INTEGRATING COMPUTATIONAL AND NEURAL FINDINGS IN VISUAL OBJECT PERCEPTION

EDITED BY: Judith C. Peters, Hans P. Op de Beeck and Rainer Goebel

PUBLISHED IN: Frontiers in Computational Neuroscience

frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-88919-873-3
DOI 10.3389/978-2-88919-873-3

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.

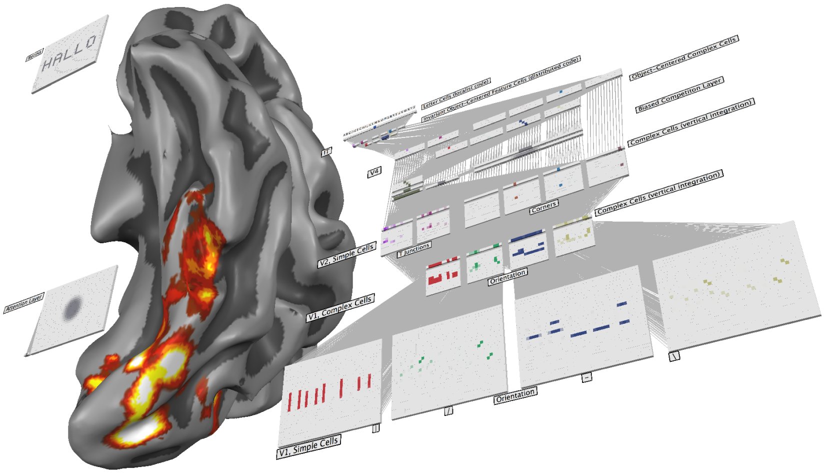
By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

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**

INTEGRATING COMPUTATIONAL AND NEURAL FINDINGS IN VISUAL OBJECT PERCEPTION

Topic Editors:
Judith C. Peters, Maastricht University and Netherlands Institute for Neuroscience, Netherlands
Hans P. Op de Beeck, University of Leuven, Belgium
Rainer Goebel, Maastricht University and Netherlands Institute for Neuroscience, Netherlands


Retina
HALLO
Attention Layer
Letter Cells (localist code)
Invariant Object-Centered Feature Cells (distributed code)
Object-Centered Complex Cells
Biased Competiton Layer
Complex Cells (vertical integration)
IT
V4
Corners
Complex Cells (vertical integration)
V2, Simple Cells
T Junctions
Orientation
V1, Complex Cells
Orientation
V1, Simple Cells
|
/
—
\

Multi-layer neural network modeling mechanisms of invariant object-recognition in the visual ventral stream and corresponding activations projected on an inflated cortical sheet (bottom view). Screenshot of Neurolator (BrainInnovation, Maastricht, The Netherlands), a neural network simulation software package in which simulated activity can be projected to the same anatomical "brain space" as empirically acquired neuroimaging data, thereby allowing direct quantitive, spatiotemporal comparisons.

The articles in this Research Topic provide a state-of-the-art overview of the current progress in integrating computational and empirical research on visual object recognition. Developments in this exciting multidisciplinary field have recently gained momentum: High performance computing enabled breakthroughs in computer vision and computational neuroscience. In parallel, innovative machine learning applications have recently become available for datamining the large-scale, high resolution brain data acquired with (ultra-high field) fMRI and dense multi-unit recordings. Finally, new techniques to integrate such rich simulated and empirical datasets for direct model testing could aid the development of a comprehensive brain model. We hope that this Research Topic contributes to these encouraging advances and inspires future research avenues in computational and empirical neuroscience.

Citation: Peters, J. C., Op de Beeck, H. P., Goebel, R., eds. (2016). Integrating Computational and Neural Findings in Visual Object Perception. Lausanne: Frontiers Media.
doi: 10.3389/978-2-88919-873-3

Table of Contents

04 ***Editorial: Integrating Computational and Neural Findings in Visual Object Perception***
Judith C. Peters, Hans P. Op de Beeck and Rainer Goebel

07 ***Finding and recognizing objects in natural scenes: complementary computations in the dorsal and ventral visual systems***
Edmund T. Rolls and Tristan J. Webb

26 ***Unsupervised invariance learning of transformation sequences in a model of object recognition yields selectivity for non-accidental properties***
Sarah M. Parker and Thomas Serre

34 ***Diversity priors for learning early visual features***
Hanchen Xiong, Antonio J. Rodríguez-Sánchez, Sandor Szedmak and Justus Piater

45 ***Correlated activity supports efficient cortical processing***
Chou P. Hung, Ding Cui, Yueh-peng Chen, Chia-pei Lin and Matthew R. Levine

61 ***Corrigendum: Correlated activity supports efficient cortical processing***
Chou P. Hung, Ding Cui, Yueh-peng Chen, Chia-pei Lin and Matthew R. Levine

62 ***Applying artificial vision models to human scene understanding***
Elissa M. Aminoff, Mariya Toneva, Abhinav Shrivastava, Xinlei Chen, Ishan Misra, Abhinav Gupta and Michael J. Tarr

76 ***Fourier power, subjective distance, and object categories all provide plausible models of BOLD responses in scene-selective visual areas***
Mark D. Lescroart, Dustin E. Stansbury and Jack L. Gallant

96 ***Optimal attentional modulation of a neural population***
Ali Borji and Laurent Itti

110 ***On the role of spatial phase and phase correlation in vision, illusion, and cognition***
Evgeny Gladilin and Roland Eils

124 ***Aesthetic perception of visual textures: a holistic exploration using texture analysis, psychological experiment, and perception modeling***
Jianli Liu, Edwin Lughofer and Xianyi Zeng

frontiers in Computational Neuroscience

EDITORIAL
published: 20 April 2016
doi: 10.3389/fncom.2016.00036

Editorial: Integrating Computational and Neural Findings in Visual Object Perception

Judith C. Peters [1, 2], Hans P. Op de Beeck [3] and Rainer Goebel [1, 2]*

[1] *Cognitive Neuroscience Department, Faculty of Psychology and Neuroscience, Maastricht University, Maastricht, Netherlands,* [2] *Neuroimaging and Neuromodeling Department, Netherlands Institute for Neuroscience, Amsterdam, Netherlands,* [3] *Laboratory of Biological Psychology, University of Leuven, Leuven, Belgium*

Keywords: object recognition, computer vision, fMRI, feature representation, ventral visual pathway, invariance

The Editorial on the Research Topic

Integrating Computational and Neural Findings in Visual Object Perception

Recognizing objects despite infinite variations in their appearance is a highly challenging computational task the visual system performs in a remarkably fast, accurate, and robust fashion. The complexity of the underlying mechanisms is reflected in the large proportion of cortical real-estate dedicated to visual processing, as well as in the difficulties encountered when trying to build models whose performance matches human proficiency.

The articles in this Research Topic provide an overview of recent advances in our understanding of the neural mechanisms underlying visual object perception, focusing on integrative approaches which encompass both computational and empirical work. Given the vast expanse of topics covered in the discipline of computational visual neuroscience, it is impossible to provide a comprehensive overview of the field's status-quo. Instead, the presented papers highlight interesting extensions to existing models and novel insights into computational principles and their neural underpinnings. Contributions could be coarsely subdivided into three different sections: Two papers focused on implementing biologically-valid learning rules and heuristics in well-established neural models of the visual pathway (i.e., "VisNet" and "HMAX") to improve flexible object recognition. Three other studies investigated the role of sparseness, selectivity, and correlation in optimizing neural coding of object features. Finally, another set of contributions focused on integrating computational vision models and human brain responses to gain more insights in the computational mechanisms underlying neural object representations.

OPEN ACCESS

Edited by:
Si Wu,
Beijing Normal University, China

Reviewed by:
Yuanyuan Mi,
Weizmann Institute of Science, Israel

****Correspondence:***
Judith C. Peters
j.peters@nin.knaw.nl

Received: *08 February 2016*
Accepted: *31 March 2016*
Published: *20 April 2016*

Citation:
Peters J, Op de Beeck H and Goebel R (2016) Editorial: Integrating Computational and Neural Findings in Visual Object Perception. Front. Comput. Neurosci. 10:36. doi: 10.3389/fncom.2016.00036

EXTENDING INVARIANT RECOGNITION CAPABILITIES OF EXISTING MODELS

A key challenge our visual system faces is a trade-off between discrimination and generalization. It should be able to discriminate an encountered object from a myriad of possible alternatives. Yet, it has to generalize across different instances of the same object, or, in other words, be invariant to so-called "identity-preserving transformations" (DiCarlo et al., 2012). Two contributions in this Research Topic propose updates to influential computational models to more adequately deal with the latter invariance constraint.

Rolls and Webb, introduce an extension of the Ventral Visual Stream (VVS) model "VisNet" (Rolls, 2012) by incorporating a bottom-up driven saliency-detection mechanism to locate items of interest in natural scenes. By adding this functionality, their model mimics the "divide-and-conquer" strategy applied by the primate visual system: the dorsal stream uses stimulus saliency to guide saccades, which then allows the VVS to successively process a set of relatively small fixated regions (instead of having to deal with a complex visual scene in its entirety), thereby reducing the computational requirements to achieve invariant object recognition. The presented results show that VisNet could reliably locate and identify a number of objects in cluttered scenes, portraying both view and translation invariance, even though training encompassed only four viewpoints and a limited range of positions per object. These findings further corroborate the notion that learning rules based on temporal continuity (i.e., exploiting the increased likelihood that consecutive retinal images belong to the same object despite slight changes in its appearance) can successfully guide the development of invariant object representations.

Likewise, Parker and Serre show that another prominent model, namely HMAX (Riesenhuber and Poggio, 1999), can be extended to learn invariant recognition across 3D-rotations (while previous instantiations were limited to 2D changes in position and scale) based on unsupervised training on short object transformation sequences. The extended model exhibited greater sensitivity to so-called "Non-Accidental Properties" (akin to infero-temporal cortical responses) and concomitantly demonstrated greater tolerance to object transformations in its input.

EFFICIENT NEURAL CODING STRATEGIES

The selectivity and sparseness observed in neural firing elicited by visual stimulation are generally considered hallmarks of an efficient coding scheme: since a given neuron only responds to a limited set of inputs, and conversely any input only triggers activity in a relatively small fraction of the neural population, redundancy is minimized. In their contribution, Xiong et al. show that both selectivity and sparseness (which need not be correlated) can simultaneously arise as properties of modeled V1 receptive fields by reinforcing diversity (i.e., minimizing similarity by mimicking neural inhibition) during the training of a restricted Boltzmann machine (a type of network routinely used in "deep learning" approaches LeCun et al., 2015).

Interestingly, the findings presented by Hung et al. actually point to a role of *correlated* neural activity in efficient visual recognition as opposed to the proposedly beneficial de-correlation that tuning selectivity might offer. Based on dense neurophysiological recordings in monkey infero-temporal cortex, the authors show that correlation strength and tuning selectivity are only weakly related and that the observed correlated activity is mainly driven by neurons in IT output layers that convey generalizable object information, which is behaviorally relevant as it predicts human visual search performance (see below). Relatedly, Gladilin and Eils discuss the behavioral and neural importance of (phase) correlation in visual input.

LINKING COMPUTATIONAL MODELS TO HUMAN BRAIN RESPONSES

Human neuropsychological and neuroimaging studies have consistently identified brain regions involved in object recognition. Nevertheless, our current understanding of ongoing computations and feature representations within these areas is rather limited.

One way forward to unravel the identified regions' inner workings is to compare the similarity across neural response patterns elicited by a given stimulus set to the similarity in output of a range of computer-vision models (with different feature extractions) when presented with the same stimulus set. Using this exploratory strategy, Aminoff et al. demonstrate that fMRI activation-patterns within scene-selective brain regions, such as the parahippocampal (PPA) and occipital place area (OPA), correlated most strongly with computer-vision models incorporating semantic features. In comparison, correlations were lower for models representing low-level features and for behavioral similarity scores. Conversely, the activation-pattern observed in the retrosplenial complex (RSC) was more in line with one of the low-level models and did correlate with subjective similarity ratings. Although encouraging, the results also clearly indicated that the overall correspondence between empirical and modeled responses was weak, suggesting that we still lack a clear grasp on cortical feature representation. One such feature, visual texture, is further explored in the contribution by Liu et al. using behavioral methods and modeling.

Another approach to gain insights into VVS feature representations is employed by Lescroart et al. They compared how well three encoding models, based on different scene-defining feature classes, could voxel-wise predict neural representations in scene-selective brain regions. The encoding models mapped a diverse set of natural images to three qualitatively different feature spaces: 2D-features related to Fourier-power, the subjective 3D—distance to salient objects in the scene, and a more abstract, semantic scene description ("object-categorization"). In line with Aminoff et al. the object-category model provided a better prediction of PPA and OPA activity compared to the other two encoding models which did not include semantic features. In addition, RSC activity was more accurately predicted by the object-category model than the Fourier-power model, but the object-category model and the 3D-distance model performed equally well. Although, results of both studies suggest a different feature representation for scenes in RSC compared to PPA and OPA, it should be noted that feature representations in all areas are more complex than captured by the applied computer-vision and encoding models. Response variance explained by the models was largely shared in the fMRI data of Lescroart et al. To which extent this reflects an actual combined representation of the model's different feature classes, or alternatively the high correlation between these feature spaces in natural images, could be further explored by follow-up studies using stimulus sets with reduced feature covariance (yet covering enough variance for real-world generalization). Furthermore, such studies might attempt to establish new encoding models based on feature spaces inspired by feature representations in high-level computer-vision models (e.g., Aminoff et al.) or deep neural nets (e.g., Güçlü and Van Gerven, 2015).

However, even the most optimal feature representations based on such approaches currently miss an important ingredient that might be essential for our fast and efficient object recognition: feature representations in the brain are dynamically influenced by task demands. We actively engage in a dynamical world, intentionally searching for and interacting with objects, rather than passively observing static sceneries. Several aforementioned contributions highlight specific aspects of such active perception, and more aspects can be distinguished. For example, to selectively process objects of interest over distracting information, we can use (c)overt spatial attention to constrain computations (see Rolls and Webb), but also non-spatial attention contributes to an efficient read-out of neural representations by altering the corresponding feature space. In particular, during visual search for objects in a movie, fronto-parietal and occipito-temporal activations become tuned toward the attended object-category, expanding representations of this and semantically related categories, at the cost of unattended categories (Çukur et al., 2013). Likewise, the work by Hung et al. revealed that proximity in a neurally defined feature space (based on monkey IT data) predicts human visual search efficiency: targets were more easily identified when subjects were previously adapted to surrounding distractors containing contrastive features represented in neighboring cortical columns. This relates to neural simulations in the contribution of Borji and Itti, suggesting that feature similarity between target and distractors affects whether attention modulates (combinations of) neural gain, shifts in tunings, or sharpening of tunings, to allow for the most informative representations of important stimulus features. Moreover, the employed attentional mechanisms were influenced by task requirements (e.g., object discrimination vs. search), providing a further demonstration of the adaptive nature of feature representations optimized for fast and efficient read-out by higher-level areas. Adding such cognitive top-down influences that warp feature spaces according to salience and relevance, employing vision models with recurrent connections, and defining specific encoding models for each processing stage remains challenging, yet appears necessary for a profound understanding of object representations in the primate brain.

CONCLUDING REMARKS

Combining computational and empirical efforts to reveal the neural mechanisms underlying visual object recognition has recently gained momentum. There has been a vast increase in studies employing encoding models to understand how input, transformed to an abstract feature space, predicts measured neural activity. The variety of models under investigation has expanded, ranging from low-level visual descriptors to models that incorporate high-level semantic features. Moreover, advances in high performance computing made it possible to move beyond predefined sets of features, to feature spaces learned from huge and diverse sets of natural world images using deep-learning techniques. Comparing different feature spaces to neural activity can be performed for each measure unit separately (e.g., for each fMRI voxel, see Lescroart et al.) or features can be compared to activation patterns in pre-localized brain regions using similarity estimates (e.g., Aminoff et al.). Recently, Khaligh-Razavi et al. (2014) showed that integrating both approaches, by reweighting and remixing model features via voxel-wise modeling, can lead to higher similarity between models and neural responses in object-selective visual cortex. Direct integration by projecting (population receptive field) voxel models and measured fMRI data in the same brain space might further facilitate comparisons by enabling the use of identical data analysis and visualization techniques for both modeled and measured data (Peters et al., 2012).

The advent of ultra-high field fMRI imaging, large-scale electrocorticographic grids, and dense electrode arrays will provide increasingly rich datasets to study neural activity-patterns with unprecedented detail, yet with sufficient coverage to track reformatting of feature representations from low- to mid- to high-level areas along the VVS. By capitalizing on these increasing opportunities to integrate advanced computer-vision models and large-scale, high-resolution neural datasets, future research can rely on an ever-expanding data mining toolbox to probe neural feature and object representations to uncover the underlying neural "vocabularies."

AUTHOR CONTRIBUTIONS

JP wrote the paper with assistance and approval from HO and RG.

ACKNOWLEDGMENTS

This work received funding from the European Research Council under grant agreement n° 269853/Human Brain Project grant agreement n° 604102. We thank Joel Reithler for insightful discussions and many useful comments on this Editorial and his invaluable contribution to this Research Topic.

REFERENCES

Çukur, T., Nishimoto, S., Huth, A. G., and Gallant, J. L. (2013). Attention during natural vision warps semantic representation across the human brain. *Nat. Neurosci.* 16, 763–770. doi: 10.1038/nn.3381

DiCarlo, J. J., Zoccolan, D., and Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron* 73, 415–434. doi: 10.1016/j.neuron.2012.01.010

Güçlü, U., and Van Gerven, M. A. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *J. Neurosci.* 35, 10005–10014. doi: 10.1523/JNEUROSCI.5023-14.2015

Khaligh-Razavi, S. M., Henriksson, L., Kay, K., and Kriegeskorte, N. (2014). Explaining the hierarchy of visual representational geometries by remixing of features from many computational vision models. *bioRxiv*. doi: 10.1101/009936

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444. doi: 10.1038/nature14539

Peters, J. C., Reithler, J., and Goebel. R. (2012). Modeling invariant object processing based on tight integration of simulated and empirical data in a Common Brain Space. *Front. Comput. Neurosci.* 6:12. doi: 10.3389/fncom.2012.00012

Riesenhuber, M., and Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nat. Neurosci.* 2, 1019–1025. doi:10.1038/14819

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

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.

frontiers in
COMPUTATIONAL NEUROSCIENCE

ORIGINAL RESEARCH ARTICLE
published: 12 August 2014
doi: 10.3389/fncom.2014.00085

Finding and recognizing objects in natural scenes: complementary computations in the dorsal and ventral visual systems

Edmund T. Rolls [1,2]* ***and Tristan J. Webb*** [1]

[1] *Department of Computer Science, University of Warwick, Coventry, UK*
[2] *Oxford Centre for Computational Neuroscience, Oxford, UK*

Edited by:
Hans P. Op De Beeck, University of Leuven (KU Leuven), Belgium

Reviewed by:
Hans P. Op De Beeck, University of Leuven (KU Leuven), Belgium
Da-Hui Wang, Beijing Normal University, China

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

Searching for and recognizing objects in complex natural scenes is implemented by multiple saccades until the eyes reach within the reduced receptive field sizes of inferior temporal cortex (IT) neurons. We analyze and model how the dorsal and ventral visual streams both contribute to this. Saliency detection in the dorsal visual system including area LIP is modeled by graph-based visual saliency, and allows the eyes to fixate potential objects within several degrees. Visual information at the fixated location subtending approximately 9° corresponding to the receptive fields of IT neurons is then passed through a four layer hierarchical model of the ventral cortical visual system, VisNet. We show that VisNet can be trained using a synaptic modification rule with a short-term memory trace of recent neuronal activity to capture both the required view and translation invariances to allow in the model approximately 90% correct object recognition for 4 objects shown in any view across a range of 135° anywhere in a scene. The model was able to generalize correctly within the four trained views and the 25 trained translations. This approach analyses the principles by which complementary computations in the dorsal and ventral visual cortical streams enable objects to be located and recognized in complex natural scenes.

Keywords: object recognition, invariance, saliency, inferior temporal visual cortex, trace learning rule, VisNet

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 that 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, 2008, 2012), 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, 2008, 2014). Here we consider how the cerebral cortex solves the major computational task of view-invariant recognition of objects in complex natural scenes, still a major challenge for computer vision approaches, as described in the Discussion.

One mechanism that the brain uses to simplify the task of recognizing objects in complex natural scenes is that the receptive fields of inferior temporal cortex neurons change from approximately 70° in diameter when tested under classical neurophysiology conditions with a single stimulus on a blank screen to as little as a radius of 8° (for a 5° stimulus) when tested in a complex natural scene (Rolls et al., 2003; Aggelopoulos and Rolls, 2005) (with consistent findings described by Sheinberg and Logothetis, 2001). This greatly simplifies the task for the object recognition system, for instead of dealing with the whole scene as in traditional computer vision approaches, the brain processes just a small fixated region of a complex natural scene at any one time, and then the eyes are moved to another part of the screen. During visual search for an object in a complex natural scene, the primate visual system, with its high resolution fovea, therefore keeps moving the eyes until they fall within approximately 8° of the target, and then inferior temporal cortex neurons respond to the target object, and an action can be initiated toward the target, for example to obtain a reward (Rolls et al., 2003). The inferior temporal cortex neurons then respond to the object being fixated with view, size, and rotation invariance (Rolls, 2012), and also need some translation invariance, for the eyes may not be fixating the center of the object when the inferior temporal cortex neurons respond (Rolls et al., 2003).

The questions then arise of how the eyes are guided in a complex natural scene to fixate close to what may be an object; and how close the fixation is to the center of typical objects for this determines how much translation invariance needs to be built into the ventral visual system. It turns out that the dorsal visual system (Ungerleider and Mishkin, 1982; Ungerleider and Haxby, 1994) implements bottom-up saliency mechanisms by guiding saccades to salient stimuli, using properties of the stimulus such as high contrast, color, and visual motion (Miller and Buschman, 2013). (Bottom-up refers to inputs reaching the visual system from the retina). One particular region, the lateral intraparietal cortex (LIP), which is an area in the dorsal visual system, seems to contain saliency maps sensitive to strong sensory inputs (Arcizet et al., 2011). Highly salient, briefly flashed, stimuli capture both behavior and the response of LIP neurons (Bisley and Goldberg, 2003, 2006; Goldberg et al., 2006). Inputs reach LIP via dorsal visual stream areas including area MT, and via V4 in the ventral stream (Soltani and Koch, 2010; Miller and Buschman, 2013). Although top-down attention using biased competition can facilitate the operation of attentional mechanisms, and is a subject of great interest (Desimone and Duncan, 1995; Rolls and Deco, 2002; Deco and Rolls, 2005a; Miller and Buschman, 2013), top-down object-based attention makes only a small contribution to visual search for an object in a complex natural unstructured scene (such as leaves on a tree), increasing the receptive field size from a radius of approximately 7.8 to approximately 9.6° (Rolls et al., 2003), and is not considered further here. Indeed, in these investigations, multiple saccades were required round the scene to find a target object (Rolls et al., 2003).

In the research described here we investigate computationally how a bottom-up saliency mechanism in the dorsal visual stream reaching for example area LIP could operate in conjunction with invariant object recognition performed by the ventral visual stream reaching the inferior temporal visual cortex to provide for invariant object recognition in natural scenes. The hypothesis is that the dorsal visual stream, in conjunction with structures such as the superior colliculus (Knudsen, 2011), uses saliency to guide saccadic eye movements to salient stimuli in large parts of the visual field, and that once a stimulus has been fixated, the ventral visual stream performs invariant object recognition on the region being fixated. The dorsal visual stream in this process knows little about invariant object recognition, so cannot identify objects in natural scenes. Similarly, the ventral visual stream cannot perform the whole process, for it cannot efficiently find possible objects in a large natural scene, because its receptive fields are only approximately 9° in radius in complex natural scenes. It is how the dorsal and ventral streams work together to implement invariant object recognition in natural scenes that we investigate here. By investigating this computationally, we are able to test whether the dorsal visual stream can find objects with sufficient accuracy to enable the ventral visual stream to perform the invariant object recognition. The issue here is that the ventral visual stream has in practice some translation invariance in natural scenes, but this is limited to approximately 9° (Rolls et al., 2003; Aggelopoulos and Rolls, 2005). The computational reason why the ventral visual stream does not compute translation invariant representations over the whole visual field as well as view, size and rotation invariance, is that the computation is too complex. Indeed, it is a problem that has not been fully solved in computer vision systems when they try to perform invariant object recognition over a large natural scene. The brain takes a different approach, of simplifying the problem by fixating on one part of the scene at a time, and solving the somewhat easier problem of invariant representations within a region of approximately 9°.

For this scenario to operate, the ventral visual stream needs then to implement view invariant recognition, but to combine it with some translation invariance, as the fixation position produced by bottom up saliency will not be at the center of an object, and indeed may be considerably displaced from the center of an object. In the model of invariant visual object recognition that we have developed, VisNet, which models the hierarchy of visual areas in the ventral visual stream by using competitive learning to develop feature conjunctions supplemented by a temporal trace or by spatial continuity or both, all previous investigations have explored either view or translation invariance learning, but not both (Rolls, 2012). Combining translation and view invariance learning is a considerable challenge, for the number of transforms becomes the product of the numbers of each transform type, and it is not known how VisNet (or any other biologically plausible approach to invariant object recognition) will perform with the large number, and with the two types of transform combined. Indeed, an important part of the research described here was to investigate how well architectures of the VisNet type generalize between both trained locations and trained views. This is important for setting the numbers of different views and translations of each object that must be trained.

The specific goals of the research and simulations described here were as follows. (1) To demonstrate with a biologically plausible model of the ventral visual system how it could operate to implement view invariant object/person identity recognition with a generic model of the dorsal visual system that produced fixations on parts of scenes that were salient. How would the combined cortical visual areas operate with the dorsal visual system not encoding object identity but only saliency; and the ventral visual system being unable to find objects efficiently in large natural scenes, but able to perform view invariant object recognition once fixation was close to an object? (2) How closely and effectively would a simple, generic, bottom-up saliency system modeling part of the functions of the dorsal visual system find objects in a complex scene, and how accurately would the center of the object be fixated? The accuracy with which the center of the object is fixated is crucial to understand, for this defines how much translation invariance must be incorporated into the ventral visual system for the whole system to work. (3) Can VisNet be trained for both view and translation invariance? This has not been attempted previously with VisNet, and for that matter view invariant object recognition is not a property of most computer vision models (see Discussion). (4) If VisNet can be trained on both view and translation invariant object identification, can it be trained with sufficient translation invariance to cover the visual angle needed given the inaccuracies of the saliency-based fixation mechanism in finding the center of an object, and yet be trained with sufficient views to provide for view-invariant object identification? (5) How well does VisNet generalize from trained views to untrained views of an object? This is important, for it influences how much training of different views is required, which could have an impact on the capacity of the system, that is on the number of objects or people that it can correctly identify with the required translation invariance. (6) How well does VisNet perform in object identification when the objects appear in natural scenes with fixation not necessarily at the trained location, and when views intermediate to those at which VisNet has been trained are presented? That is, how well under the natural scene conditions can VisNet ignore the background and identify a trained object despite it being presented in a view and position that were not trained?

2. METHODS

2.1. SALIENCY

We chose a bottom up saliency algorithm that is one of the standard ones that has been developed, which adopts the Itti and Koch (2000) approach to visual saliency, and implements it by graph-based visual saliency (GBVS) algorithms (Harel et al., 2006a,b). This system performs well, that is similarly to humans, in many bottom-up saliency tasks. The particular algorithm used for the bottom-up saliency was not crucial to the present research, so we chose a generically representative algorithm[1]. We used static images, so motion was not used to detect saliency. Of course in the human brain, and in a computer application, performance could be made better than described here by using many different cues that can influence saliency, including also color which was disabled in the current algorithm, as VisNet works with grayscale images to help ensure that object shape is being processed, and not a simple feature such as color (Rolls, 2012).

2.2. ARCHITECTURE OF THE VENTRAL VISUAL STREAM MODEL, VisNet

The architecture of VisNet has been described previously (Rolls, 2008, 2012), and is summarized briefly next, with a full description provided in the Appendix. Extensions important for the present research included training in both view and translation invariance, together with careful specification of the learning rate during the presentation of each transform, as there were typically 100 or more transforms of every object to be learned.

Fundamental elements of Rolls' 1992 theory for how cortical networks might implement invariant object recognition are described in detail elsewhere (Rolls, 2008, 2012). They provide the basis for the design of VisNet, which 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 using competitive learning (Rolls, 2008), ensuring that higher order spatial properties of the input stimuli are represented in the network. In VisNet, layer 1 corresponds to V2, layer 2 to V4, layer 3 to posterior inferior temporal visual cortex, and layer 4 to anterior inferior temporal cortex. Layer one is preceded by a simulation of the Gabor-like receptive fields of V1 neurons produced by each image presented to VisNet (Rolls, 2012).
- A convergent series of connections from a localized population of neurons in the preceding layer to each neuron of the following layer, thus allowing the receptive field size of neurons to increase through the visual processing areas or layers, as illustrated in **Figure 1**.
- A modified associative (Hebb-like) learning rule incorporating a temporal trace of each neuron's previous activity, which, it has been shown (Földiák, 1991; Rolls, 1992; Wallis et al., 1993; Wallis and Rolls, 1997; Rolls and Milward, 2000; Rolls, 2012), enables the neurons to learn transform invariances.

The learning rates for each of the four layers were 0.05, 0.03, 0.005, and 0.005, as these rates were shown to produce convergence of the synaptic weights after 15–50 training epochs. 50 training epochs were run.

The developments to VisNet that facilitated this principled approach to the learning rate, combined view and translation invariance learning, etc, and the parameters used, are described in the Appendix.

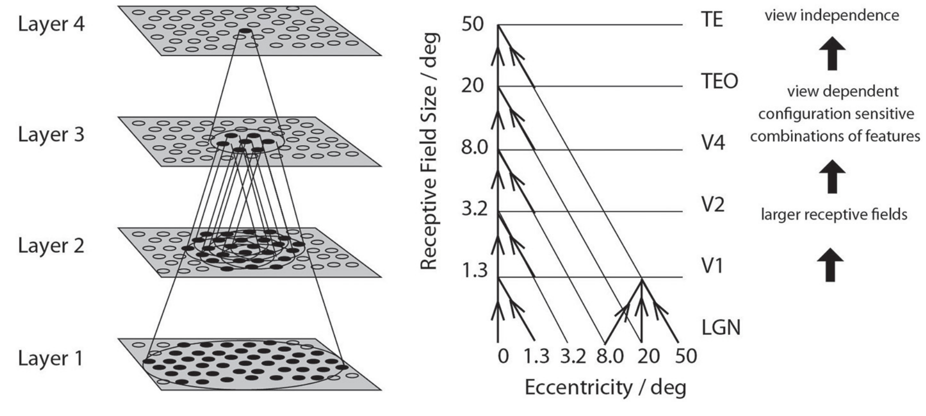


FIGURE 1 | Convergence in the visual system. Right: As it occurs in the brain. V1, visual cortex area V1; TEO, posterior inferior temporal cortex; TE, inferior temporal cortex (IT). **Left:** As implemented in VisNet. Convergence through the network is designed to provide fourth layer neurons with information from across the entire input retina.

[1] GBVS was used with its default parameters, except as follows: channels = CIO; gaborangles 0, 30, 60, 90, 120, 150; onCenterBias = 1; levels 2 3; sigma_frac_act = 0.35; sigma_frac_norm = 0.26.

2.3. INFORMATION MEASURES OF PERFORMANCE

The performance of VisNet was measured by Shannon information-theoretic measures that are essentially identical to those used to quantify the specificity and selectiveness of the representations provided by neurons in the brain (Rolls and Milward, 2000; Rolls and Treves, 2011; Rolls, 2012). A single cell information measure indicated how much information was conveyed by a single neuron about the most effective stimulus. A multiple cell information measure indicated how much information about every stimulus was conveyed by small populations of neurons, and was used to ensure that all stimuli had some neurons conveying information about them. Details are provided in the Appendix.

2.4. TRAINING

VisNet was trained on four views spaced 45° apart of each of the 4 objects as illustrated in **Figure 2**. The images of each object were generated from a 3D model using Blender (The Blender Foundation, www.blender.org) so that lighting could be carefully controlled. Each grayscale image of an object was 256 × 256 pixels, with the intensity scaled to be in the range 0–255, and the background approximately 127. The object images were pasted into a 512 × 512 gray image to prevent wrap-around effects, prior to the spatial frequency filtering to produce neurons with Gabor-like receptive fields in an emulation of V1 neurons that provided the input to the first layer of VisNet (see Appendix). [We have previously shown that the training need not be on a blank background, provided that the background is not constant across transforms and objects, as will be the case in the natural world (Stringer et al., 2007; Stringer and Rolls, 2008)]

Each training image was trained in 25 locations set out in a 5 ×5 rectangular grid with these locations separated by 8 pixels in the training image. To provide an indication of the range of this translation invariance training, the grid extended between the centers of the headlights in the front view of the jeep shown in **Figure 2**. This resulted in 100 transforms of each object to be learned. To enable VisNet to learn invariant representations with the trace synaptic learning rule, all the transforms of one object were shown in a random permuted sequence, the trace was reset, and the procedure was repeated with each of the other objects. 50 training epochs were run, as this was sufficient to produce gradual convergence of the synaptic weights over 15–50 epochs, as described in the Appendix.

2.5. TESTING INVARIANT OBJECT RECOGNITION IN NATURAL SCENES

Eight of the 12 test scenes are illustrated in **Figure 3A**. Each scene had each of the objects in one of the four poses. The aim of the combined visual processing was for the dorsal visual stream to detect the salient regions in these 12 scenes, and then for the salient regions to be passed to VisNet to perform the view (and translation) invariant object recognition for every object in the scene. VisNet had been trained on the 4 objects in each of the 4 views, but not on the background scenes, and it was part of the task of VisNet to identify each of the four objects in every scene without being affected by the background clutter of each scene (Stringer and Rolls, 2000). The objects used in this

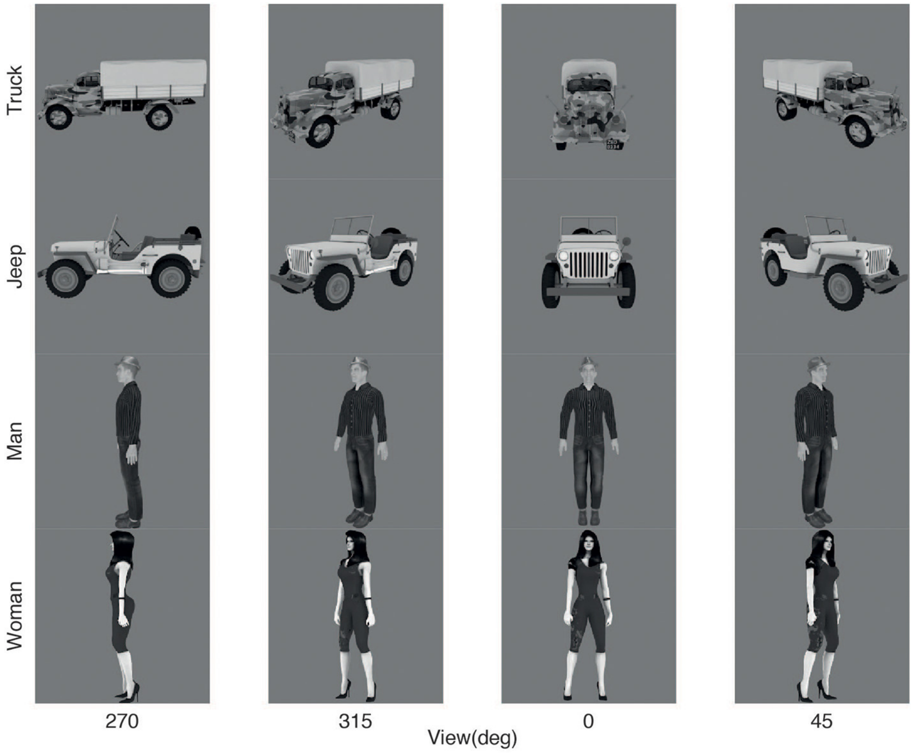


FIGURE 2 | Training images: 4 views of each of 4 objects. Each image was 256 × 256 pixels.

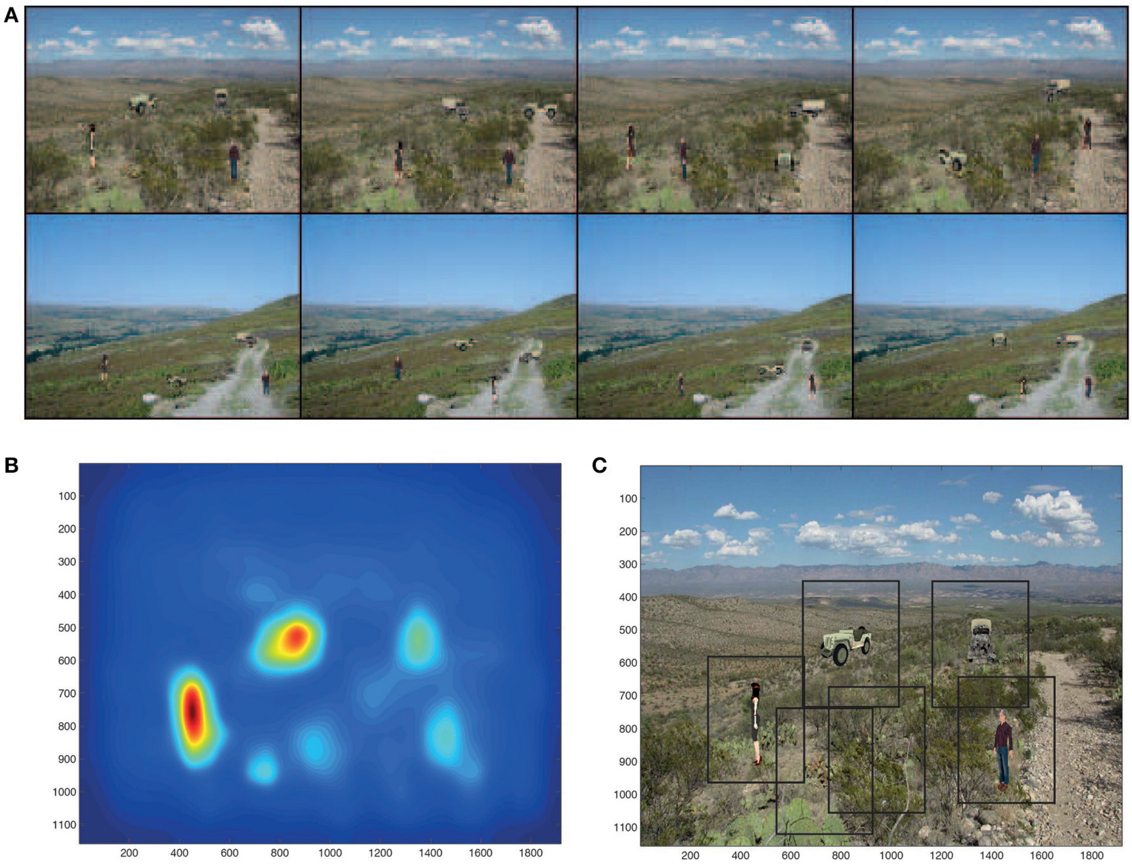


FIGURE 3 | (A) Eight of the 12 test scenes. Each scene has 4 objects, each in one of its four views. **(B)** The bottom up saliency map generated by the GBVS code for one of the scenes. The highest levels in the saliency map are red, and the lowest blue. **(C)** Rectangles (384 × 384 pixels) placed around each peak in the scene for which the bottom-up saliency map is illustrated in **(B)**.

investigation were common types of object with which the human visual system performs good view invariant identification, people and vehicles. Two people and two vehicles were chosen to provide evidence on how the system might operate with typical stimuli for which view-invariant identification is necessary and is performed by the human visual system.

3. RESULTS

3.1. THE OPERATION OF THE SALIENCY PROCESSING

The bottom up saliency map generated by the GBVS code (acting as a surrogate for the dorsal visual system) for one of the scenes is illustrated in **Figure 3B**. The saliency map has of course no indication of which peak is a trained object, nor of which object it might be.

The saliency maps generated by GBVS correspond closely to the saccades and resulting fixations of humans (Itti and Koch, 2000; Harel et al., 2006a,b). We therefore extracted images from the scene that were at the center of each peak of the saliency map. A weighted centroid was used, as implemented in MATLAB. Each extracted image centered on a peak in the saliency map was 384 × 384 pixels (not the originally trained 256 × 256 size of a training image), because sometimes a saliency peak was not well centered on an object, and we wished to be sure that the whole object was in the image presented to VisNet. **Figure 3C** shows rectangles produced in this way round the 6 most salient regions in the test scene for which the saliency map is shown in **Figure 3B**. Four of the saliency peaks and therefore the rectangles contained trained objects, and two extracted images just salient parts of the background scene in which the trained objects appeared.

The extracted ("foveated") images of the objects to be presented to VisNet based on saliency are not always well-centered in the 384 × 384 extracted image, and this is clear for one of the objects, the man, as shown in **Figure 3C**.

To provide evidence on the degree of translation invariance that would be required of VisNet given that the center of each image was not always at the peak of the saliency map, so that the extracted image would be offset from a central trained location, the offsets of the saliency peaks from the center of each object image are shown in **Figure 4**. While it is clear that the majority of the offsets of the saliency peak from the center of the object were in the range 0–32 pixels, some were beyond this. For this reason, we do not necessarily expect that VisNet, trained on a grid with an offset up to 32 would achieve 100% correct object recognition. The evidence shown in **Figure 4** does provide though the useful indication that training to allow for offsets up to 64 for a 256 × 256 image might improve performance.

