range of locations near location  $j$  in the right image, for which sub-template similarities are computed, such that

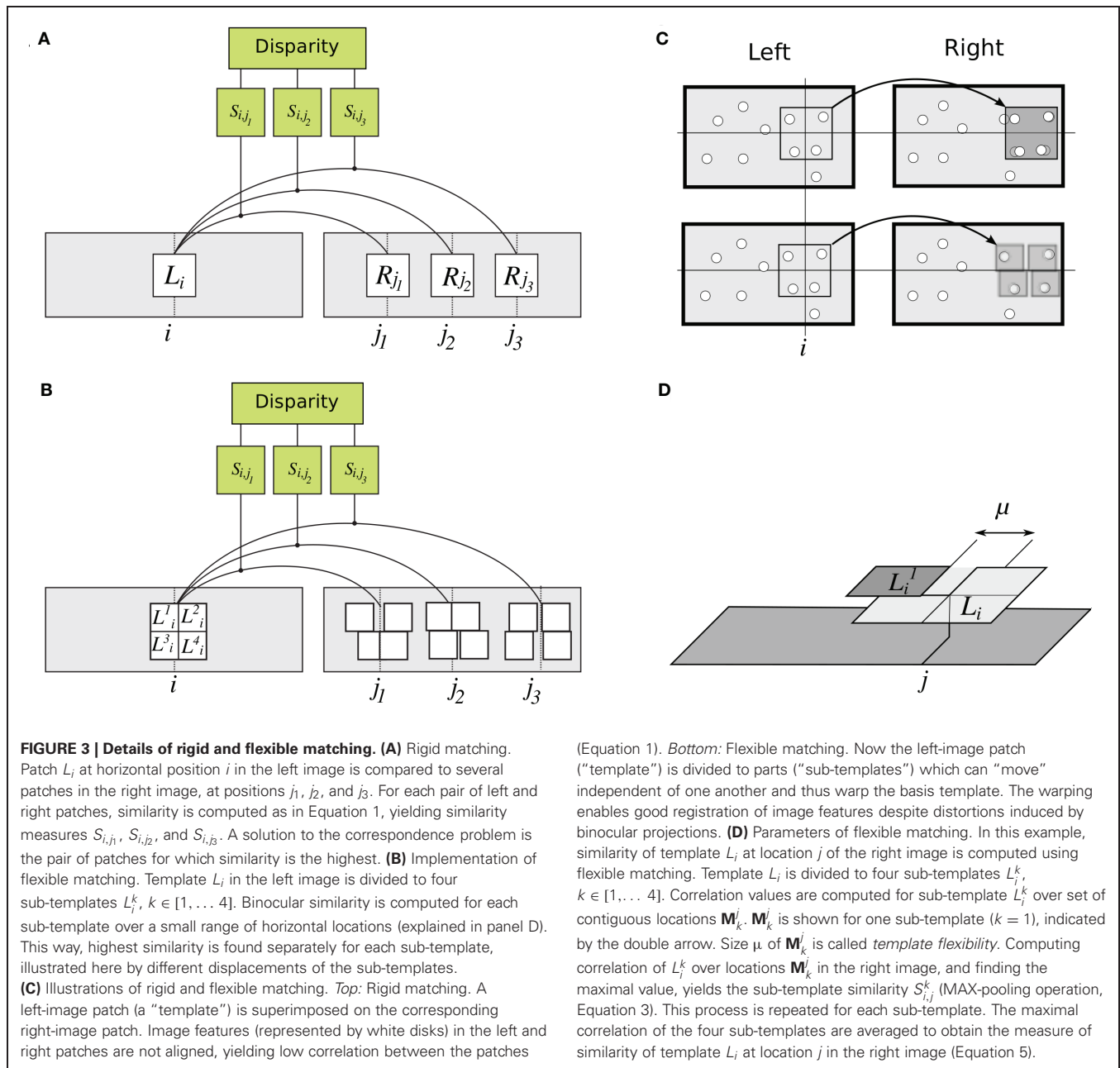
$$\mu = \max(\mathbf{M}_k^j) - \min(\mathbf{M}_k^j) + 1. \quad (4)$$

Template flexibility determines the range of inter-ocular distortions tolerated by the matching procedure. (In these experiments, all sub-templates had the same flexibility  $\mu$ .)

2. Results of MAX-pooling are combined across sub-templates:

$$S_{i,j}^{\text{flex}} = \frac{1}{T} \sum_{k=1}^T S_{i,j}^k. \quad (5)$$

This way, best match is found for each sub-template—over a small image vicinity, independent of other sub-templates, and without computing disparities for each sub-template—possibly “warping” the template. The maximal amount of warping depends on template flexibility  $\mu$ . (As explained in section *Computation of tilt* below, visual systems may automatically select



the magnitude of  $\mu$  that is most suitable for the local slant in the stimulus.)

### COMPUTATION OF DISPARITY

In both rigid and flexible methods, inter-ocular correspondences are found by computing similarity ( $S$ ) between multiple parts of the left and right images of the scene (Figures 1, 2). Suppose a small part of the left image, centered on location  $i$ , is compared to multiple parts of the right image, at locations  $j$  (Figure 3A). (For simplicity, we consider only image parts at the same height in the two images, i.e., we assume the *epipolar constraint*; Hartley and Zisserman, 2003). Thus,  $S_{i,j}$  is the similarity between image patches at locations  $i$  and  $j$ , in the left and right

images, respectively. The patch at  $j^*$  that is most similar to the patch at  $i$  is a solution to the correspondence problem:

$$j^* = \arg \max_j S_{i,j}, \quad (6)$$

such that the estimated binocular disparity at  $i$  is

$$\delta_i = j^* - i. \quad (7)$$

### COMPUTATION OF TILT

We compared how efficiently the rigid and flexible matching methods estimated inter-ocular similarity using a winner-take-all

(WTA) computation (which is believed to be widely implemented in cortical circuits, e.g., Abeles, 1991; Sakurai, 1996; Lee et al., 1999; Flash and Sejnowski, 2001). Assuming that different magnitudes of template flexibility  $\mu$  correspond to different sizes of respective fields in complex cells, the WTA computation amounts to the competition between complex cells with respective fields of different sizes.

We simulated estimation of tilt at point  $P$  in the left image using several samples of disparity  $\delta_i$ : six points  $X_i$  forming vertices of a regular hexagon centered on  $P$  (**Figure 4A**). Disparities  $\delta_i$  were computed as in Equations 6–7 for each sampling point. The similarity measure of Equation 6 was implemented separately for each matching method—rigid matching, flexible matching with fixed  $\mu$ , and flexible matching with variable, “adaptive”  $\mu$ —each leading to a separate estimate of tilt, as follows.

We took advantage of the fact that the sum of vectors  $\vec{PX}_i$ , weighted by disparities  $\delta_i$ :

$$\mathbf{g} = \sum \delta_i \vec{PX}_i, \quad (8)$$

is proportional to surface gradient at  $P$ . Tilt  $\theta$  at point  $P$ , computed separately for each matching method, therefore is

$$\theta = \arctan \frac{g_y}{g_x}, \quad (9)$$

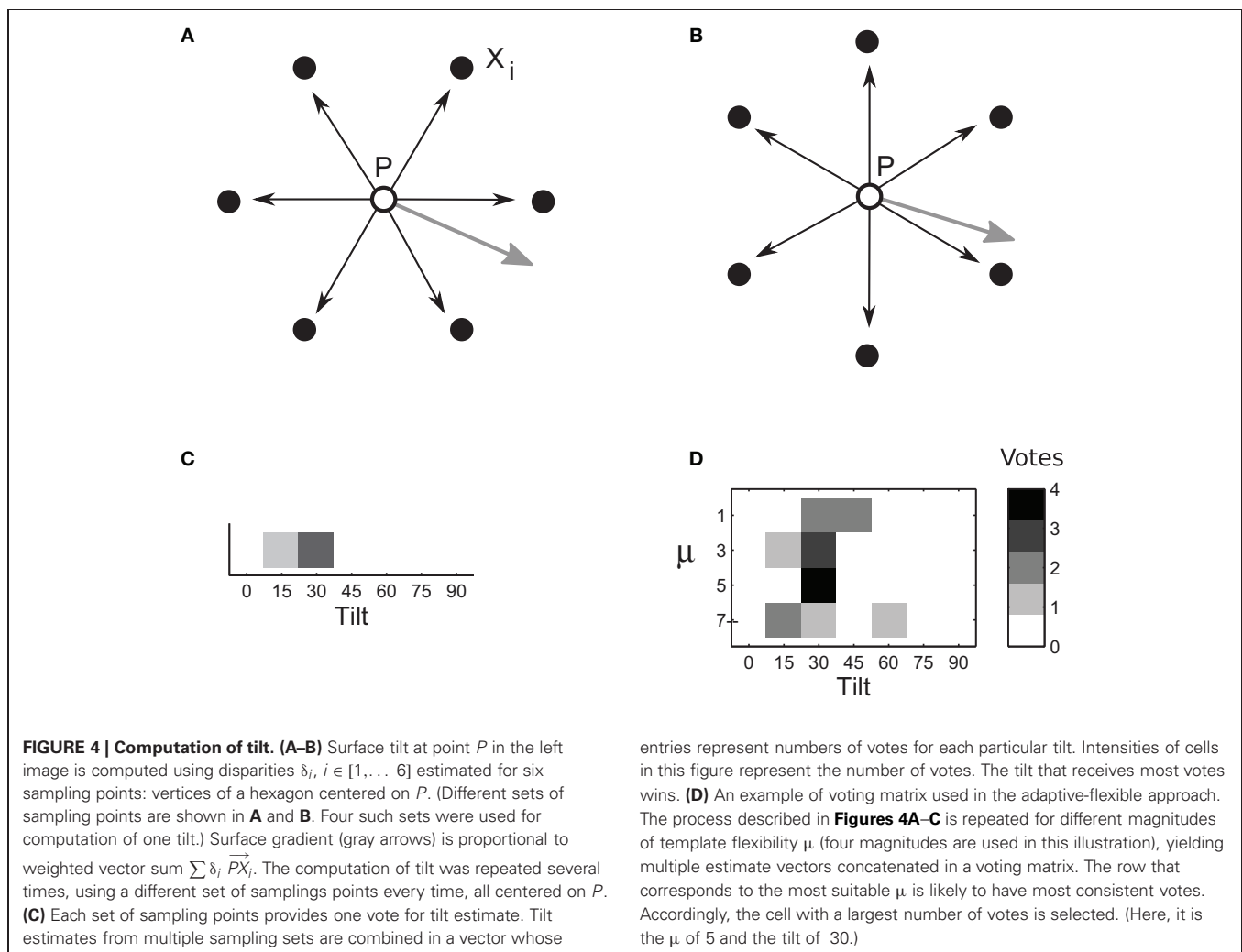
where  $\mathbf{g} = [g_x, g_y]^T$ . The relation between disparity gradient  $\mathbf{g}$ , inter-ocular distance  $I$ , slant  $s$ , and viewing distance  $d$  is (Pollard et al., 1986):

$$|\mathbf{g}| = \frac{I}{d} \arctan(s). \quad (10)$$

Final estimates of tilt were derived by way of population vote, in which several sets of sampling points were used to provide independent estimates.

#### Population vote for rigid matching

In rigid matching, tilt at point  $P$  was estimated using four different sets of sampling points, yielding four tilt estimates. Each set contained six different points, all centered on  $P$  (**Figure 4B**). The four estimates were assembled in a one-dimensional *voting*



matrix, whose entries were cumulative counts of “votes” supporting a particular tilt (**Figure 4C**). (In a separate experiment, we determined that performance of the voting method, using several sampling sets, was better than performance based on the same number of sampling points in one large set.)

### Population vote for flexible matching

In flexible matching, the voting matrix was two-dimensional. Tilt estimates were obtained: for different sampling sets, as in rigid matching, but also for different magnitudes of template flexibility  $\mu$  (**Figure 4D**). The different entries in the matrix represented different hypotheses about the tilt. As in rigid matching, the entry with the largest number of votes was taken as the indicator of tilt. Since flexible matching with a fixed magnitude of  $\mu$  favored a particular range of slants (**Figure 7**), this procedure found the magnitude of  $\mu$  that was most useful for the present stimulus.

We summarize the WTA computation in pseudo-code:

1. Initialize a  $4 \times 7$  voting matrix to 0,
2. For each magnitude of template flexibility  $\mu$  (e.g.,  $\mu \in [1\ 3\ 5\ 7]$ ):
  - For each set of sampling points (four sets of six points each):
    - i. compute disparity (Equations 6–7),
    - ii. compute tilt (**Figure 4A**),
    - iii. increment the voting matrix cell that corresponds to the estimated tilt and the magnitude of template flexibility (**Figure 4D**).
3. Select the tilt indicated by the cell with a highest number of votes.

Each cell in the voting matrix contained the number of times a particular tilt was voted for, using particular template flexibility  $\mu$  (**Figure 4D**). The winning tilt was the one that received most votes. We call this computation “adaptive” because it selects a magnitude of  $\mu$  that is most suitable for current stimulation. We refer to computations that use a single magnitude of  $\mu$ , i.e., where the voting matrix consists of a single row, as “flexible matching with fixed  $\mu$ ”.

We performed two experiments. In Experiment 1, each stimulus represented a planar surface and thus it was characterized by a single tilt (of seven possible tilts), such that a single voting matrix was used for each stimulus (with 28 entries generated by four magnitudes of template flexibility and seven tilts). We also tested larger magnitudes of  $\mu$  and larger numbers of sub-templates, as described in Results.

In Experiment 2, the stimulus represented a concentric sinusoidal surface whose tilts spanned the range of  $0$ – $360^\circ$ . A voting matrix of  $4 \times 360$  was derived for every location in the stimuli. The resulting matrices were each filtered using a  $1 \times 20$  Gaussian kernel, to ensure additive contribution of the nearby votes.

Notably, the computation of tilt made no commitment to particular magnitudes of template flexibility, and consequently no commitment to particular magnitudes of binocular disparity. Multiple hypotheses about template flexibility and binocular disparity coexisted, yielding a single estimate of tilt.

## STIMULI

Stereoscopic stimuli were generated using two types of luminance patterns and they depicted two types of surfaces.

### Luminance patterns

Images of stimulus stereograms contained either textures with a  $1/f$  luminance power spectrum or random-dot textures. The former reproduced the scale invariant property of natural scenes (Ruderman and Bialek, 1994). The latter are commonly used in psychophysical and computational studies of stereopsis. In both cases, the image pairs were obtained by first generating a *source image* (random-dot or  $1/f$ ) and then displacing pixels by half the disparity signal in opposite directions, to obtain the left and right images (as in Banks et al., 2004). In random-dot sources images, the dots formed a perturbed hexagonal grid of  $40 \times 40$  dots. Dots were displaced from positions in a hexagonal grid in random directions, uniformly in all directions, and for a random distance of up to half of inter-dot distance. The  $1/f$  source images were obtained by first generating a white-noise image, whose Fourier amplitude was then modified to obtain the desired power spectrum. Images of both kinds were  $512 \times 512$  pixels. Left and right images were blurred using a Gaussian kernel of size  $6 \times 6$  pixels and standard deviation of 1.5 pixels, to emulate the effect of the optical point-spread function (Campbell and Gubisch, 1966; Banks et al., 2004).