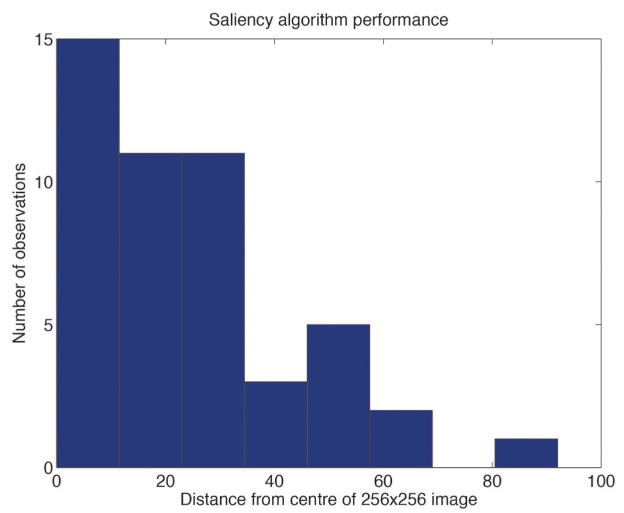


FIGURE 4 | Distribution of the offsets of the saliency peaks from the center of each object. The data were obtained for 48 images (different views of the different objects) presented in 3 backgrounds. An example of one of the backgrounds containing one view of each of four objects is illustrated in **Figure 3C**.

3.2. TESTS OF VisNet ON VIEW AND TRANSLATION INVARIANCE

Although VisNet had been trained on a 25-location grid with size 64 × 64 with spacing of 16 pixels, and with 4 different views of each object, we did not know how well VisNet would perform on this task as this has never been tested before, nor whether performance would generalize to intermediate locations in the 64 × 64 grid, given that there were only 25 training locations spaced 16 pixels apart. An analysis is shown in **Figure 5A** which covers the 4096 locations in the 64 × 64 grid. This indicates that the performance (on the view invariant object recognition) peaks at the trained locations (0, 16, and 32 in this Figure), but also that there is reasonable performance at intermediate locations between the training locations. (The chance performance with 4 objects is 25% correct.) This is an important new result, which adds to previous evidence that smaller versions of VisNet with 32 × 32 neurons in each of 4 layers can generalize reasonably across intermediate untrained locations in scenes with blank backgrounds (Wallis and Rolls, 1997). The performance was measured with a pattern associator trained on layer 4 of VisNet, with four output neurons (one for each object), and the 25 most selective cells for each object identified using the single cell information measure (see Appendix). The best cells were quite selective for one of the objects, and quite invariant in their response over the 100 transforms (4 views and 25 locations), as illustrated in **Figure 5B**.

3.3. TESTS OF THE WHOLE SALIENCY PLUS VIEW INVARIANCE SYSTEM

With 48 images extracted from the the 12 test scenes (8 illustrated in **Figure 3A**), performance was 90% correct (43 correct/48), where chance with the four objects is 25% (Fisher test $p \ll 0.0001$).

It is important that this good performance on this identification task was found when the images extracted for presentation to VisNet had background parts of the scene included (e.g., **Figure 3C**). These background features did not produce large decreases in the performance of VisNet, given that VisNet had been trained on the objects but not on the backgrounds (Stringer and Rolls, 2000). This is important for the processes of invariant visual object identification in novel complex natural scenes described here. Further, if there was a low amplitude saliency peak containing only part of the background scene and not an object, then VisNet did not respond to this as a trained object. When

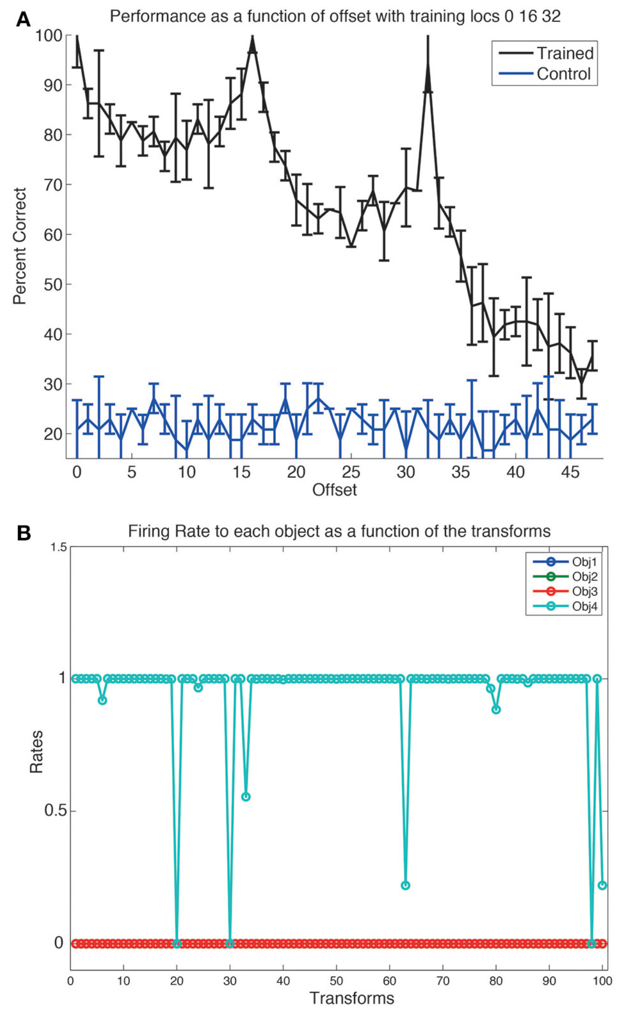


FIGURE 5 | (A) The performance on the view invariant object recognition tested with images at the 15 trained locations on the 64 × 64 training grid, and at intermediate locations. The ordinate shows the distance from the central line in the training grid, and trained locations thus correspond to offsets of 0, 16, and 32. The mean and standard deviation are shown for each data point. The standard deviation was measured by performing the training ten times each with a different random seed to generate the connectivity of VisNet. Performance decreases beyond an offset of 32, because there was no translation invariant training beyond this. **(B)** A neuron in layer 4 of VisNet that responded to almost all transforms of one object (4), and to no transform of any other object (1–3). There were 25 location transforms on a grid of size 64 with a spacing of 16, and 4 views of each object at each location. The stimulus-specific information or surprise was 2 bits, as there were 4 objects.

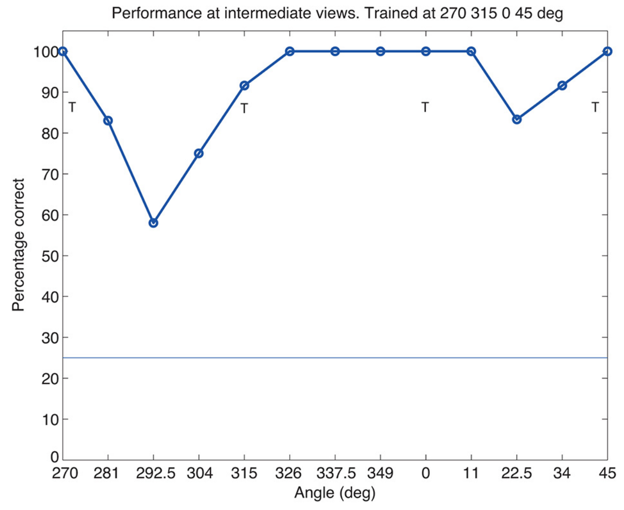


FIGURE 6 | Performance of VisNet at views intermediate to the trained views of 270, 315, 0, and 45°, which are indicated by T. Performance was tested at 6 intermediate views between each trained view, and then for illustrative purposes the results for the 6 intermediate views were averaged using adjacent views. Each data point shown is the average of 12 observations. The chance level of performance, 25%, is indicated.

errors were made by VisNet on the object identification, the confusions were as frequent between the classes of people and vehicle as within these classes.

3.4. TESTS OF VIEW PLUS TRANSLATION INVARIANCE AT INTERMEDIATE VIEWS

The training images had four views of each object separated by 45° as illustrated in **Figure 2**. To assess whether these views were sufficiently close to allow for generalization between the trained views, we tested VisNet with 6 intermediate views (presented on plain backgrounds) between each trained view. As shown in **Figure 6**, performance is reasonable at the untrained intermediate views. The important implication is that VisNet does not need to be trained on a large set of closely spaced views, and this helps the rapid learning of new objects, and also may help to increase the capacity of VisNet, as only few views of each new object need to be learned.

4. DISCUSSION

By combining in a simulation the operation of the dorsal and ventral visual systems in the identification of objects in complex natural scenes, we believe that important progress has been made, in a biologically inspired approach not attempted in other including computer-based approaches. The models simulated show how the brain may solve this major computational problem by moving the eyes to fixate close to objects in a natural scene using bottom-up saliency implemented in the dorsal visual system, and then performs objects recognition successively for each of the fixated regions using the ventral visual system. The research described here emphasizes that because the eyes do not locate the center of objects based on saliency, then translation invariance as well as view, size etc invariance needs to be implemented in the ventral visual system. We show how a model of invariant object recognition in the ventral visual system, VisNet, can perform the required combination of translation and view invariant recognition, and moreover can generalize between views of objects that are 45° apart during training, and can also generalize to intermediate locations when trained in a coarse training grid with the spacing between trained locations equivalent to 1–3°.

We emphasize that the model is closely linked to neurophysiological research on visual object recognition in natural scenes, and explicitly models how the system could operate computationally to achieve the degree of translation invariance shown in complex natural scenes by inferior temporal cortex neurons (Rolls et al., 2003; Aggelopoulos and Rolls, 2005) as well as the view invariance that is combined with this (Hasselmo et al., 1989; Booth and Rolls, 1998). Moreover, the deformation or pose invariance that can be shown by inferior temporal cortex neurons is also a property that can be learned by this functional architectural computational model of object recognition in the ventral visual system, VisNet (Webb and Rolls, 2014).

We note that in the underlying neurophysiological experiments, the objects were small and were presented in an unstructured scene, which was the leaves of trees (Rolls et al., 2003). In this type of scene, objects can only be found by repeated saccades round the scene until the eyes become sufficiently close for the object to fall within the inferior temporal visual cortex neuronal receptive fields which become dynamically reduced to a few degrees in such scenes (Rolls et al., 2003). The receptive fields of inferior temporal cortex neurons are thus small, a few degrees, in complex natural scenes (Rolls et al., 2003; Aggelopoulos and Rolls, 2005). In previous research, sometimes large receptive fields have been reported (Gross et al., 1969), and sometimes small, a few degrees (Op de Beeck and Vogels, 2000; DiCarlo and Maunsell, 2003). We showed that an important factor in the receptive field size is the background. If the receptive fields are measured as in traditional visual neurophysiology against a blank background, then the receptive fields can be as large as 70°, whereas in a complex cluttered natural scene the receptive fields can be as small as a few degrees (Rolls et al., 2003). Moreover, we went on to show that the underlying dynamical mechanism for receptive field size adjustment is probably competition between neurons operating with neurons that have more input from objects close to the fovea (Trappenberg et al., 2002). If objects can be recognized by humans rapidly without the need for multiple fixations round the scene (Thorpe, 2009), then one has to assume that the scene has properties including probably some structure or contrast or color or other low-level feature (Crouzet and Thorpe, 2011), that enables the object to pop out using lower-level processing that does not engage the invariant representations provided by inferior temporal cortex neurons (Rolls, 2012).

The operation of VisNet coupled with the saliency model of the dorsal visual system described here for the identification of multiple objects at different positions in a natural scene with view invariance is now compared with that of other systems and approaches. First, VisNet provides a theory and model of how object identification with view (Stringer and Rolls, 2002), size (Wallis and Rolls, 1997), isomorphic rotation, translation (Stringer and Rolls, 2000; Perry et al., 2010), contrast, illumination (Rolls and Stringer, 2006), and spatial frequency invariance is performed in the cerebral cortex (Rolls, 2012). The approach is addressing fundamental issues about how the cerebral cortex functions. VisNet models four stages of visual processing beyond V1, and simulates V1; it uses local, biologically plausible, synaptic learning rules; it produces neurons in its layer 4 that are comparable to neurons recorded in the inferior temporal visual cortex (IT) (Rolls and Treves, 2011; Rolls, 2012) in terms of their receptive fields and how they are influenced by multiple items in a scene and by top-down attention (Trappenberg et al., 2002; Rolls et al., 2003); in terms of the neuronal tuning to different objects (though VisNet has somewhat more binary neurons that IT neurons) (Rolls, 2008, 2012; Rolls and Treves, 2011); and in terms of size, view, translation, spatial frequency, and contrast invariance (Rolls, 2012). We know of no other biologically plausible model that performs view invariant as well as other types of transform invariant object identification, and that can do this with multiple different objects in complex natural scenes, as demonstrated here.

We provide now (following a suggestion) an account of how VisNet is able to solve the type of invariant object recognition problem described here when an image is presented to it, with more detailed accounts available elsewhere (Wallis and Rolls, 1997; Rolls, 2008, 2012). VisNet is a 4-layer network with feedforward convergence from stage to stage that enables the small receptive fields present in its V1-like Gabor filter inputs of approximately 1° to increase in size so that by the fourth layer a single neuron can potentially receive input from all parts of the input space (**Figure 1**). The feedforward connections between layers are trained by competitive learning, which is an unsupervised form of learning (Rolls, 2008), that allows neurons to learn to respond to feature combinations. As one proceeds up though the hierarchy, the feature combinations become combinations of feature combinations (see Rolls, 2008 Figure 4.20 and Elliffe et al., 2002). Local lateral inhibition within each layer allows each local area within a layer to respond to and learn whatever is present in that local region independently of how much information and contrast there may be in other parts of a layer, and this, together with the non-linear activation function of the neurons, enables a sparse distributed representation to be produced. In the sparse distributed representation, a small proportion of neurons is active at a high rate for the input being presented, and most of the neurons are close to their spontaneous rate, and this makes the neurons of VisNet (Rolls, 2008, 2012) very similar to those recorded in the visual system (Rolls, 2008; Rolls and Treves, 2011). A key property of VisNet is the way that it learns whatever can be learned at every stage of the network that is invariant as an image transforms in the natural world, using the temporal trace learning rule. This learning rule enables the firing from the preceding few items to be maintained, and given the temporal statistics of visual inputs, these inputs are likely to be from the same object. (Typically primates including humans look at one object for a short period during which it may transform by translation, size, isomorphic rotation, and/or view, and all these types of transform can therefore be learned by VisNet.) Effectively, VisNet uses as a teacher the temporal and spatial continuity of objects as they transform in the world to learn invariant representations. (An interesting example is that representations of individual people or objects invariant with respect to pose (e.g., standing, sitting, walking) can be learned by VisNet, or representations of pose invariant with respect to the individual person or object can be learned by VisNet depending on the order in which the identical images are presented during training Webb and Rolls, 2014.) Indeed, we developed these hypotheses (Rolls, 1992, 1995, 2012; Wallis et al., 1993) into a model of the ventral visual system that can account for translation, size, view, lighting, and rotation invariance (Wallis and Rolls, 1997; Rolls and Milward, 2000; Stringer and Rolls, 2000, 2002, 2008; Rolls and Stringer, 2001, 2006, 2007; Elliffe et al., 2002; Perry et al., 2006, 2010; Stringer et al., 2006, 2007; Rolls, 2008, 2012). Consistent with the hypothesis, we have demonstrated these types of invariance (and spatial frequency invariance) in the responses of neurons in the macaque inferior temporal visual cortex (Rolls et al., 1985, 1987, 2003; Rolls and Baylis, 1986; Hasselmo et al., 1989; Tovee et al., 1994; Booth and Rolls, 1998). Moreover, we have tested the hypothesis by placing small 3D objects in the macaque's home environment, and showing that in the absence of any specific rewards being delivered, this type of visual experience in which objects can be seen from different views as they transform continuously in time to reveal different views leads to single neurons in the inferior temporal visual cortex that respond to individual objects from any one of several different views, demonstrating the development of view-invariance learning (Booth and Rolls, 1998). (In control experiments, view invariant representations were not found for objects that had not been viewed in this way.) The learning shown by neurons in the inferior temporal visual cortex can take just a small number of trials (Rolls et al., 1989). The finding that temporal contiguity in the absence of reward is sufficient to lead to view invariant object representations in the inferior temporal visual cortex has been confirmed (Li and DiCarlo, 2008, 2010, 2012). The importance of temporal continuity in learning invariant representations has also been demonstrated in human psychophysics experiments (Perry et al., 2006; Wallis, 2013). Some other simulation models are also adopting the use of temporal continuity as a guiding principle for developing invariant representations by learning (Wiskott and Sejnowski, 2002; Wiskott, 2003; Wyss et al., 2006; Franzius et al., 2007), and the temporal trace learning principle has also been applied recently (Isik et al., 2012) to HMAX (Riesenhuber and Poggio, 2000; Serre et al., 2007c).

We now compare this VisNet approach to invariant object recognition to some other approaches that seek to be biologically plausible. One such approach is HMAX (Riesenhuber and Poggio, 2000; Serre et al., 2007a,b,c; Mutch and Lowe, 2008), which is a hierarchical feedforward network with alternating simple cell-like (S) and complex cell-like (C) layers. The simple cell-like layers respond to a similarity function of the firing rates of the input neuron to the synaptic weights of the receiving neuron (used as an alternative to the more usual dot product), and the complex cells to the maximum input that they receive from a particular class of simple cell in the preceding layer. The classes of simple cell are set to respond maximally to a random patch of a training image (by presenting the image, and setting the synaptic weights of the S cells to be the firing rates of the cells from it receives), and are propagated laterally, that is there are exact copies throughout a layer, which is of course a non-local operation and not biologically plausible. The hierarchy receives inputs from Gabor-like filters (which is like VisNet). The result of this in HMAX is that in the hierarchy there is no learning of invariant representations of objects; and that the output firing in the final C layer (for example the second C layer in a four-layer S1-C1-S2-C2 hierarchy) is high for almost all neurons to most stimuli, with almost no invariance represented in the output layer of the hierarchy, in that two different views of the same object may be as different as a view of another object, measured using the responses of a single neuron or of all the neurons (Robinson and Rolls, 2014). The neurons in the output C layer are thus quite unlike those in VisNet or in the inferior temporal cortex, where there is a sparse distributed representation, and where single cells convey much information in their firing rates, and populations of single cells convey much information that can be decoded by biologically plausible dot product decoding such as might be performed by a pattern association network in the areas that receive from the inferior temporal visual cortex, such as the orbitofrontal cortex and amygdala (Rolls, 2008, 2012; Rolls and Treves, 2011). HMAX therefore must resort to a very powerful classification algorithm, in practice typically a Support Vector Machine (SVM), which is not biologically plausible, to learn to classify all the outputs of the final layer that are produced by the different transforms of one object to be of the same object, and different to those of other objects. Thus HMAX does not learn invariant representations by its output layer of the S–C hierarchy, but instead uses a SVM to perform the classification that the SVM is taught. This is completely unlike the output of VisNet and of inferior temporal cortex neuron firing, which by responding very similarly in terms of firing rate to the different transforms of an object show that the invariance has been learned in the hierarchy (Rolls, 2008, 2012). Another way that the output of HMAX may be assessed is by the use of View-Tuned Units (VTUs), each of which is set to respond to one view of a class or object by setting its synaptic weights from each C unit to the value of the firing of the C unit to one view or exemplar of the object or class (Serre et al., 2007b). Because there is little invariance in the C units, many different VTUs are needed, with one for each training view or exemplar. Because the VTUs are different to each other for the different views of the same object or class, a further stage of training is then needed to classify the VTUs into object classes, and the type of learning is least squares error minimization (Serre et al., 2007b), equivalent to a delta-rule one-layer perceptron which again is not biologically plausible for neocortex (Rolls, 2008). Thus HMAX does not generate invariant representations in its S–C hierarchy, and in the VTU approach uses two layers of learning after the S–C hierarchy, the second involving least squares learning, to produce classification. This is unlike VisNet, which learns invariant representations in its hierarchy, and produces view invariant neurons (similar to those for faces (Hasselmo et al., 1989) and objects (Booth and Rolls, 1998) in the inferior temporal visual cortex) that can be read by a biologically plausible pattern associator (Rolls, 2008, 2012).

Another difference of HMAX from VisNet is in the way that VisNet is trained, which is a fundamental aspect of the VisNet approach. HMAX has traditionally been tested with benchmarking databases such as the CalTech-101 and CalTech-256 (Griffin et al., 2007) in which sets of images from different categories are to be classified. The Caltech-256 dataset 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. A network is supposed to classify these correctly into classes such as hats and bears (Rolls, 2012; Robinson and Rolls, 2014). The problem is that examples of each class of object transforming continuously though different positions on the retina, size, isomorphic rotation, and view are not provided to help the system learn about how a given type of object transforms in the world. The system just has to try to classify based on a set of often quite different exemplars that are not transforms of each other. Thus a system trained in this way is greatly hindered in generating transform invariant representations by the end of the hierarchy, and such a system has to rely on a powerful classifier such as a SVM to perform a classification that is not based on transform invariance learned in the hierarchical network. In contrast, VisNet is provided during training with systematic transforms of objects of the type that would be seen as objects transform in the world, and has a well-posed basis for learning invariant representations. It is important that with VisNet, the early layers may learn what types of transform can be produced in small parts of the visual field by different classes of object, so that when a new class of object is introduced, rapid learning in the last layer and generalization to untrained views can occur without the need for further training of the early layers (Stringer and Rolls, 2002).

Some other approaches to biologically plausible invariant object recognition are being developed with hierarchies that may be allowed unsupervised learning (Pinto et al., 2009; DiCarlo et al., 2012; Yamins et al., 2014). For example, a hierarchical network has been trained with unsupervised learning, and with many transforms of each object to help the system to learn invariant representations in an analogous way to that in which VisNet is trained, but the details of the network architecture are selected by finding parameter values for the specification of the network structure that produce good results on a benchmark classification task (Pinto et al., 2009). However, formally these are convolutional networks, so that the neuronal filters for one local region are replicated over the whole of visual space, which is computationally efficient but biologically implausible. Further, a general linear model is used to decode the firing in the output level of the model to assess performance, so it is not clear whether the firing rate representations of objects in the output layer of the model are very similar to that of the inferior temporal visual cortex. In contrast, with VisNet (Rolls and Milward, 2000; Rolls, 2012) the information measurement procedures that we use (Rolls et al., 1997a,b) are the same as those used to measure the representation that is present in the inferior temporal visual cortex (Tovee et al., 1993; Rolls and Tovee, 1995; Tovee and Rolls, 1995; Abbott et al., 1996; Baddeley et al., 1997; Rolls et al., 1997a,b, 2004, 2006; Panzeri et al., 1999; Treves et al., 1999; Franco et al., 2004, 2007; Aggelopoulos et al., 2005; Rolls and Treves, 2011).

We turn next to compare the operation of VisNet, as a model of cerebral cortical mechanisms involved in view-invariant object identification, with artificial, computer vision, approaches to object identification. However, we do emphasize that our aim in the present research is to investigate how the cerebral cortex operates in vision, not how computer vision attempts to solve similar problems. Within computer vision, we note that many approaches start with using independent component analysis (ICA) (Kanan, 2013), sparse coding (Kanan and Cottrell, 2010), and other mathematical approaches (Larochelle and Hinton, 2010) to derive what may be suitable "feature analyzers," which are frequently compared to the responses of V1 neurons. Computer vision approaches to object identification then may take combinations of these feature analyzers, and perform statistical analyses using computer-based algorithms that are not biologically plausible such as Restricted Boltzmann Machines (RBMs) on these primitives to statistically discriminate different objects (Larochelle and Hinton, 2010). Such a system does not learn view invariant object recognition, for the different views of an object may have completely different statistics of the visual primitives, yet are the different views of the same object. (Examples might include frontal and profile views of faces, which are well tolerated for individual recognition by some inferior temporal cortex neurons (Hasselmo et al., 1989); very different views of 3D object which are identified correctly as the same object by IT neurons after visual experience with the objects to allow for view-invariant learning (Booth and Rolls, 1998); and many man-made tools and objects which may appear quite different in 2D image properties from different views.) Part of the difficulty of computer vision lay in attempts to parse a whole scene at one time (Marr, 1982). However, the biological approach is to place the fovea on one part of a scene, perform image analysis/object identification there, and then move the eyes to fixate a different location in a scene (Trappenberg et al., 2002; Rolls et al., 2003). This is a divide-and-conquer strategy used by the real visual system, to simplify the computational problem into smaller parts performed successively, to simplify the representation of multiple objects in a scene, and to facilitate passing the coordinates of a target object for action by using the coordinates of the object being fixated (Ballard, 1990; Rolls and Deco, 2002; Rolls et al., 2003; Aggelopoulos and Rolls, 2005; Rolls, 2008, 2012). This approach has now been adopted by some computer vision approaches (Denil et al., 2012).

Important issues are raised for future research.

First, how well does this approach scale up? At present there are 128 × 128 neurons in each of 4 layers of VisNet, that is 65,536 neurons. This is small compared to the number of neurons in the ventral visual stream, which number tens of millions of neurons (Rolls, 2008). If this is indeed a good model of the processing in the ventral visual system, as we hypothesize and on which VisNet is based (Rolls, 2012), then the system should scale up appropriately, that is, probably linearly. There are a number of different aspects that need to scale up. One is the number of objects that can be trained. A second is the number of views that can be trained. A third is the number of locations in which the system is trained, both because saliency mechanisms are not as accurate as the range of 32 pixels from the fovea over which we trained here (**Figure 4**), and because it may be advantageous to train at intermediate locations (**Figure 5**). We propose to scale up VisNet by 16 times, from 128 × 128 neurons per layer to 512 × 512 neurons per layer, and to simultaneously address all these issues.

Second, we have used a generically sound and well-known approach to bottom-up saliency, an approach developed by Koch, Itti, Harel and colleagues (Itti and Koch, 2000; Harel et al., 2006a,b). However, it is possible to tune saliency algorithms so that they are more likely to detect objects of certain classes, such as faces or cars. This may greatly increase the capability of the approach described here, and we plan to test how much improvement in performance for the detection and then identification of certain classes of objects can be obtained by incorporating more specialized saliency algorithms. Many saliency approaches and algorithms that are of interest for future research are available (Bruce and Tsotsos, 2006; Achanta et al., 2008; Zhang et al., 2008; Kootstra et al., 2010; Goferman et al., 2012; Riche et al., 2012; Jia et al., 2013; Li et al., 2013). For example, contextual information may be useful, such as the fact that sofas are not usually found in the sky, and that people are usually tall, skinny objects on the ground (though see Webb and Rolls, 2014), and contextual guidance models have been combined with bottom-up saliency models (Oliva and Torralba, 2006; Torralba et al., 2006; Ehinger et al., 2009; Kanan et al., 2009). We emphasize that in the system described here, only one fixation is assumed for each object in a scene, consistent with the fact that single neurons in the inferior temporal visual cortex provide sufficient information for object and face identification during a single fixation and in only 20–50 ms of neuronal firing, as shown by information theoretic analyses of neuronal activity and by backward masking (Rolls et al., 1994; Rolls and Tovee, 1994; Tovee and Rolls, 1995). [More detailed information may become available with repeated fixations on different parts of an object, and this has been investigated in computer vision (Barrington et al., 2008; Kanan and Cottrell, 2010; Larochelle and Hinton, 2010).]

Third, we have not utilized top-down attention in the developments described here. Top-down attention, whereby an object or set of objects is held active in a short term memory which biases the competitive networks in VisNet, can in principle improve performance considerably (Rolls and Deco, 2002; Deco and Rolls, 2005b; Rolls, 2008). Indeed, we have developed and successfully tested a reduced version of VisNet in which top-down attention does facilitate processing (Deco and Rolls, 2004), and this approach has also been used in computer vision (Walther et al., 2002). Another type of top-down effect is that task requirements can influence fixations in a scene (Hayhoe and Ballard, 2005). We plan in future to incorporate top-down attention into the full, current, version of VisNet, to investigate how this is likely to improve performance, especially for certain selected classes of object.

Fourth, it will be useful to investigate in future the incorporation of more powerful synaptic learning rules when training with the large number of transforms needed when learning invariance for both view and translation transforms of objects. With VisNet, we have so far used an associative (Hebbian) synaptic modification rule (with a trace of previous firing in the postsynaptic term), for biological plausibility (Rolls, 2012). However, to explore further the potential of the overall architecture of VisNet, it will be of interest to investigate how much performance improves when error correction of the post-synaptic firing with respect to the trace of previous neuronal activity is incorporated to implement gradient descent. Gradient descent (Einhauser et al., 2005; Wyss et al., 2006) or optimized slow learning (Wiskott and Sejnowski, 2002; Wiskott, 2003) have been found useful with different architectures.

Fifth, if a strong saliency peak occurs due to something in the background scene that is close to an object, or due to another trained object, how will the system respond? We suggest that the general answer is that the asymmetry that is present in the receptive fields of inferior temporal cortex neurons in cluttered scenes (Aggelopoulos and Rolls, 2005) that is related to the asymmetries caused by the sparse probabilistic forward connections of each neuron (Rolls et al., 2008) and that enables two instances of the same object close together to be correctly identified in terms of both object and position (Rolls et al., 2008) provides the solution, but it will be of interest to investigate this in detail.

Part of the value of the research described here is that it tests, and investigates the operation of, a theory of how view invariant object identification could be implemented by the cerebral cortex. Some predictions of the simulations are (1) that learning will need to be part of the process involved in view-invariant object identification, as the views of an object can be very different; (2) that for at least views of people, a few well-spaced views (we used 45°) should suffice; (3) that translation invariance in complex unstructured crowded scenes may need to be over just a few degrees, for fixation guided by bottom-up saliency has precision of that order at least for the types of object considered here, and repeated saccades are necessary to reach sufficiently close to an object in a large scene for the invariance available to be able to operate in object identification (Rolls et al., 2003; Aggelopoulos and Rolls, 2005); and (4) that just a single fixation of each object will in general suffice for object/person identification, because of the speed of cortical processing (Rolls and Treves, 2011; Rolls, 2012).

ACKNOWLEDGMENTS

The authors acknowledge with thanks the use of the GBVS software (Harel et al., 2006a,b) (http://www.vision.caltech.edu/~harel/share/gbvs.php). The images shown in **Figure 2** were created with Blender from models available at www.blendswap.com, and acknowledged as follows: truck—Opel Blitz by orokrhus; jeep by Jay-Artist; woman by Gerardus. The man was generated using MakeHuman available at www.makehuman.org.

REFERENCES

Abbott, L. F., Rolls, E. T., and Tovee, M. J. (1996). Representational capacity of face coding in monkeys. *Cereb. Cortex* 6, 498–505. doi: 10.1093/cercor/6.3.498

Achanta, R., Estrada, F., Wils, P., and Süsstrunk, S. (2008). Salient region detection and segmentation. *Comput. Vis. Syst.* 5008, 66–75. doi: 10.1007/978-3-540-79547-6_7

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. doi: 10.1152/jn.00553.2004

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. doi: 10.1111/j.1460-9568.2005.04487.x

Arcizet, F., Mirpour, K., and Bisley, J. W. (2011). A pure salience response in posterior parietal cortex. *Cereb. Cortex* 21, 2498–2506. doi: 10.1093/cercor/bhr035

Baddeley, R. J., Abbott, L. F., Booth, M. J. A., Sengpiel, F., Freeman, T., Wakeman, E. A., et al. (1997). Responses of neurons in primary and inferior temporal visual cortices to natural scenes. *Proc. R. Soc. B* 264, 1775–1783. doi: 10.1098/rspb.1997.0246

Ballard, D. H. (1990). "Animate vision uses object-centred reference frames," in *Advanced Neural Computers*, ed R. Eckmiller (North-Holland, Amsterdam: Elsevier), 229–236.

Barrington, L., Marks, T. K., Hsiao, J. H., and Cottrell, G. W. (2008). NIMBLE: a kernel density model of saccade-based visual memory. *J. Vis.* 8:17. doi: 10.1167/8.14.17

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. doi: 10.1016/0006-8993(85)91356-3

Bisley, J. W., and Goldberg, M. E. (2003). Neuronal activity in the lateral intraparietal area and spatial attention. *Science* 299, 81–86. doi: 10.1126/science.1077395

Bisley, J. W., and Goldberg, M. E. (2006). Neural correlates of attention and distractibility in the lateral intraparietal area. *J. Neurophysiol.* 95, 1696–1717. doi: 10.1152/jn.00848.2005

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. doi: 10.1093/cercor/8.6.510

Bruce, N. D. B., and Tsotsos, J. K. (2006). "Saliency based on information maximization," in *Advances in Neural Information Processing Systems 18: Proceedings of the 2005 Conference*, Vol. 18 (Cambridge, MA: MIT Press), 155.

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.

Crouzet, S. M., and Thorpe, S. J. (2011). Low-level cues and ultra-fast face detection. *Front. Psychol.* 2:342. doi: 10.3389/fpsyg.2011.00342

Daugman, J. (1988). Complete discrete 2D-Gabor transforms by neural networks for image analysis and compression. *IEEE Trans. Acoust. Speech Signal Process.* 36, 1169–1179. doi: 10.1109/29.1644

Deco, G., and Rolls, E. T. (2004). A neurodynamical cortical model of visual attention and invariant object recognition. *Vision Res.* 44, 621–644. doi: 10.1016/j.visres.2003.09.037

Deco, G., and Rolls, E. T. (2005a). Attention, short term memory, and action selection: a unifying theory. *Prog. Neurobiol.* 76, 236–256. doi: 10.1016/j.pneurobio.2005.08.00

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. doi: 10.1152/jn.01095.2004

Denil, M., Bazzani, L., Larochelle, H., and de Freitas, N. (2012). Learning where to attend with deep architectures for image tracking. *Neural Comput.* 24, 2151–2184.

Desimone, R., and Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annu. Rev. Neurosci.* 18, 193–222. doi: 10.1146/annurev.ne.18.030195.001205

De Valois, R. L., and De Valois, K. K. (1988). *Spatial Vision*. New York, NY: Oxford University Press .

DeWeese, M. R., and Meister, M. (1999). How to measure the information gained from one symbol. *Network* 10, 325–340. doi: 10.1088/0954-898X/10/4/303

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. doi: 10.1152/jn.00358.2002

DiCarlo, J. J., Zoccolan, D., and Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron* 73, 415–434. doi: 10.1016/j.neuron.2012.01.010

Ehinger, K. A., Hidalgo-Sotelo, B., Torralba, A., and Oliva, A. (2009). Modeling search for people in 900 scenes: a combined source model of eye guidance. *Vis. Cogn.* 17, 945–978. doi: 10.1080/13506280902834720

Einhauser, W., Eggert, J., Korner, E., and Konig, P. (2005). Learning viewpoint invariant object representations using a temporal coherence principle. *Biol. Cybern.* 93, 79–90. doi: 10.1007/s00422-005-0585-8

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. doi: 10.1007/s004220100284

Földiák, P. (1991). Learning invariance from transformation sequences. *Neural Comput.* 3, 193–199. doi: 10.1162/neco.1991.3.2.194

Földiák, P. (1992). *Models of Sensory Coding.* Technical Report CUED/F–INFENG/TR 91, Cambridge: University of Cambridge.

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. Cybernet.* 96, 547–560. doi: 10.1007/s00422-007-0149-1

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. doi: 10.1007/s00221-003-1737-5

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. Cybernet.* 36, 193–202. doi: 10.1007/BF00344251

Garthwaite, J. (2008). Concepts of neural nitric oxide-mediated transmission. *Eur. J. Neurosci.* 27, 2783–3802. doi: 10.1111/j.1460-9568.2008.06285.x

Goferman, S., Zelnik-Manor, L., and Tal, A. (2012). Context-aware saliency detection. *Pattern Anal. Mach. Intel. IEEE Trans.* 34, 1915–1926. doi: 10.1109/TPAMI.2011.272

Goldberg, M. E., Bisley, J. W., Powell, K. D., and Gottlieb, J. (2006). Saccades, salience and attention: the role of the lateral intraparietal area in visual behavior. *Prog. Brain Res.* 155, 157–175. doi: 10.1016/S0079-6123(06)55010-1

Griffin, G., Holub, A., and Perona, P. (2007). *The Caltech-256*. *Caltech Technical Report*. Los Angeles, CA: California Institute of Technology.

Gross, C., Bender, D., and Rocha-Miranda, C. (1969). Visual receptive fields of neurons in inferotemporal cortex of the monkey. *Science* 166, 1303–1306. doi: 10.1126/science.166.3910.1303

Harel, J., Koch, C., and Perona, P. (2006a). A Saliency Implementation in MATLAB. Available online at: http://www.vision.caltech.edu/∼harel/share/gbvs.php

Harel, J., Koch, C., and Perona, P. (2006b). Graph-based visual saliency. *Adv. Neural Inf. Process. Syst.* 545–552.

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. doi: 10.1007/BF00247948

Hawken, M. J., and Parker, A. J. (1987). Spatial properties of the monkey striate cortex. *Proc. R. Soc. Lond. B* 231, 251–288. doi: 10.1098/rspb.1987.0044

Hayhoe, M., and Ballard, D. (2005). Eye movements in natural behavior. *Trends Cogn. Sci.* 9, 188–194. doi: 10.1016/j.tics.2005.02.009

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. doi: 10.1016/0896-6273(90)90162-9

Hummel, J. E., and Biederman, I. (1992). Dynamic binding in a neural network for shape recognition. *Psychol. Rev.* 99, 480–517. doi: 10.1037/0033-295X.99.3.480

Isik, L., Leibo, J. Z., 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

Itti, L., and Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vis. Res.* 40, 1489–1506. doi: 10.1016/S0042-6989(99)00163-7

Jia, C., Hou, F., and Duan, L. (2013). Visual saliency based on local and global features in the spatial domain. *Int. J. Comput. Sci.* 10, 3, 713–719.

Kanan, C. (2013). Active object recognition with a space-variant retina. *ISRN Mach. Vis.* 2013:138057. doi: 10.1155/2013/138057

Kanan, C., and Cottrell, G. W. (2010). "Robust classification of objects, faces, and flowers using natural image statistics," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on* (IEEE), 2472–2479. doi: 10.1109/CVPR.2010.5539947

Kanan, C., Tong, M. H., Zhang, L., and Cottrell, G. W. (2009). SUN: Top-down saliency using natural statistics. *Vis. Cognit.* 17, 979–1003. doi: 10.1080/13506280902771138

Knudsen, E. I. (2011). Control from below: the role of a midbrain network in spatial attention. *Eur. J. Neurosci.* 33, 1961–1972. doi: 10.1111/j.1460-9568.2011.07696.x

Kootstra, G., Bergstrom, N., and Kragic, D. (2010). "Fast and automatic detection and segmentation of unknown objects," in *Humanoid Robots (Humanoids), 2010 10th IEEE-RAS International Conference* (IEEE), 442–447. doi: 10.1109/ICHR.2010.5686837

Larochelle, H., and Hinton, G. E. (2010). Learning to combine foveal glimpses with a third-order Boltzmann machine. *Adv. Neural Inf. Process. Syst.* 1, 1243–1251.

Lee, T. S. (1996). Image representation using 2D Gabor wavelets. *IEEE Trans. Patt. Anal. Mach. Intell.* 18, 959–971. doi: 10.1109/34.541406

Li, J., Levine, M. D., An, X., Xu, X., and He, H. (2013). Visual saliency based on scale-space analysis in the frequency domain. *IEEE Trans. Patt. Anal. Mach. Intell.* 35, 996–1010. doi: 10.1109/TPAMI.2012.147

Li, N., and DiCarlo, J. J. (2008). Unsupervised natural experience rapidly alters invariant object representation in visual cortex. *Science* 321, 1502–1507. doi: 10.1126/science.1160028

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. doi: 10.1016/j.neuron.2010.08.029

Li, N., and DiCarlo, J. J. (2012). Neuronal learning of invariant object representation in the ventral visual stream is not dependent on reward. *J. Neurosci.* 32, 6611–6620. doi: 10.1523/JNEUROSCI.3786-11.2012

Malsburg, C. V. D. (1973). Self-organization of orientation-sensitive columns in the striate cortex. *Kybernetik* 14, 85–100. doi: 10.1007/BF00288907

Marr, D. (1982). *Vision*. (San Francisco, CA: Freeman).

Miller, E. K., and Buschman, T. J. (2013). Cortical circuits for the control of attention. *Curr. Opin. Neurobiol.* 23, 216–222. doi: 10.1016/j.conb.2012.11.011

Miyashita, Y. (1988). Neuronal correlate of visual associative long-term memory in the primate temporal cortex. *Nature* 335, 817–820. doi: 10.1038/335817a0

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. doi: 10.1093/cercor/1.3.199

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. doi: 10.1007/s11263-007-0118-0

Oja, E. (1982). A simplified neuron model as a principal component analyzer. *J. Math. Biol.* 15, 267–273. doi: 10.1007/BF00275687

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. doi: 10.1016/S0079-6123(06)55002-2

Op de Beeck, H., and Vogels, R. (2000). Spatial sensitivity of macaque inferior temporal neurons. *J. Comp. Neurol.* 426, 505–518. doi: 10.1002/1096-9861(20001030)426:4<505::AID-CNE1>3.0.CO;2-M

Panzeri, S., Treves, A., Schultz, S., and Rolls, E. T. (1999). On decoding the responses of a population of neurons from short time epochs. *Neural Comput.* 11, 1553–1577. doi: 10.1162/089976699300016142

Perrett, D. I., Oram, M. W., Harries, M. H., Bevan, R., Hietanen, J. K. and Benson, P. J. (1991). Viewer–centered and object centered coding of heads in the macaque temporal cortex. *Exp. Brain Res.* 86, 159–173. doi: 10.1007/BF00231050

Perry, G., Rolls, E. T., and Stringer, S. M. (2006). Spatial vs temporal continuity in view invariant visual object recognition learning. *Vis. Res.* 46, 3994–4006. doi: 10.1016/j.visres.2006.07.025

Perry, G., Rolls, E. T., and Stringer, S. M. (2010). Continuous transformation learning of translation invariant representations. *Exp. Brain Res.* 204, 255–270. doi: 10.1007/s00221-010-2309-0

Pinto, N., Doukhan, D., DiCarlo, J. J., and Cox, D. D. (2009). A high-throughput screening approach to discovering good forms of biologically inspired visual representation. *PLoS Comput. Biol.* 5:e1000579. doi: 10.1371/journal.pcbi.1000579

Pollen, D., and Ronner, S. (1981). Phase relationship between adjacent simple cells in the visual cortex. *Science* 212, 1409–1411. doi: 10.1126/science.7233231

Rhodes, P. (1992). The open time of the NMDA channel facilitates the self-organisation of invariant object responses in cortex. *Soc. Neurosci. Abstr.* 18, 740.