### Surfaces

In Experiment 1, stimuli depicted flat surfaces at different slants and tilts (**Figures 1A,B**), using both random-dot and  $1/f$  luminance textures. For each combination of slant and tilt, we generated 100 random-dot stimuli and 100 naturalistic stimuli. The tilts ranged from  $0$  to  $90^\circ$ , and surface disparity gradients (Equation 10) ranged from  $0$  to  $0.95$  (**Figure 5**). Tilt estimates were derived for stimulus center using Equation 10. For each slant, tilt, and stimulus type (random-dot or  $1/f$ ), we computed accuracy of tilt discrimination using the rigid and flexible matching methods. (Accuracy is the frequency of cases where the estimated tilt was equal to the true tilt). **Figures 7–8** are summaries of accuracy, plotted as a function of slant for the two matching methods, using different luminance patterns in the stimulus.

In Experiment 2, the stimuli were generated using only  $1/f$  luminance textures, depicting a surface whose depth was modulated according to a concentric sinusoidal function, illustrated in **Figure 6**.

The slope of this surface is the disparity gradient. The larger the slope, the stronger the inter-ocular dissimilarity, and so a larger template flexibility is needed to attain accurate binocular matching.

## RESULTS

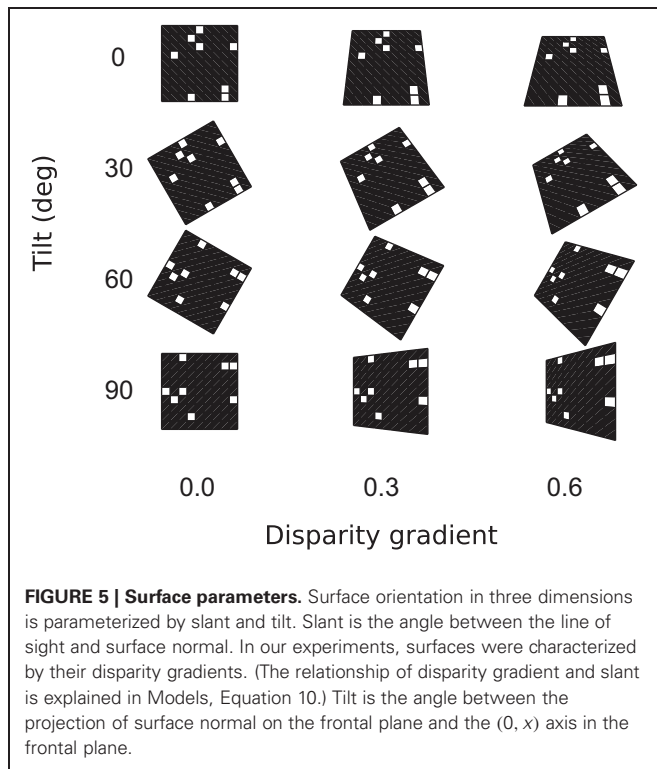
### EXPERIMENT 1

We measured accuracy of tilt estimation as a function of slant using different matching methods:

#### Rigid matching

Outcomes of rigid matching in Experiment 1 are represented by the black curve in **Figure 7**, for  $1/f$  stimuli in panel A and for





random-dot stimuli in panel B. For 1/f stimuli, performance of the rigid procedure peaked at the disparity gradients of 0.1–0.4. For random-dot stimuli, performance peaked near the disparity gradient of 0.16 and then abruptly decreased, falling to half of its peak performance at the disparity gradient of 0.2. For disparity gradients larger than 0.4 in 1/f stimuli, and larger than 0.16 in random-dot stimuli, the inter-ocular distortion of corresponding patches was too large for the rigid procedure to find correct matches, which explains the sharp decrease in performance.

#### Flexible matching with fixed flexibility

Outcomes of flexible matching with fixed magnitudes of  $\mu$  are represented by the colored curves in Figure 7, for 1/f stimuli in

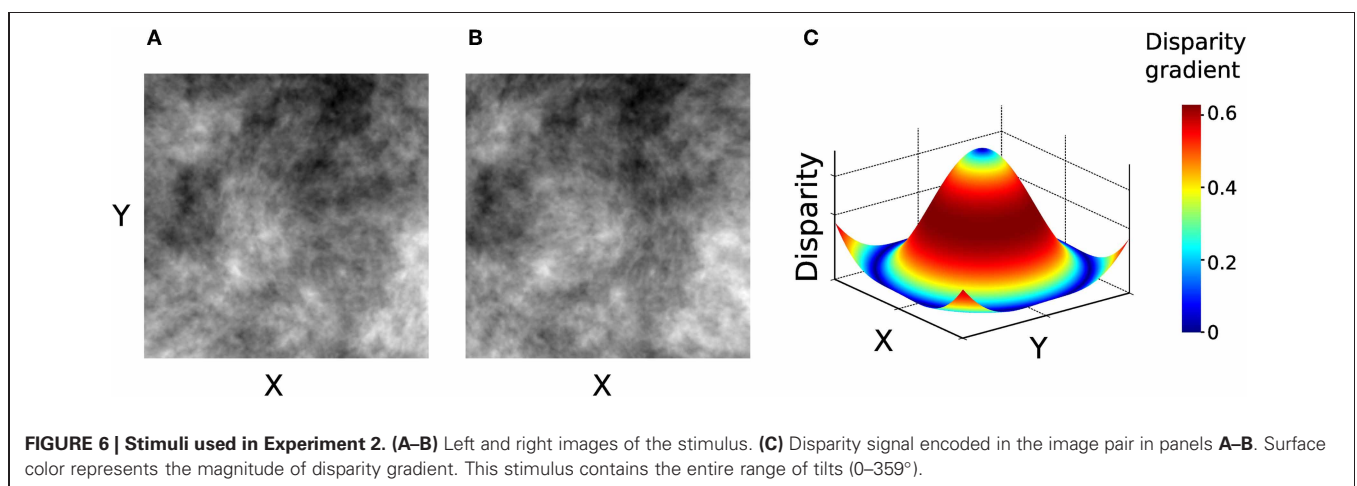
panel A and for random-dot stimuli in panel B. (The black curve represents outcomes of rigid matching.) As template flexibility increased, the peak of performance shifted toward the higher disparity gradients for both 1/f and random-dot stimuli. Maximal performance was high for small and intermediate magnitudes of  $\mu$ , but it deteriorated at the large magnitudes of  $\mu$  (9 and 13).

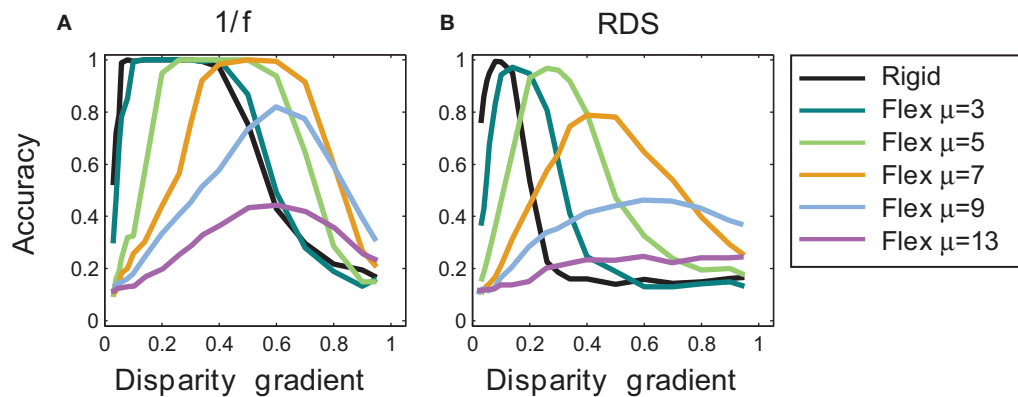
The preference for higher disparity gradients at larger magnitudes of  $\mu$  is expected because large template flexibility entails high tolerance to dissimilarity of corresponding image patches. But as flexibility  $\mu$  is increased yet further, the matching is increasingly afflicted by spurious matches, which explains the drop of performance at the two largest magnitudes of  $\mu$ .