Riche, N., Mancas, M., Gosselin, B., and Dutoit, T. (2012). "Rare: a new bottom-up saliency model," in *Image Processing, 2012 19th IEEE Conference on* (IEEE), 641–644. doi: 10.1109/ICIP.2012.6466941

Riesenhuber, M., and Poggio, T. (2000). Models of object recognition. *Nat. Neurosci. Suppl.* 3, 1199–1204. doi: 10.1038/81479

Robinson, L., and Rolls, E. T. (2014). Invariant visual object recognition: the biological plausibility of two approaches.

Rolls, E. T. (1992). Neurophysiological mechanisms underlying face processing within and beyond the temporal cortical visual areas. *Philos. Trans. R. Soc.* 335, 11–21. doi: 10.1098/rstb.1992.0002

Rolls, E. T. (1995). Learning mechanisms in the temporal lobe visual cortex. *Behav. Brain Res.* 66, 177–185. doi: 10.1016/0166-4328(94)00138-6

Rolls, E. T. (2000). Functions of the primate temporal lobe cortical visual areas in invariant visual object and face recognition. *Neuron* 27, 205–218. doi: 10.1016/S0896-6273(00)00030-1

Rolls, E. T. (2007). The representation of information about faces in the temporal and frontal lobes of primates including humans. *Neuropsychologia* 45, 124–143. doi: 10.1016/j.neuropsychologia.2006.04.019

Rolls, E. T. (2008). *Memory, Attention, and Decision-Making. A Unifying Computational Neuroscience Approach*. Oxford: Oxford University Press.

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

Rolls, E. T. (2014). *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. doi: 10.1007/s00422-003-0451-5

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. doi: 10.1007/BF00243828

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. doi: 10.1016/0042-6989(87)90081-2

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. doi: 10.1016/0042-6989(85)90091-4

Rolls, E. T., and Deco, G. (2002). *Computational Neuroscience of Vision*. Oxford: Oxford University Press.

Rolls, E. T., Franco, L., Aggelopoulos, N. C., and Jerez, J. M. (2006). 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. doi: 10.1016/j.visres.2006.07.026

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. doi: 10.1162/089976600300014845

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. doi: 10.1080/net.12.2.111.129

Rolls, E. T., and Stringer, S. M. (2006). Invariant visual object recognition: a model, with lighting invariance. *J. Physiol. Paris* 100, 43–62. doi: 10.1016/j.jphysparis.2006.09.004

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. doi: 10.1162/neco.2007.19.1.139

Rolls, E. T., and Tovee, M. J. (1994). Processing speed in the cerebral cortex and the neurophysiology of visual masking. *Proc. R. Soc. B* 257, 9–15. doi: 10.1098/rspb.1994.0087

Rolls, E. T., and Tovee, M. J. (1995). Sparseness of the neuronal representation of stimuli in the primate temporal visual cortex. *J. Neurophysiol.* 73, 713–726.

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. doi: 10.1007/BF00227340

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. doi: 10.1016/j.pneurobio.2011.08.002

Rolls, E. T., Treves, A., and Tovee, M. J. (1997a). 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. (1997b). Information in the neuronal representation of individual stimuli in the primate temporal visual cortex. *J. Comput. Neurosci.* 4, 309–333.

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. doi: 10.1111/j.1460-9568.2008.06486.x

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. doi: 10.1016/S0079-6123(06)65004-8

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. doi: 10.1073/pnas.0700622104

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. doi: 10.1109/TPAMI.2007.56

Sheinberg, D. L., and Logothetis, N. K. (2001). Noticing familiar objects in real world scenes: the role of temporal cortical neurons in natural vision. *J. Neurosci.* 21, 1340–1350.

Soltani, A., and Koch, C. (2010). Visual saliency computations: mechanisms, constraints, and the effect of feedback. *J. Neurosci.* 30, 12831–12843. doi: 10.1523/JNEUROSCI.1517-10.2010

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.

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. doi: 10.1007/s00422-005-0030-z

Stringer, S. M., and Rolls, E. T. (2000). Position invariant recognition in the visual system with cluttered environments. *Neural Netw.* 13, 305–315. doi: 10.1016/S0893-6080(00)00017-4

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. doi: 10.1162/089976602760407982

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. doi: 10.1016/j.neunet.2007.11.004

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. doi: 10.1080/09548980701556055

Sutton, R. S., and Barto, A. G. (1981). Towards a modern theory of adaptive networks: expectation and prediction. *Psychol. Rev.* 88, 135–170. doi: 10.1037/0033-295X.88.2.135

Thorpe, S. J. (2009). The speed of categorization in the human visual system. *Neuron* 62, 168–170. doi: 10.1016/j.neuron.2009.04.012

Torralba, A., Oliva, A., Castelhano, M. S., and Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychol. Rev.* 113, 766–786. doi: 10.1037/0033-295X.113.4.766

Tovee, M. J., and Rolls, E. T. (1995). Information encoding in short firing rate epochs by single neurons in the primate temporal visual cortex. *Visual Cogn.* 2, 35–58. doi: 10.1080/13506289508401721

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., 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. Gharamani (Cambridge, MA: MIT Press), 293–300.

Treves, A., Panzeri, S., Rolls, E. T., Booth, M., and Wakeman, E. A. (1999). Firing rate distributions and efficiency of information transmission of inferior temporal cortex neurons to natural visual stimuli. *Neural Comput.* 11, 601–631. doi: 10.1162/089976699300016593

Ungerleider, L. G., and Haxby, J. V. (1994). "What" and "Where" in the human brain. *Curr. Opin. Neurobiol.* 4, 157–165. doi: 10.1016/0959-4388(94)90066-3

Ungerleider, L. G., and Mishkin, M. (1982). "Two cortical visual systems," in *Analysis of Visual Behaviour*, eds D. Ingle, M. A. Goodale, and R. J. W. (Cambridge, MA: Mansfield MIT Press), 549–586.

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. doi: 10.1126/science.1734518

Wallis, G. (2013). Toward a unified model of face and object recognition in the human visual system. *Front. Psychol.* 4:497. doi: 10.3389/fpsyg.2013.00497

Wallis, G., and Rolls, E. T. (1997). Invariant face and object recognition in the visual system. *Prog. Neurobiol.* 51, 167–194. doi: 10.1016/S0301-0082(96)00054-8

Wallis, G., Rolls, E. T., and Földiák, P. (1993). Learning invariant responses to the natural transformations of objects. *Int. Joint Conf. Neural Netw.* 2, 1087–1090.

Walther, D., Itti, L., Riesenhuber, M., Poggio, T., and Koch, C. (2002). Attentional selection for object recognition–a gentle way. *Biol. Mot. Comput. Vis.* 472–479.

Webb, T. J., and Rolls, E. T. (2014). Deformation-specific and deformation-invariant visual object recognition: pose vs identity recognition of people and deforming objects. *Front. Comput. Neurosci.* 8:37. doi: 10.3389/fncom.2014.00037

Wiskott, L. (2003). Slow feature analysis: a theoretical analysis of optimal free responses. *Neural Comput.* 15, 2147–2177. doi: 10.1162/089976603322297331

Wiskott, L., and Sejnowski, T. J. (2002). Slow feature analysis: unsupervised learning of invariances. *Neural Comput.* 14, 715–770. doi: 10.1162/089976602317318938

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

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D. and DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl. Acad. Sci. U.S.A.* 111, 8619–8624. doi: 10.1073/pnas.1403112111

Zhang, L., Tong, M. H., Marks, T. K., Shan, H., and Cottrell, G. W. (2008). SUN: A Bayesian framework for saliency using natural statistics. *J. Vis.* 8:32. doi: 10.1167/8.7.32

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: 21 May 2014; accepted: 16 July 2014; published online: 12 August 2014.
Citation: Rolls ET and Webb TJ (2014) Finding and recognizing objects in natural scenes: complementary computations in the dorsal and ventral visual systems. Front. Comput. Neurosci. ***8****:85. doi: 10.3389/fncom.2014.00085*
This article was submitted to the journal Frontiers in Computational Neuroscience.

A. APPENDIX: THE ARCHITECTURE OF VISNET

This Appendix describes the functional architecture, operation, and testing of VisNet as used in this paper. VisNet is a hierarchical feedforward 4-layer network that models properties of the ventral visual system involved in invariant visual object recognition (Rolls, 2008, 2012).

A.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), Wallis and Rolls (1997), and Rolls (2012). 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; Rolls and Deco, 2002; Rolls, 2008, 2012).

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 temporally extended cell firing in a local cortical attractor could implement a short-term memory of previous neuronal firing (Rolls, 2008).]

- 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, 2008). [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 (Földiák, 1992; Rolls, 1992; Rhodes, 1992; Spruston et al., 1995; Hestrin et al., 1990).
- 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 (Földiák, 1992; Montague et al., 1991; 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:

$$\delta w_j = \alpha \bar{y}^{\tau} x_j \tag{A1}$$

where

$$\bar{y}^{\tau} = (1-\eta) y^{\tau} + \eta \bar{y}^{\tau-1} \tag{A2}$$

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 τ.	α:	Learning rate.
w_j:	Synaptic weight between j^{th} input and the neuron.	η:	Trace value. The optimal value varies with presentation sequence length.

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 with a modified Hebbian trace learning rule (Equation A3) 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

$$\delta w_j = \alpha \bar{y}^{\tau-1} x_j^{\tau}. \tag{A3}$$

The trace shown in Equation (A3) is in the postsynaptic term. The crucial difference from the earlier rule (see Equation A1) 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. This has the effect of updating the weights based on the preceding activity of the neuron, which is likely given the spatio-temporal statistics of the visual world to be from previous transforms of the same object (Rolls and Milward, 2000; Rolls and Stringer, 2001). This is biologically not at all implausible, as considered in more detail elsewhere (Rolls, 2008, 2012), and this version of the trace rule was used in this investigation.

The optimal value of η in the trace rule is likely to be different for different layers of VisNet. For early layers with small receptive fields, few successive transforms are likely to contain similar information within the receptive field, so the value for η might be low to produce a short trace. In later layers of VisNet, successive transforms may be in the receptive field for longer, and invariance may be developing in earlier layers, so a longer trace may be beneficial. In practice, after exploration we used η values of 0.6 for layer 2, and 0.8 for layers 3 and 4. In addition, 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 (Rolls, 2008, 2012). For this reason, purely associative learning with no temporal trace was used in layer 1 of VisNet (Rolls and Milward, 2000).

The following principled method was introduced to choose the value of the learning rate α for each layer. The mean weight change from all the neurons in that layer for each epoch of training was measured, and was set so that with slow learning over 15–50 trials, the weight changes per epoch would gradually decrease and asymptote with that number of epochs, reflecting convergence. Slow learning rates are useful in competitive nets, for if the learning rates are too high, previous learning in the synaptic weights will be overwritten by large weight changes later within the same epoch produced if a neuron starts to respond to another stimulus (Rolls, 2008). If the learning rates are too low, then no useful learning or convergence will occur. It was found that the following learning rates enabled good operation with the 100 transforms of each of 4 stimuli used in each epoch in the present investigation: Layer 1 $\alpha = 0.05$; Layer 2 $\alpha = 0.03$ (this is relatively high to allow for the sparse representations in layer 1); Layer 3 $\alpha = 0.005$; Layer 4 $\alpha = 0.005$.

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

A.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 (Van Essen et al., 1992; Rolls, 1992, 2008, for example) 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 A1** which will contain approximately 67% of the connections from the preceding layer. **Table A1** shows the dimensions for the research described here, a (16×) larger version than the version of VisNet used in most of our previous investigations, which utilized 32 × 32 neurons per layer. For the research on view and translation invariance learning described here, we decreased the number of connections to layer 1 neurons to 100 (from 272), in order to increase the selectivity of the network between objects. We increased the number of connections to each neuron in layers 2–4 to 400 (from 100), because this helped layer 4 neurons to reflect evidence from neurons in previous layers about the large number of transforms (typically 100 transforms, from 4 views of each object and 25 locations) each of which corresponded to a particular object.

Table A1 | VisNet dimensions.

	Dimensions	# Connections	Radius
Layer 4	128 × 128	400	48
Layer 3	128 × 128	400	36
Layer 2	128 × 128	400	24
Layer 1	128 × 128	100	24
Input layer	256 × 256 × 16	–	–

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 elsewhere (Rolls, 2008, 2012).

A.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, 2008). 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). The lateral inhibition used in this investigation used the parameters for σ shown in **Table A3**.

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 elsewhere (Rolls, 2008, 2012). The lateral inhibition was implemented by convolving the activation of the neurons in a layer with a spatial filter, I, where δ controls the contrast and σ controls the width, and a and b index the distance away from the center of the filter

$$I_{a,b} = \begin{cases} -\delta e^{-\frac{a^2+b^2}{\sigma^2}} & \text{if } a \neq 0 \text{ or } b \neq 0, \\ 1 - \sum\limits_{a\neq 0, b\neq 0} I_{a,b} & \text{if } a = 0 \text{ and } b = 0. \end{cases} \tag{A4}$$

This is a filter that leaves the average activity unchanged.

The second stage involves contrast enhancement. A sigmoid activation function was used in the way described previously (Rolls and Milward, 2000):

$$y = \mathrm{f}^{\mathrm{sigmoid}}(r) = \frac{1}{1 + e^{-2\beta(r-\alpha)}} \tag{A5}$$

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 β is the slope or gain and α 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; Franco et al., 2007; Rolls, 2008; Rolls and Treves, 2011) as:

$$a = \frac{(\sum_i y_i/n)^2}{\sum_i y_i^2/n} \tag{A6}$$

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.

The sigmoid activation function was used with parameters (selected after a number of optimization runs) as shown in **Table A2**.

In addition, the lateral inhibition parameters are as shown in **Table A3**.

A.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 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 (Hummel and Biederman, 1992; Buhmann et al., 1991; Fukushima, 1980), 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 were computed by Gabor filters. Each individual filter is tuned to spatial frequency (0.0626 to 0.5 cycles / pixel over four octaves); orientation (0° to 135° in steps of 45°); and sign (±1). Of the 100 layer 1 connections, the number to each group in VisNetL is as shown in **Table A4**. 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. 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.

The Gabor filters used were similar to those used previously (Deco and Rolls, 2004). 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 to 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).

Table A2 | Sigmoid parameters for the runs with 25 locations by Rolls and Milward (2000).

Layer	1	2	3	4
Percentile	99.2	98	88	95
Slope β	190	40	75	26

Table A3 | Lateral inhibition parameters for the 25-location runs.

Layer	1	2	3	4
Radius, σ	1.38	2.7	4.0	6.0
Contrast, δ	1.5	1.5	1.6	1.4

Table A4 | VisNet Layer 1 Connectivity.

Frequency	0.5	0.25	0.125	0.0625
# Connections	74	19	5	2

The frequency is in cycles per pixel.

In more detail, the Gabor filters are constructed as follows (Deco and Rolls, 2004). We consider a pixelized grey-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 grey level brightness value coded in a scale between 0 (black) and 255 (white). The first step in the preprocessing consists of removing the DC component of the image (i.e., the mean value of the grey-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 Γ_{ij} defined by the equation

$$\Gamma_{ij} = \Gamma_{ij}^{\text{orig}} - \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \Gamma_{ij}^{\text{orig}}. \tag{A7}$$

Feedforward 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

$$G_{pqkl}(x, y) = a^{-k} \Psi_{\Theta_l}(a^{-k}(x - 2p), a^{-k}(y - 2q)) \tag{A8}$$

where

$$\Psi_{\Theta_l} = \Psi(x \cos(l\Theta_0) + y \sin(l\Theta_0), -x \sin(l\Theta_0) + y \cos(l\Theta_0)), \tag{A9}$$

and the mother wavelet is given by

$$\Psi(x, y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{8}(4x^2+y^2)} [e^{i\kappa x} - e^{-\frac{\kappa^2}{2}}]. \tag{A10}$$

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 used symmetric filters with the angular spacing between the different orientations set to 45 degrees; and with 4 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 A4**.

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.

A.5 MEASURES FOR NETWORK PERFORMANCE

A.5.1 Information theory measures

A neuron can be said to have learnt an invariant representation if it discriminates one set of stimuli from another set, across all transforms. 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., 1997a; Rolls and Treves, 1998; Rolls and Deco, 2002; Franco et al., 2007; Rolls, 2007, 2008; Rolls and Treves, 2011). We used measures of network performance (Rolls and Milward, 2000) based on information theory and similar to those used in the analysis of the firing of real neurons in the brain (Rolls, 2008; Rolls and Treves, 2011). A single cell information measure was introduced which is the maximum amount of information the cell has about any one object independently of which transform (here position on the retina and view) 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) also introduced a multiple cell information measure used here, 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, 2008 and 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., 1997b and Rolls and Milward, 2000). $I(s, R)$ is given by

$$I(s, R) = \sum_{r \in R} P(r|s) \log_2 \frac{P(r|s)}{P(r)} \tag{A11}$$

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. (1997a) 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. (1997a) 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

$$I(S, S') = \sum_{s,s'} P(s, s') \log_2 \frac{P(s, s')}{P(s)P(s')} \tag{A12}$$

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).

A.5.2 Pattern association decoding

The output of the inferior temporal visual cortex reaches structures such as the orbitofrontal cortex and amygdala, where associations to other stimuli are learned by a pattern association network with an associative (Hebbian) learning rule (Rolls, 2008, 2014). We therefore used a one-layer pattern association network (Rolls, 2008) to measure how well the output of VisNet could be classified into one of the objects. The pattern association network had four output neurons, one for each object. The inputs were the ten neurons from layer 4 of VisNet for each of the four objects with the best single cell information, making 40 inputs to each neuron. The network was trained with the Hebb rule:

$$\delta w_{ij} = \alpha y_i x_j \tag{A13}$$

where δw_{ij} is the change of the synaptic weight w_{ij} that results from the simultaneous (or conjunctive) presence of presynaptic firing x_j and postsynaptic firing or activation y_i, and α is a learning rate constant that specifies how much the synapses alter on any one pairing. The pattern associator was trained for one trial on the output of VisNet produced by every transform of each object.

Performance on the test images extracted from the scenes was tested by presenting an image to VisNet, and then measuring the classification produced by the pattern associator. Performance was measured by the percentage of the correct classifications of an image as the correct object.

This approach to measuring the performance is very biologically appropriate, for it models the type of learning thought to be implemented in structures that receive information from the inferior temporal visual cortex such as the orbitofrontal cortex and amygdala (Rolls, 2008, 2014). The small number of neurons selected from layer 4 of VisNet might correspond to the most selective for this stimulus set in a sparse distributed representation (Rolls, 2008; Rolls and Treves, 2011). The method would measure whether neurons of the type recorded in the inferior temporal visual cortex with good view and position invariance are developed in VisNet. In fact, an appropriate neuron for an input to such a decoding mechanism might have high firing rates to all or most of the view and position transforms of one of the stimuli, and smaller or no responses to any of the transforms of other objects, as found in the inferior temporal cortex for some neurons (Hasselmo et al., 1989; Perrett et al., 1991; Booth and Rolls, 1998), and as illustrated for VisNet layer 4 neuron in this investigation in **Figure 5B**. Moreover, it would be inappropriate to train a device such as a support vector machine or even an error correction perceptron on the outputs of all the neurons in layer 4 of VisNet to produce 4 classifications, for such learning procedures, not biologically plausible (Rolls, 2008), could map the responses produced by a multilayer network with untrained random weights to obtain good classifications.

frontiers
in Computational Neuroscience

ORIGINAL RESEARCH
published: 07 October 2015
doi: 10.3389/fncom.2015.00115

Unsupervised invariance learning of transformation sequences in a model of object recognition yields selectivity for non-accidental properties

Sarah M. Parker[1] *and Thomas Serre*[1, 2]*

[1] *Department of Cognitive, Linguistic, and Psychological Sciences, Brown University, Providence, RI, USA,* [2] *Brown Institute for Brain Sciences, Providence, RI, USA*

Non-accidental properties (NAPs) correspond to image properties that are invariant to changes in viewpoint (e.g., straight vs. curved contours) and are distinguished from metric properties (MPs) that can change continuously with in-depth object rotation (e.g., aspect ratio, degree of curvature, etc.). Behavioral and electrophysiological studies of shape processing have demonstrated greater sensitivity to differences in NAPs than in MPs. However, previous work has shown that such sensitivity is lacking in multiple-views models of object recognition such as HMAX. These models typically assume that object processing is based on populations of view-tuned neurons with distributed symmetrical bell-shaped tuning that are modulated at least as much by differences in MPs as in NAPs. Here, we test the hypothesis that unsupervised learning of invariances to object transformations may increase the sensitivity to differences in NAPs vs. MPs in HMAX. We collected a database of video sequences with objects slowly rotating in-depth in an attempt to mimic sequences viewed during object manipulation by young children during early developmental stages. We show that unsupervised learning yields shape-tuning in higher stages with greater sensitivity to differences in NAPs vs. MPs in agreement with monkey IT data. Together, these results suggest that greater NAP sensitivity may arise from experiencing different in-depth rotations of objects.

Keywords: inferotemporal cortex, ventral stream, HMAX, invariance, object constancy, object recognition, learning

Edited by:
Hans P. Op De Beeck,
University of Leuven, Belgium

Reviewed by:
Michael J. Tarr,
Carnegie Mellon University, USA
Irving Biederman,
University of Southern California, USA

***Correspondence:**
Thomas Serre,
Department of Cognitive, Linguistic,
and Psychological Sciences, Brown
University, 190 Thayer st., Providence,
RI 02912, USA
thomas_serre@brown.edu

Received: *18 March 2015*
Accepted: *07 September 2015*
Published: *07 October 2015*

Citation:
Parker SM and Serre T (2015)
Unsupervised invariance learning of
transformation sequences in a model
of object recognition yields selectivity
for non-accidental properties.
Front. Comput. Neurosci. 9:115.
doi: 10.3389/fncom.2015.00115

1. Introduction

Invariant object recognition is a notoriously challenging computational problem (Marr, 1982). Our visual system has to deal with large intra-class variations owing to the effect of 2D and 3D transformations (including translation, scaling and rotation) because small changes in an object's 3D view may yield large changes on its 2D projection on our retinas. Yet, despite these large intra-class variations, primates are capable of robustly and effortlessly recognizing objects (Thorpe et al., 1996), vastly outperforming the best existing computer vision systems.

Object constancy requires the development of visual representations that remain stable across object transformations (Földiák, 1998). In particular, one may distinguish between those object properties that will remain stable across changes in viewpoint and those that will not (see **Figure 1**, for an illustration). Properties such as the degree of curvature of an object's contours, its length, or the amount of expansion of a cross section are examples of properties that will be affected by changes in viewpoint. Conversely, there also exist qualitative shape properties that remain stable across changes in viewpoint, e.g., whether an edge is straight or curved, whether a surface is convex or concave, or whether a cross section ends at a point vs. a side. These qualitative properties are known as non-accidental properties (NAPs) and need to be contrasted with their quantitative counterparts known as metric properties (MPs).

There is a long history of studies related to NAPs in computational vision (see Lowe, 1984, for review): From a theoretical point of view, a visual system needs to focus on the detection of image structures that are unlikely to have arisen by accident. For instance, the probability of a curved edge to appear straight because of projection is extremely small and would happen as an "accident" of viewpoint (Richards et al., 1996). The stability of NAPs over viewpoints makes them useful for achieving object constancy. Indeed, NAPs have been the focus of a prominent psychological theory of object recognition called the Recognition-by-Components (RBC) theory (Biederman, 1987). Briefly, this *structural-description* theory states that the visual system may encode a finite visual vocabulary of basic 3D shapes called geons. These geons can be differentiated on the basis of differences in NAPs, and generic object categories can be represented as compositions of geons. This theory has motivated the design of a number of experimental studies and it is now relatively well established that our visual system exhibit greater sensitivity to differences in NAPs compared to MPs (see Biederman, 2007, for review).

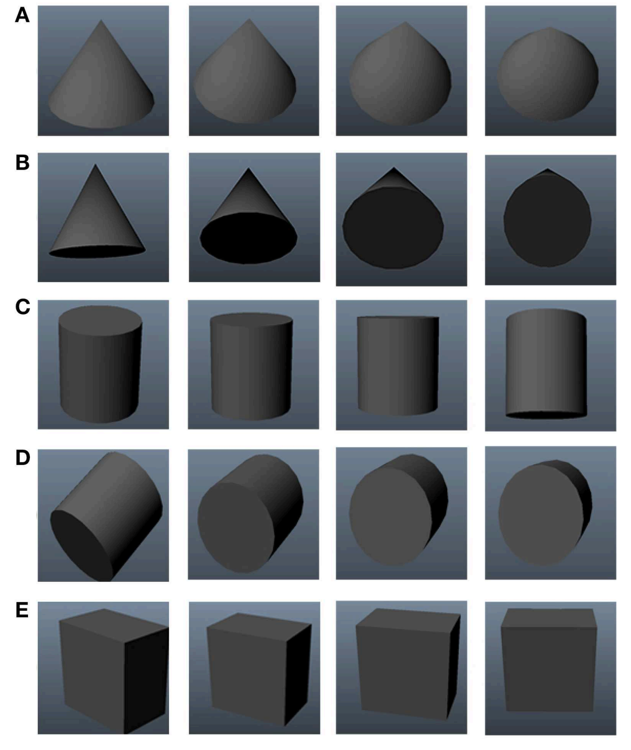


FIGURE 1 | Representative appearance changes undergone by objects during out-of-plane rotations. Variations of metric properties here include: **(A)** increasing angle at a point and **(B)** increasing size, shape and curvature of cross section of a cone **(C)** increasing size, shape and curvature of cross section of a cylinder **(D)** decreasing length of a cylinder and **(E)** decreasing area of cross-section and increasingly skewness of the edges of a cube.

Behaviorally, it has been shown that participants can more accurately distinguish between two objects that differ along an NAP vs. an MP (Biederman and Bar, 1999). Furthermore, when trained to recognize novel object categories where two NAPs (the degree of curvature and the degree of parallelism) are systematically varied, adult participants are more likely to treat a change in NAP as categorical (as opposed to within-category variation) compared to a similar change in MP (Abecassis et al., 2001). When a more sensitive paradigm is employed, preschool children, like adults, find it easier to discriminate NAPs vs. MPs (Amir et al., 2014). In addition, both adults and 4 month olds exhibit a saccadic preference for NAPs vs. MPs (Amir et al., 2011).

The neural basis of NAP selectivity was more directly studied by Kayaert et al. (2003) who recorded neuronal responses in the inferior temporal cortex (ITC) of the macaque. It was shown that neural responses are more strongly modulated by changes in NAPs than by equally large pixel-wise changes in MPs (Kayaert et al., 2003).

Further work later showed that such increased NAP sensitivity is incompatible with *multiple-views* models of object recognition such as the HMAX (see Riesenhuber and Poggio, 1999; Serre, 2014, for reviews), which assume that shape processing is based on broadly-tuned neuronal populations with distributed symmetric bell-shaped tuning: Shape-tuned units in these models are modulated at least as much by differences in MPs as in NAPs (Amir et al., 2012). It remains an open question—if and how—HMAX can be modified to account for the increased NAP sensitivity found both behaviorally and electrophysiologically.

Here, we test the hypothesis that mechanisms for learning transformation sequences may increase the model sensitivity to differences in NAPs vs. MPs. Given that MP changes result in part from generic object transformations (3D rotation), and given the focus of the original model on 2D transformations, we reasoned that learning invariances to natural object transformations should yield a decrease in the sensitivity of model units to MPs compared to NAPs (see Tarr and Kriegman, 2001, for a similar argument). To test our hypothesis, we created a database of video sequences with objects slowly rotating in depth in an attempt to mimic sequences viewed during object manipulation by young children during early developmental stages (**Figure 3**).

Several algorithms have been proposed for learning transformation sequences (e.g., Perrett et al., 1984; Foldiak, 1991; Hietanen et al., 1992; Wallis et al., 1993; Einhäuser et al., 2002; Wiskott and Sejnowski, 2002; Spratling, 2005; Stringer et al., 2006; Masquelier et al., 2007). Here, we consider a simple form of sequence learning via a "temporal pooling" mechanism similar to that used in the Hierarchical Temporal Memory algorithm (Hawkins and Blakeslee, 2004). The basic idea is to incorporate invariance pooling mechanisms in intermediate stages of the HMAX to include more generic object transformations (such as 3D rotation).

In the original model, IT-like units in the the last stage are organized in feature columns (**Figure 2A**) modeled after those found in cortex (Tanaka, 2003): Each feature column is characterized by its tuning for a distinct visual feature over a range of positions and scales. Feature selectivity is learned from individual object views (Serre et al., 2007b) and each column activity reflects the degree of similarity between an input stimulus and the corresponding preferred feature. Assuming *N* feature columns, the resulting population activity encodes an input stimulus as an *N*-dimensional pattern of activity (**Figure 2B**). The difference in the pattern of activity associated with two distinct input stimuli reflects the visual dissimilarity between the two stimuli and does not distinguish between an MP vs. NAP change ($\Delta_{MP-Base} \approx \Delta_{NAP-Base}$; **Figure 2C**).

In the extended model, feature columns include multiple views of the same feature sampled from short object transformation sequences (~300 ms). The responses of features within a column are then combined via a max operation (as done in the original model for invariance to position and scale; **Figures 2A,B**). Such unsupervised learning mechanism is consistent with both human behavioral (Wallis and Bülthoff, 2001; Cox et al., 2005) and nonhuman primate (Li et al., 2008, 2010) studies which suggest that tolerance to object transformations is at least partly supported by the natural temporal contiguity of visual experience. As we will show, the proposed pooling mechanism yields a visual representation which exhibits greater tolerance to object transformations and, as a result, a greater sensitivity for NAP compared to MP changes ($\Delta_{MP-Base} < \Delta_{NAP-Base}$) in agreement with neurophysiological data.

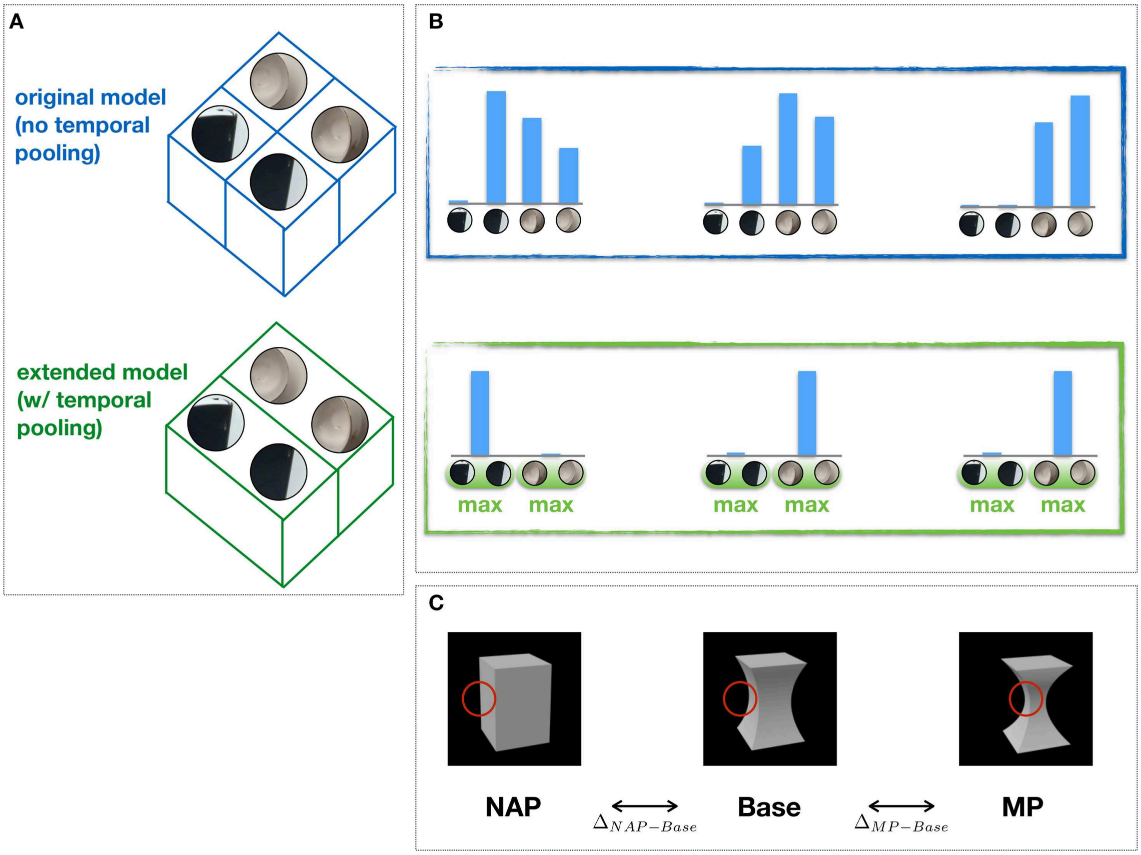


FIGURE 2 | Feature columns, invariance to object transformations and NAP sensitivity. (A) Feature columns in the extended (bottom) vs. the original (top) model (i.e., w/ and w/o temporal pooling). One of the key computational mechanisms in the HMAX builds on the proposal by Hubel and Wiesel (1962) to achieve tolerance of 2D transformations via a selective pooling mechanism (at the level of complex cells) over afferent units with the same preferred selectivity (feature) but slightly different positions and scales (not shown). Here, we propose a simple extension of this idea to include a more general form of pooling, i.e., over a transformation sequence of the preferred stimulus learned through visual experience. This pooling is done within feature columns which include different views of the same feature learned from object transformation sequences. **(B)** Shown are the corresponding patterns of (column) activity for the original and the extended model. **(C)** Sample stimuli used to probe the selectivity for MP ($\Delta_{MP-Base}$) vs. NAP ($\Delta_{NAP-Base}$) changes from a Base stimulus as done in Kayaert et al. (2003). Whereas the original model fails to exhibit any sensitivity to NAP vs. MP changes ($\Delta_{NAP-Base} \approx \Delta_{MP-Base}$), the extended model exhibits greater tolerance to object transformation through the "temporal pooling" mechanism and, as a result, greater sensitivity to NAP vs. MP changes ($\Delta_{NAP-Base} > \Delta_{MP-Base}$). Shown in red is the hypothetical stimulus location driving the unit response.

2. Materials and Methods

2.1. Video Database

We used a consumer-grade camera to collect short video sequences (30 Hz) with the aim to mimic object manipulations (**Figure 3**). Everyday objects were placed in diverse environments and the camera was moved slowly around the object to create 3–5 s long videos of the object undergoing a transformation (combination of small translation, scaling, and in-depth rotation). The video database included 12 common objects routinely found in a dorm room with at least 20 video sequences per category for a total of about 240 video sequences. For each category, the object background, initial viewpoint, and magnitude of the rotation was varied as much as possible.

2.2. The HMAX Model

Here, for convenience, we used a somewhat simplified implementation of the HMAX, which includes only four processing stages (Serre et al., 2007b). We only very briefly review the model architecture as details of the implementation have been described elsewhere (see Serre et al., 2007b; Serre, 2014, for details) and source code for the model is publicly available at: http://serre-lab.clps.brown.edu/resources.

The HMAX model of object recognition combines a hierarchical build-up of invariance and selectivity (inspired by Fukushima, 1980) with the idea of multiple-views (view-based) recognition of 3D objects (Riesenhuber and Poggio, 1999, 2000). Over the years, several related hierarchical models have been developed (Mel, 1997; Wallis and Rolls, 1997; LeCun et al., 1998; Riesenhuber and Poggio, 1999; Ullman et al., 2002; Amit and Mascaro, 2003; Wersing and Köerner, 2003; Masquelier and Thorpe, 2007; Mutch and Lowe, 2008; Jarrett et al., 2009; Pinto et al., 2011; Saxe et al., 2011). We focus here on the HMAX because the underlying parameters of the architecture were explicitly derived from available neuroscience data and because this was the model originally tested for NAP modulation and compared against IT data by Amir et al. (2012). Without loss of generality, we expect related models to exhibit similar trends.

Each processing stage in the HMAX model is organized in columns. Each column contains a complete dictionary of S unit selectivities for that particular layer. For instance, a column in the first S_1 stage (modeled after simple cells in striate cortex; see Lades et al., 1993, for an early system using Gabor filters for face recognition) contains a complete range of orientation and spatial frequency tuning and a column of $S2$ units (corresponding to units in intermediate areas of the ventral stream of the visual cortex) to a complete dictionary of shape-tuned units (see later). Simple units pool over afferent units using a Gaussian-like tuning operation. That is, the response y of a simple unit, receiving the pattern of inputs $\mathbf{x}$ from the previous layer is given by $y = \exp -\gamma ||\mathbf{w} - \mathbf{x}||^2$, where γ defines the sharpness of the tuning around the preferred stimulus of the unit corresponding to the weight vector $\mathbf{w}$. These columns are then replicated at different positions and scales, which is the key mechanism by which the model gains its tolerance to 2D transformations (position and scale) at the level of C units. The pooling operation at the level of complex units is a max operation over afferent units. That is, the response y of a complex unit from the previous layer is given by $y = \max_{j \in pool} x_j$. The parameters governing the invariance properties of the C units (i.e., the size of the pooling range over

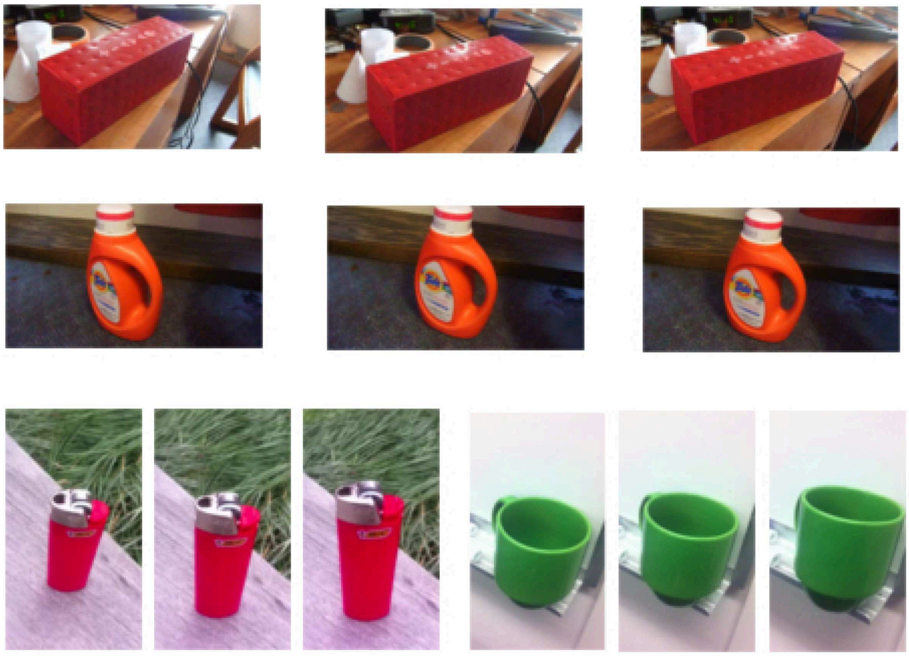


FIGURE 3 | Representative frames sampled from a collected video database of everyday objects undergoing 3D transformations (i.e., combination of translation, scaling, and in-depth rotation).

position and scale) is constrained by available physiology data (Serre et al., 2007a).

In the original model, the only learning that takes place is at the level of the dictionary of S_2 units. This is done via an *imprinting* learning rule whereby during the training procedure, units store patterns of neural activity associated with the presentation of patches of natural scenes that are presented in their receptive field (see Serre et al., 2007b, for details). More sophisticated algorithms have been proposed for learning intermediate visual features (e.g., Shams and von der Malsburg, 2002; Ullman et al., 2002; Masquelier and Thorpe, 2007; Hu et al., 2014). Here, without loss of generality, we used the simple imprinting learning rule to stay as faithful as possible to the original model but it is expected that other algorithms would yield qualitatively similar results.

2.3. Measuring NAP Selectivity

Here we conducted *in silico* experiments on the HMAX model with the aim to mimic the experimental methods described in the original studies (Kayaert et al., 2003; Amir et al., 2012) as closely as possible. The stimulus set consisted in the 36 basic shapes used in Kayaert et al. (2003). Each of the 36 stimuli exhibited five level of variations along a single dimension: four metric variations of increasing amplitude (denoted MP1–MP4) and one non-accidental variation (denoted NAP). The NAP variation was calibrated so that the resulting change from the base shape (measured by the euclidean distance directly on pixel intensities) was equal or less than the change associated with MP2. Sample stimuli are shown on **Figure 4**. The NAP/MP percent modulation for model units was computed using the same formula as described in the original study by Kayaert et al. (2003): (response basic shape—response to object variation)/(response basic shape)*100.