In other words, Figure 7 captures a tradeoff between effects of different magnitudes of template flexibility. Flexible matching with low magnitudes of  $\mu$  favors matching of similar image patches, making the matching procedure miss the corresponding patches under high inter-ocular deformation at large disparity gradients. Flexible matching with high magnitudes of  $\mu$  does not miss the correspondences under high inter-ocular deformation, but it is prone to register spurious matches. In effect, performance curves for flexible matching with fixed magnitudes of  $\mu$  shift along the dimension of disparity gradient: the larger  $\mu$  the farther the shift toward large disparity gradients.

#### Flexible matching with variable flexibility

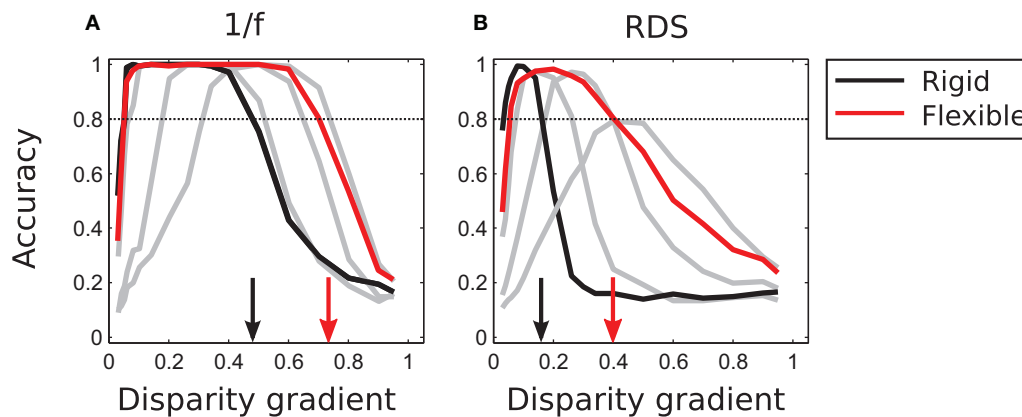
As demonstrated in Figure 7, a fixed amount of template flexibility favors a particular range of slants. A system employing different magnitudes of template flexibility would be able to take advantage of the degree of flexibility that is most suitable for current stimulus and thus yield reliable performance for a large range of slants. Performance of such an “adaptive” system (described in section “Population vote for flexible matching” in “Models and Methods”) is represented by the red curve in Figure 8. (The black curve is the same as in Figure 7; it represents outcomes of rigid matching.) For 1/f stimuli, maximal performance of adaptive matching was reached for disparity gradients in the range of 0.1–0.6. For the random-dot stimuli, performance of adaptive matching peaked at the disparity gradient of 0.2. The red curve in Figure 8 effectively circumscribes the pertinent curves of Figure 7. (Very large magnitudes of template flexibility did





**FIGURE 7 | Tilt discrimination performance in Experiment 1: rigid matching vs. flexible matching with fixed flexibility.** (A) Results for 1/f stimuli, using rigid matching (black curve) and flexible matching with fixed magnitudes of template flexibility  $\mu$  (colored curves). Accuracy of tilt estimation is plotted as a function of surface slant. (Perfect

performance is 1 and random performance is 0.14.) In flexible matching, performance depends on template flexibility  $\mu$ : the higher the template flexibility, the larger the slant at which performance peaks. (B) Results for random-dot stereograms, using the same convention as in panel A.



**FIGURE 8 | Tilt discrimination performance in Experiment 1: rigid matching vs. flexible matching with adaptive selection of template flexibility.** (A) Results for 1/f stimuli. Flexible matching with variable template flexibility (red curve) attained a much larger range of correct classification than rigid matching (black curve). (The gray curves represent performance of flexible matching with different fixed magnitudes

of template flexibility  $\mu$ , the same as those rendered as colored curves in Figure 7, but excluding  $\mu$  of 9 and 13.) (B) Results for random-dot stimuli, using the same convention as in panel A. In both panels, the arrows mark the magnitudes of disparity gradient at which the descending arms of performance curves crossed the 0.8 level of accuracy.

not affect performance of the adaptive process, because matching performance at the large magnitudes of  $\mu$ —here  $\mu \geq 9$ —was crippled by spurious matches.)

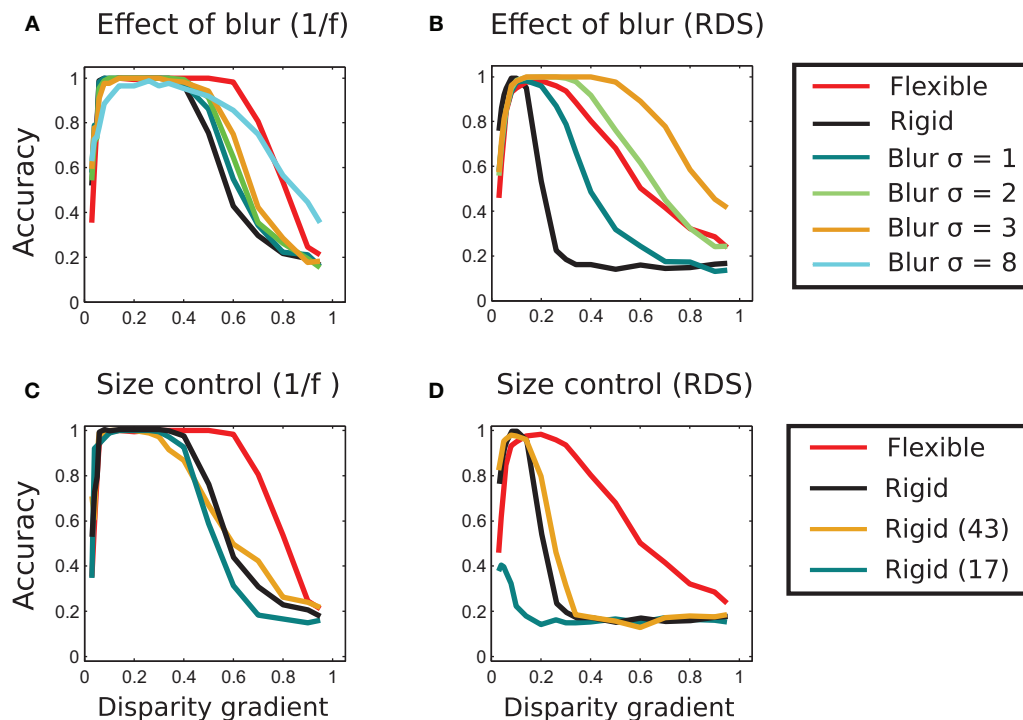
To summarize, flexible matching yields much better performance than rigid matching at large disparity gradients, explained by the capability of flexible matching to identify corresponding image parts distorted due to the viewing geometry. Provided multiple degrees of flexibility, flexible matching is also capable of reliable performance at a much larger range of disparity gradients than rigid matching.