3. Results

We first reproduced the results by Amir et al. (2012) demonstrating that the original HMAX failed to exhibit a greater sensitivity for NAPs vs. MPs. We trained a baseline model with the object video dataset (Section 2; **Figure 3**). As in the original electrophysiology study, units were selected based on their visual responsiveness to the base images in the stimulus dataset used for electrophysiology (see Kayaert et al., 2003, for details) which yielded 243 NAP-MP comparisons. For each model unit, we computed the NAP and MP percent modulation for its preferred stimulus (Section 2). **Figure 5A** shows the MP percent modulation vs. NAP percent modulation for each unit in the original model. We found an average of 20% NAP modulation, compared to an average 22% MP modulation from the base object. A wilcoxon signed-rank test confirmed no significant NAP vs. MP modulation ($p = 0.76$). Overall, only 49% of the units had a greater NAP modulation, as compared to the 63% found in IT (Kayaert et al., 2003).

We then proceeded to extend the model to learning invariances from transformation sequences. In the original model, IT-like units are organized in feature columns whose

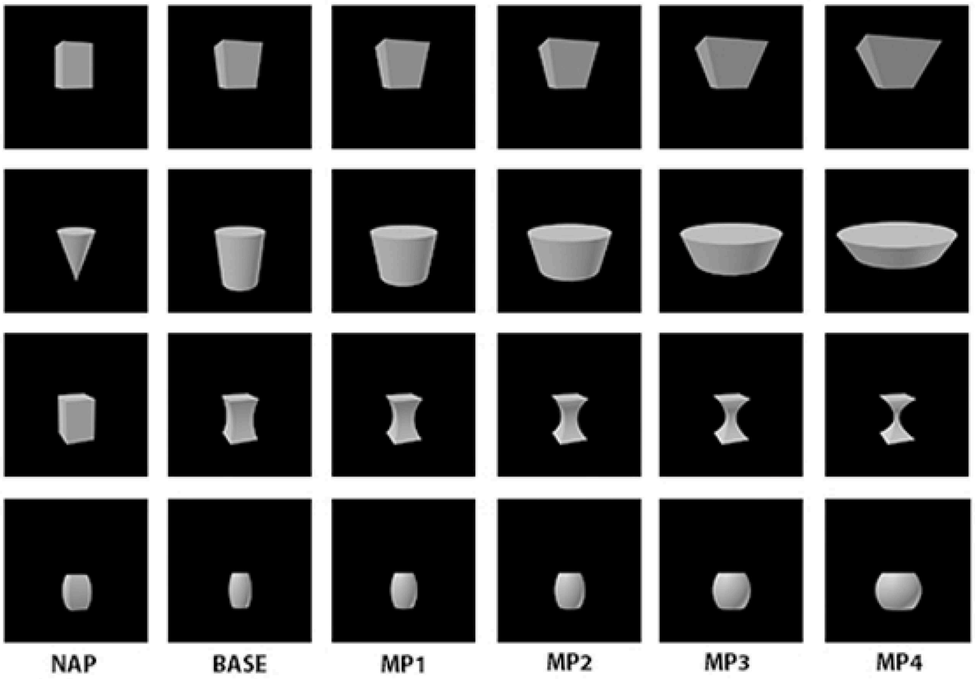


FIGURE 4 | Sample stimuli from the study by Kayaert et al. (2003). The column labeled BASE corresponds to a reference image. The column labeled NAP corresponds to a transformation of the base image where an NAP was changed. MP1, MP2, MP3, and MP4 correspond to a transformation of the base image with an MP of increasing magnitude.

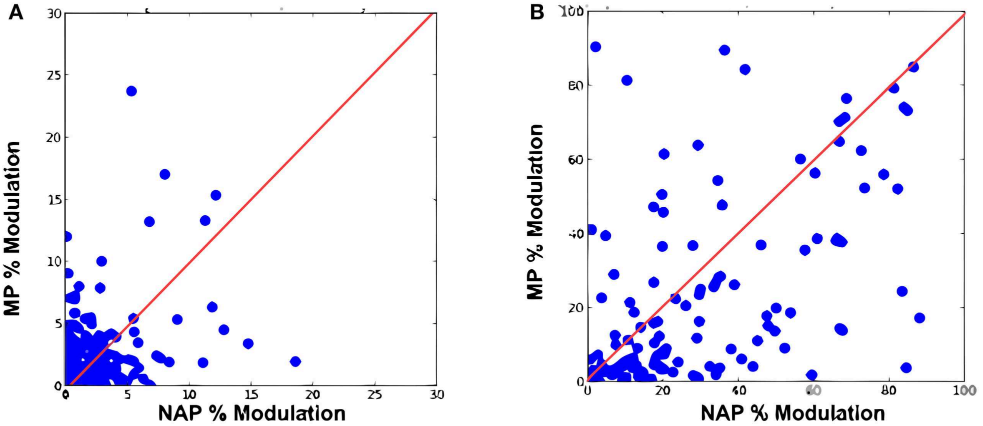


FIGURE 5 | Percent modulation of C_2 units from the base image to MP2 vs. percent modulation of C_2 units from the base image to NAP for (A) original HMAX (B) the extended model.

selectivity is determined by a simple imprinting learning rule (Section 2). Each feature column (C_2 unit) is hard-coded by considering afferent (S_2) units tuned to the same preferred feature but with receptive fields at different locations (and scales), yielding a visual representation which is tolerant to 2D transformations. However, no mechanism for invariance to 3D transformations is present in the original model yielding a "salt-and-pepper" organization of feature columns for changes in viewpoint (**Figure 2A**).

Here, we extended the invariance pooling mechanism to also include different views of a feature undergoing a 3D transformation during a relatively small (~300 ms) time window. This was done by considering feature columns which include multiple units with a selectivity for different views of the same feature occurring in close temporal proximity.

Visual responsiveness for this new set of C_2 model units was assessed as for the original model which yielded 159 NAP-MP comparisons. As shown on **Figure 5B**, this model extension yielded a dramatic increase in NAP vs. MP modulation with an average 35% NAP modulation vs. a 24% MP modulation. A Wilcoxon test showed a significant modulation for NAP vs. MP ($p < 0.01$). We further observed that 71% of the new model units were now more strongly modulated by a change in NAP vs. MP. As seen in **Figure 5B**, the majority of data points now fell below the diagonal, illustrating a greater sensitivity to NAP change. **Table 1** summarizes these findings and provides a comparison to IT data reported in Kayaert et al. (2003).

Interestingly, we also found that learning transformation sequences yielded a significant improvement in object recognition classification accuracy over changes in viewpoint. We used the scikit-learn toolbox (Pedregosa et al., 2011) to train and test a multi-class linear SVM on the original and extended model outputs using a random split procedure of the video dataset ($n = 15$). The regularization parameter was optimized using a cross-validation procedure. We found an overall significantly higher accuracy for the extended model (95.2 ± 2.1%, chance level: 8.3%) vs. the original model (85.6 ± 1.8%, $p < 0.01$) suggesting that the proposed unsupervised invariance learning algorithm does indeed yield a model with greater generalization to changes in viewpoint.

TABLE 1 | Comparison between IT Data (Kayaert et al., 2003), the original as well as the extended HMAX.

	% NAP Modulation from base	% MP Modulation from base	Sample size	Wilcoxon *p*-value	% units NAP>MP Modulation
IT Data	33	21–26	$n = 243$	$p < 2e\text{-}06$	63
Original HMAX	20	22	$n = 243$	$p = 0.7645$	49
Extended HMAX	35	24	$n = 159$	$p = 1.2e\text{-}05$	71

Sample sizes correspond to the number of NAP-MP comparisons as done in the original study.

4. Discussion

We have described a simple extension of a hierarchical model of object recognition (HMAX) which enables the network to learn transformation sequences. The original model includes mechanisms for building tolerance to 2D transformations (position and scale). We have shown that the proposed extension yields a model with better generalization capability for more complex transformation sequences which also include 3D rotations. Most importantly, we have shown that the resulting model exhibits greater sensitivity for NAPs vs. MPs in better agreement with IT data (Kayaert et al., 2003).

While our study has focused on the HMAX model, we expect our main results to apply broadly to the general class of feedforward hierarchical models (see Serre, 2014, for review). Despite differences in their specific wiring and detailed architecture, tolerance to object transformations in these models arise from Hubel-Wiesel types of pooling mechanisms and we thus expect our results to generalize to this broad class of models. Similarly, we also expect different learning rules to yield qualitatively similar results. While the present learning rule yielded NAP modulation in excellent agreement with IT data, it remains an open question whether other learning rules would provide similar or better fit to data.

Overall, our study suggests that the greater sensitivity for NAPs over MPs, as reported in several behavioral and electrophysiological studies (see Biederman, 2007, for review) may be driven by computational mechanisms for invariant object recognition.

Author Contributions

SP and TS conceived the research. SP performed the research. SP and TS wrote the manuscript and approved the final version for submission.

Funding

This work was supported by DARPA young faculty award [grant number YFA N66001-14-1-4037] and NSF early career award [grant number IIS-1252951]. Additional support was provided by ONR [grant number N000141110743].

Acknowledgments

We would like to thank David Reichert for his initial contribution to the early phase of this work. This work appeared in abstract form as (Parker S., Reichert D., and Serre, T., Selectivity for non-accidental properties emerges from learning object transformation sequences; Abstract 52.24. Presented at the Vision Science Society, May 20, 2014, St. Pete Beach, Florida).

References

Abecassis, M., Sera, M. D., Yonas, A., and Schwade, J. (2001). What's in a shape? children represent shape variability differently than adults when naming objects. *J. Exp. Child Psychol.* 78, 213–239. doi: 10.1006/jecp.2000.2573

Amir, O., Biederman, I., and Hayworth, K. J. (2011). The neural basis for shape preferences. *Vis. Res.* 51, 2198–2206. doi: 10.1016/j.visres.2011.08.015

Amir, O., Biederman, I., and Hayworth, K. J. (2012). Sensitivity to nonaccidental properties across various shape dimensions. *Vis. Res.* 62, 35–43. doi: 10.1016/j.visres.2012.03.020

Amir, O., Biederman, I., Herald, S. B., Shah, M. P., and Mintz, T. H. (2014). Greater sensitivity to nonaccidental than metric shape properties in preschool children. *Vis. Res.* 97, 83–88. doi: 10.1016/j.visres.2014.02.006

Amit, Y., and Mascaro, M. (2003). An integrated network for invariant visual detection and recognition. *Vis. Res.* 43, 2073–2088. doi: 10.1016/S0042-6989(03)00306-7

Biederman, I. (1987). Recognition-by-components: a theory of human image understanding. *Psychol. Rev.* 94, 115–147. doi: 10.1037/0033-295X.94.2.115

Biederman, I. (2007). Recent psychophysical and neural research in shape recognition. *Object Recognit. Atten. Action*. doi: 10.1007/978-4-431-73019-4/6

Biederman, I., and Bar, M. (1999). One-shot viewpoint invariance in matching novel objects. *Vis. Res.* 39, 2885–2899. doi: 10.1016/S0042-6989(98)00309-5

Cox, D. D., Meier, P., Oertelt, N., and DiCarlo, J. J. (2005). 'Breaking' position-invariant object recognition. *Nat. Neurosci.* 8, 1145–1147. doi: 10.1038/nn1519

Einhäuser, W., Kayser, C., König, P., and Körding, K. P. (2002). Learning the invariance properties of complex cells from their responses to natural stimuli. *Eur. J. Neurosci.* 15, 475–486. doi: 10.1046/j.0953-816x.2001.01885.x

Foldiak, P. (1991). Learning invariance from transformation sequences. *Neural Comput.* 3, 194–200. doi: 10.1162/neco.1991.3.2.194

Földiák, P. (1998). "Learning constancies for object perception," in *Perceptual Constancy: Why Things Look as They Do*, eds V. Walsh and J. J. Kulikowski (Cambridge, UK: Cambridge University Press), 144–172.

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. doi: 10.1007/BF00344251

Hawkins, J., and Blakeslee, S. (2004). *On Intelligence*. New York, NY: Henry Holt and Company, LLC.

Hietanen, J. K., Perrett, D. I., Oram, M. W., Benson, P. J., and Dittrich, W. H. (1992). The effects of lighting conditions on responses of cells selective for face views in the macaque temporal cortex. *Exp. Brain Res.* 89, 157–171. doi: 10.1007/BF00229013

Hu, X., Zhang, J., Li, J., and Zhang, B. (2014). Sparsity-regularized HMAX for visual recognition. *PLoS ONE* 9:e81813. doi: 10.1371/journal.pone.0081813

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. doi: 10.1113/jphysiol.1962.sp006837

Jarrett, K., Kavukcuoglu, K., Ranzato, M., and LeCun, Y., (2009). "What is the best multi-stage architecture for object recognition?," in *Computer Vision, 2009 IEEE 12th International Conference on* (Kyoto), 2146–2153. doi: 10.1109/ICCV.2009.5459469

Kayaert, G., Biederman, I., and Vogels, R. (2003). Shape tuning in macaque inferior temporal cortex. *J. Neurosci.* 23, 3016–3027.

Lades, M., Vorbruggen, J. C., Buhmann, J., Lange, J., Von der Malsburg, C., Wurtz, R. P., et al. (1993). "Distortion invariant object recognition in the dynamic link architecture," in *Computers, IEEE Transactions on*, Vol. 42 (Washington, DC), 300–311. doi: 10.1109/12.210173

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 2278–2324. doi: 10.1109/5.726791

Li, L., Qin, W., Bai, L., and Tian, J. (2010). Exploring vision-related acupuncture point specificity with multivoxel pattern analysis. *Magn. Reson. Imaging* 28, 380–387. doi: 10.1016/j.mri.2009.11.009

Li, Y., Van Hooser, S. D., Mazurek, M., White, L. E., and Fitzpatrick, D. (2008). Experience with moving visual stimuli drives the early development of cortical direction selectivity. *Nature* 456, 952–956. doi: 10.1038/nature07417

Lowe, D. G. (1984). *Perceptual Organization and Visual Recognition*. Ph.D. thesis, Department of Computer Science, Stanford University, Stanford, CA.

Marr, D. (1982). *Vision: A Computational Investigation Into the Human Representation and Processing of Visual Information*. San Francisco, CA: W.H.Freeman & Co Ltd.

Masquelier, T., Serre, T., and Poggio, T. (2007). Learning complex cell invariance from natural videos : a plausibility proof. Technical report, Massachusetts Institute of Technology, Cambridge MA.

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

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. doi: 10.1162/neco.1997.9.4.777

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. doi: 10.1007/s11263-007-0118-0

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* 12, 2825–2830.

Perrett, D. I., Smith, P. A., Potter, D. D., Mistlin, A. J., Head, A. S., Milner, A. D., et al. (1984). Neurones responsive to faces in the temporal cortex: studies of functional organization, sensitivity to identity and relation to perception. *Hum. Neurobiol.* 3, 197–208.

Pinto, N., Barhomi, Y., Cox, D. D., and DiCarlo, J. J. (2011). "Comparing state-of-the-art visual features on invariant object recognition tasks," in *Applications of Computer Vision (WACV), 2011 IEEE Workshop on* (Kona, HI), 463–470. doi: 10.1109/WACV.2011.5711540

Richards, W., Jepson, A., and Feldman, J. (1996). "Priors, preferences and categorical percepts," in *Perception as Bayesian Inference*, ed D. C. Knill and W. Richards (New York, NY: Cambridge University Press), 93–122.

Riesenhuber, M., and Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nat. Neurosci.* 2, 1019–1025. doi: 10.1038/14819

Riesenhuber, M., and Poggio, T. (2000). Models of object recognition. *Nat. Neurosci.* 3, 1199–1204. doi: 10.1038/81479

Saxe, A., Koh, P., Chen, Z., Bhand, M., Suresh, B., and Ng, A. (2011). "On random weights and unsupervised feature learning," in *Proceedings of the 28th International Conference on Machine Learning* (Bellevue, WA).

Serre, T. (2014). *Hierarchical Models of the Visual System*. New York, NY: Springer Science+Business Media. doi: 10.1007/978-1-4614-7320-6_345-1

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. doi: 10.1073/pnas.0700622104

Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., and Poggio, T. (2007b). Robust object recognition with cortex-like mechanisms. *IEEE Trans. Pattern Anal. Mach. Intell.* 29, 411–426. doi: 10.1109/TPAMI.2007.56

Shams, L., and von der Malsburg, C. (2002). Acquisition of visual shape primitives. *Vis. Res.* 42, 2105–2122. doi: 10.1016/S0042-6989(02)00130-X

Spratling, M. W. (2005). Learning view-point invariant perceptual representations from cluttered images. *IEEE Trans. Pattern Anal. Mach. Intell.* 27, 753–761. doi: 10.1109/TPAMI.2005.105

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. doi: 10.1007/s00422-005-0030-z

Tanaka, K. (2003). Columns for complex visual object features in the inferotemporal cortex: clustering of cells with similar but slightly different stimulus selectivities. *Cereb. Cortex* 13, 90–99. doi: 10.1093/cercor/13.1.90

Tarr, M. J., and Kriegman, D. J. (2001). What defines a view? *Vis. Res.* 41, 1981–2004. doi: 10.1016/S0042-6989(01)00024-4

Thorpe, S., Fize, D., and Marlot, C. (1996). Speed of processing in the human visual system. *Nature* 381, 520–522. doi: 10.1038/381520a0

Ullman, S., Vidal-Naquet, M., and Sali, E. (2002). Visual features of intermediate complexity and their use in classification. *Nat. Neurosci.* 5, 682–687. doi: 10.1038/nn870

Wallis, G., and Bülthoff, H. H. (2001). Effects of temporal association on recognition memory. *Proc. Natl. Acad. Sci. U.S.A.* 98, 4800–4804. doi: 10.1073/pnas.071028598

Wallis, G., and Rolls, E. T. (1997). A model of invariant object recognition in the visual system. *Prog. Neurobiol.* 51, 167–194. doi: 10.1016/S0301-0082(96)00054-8

Wallis, G., Rolls, E. T., and Földiák, P. (1993). "Learning invariant responses to the natural transformations of objets," in *Proceedings of the International Joint Conference on Neural Networks*, Vol. 2 (Nagoya), 1087–1090. doi: 10.1109/IJCNN.1993.716702

Wersing, H., and Köerner, E. (2003). Learning optimized features for hierarchical models of invariant recognition. *Neural Comp.* 15, 1559–1588. doi: 10.1162/089976603321891800

Wiskott, L., and Sejnowski, T. J. (2002). Slow feature analysis: unsupervised learning of invariances. *Neural. Comput.* 14, 715–770. doi: 10.1162/089976602317318938

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.

frontiers in Computational Neuroscience

ORIGINAL RESEARCH
published: 12 August 2015
doi: 10.3389/fncom.2015.00104

Diversity priors for learning early visual features

*Hanchen Xiong *, Antonio J. Rodríguez-Sánchez, Sandor Szedmak and Justus Piater*

Intelligent and Interactive Systems Group, Institute of Computer Science, University of Innsbruck, Innsbruck, Austria

This paper investigates how utilizing diversity priors can discover early visual features that resemble their biological counterparts. The study is mainly motivated by the sparsity and selectivity of activations of visual neurons in area V1. Most previous work on computational modeling emphasizes selectivity or sparsity independently. However, we argue that selectivity and sparsity are just two epiphenomena of the diversity of receptive fields, which has been rarely exploited in learning. In this paper, to verify our hypothesis, restricted Boltzmann machines (RBMs) are employed to learn early visual features by modeling the statistics of natural images. Considering RBMs as neural networks, the receptive fields of neurons are formed by the inter-weights between hidden and visible nodes. Due to the conditional independence in RBMs, there is no mechanism to coordinate the activations of individual neurons or the whole population. A diversity prior is introduced in this paper for training RBMs. We find that the diversity prior indeed can assure simultaneously sparsity and selectivity of neuron activations. The learned receptive fields yield a high degree of biological similarity in comparison to physiological data. Also, corresponding visual features display a good generative capability in image reconstruction.

Keywords: restricted Boltzmann machine, diversity prior, V1 simple cell, inhibition, Markov networks

OPEN ACCESS

Edited by:
Judith Peters,
Netherlands Institute for Neuroscience - Royal Netherlands Academy of Arts and Sciences, Netherlands

Reviewed by:
Joel Zylberberg,
University of Washington, USA
Rajat Mani Thomas,
Netherlands Institute for Neuroscience - Royal Netherlands Academy of Arts and Sciences, Netherlands

****Correspondence:***
Hanchen Xiong,
Institute of Computer Science,
University of Innsbruck,
Technikerstrasse 21a, A-6020
Innsbruck, Austria
hanchen.xiong@uibk.ac.at

Received: *30 March 2015*
Accepted: *27 July 2015*
Published: *12 August 2015*

Citation:
Xiong H, Rodríguez-Sánchez AJ, Szedmak S and Piater J (2015) Diversity priors for learning early visual features.
Front. Comput. Neurosci. 9:104.
doi: 10.3389/fncom.2015.00104

1. Introduction

Much has been advanced in the knowledge of the brain in the last century since the foundation of modern neuroanatomy by Ramón y Cajal (Ramón y Cajal, 1888, 1904; Jones, 2007). The work of HUBEL and WIESEL (1959) was the first breakthrough in the understanding of simple cells in area V1 of the visual cortex. V1 simple cells perform an early stage processing of the visual input from the retina and the lateral geniculate nucleus (LGN). One important property of V1 simple cells is that their receptive fields are *selective* in terms of location, orientation, and frequency, which can be modeled by Gabor filters. Another characteristic on V1 simple cells is that their activation pattern—when analyzed as a population—is sparse (Field, 1994). Selectivity (also referred to as "lifetime sparseness" by Willmore and Tolhurst, 2001) is related to a neuron having a response only to a small number of different (although similar) stimuli and providing a much lower response to other (usually very different) stimuli. Sparsity (or "population sparseness" by Willmore and Tolhurst, 2001) is a term expressing that the fraction of neurons from a population that is activated by a certain stimulus should be relatively small. Selectivity and sparsity would be due to a redundancy-reduction mechanism, where the visual cortex has evolved to encode visual information as efficiently as possible (Barlow, 1989). This *sparse coding* would then enhance coding efficiency, and when tested, leads in fact to Gabor-like representations (Olshausen and Field, 1996). Although sparse coding has been very successful at generating receptive fields similar to those of simple cells, sparsity does not necessarily imply selectivity (Willmore and Tolhurst, 2001). In addition to this, recent multi-unit neurophysiological recordings found that just maximizing sparsity does not correlate with visual experience, suggesting that coding efficiency is also due to lateral, recurrent and feedback connections for the purpose of resolving ambiguities (Berkes et al., 2009). In order to show the (lack of) relationship between sparsity and selectivity, we illustrate these concepts in **Figure 1A**. Each row (red) in this figure represents how one neuron selectively responds to different visual stimuli while each column (blue) describes how many neurons are activated by one stimulus. Although selectivity and sparsity can be related at their average values, they are not necessarily correlated: Selective neurons do not ensure sparse neuron coding (**Figure 1C**); similarly, sparsely activated neurons are not necessarily narrowly selective (**Figure 1D**).

Another hypothesis on how to achieve coding efficiency is dependence minimization, which can be achieved applying *independent component analysis* (ICA) (Hyvärinen and Oja, 2000). ICA is a dimensionality reduction methodology widely used in signal processing for decomposing a compound signal into their components (or so-called bases) that are as independent as possible. In ICA, independence maximization is achieved by pursuing extrema of the kurtosis (a measure of function "peakedness") of each components' distribution. Applying ICA on natural images has also produced receptive fields like those of V1 simple cells (Bell and Sejnowski, 1997; van Hateren and van der Schaaf, 1998). Be either ICA or sparse coding, in the end, they are two successful learning strategies that can learn primary visual cortex-like receptive fields (Olshausen and Field, 1996; van Hateren and van der Schaaf, 1998). Another successful learning strategy at emulating the hierarchical architecture of the brain is *deep learning* (Bengio, 2009; LeCun et al., 2015), which is usually constructed with a stack of restricted Boltzmann machines (RBMs). RBMs have recently attracted increasing attention due to its successes in learning representations (Hinton, 2002; Hinton and Salakhutdinov, 2006). In RBMs, there is no connection among hidden units (**Figure 2**), which makes inference and learning of RBMs quite easy and fast. That means that given some visible data, all hidden units are conditionally independent from each other (see Section 2.2). Even so, RBMs provide a nonlinear coding of natural images, which goes beyond sparse coding or ICA. However, the capability of RBMs is still limited when learning receptive fields similar to those of V1 simple cells. When RBMs are trained on natural images, many learned features can be rather

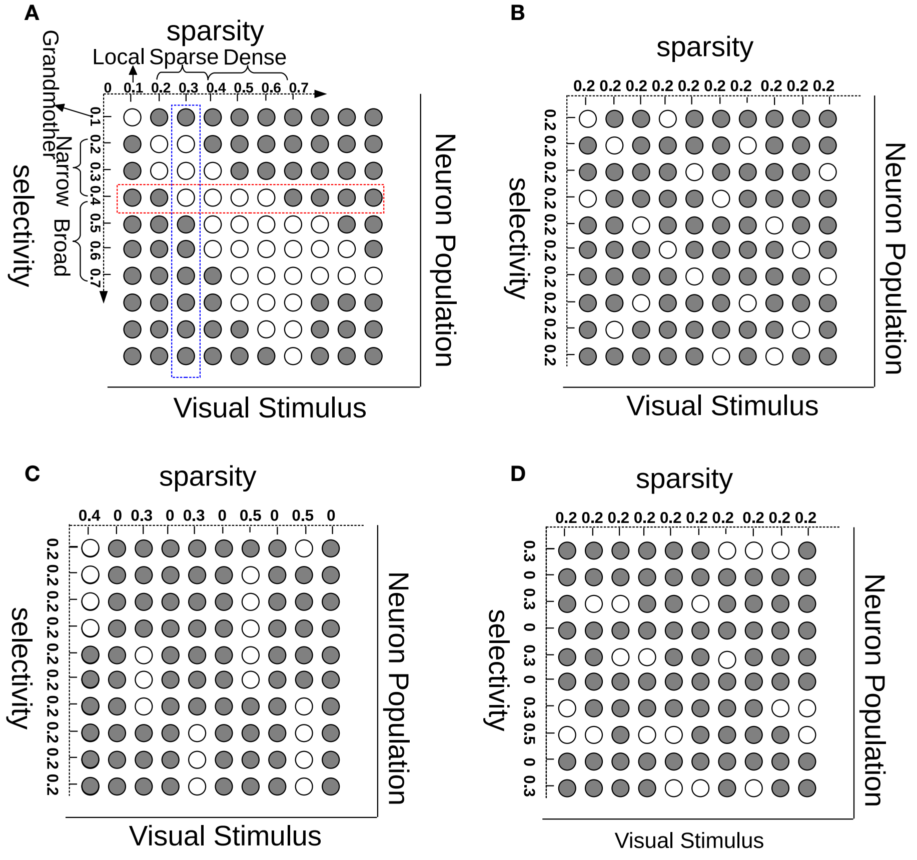


FIGURE 1 | Understanding sparsity and selectivity. White circles indicate activations while gray circles denote inactivations. **(A)** Explaining the concepts of sparsity and selectivity; **(B)** An example of good sparsity and good selectivity; **(C)** An example of good selectivity but bad sparsity; **(D)** An example of good sparsity but bad selectivity. See text for further description.

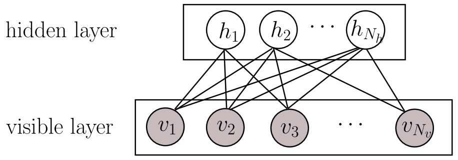


FIGURE 2 | A graphical model of a restricted Boltzmann machine (RBM). Gray circles represent observed variables while empty circles are hidden variables.

distributed, unlocalized and repeated, which is far from the (selective and sparse) nature of the learning task. Prior work has exploited different strategies to adapt RBMs toward learning selective or sparsely-activated neurons (Lee et al., 2007; Goh et al., 2010; Luo et al., 2011) on visual inputs. Meanwhile, most of those works focus on either one property, thus not ensuring sparsity and selectivity simultaneously in the resulting emulated neurons, which as mentioned before may be suboptimal for coding efficiency.

Empirically, neither sparse coding nor ICA can yield both, good selectivity and sparsity simultaneously (Willmore and Tolhurst, 2001). In this paper, we propose a novel hypothesis to interpret the selectivity and sparsity of neuron activations through the *diversity of neurons' receptive fields*. Based on the analysis exposed above, we can see that the effect of sparsity is to better differentiate neurons, while the goal of selectivity is to avoid "over-tolerant" neurons, thus both aimed at reducing ambiguities. We propose that, in order to reach bot—high degrees of neural population sparsity and individual neuronal selectivity—we need one condition: diverse receptive fields. To the best of our knowledge, the diversity of receptive fields (features) has rarely been exploited to guide learning, even though it has been achieved unintentionally in several existing models. By contrast to conventional models, we use diversity as a starting point instead of as a result. An earlier pioneering work focusing on the importance of diversity in neural coding was presented by Padmanabhan and Urban (2010).

We argue that selectivity and sparsity of neurons' activations can be seen as two epiphenomena of the diversity of receptive fields. To verify this hypothesis, we impose a *diversity prior* on the inter-weights within the RBMs when learning simple neurons' receptive fields from natural images. This prior will introduce a bias over the inter-weights toward higher degrees of *sum similarity minimization*. The prior indirectly coordinates neurons' activations by diversifying the inter-weights within the RBMs, which would mimic the effect of inhibition. It is worth noting that the prior is only employed in the learning phase, yet its implicit effect on coordinating neurons' activations will remain after learning. In this sense, the diversity prior is in line with the influence of inhibitory interneurons (King et al., 2013) (see Section 2.3 for more details). It should be finally noted that we do not consider an RBM (even if trained with diversity priors) as a full biologically-plausible model of V1 simple cells, since we are not considering many other aspects and properties of simple cells, e.g., contrast normalization, contrast adaption, etc. The purpose of our study is to verify and advocate for using *diversity* as a new principle in order to guide the learning of more similar primary visual cortex cell receptive fields.

2. Materials and Methods

In this section, we describe our basic experimental setup, which includes the construction of visual stimulus data, the restricted Boltzmann machine (RBM), and the proposed prior for training. For the RBM, a brief introduction of the model and its probabilistic properties is provided in Section 2.2. Readers are referred to Hinton (2002) for a more detailed and deeper study.

2.1. Images

The benchmark database from Olshausen and Field (1996)[1] was used in this paper. This database consists of 10 natural images, which were preprocessed with a pseudo-whitening filter, which flattens the spectrum of natural images by rescaling Fourier coefficients. This step is commonly applied (Olshausen and Field, 1996; Willmore and Tolhurst, 2001), and to some extent is similar to retinal processing. Alternatively, a similar preprocessing function is the log transform, which is more often used in ICA (van Hateren and van der Schaaf, 1998). Then, 100,000 small patches (size 14 × 14) were extracted from random positions of the 10 whitened images. Furthermore, a sigmoid function was applied to the pixel intensities to fit their values into the range [0, 1]. In addition, the patches with variances smaller than 0.1 were filtered out in order to accelerate training.

2.2. Restricted Boltzmann Machines

The restricted Boltzmann machine (RBM) is a two-layer, bipartite Markov network, which is a "restricted version" of the Boltzmann machine with only inter-connections between a hidden layer and a visible layer. RBMs have been recently rather popular in constructing deep neural networks (DNNs) (Hinton and Salakhutdinov, 2006). A graphical model of an RBM is presented in **Figure 2**. Input data is binary and N_v dimensional; they are fed into N_v units in the visible layer $\mathbf{v}$. The N_h units in the hidden layer $\mathbf{h}$ are stochastic binary variables, i.e., $\mathbf{v} \in \{0, 1\}^{N_v}$, $\mathbf{h} \in \{0, 1\}^{N_h}$. The joint probability of $\{\mathbf{v}, \mathbf{h}\}$ is:

$$p(\mathbf{v}, \mathbf{h}) = \frac{1}{\mathbf{Z}} \exp(-E(\mathbf{v}, \mathbf{h})) \qquad E(\mathbf{v}, \mathbf{h}) = -\mathbf{v}^\top \mathbf{W} \mathbf{h} - \mathbf{h}^\top \mathbf{b} - \mathbf{v}^\top \mathbf{c} \tag{1}$$

where $\mathbf{W} \in \mathbb{R}^{N_v \times N_h}$ is the matrix of symmetric weights, $\mathbf{b} \in \mathbb{R}^{N_h \times 1}$ and $\mathbf{c} \in \mathbb{R}^{N_v \times 1}$ are biases for hidden units and visible units, respectively. $\mathbf{Z} = \sum_{\mathbf{v},\mathbf{h}} \exp(-E(\mathbf{v}, \mathbf{h}))$ is the partition function for normalization. In our experiment, to fit the size of small image patches, N_v is equivalent to 196, and N_h is 200, i.e., 200 hidden units. Because of the restricted connections in RBMs, hidden units h_j are conditionally independent of each other given the visible data $\mathbf{v}$,

$$p(\mathbf{h}|\mathbf{v}) = \prod_j p(h_j|\mathbf{v}) \qquad p(h_j = 1|\mathbf{v}) = \mathcal{S}(\mathbf{v}^\top \mathbf{W}_{\cdot j} + b_j) \tag{2}$$

[1] Available on http://redwood.berkeley.edu/bruno/sparsenet/.

and similarly, visible units v_i are conditionally independent of each other given $\mathbf{h}$.

$$p(\mathbf{v}|\mathbf{h}) = \prod_i p(v_i|\mathbf{h}) \qquad p(v_i = 1|\mathbf{h}) = \mathcal{S}(\mathbf{W}_{i\cdot}\mathbf{h} + c_i) \quad (3)$$

where $\mathbf{W}_{i\cdot}$ and $\mathbf{W}_{\cdot j}$ denote the ith row and jth column of matrix $\mathbf{W}$, b_j and c_i are the jth and ith entry of vector $\mathbf{b}$ and $\mathbf{c}$, respectively. $\mathcal{S}(\cdot)$ is the logistic function $\mathcal{S}(x) = \frac{1}{1+\exp(-x)}$. Given training data $\mathcal{D} = \{\mathbf{v}^{(l)}\}_{l=1}^{L}$, an RBM can be learned by maximizing the average log-likelihood of $\mathcal{D}$:

$$\mathbf{W}^* = \arg\max_{\mathbf{W}} \mathcal{L}(\mathcal{D}) = \arg\max_{\mathbf{W}} \frac{1}{L}\sum_{l=1}^{L}\left(\log\sum_{\mathbf{h}} p(\mathbf{v}^{(l)}, \mathbf{h})\right) \quad (4)$$

Since the log-likelihood is concave with respect to $\mathbf{W}, \mathbf{b}, \mathbf{c}$ (Koller and Friedman, 2009, Chapter 20), based on Equation (1), *gradient ascent* can be applied on Equation (4) by computing the gradient of $\mathcal{L}(\mathcal{D})$ with respect to $\mathbf{W}, \mathbf{b}, \mathbf{c}$ as:

$$\nabla_{\mathbf{W}}\mathcal{L}(\mathcal{D}) = \frac{1}{L}\sum_{l=1}^{L}\left[\mathbb{E}_{\mathbf{v}^{(l)}\in\mathcal{D}, \mathbf{h}\sim p(\mathbf{h}|\mathbf{v}^{(l)})}(\mathbf{v}^{(l)}\mathbf{h}^{\top}) - \mathbb{E}_{\mathbf{v},\mathbf{h}\sim p(\mathbf{v},\mathbf{h})}(\mathbf{v}\mathbf{h}^{\top})\right] \quad (5)$$

$$\nabla_{\mathbf{b}}\mathcal{L}(\mathcal{D}) = \frac{1}{L}\sum_{l=1}^{L}\left[\mathbb{E}_{\mathbf{h}\sim p(\mathbf{h}|\mathbf{v}^{(l)})}(\mathbf{h}) - \mathbb{E}_{\mathbf{h}\sim p(\mathbf{v},\mathbf{h})}(\mathbf{h})\right] \quad (6)$$

$$\nabla_{\mathbf{c}}\mathcal{L}(\rfloor) = \frac{1}{L}\sum_{l=1}^{L}\left[\mathbb{E}_{\mathbf{v}^{(l)}\in\mathcal{D}}(\mathbf{v}^{(l)}) - \mathbb{E}_{\mathbf{v}\sim p(\mathbf{v},\mathbf{h})}(\mathbf{v})\right] \quad (7)$$

where $\mathbb{E}_p(\cdot)$ denotes the expected values with respect to p. Obviously, the first terms in Equations (5–7) are easy to compute with $\mathbf{v}^{(l)}$ from $\mathcal{D}$ and $\mathbf{h}$ inferred using Equation (2). However, the sampling $\mathbf{v}, \mathbf{h} \sim p(\mathbf{v}, \mathbf{h})$ in the second term of Equation (5) makes learning practically infeasible because it requires a large number of Markov chain Monte Carlo (MCMC) iterations to reach equilibrium. Fortunately, we can compute an efficient approximation to the exact gradient: contrastive divergence (CD), which works well in practice (Hinton and Salakhutdinov, 2006). By using CD$_k$, only a small number of k steps are run in block Gibbs sampling (usually $k = 1$), and Equation (5) can finally be approximated as

$$\nabla_{\mathbf{W}}\hat{\mathcal{L}}(\mathcal{D}) = \frac{1}{L}\sum_{l=1}^{L}\Big[\mathbf{v}^{(l)}p(\mathbf{h}^{(l)+}|\mathbf{v}^{(l)})^{\top} - p(\mathbf{v}^{(l)-}|\mathbf{h}^{(l)+}) \\ p(\mathbf{h}^{(l)-}|\mathbf{v}^{(l)-})^{\top}\Big] \quad (8)$$

$$\nabla_{\mathbf{b}}\hat{\mathcal{L}}(\mathcal{D}) = \frac{1}{L}\sum_{l=1}^{L}\left[p(\mathbf{h}^{(l)+}|\mathbf{v}^{(l)}) - p(\mathbf{h}^{(l)-}|\mathbf{v}^{(l)-})\right] \quad (9)$$

$$\nabla_{\mathbf{c}}\hat{\mathcal{L}}(\mathcal{D}) = \frac{1}{L}\sum_{l=1}^{L}\left[\mathbf{v}^{(l)} - p(\mathbf{v}^{(l)-}|\mathbf{h}^{(l)+})\right] \quad (10)$$

where $\mathbf{h}^{(l)+}$ denotes the inferred hidden vector from the lth observed data point $\mathbf{v}^{(l)}$ (using Equation 2), and $\mathbf{v}^{(l)-}, \mathbf{h}^{(l)-}$ are vectors after one-step block Gibbs sampling (using Equations 2, 3 and again Equation 2).