#### REDUCTION OF INTER-OCULAR DISTORTIONS BY IMAGE BLUR

A method previously proposed to facilitate binocular matching and overcome inter-ocular distortions is to blur images. Blurring

by the front-end (optical and post-optical) stages of the biological visual process (Campbell and Gubisch, 1966; Geisler, 1989) scatters luminance of monocular image features that do not align across the left and right images, thus improving inter-ocular registration of the features (e.g., Berg and Malik, 2001).

We applied Gaussian horizontal blur to each stimulus image of our stimuli. In Figures 9A,B we plot tilt discrimination performance using different amounts of blur, parameterized by size  $\sigma$  of the blurring kernel, for 1/f stimuli in panel A and random-dot stimuli in panel B. For 1/f stimuli, blur marginally improved performance of rigid matching, using  $\sigma \in [1\ 2\ 3]$ . (Results of rigid and flexible matching without blur are also shown, using the same black and red curves as in Figure 8). Increasing  $\sigma$  further reduced



**FIGURE 9 | Effect of blur and template size. (A)** Effect of image blur on matching performance in 1/f stereograms. The blue, green, and orange curves represent results of matching using different strengths of blur. The blurring marginally increases the range of perceived slants and performance of rigid matching. Even at very large blur ( $\sigma = 8$  in the figure) the range of high performance is wide, but the maximal performance of unity is never reached. The curves representing performance of adaptive (red) and rigid (black) matching (using  $33 \times 33$  pixel templates) are copied from **Figure 8A** for reference. **(B)** Effect of blur in random-dot stereograms. Here, the blurring significantly improves performance of the rigid method. For  $\sigma = 2$  and 3, the range of slants for correctly identified tilts is wider than in the adaptive-flexible approach (red curve, as in **Figure 8B**.) Result for the blur of

$\sigma = 8$  is not shown here to avoid clutter, as performance of rigid matching with the blur of  $\sigma = 3$  already exceeds performance of flexible matching. **(C)** Effect of template size in rigid matching with 1/f stimuli. Rigid matching using templates smaller ( $17 \times 17$  pixels, blue curve) and larger ( $45 \times 45$ , orange) than the original size ( $33 \times 33$ , black) yielded approximately the same performance as the templates used in the rest of the study. **(D)** Effect of template size in rigid matching with random-dot stimuli. Performance of the larger template size ( $45 \times 45$  pixels, orange) is approximately the same as performance of the original size (**Figure 7B**). Performance is significantly reduced for smaller templates ( $17 \times 17$  pixels, blue). The curves representing performance of adaptive (red) and rigid (black) matching with the  $33 \times 33$  pixels templates are copied from **Figure 8B**.

the peak performance of rigid matching, such that it failed to reach accuracy of 1 (shown for  $\sigma = 8$  in panel A).

For random-dot stimuli, however, blur significantly improved performance of rigid matching, yielding better results than flexible matching. That is, advantages of flexible matching hold for the naturalistic stimuli and not for the random-dot stimuli.

### ROLE OF TEMPLATE SIZE

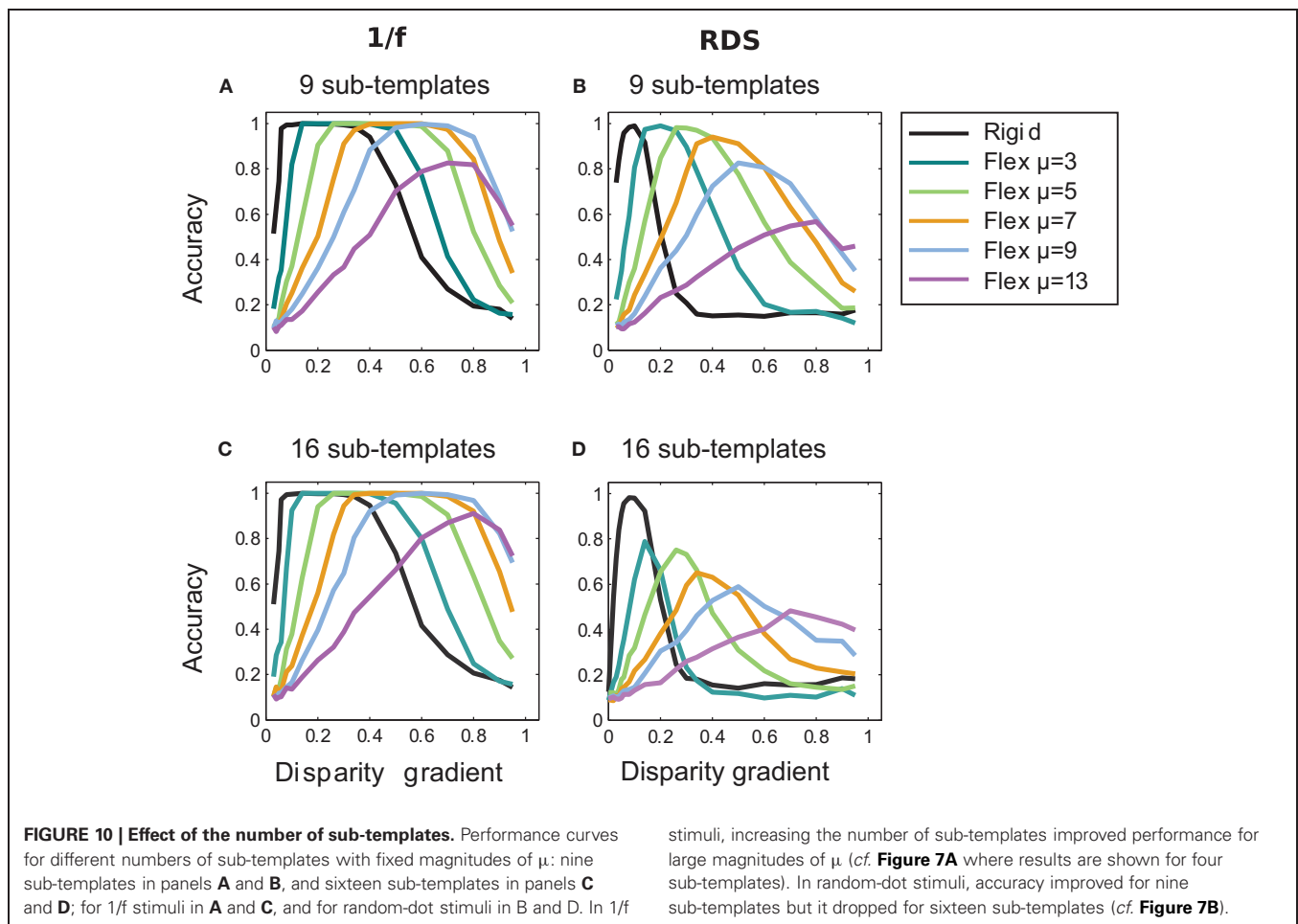
We ruled out the possibility that the better performance of flexible matching can be accounted for by a particular choice of template size. We did so by evaluating performance of a rigid matching procedure with template sizes  $17 \times 17$  and  $43 \times 43$  pixels (original size:  $33 \times 33$  pixels). The results are plotted in **Figure 9**: for 1/f stimuli in panel C and for random-dot stimuli in panel D. The plots indicate that flexible matching (red curve, also shown in **Figure 8A**) performs significantly better than rigid matching with the other template sizes. We also plot performance of rigid matching using the (original) template size

of  $33 \times 33$  pixels (black curve, for comparison). Performance of the flexible model for template sizes  $17 \times 17$  and  $43 \times 43$  pixels (not shown in this figure to avoid clutter) was similar to performance of the adaptive procedure with template size  $33 \times 33$  used in Experiment 1. Notably, performance of rigid matching is worse than that of flexible matching when the size of rigid templates is the same as the size of sub-templates of flexible matching (**Figures 9C,D**).