2.3. Imposing a Diversity Prior

In RBMs, columns of $\mathbf{W}$ are basis images, with which $\mathbf{v}$ can be reconstructed from $\mathbf{h}$. To some extent, they can also represent neurons' receptive fields. To this end, a natural choice of biasing parameters is to diversify the columns of $\mathbf{W}$ as much as possible. The way in which we approach diversification is minimizing *square cosine similarities* among columns of $\mathbf{W}$:

$$\arg\min_{\mathbf{W}} \sum_{j=1}^{N_h}\sum_{k\neq j}^{N_h}\left\|\frac{\mathbf{W}_{\cdot,j}^{\top}\mathbf{W}_{\cdot,k}}{||\mathbf{W}_{\cdot,j}||||\mathbf{W}_{\cdot,k}||}\right\|^2 \quad (11)$$

Note that the denominator in Equation (11) is necessary, because eliminating it will generate many "dead" neurons. This repulsive design among $\mathbf{W}_{\cdot,j}$ was also employed in the local competition algorithm (LCA) (Rozell et al., 2008). Zylberberg et al. (2011) also found that inhibition between two neurons are proportional to the similarity (measured by the vector dot product) between their receptive fields. Here, in order to gain a more clear understanding on how the diversity prior can replicate the effect of neural inhibition, an illustrating example is presented in **Figure 3**. In particular, for computing the gradient with respect to $\mathbf{W}$, Equation (8) needs to infer the activations of the hidden units. The prior, which can bias the columns of $\mathbf{W}$ toward a more diverse population will indirectly coordinate the activations by suppressing the emergence of similar receptive fields, and therefore leads to a similar effect neural inhibition has during learning. Also, the effect from the prior will remain after learning with the learned diverse $\mathbf{W}$. An extreme case is that the activation probabilities of neurons are exclusive to each other. Sparsity and selectivity are expected to be enhanced simultaneously by using this diversity-induced bias (Equation 11) (**Figure 1B**). We can define the prior probability distribution over parameters $p(\mathbf{W})$ as

$$p(\mathbf{W}) \propto \exp\left(-\lambda\cdot\sum_{j=1}^{N_h}\sum_{k\neq j}^{N_h}\left\|\frac{\mathbf{W}_{\cdot,j}^{\top}\mathbf{W}_{\cdot,k}}{||\mathbf{W}_{\cdot,j}||||\mathbf{W}_{\cdot,k}||}\right\|^2\right). \quad (12)$$

Then, the parameters can be estimated via maximum a posteriori (MAP):

$$\mathbf{W}^* = \arg\max_{\mathbf{W}} p(\mathbf{W}|\mathcal{D}) = \arg\max_{\mathbf{W}} p(\mathbf{W})\prod_{l=1}^{L}\sum_{\mathbf{h}} p(\mathbf{v}^{(l)}, \mathbf{h}|\mathbf{W}) \quad (13)$$

In our previous work (Xiong et al., 2014), we used absolute cosine similarities, of which the derivative cannot be analytically computed and therefore we had to resort to MCMC-based simulated annealing to conduct MAP. However, here by using the square cosine similarity, Equation (13) can be converted to a constrained concave optimization:

$$\mathbf{W}^* = \arg\max_{\mathbf{W}} \mathcal{L}(\mathcal{D}) - \lambda\sum_{j=1}^{N_h}\sum_{k\neq j}^{N_h}(\mathbf{W}_{\cdot,j}^{\top}\mathbf{W}_{\cdot,k})^2 \quad (14)$$

$$s.t. \qquad \forall j \in [1, N_h], ||\mathbf{W}_{\cdot,j}|| = 1$$

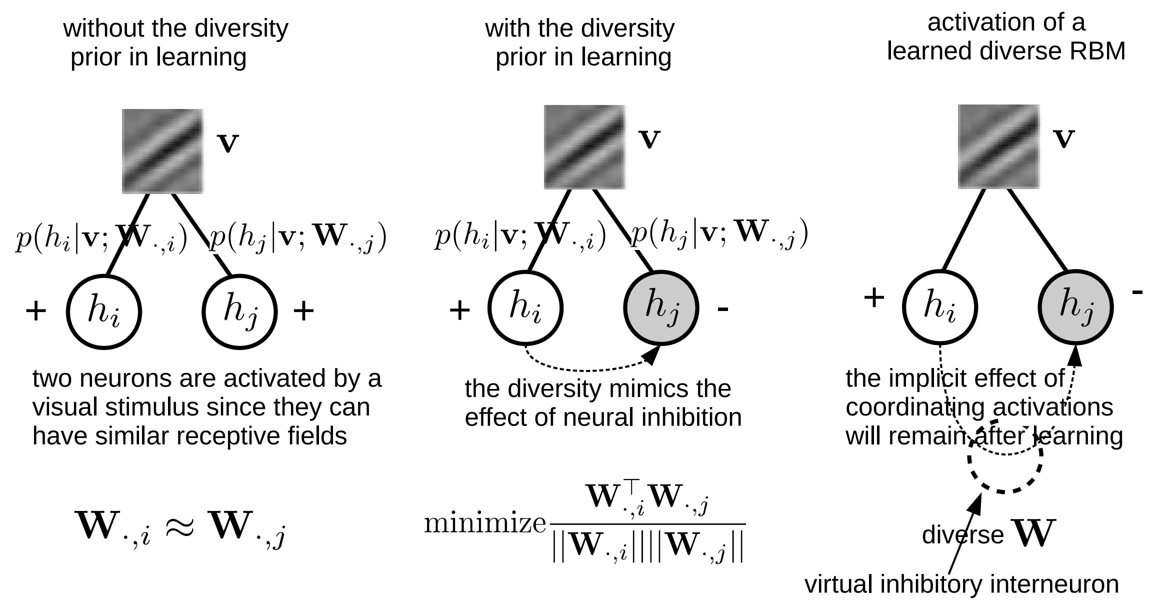

without the diversity prior in learning
with the diversity prior in learning
activation of a learned diverse RBM
$\mathbf{v}$
$p(h_i|\mathbf{v};\mathbf{W}_{\cdot,i})$
$p(h_j|\mathbf{v};\mathbf{W}_{\cdot,j})$
+
h_i
h_j
-
two neurons are activated by a visual stimulus since they can have similar receptive fields
the diversity mimics the effect of neural inhibition
the implicit effect of coordinating activations will remain after learning
$\mathbf{W}_{\cdot,i} \approx \mathbf{W}_{\cdot,j}$
$\text{minimize}\frac{\mathbf{W}_{\cdot,i}^{\top}\mathbf{W}_{\cdot,j}}{||\mathbf{W}_{\cdot,i}||||\mathbf{W}_{\cdot,j}||}$
diverse $\mathbf{W}$
virtual inhibitory interneuron

FIGURE 3 | Left, Middle: An illustrating example shows how the diversity prior would mimic the effect of inhibition among neurons' activations during learning. Empty circles denote activated neurons while gray circles are inactivated ones. **Right**: although the diversity prior is only employed in the learning phase, its implicit effect on coordinating neurons' activations will remain after learning. Intuitively, it can be considered as if there would exist virtual inhibitory interneurons which are induced by the diversity prior.

In this paper, since the above optimization problem is concave with respect to **W**, we employed gradient ascent to solve it (see the Appendix for details), and derived an iterative update of **W** as

$$\mathbf{W}_{\cdot,j}^{t+1} = \mathbf{W}_{\cdot,j}^{t} + \nabla_{\mathbf{W}}\hat{\mathcal{L}}(\mathcal{D}) - 2\lambda\left(\sum_{k\neq j}^{N_h}(\mathbf{W}_{\cdot,k}\otimes\mathbf{W}_{\cdot,k}) + C\frac{||\mathbf{W}_{\cdot,j}||-1}{||\mathbf{W}_{\cdot,j}||}\mathbf{I}_{N_v}\right)\mathbf{W}_{\cdot,j} \quad (15)$$

where $\otimes$ denotes the outer product between vectors, and $\mathbf{I}_{N_v}$ is a $N_v \times N_v$ identity matrix. In Equation (15) we can see that the iterative update of **W** is composed of two parts, where the first is the gradient of the log-likelihood while the second is the gradient of the log prior.

3. Results

In this section the learned receptive fields are shown, with which we measure the selectivity and sparsity of neurons' activations. We also compare the learned receptive fields with physiological data. Finally, we test the learned receptive fields in an image reconstruction experiment. The training dataset, the code of learning RBM, the learned diverse RBM and other materials used in our experiments are available at: https://iis.uibk.ac.at/public/xiong/resources.html#Diverse_RBM. Following Hinton (2002), we conducted training on mini-batches at one epoch. In all 400 epochs were run and it takes around 18 h with our Matlab code on an Intel core i7 laptop.

3.1. Basis Images

In **Figure 4**, a subset of basis images (i.e., columns of **W**) of RBMs trained with the diversity prior are shown. They look quite similar to the receptive fields of simple cells in macaque monkey V1 (Zylberberg et al., 2011, **Figure 3**). Rigorously speaking, basis images cannot be directly considered as receptive fields since they are internal connections or representations instead of response characteristics. The receptive fields of ICA are usually estimated as the inverse of the weight matrix (van Hateren and van der Schaaf, 1998), while in sparse coding reverse correlation is used for receptive fields (Olshausen and Field, 1996). Here, we employed a reverse correlation method similar to Hosoya (2012) who also developed a probabilistic model. For each hidden unit, its receptive field is estimated as

$$RF = \sum_{s=1}^{S} p(h_j = 1|\mathbf{v}_s)\mathbf{v}_s, \quad (16)$$

where $p(h_j = 1|\mathbf{v}_s)$ is computed as in Equation (2), while $\{\mathbf{v}_s\}_{s=1}^{S}$ is a set of visual stimuli which are randomly selected in the training database. This is a little different from the procedure by Hosoya (2012), since they generate synthetic $\mathbf{v}_s$ from a Gaussian distribution. Meanwhile, we arrived at a finding similar to Hosoya (2012). By linearly fitting the *RF* of each unit to its corresponding basis images, we found that our basis images are almost identical to their corresponding receptive fields.

3.2. Selectivity and Sparsity

There exist several ways to measure selectivity and sparsity, out of which *kurtosis* and *Treves-Rolls sparseness* are popularly used

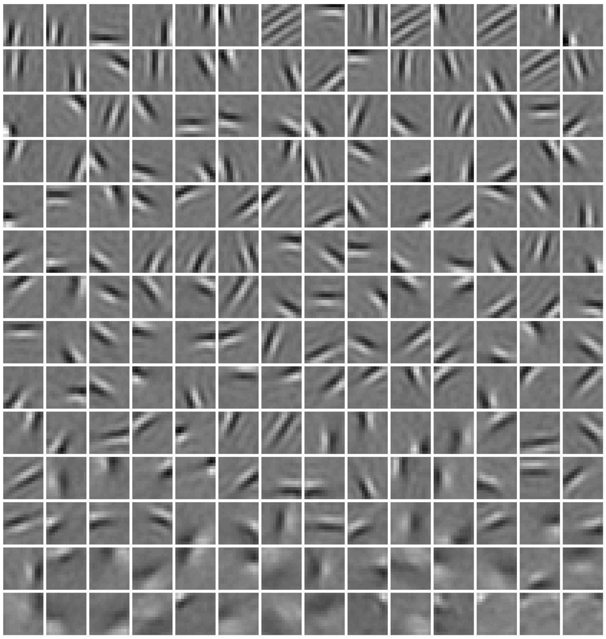


FIGURE 4 | Basis images (columns of W) learned in our model. They can be also considered as receptive fields since we found that they are almost identical.

(Willmore and Tolhurst, 2001). Willmore and Tolhurst (2001) empirically proved that there exists a high correlation between these two measures. In other words, there would be no difference in using these two measures to quantify neurons' activations. Here, we use Treves-Rolls sparseness.

For a neuron, its selectivity is computed across all L input visual stimuli:

$$selectivity = 1 - \frac{(\sum_{l=1}^{L} r_l/L)^2}{(\sum_{l=1}^{L} r_l^2/L)} \tag{17}$$

where r_l is the activation probability of the neuron given the lth stimulus, computed as in Equation (2).

The sparsity of population activations by one stimulus is computed across all N_h neurons:

$$sparsity = 1 - \frac{(\sum_{j=1}^{N_h} r_j/N_h)^2}{(\sum_{j=1}^{N_h} r_j^2/N_h)} \tag{18}$$

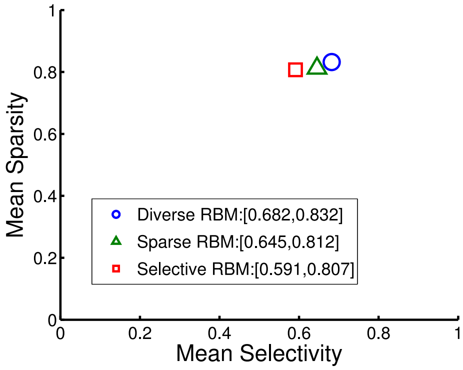


FIGURE 5 | Mean sparsity and mean selectivity of neurons' activations in diverse RBM, sparse RBM and selective RBM, respectively.

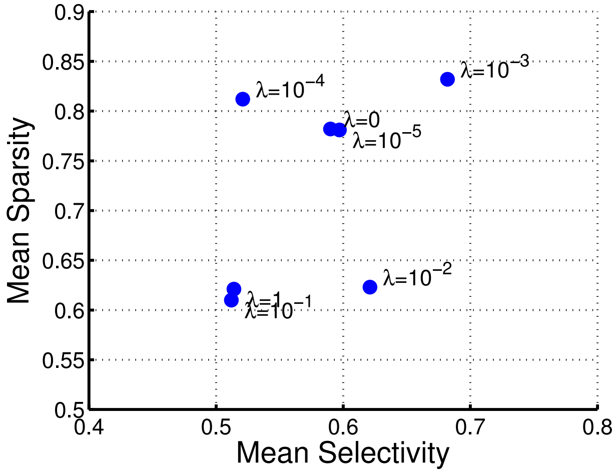


FIGURE 6 | Selectivies and sparsities when using different λ-values in the diverse RBM.

where r_j denotes the activation probability of the jth neuron by the stimulus. We computed the mean selectivity of all 200 neurons and the mean sparsity on all training small patches. The results are plotted in **Figure 5**. Two relevant models (selective RBM and sparse RBM, see Section 4.1) were tested as well for comparison. It can be seen that using the diversity prior in learning can result in comparable selectivity and sparsity as using selectivity prior or sparse prior. Meanwhile, the diversity prior should be preferred since it generates a much smaller number of "dead" neurons (see Section 4.1). In our experiment, $\lambda = 10^{-3}$ was used to obtain the above result. To check how λ value affects sparsity and selectivity, in **Figure 6** a plot with several λ is presented. When λ is small, e.g., $0, 10^{-5}, 10^{-4}$, the effect of the diversity prior is weak or totally removed and both selectivity and sparsity decrease (**Figure 6**). The receptive fields of a diverse RBM trained with $\lambda = 10^{-5}$ are shown in **Figure 7A**. If we use a big value of λ, e.g., $10^{-2}, 10^{-1}, 1$, the iterative update of **W** Equation (15) is greatly dominated by the prior part, and therefore the fitness to the training data $\mathcal{D}$ deteriorates. It can be seen that selectivity and sparsity also decrease (even to a larger degree) using relatively large λs (**Figure 6**). The receptive fields learned with $\lambda = 1$ are displayed in **Figure 7B**.

3.3. Comparison with Biological Data

To better compare our receptive fields against physiological results (Ringach, 2002), we first fitted our receptive fields to Gabor filters:

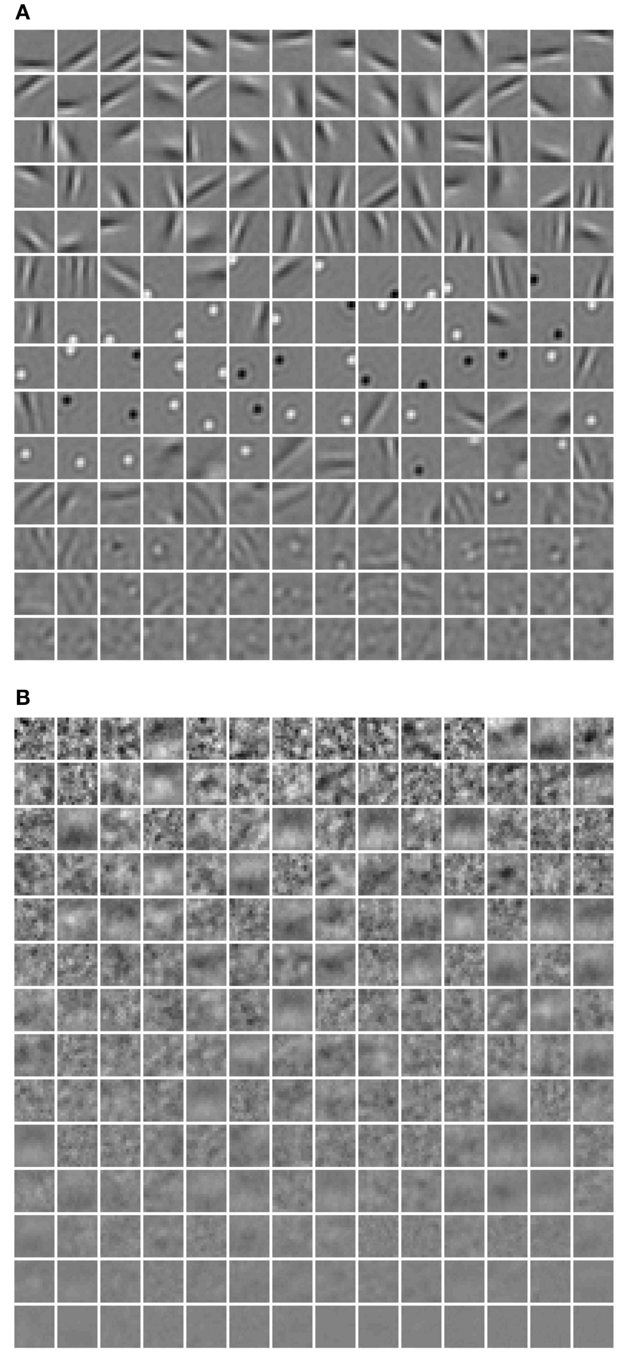


FIGURE 7 | The receptive fields learned using (A) λ = 10^{-5}, (B) λ = 1.

$$
\begin{aligned}
G(x, y; x_0, y_0, A, \sigma_x, \sigma_y, \theta, f, \phi) &= A\cos(2\pi f x' + \phi) \\
&\quad \exp\left(-\frac{x'^2}{2\sigma_x^2} - \frac{y'^2}{2\sigma_y^2}\right) \\
x' &= (x - x_0)\cos\theta + (y - y_0)\sin\theta \\
y' &= -(x - x_0)\sin\theta + (y - y_0)\cos\theta
\end{aligned} \tag{19}
$$

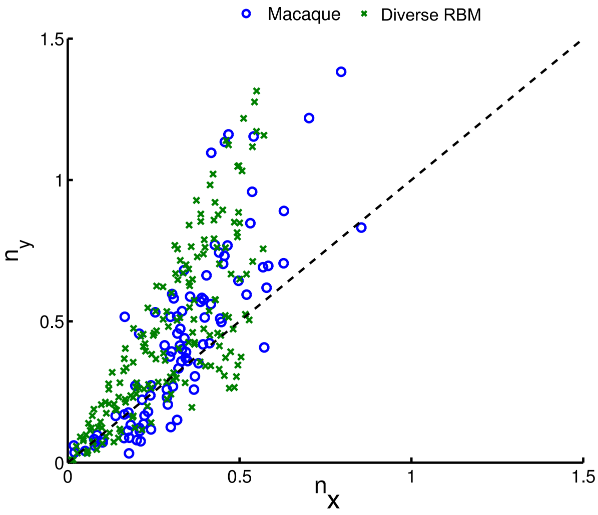


FIGURE 8 | After fitting receptive fields with Gabor filters, we pooled their shape profiles (n_x, n_y), for comparison to physiological data of macaque monkeys (Ringach, 2002).

whose parameters are the center position (x_0, y_0), amplitude A, size (σ_x, σ_y), orientation θ, spatial frequency f and phase ϕ. The fitting is done via the *Nelder-Mead Simplex* method, and therefore is not very reliable. Similar to Hosoya (2012) and Zylberberg and DeWeese (2013), we conducted quality control by filtering out some receptive fields which were poorly fitted. First, we compared our receptive fields with those of macaque monkey V1 cells[2] (Ringach, 2002) in units of the sinusoidal wavelength: $(n_x, n_y) = (\sigma_x f, \sigma_y f)$. In **Figure 8**, we pooled (n_x, n_y) of our receptive fields as well as the data from Zylberberg and DeWeese (2013). We found that they don't deviate very much although they slightly differ from each other. We also checked the statistics of aspect ratios within receptive fields: $\frac{n_y}{n_x}$. In **Figure 9** two histograms are displayed, which are global distributions of aspect ratios from our receptive fields and from the macaque monkey V1 cells, respectively. We can see that they are also quite close.

3.4. Image Reconstruction

Reconstruction using RBMs is quite straightforward. First, small, non-overlapping patches (size 14 × 14) were extracted from a preprocessed image. For each small patch $\mathbf{v}$, the activation probability of each neuron $p(h_j|\mathbf{v})$ can be computed as in Equation (2). Then, instead of using binary states of h_j, $p(h_j|\mathbf{v})$ is used for recovering $\mathbf{v}$ by using Equation (3). It is worth noting that although RBMs are probabilistic models, we use the value of $p(v_i|\mathbf{h})$ to recover the intensity of each pixel and thus the reconstruction is deterministic.

Out of the 10 images in the original database, 8 were used for training and the remaining 2 were used for testing the image reconstruction. The two test images were whitened and sigmoid-mapped using the same preprocessing procedure as the training images. They are shown in the left panel of **Figure 10**. In the right

[2] Data are available at: http://www.ringachlab.net/lab/Data.html.

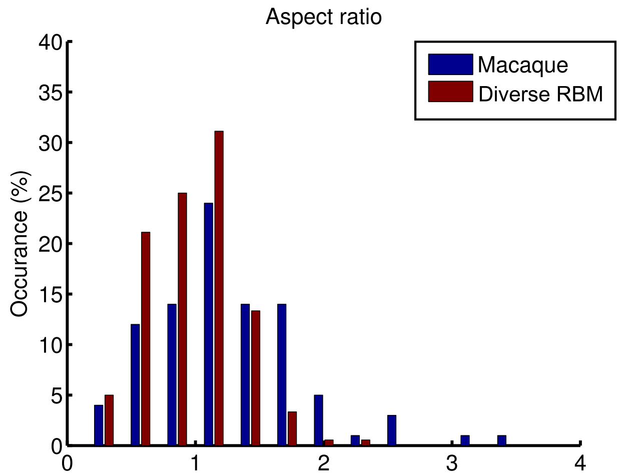


FIGURE 9 | A comparison of the histograms of aspect ratios (n_x/n_y) within macaque monkey V1 neurons and the learned diverse RBM neurons.

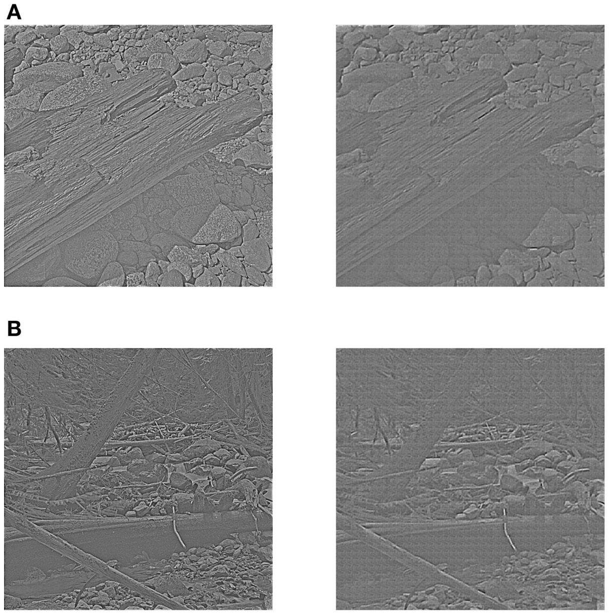


FIGURE 10 | Reconstruction using receptive fields of the learned diverse RBM. The left panel of **(A,B)** shows two test images after preprocessing, while the right panel of **(A,B)** shows two corresponding reconstructions.

panel of **Figure 10** the reconstructions of the two test images are presented. It can be seen that the reconstructions look very good in qualitative terms.

4. Discussion

4.1. Sparsity and Selectivity Prior on RBM

There are previous studies that learn simple cell receptive fields through the use of RBMs, either enforcing sparsity or selectivity. One recent example of the former is the sparse group restricted Boltzmann machine (SGRBM) (Luo et al., 2011), an RBM trained with the CD algorithm plus an $l1/l2$ norm regularization on the activations of the neuron population. At each iteration, given a visual stimulus, and after computing the activation probabilities of the whole neuron set, SGRBM attempts to minimize the $l1/l2$ norm of the set of activation probabilities. Although $l1/l2$ norm regularization can ensure sparsity, it can also lead to many "dead" (never responding) and "potential over-tolerant" (always responding) neurons (see **Figure 1D**). In the case of the latter, a study that enforces selectivity is the one from Lee et al. (2007) which uses a selectivity-induced regularization that suppresses the average activation probability of each neuron to all training stimuli.

One limitation of this strategy, as argued by Goh et al. (2010), is that decreasing average activation probabilities cannot guarantee selectivity. Instead, it will result in many similar neurons with uniformly low activation probabilities to all types of visual stimuli, which are prone to be "dead" as well. Following this line of thought and in order to prove the validity of our diverse RBM, two additional RBMs were trained using the CD algorithm with sparse regularization (sparse RBM) (Luo et al., 2011) and the CD algorithm with selectivity regularization (selective RBM) (Lee et al., 2007). For both of them, 200 hidden neurons were learned and their receptive fields are presented in **Figure 11**. We can see that the neurons' receptive fields learned in sparse RBM and selective RBM look similar to those of our RBM trained with a diversity prior. However, both sparse CD and selective CD led to many useless, "dead" neurons. We estimated the rough number of "dead" neurons by counting the number of neurons whose maximal activation probabilities to all training stimuli is smaller than 0.1, and the results are shown in **Figure 12**. Furthermore, we also computed the mean selectivity and the mean sparsity of neurons in sparse RBM and selective RBM in the same way as we did for the diverse RBM; their results are also shown in **Figure 5**.

4.2. The Equivalent to a Diversity Prior in Biological Systems

Knowing about how neuron receptive field properties arise is of great importance in visual neuroscience in order to hypothesize the circuits and connections that give rise to those properties. On one hand, one of the characteristics of simple cells in V1 is selectivity to oriented stimuli. These can be obtained through placing some constraint in learning from natural images. An example is the influential work by Olshausen and Field (1996). A set of coefficients is then formed such that they have a cost associated to them depending on how the activity is distributed. The aim is to increase sparsity, meaning lower cost. This approach leads to V1-like simple-cell receptive fields through the learning of a set of weights that correspond to the connections of the input layer with simple neurons in area V1.

On the other hand, inhibition seems to play a central role in the shaping of simple-cell receptive fields. We can consider three types of inhibitory inputs: feedforward, lateral and feedback (also known as recurrent). Feedforward inhibition is regarded as the main source of orientation selectivity in simple cells

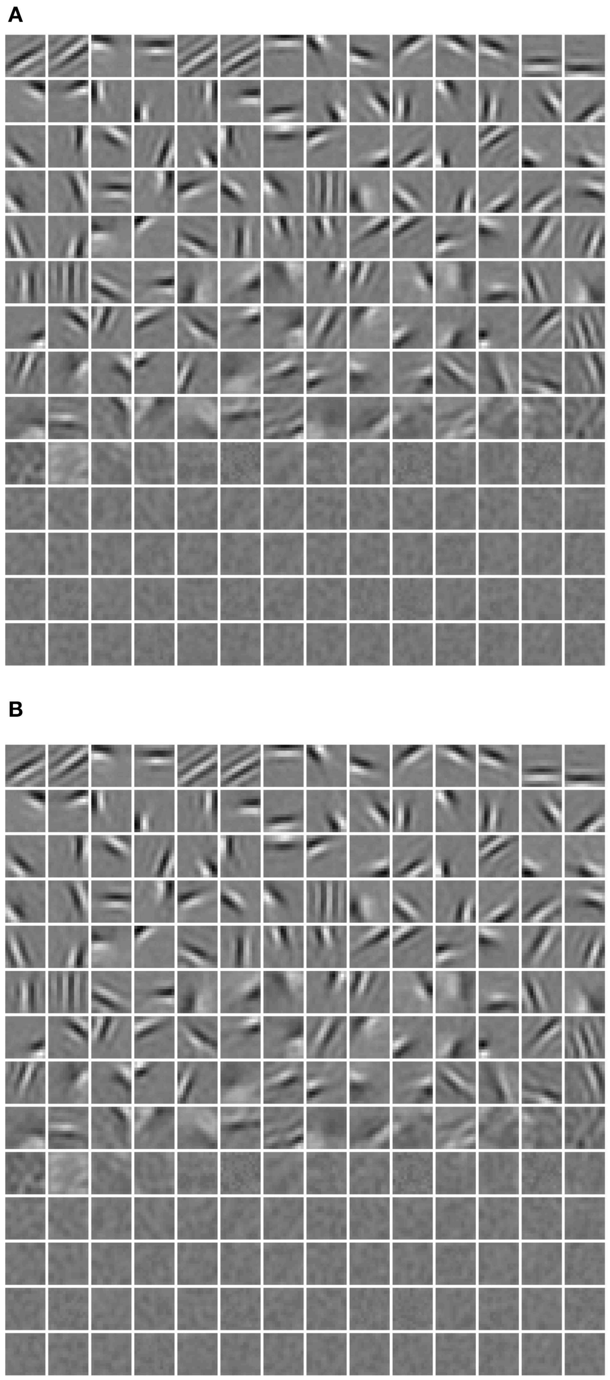


FIGURE 11 | The receptive fields learned in the learned (A) sparse RBM, (B) selective RBM.

by some researchers (Heggelund, 1981; Celebrini et al., 1993; Ferster and Miller, 2000) and has been modeled by others, e.g., (Azzopardi et al., 2014). The classical role of feedback connections was the enhancement of receptive-field responses to top-down modulations (Ito and Gilbert, 1999; Treue, 2003), which have been successfully modeled for attention (Rodriguez-Sanchez et al., 2007) and contour integration (Neumann and Sepp, 1999; Tschechne and Neumann, 2014). But other studies are in support of feedback connections as the source of simple-cell selectivity through recurrent connections, most recently from Angelucci and Bressloff (2006). The appearance of orientation selectivity this way has also been proposed in models of recurrent inhibition, e.g., (Sabatini, 1996; Carandini and Ringach, 1997). Finally, even though there is an alive discussion regarding if orientation selection is achieved through feedforward or recurrent connections, it is interesting to note that none of them rule out that lateral inhibition can at least be partially blamed for this selectivity, e.g., (Celebrini et al., 1993; Angelucci and Bressloff, 2006). Lateral connections have in fact being made explicit into recent sparse coding models (Garrigues and Olshausen, 2008; King et al., 2013).

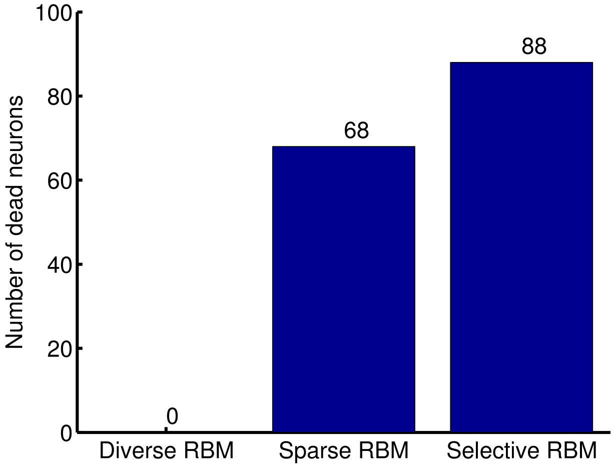


FIGURE 12 | Number of dead neurons in the learned diverse RBM, sparse RBM and selective RBM, respectively.

The common ground of all the aforementioned works is that inhibition is fundamental to the selectivity properties of simple cells, irrespective of where that inhibition comes from. Inhibition is also linked to the appearance of sparse sensory coding (Vinje and Gallant, 2000; Haider et al., 2010). We can conclude then, that inhibition would generate RF diversity, since as we have shown in this work (**Figure 1**), imposing diversity generates both selective and sparse neural populations. By explicitly favoring diversity in our model, we would be mimicking the effect that inhibition should have on feature learning in a biological system.

5. Conclusion

We test a recent new concept, that of diversity (Padmanabhan and Urban, 2010; O'Donnell and Nolan, 2011), by applying diversification on the columns of **W** when using a RBM to learn receptive fields. This diversification has the implication of providing a set of neurons that is at the same time sparse and selective, which, as mentioned in the introduction, is not always the case for sparse models. Imposing diversity is thus a more general condition to achieve both, sparsity and selectivity.

References

Angelucci, A., and Bressloff, P. C. (2006). Contribution of feedforward, lateral and feedback connections to the classical receptive field center and extra-classical receptive field surround of primate v1 neurons. *Progr. Brain Res.* 154, 93–120. doi: 10.1016/S0079-6123(06)54005-1

Azzopardi, G., Rodríguez-Sánchez, A., Piater, J., and Petkov, N. (2014). A push-pull corf model of a simple cell with antiphase inhibition improves snr and contour detection. *PLoS ONE* 9:e98424. doi: 10.1371/journal.pone.0098424

Barlow, H. B. (1989). Unsupervised learning. *Neural Comput.* 1, 295–311.

Bell, A., and Sejnowski, T. (1997). The "independent components" of natural scenes are edge filters. *Vis. Res.* 37, 3327–3338.

Bengio, Y. (2009). Learning deep architectures for AI. *Found. Trends Mach. Learn.* 2, 1–127. doi: 10.1561/2200000006

Berkes, P., White, B., and Fiser, J. (2009). "No evidence for active sparsification in the visual cortex," in *Neural Information Processing Systems (NIPS)* (Montreal, QC), 108–116.

Carandini, M., and Ringach, D. L. (1997). Predictions of a recurrent model of orientation selectivity. *Vis. Res.* 37, 3061–3071.

Celebrini, S., Thorpe, S., Trotter, Y., and Imbert, M. (1993). Dynamics of orientation coding in area v1 of the awake primate. *Vis. Neurosci.* 10, 811–825.

Ferster, D., and Miller, K. D. (2000). Neural mechanisms of orientation selectivity in the visual cortex. *Annu. Rev. Neurosci.* 23, 441–471. doi: 10.1146/annurev.neuro.23.1.441

Field, D. (1994). What is the goal of sensory coding? *Neural Comput.* 6, 559–601.

Garrigues, P., and Olshausen, B. A. (2008). "Learning horizontal connections in a sparse coding model of natural images," in *Neural Information Processing Systems (NIPS)* (Vancouver, BC), 505–512.

Goh, H., Thome, N., and Cord, M. (2010). "Biasing restricted Boltzmann machines to manipulate latent selectivity and sparsity," in *Neural Information Processing Systems (NIPS) Workshop on Deep Learning and Unsupervised Feature Learning* (Vancouver, BC).

Haider, B., Krause, M. R., Duque, A., Yu, Y., Touryan, J., Mazer, J. A., et al. (2010). Synaptic and network mechanisms of sparse and reliable visual cortical activity during nonclassical receptive field stimulation. *Neuron* 65, 107–121. doi: 10.1016/j.neuron.2009.12.005

Heggelund, P. (1981). Receptive field organization of simple cells in cat striate cortex. *Exp. Brain Res.* 42, 89–98.

Hinton, G. E. (2002). Training products of experts by minimizing contrastive divergence. *Neural Comput.* 14, 1771–1800. doi: 10.1162/089976602760128018

Hinton, G. E., and Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science* 313, 504–507. doi: 10.1126/science.1127647

Hosoya, H. (2012). Multinomial bayesian learning for modeling classical and nonclassical receptive field properties. *Neural Comput.* 24, 2119–2150. doi: 10.1162/NECO_a_00310

HUBEL, D., and WIESEL, T. (1959). Receptive fields of single neurones in the cat's striate cortex. *J. Physiol.* 148, 574–591.

Hyvärinen, A., and Oja, E. (2000). Independent component analysis: algorithms and applications. *Neural Netw.* 13, 411–430. doi: 10.1016/S0893-6080(00)00026-5

Ito, M., and Gilbert, C. D. (1999). Attention modulates contextual influences in the primary visual cortex of alert monkeys. *Neuron* 22, 593–604.

Jones, E. (2007). Neuroanatomy: cajal and after cajal. *Brain Res. Rev.* 55, 248–255. doi: 10.1016/j.brainresrev.2007.06.001

King, P. D., Zylberberg, J., and DeWeese, M. R. (2013). Inhibitory interneurons decorrelate excitatory cells to drive sparse code formation in a spiking model of v1. *J. Neurosci.* 33, 5475–5485. doi: 10.1523/JNEUROSCI.4188-12.2013

Koller, D., and Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*. Cambridge, MA: MIT Press.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444. doi: 10.1038/nature14539

Lee, H., Ekanadham, C., and Ng, A. Y. (2007). "Sparse deep belief net model for visual area v2," in *Neural Information Processing Systems (NIPS)*, eds J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis (Vancouver, BC: Curran Associates, Inc.), 873–880.

Luo, H., Shen, R., Niu, C., and Ullrich, C. (2011). "Sparse group restricted boltzmann machines," in *AAAI Conference on Artificial Intelligence (AAAI)* (Granada).

Neumann, H., and Sepp, W. (1999). Recurrent v1–v2 interaction in early visual boundary processing. *Biol. Cybern.* 81, 425–444.

O'Donnell, C., and Nolan, M. F. (2011). Tuning of synaptic responses: an organizing principle for optimization of neural circuits. *Trends Neurosci.* 34, 51–60. doi: 10.1016/j.tins.2010.10.003

Olshausen, B., and Field, D. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* 381, 607–609.

Padmanabhan, K., and Urban, N. N. (2010). Intrinsic biophysical diversity decorrelates neuronal firing while increasing information content. *Nat. Neurosci.* 13, 1276–1282. doi: 10.1038/nn.2630

Ramón y Cajal, S. (1888). Sobre las fibras nerviosas de la capa molecular del cerebelo. *Rev. Trim. Histol. Norm. Patol.* 1, 33–49.

Ramón y Cajal, S. (1904). Variaciones morfologicas, normales y patologicas del reticulo neurofibrilar. *Trab. Lab. Investig. Biol. Madrid* 3, 9–15.

Ringach, D. L. (2002). Spatial structure and symmetry of simple-cell receptive fields in macaque primary visual cortex. *J. Neurophysiol.* 88, 455–463.

Rodriguez-Sanchez, A. J., Simine, E., and Tsotsos, J. K. (2007). Attention and visual search. *Int. J. Neural Syst.* 17, 275–288. doi: 10.1142/S0129065707001135

Rozell, C. J., Johnson, D. H., Baraniuk, R. G., and Olshausen, B. A. (2008). Sparse coding via thresholding and local competition in neural circuits. *Neural Comput.* 20, 2526–2563. doi: 10.1162/neco.2008.03-07-486

Sabatini, S. P. (1996). Recurrent inhibition and clustered connectivity as a basis for gabor-like receptive fields in the visual cortex. *Biol. Cybern.* 74, 189–202.

Treue, S. (2003). Visual attention: the where, what, how and why of saliency. *Curr. Opin. Neurobiol.* 13, 428–432. doi: 10.1016/S0959-4388(03)00105-3

Tschechne, S., and Neumann, H. (2014). Hierarchical representation of shapes in visual cortex from localized features to figural shape segregation. *Front. Comput. Neurosci.* 8:93. doi: 10.3389/fncom.2014.00093

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. R. Soc. B. Biol. Sci.* 265, 359–366.

Vinje, W. E., and Gallant, J. L. (2000). Sparse coding and decorrelation in primary visual cortex during natural vision. *Science* 287, 1273–1276. doi: 10.1126/science.287.5456.1273

Willmore, B., and Tolhurst, D. (2001). Characterising the sparseness of neural codes. *Netw. Comput. Neural Syst.* 12, 255–270. doi: 10.1080/713663277

Xiong, H., Szedmak, S., Rodríguez-Sánchez, A., and Piater, J. (2014). "Towards sparsity and selectivity: bayesian learning of restricted boltzmann machine for early visual features," in *24th International Conference on Artificial Neural Networks (ICANN14)* (Hamburg).

Zylberberg, J., and DeWeese, M. R. (2013). Sparse coding models can exhibit decreasing sparseness while learning sparse codes for natural images. *PLoS Comput. Biol.* 9:e1003182. doi: 10.1371/journal.pcbi.1003182

Zylberberg, J., Murphy, J. T., and DeWeese, M. R. (2011). A sparse coding model with synaptically local plasticity and spiking neurons can account for the diverse shapes of v1 simple cell receptive fields. *PLoS Comput. Biol.* 7:e1002250. doi: 10.1371/journal.pcbi.1002250

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.

Appendix

A1. MAP optimization with the Sum Similarity Minimization Prior

The optimization problem (Equation 14) can be rewritten as

$$\max_{\mathbf{W}} \mathcal{L}(\mathcal{D}) - \lambda \underbrace{\left[\sum_{j=1}^{N_h} \sum_{k \neq j}^{N_h} (\mathbf{W}_{\cdot,j}^{\top} \mathbf{W}_{\cdot,k})^2 + C \sum_{j=1}^{N_h} (||\mathbf{W}_{\cdot,j}|| - 1)^2 \right]}_{O}, \tag{A1}$$

where C is an extra parameter that is set relatively large to guarantee the satisfaction of the constraints in Equation (14). In our experiment, C is equivalent to 10^4. In this way, the constrained optimization problem is converted to an unconstrained one. It was already shown that the gradient ascent can used to maximize $\mathcal{L}(\mathcal{D})$. It is easy to see that O is also convex with respect to $\mathbf{W}$; therefore, the same gradient ascent can be also applied on $-\lambda O$. The gradient of O with respect to $\mathbf{W}$ can be computed as

$$\frac{\partial O}{\partial \mathbf{W}_{\cdot,j}} = 2 \sum_{k \neq j}^{N_h} (\mathbf{W}_{\cdot,j}^{\top} \mathbf{W}_{\cdot,k}) \mathbf{W}_{\cdot,k} + 2C(||\mathbf{W}_{\cdot,j}|| - 1) \frac{\mathbf{W}_{\cdot,j}}{||\mathbf{W}_{\cdot,j}||} \tag{A2}$$

$$= 2 \sum_{k \neq j}^{N_h} (\mathbf{W}_{\cdot,k} \otimes \mathbf{W}_{\cdot,k}) \mathbf{W}_{\cdot,j} + 2C(||\mathbf{W}_{\cdot,j}|| - 1) \frac{\mathbf{W}_{\cdot,j}}{||\mathbf{W}_{\cdot,j}||} \tag{A3}$$

$$= 2 \left(\sum_{k \neq j}^{N_h} (\mathbf{W}_{\cdot,k} \otimes \mathbf{W}_{\cdot,k}) + C \frac{||\mathbf{W}_{\cdot,j}|| - 1}{||\mathbf{W}_{\cdot,j}||} \mathbf{I}_{N_v} \right) \mathbf{W}_{\cdot,j}, \tag{A4}$$

where $\otimes$ denotes the outer product between vectors and $\mathbf{I}_{N_v}$ is a $N_v \times N_v$ identity matrix.

frontiers in
COMPUTATIONAL NEUROSCIENCE

ORIGINAL RESEARCH ARTICLE
published: 06 January 2015
doi: 10.3389/fncom.2014.00171

Correlated activity supports efficient cortical processing

Chou P. Hung*[1,2], **Ding Cui**[1], **Yueh-peng Chen**[2], **Chia-pei Lin**[2] **and Matthew R. Levine**[1]

[1] *Department of Neuroscience, Georgetown University, Washington, D.C., USA*
[2] *Institute of Neuroscience, National Yang-Ming University, Taipei, Taiwan*

Edited by:
Hans P. Op De Beeck, KU Leuven, Belgium

Reviewed by:
Ben Vermaercke, KU Leuven, Belgium
Bram-Ernst Verhoef, University of Chicago, USA

****Correspondence:***
Chou P. Hung, Department of Neuroscience, Georgetown University, NRB EP-04, 3970 Reservoir Rd. NW, Washington, 20007, D.C., USA
e-mail: ch486@georgetown.edu

Visual recognition is a computational challenge that is thought to occur via efficient coding. An important concept is sparseness, a measure of coding efficiency. The prevailing view is that sparseness supports efficiency by minimizing redundancy and correlations in spiking populations. Yet, we recently reported that "choristers", neurons that behave more similarly (have correlated stimulus preferences and spontaneous coincident spiking), carry more generalizable object information than uncorrelated neurons ("soloists") in macaque inferior temporal (IT) cortex. The rarity of choristers (as low as 6% of IT neurons) indicates that they were likely missed in previous studies. Here, we report that correlation strength is distinct from sparseness (choristers are not simply broadly tuned neurons), that choristers are located in non-granular output layers, and that correlated activity predicts human visual search efficiency. These counterintuitive results suggest that a redundant correlational structure supports efficient processing and behavior.