### EFFECT OF THE NUMBER OF SUB-TEMPLATES

We repeated the above experiments using a larger numbers of sub-templates: nine and 16, using respectively  $3 \times 3$  and  $4 \times 4$  square sub-templates, 11 pixels wide for nine sub-templates and 9 pixels wide for 16 sub-templates. (Sub-templates slightly overlapped in the latter case since the 33-pixel templates did not evenly divide to the 9-pixel sub-templates.)

Results of matching with the larger number of sub-templates for fixed  $\mu$  are shown in **Figure 10** for random-dot and 1/f stimuli. In comparison to results for the four sub-templates



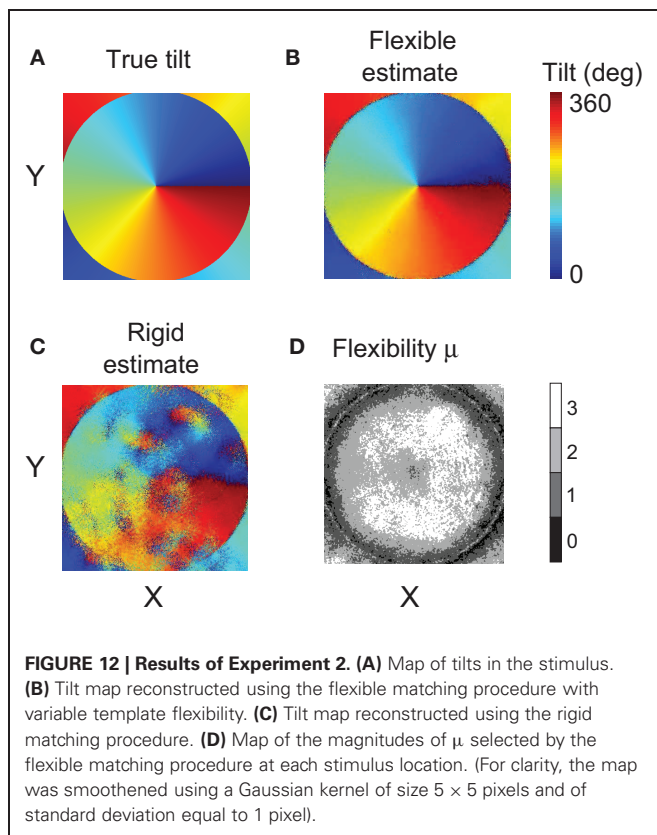
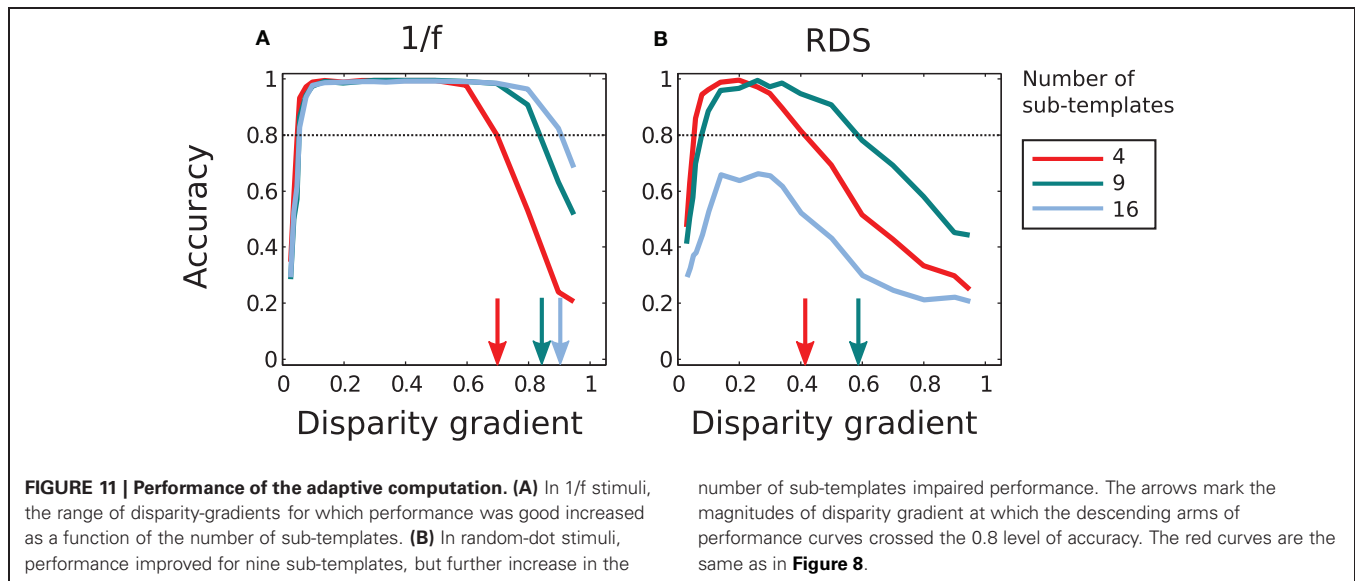
(**Figure 7**), the larger number of sub-templates improved performance at high magnitudes of  $\mu$  (9 and 13) in the  $1/f$  stimuli (**Figures 10A** and **C**), consistent with the view that the increased flexibility of matching has a larger tolerance to inter-ocular distortions. In the random-dot stimuli, performance improved for nine sub-templates but did not improve for sixteen sub-templates (**Figures 10B** and **D**), indicating that for the scarce luminance distribution in the random-dot stimuli, the additional flexibility of matching was beneficial up to a point at which the smaller sub-templates failed to capture patterns of luminance sufficiently unique to support reliable matching.

**Figure 11** summarizes performance of the adaptive system that employs different numbers of sub-templates. Increasing the number of sub-templates improved performance, in particular for  $1/f$  stimuli. (Now all magnitudes of  $\mu$  were used in the adaptive computation since performance improved at large  $\mu$  with nine and sixteen sub-templates, in contrast to the lack of such improvement with four sub-templates.) The range of disparity gradients at which performance was high increased with the number of sub-templates in the  $1/f$  stimuli (panel **A**). But in random-dot stimuli performance improved with nine sub-templates while it was impaired with sixteen sub-templates, as explained in the previous paragraph (**Figure 11A**).

## EXPERIMENT 2

In Experiment 2 we investigate the ability of flexible matching to tolerate different amounts of inter-ocular distortion in different parts of the stimulus. Now we used a complex stimulus that contains multiple slants (**Figure 6**). We applied rigid and flexible matching procedures at all locations in this stimulus yielding maps of estimated tilt. Flexible matching employed four sub-templates. Instead of the hexagonal sampling used in Experiment 1, now positions of the sampling points were randomized (or else the regular placement of sampling points created artifacts in maps of estimated tilt) while care was taken that the arrangement of sampling points did not introduce a directional bias (i.e., that the covariance matrix of sample-point coordinates was proportional to the identity matrix and so Equation 8 held).

**Figure 12** presents the map of true tilt in panel **A**, and the maps computed using different matching methods in panels **B** and **C**. Visual inspection of the maps makes it clear that flexible matching yielded a consistently more accurate tilt estimation than rigid matching. In particular, rigid matching performed poorly where the disparity gradient was large: on the flanks of the central peak of disparity. The tilt map by flexible matching is significantly more similar to the map of true tilt.



In **Figure 12D** we plot the magnitudes of template flexibility  $\mu$  selected by the flexible matching procedure with variable template flexibility at each location in the stimuli. The plot shows that high magnitudes of  $\mu$  were preferred where the disparity gradient was high (on the flanks of the disparity peak) and low magnitudes of  $\mu$  were preferred where the gradient was low. The light ring in the

periphery corresponds to the trough of disparity, where disparity gradient was zero and surface tilt was undefined. At these points, no particular magnitude of  $\mu$  was preferred.