Keywords: object recognition, inferior temporal cortex, macaque, visual search, efficient coding

INTRODUCTION

Visual recognition engages neural mechanisms that are essential to our ability to learn and process complex information (Poggio and Bizzi, 2004). The key challenge of recognition is generalization, which requires that the representation is both object-specific and invariant to changes such as illumination and pose, even for novel objects. This is thought to occur via a hierarchy of cortical areas along the ventral visual pathway, ending in the inferior temporal (IT) cortex (Miyashita, 1993; Logothetis and Sheinberg, 1996; Tanaka, 1996; Tootell et al., 2003), but the underlying computations remain poorly understood (DiCarlo and Cox, 2007; DiCarlo et al., 2012). Current models and theories of recognition (Riesenhuber and Poggio, 1999; Masquelier and Thorpe, 2007; Mutch and Lowe, 2008; Bengio, 2009; Krizhevsky et al., 2012; Le et al., 2012; Zeiler and Fergus, 2014; Cadieu et al., 2014) are based on the idea that a hierarchy of simple and complex cells combine to increase specificity and invariance. To improve these models, it is necessary to understand the computations of local populations of neurons at an intermediate level of abstraction (DiCarlo et al., 2012).

A key concept is sparseness, a measure of coding efficiency. The current thinking is that sparseness increases efficiency by minimizing redundancy, correlation, and noise (Gawne and Richmond, 1993; Zohary et al., 1994; Vinje and Gallant, 2000; Olshausen and Field, 2004; Ecker et al., 2010; Renart et al., 2010; Xing et al., 2011; Hansen et al., 2012; King et al., 2013). Yet, reports in V1 slices and *in vivo* have shown the existence of neural ensembles that fire reliably in concert during spontaneous activity (Sadovsky and Maclean, 2014), and the same ensembles are active both without stimulation and in response to stimulation (Chu et al., 2014; Miller et al., 2014). We recently reported (Lin et al., 2014) that in macaque IT, correlated neurons "choristers" (Kenet et al., 2005; Carandini, 2014), neurons that have similar stimulus tuning and coincident spike timing, even during spontaneous activity, carry more generalizable object information than uncorrelated neurons ("soloists"). This surprising result hints that, counterintuitively, correlation *supports* efficient coding and that current thinking focused on sparsening, decorrelation, and denoising may be flawed.

The idea that the correlational structure, i.e., the spatial pattern of homogeneity vs. heterogeneity within a local population of neurons, may support efficient coding has been postulated in theory (Abbott and Dayan, 1999; Sompolinsky et al., 2001; Wu et al., 2002; Dehaene and Changeux, 2005; Averbeck et al., 2006; Cohen and Kohn, 2011; Ecker et al., 2011; Eyherabide and Samengo, 2013; Shamir, 2014), but it has received little experimental support. Three novel aspects of our study allowed us to explore this hypothesis.

First, we used dense electrode arrays (64 sites across roughly two cortical columns, 0.2 mm resolution horizontally and in depth, **Figure 1A**) to characterize the correlational structure. High-density arrays allowed us to record neurons that have similar tuning, to measure redundancy as "Average Correlation Strength" (a site's average pairwise tuning similarity with all other sites in the array, where the tuning similarity between two sites is the Pearson correlation of their z-normalized stimulus responses, related to the concept of "population sparseness" Willmore et al., 2011). Because previous reports of efficient coding had insufficient sampling density to measure population sparseness, they instead measured "sparseness" as tuning sharpness, the selectivity of a neuron's response across stimuli, under the assumption that "sparseness" and "population sparseness" are interchangeable (that sparseness and correlation strength are inversely related) (Rolls and Tovee, 1995; Vinje and Gallant, 2000; Zoccolan et al., 2007; Willmore et al., 2011). When studies did examine functional correlation, it was in terms

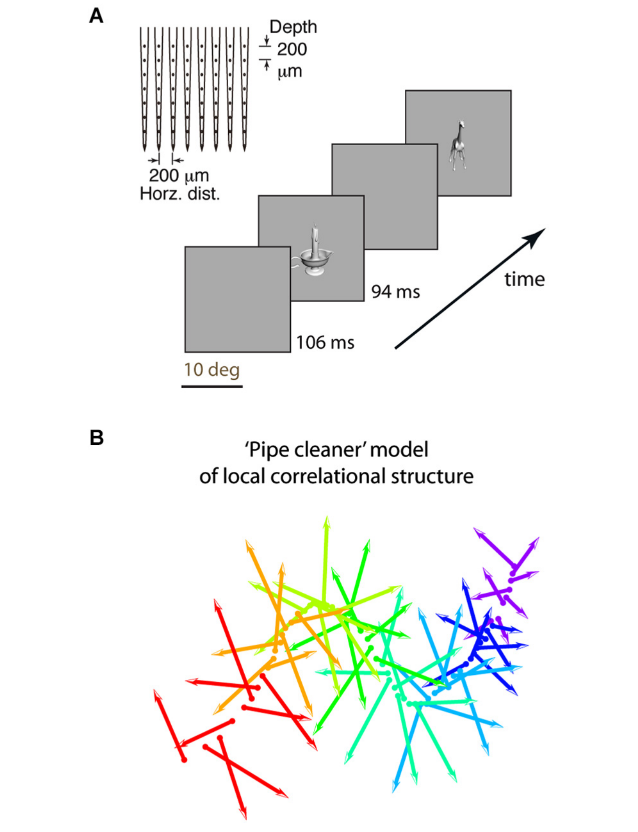


FIGURE 1 | Experimental design and "pipe cleaner" model. (A) We inserted a dense multi-depth array (64 sites across ~2 cortical columns) in macaque lateral IT (A16) and recorded spiking responses under light neurolept anesthesia. Stimuli were presented via rapid serial visual presentation, for 94 ms ON and 106 ms OFF (5 Hz), in pseudorandom order for 10 repetitions. Spike count from 100 to 200 ms post stimulus onset was averaged across repetitions. **(B)** A "pipe cleaner" model linking local correlational structure in neighboring columns to invariant representation. Most neurons are weakly correlated "soloists" (the bristles), tied to an underlying structure of correlated neurons ("choristers," the spine). Sampling a few points along the spine (a few choristers) is sufficient to reconstruct the overall structure. The model predicts that generalizable object information is carried by the choristers, and that the heterogeneity of the soloists may help to fine-tune the choristers to support generalization.

of individual pairs of neurons, without comparing the relationship between sparseness (tuning sharpness) and correlation (Gawne and Richmond, 1993; Sato et al., 2009; Takeuchi et al., 2011), or the comparison was limited to layers 2/3 (Tamura et al., 2014). Whether "correlation strength" and "sparseness" are related for a diverse sample of IT neurons remains untested (Willmore et al., 2011), and answering this question is important for understanding how local architecture relates to coding efficiency.

Second, the dense multi-depth arrays allowed us to examine layer specificity, which can tell us about input-output relationships. We and others (Sato et al., 2009; Lin et al., 2014; Tamura et al., 2014) previously reported that local IT populations have a correlational structure in which most neurons are weakly correlated and few neurons have strong tuning correlation and significant spontaneous coincident spiking (~6% of neuronal pairs in IT, vs. ~50% of pairs in V1; Chu et al., 2014). Yet, these rare IT choristers are also highly efficient. Just 4–5 choristers per array (the top 6% as defined by k-means clustering, or 8% as defined by average pairwise tuning correlation) have the same object coding capability, for within-category generalization, as the entire array population (no more are needed given their object coding efficiency; Figure 7C of Lin et al., 2014). Based on this correlational structure and the much-better object coding capability of choristers vs. soloists, we previously proposed a "pipe cleaner" model (**Figure 1B**, a "fiber bundle" in mathematical terminology) in which the choristers (the spine) are the substrate of IT's output, encoding an invariant representation that supports generalization and recognition, and in which the soloists (the bristles) are IT's inputs, acting as heterogeneous tensors that fine-tune this high-dimensional representation (in the parlance of DiCarlo et al. (2012), to support "cortically local subspace untangling" and to "flatten object manifolds"). If so, choristers and soloists should be layer specific, with soloists tending to be in input layers and choristers tending to be in output layers. Such layer-specificity would be consistent with reports of decorrelated responses near layer 4 of V1 (Ecker et al., 2010; Hansen et al., 2012) and with reports that tolerance but not selectivity (sparseness) increases along the ventral visual pathway (Rust and Dicarlo, 2010; Willmore et al., 2011).

Third, we tested whether local correlated activity can predict visual search efficiency for complex naturalistic object stimuli. Previous reports have linked IT neuronal tuning to visual perception (Logothetis and Schall, 1989; Op de Beeck et al., 2001; Baker et al., 2002; Sigala and Logothetis, 2002; Mruczek and Sheinberg, 2007a; Sripati and Olson, 2010; Verhoef et al., 2012) and have linked perception to topography in V1 (Michel et al., 2013), but the interpretation was not linked to correlational structure. If local correlational structure, e.g., from short-range lateral inhibition in IT, predicts search efficiency, it would support that correlated activity and topography are linked to complex shape perception. It would also support recent reports that abnormal correlated activity and excitatory/inhibitory balance in object areas are linked to abnormal perception in autistics, linking these findings to spiking activity (Jiang et al., 2013; Robertson et al., 2013). Here, we asked whether local correlated activity predicts visual search for combinations of naturalistic objects. To avoid effects that might be driven by spatial attention or processes earlier in visual cortex, we used brief presentations at random locations followed by masking, and we equalized the stimuli for low level visual properties such as Fourier energy. Also, our stimuli were object combinations defined by local correlated activity in IT ("neurally defined features"; each "feature" is a set of objects), rather than abstract human-defined shapes as in previous reports, so that the predictions are specifically tied to contrastive coding of complex features by neighboring IT columns (e.g., from lateral inhibition).

Together with our previous report (Lin et al., 2014), these tests provide additional support for the hypothesis that correlated activity supports efficient processing and behavior. We report that although the concepts of sparseness and decorrelation are often conflated, correlation strength and sparseness (when measured as tuning sharpness) should be considered as separate factors. We also provide additional support for choristers as the output neurons of IT, based on their cortical depth. Finally, we show that correlated activity in macaque IT predicts human visual search performance in a task with complex shapes.

METHODS

NEUROPHYSIOLOGY AND STIMULUS PRESENTATION

All experimental procedures in monkeys (*Macaca cyclopis*) were performed in accordance with the National Institutes of Health Guide for the Care and Use of Laboratory Animals and were approved by the Institutional Animal Care and Use Committee of National Yang-Ming University. The procedures for the experiments were described in Lin et al. (2014) and are briefly summarized here. We inserted dense microelectrode arrays that had 64 sites (8 shanks and 8 contacts per shank, at 0.2 mm spacing spanning 1.4 × 1.4 mm horizontally and in depth, NeuroNexus A8×8-5mm-200-200-413) spanning all cortical depths and ∼2–4 neighboring cortical columns (Figure 5 of Lin et al. (2014)). Recordings were made from 5 arrays, where each array was a separate insertion in a separate recording session, across 4 monkeys.

Initial surgery for headpost, EEG, and recording chamber implant was under isoflurane anesthesia, followed by repeated recording sessions under light neurolept anesthesia (Fujita et al., 1992; Wang et al., 2000; Tsunoda et al., 2001; Yamane et al., 2006; Sato et al., 2009, 2013; Brown et al., 2011) (0.9 μg/kg/hr i.v. Fentanyl, 70%/30% N_2O/O_2, 0.25 mg/kg i.m. droperidol, and 0.3–0.5% isoflurane) and muscle relaxation (1.2 mg/kg/hr i.v. rocuronium bromide). The fentanyl concentration is 100× lower than in a recent report that contrasted awake vs. anesthetized signals (Ecker et al., 2014), and 10× lower than in reports that did not find an effect on neuronal dynamics (Loughnan et al., 1987; Constantinople and Bruno, 2011). Our signals also lacked artifacts such as prolonged responses and up/down fluctuations reported with other anesthetics (Contreras et al., 1997; Haider et al., 2013). Compared to awake recordings, light anesthesia and muscle relaxation have the advantage of excluding potential effects from task-related top-down signals (Sigala and Logothetis, 2002; Maier et al., 2007; Ruff and Cohen, 2014) or eye movements (Rajkai et al., 2008; Ito et al., 2011), and a recent report suggests that activity during running resembles activity under anesthesia and is dissimilar to "visually detached" activity during quiet wakefulness (Froudarakis et al., 2014).

Single units were analyzed for coincident spiking and to remove cases of multiple detection of the same neuron across different contacts. All "site" responses were based on multi-unit activity (MUA) pooled from isolated single units at the same contact. We report only "site" responses because lower spike counts and the possibility of oversorting with single unit activity (SUA) can artificially weaken correlation measurements (Cohen and Kohn, 2011), and because the conclusions were the same as for MUA. The stimuli were 240 grayscale rendered objects or 113 colored photographed objects presented via rapid serial visual presentation (94 ms ON/ 106 ms OFF).

ANALYSES

Analyses were based on spike count from 100 to 200 ms after stimulus onset. Each site's tuning function was calculated as its trial-averaged response, z-normalized across stimuli. The same matrix of trial-averaged and z-normalized tuning responses across the array (i.e., a 240 × 64 matrix for 240 stimuli and 64 sites) was used as the input for correlation analysis, k-means, principal component analysis (PCA), and classifier analysis as described in Lin et al. (2014).

We classified each site as a "chorister" or "soloist" based on the site's average pairwise tuning correlation with other sites from the same array (the same calculation as in Lin et al. (2014) Figure 7C, brown line, but here "choristers" are random sites in the top 30%ile instead of rank-ordered sites in the top 8%ile). This top 30%ile corresponds to 16 sites per array for arrays 1–3 and 8 sites per array for arrays 4 and 5 that had more inactive sites. "Soloists" are the remaining sites (**Figure 2A** black dots). Choristers and soloists lie along a continuum of average pairwise tuning correlation strengths (**Figure 2A**). For object classification (**Figure 2B**) and noise covariation analyses (**Figure 2C**), we compared choristers (top 30%ile) against soloists in the 45–65%ile. For cortical depth (**Figure 3B**), the soloists are the bottom 30%ile. Layer-specificity is not seen for soloists in the 45–65%ile.

We used a linear support vector machine classifier to estimate the ability of a hypothetical downstream neuron (e.g., in prefrontal cortex) to read out the category of an untrained object (within-category generalization) based on the population activity in IT. The classifier output is based on the weighted sum of spiking activity from a set of IT neurons followed by a decision threshold. Because there were 8 possible categories, the classifier learned a one-vs.-all decision hyperplane for each of 8 categories and output the category that had the highest certainty.

Sparseness was calculated according to Vinje and Gallant (2000) and Zoccolan et al. (2007) as:

$$S = \left(1 - \frac{\left(\sum \frac{R_i}{n}\right)^2}{\sum \frac{R_i^2}{n}}\right) \Big/ \left(1 - \frac{1}{n}\right),$$

where R_i is the site response to the i-th stimulus and n is the number of stimuli in the set.

We could estimate the cortical depth because we were able to visually see individual sites disappear into the brain during insertion and, because of the small footprint of the array shanks (15 μm thick, 33 μm wide), we could also track individual units as they transitioned from the deepest to the most superficial sites during array insertion. Anatomical confirmation of depth was impossible due to damage from later recording sessions. However, we estimate that the deviation of the array from vertical was less than 8 deg (less than 0.2 mm horizontal offset at the deepest site), based on anatomical confirmation of our V1 recordings using the same arrays (Supplemental Figure 1 in Chu et al., 2014).

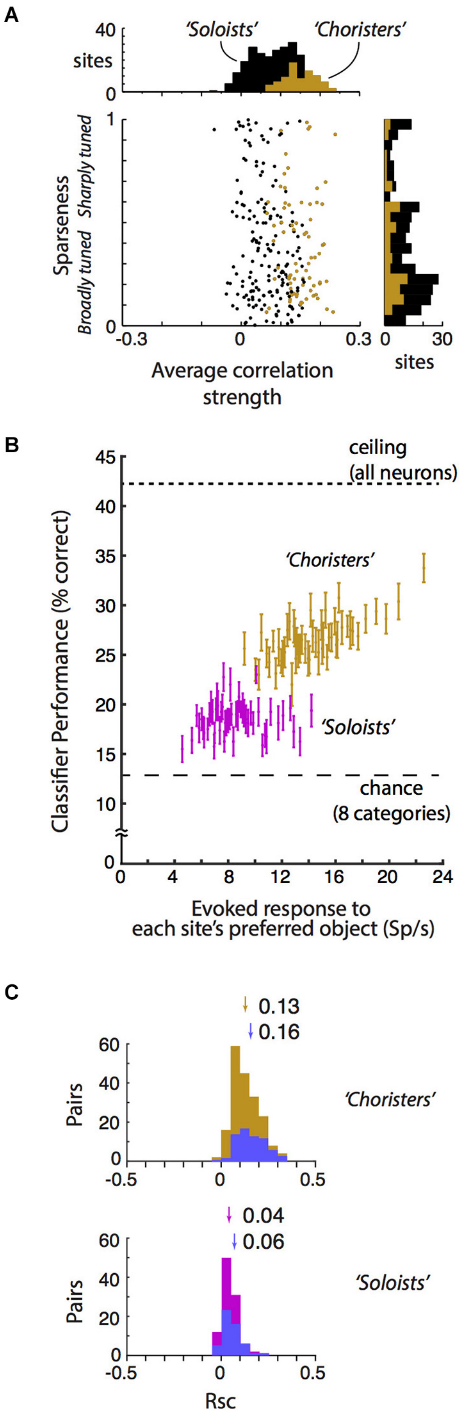


FIGURE 2 | Local correlational structure, sparseness, and generalization performance. (A) Sparseness (tuning width) and average correlation strength were only weakly related across 250 sites in anesthetized IT ($r = -0.09$, $p = 0.04$). Sparseness was calculated according to Zoccolan et al. (2007) and Vinje and Gallant (2000). Average correlation strength was the average site-to-site tuning correlation between each site and all other sites in the same array. Sparseness and average correlation strength were each highly consistent across two stimulus sets ($r = 0.72$ and 0.70, $p < 10^{-37}$ and $p < 10^{-39}$ resp.). Choristers (brown) are the 30%ile of sites with the highest average correlation strength per array, and soloists (black) are the remaining sites. **(B)** Visual responsiveness vs. within-category generalization performance for choristers (top 30%ile, brown) vs. soloists (45–65%ile, red), for 2 sites per array, with at least 600 μm horizontal distance between sites. Visual responsiveness was calculated as the evoked (baseline-subtracted) response to each site's preferred object, shown as the median across 10 sites (2 sites per array, 5 array insertions across 4 monkeys). Chance is 12.5% for 8 categories, and ceiling performance is based on all sites. Choristers and soloists were defined without test stimuli. Compare with Figure 7C of Lin et al. (2014). **(C)** Noise correlation (Rsc) of choristers vs. soloists (same colors and definitions as in **(B)** also with at least 600 μm horizontal distance). To control for visual drive, we also show Rsc for pairs of sites that have mean evoked response to each site's preferred object between 10 and 30 spikes/s (blue). Arrows and numbers indicate mean Rsc.

For PCA, each PC consists of relative site activities (e.g., 1 × 64 matrix of coefficients for 64 sites, normalized to unit length) and stimulus-related scores (e.g., 1 × 240 matrix of weights for 240 stimuli) for that PC. The z-normalized response of a site to a stimulus can be back-calculated by summing, across all PCs, the product of the site's coefficient for each PC and the stimulus's score for that PC.

HUMAN TESTING

Observers

Procedures were approved by the Institutional Review Board of Georgetown University and informed consent was obtained from all observers. Six observers (3 male, 3 female, including the second author) participated in the experiments. All observers had normal or corrected-to-normal vision. Apart from the second author, observers were naïve as to the purpose of the experiment and were paid for participation.

Apparatus

Stimuli were controlled by computer using Matlab and Psychtoolbox 3 (Kleiner et al., 2007) and displayed on a 17″ cathode ray tube (CRT) (Sony Trinitron Multiscan 17sfII) with spatial resolution 1024 × 768 pixels and refresh rate of 60 Hz. Eye-screen distance was 57 cm, so that each pixel subtended approximately 0.03°. Ambient illumination was <4 Cd/m^2.

Stimuli

Object stimuli belonging to neurally defined features (grayscale rendered objects, "Set 1") were resized to 64 × 64 pixels (1.9° × 1.9°) and convolved with a 3 × 3 pixel Difference-of-Gaussians filter to match the background gray. Because the IT correlational structure is slightly more stable across stimulus sets for z-normalized responses than for raw responses, we constructed stimuli using the neurally defined features from z-normalized responses. Object stimuli were then equated for low-level image properties using the SHINE toolbox (Willenbockel et al., 2010). Groups of object stimuli were then randomly tiled to create the background (5 different objects), target (3 different objects), and distractors (3 different objects). Tiling position combinations were restricted to avoid lines of the same object. Target and distractors luminances

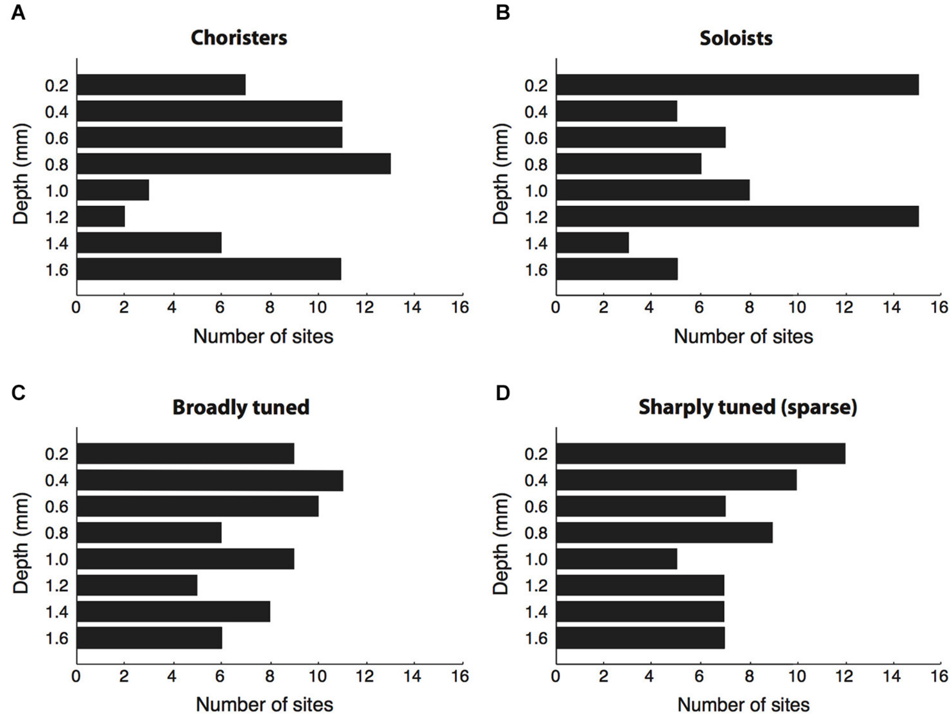


FIGURE 3 | Cortical depth vs. correlation strength and sparseness. (A,B) We sorted sites by their average strength of tuning correlation with other sites in the same array, then grouped the top ~30% of sites per array as "choristers" and the bottom ~30% as "soloists". Choristers are rarer in layer 4 (1.0–1.2 mm depth), whereas soloists are more common at 0.2 and 1.2 mm depth. The number of sites selected per group was higher for arrays 1–3 (16 choristers and 16 soloists per array) than for arrays 4 and 5 (8 choristers/soloists per array), because arrays 4 and 5 had fewer active channels. Average correlation strengths of choristers and soloists were 0.15 ± 0.04 and 0.02 ± 0.03, resp. **(C,D)** Same analysis based on sparseness (tuning sharpness). Average sparseness of broadly tuned and sharply tuned (sparse) sites were 0.12 ± 0.08 and 0.69 ± 0.25, resp.

were darkened by 5%, to make them more visible against the background. Mask stimuli were specific for each trial, created by scrambling the background (without target and distractors) at 0.24° resolution. Fixation point was a black square of 0.45° × 0.45°.

General procedure

Each block consisted of 144 trials comprising 48 "opposite", 48 "related", and 48 "unrelated" conditions, all with the same stimulus onset asynchrony (SOA) and adaptation duration. To minimize effects spanning across trials, each trial was preceded by an inter-trial interval (trial was initiated by key press), a blank fixation screen (1.5 s), and an adapting background of up to 8 s. In addition, objects were balanced across targets, distractors, and background and across conditions.

RESULTS

The neurophysiological data here is based on reanalysis of a previously reported dataset collected in monkeys under light neurolept anesthesia (Lin et al., 2014). Briefly, analyses are based on trial-averaged z-normalized responses (250 multi-unit "sites" and 6462 site pairs from 359 neurons) to stimuli that were presented via rapid serial visual presentation (**Figure 1A**). We begin by addressing a few concerns about our previous report: that the 6% cutoff of choristers is arbitrary and that in fact "choristers" and "soloists" are not two types of neurons, and that perhaps the better object coding performance of choristers is due to multiple detection of the same neuron across contacts, or because soloists are less visually driven. In fact, the distribution of average correlation strengths is continuous, and the separation into "choristers" and "soloists" is merely for convenience of comparison, not to say that there are two distinct cell types. Correlation strength and within-category generalization performance both decline smoothly, so missing a few top "choristers" during sampling should not affect the resulting structure very much. To increase the population size for testing the effect of average correlation strength, we relaxed the definition of choristers as random sites in the top 30%ile of average pairwise correlation strength per array (**Figure 2A**, brown), and of soloists as random sites in the median 30%ile (45–65%ile; black is lower 70%ile). This 30%ile threshold for choristers corresponds to minimum average correlation strengths of 0.12, 0.15, 0.16, 0.09, and 0.06 for the 5 arrays. These thresholds are similar for sites separated by at least 0.6 mm horizontal distance (0.09, 0.11, 0.14, 0.07, and 0.05). Although correlated neurons do tend to be more visually driven than uncorrelated neurons (c.f. Figure 5C of Tamura et al. (2014)), we still observed higher performance for choristers when choristers and soloists were matched for visual drive (**Figure 2B**, ~12 Sp/s baseline-subtracted response to each site's preferred stimulus; based on 5 arrays and 2 sites per array, at least 0.6 mm horizontal distance between sites).

SPARSENESS AND CORRELATION STRENGTH ARE MOSTLY UNRELATED

In previous reports, sparseness (measured as tuning sharpness) was thought to support efficient coding by reducing correlated activity (Young and Yamane, 1992; Rolls and Tovee, 1995; Baddeley, 1996; Olshausen and Field, 1996; Bell and Sejnowski, 1997; Vinje and Gallant, 2000; Zoccolan et al., 2007). This would predict that soloists should have better object coding capability (whereas our results suggest that choristers have better object coding, at least for within-category generalization) and that soloists should be sharply tuned. Conversely, a trivial explanation of the better object coding capability of choristers is that perhaps choristers are broadly tuned and therefore have better tolerance to stimulus variations.

We report that neither prediction is correct. Sparseness (measured as the modified sparseness index of Vinje and Gallant (2000), Lin et al. (2014) and average correlation strength are mostly uncorrelated across sites. Within each of 5 arrays (5 separate array insertions across 4 monkeys), the relationship between sparseness and average correlation strength was non-significant, and it was weak and barely significant when pooled across all arrays (Pearson $r = -0.09$, $p = 0.04$, $N = 250$ sites, **Figure 2A**). This weakness was not due to noise in either measurement, because sparseness and average correlation strength were each highly consistent across two stimulus sets ($r = 0.72$ and 0.70, $p < 10^{-37}$ and $p < 10^{-39}$, resp.). This dissociation between sparseness and correlation is consistent with a previous conjecture that these measures are unrelated (Willmore et al., 2011) and with a recent report that found a weak (albeit positive, $r = 0.07$, $p < 0.001$, rather than negative) dissociation in layer 2/3 (Tamura et al., 2014). Our data show that the dissociation also holds for a wider sample of IT neurons across supragranular, granular, and infragranular depths.

CORRELATED NEURONS ARE MOSTLY IN OUTPUT LAYERS

A key issue in linking neural activity to models is the cortical layer of different functional elements. An ongoing debate is whether neurons are correlated or uncorrelated, and whether these are in input or output layers. In V1, a recent study suggested that noise correlations are much lower than previously thought (Ecker et al., 2010), but alternatively it has been reported that noise correlation is layer-dependent and is lower, with better coding efficiency, in the granular layer (Hansen et al., 2012).

We suggest that neither view is entirely correct in IT. Here, we report that correlated neurons (choristers, with more efficient coding) are almost exclusively found in supragranular and infragranular layers. In IT, signal (tuning) correlation and noise correlation are related, and choristers tend to have stronger noise correlation (choristers (brown): Rsc = 0.13; soloists (45–65%ile, red): Rsc = 0.04; $p < 10^{-22}$, unpaired t-test; **Figure 2C**), including pairs separated by at least 0.6 mm horizontal distance and with similar visual drive (mean baseline-subtracted response to preferred stimulus of each cell is between 10 and 30 spikes/sec) (choristers: Rsc = 0.16; soloists: Rsc = 0.06; $p < 10^{-13}$; blue).

Of the 64 choristers, most were in supragranular and infragranular layers and only five were between 1.0–1.2 mm depth, near layer 4 (**Figure 3A**). Conversely, the most uncorrelated soloists (the ~30% of sites with the lowest correlation strength per array) were more prevalent at 0.2 and 1.2 mm depth (layers 1 and 4), although roughly half were in supragranular and infragranular layers (**Figure 3B**). The result was similar for single-unit activity. The proportion of choristers vs. soloists was significantly lower in the granular layer (1.0–1.2 mm) compared to supragranular and infragranular layers ($p = 0.0007$ and $p = 0.0003$, two-sided Fisher's test), and the difference between supragranular and infragranular layers was non-significant. This layer-specificity is consistent with a recent report in V1 that also measured correlated variability (Hansen et al., 2012). In contrast to correlation strength, sparseness, a measure of coding efficiency that is commonly based on tuning sharpness (Young and Yamane, 1992; Rolls and Tovee, 1995; Vinje and Gallant, 2000; Zoccolan et al., 2007), was not layer specific (**Figures 3C,D**, n.s. for all comparisons).

THE LOW DIMENSIONAL CORRELATIONAL STRUCTURE IS ALSO IN OUTPUT LAYERS

A recent perspective article (DiCarlo et al., 2012) highlighted the need to understand the processing of local populations of neurons at an intermediate level of abstraction. Covariation analysis (e.g., k-means clustering and PCA) is a useful form of abstraction because it directly ties the correlational structure to the idea of a low-dimensional manifold representation of object features (DiCarlo et al., 2012) and to our pipe cleaner model (Lin et al., 2014). We often encounter novel objects and novel environments (Vaziri et al., 2014) that must be categorized, and it is thought that the visual system learns useful shape statistics of the animal's environment (Srihasam et al., 2014). A key concept of the model is that the invariant representation, which supports generalization across rotation-in-depth, changes in illumination, and variations within an object category (studied here), has a spatial organization that is concentrated in a low-dimensional correlational structure. Such a low-dimensional correlational structure could be very useful for decoding by downstream neurons and for generalization learning, by providing a smoothly differentiable structure that is stable across categories, by reducing the number of inputs that must be pooled (instead of listening to all neurons, a downstream neuron could conceivably identify the most useful neurons in a population based solely on coincident timing, even during spontaneous activity), and by supporting robustness for noisy spiking populations.

We previously used k-means clustering (Lin et al., 2014) to identify clusters of sites that behaved more similarly across stimuli. Note that this is different from the typical approach, where the same data is clustered as groups of stimuli according to their response similarity (Kiani et al., 2007). Also, to focus on the local correlational structure, we focused our analysis specifically within each array, rather than pooling across the entire population (all arrays). Here, we extend our approach to PCA, to tie the low-dimensional structure to output layers and to behavior. Because many of the conclusions drawn from PCA regarding

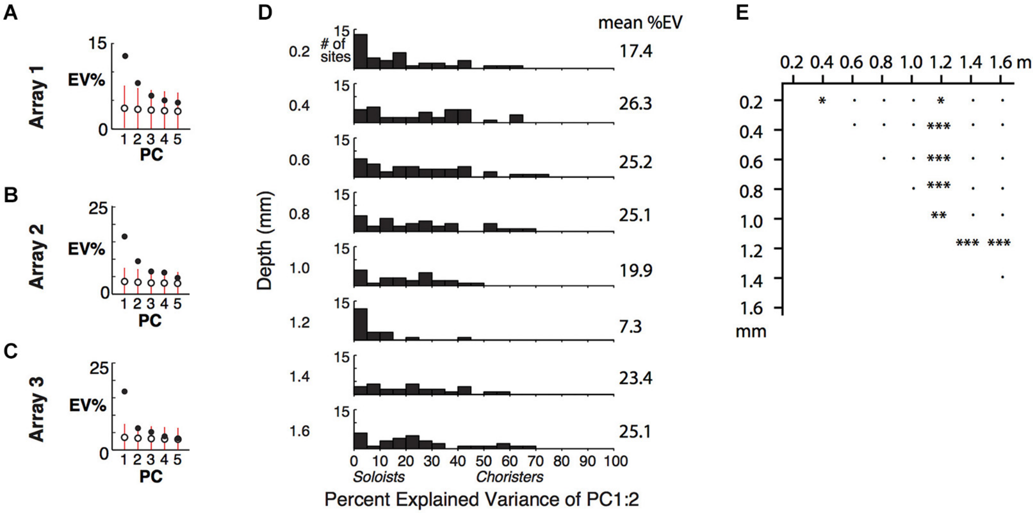


FIGURE 4 | Cortical depth vs. low-dimensional correlational structure. (A) Percent of response variance explained by PCs 1–5, based on z-normalized responses. Chance and 5–95%ile distributions are indicated by open circles and red bars, based on shuffling of response IDs across trials. **(B,C)** Explained variance for Arrays 2 and 3, from two separate array insertions (separate recording sessions) in monkey 2. **(D)** Cortical depth vs. percent explained variance of PCs 1 and 2 across 5 arrays. **(E)** Comparison of distributions in **(D)** among different depths. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, •p = n.s.

spatial structure are similar to those from k-means clustering and from pairwise site correlations, we will discuss only the highlights below.

Compared to k-means clustering, PCA has the advantage of explaining more of the variance using fewer dimensions, because each PC is exactly aligned to maximally explain the remaining variance, whereas k-means clustering will include sites that are uncorrelated (soloists). Thus, whereas k-means clustering tends to highlight columnar organization (changes in tuning across cortex), PCA can reveal the layer-specificity of the low-dimensional correlational structure. For these dense arrays, the spatial patterns of site covariation were nearly identical between the lower PCs and k-means clustering and corresponded roughly to the differential activation of neighboring columns (at higher PCs, it is less likely that the orthogonal PCs are relatable to biological processes). Specifically, PC1 (the dimension of maximum variance within an array) corresponds roughly to activation vs. suppression of most sites in the array, which may contribute to invariant representation by encoding how strongly a feature or feature contrast is present within an object category. PC2 (the dimension that best explains the remaining variance) corresponds roughly to the differential activation of two neighboring cortical columns (i.e., the sign and relative strength of a feature contrast) and appears virtually identical to k-means clustering at $k = 2$ (Figure 5 of Lin et al., 2014).

By examining how well the lowest PCs explain the variance of individual sites, we can determine the layer-specificity of the low-dimensional correlational structure. Because of the high proportion of soloists and the rarity of choristers (**Figure 2A** and Gawne and Richmond, 1993; Sato et al., 2009; Lin et al., 2014), the first few PCs explained only a small fraction of the total variance within each array (**Figures 4A–C**, example arrays 1–3; responses were averaged across repetitions and z-normalized), even though the tuning of individual sites was highly consistent across even vs. odd trials ($r \sim 0.6$–0.9, Figure 2A of Lin et al. (2014)). The first two PCs explained only about 25% of the response variance across stimuli. A previous study reported 15% explained variance for two PCs, based on recordings from random penetrations across IT (Baldassi et al., 2013), and another study reported ~70% for two PCs (~60% for PC1) based on recordings from electrode bundles targeted to the centers of IT optical imaging domains (Figures 10, 16 of Sato et al. (2009)). A possible explanation for the large difference across studies is that in addition to layer-specific heterogeneity of choristers vs. soloists, there is also topographical heterogeneity that cannot be attributed to bias from image guided electrode targeting.

For individual sites, the percent of variance explained (%EV) by the first two PCs varied widely and was lower for sites in layers 1 and 4 (0.2 and 1.2 mm depths) (**Figures 4D,E**). It was more widely distributed, with up to 77%EV, for sites in supragranular and infragranular layers. The specificity of the correlational structure to output layers hints that it is shaped by local networks, rather than by feedforward or thalamic input to layer 4 or feedback from higher areas to layer 1. The lower explained variance at 0.2 mm and 1.2 mm depths (many sites are <5%EV), despite good even-vs.-odd trial tuning consistency at all depths (Figure 2A of Lin et al. (2014)), suggests that the inputs to layers 1 and 4 are nearly orthogonal to the correlated activity, consistent with our "pipe cleaner" model. In comparison, scrambling the stimulus IDs of the scores of the first two PCs, without altering the PCA coefficients, resulted in −23%EV on average (i.e., the difference between the scrambled prediction and the actual response has a total variance that is on average 123% that of each site's actual total variance, **Figure 4D**). The fact that the EVs of scrambled predictors are negative, and not zero, for sites in all layers further supports that these sites are visually driven and selective. Only layer 4 (1.2 mm depth) had average response below baseline (−0.5 spikes/s), consistent with suppression by local or feedforward inhibition. Overall, this result extends upon the layer-specificity of choristers and soloists as measured by average correlation strength by showing that the correlational structure is concentrated in a few dimensions (a low-dimensional manifold), mainly in a subpopulation of neurons in output layers.

NEURALLY DEFINED FEATURES BASED ON CORRELATED ACTIVITY

To link the correlational structure to behavior, we constructed "neurally defined features" based on the tuning of neighboring IT columns. Previous reports used a variety of methods (feature reduction, k-means clustering, PCA, or simply averaging the tuning along a penetration) to characterize IT tuning (Young and Yamane, 1992; Tsunoda et al., 2001; Baldassi et al., 2013; Sato et al., 2013). However, because their analyses focused on the tuning of single neurons or random IT populations, behavior has not been tied to the concept of a cortically local low-dimensional manifold and lateral inhibition.

Here, to focus on the differential coding by local populations, we defined "neurally defined features" as sets of stimuli determined by PCA of each array (the same PCs as in **Figure 4**, e.g., computed from a 240 stimuli × 64 site matrix of z-normalized tuning responses). The "neurally defined features" are sets of stimuli with extreme PCA scores, treated collectively without altering or blending the images. For example, feature "Array 1 PC1+" is the set of 10 stimuli with the most positive PC1 scores for Array 1, allowing that some of the same stimuli may also belong to PC2+ or PC2− or to features of another array (however, for behavioral testing we did not allow "reuse" of stimuli across background, target, and distractor within a trial). Although one "Array 1 PC1+" stimulus may have a higher PC1 score than another "Array 1 PC1+" stimulus, we treat them equally because it is the collective effect of the set of "Array 1 PC1+" stimuli that dilutes away stimulus-specific effects, to increase the feature's specificity to that neuronal population (e.g., vs. populations in early visual cortex or elsewhere in IT). In each array, PC2+ and PC2− correspond to differential activation of neighboring cortical columns, and PC1+ and PC1− correspond to co-activation and co-suppression of most neurons in the array (differential activation at a larger spatial scale).

Figure 5A shows examples of stimulus responses along PCs 1 and 2 for Array 1. The red and blue matrices show examples of baseline-subtracted firing rates across the 8 × 8 array to specific stimuli. Across different levels of overall activation and suppression (different PC1 scores), stimuli that differentially activated the left column more than the right column (PC2− stimuli) tended to be objects with protrusions, whereas stimuli that differentially activated the right column more than the left column (PC2+ stimuli) tended to be objects with internal features. Although these semantic descriptions are qualitative and are not part of the feature definition, the positive and negative PC extremes appeared to prefer contrastive features ("rumpled" vs. "smooth", "upward" vs. "downward vertex with gradient") that were consistent across two stimulus sets (grayscale rendered objects and color/grayscale/silhouette photographed objects). Therefore, we assigned the positive and negative extremes to separate features, resulting in 12 neurally defined features derived from 6 PC feature dimensions (PCs 1 and 2 from 3 arrays, i.e., 3 separate recording sessions across 2 monkeys, **Figure 5B**).

Why use PCA, instead of k-means clustering or penetration averaging? Because neighboring columns have correlated tuning and because k-means clustering does not distinguish soloists from choristers, features from one k-means cluster are less visually distinguishable from those of neighboring clusters. Unlike PCA, k-means clustering or penetration averaging would have ordered stimuli according to how strongly they activated each column, causing stimuli that appear very different to group together (e.g., the bike (#46) and the fence (#11) for the right column, or the shears (#128) and the couch (#87) for the left column). PCA features appear more different, particularly PC2+ vs. PC2−. We note that this advantage of PCA may be specific to local populations sampled by densely spaced electrode arrays. The differential coding along PC2 is consistent with previous reports of a "shape-contrast" effect in perception (Suzuki and Cavanagh, 1998) and in IT responses (Leopold et al., 2006), although it is distinct from the idea of norm-based encoding (Valentine, 1999) because it is primarily driven by shape rather than by semantic category or low-level properties such as color and texture (Baldassi et al., 2013; Lin et al., 2014).

These PC feature dimensions were uncorrelated across arrays (Pearson correlations of PC scores were non-significant), even the features measured from different sessions 3 mm apart in the same monkey (M2), indicating that the features are not simply due to familiarity (Mruczek and Sheinberg, 2005, 2007b; Hein et al., 2007; Anderson et al., 2008) or coarse topography (Op de Beeck et al., 2007; Sato et al., 2013). Also, the monkeys had never seen these stimuli previously.

NEURALLY DEFINED FEATURES PREDICT VISUAL SEARCH EFFICIENCY

To link correlated activity to behavior, we designed a human visual search task in which the target, distractors, and background were disjoint sets of objects from monkey neurally defined features. Previous reports based on simple features such as orientation, color, and size hint that visual performance is associated with horizontal processes and lateral inhibition in early visual cortex (Butler et al., 2008; Yoon et al., 2010; Michel et al., 2013). Here, we asked whether lateral inhibition among complex feature representations in IT might also predict visual performance. Because of the short horizontal range of lateral inhibition in macaque IT, which may translate to longer-range inhibition in humans if similar statistical feature mechanisms are useful for representation, we tested whether the target would be more salient from the background if they were contrastive ("opposite" sign) features from the same array (differentially activating neighboring columns, e.g., Array 2 PC2+ target vs. PC2− background), than if they were "related" features (different PCs of the same array, activating the same column at different scales, e.g., Array 2 PC2+ target vs. PC1+ background) or "unrelated" features (PCs from different arrays, activating distant

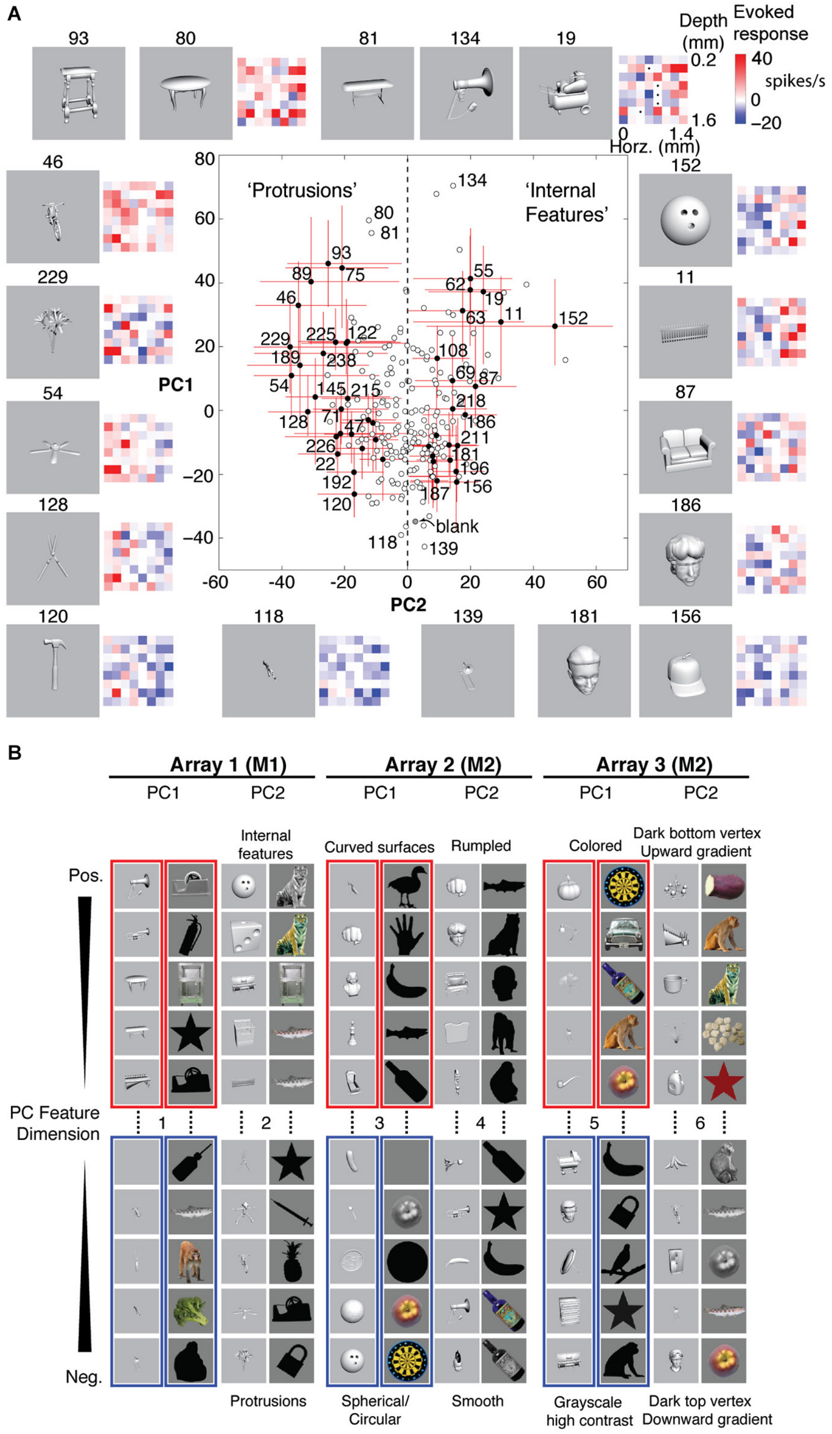


FIGURE 5 | Neurally defined features. (A) Neurally defined features based on Array 1's PC1 and PC2 scores. Each red and blue 8 × 8 matrix shows baseline-subtracted response to one stimulus across the 64 sites, spanning all depths and neighboring IT columns. PC1+ stimuli activated most sites. PC2+ and PC2− stimuli differentially activated sites on the right and left sides of the array. Numbers indicate stimulus IDs. Black dots in stimulus 19's matrix indicate inactive sites. Red lines indicate 5–95%ile, and filled circles indicate stimuli with significant PC2. "Protrusions" and "Internal Features" are labels to help see the pattern of PC2− and PC2+ stimuli, but the labels are not part of the feature definition. **(B)** Features from 3 array insertions (3 recording sessions) in two monkeys and two stimulus sets. Only the stimuli with the most extreme scores are shown, out of 240 object stimuli for set 1 (grayscale rendered 3D objects) and 113 stimuli for set 2 (color, grayscale, and silhouette photographs). The slight difference between panels **(A)** and **(B)** is because the scores in A are calculated from unnormalized responses, to scale with the baseline-subtracted firing rates in the matrices, whereas the scores in **(B)** are from z-normalized responses, for better consistency of spatial covariation patterns across stimulus sets.

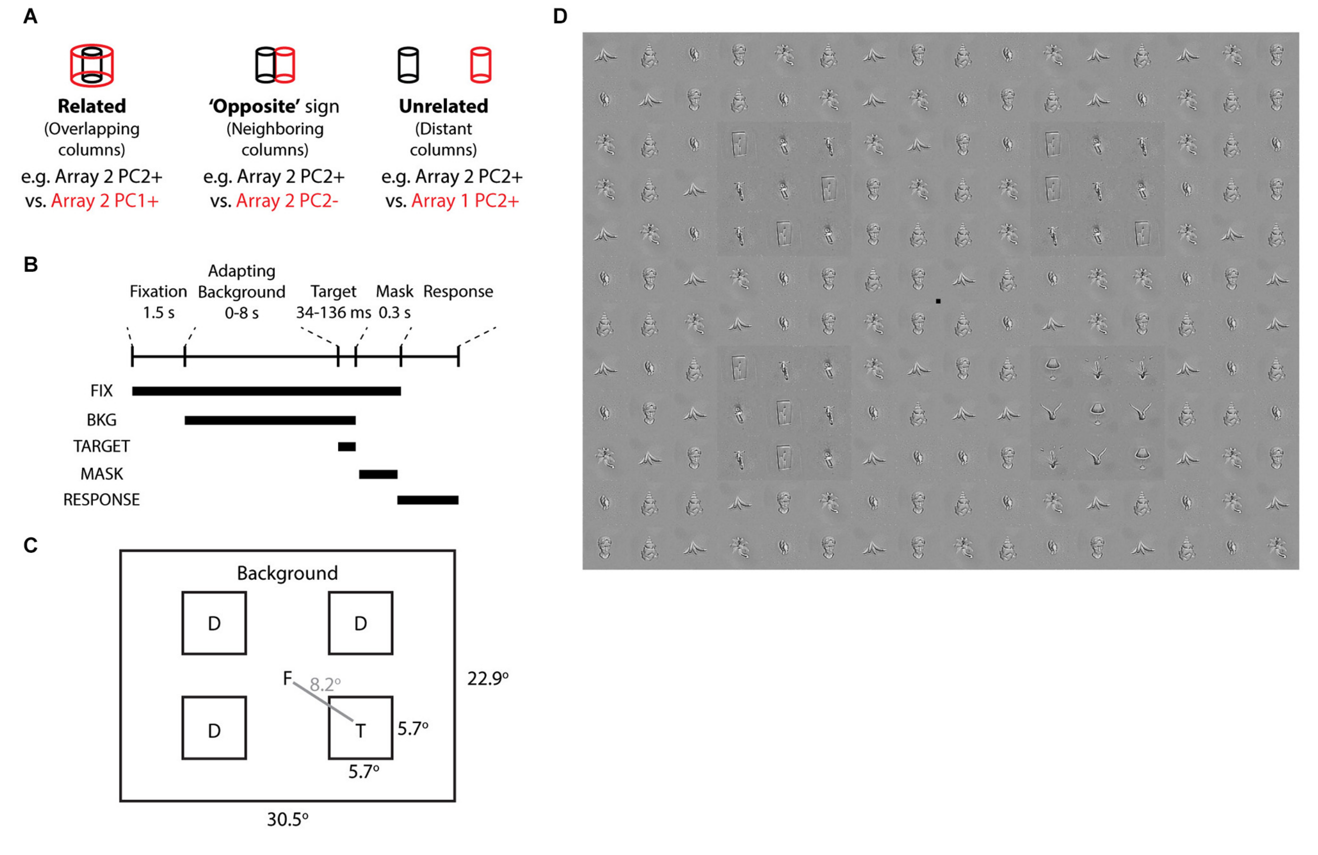


FIGURE 6 | Visual search task based on neurally defined features. **(A)** "Related", "Opposite", and "Unrelated" conditions are tied to the differential activation of overlapping, neighboring, and distant IT columns by neurally defined features. In each condition, the target objects belong to one feature (e.g., Array 2 PC2+) and the distractor and background objects are disjoint sets belonging to the other feature (e.g., Array 2 PC2−). "Related" features are from different PCs of the same array. "Opposite" features are from opposite signs of the same PC of the same array. "Unrelated" features are from different arrays. **(B)** Time course of each trial. Following Fixation screen and Adapting Background (0–8 s), Target and Distractors appeared for 34–136 ms, followed by a Mask with tile-scrambled background images. After disappearance of the fixation point, subjects reported via keypress the target quadrant. **(C)** Each trial consisted of one target and 3 distractors at four possible locations. Objects and target locations were balanced across all conditions. **(D)** Example stimulus from "opposite" condition, with target in quadrant 4. Distractors and background are from Array 3 PC2−. Target is from Array 3 PC2+. All object stimuli were matched for low level image properties via the SHINE toolbox. To aid target localization, a luminance pedestal was added to target and distractors.

columns >3 mm apart, e.g., Array 2 PC2+ target vs. Array 1 PC2+ background) (**Figure 6A**). The distractors and background were disjoint sets of objects sharing the same neurally defined feature.

To induce a temporary visuoperceptual distortion as in Leopold et al. (2001), we adapted the subject to the background for up to 8 s, followed by a brief (34–136 ms) presentation of the target and distractors and then a mask (scrambled background) (**Figures 6B,C**). Subjects indicated via key press the quadrant in which the target appeared. Subjects were instructed to search for the quadrant whose pattern appeared different from the other three quadrants. We measured visual search efficiency as the reporting accuracy of the target quadrant (chance = 25% correct). Such visual search displays are commonly used to study early perceptual processes and have only recently been applied to neurally related complex shapes (Sripati and Olson, 2010). A strength of the task is that the brief stimulus appearance and the target location randomization preclude artifacts from differences in spatial attention or eye position. To focus the task on complex shapes rather than early visual processes, we used the SHINE toolbox (Willenbockel et al., 2010) to equalize the objects in terms of low-level cues including luminance, contrast, and orientation-specific Fourier power (including spatial frequency) (**Figure 6D**).

We began by comparing, in one subject, how performance depended on stimulus condition, target duration (stimulus onset asynchrony "SOA" between target/distractors and mask) and adaptation duration. At 34 ms SOA, performance was consistently higher across different durations of adaptation when the target feature was "opposite" in sign to the distractors and background (i.e., when target and background were contrastive features that differentially drive neighboring IT columns) (**Figure 7A**, blue) than when target and background were "related" features (red) (p = 0.038, Cochran-Mantel-Haenszel test). Although performance was slightly higher at 2 s adaptation, the odds ratios were not heterogeneous across different levels of adaptation

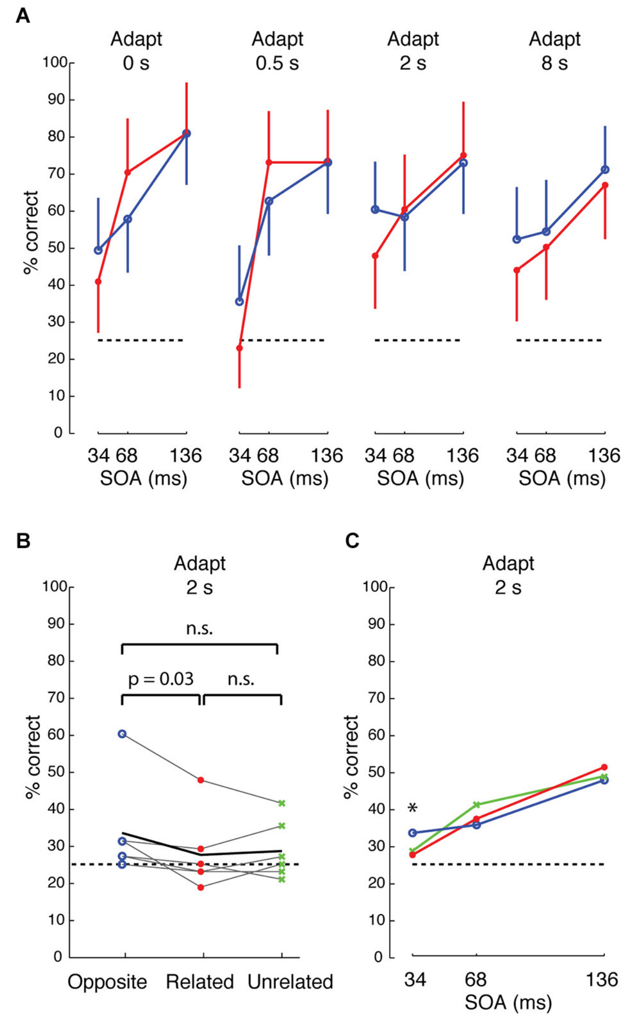


FIGURE 7 | Visual search performance for neurally defined features. **(A)** Performance of one subject for target whose neurally defined feature is "opposite" (blue) or "related" (red) to that of the distractors and background, across different adaptations and different stimulus onset asynchrony (SOA) between stimulus and mask. Accuracy is higher for "opposite" at 34 ms SOA, and the difference between "opposite" and "related" is more consistent at longer adaptation. Error bars show 95% CI, based on 48 trials per condition. **(B)** Performance across 6 subjects at 2 s adaptation and 34 ms SOA for "opposite" (blue), "related" (red), and "unrelated" (green) conditions. Black line indicates average performance across subjects. The difference between "opposite" and "related" is significant at $p = 0.03$, based on Wilcoxon signed ranks test. **(C)** Average performance across 6 subjects at 34, 68, and 136 ms SOA, 2 s adaptation.

(p = n.s., Breslow-Day test). An opposite effect was seen at 68 ms SOA at 0 and 0.5 s adaptation, but the effect reversed with longer adaptation. At the longest adaptation (8 s), performance was higher for "opposite" than for "related" features at all SOAs. Based on this pattern, we surmised that the effect was most reliably consistent with the prediction at the shortest SOA (34 ms) and with longer adaptation. To avoid tiring the subjects with the 8 s long adaptation or a possible flooring effect at shorter adaptation (e.g., 0.5 s), we tested all subjects at 2 s adaptation.

Across 6 subjects, visual search efficiency was consistently higher for "opposite" features than for "related" features at 34 ms SOA and 2 s adaptation (**Figure 7B**, $p = 0.03$, Wilcoxon signed rank test). As with the first subject, the higher performance for "opposite" features was only observed at the shortest SOA of 34 ms across the 6 subjects (**Figure 7C**). The persistence of the effect across different durations of adaptation at the short 34 ms SOA hints that the effect is likely driven by feed-forward processing and short-range lateral interactions in IT, because 34 ms is likely too brief for feedback (Bansal et al., 2014; Scholl et al., 2014) or long-range lateral interactions (Singer and Kreiman, 2014; Tang et al., 2014). We suggest that the mechanism is associated with short-range lateral inhibition (e.g., between neighboring columns) in IT, similar to reports of lateral interactions in early visual cortex (Das and Gilbert, 1999; Michel et al., 2013), rather than a distance-dependent effect in IT, because the search efficiency of "related" (0 mm cortical separation) and "unrelated" (>3 mm separation) features was not significantly different. Also, "opposite" features had higher search performance than "unrelated" features in 4 of 6 subjects, but this difference did not reach significance.

DISCUSSION

Our results suggest that correlated activity contributes to efficient coding and human visual search efficiency. The main findings are that correlation strength and sparseness are only weakly related and should be considered as separate factors, that correlated activity is primarily located in output layers, and that correlated activity in monkey IT predicts human visual search efficiency. Together, these results suggest that correlated activity may be the substrate of IT's output and that, contrary to previous reports, correlated activity contributes to coding efficiency.

"POPULATION SPARSENESS" VS. "SPARSENESS" IN EFFICIENT CODING

These results suggest that a fundamental shift is needed in our approach to understanding efficient coding. Previous reports of efficient coding assumed that population sparseness and tuning sharpness (conventionally termed "sparseness"; Rolls and Tovee, 1995; Vinje and Gallant, 2000; Zoccolan et al., 2007; Willmore et al., 2011) are interchangeable. Instead, our results suggest that correlation strength (inversely related to population sparseness) is better than tuning sharpness as a measure of population redundancy, and that these two measures are mostly unrelated. Surprisingly, they show that the representation is more correlated in output layers than in input layers, which is opposite to the expectation that increasing sparseness supports efficient coding. This layer-specific increase in correlation is unrelated to tuning sharpness. Together with our previous report showing the better object coding capability of choristers vs. soloists, these results highlight the role of correlation in efficient coding.

WHY ARE CORRELATION STRENGTH AND SPARSENESS UNRELATED?

The main reason for this apparent discrepancy is that previous studies did not measure population sparseness. Their wider electrode spacing meant that neuronal tuning was too dissimilar to compute population sparseness. Dense sampling, on the order of 64 neurons per mm^3, is necessary to measure population sparseness, because neuronal tuning is heterogeneous even within a cortical column and because choristers are rare (Sato et al., 2009; Lin et al., 2014). It is unclear what mechanism might enable correlation of sharply tuned neurons in output layers and decorrelation of broadly tuned neurons in input layers. The prevalence of soloists in input layers 1 and 4 suggests that the feedforward and feedback inputs to IT are already decorrelated, or that they are actively decorrelated by inhibition. Conversely, our finding that correlated activity is mostly in output layers is consistent with the layer specificity of local circuits and horizontal fibers. However, the consistency of our V1 and IT results in terms of tuning and spike timing correlational structure suggest that they are probably driven more by local circuitry than by long range fibers, which have different patterns in V1 vs. IT (Tanigawa et al., 2005).

IMPLICATIONS FOR VISUAL SEARCH EFFICIENCY

Overall, these results support that a human homolog of IT, previously shown by many studies (Grill-Spector et al., 2001; Tootell et al., 2003; Orban et al., 2004; Kriegeskorte et al., 2008), guides search based on complex features. In relation to classical theories of visual search based on feature integration theory (FIT; Treisman and Gelade, 1980), these results differ in two key aspects. First, whereas FIT posits that fast visual search relies on early visual areas, our results support an accumulating body of evidence that later visual areas also contribute to fast visual search (Hochstein and Ahissar, 2002). Second, FIT posits that preattentive, parallel search is more efficient for low-level features than for feature conjunctions. Our results show that preattentive, parallel search is also more efficient for specific types of complex features, contextually dependent on the complex features present in the background, and that this contextual dependency is specifically linked to cortical neighborhood relationships and correlated activity in IT. This supports a model by Duncan and Humphreys that all search is parallel and depends on representational similarity and competition for resources across multiple levels of the visual system (Duncan and Humphreys, 1989, 1992).

Consistent with a previous report that linked macaque IT responses to human visual search efficiency (Sripati and Olson, 2010), our results suggest that this context dependency is due to a stimulus-specific competition for resources that can be explained by local contrastive mechanisms such as lateral inhibition (Wang et al., 2000; Leopold et al., 2001). Our results strengthen the case that this mechanism is tied to competition for local resources in IT, vs. in earlier areas, because it depends on IT cortical proximity. Also, by linking search efficiency to correlational structure, our results support an assumption in the previous report (Sripati and Olson, 2010), that a population of heterogeneous neurons (e.g., within an IT column) can be modeled by the discriminative capacity of their correlated activity, as the activity of a few neurons (choristers). One difference from the previous report (Sripati and Olson, 2010) is that their behavioral and neural responses were predicted by the coarse footprint difference of the objects, i.e., the spatial overlap of the blurred images, whereas in our data the coarse footprint difference does not predict better performance in the "opposite" condition (unpaired *t*-test of distributions of coarse footprint index in correct vs. incorrect trials was non-significant). This difference, together with our use of brief presentations and masking, further supports that low-level features are insufficient to account for our results. Also, because the correlational structure was tied to shape rather than semantic category (Lin et al., 2014), and because there was no difference in category overlap across stimulus conditions, our results support that the contrastive mechanism was feature-based, not semantically-based.

IMPLICATIONS FOR COMPUTATIONAL MODELS OF RECOGNITION

An ongoing debate in computational modeling of recognition and generalization learning is how to design the architecture, e.g., whether it is necessary to simulate populations of binary spiking neurons (Masquelier and Thorpe, 2007; Chan et al., 2011; Merolla et al., 2014), or whether convolutional networks are sufficient or even superior. Although convolutional networks outperform spiking networks on datasets like ImageNet, and their performance approximates ideal observers on object categorization and exceeds that of randomly sampled IT neurons (surprising because IT is the last stage of the ventral pathway) (Krizhevsky et al., 2012; Cadieu et al., 2013, 2014; Zeiler and Fergus, 2014), their performance remains far worse than that of humans on real-world vision. In a recent model that approximates IT and ideal observers (Yamins et al., 2014), the approximation to IT is as low as 20%EV for single sites (mean 48.5%EV), and IT split-half data still outperforms the best model on predicting representational dissimilarity for image generalization, object generalization, and category generalization. Our results suggest that part of this gap is due to the much poorer coding capability of soloists vs. choristers and due to the rarity of choristers. We suggest that the comparison (for both convolutional and spiking networks) should be against IT choristers, rather than a random pool of IT neurons. Also, our layer-specific correlation results show that correlation strength increases from input to output layers within a cortical column. Increasing the cell and layer specificity of modeling could in principle favor spiking network models with individual cores that simulate computations within cortical columns, as in Merolla et al. (2014).

Another aspect that may favor spiking network models is the relationship between correlational structure and learning. Whereas learning in convolutional networks occurs via genetic algorithms that guide connection patterns based on overall performance, learning in spiking networks is more local and can in principle be tied to our "pipe cleaner" model and spike timing dependent plasticity (STDP). This difference in approach manifests in convolutional networks as a gradual increase in performance along the hierarchy (Serre et al., 2007), whereas in our data the near-chance performance of soloists hints that performance increase may be staggered along the hierarchy, alternating between low performing input layers and high performing output layers for each cortical "area". This alternation between low-performing soloists and high-performing choristers may be critical to learning and maintaining an invariant representation (cortically local subspace untangling; DiCarlo et al., 2012). Another aspect that could be modeled is that choristers are rarer in IT (Lin et al., 2014) than in V1 (Chu et al., 2014). We speculate that the increasing rarity of choristers is because of increasing complexity (increasingly high-dimensional feature spaces) along the visual hierarchy, which requires ever larger and more heterogeneous populations of soloists within each column to develop and maintain an invariant low-dimensional (manifold) representation. These emphases on local learning could also benefit from architectures based on multicore networks.

HOW MIGHT THE CORRELATIONAL STRUCTURE FUNCTION ALGORITHMICALLY?

How might a homogeneous/heterogeneous spiking network support generalization learning? Our conceptual "pipe cleaner" model (Lin et al., 2014) predicts that the feedforward and feedback inputs to IT may act as tensors, enabling the fine adjustments that may be necessary to build and maintain an invariant representation. The near-orthogonality of the inputs vs. the manifold (**Figure 4D**) indicates that they are optimally tuned to alter the manifold (i.e., co-alignment with the manifold would be inefficient and could conceivably result in uneven coverage). Such adjustments could occur via STDP, because soloists (mostly in the input layers) that are better tuned to the feedforward (environmental) and feedback (behavioral context) input statistics will spike more quickly, shaping the tuning of the choristers that support the invariant representation. This prediction is consistent with a recent report that found layer-specific temporal sequencing in perirhinal cortex (Takeuchi et al., 2011). Because the invariant representation in IT is based on shape rather than semantic category, invariance training on any category would also improve invariance to other categories that share the same feature, supporting generalization from few examples. We speculate that the combination of heterogeneity (population sparseness) in the input layers and redundancy/smoothness (overlap in tuning) in the output layers may be important for populations of spiking neurons, to achieve sufficient bit resolution from binary spiking neurons operating in high dimensional feature space. This problem of poor bit resolution has been criticized as a fundamental weakness of spiking network models vs. convolutional network models of recognition, and homogeneous/heterogeneous networks may be a key part of the brain's solution.

ON TECHNICAL APPROACHES TO STUDY EFFICIENT CODING

Recent technical advances have improved cell-specificity, sampling density, and anatomical co-registration. Our results suggest that, to better understand how the local correlational structure contributes to efficient coding, simultaneous sampling across multiple depths down to at least 1.0–1.2 mm is essential, to map both the inputs and the outputs within a column. Sampling density on the order of 64 neurons/mm^3 is also critical, to measure correlational structure (not just tuning sharpness) and to detect choristers that are rare in IT. Finally, high temporal resolution is necessary, to link the correlational structure to mechanisms such as STDP and to learning behavior. Currently, dense electrode arrays (e.g., the NeuroNexus Matrix Array) are the only technology that meets these design requirements in terms of deep sampling and sampling individual spikes *in-vivo* in behaving mammals. Unlike other technologies that are still in development, this technology is available today, and its potential for transforming neuroscience remains largely untapped.

ACKNOWLEDGMENTS

We thank M. Riesenhuber and B. Taba for helpful discussions. Supported by Georgetown University Medical Center GD4235619 and by the Taiwan Ministry of Education Five-Year Aim for the Top University Plan, NSC-96-2811-B-010-501 and NSC-97-2321-B-010-007.

REFERENCES

Abbott, L. F., and Dayan, P. (1999). The effect of correlated variability on the accuracy of a population code. *Neural Comput.* 11, 91–101. doi: 10.1162/089976699300016827

Anderson, B., Mruczek, R. E., Kawasaki, K., and Sheinberg, D. (2008). Effects of familiarity on neural activity in monkey inferior temporal lobe. *Cereb. Cortex* 18, 2540–2552. doi: 10.1093/cercor/bhn015

Averbeck, B. B., Latham, P. E., and Pouget, A. (2006). Neural correlations, population coding and computation. *Nat. Rev. Neurosci.* 7, 358–366. doi: 10.1038/nrn1888

Baddeley, R. (1996). Searching for filters with 'interesting' output distributions: an uninteresting direction to explore? *Network* 7, 409–421. doi: 10.1088/0954-898x/7/2/021

Baker, C. I., Behrmann, M., and Olson, C. R. (2002). Impact of learning on representation of parts and wholes in monkey inferotemporal cortex. *Nat. Neurosci.* 5, 1210–1216. doi: 10.1038/nn960

Baldassi, C., Alemi-Neissi, A., Pagan, M., Dicarlo, J. J., Zecchina, R., and Zoccolan, D. (2013). Shape similarity, better than semantic membership, accounts for the structure of visual object representations in a population of monkey inferotemporal neurons. *PLoS Comput. Biol.* 9:e1003167. doi: 10.1371/journal.pcbi.1003167

Bansal, A. K., Madhavan, R., Agam, Y., Golby, A., Madsen, J. R., and Kreiman, G. (2014). Neural dynamics underlying target detection in the human brain. *J. Neurosci.* 34, 3042–3055. doi: 10.1523/jneurosci.3781-13.2014

Bell, A. J., and Sejnowski, T. J. (1997). The "independent components" of natural scenes are edge filters. *Vision Res.* 37, 3327–3338. doi: 10.1016/s0042-6989(97)00121-1

Bengio, Y. (2009). Learning deep architectures for AI. *Found. Trends Mach. Learn.* 2, 1–127. doi: 10.1561/2200000006

Brown, E. N., Purdon, P. L., and Van Dort, C. J. (2011). General anesthesia and altered states of arousal: a systems neuroscience analysis. *Annu. Rev. Neurosci.* 34, 601–628. doi: 10.1146/annurev-neuro-060909-153200

Butler, P. D., Silverstein, S. M., and Dakin, S. C. (2008). Visual perception and its impairment in schizophrenia. *Biol. Psychiatry* 64, 40–47. doi: 10.1016/j.biopsych.2008.03.023

Cadieu, C. F., Hong, H., Yamins, D. L. K., Pinto, N., Ardila, D., Solomon, E. A., et al. (2014). Deep neural networks rival the representation of primate IT cortex for core visual object recognition. *arXiv:*1406.3284.

Cadieu, C. F., Hong, H., Yamins, D., Pinto, N., Majaj, N. J., and Dicarlo, J. J. (2013). "The neural representation benchmark and its evaluation on brain and machine," in *International Conference on Learning Representations (ICLR)*, (Scottsdale, AZ). *arXiv:*1301.3530.

Carandini, M. (2014). "Soloists and choristers in a cortical population," in Computational and Systems Neuroscience Workshop: Scalable models for high-dimensional neural data (Snowbird, UT).

Chan, V. H., Hunzinger, J. F., and Behabadi, B. F. (2011). *Method and Apparatus for Neural Temporal Coding, Learning and Recognition.* USA patent application 13/211,091.

Chu, C. C. J., Chien, P. F., and Hung, C. P. (2014). Tuning dissimilarity explains short distance decline of spontaneous spike correlation in macaque V1. *Vision Res.* 96, 113–132. doi: 10.1016/j.visres.2014.01.008

Cohen, M. R., and Kohn, A. (2011). Measuring and interpreting neuronal correlations. *Nat. Neurosci.* 14, 811–819. doi: 10.1038/nn.2842

Constantinople, C. M., and Bruno, R. M. (2011). Effects and mechanisms of wakefulness on local cortical networks. *Neuron* 69, 1061–1068. doi: 10.1016/j.neuron.2011.02.040

Contreras, D., Destexhe, A., Sejnowski, T. J., and Steriade, M. (1997). Spatiotemporal patterns of spindle oscillations in cortex and thalamus. *J. Neurosci.* 17, 1179–1196.

Das, A., and Gilbert, C. D. (1999). Topography of contextual modulations mediated by short-range interactions in primary visual cortex. *Nature* 399, 655–661.

Dehaene, S., and Changeux, J. P. (2005). Ongoing spontaneous activity controls access to consciousness: a neuronal model for inattentional blindness. *PLoS Biol.* 3:e141. doi: 10.1371/journal.pbio.0030141

DiCarlo, J. J., and Cox, D. D. (2007). Untangling invariant object recognition. *Trends Cogn. Sci.* 11, 333–341. doi: 10.1016/j.tics.2007.06.010

DiCarlo, J. J., Zoccolan, D., and Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron* 73, 415–434. doi: 10.1016/j.neuron.2012.01.010

Duncan, J., and Humphreys, G. W. (1989). Visual search and stimulus similarity. *Psychol. Rev.* 96, 433–458. doi: 10.1037//0033-295x.96.3.433

Duncan, J., and Humphreys, G. (1992). Beyond the search surface: visual search and attentional engagement. *J. Exp. Psychol. Hum. Percept. Perform.* 18, 578–588. doi: 10.1037//0096-1523.18.2.578

Ecker, A. S., Berens, P., Cotton, R. J., Subramaniyan, M., Denfield, G. H., Cadwell, C. R., et al. (2014). State dependence of noise correlations in macaque primary visual cortex. *Neuron* 82, 235–248. doi: 10.1016/j.neuron.2014.02.006

Ecker, A. S., Berens, P., Keliris, G. A., Bethge, M., Logothetis, N. K., and Tolias, A. S. (2010). Decorrelated neuronal firing in cortical microcircuits. *Science* 327, 584–587. doi: 10.1126/science.1179867

Ecker, A. S., Berens, P., Tolias, A. S., and Bethge, M. (2011). The effect of noise correlations in populations of diversely tuned neurons. *J. Neurosci.* 31, 14272–14283. doi: 10.1523/jneurosci.2539-11.2011

Eyherabide, H. G., and Samengo, I. (2013). When and why noise correlations are important in neural decoding. *J. Neurosci.* 33, 17921–17936. doi: 10.1523/jneurosci.0357-13.2013

Froudarakis, E., Berens, P., Ecker, A. S., Cotton, R. J., Sinz, F. H., Yatsenko, D., et al. (2014). Population code in mouse V1 facilitates readout of natural scenes through increased sparseness. *Nat. Neurosci.* 17, 851–857. doi: 10.1038/nn.3707

Fujita, I., Tanaka, K., Ito, M., and Cheng, K. (1992). Columns for visual features of objects in monkey inferotemporal cortex. *Nature* 360, 343–346. doi: 10.1038/360343a0

Gawne, T. J., and Richmond, B. J. (1993). How independent are the messages carried by adjacent inferior temporal cortical neurons? *J. Neurosci.* 13, 2758–2771.

Grill-Spector, K., Kourtzi, Z., and Kanwisher, N. (2001). The lateral occipital complex and its role in object recognition. *Vision Res.* 41, 1409–1422. doi: 10.1016/s0042-6989(01)00073-6

Haider, B., Häusser, M., and Carandini, M. (2013). Inhibition dominates sensory responses in the awake cortex. *Nature* 493, 97–100. doi: 10.1038/nature11665

Hansen, B. J., Chelaru, M. I., and Dragoi, V. (2012). Correlated variability in laminar cortical circuits. *Neuron* 76, 590–602. doi: 10.1016/j.neuron.2012.08.029

Hein, G., Doehrmann, O., Müller, N. G., Kaiser, J., Muckli, L., and Naumer, M. J. (2007). Object familiarity and semantic congruency modulate responses in cortical audiovisual integration areas. *J. Neurosci.* 27, 7881–7887. doi: 10.1523/jneurosci.1740-07.2007

Hochstein, S., and Ahissar, M. (2002). View from the top: hierarchies and reverse hierarchies in the visual system. *Neuron* 36, 791–804. doi: 10.1016/S0896-6273(02)01091-7

Ito, J., Maldonado, P., Singer, W., and Grun, S. (2011). Saccade-related modulations of neuronal excitability support synchrony of visually elicited spikes. *Cereb. Cortex* 21, 2482–2497. doi: 10.1093/cercor/bhr020

Jiang, X., Bollich, A., Cox, P., Hyder, E., James, J., Gowani, S. A., et al. (2013). A quantitative link between face discrimination deficits and neuronal selectivity for faces in autism. *Neuroimage Clin.* 2, 320–331. doi: 10.1016/j.nicl.2013.02.002

Kenet, T., Arieli, A., Tsodyks, M., and Grinvald, A. (2005). "Are single cortical neurons soloists or are they obedient members of a huge orchestra?," in *Problems in Systems Neuroscience*, eds J. L. Van Hemmen and T. J. Sejnowski (New York: Oxford University Press), 160–181.

Kiani, R., Esteky, H., Mirpour, K., and Tanaka, K. (2007). Object category structure in response patterns of neuronal population in monkey inferior temporal cortex. *J. Neurophysiol.* 97, 4296–4309. doi: 10.1152/jn.00024.2007

King, P. D., Zylberberg, J., and Deweese, M. R. (2013). Inhibitory interneurons decorrelate excitatory cells to drive sparse code formation in a spiking model of V1. *J. Neurosci.* 33, 5475–5485. doi: 10.1523/jneurosci.4188-12.2013

Kleiner, M., Brainard, D., and Pelli, D. (2007). What's new in Psychtoolbox-3? *Perception 36 ECVP Abstract Supplement.*

Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., et al. (2008). Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron* 60, 1126–1141. doi: 10.1016/j.neuron.2008.10.043

Krizhevsky, A., Sutskever, I., and Hinton, G. (2012). ImageNet classification with deep convolutional neural networks. *Adv. Neural. Inf. Process. Syst.* 25, 1106–1114.

Le, Q. V., Ranzato, M. A., Monga, R., Devin, M., Chen, K., Corrado, G. S., et al. (2012). "Building high-level features using large scale unsupervised learning," in Proceedings of the 29th International Conference on Machine Learning, (Edinburgh, Scotland), 11.

Leopold, D. A., Bondar, I. V., and Giese, M. A. (2006). Norm-based face encoding by single neurons in the monkey inferotemporal cortex. *Nature* 442, 572–575. doi: 10.1038/nature04951

Leopold, D. A., O'toole, A. J., Vetter, T., and Blanz, V. (2001). Prototype-referenced shape encoding revealed by high-level aftereffects. *Nat. Neurosci.* 4, 89–94. doi: 10.1038/82947

Lin, C.-P., Chen, Y.-P., and Hung, C. P. (2014). Tuning and spontaneous spike time synchrony share a common structure in macaque inferior temporal cortex. *J. Neurophysiol.* 112, 856–869. doi: 10.1152/jn.00485.2013

Logothetis, N. K., and Schall, J. D. (1989). Neuronal correlates of subjective visual perception. *Science* 245, 761–763. doi: 10.1126/science.2772635

Logothetis, N. K., and Sheinberg, D. L. (1996). Visual object recognition. *Annu. Rev. Neurosci.* 19, 577–621.

Loughnan, B. L., Sebel, P. S., Thomas, D., Rutherfoord, C. F., and Rogers, H. (1987). Evoked potentials following diazepam or fentanyl. *Anaesthesia* 42, 195–198. doi: 10.1111/j.1365-2044.1987.tb02999.x

Maier, A., Logothetis, N. K., and Leopold, D. A. (2007). Context-dependent perceptual modulation of single neurons in primate visual cortex. *Proc. Natl. Acad. Sci. U S A* 104, 5620–5625. doi: 10.1073/pnas.0608489104

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.eor

Merolla, P. A., Arthur, J. V., Alvarez-Icaza, R., Cassidy, A. S., Sawada, J., Akopyan, F., et al. (2014). Artificial brains. A million spiking-neuron integrated circuit with a scalable communication network and interface. *Science* 345, 668–673. doi: 10.1126/science.1254642

Michel, M. M., Chen, Y., Geisler, W. S., and Seidemann, E. (2013). An illusion predicted by V1 population activity implicates cortical topography in shape perception. *Nat. Neurosci.* 16, 1477–1483. doi: 10.1038/nn.3517

Miller, J. E., Ayzenshtat, I., Carrillo-Reid, L., and Yuste, R. (2014). Visual stimuli recruit intrinsically generated cortical ensembles. *Proc. Natl. Acad. Sci. U S A* 111, E4053–E4061. doi: 10.1073/pnas.1406077111

Miyashita, Y. (1993). Inferior temporal cortex: where visual perception meets memory. *Annu. Rev. Neurosci.* 16, 245–263. doi: 10.1146/annurev.neuro.16.1.245

Mruczek, R. E., and Sheinberg, D. L. (2005). Distractor familiarity leads to more efficient visual search for complex stimuli. *Percept. Psychophys.* 67, 1016–1031. doi: 10.3758/bf03193628

Mruczek, R. E., and Sheinberg, D. L. (2007a). Activity of inferior temporal cortical neurons predicts recognition choice behavior and recognition time during visual search. *J. Neurosci.* 27, 2825–2836. doi: 10.1523/jneurosci.4102-06.2007

Mruczek, R. E., and Sheinberg, D. L. (2007b). Context familiarity enhances target processing by inferior temporal cortex neurons. *J. Neurosci.* 27, 8533–8545. doi: 10.1523/jneurosci.2106-07.2007

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. doi: 10.1007/s11263-007-0118-0

Olshausen, B. A., and Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* 381, 607–609. doi: 10.1038/381607a0

Olshausen, B. A., and Field, D. J. (2004). Sparse coding of sensory inputs. *Curr. Opin. Neurobiol.* 14, 481–487. doi: 10.1016/j.conb.2004.07.007

Op de Beeck, H. P., Deutsch, J. A., Vanduffel, W., Kanwisher, N. G., and Dicarlo, J. J. (2007). A stable topography of selectivity for unfamiliar shape classes in monkey inferior temporal cortex. *Cereb. Cortex* 18, 1676–1694. doi: 10.1093/cercor/bhm196

Op de Beeck, H., Wagemans, J., and Vogels, R. (2001). Inferotemporal neurons represent low-dimensional configurations of parameterized shapes. *Nat. Neurosci.* 4, 1244–1252. doi: 10.1038/nn767

Orban, G. A., Van Essen, D., and Vanduffel, W. (2004). Comparative mapping of higher visual areas in monkeys and humans. *Trends Cogn. Sci.* 8, 315–324. doi: 10.1016/j.tics.2004.05.009

Poggio, T., and Bizzi, E. (2004). Generalization in vision and motor control. *Nature* 431, 768–774. doi: 10.1038/nature03014

Rajkai, C., Lakatos, P., Chen, C. M., Pincze, Z., Karmos, G., and Schroeder, C. E. (2008). Transient cortical excitation at the onset of visual fixation. *Cereb. Cortex* 18, 200–209. doi: 10.1093/cercor/bhm046

Renart, A., De La Rocha, J., Bartho, P., Hollender, L., Parga, N., Reyes, A., et al. (2010). The asynchronous state in cortical circuits. *Science* 327, 587–590. doi: 10.1126/science.1179850

Riesenhuber, M., and Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nat. Neurosci.* 2, 1019–1025.