We computed mean errors of tilt estimated using the different matching methods: rigid, flexible with fixed magnitudes of  $\mu$ , and flexible with variable magnitudes of  $\mu$ . The mean error of tilt estimation was the mean absolute difference of the estimated and true tilts, modulo  $180^\circ$ , across all stimulus pixels. The mean error was below  $5^\circ$  for flexible matching, and it was larger than  $30^\circ$  for rigid matching.

## DISCUSSION

We investigated how the well-known capacity of binocular complex cells for spatially invariant computation may benefit stereoscopic vision. We compared two approaches to binocular matching. One approach uses computations implicit in the standard model of binocular matching. We call this approach “rigid matching” because it favors identical left and right images. The other approach uses spatially invariant computations. It is “flexible” in the sense it allows for small independent displacements of fragments of left and right image parts, locally warping the images, thus helping to find corresponding image parts distorted by binocular projection. We modeled flexible matching using the computational framework of MAX-pooling (Riesenhuber and Poggio, 1999; Ullman et al., 2002; Serre et al., 2007a,b; Ullman, 2007).

Differences of outcomes from rigid and flexible matching were striking. Flexible matching was able to support efficient matching for a much larger range of slants than rigid matching, both in random-dot stereograms and in stimuli with naturalistic (1/f) luminance distributions (**Figure 8**). We found that performance of rigid matching significantly improved when combined with image blur (Berg and Malik, 2001) (our **Figures 9A,B**), but this result held only in random-dot stimuli. In stimuli with naturalistic luminance distributions, blurring did not improve



performance of rigid matching, indicating that the spatially invariant computation is suited for perception of the natural visual environment.

In flexible matching, the amount of inter-ocular distortion tolerated by the matching process depends on the parameter we called template flexibility ( $\mu$ , Equation 4) which represents different receptive field sizes of binocular complex cells. We showed that the amount of template flexibility most suitable for the current stimulus could be determined automatically, by WTA competition between cells with respective fields of different sizes. This competition may proceed concurrently and independently at many different stimulus locations, making binocular matching highly adaptive to the diverse scene geometry (**Figure 12**). It is possible that adaptive blurring can further improve performance: further studies should explore how adaptive blurring and adaptive flexible matching can be combined optimally.

Tanabe et al. (2004) found evidence of competition between hypotheses about binocular correspondence in cortical area V4. Such competition is akin to the process of “voting” in our study, which insured that the most suitable amount of matching flexibility was used at every location in the stimulus. Yet physiological studies have shown that the mechanisms that encode surface shape span many cortical areas from primary to inferotemporal cortical areas (Burkhalter and Essen, 1986; Uka et al., 2000, 2005; Qiu and von der Heydt, 2005; Sanada and Ohzawa, 2006), making it difficult to localize the neural substrate for these mechanisms. Indeed, it is likely that these mechanisms are distributed across several cortical areas.

We have focused on one component of binocular matching: the computation of inter-ocular similarity. We have shown that spatially-invariant computation of similarity is useful for discovering the corresponding image parts distorted by binocular projection. Since spatially-invariant computation is believed to be performed by binocular complex cells, we consider implications of our study for understanding the role of these cells in biological stereopsis.

The standard view is that binocular complex cells play the role of “disparity detectors”—i.e., they compute binocular disparity (Qiang, 1994; Ohzawa, 1998; Anzai et al., 1999). Our study suggests a different picture, that binocular complex cells cooperate in the computation of inter-ocular similarity. Indeed, receptive fields of individual complex cells are often too small to sufficiently represent the spatial-frequency content of the stimulus, which is essential for identifying corresponding image parts (as Banks et al., 2004, pointed out). We propose that inter-ocular similarity is computed by populations of complex cells with retinotopically adjacent respective fields of different sizes. This arrangement will have sufficient flexibility for finding corresponding image parts of variable size and under variable amount of image distortion.

Our results also suggest that binocular visual systems may do well by avoiding an early commitment to binocular disparity. Models of stereopsis commonly derive a single map of binocular disparity as soon as inter-ocular similarities are computed. In our framework, multiple disparity maps are computed using

different magnitudes of template flexibility, simulating computations by binocular complex cells with receptive fields of different size. The alternative disparity maps coexist up to the stage where a higher-order stimulus property (such as tilt) is computed, taking advantage of the information that would be lost had the system committed to a single map of disparity early on. Computational studies of other sensory processes showed that preserving ambiguity about stimulus parameters until late stages of the sensory process can benefit system performance: in models of feedforward computations (e.g., Serre et al., 2007a,b and as implemented here) and also in models that involve feedback (e.g., Epshtein et al., 2008), where outcomes of computations at a late stage help to disambiguate results of early computations.

Our results indicate that the choice of stimulus for probing the computation of inter-ocular similarity is significant. Spatially invariant computations were more beneficial for stimuli with naturalistic distribution of luminance than for random-dot stimuli. The advantage was more pronounced as the flexibility of matching increased, both in terms of the spatial range of inter-ocular comparisons (**Figures 7, 8**) and in terms of the number of sub-templates (e.g., **Figure 11**). A likely reason for the stimulus effect is the fact that correlation measures of image similarity are highly sensitive to statistics of luminance in the images (Sharpee et al., 2006; Vidal-Naquet and Tanifuji, 2007). These findings suggest that results of studies of biological stereopsis that involved random-dot luminance patterns may need to be revisited. Also, the possibility should be considered that matching is adaptive and so changes in luminance statistics may yield a different outcomes of matching.

For example, Allenmark and Read (2010) found that rigid matching failed to account for human perception of slanted surfaces in random-dot stimuli. Allenmark and Read (2011) proposed that the inconsistency between outcomes of rigid matching and human performance could be resolved by adaptively increasing the size of the correlation window: the larger the disparity the larger the window (*cf.* Kanade and Okutomi, 1994). Future studies should compare human performance and performance of the alternative methods of matching using stimuli with naturalistic distribution of luminance. Moreover, a combination of the adaptive use of spatial invariance (as in our study) and adaptive use of the size of correlation window (as in Kanade and Okutomi, 1994 and Allenmark and Read, 2011) is likely to be most beneficial, such that a full model of the biological computation of inter-ocular similarity will incorporate adaptive spatially-invariant matching on multiple spatial scales, helping to explain the fact that biological vision is capable of reliable performance at yet higher disparity gradients (Tyler, 1973; Burt and Julesz, 1980; Allenmark and Read, 2010, 2011) than observed in the present study.

## ACKNOWLEDGMENTS

Michel Vidal-Naquet was supported by the RIKEN Foreign Postdoctoral Researcher program. Sergei Gepshtein was supported by the Swartz Foundation and grants from NSF (#1027259) and NIH (R01 EY018613).

## REFERENCES

- Abeles, M. (1991). *Corticonics: Neural Circuits of the Cerebral Cortex*. New York, NY: Cambridge University Press.
- Allenmark, E., and Read, J. C. A. (2010). Detectability of sine-versus square-wave disparity gratings: a challenge for current models of depth perception. *J. Vis.* 10, 1–16.
- Allenmark, E., and Read, J. C. A. (2011). Spatial stereoresolution for depth corrugations may be set in primary visual cortex. *PLoS Comput. Biol.* 7:e1002142. doi: 10.1371/journal.pcbi.1002142
- Anzai, A., Ohzawa, I., and Freeman, R. (1999). Neural mechanisms for processing binocular information II. Complex cells. *J. Neurophysiol.* 82, 909–924.
- Banks, M., Gepshtein, S., and Landy, M. (2004). Why is spatial stereoresolution so low? *J. Neurosci.* 24, 2077–2089.
- Banks, M., Gepshtein, S., and Rose, H. F. (2005). “Local cross-correlation model of stereo correspondence,” in *Proceedings of SPIE: Human Vision and Electronic Imaging*, (San Jose, CA), 53–61.
- Berg, A., and Malik, J. (2001). “Geometric blur for template matching,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 1, (Kauai, HI), 607–614.
- Burkhalter, A., and Essen, D. V. (1986). Processing of color, form and disparity information in visual areas vp and v2 of ventral extrastriate cortex in the macaque monkey. *J. Neurosci.* 6, 2327–2351.
- Burt, P., and Julesz, B. (1980). A disparity gradient limit for binocular fusion. *Science* 208, 615–617.
- Campbell, F. W., and Gubisch, R. W. (1966). Optical quality of the human eye. *J. Physiol.* 186, 558–578.
- Cormack, L., Stevenson, S., and Schor, C. (1991). Interocular correlation, luminance contrast and cyclopean processing. *Vision Res.* 31, 2195–2207.
- Cumming, B., and DeAngelis, G. (2001). The physiology of stereopsis. *Annu. Rev. Neurosci.* 24, 203–238.
- Cumming, B. G., and Parker, A. J. (1997). Responses of primary visual cortical neurons to binocular disparity without depth perception. *Nature* 389, 280–283.
- Epshtein, B., Lifshitz, I., and Ullman, S. (2008). Image interpretation by a single bottom-up top-down cycle. *Proc. Natl. Acad. Sci. U.S.A.* 105, 14298–14303.
- Filippini, H., and Banks, M. (2009). Limits of stereopsis explained by local cross-correlation. *J. Vis.* 9, 1–18.
- Flash, T., and Sejnowski, T. (2001). Computational approaches to motor control. *Curr. Opin. Neurobiol.* 11, 655–662.
- Fleet, D., Wagner, H., and Heeger, D. (1996). Neural encoding of binocular disparity: energy model, position shifts and phase shifts. *Vision Res.* 36, 1839–1857.
- Geisler, W. S. (1989). Sequential ideal-observer analysis of visual discrimination. *Psychol. Rev.* 96, 267–314.
- Haefner, R., and Cumming, B. (2008). Adaptation to natural binocular disparities in primate v1 explained by a generalized energy model. *Neuron* 57, 147–158.
- Hartley, R., and Zisserman, A. (2003). *Multiple View Geometry in Computer Vision*. Cambridge, UK: Cambridge University Press.
- Hyvarinen, A., and Hoyer, P. (2000). Emergence of phase and shift invariant features by decomposition of natural images into independent feature subspaces. *Neural Comput.* 12, 1705–1720.
- Kanade, T., and Okutomi, M. (1994). A stereo matching algorithm with an adaptive window: theory and experiment. *IEEE Trans. Pattern Anal. Mach. Intell.* 16, 920–942.
- Karklin, Y., and Lewicki, M. S. (2009). Emergence of complex cell properties by learning to generalize in natural scenes. *Nature* 457, 83–86.
- Lee, D. K., Itti, L., Koch, C., and Braun, J. (1999). Attention activates winner-takes-all competition among visual filters. *Nat. Neurosci.* 2, 375–381.
- Ohzawa, I. (1998). Mechanisms of stereoscopic vision: the disparity energy model. *Curr. Opin. Neurobiol.* 8, 509–515.
- Ohzawa, I., DeAngelis, G., and Freeman, R. (1990). Stereoscopic depth discrimination in the visual cortex: neurons ideally suited as disparity detectors. *Science* 249, 1037–1041.
- Pollard, S., Porrill, J., Mayhew, J. E., and Frisby, J. (1986). “Disparity gradient, lipschitz continuity, and computing binocular correspondences,” in *Robotics Research: The Third International Symposium*, eds O. D. Faugeras and D. Gi-ralt (Gouvieux-Chantilly, France: MIT Press), 19–26.
- Qian, N., and Zhu, Y. (1997). Physiological computation of binocular disparity. *Vision Res.* 37, 1811–1827.
- Qiang, N. (1994). Computing stereo disparity and motion with known binocular cell properties. *Neural Comput.* 6, 390–404.
- Qiu, F. T., and von der Heydt, R. (2005). Figure and ground in the visual cortex: v2 combines stereoscopic cues with gestalt rules. *Neuron* 47, 155–166.
- Riesenhuber, M., and Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nat. Neurosci.* 2, 1019–1025.
- Ruderman, D. L., and Bialek, W. (1994). Statistics of natural images: scaling in the woods. *Phys. Rev. Lett.* 73, 814–817.
- Sakurai, Y. (1996). Population coding by cell assemblies what it really is in the brain. *Neurosci. Res.* 26, 1–16.
- Sanada, T., and Ohzawa, I. (2006). Encoding of three-dimensional surface slant in cat visual areas 17 and 18. *J. Neurophysiol.* 95, 2768–2786.
- Serre, T., Oliva, A., and Poggio, T. (2007a). A feedforward architecture accounts for rapid categorization. *Proc. Natl. Acad. Sci. U.S.A.* 104, 6424–6429.
- Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., and Poggio, T. (2007b). Object recognition with cortex-like mechanisms. *IEEE Trans. Pattern Anal. Mach. Intell.* 29, 411–426.
- Sharpee, T., Sugihara, H., Kurgansky, A., Rebrik, S., Stryker, M., and Miller, K. (2006). Adaptive filtering enhances information transmission in visual cortex. *Nature* 439, 936–942.
- Tanabe, S., Umeda, K., and Fujita, I. (2004). Rejection of false matches for binocular correspondence in macaque visual cortical area v4. *J. Neurosci.* 24, 8170–8180.
- Tyler, C. (1973). Stereoscopic vision: cortical limitations and a disparity scaling effect. *Science* 181, 276–278.
- Tyler, C., and Julesz, B. (1978). Binocular cross-correlation in time and space. *Vision Res.* 18, 101–105.
- Uka, T., Tanabe, S., Watanabe, M., and Fujita, I. (2005). Neural correlates of fine depth discrimination in monkey inferior temporal cortex. *J. Neurosci.* 25, 10796–10802.
- Uka, T., Yoshiyama, K., Kato, M., and Fujita, I. (2000). Disparity selectivity of neurons in monkey inferior temporal cortex. *J. Neurophysiol.* 84, 120–132.
- Ullman, S. (2007). Object recognition and segmentation by a fragment-based hierarchy. *Trends Cogn. Sci.* 11, 58–64.
- Ullman, S., Vidal-Naquet, M., and Sali, E. (2002). Visual features of intermediate complexity and their use in classification. *Nat. Neurosci.* 5, 682–687.
- Vidal-Naquet, M., and Tanifuji, M. (2007). “The effective resolution of correlation filters applied to natural scenes,” in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, Beyond Patches Workshop*, (Minneapolis, MN), 1–6.
- Yu, A. J., Giese, M. A., and Poggio, T. (2002). Biophysiological plausible implementations of the maximum operation. *Neural Comput.* 14, 2857–2881.

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 11 December 2011; accepted: 26 June 2012; published online: 16 July 2012.

Citation: Vidal-Naquet M and Gepshtein S (2012) Spatially invariant computations in stereoscopic vision. *Front. Comput. Neurosci.* 6:47. doi: 10.3389/fncom.2012.00047

Copyright © 2012 Vidal-Naquet and Gepshtein. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in other forums, provided the original authors and source are credited and subject to any copyright notices concerning any third-party graphics etc.