Robertson, C. E., Kravitz, D. J., Freyberg, J., Baron-Cohen, S., and Baker, C. I. (2013). Slower rate of binocular rivalry in autism. *J. Neurosci.* 33, 16983–16991. doi: 10.1523/jneurosci.0448-13.2013

Rolls, E. T., and Tovee, M. J. (1995). Sparseness of the neuronal representation of stimuli in the primate temporal visual cortex. *J. Neurophysiol.* 73, 713–726.

Ruff, D. A., and Cohen, M. R. (2014). Attention can either increase or decrease spike count correlations in visual cortex. *Nat. Neurosci.* 17, 1591–1597. doi: 10.1038/nn.3835

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. doi: 10.1523/jneurosci.0179-10.2010

Sadovsky, A. J., and Maclean, J. N. (2014). Mouse visual neocortex supports multiple stereotyped patterns of microcircuit activity. *J. Neurosci.* 34, 7769–7777. doi: 10.1523/jneurosci.0169-14.2014

Sato, T., Uchida, G., Lescroart, M. D., Kitazono, J., Okada, M., and Tanifuji, M. (2013). Object representation in inferior temporal cortex is organized hierarchically in a mosaic-like structure. *J. Neurosci.* 33, 16642–16656. doi: 10.1523/jneurosci.5557-12.2013

Sato, T., Uchida, G., and Tanifuji, M. (2009). Cortical columnar organization is reconsidered in inferior temporal cortex. *Cereb. Cortex* 19, 1870–1888. doi: 10.1093/cercor/bhn218

Scholl, C. A., Jiang, X., Martin, J. G., and Riesenhuber, M. (2014). Time course of shape and category selectivity revealed by EEG rapid adaptation. *J. Cogn. Neurosci.* 26, 408–421. doi: 10.1162/jocn_a_00477

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. doi: 10.1109/tpami.2007.56

Shamir, M. (2014). Emerging principles of population coding: in search for the neural code. *Curr. Opin. Neurobiol.* 25, 140–148. doi: 10.1016/j.conb.2014.01.002

Sigala, N., and Logothetis, N. K. (2002). Visual categorization shapes feature selectivity in the primate temporal cortex. *Nature* 415, 318–320. doi: 10.1038/415318a

Singer, J. M., and Kreiman, G. (2014). Short temporal asynchrony disrupts visual object recognition. *J. Vis.* 14:7. doi: 10.1167/14.5.7

Sompolinsky, H., Yoon, H., Kang, K., and Shamir, M. (2001). Population coding in neuronal systems with correlated noise. *Phys. Rev. E Stat. Nonlin. Soft. Matter Phys.* 64:051904. doi: 10.1103/physreve.64.051904

Srihasam, K., Vincent, J. L., and Livingstone, M. S. (2014). Novel domain formation reveals proto-architecture in inferotemporal cortex. *Nat. Neurosci.* 17, 1776–1783. doi: 10.1038/nn.3855

Sripati, A. P., and Olson, C. R. (2010). Global image dissimilarity in macaque inferotemporal cortex predicts human visual search efficiency. *J. Neurosci.* 30, 1258–1269. doi: 10.1523/jneurosci.1908-09.2010

Suzuki, S., and Cavanagh, P. (1998). A shape-contrast effect for briefly presented stimuli. *J. Exp. Psychol. Hum. Percept. Perform.* 24, 1315–1341. doi: 10.1037//0096-1523.24.5.1315

Takeuchi, D., Hirabayashi, T., Tamura, K., and Miyashita, Y. (2011). Reversal of interlaminar signal between sensory and memory processing in monkey temporal cortex. *Science* 331, 1443–1447. doi: 10.1126/science.1199967

Tamura, H., Mori, Y., and Kaneko, H. (2014). Organization of local horizontal functional interactions between neurons in the inferior temporal cortex of macaque monkeys. *J. Neurophysiol.* 111, 2589–2602. doi: 10.1152/jn.00336.2013

Tanaka, K. (1996). Inferotemporal cortex and object vision. *Annu. Rev. Neurosci.* 19, 109–139. doi: 10.1146/annurev.neuro.19.1.109

Tang, H., Buia, C., Madhavan, R., Crone, N. E., Madsen, J. R., Anderson, W. S., et al. (2014). Spatiotemporal dynamics underlying object completion in human ventral visual cortex. *Neuron* 83, 736–748. doi: 10.1016/j.neuron.2014.06.017

Tanigawa, H., Wang, Q., and Fujita, I. (2005). Organization of horizontal axons in the inferior temporal cortex and primary visual cortex of the macaque monkey. *Cereb. Cortex* 15, 1887–1899. doi: 10.1093/cercor/bhi067

Tootell, R. B. H., Tsao, D., and Vanduffel, W. (2003). Neuroimaging weighs in: humans meet macaques in "primate" visual cortex. *J. Neurosci.* 23, 3981–3989.

Treisman, A. M., and Gelade, G. (1980). A feature-integration theory of attention. *Cogn. Psychol.* 12, 97–136.

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. doi: 10.1038/90547

Valentine, T. (1999). *Face-space Models of Face Recognition.* Hillsdale, New Jersey: Lawrence Erlbaum Associates, Inc.

Vaziri, S., Carlson, E. T., Wang, Z., and Connor, C. E. (2014). A channel for 3D environmental shape in anterior inferotemporal cortex. *Neuron* 84, 55–62. doi: 10.1016/j.neuron.2014.08.043

Verhoef, B. E., Vogels, R., and Janssen, P. (2012). Inferotemporal cortex subserves three-dimensional structure categorization. *Neuron* 73, 171–182. doi: 10.1016/j.neuron.2011.10.031

Vinje, W. E., and Gallant, J. L. (2000). Sparse coding and decorrelation in primary visual cortex during natural vision. *Science* 287, 1273–1276. doi: 10.1126/science.287.5456.1273

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. doi: 10.1038/77712

Willenbockel, V., Sadr, J., Fiset, D., Horne, G. O., Gosselin, F., and Tanaka, J. W. (2010). Controlling low-level image properties: the SHINE toolbox. *Behav. Res. Methods* 42, 671–684. doi: 10.3758/BRM.42.3.671

Willmore, B. D., Mazer, J. A., and Gallant, J. L. (2011). Sparse coding in striate and extrastriate visual cortex. *J. Neurophysiol.* 105, 2907–2919. doi: 10.1152/jn.00594.2010

Wu, S., Amari, S., and Nakahara, H. (2002). Population coding and decoding in a neural field: a computational study. *Neural Comput.* 14, 999–1026. doi: 10.1162/089976602753633367

Xing, D., Ringach, D. L., Hawken, M. J., and Shapley, R. M. (2011). Untuned suppression makes a major contribution to the enhancement of orientation selectivity in macaque v1. *J. Neurosci.* 31, 15972–15982. doi: 10.1523/jneurosci.2245-11.2011

Yamane, Y., Tsunoda, K., Matsumoto, M., Phillips, A. N., and Tanifuji, M. (2006). Representation of the spatial relationship among object parts by neurons in macaque inferotemporal cortex. *J. Neurophysiol.* 96, 3147–3156. doi: 10.1152/jn.01224.2005

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., and Dicarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl. Acad. Sci. U S A* 111, 8619–8624. doi: 10.1073/pnas.1403112111

Yoon, J. H., Maddock, R. J., Rokem, A., Silver, M. A., Minzenberg, M. J., Ragland, J. D., et al. (2010). GABA concentration is reduced in visual cortex in schizophrenia and correlates with orientation-specific surround suppression. *J. Neurosci.* 30, 3777–3781. doi: 10.1523/JNEUROSCI.6158-09.2010

Young, M. P., and Yamane, S. (1992). Sparse population coding of faces in the inferotemporal cortex. *Science* 256, 1327–1331. doi: 10.1126/science.1598577

Zeiler, M. D., and Fergus, R. (2014). "Visualizing and understanding convolutional networks," in *ECCV 2014, Part I, LNCS 8689*, eds D. Fleet et al. (Switzerland: Springer International Publishing), 818–833.

Zoccolan, D., Kouh, M., Poggio, T., and Dicarlo, J. J. (2007). Trade-off between object selectivity and tolerance in monkey inferotemporal cortex. *J. Neurosci.* 27, 12292–12307. doi: 10.1523/jneurosci.1897-07.2007

Zohary, E., Shadlen, M. N., and Newsome, W. T. (1994). Correlated neuronal discharge rate and its implications for psychophysical performance. *Nature* 370, 140–143. doi: 10.1038/371358c0

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: 28 August 2014; accepted: 09 December 2014; published online: 06 January 2015.

Citation: Hung CP, Cui D, Chen Y-p, Lin C-p and Levine MR (2015) Correlated activity supports efficient cortical processing. Front. Comput. Neurosci. ***8****:171. doi: 10.3389/fncom.2014.00171*

This article was submitted to the journal Frontiers in Computational Neuroscience.

frontiers in
COMPUTATIONAL NEUROSCIENCE

GENERAL COMMENTARY
published: 24 February 2015
doi: 10.3389/fncom.2015.00025

Corrigendum: Correlated activity supports efficient cortical processing

Chou P. Hung[1,2]*, Ding Cui[1], Yueh-peng Chen[2], Chia-pei Lin[2] and Matthew R. Levine[1]

[1] Department of Neuroscience, Georgetown University, Washington, DC, USA
[2] Institute of Neuroscience, National Yang-Ming University, Taipei, Taiwan
*Correspondence: ch486@georgetown.edu

Edited and reviewed by:
Hans P. Op De Beeck, University of Leuven (KU Leuven), Belgium

Keywords: object recognition, inferior temporal cortex, macaque, visual search, efficient coding

A corrigendum on

Correlated activity supports efficient cortical processing
by Hung, C. P., Cui, D., Chen, Y.-p., Lin, C.-p., and Levine, M. R. (2014). Front. Comput. Neurosci. 8:171. doi: 10.3389/fncom.2014.00171

In the original article, in the second paragraph of the Analyses section, there was a typo in the description of "soloists" that are based on the median 30%ile. They are the 35–65%ile, not the 45–65%ile as previously reported. In addition to the 2nd paragraph of the Analyses section, the typo also appears in 3 other places: Figure 2 legend, 1st paragraph of Results, and in the second paragraph of "Correlated Neurons are Mostly in Output Layers."

The authors apologize for this error.

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: 05 February 2015; accepted: 09 February 2015; published online: 24 February 2015.

Citation: Hung CP, Cui D, Chen Y-p, Lin C-p and Levine MR (2015) Corrigendum: Correlated activity supports efficient cortical processing. Front. Comput. Neurosci. ***9****:25. doi: 10.3389/fncom.2015.00025*

This article was submitted to the journal Frontiers in Computational Neuroscience.

frontiers in
COMPUTATIONAL NEUROSCIENCE

ORIGINAL RESEARCH ARTICLE
published: 04 February 2015
doi: 10.3389/fncom.2015.00008

Applying artificial vision models to human scene understanding

Elissa M. Aminoff[1,2]*, *Mariya Toneva*[1,3], *Abhinav Shrivastava*[4], *Xinlei Chen*[4], *Ishan Misra*[4], *Abhinav Gupta*[4] and *Michael J. Tarr*[1,2]

[1] Center for the Neural Basis of Cognition, Carnegie Mellon University, Pittsburgh, PA, USA
[2] Department of Psychology, Carnegie Mellon University, Pittsburgh, PA, USA
[3] Department of Machine Learning, Carnegie Mellon University, Pittsburgh, PA, USA
[4] Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, USA

Edited by:
Hans P. Op De Beeck, University of Leuven, Belgium

Reviewed by:
Dwight Kravitz, National Institutes of Health, USA
Sean MacEvoy, Boston College, USA

***Correspondence:**
Elissa M. Aminoff, Department of Psychology, Carnegie Mellon University, 352 Baker Hall, 5000 Forbes Ave., Pittsburgh, PA 15213, USA
e-mail: eaminoff@cmu.edu

How do we understand the complex patterns of neural responses that underlie scene understanding? Studies of the network of brain regions held to be scene-selective—the parahippocampal/lingual region (PPA), the retrosplenial complex (RSC), and the occipital place area (TOS)—have typically focused on single visual dimensions (e.g., size), rather than the high-dimensional feature space in which scenes are likely to be neurally represented. Here we leverage well-specified artificial vision systems to explicate a more complex understanding of how scenes are encoded in this functional network. We correlated similarity matrices within three different scene-spaces arising from: (1) BOLD activity in scene-selective brain regions; (2) behavioral measured judgments of visually-perceived scene similarity; and (3) several different computer vision models. These correlations revealed: (1) models that relied on mid- and high-level scene attributes showed the highest correlations with the patterns of neural activity within the scene-selective network; (2) NEIL and SUN—the models that best accounted for the patterns obtained from PPA and TOS—were different from the GIST model that best accounted for the pattern obtained from RSC; (3) The best performing models outperformed behaviorally-measured judgments of scene similarity in accounting for neural data. One computer vision method—NEIL ("Never-Ending-Image-Learner"), which incorporates visual features learned as statistical regularities across web-scale numbers of scenes—showed significant correlations with neural activity in all three scene-selective regions and was one of the two models best able to account for variance in the PPA and TOS. We suggest that these results are a promising first step in explicating more fine-grained models of neural scene understanding, including developing a clearer picture of the division of labor among the components of the functional scene-selective brain network.

Keywords: scene processing, parahippocampal place area, retrosplenial cortex, transverse occipital sulcus, computer vision

INTRODUCTION

The past several decades have given us an unprecedented view of the inner workings of the human brain, allowing us to measure localized neural activity in awake, behaving humans. As cognitive neuroscientists, our challenge is to make sense of this rich source of data, connecting the activity we observe to mental mechanisms and behavior. For those of us who study high-level vision, making this connection is particularly difficult—vision scientists have not yet articulated any clear theories about what constitutes a "vocabulary" of intermediate visual features or what are the underlying building blocks of scene or object representation. Here we begin to address this issue by taking a different path to articulating a candidate set of features for visual representation: using a variety of extant computer vision models that make different commitments as to what counts as a visual feature as proxies for models of biological vision. We suggest that, to the extent that computer vision models and biological vision systems have similar end goals, the two domains will overlap in both their representations and processing assumptions.

To explore this issue, we had participants view 100 different scenes while we measured their brain activity, using functional Magnetic Resonance Imaging ("fMRI"), in regions that are known to be preferentially involved in scene processing. In particular, we hold that meaningful information can be extracted from the reliable patterns of activity that occur within scene selective regions: the parahippocampal/lingual region (the parahippocampal place area, "PPA"), the retrosplenial complex ("RSC"), and the occipital place area (also referred to as the transverse occipital sulcus, "TOS"). However, due to a lack of any fine-grained theories of scene understanding, it is unclear as to how one goes about interpreting the complex meaning inherent in these neural patterns. As alluded to above, we turn to models of computer vision to help us unravel how the human brain encodes and represents visual scenes, directly comparing the representations of scenes within these artificial vision systems to our obtained patterns of BOLD activity as measured by fMRI. The application of models derived from computer vision has one significant advantage: the models are well specified. As such, any particular model makes clear and explicit assumptions regarding representation and correspondence between a model and human neural responses or behavior allows us to infer that the two work similarly. Hence our emphasis on comparing a large number of models that all work somewhat differently from one another. In adopting modern computer vision models, we also note that these systems are built to understand the same complex visual world we deal with everyday (i.e., in contrast to earlier models that relied on "toy" worlds or highly-restricted visual domains). In particular, some of the models we include leverage large-scale/"web-scale" image datasets that may more accurately learn informative visual regularities embedded in the natural environment.

In that we have no strong *a priori* knowledge as to which of several very different models might be most effective with respect to accounting for neural data, our primary goal is to test whether we observe some correspondence between the patterns of neural activity elicited in high-order visual scene regions (i.e., PPA, RSC, and TOS) and the patterns of scene similarity as defined by these varying artificial vision models, and, specifically, which of these models does the best job at accounting for the neural data. We are also interested in the correspondence between artificial and biological vision systems, as well as the correspondence between the patterns of similarity obtained from neural responses and from behaviorally-measured explicit perceptual ratings.

We should note that our focus on accounting for neural responses in three specific brain regions of interest—the PPA, RSC, and TOS—is based on several decades of research describing the neural responses of these particular regions. Each has been shown to be selectively responsive to and optimized for processing scenes as compared to other visual stimuli, for example, single objects, faces, and meaningless visual patterns. It is also the case that all three of these regions are involved both in scene perception and spatial navigation; however, the PPA tends to be preferentially involved in scene recognition and the RSC tends to be preferentially involved in processing the larger spatial environment (Epstein and Higgins, 2007). These regions have also been sensitive to scene parts: both objects and spatial relations (Harel et al., 2013; Park et al., 2014); as well as more global properties of a scene such as the spatial boundary (Kravitz et al., 2011; Park et al., 2011; Watson et al., 2014). Finally, PPA, RSC and TOS have been shown to carry information regarding the statistical significance of objects occurring with specific scene categories (Stansbury et al., 2013) and the PPA has been shown to be sensitive to mid-level visual features, for example, recurring textures (Cant and Goodale, 2011; Cant and Xu, 2012). However, despite this array of empirically-demonstrated sensitivities to properties of the visual world, the specific computations that give rise to these functional responses are not well understood.

Here we use models originating from the field of computer vision to help reveal the computational processes that may be realized within these scene-selective brain regions. Given that scenes are complex visual stimuli that carry useful information within low-level visual features (e.g., oriented lines, edges, junctions, etc.), mid-level features (e.g., groupings and divisions of features that are superordinate to the low-level features), and high-level features (e.g., semantic meaning, categorization) we apply several different computer vision methods to capture these multiple levels. In particular, we attempt to account for variation in our neuroimaging data collected while participants are viewing a wide variety of different scenes using both high-level semantic feature-based models (e.g., SUN semantic attributes; Patterson and Hays, 2012) and low-level visual feature-based models (e.g., SIFT, HOG; Lowe, 2004; Dalal and Triggs, 2005). We predict that low-level features will be encoded in brain areas that selectively process scenes, but are also encoded in non-scene-selective regions such as early visual areas. In contrast, as discussed below, mid- and high-level features that capture the inherent meaning of a scene are predicted to be specifically encoded in scene-selective brain regions exclusively.

In studying scene or object understanding, the field faces a significant challenge: between visual input and semantics there is a significant gap in knowledge with respect to any detailed account of the mid- and high-level visual features that form the representation of visual information. That is, almost all theories of mid- and high-level visual representation rest on human intuition, providing little formal method for articulating the features underlying visual semantics or its precursors: mid-level visual features that are compositional in nature (Barenholtz and Tarr, 2007). For example, for us, distinguishing between a manmade and a natural scene is trivial and we typically account for our judgments by referring to semantic features within a scene (e.g., trees, buildings). However, there are also mid-level features (e.g., rectangular shapes) that are highly correlated with a scene's high-level semantics that may provide some insight into how the visual system can so readily understand the difference between manmade and natural. As one example, recent work suggests that the PPA responds preferentially to both simple rectilinear features and objects comprised of a predominantly rectilinear features (Nasr et al., 2014). This and other results hint that focusing on high-level semantics exclusively may miss critical elements of how scenes are selectively processed in the human brain. Relying on human intuition also suffers from the Titchenerian problem that introspection alone does not have access to the unconscious processing that makes up the bulk of our cognition. Thus, theories based largely on intuition almost surely miss identifying the bulk of visual features (or parts) that are critical in the neural representation of scenes. To address the need for mid-level, non-intuition-based visual features, one of the primary (and most interesting) computer vision models we apply is NEIL, the "Never Ending Image Learner" (www.neil-kb.com; Chen et al., 2013). NEIL is a large-scale ("web-scale") image-analysis system that, using only "weak supervision," automatically extracts underlying statistical regularities (e.g., both mid-level and high-level visual attributes) from natural scene images and constructs intuitively-correct scene categories. In doing so, NEIL both limits the need for the application of human intuition and allows for the simultaneous exploration of features at multiple levels of scene representation (i.e., low- to high-level). In applying NEIL, we asked whether the attributes that NEIL learns to characterize scenes give rise to a scene similarity space that correlates with a neurally-derived scene similarity space. Good correspondence between the two domains representing the same scenes would suggest that cortical vision is sensitive to some of the same statistical regularities—at a variety of levels—NEIL extracts to build a category structure for scenes.

In the past few years, a small number of studies have applied models drawn from computer vision to the question of neural representation in visual cortex. For the most part, this approach has focused on object recognition and examined a wide region of visual cortex, including low-level regions, V1–V3, mid-level regions, V4, and high-level regions, IT (Baldassi et al., 2013; Leeds et al., 2013; Khaligh-Razavi and Kriegeskorte, 2014; Yamins et al., 2014). However, to our knowledge, only one study has combined computer vision methods with neural scene understanding. In particular, Watson et al. (2014) examined how well low-level scene features derived from GIST, a descriptor that analyzes orientation energy at different spatial frequencies and spatial positions (Oliva and Torralba, 2001), might account for the fMRI-derived neural patterns associated with scene processing in the human ventral stream. They found that scene-specific regions (PPA, RSC, TOS) elicited patterns of activity that were better accounted for by low-level (GIST) properties as compared to semantic categories for scenes. However, Watson et al.'s study is limited by its "jump" from very low-level features (GIST) to very high-level semantic categories and their use of only four scene categories. Here we build on this result by looking at different metrics at different levels of representation and expanding the space of stimuli to 100 different scenes across 50 different scene categories, asking how well this range of computationally-motivated metrics can account for the complex neurally-derived scene space we measure in PPA, RSC, and TOS.

At the same time we explore representational metrics derived from computer vision, we also consider human behavior directly, examining the scene space derived from how humans judge two visually-presented scenes as similar. *A priori*, if two scenes are judged as similar, we might expect that the two scenes would elicit similar neural response patterns in scene selective brain regions. Of course, as noted earlier, explicit intuitions about cognitive processing are unreliable indicators of the complex underlying mechanisms supporting such processing. As such, it is unclear as to whether conscious behavior is a good predictor of neural representation. Thus, models of representation arising from computer vision may actually reveal more subtle information about neural encoding that cannot be inferred using behavioral methods. This empirical question—how well does behavior explain the neural activity elicited by scene understanding—is included in our study as a benchmark against which we can measure the performance of the computer vision models we apply to our data.

More generally, it is worth considering what we might be able to infer from our present methods. Particularly given our background emphasis on explicating better-specified accounts of mid- and high-level features, we might hope that a fine-grained analysis of our results would reveal the *specific* nature of representational features (e.g., a catalog of some sort). Unfortunately, such a detailed account is beyond what is realistically possible in our present study given: (1) the power limitations arising from the low number of observations we can collect from individual participants in an fMRI session; (2) the low SNR of BOLD responses; and (3) the middling spatial and poor temporal resolution of fMRI. To be clear, we view our present study as a first step in working toward such detailed accounts, but, realistically, such an account is not obtainable without many refinements in methods and theories. That being said, we hold that our present study does allow important inferences about the neural representation of scenes. More specifically, as discussed earlier, each of the computer vision models employed here makes assumptions regarding how it encodes visual scene information. Although the similarity metrics we use do not allow us to break down these assumptions to the level of specific features, they do help us choose between different models. Such model selection is common in some areas of science, but less so in the cognitive neurosciences where there are often few options from which to select (which is our point about the current state of knowledge regarding mid- and high-level visual representation). Our approach is to adopt a range of models from computer vision to enable a more comprehensive search space that encompasses a wider range of representational assumptions, including assumptions that might not be inferred through intuition. In the end, we learn something about which representational assumptions appear most promising for further investigation, thereby laying the groundwork for studies in which we specifically manipulate features derived from the most effective models.

A separate concern relates to a potential confound between receptive field (RF) size and feature complexity. At issue is the fact that more complex features tend to encompass more of the visual field and, therefore, are more likely to produce responses in the extrastriate scene-selective regions that are known to have larger RF sizes. However, we are less than certain as to how one would tractably partial out RF size from feature complexity. For example, if more complex features are more complex precisely because they are more global and reflect the relations between constituent parts, then—by definition—they are also captured in larger RFs. This is similar to the confound in the face recognition literature between RF size and "holistic" or "configural" processing (see for example, Nestor et al., 2008). Researchers argue that a particular effect is holistic, when, in fact, it is also the case that it is captured by larger RFs. Indeed, it may be that much of what we think of in the ventral pathway with respect to complexity is reasonably equivalent to RF size. We view trying to tease these two dimensions apart as an important question, but one that is beyond our present study.

More concretely, our study empirically examines human visual scene processing by way of scene similarity across three different domains: neuroimaging data, behavior, and computer vision models. In particular, we used fMRI with a slow event-related design to isolate the patterns of neural activity elicited by 100 different visual scenes. Using a slow event-related design we were able to analyze the data on a trial-by-trial/scene-by-scene basis, therefore allowing us to associate a specific pattern of BOLD activity with each individual scene. We then constructed a correlation matrix representing "scene-space" based on this neural data, performing all pairwise correlations between measured neural patterns within the brain regions of interest. This neurally-defined scene-space was then correlated with scene-spaces arising from a range of computer vision models [see Section Computer Vision (CV) Metrics]—each one providing a matrix of pairwise scene similarities of the same dimensionality as our neural data. At the same time, to better understand how the neural representation of scenes relates to behavioral judgments of scene similarity, we also ran a study using Amazon Mechanical Turk in which participants rated the similarity, on a seven-point scale, between two visually-presented scenes (4950 pairwise similarity comparisons).

MATERIALS AND METHODS

STIMULI

Scene stimuli were 100 color photographs from the NEIL database (www.neil-kb.com) (Chen et al., 2013) depicting scenes from 50 different scene categories as defined by NEIL—two exemplars from each category were used. Categories ranged from indoor to outdoor and manmade to natural in order to achieve good coverage of scene space. See Supplemental Material for a list of categories and Figure S1 for images of stimuli used. Scene images were square 600 × 600 pixels, and were presented at a 7° × 7° visual angle.

fMRI EXPERIMENT

Localizer stimuli

Stimuli used in the independent scene "localizer" consisted of color photographs of scenes, objects, and phase-scrambled pictures of the scenes. The objects used were not strongly associated with any context, and therefore were considered weak contextual objects (e.g., a folding chair) (Bar and Aminoff, 2003). Pictures were presented at 5° × 5° visual angle. There were 50 unique stimuli in each of the three stimulus conditions.

Participants

Data from nine participants in the fMRI portion of the study were analyzed (age: $M = 23$, 20–29; two left handed; five female). One additional participant (i.e., $N = 10$) was excluded from the data analysis due to falling asleep and missing a significant number of trial responses. Data from one other participant only had half the dataset included in the analysis due to severe movement issues in one of the two sessions. All participants had normal, or corrected-to-normal vision, and were not taking any psychoactive medication. Written informed consent was obtained from all participants prior to testing in accordance with the procedures approved by the Institutional Review Board of Carnegie Mellon University. Participants were financially compensated for their time.

Procedure

Each individual participated in two fMRI sessions in order to acquire sufficient data to examine the responses associated with individual scenes. Both sessions used the same procedure. The average time between the two sessions was 3.6 days, ranging from 1 to 7 days. Each fMRI session included six scene processing runs, a high resolution mprage anatomical scan run after the third scene processing run, and at the end of the session, one or two runs of a functional scene localizer.

During fMRI scanning, images were presented to the participants via 24 inch MR compatible LCD display (BOLDScreen, Cambridge Research Systems LTD., UK) presented at the head of the bore and reflected through a head coil mirror to the participant. Each functional scan began and ended with 12 s of a white fixation cross ("+") presented against a black background. For the scene processing runs, there were 50 picture trials—one exemplar from each of the 50 categories. The paradigm was a slow event-related design and order of the stimuli were random within the run. Two runs were required to get through the full set of 100 scenes, with no scene category repeating within the run. There were three presentations of each stimuli in each session (i.e., six functional runs) and across the two sessions, there were data for a total of six trials per a unique stimulus. Stimuli were presented for 1 s, followed by 7 s of fixation. On a random eight of the 50 trials of a run, the image rotated a half a degree to the right and then back to center, which took a total of 250 ms. Participants were asked to press a button when a pictured "jiggled." Participants performed on average 96% correct.

After all six of the scene processing runs, a functional scene localizer was administered in order to independently define scene selective areas of the cortex (PPA, RSC, and TOS). The localizer was a block design such that 12 stimuli of the same condition (either scenes, objects, or phase scrambled scenes) were presented in row. Each stimulus was presented for 800 ms with a 200 ms ISI. Between stimuli blocks, there were 8 s of a fixation cross presentation. There were six blocks per condition, and 18 blocks across conditions per run. The participant's task was to press a button if the picture immediately repeated (1-back task), of which there were two per block. Thus, in each block there were 10 unique stimuli presented, with two stimuli repeated once. Based on time of the scan session and energy of the participant, either one or two localizer runs were administered.

Before the participant went into the MRI scanner, they were told to remember the images as best as possible for a memory test. Once the participant concluded the fMRI portion of the session they performed a memory test outside the scanner. In the memory test, there were two trials for each of the 50 scene categories, with one trial presenting an image from the MRI session and the other trial presenting a new exemplar. For each trial, the participant had a maximum of 3 s to respond, with the picture on the screen for the entire time. The picture was removed from the screen as soon as the participant responded and the next trial began. Participants were 81% correct on average. The memory test was used to motivate the participants to pay attention, and was not used in any of the analyses.

fMRI data acquisition

Functional MRI data was collected on a 3T Siemens Verio MR scanner at the Scientific Imaging and Brain Research Center at Carnegie Mellon University using a 32-channel head coil. Functional images were acquired using a T2*-weighted echo-planar imaging pulse sequence (31 slices aligned to the AC/PC, in-plane resolution 2 × 2 mm, 3 mm slice thickness, no gap, *TR*

= 2000 ms, $TE = 29$ ms, flip angle = 79°, GRAPPA = 2, matrix size 96 × 96, field of view 192 mm, reference lines = 48, descending acquisition). Number of acquisitions per run was 209 for the main experiment, and 158 for the scene localizer. High-resolution anatomical scans were acquired for each participant using a T1-weighted MPRAGE sequence (1 × 1 × 1 mm, 176 sagittal slices, $TR = 2.3$ s, $TE = 1.97$ ms, flip angle = 9°, GRAPPA = 2, field of view = 256).

fMRI data analysis

All fMRI data were analyzed using SPM8 (http://www.fil.ion.ucl.ac.uk/spm/) and in-house Matlab scripts. Data across the two sessions were realigned to correct for minor head motion by registering all images to the mean image.

Functional scene localizer. After motion correction, the data of the scene functional localizer was smoothed using an isotropic Gaussian kernel (FWHM = 4 mm). The data was then analyzed as a block design using a general linear model and a canonical hemodynamic response function. A high pass filter using 128 s was implemented. The general linear model incorporated a robust weighted least squares (rWLS) algorithm (Diedrichsen and Shadmehr, 2005). The model simultaneously estimated the noise covariates and temporal auto-correlation for later use as covariates within the design matrix. The six motion parameter estimates that output from realignment were used as additional nuisance regressors. Data were collapsed across all localizer runs, with each run used as an additional regressor. The design matrix modeled three conditions: scenes, weak contextual objects, and phase scrambled scenes. The main contrast of interest was examining the differential activity that was greater for scenes as compared with objects and phase-scrambled scenes.

Event-related scene data. After motion correction, the data from the scene task runs were analyzed using a general linear model. Motion corrected data from a specific region of interest was extracted and nuisance regressors from the realignment were applied. The data was subjected to a 128 s high pass filter and was subjected to correction from rWLS, as well as a regressor represented each of the different runs. The data for the entire event window (8 s) was extracted for each scene stimulus, for each voxel within the region of interest, and averaged across the number of repetitions. Data in the 6–8 s time frame was used for all further analysis. This was the average peak activity in the time course across all trials for all participants. All six presentations of the stimulus were averaged together, including those that "jiggled" for the 250 ms. A similarity matrix of all the scenes (100 × 100) was then derived by cross-correlating the data for each scene across the voxels in the brain regions of interest within each individual. R-values from each of the cells in the similarity matrix were then averaged across participants for a group average.

Region of interest (ROI) analysis

All regions of interest analyzed were defined at the individual level using the MarsBaR toolbox (http://marsbar.sourceforge.net/index.html). Scene-selective regions (PPA, RSC, and TOS) were defined using the localizer data in the contrast of scenes greater than objects and phase-scrambled scenes. Typically, a threshold of FWE $p < 0.001$ was used to define the set of voxels. Size of ROIs were *a priori* chosen to have a goal of 100–200 voxels, or as close to that as possible. Two control non-scene selective ROIs were also chosen. One was a region in very early visual cortex along the left hemisphere calcarine sulcus defined in the localizer data as phase-scrambled greater than objects. The right hemisphere dorsolateral prefrontal cortex (DLPFC) was also chosen as a control region, which was defined using the localizer data in an all task (collapsed across all three conditions) greater than baseline comparison. Typically the threshold for the DLPFC ROI was lower than the other ROIs—FWE $p < 0.01$, or $p < 0.00001$ uncorrected, if not enough voxels survived the correction. Control ROIs were defined in all participants.

AMAZON MECHANICAL TURK (MTurk)

Behavioral judgments of similarity for each pairwise comparison of scenes were acquired through the use of study administered on MTurk.

Participants

Participants were voluntarily recruited through the human intelligence task (HIT) directory on MTurk. Enough individuals were recruited to satisfy reaching 20 observations for each of the 4950 pairwise scene comparisons. This resulted in 567 individuals participating in at least one HIT (10 scene pairs). An individual participated in an average of 17.2 HITs, and the range was from 1 to 174. All participants reported they were over the age of 18, with normal or corrected to normal vision, and located within the United States. Participants were financially compensated for each HIT completed. Participants read an online consent form prior to testing in accordance with the procedures approved by the Institutional Review Board of Carnegie Mellon University.

Procedure and data analysis

Each HIT contained 11 comparisons. Pairs of scenes were presented side-by-side, and the participant was asked to rate the similarity of the two scenes on 1–7 scale (1 = completely different; 7 = very similar), there was also an option of 8 for identical. The scale was presented below the pair of images with both the number and the description by each response button. In each HIT there was one pair that was identical for use as a catch trial. Participants were encouraged to use the entire scale. A participant's data were removed from the analysis if he/she did not respond correctly on the catch trials. If the participant missed a number of catch trials (over the course of several HITs) and exclusively used only 1 and 8 on the scale, that participant's entire data was removed from the analysis due to ambiguity as to whether she/he was actually completing the task, or just pressing 1 and 8. All valid data was then log transformed due to a preponderance of different judgments relative to any other response; skewness of 2.67 ($SE = 0.04$) and kurtoisis of 8.76 ($SE = 0.07$). The data were then used to construct a similarity matrix of the scenes (100 × 100) with the value of each cell determined by the average response for the pairwise comparison across the ∼20 observations. Some comparisons had missing responses due to removal of ambiguous data.

COMPUTER VISION (CV) METRICS

Each of the 100 scenes was analyzed by several different computer vision methods. The vector of features for each scene within each model was cross-correlated across all pairwise scene correlations to generate the similarity matrix defining the scene-space for that technique. We chose a wide variety of computer vision models that implement features that can roughly be divided in two categories: mid- and high-level attribute-based (NEIL, SUN semantic attributes, GEOM) and low-level (GIST, SIFT, HOG, SSIM, color). The former, attribute-based features, capture semantic aspects in the image, for example, highways, fountains, canyons, sky, porous etc. Low-level features such as GIST, SIFT, and HOG capture distributions of gradients and edges in the image. Gradients are defined as changes/derivatives of pixel values in the X and the Y direction in the image and edges are obtained after post-processing of these gradients. Note that for the purposes of this paper, we will use the terms gradients and edges interchangeably. Finally, models such as SSIM encode geometric layout of low-level features and shape information in the image. Local self-similarities in edge and gradient distributions complement low-level features such as those in SIFT. Critically, all of these models have a proven track record for effective scene classification (Oliva and Torralba, 2006; Vedaldi et al., 2010; Xiao et al., 2010). We now describe each of these models in more detail.

NEIL

The Never-Ending Image Learner (Chen et al., 2013) is a system that continuously crawls the images on the internet to automatically learn visual attributes, objects, scenes and common sense knowledge (i.e., the relationships between them). NEIL's strength comes from the large-scale data it analyzes in which it learns this knowledge; and by using commonsense relationships in this knowledge base to constrain its classifiers. NEIL's list of visual attributes were generated using the following mechanism (Chen et al., 2013): first, an exhaustive list of attributes used in the computer vision community were compiled, which included semantic scene attributes (SUN) (Patterson and Hays, 2012; Shrivastava et al., 2012), object attributes (Farhadi et al., 2009; Lampert et al., 2013) and generic attributes used for multimedia retrieval (Naphade et al., 2006; Yu et al., 2012). This exhaustive list was then pruned to only include attributes that represented adjectival properties of scenes and objects (e.g., red, circle shape, vertical lines, grassy texture). At the time of our study, the scene classifiers learned by NEIL were based on a scene space defined by 84 of these visual attributes, encompassing low-, mid- and high-level visual information of the scenes. For each scene there is a vector of scores, one for each attribute, of how confidently that attribute can be identified in that scene image. For each attribute classifier, we computed the variance of its scores across all scene categories used within the experiment, and used exponentiated variance for re-weighing the scores of each attribute individually. This normalization increases the weights on attributes that are more effective for distinguishing between scene categories and down weights the attributes that are less effective. The similarity matrix for NEIL was constructed as a cross-correlation of these scores.

Semantic scene attributes (SUN)

We use the set of 102 high-level SUN attributes as proposed in Patterson and Hays (2012), which were originally defined through crowd-sourcing techniques specifically intended to characterize scenes. These attributes were classified under five different categories: materials (e.g., vegetation), surface properties (e.g., sunny), functions or affordances (e.g., biking), spatial envelope (e.g., man-made), and object presence (e.g., tables). For each attribute, we have a corresponding image classifier as trained in Patterson and Hays (2012). The scores of these 102 classifiers were then used as features. These scores represent the confidence of each classifier in predicting the presence of the attribute in the image. The similarity matrix for SUN was constructed as a cross-correlation of these semantic attribute scores.

GEOM

Geometric class probabilities (Hoiem et al., 2007) for image regions—ground (gnd), vertical (vrt), porous (por), sky, and all were used. The probability maps for each class are further reduced to 8 × 8 matrix, where each entry represents the probability of the geometric class in a region of the image (Xiao et al., 2010). The similarity matrix for GEOM for each subset definition (e.g., vrt) was constructed as a cross-correlation of the probability scores for each region of the picture.

GIST

GIST (Oliva and Torralba, 2006) captures spatial properties of scenes (e.g., naturalness, openness, symmetry etc.) using low-level filters. The magnitude of these low-level filters encodes information about horizontal and vertical lines in an image, thus encoding the global spatial structure. As a byproduct, it also encodes semantic concepts like horizon, tall buildings, coastal landscapes etc., which are highly correlated with distribution of horizontal/vertical edges in an image. The GIST descriptor is computed using 24 Gabor-like filters tuned to 8 orientations at 4 different scales. The squared output of each filter is then averaged on a 4 × 4 grid (Xiao et al., 2010). The similarity matrix for GIST was constructed as a cross-correlation of these averaged filter outputs (512 dimensions).

HOG 2 × 2 *(L0–L2)*

Histogram of oriented gradients (HOG) (Dalal and Triggs, 2005) divides an image into a grid of 8 × 8 pixel cells and computes histogram statistics of edges/gradients in each cell. These statistics capture the rigid shape of an image and are normalized in different ways to include contrast sensitive, contrast insensitive and texture distributions of edges. For HOG 2 × 2 (Felzenszwalb et al., 2010; Xiao et al., 2010), the HOG descriptor is enhanced by stacking spatially overlapping HOG features, followed by quantization and spatial histograms. The spatial histograms are computed at three levels on grids of 1 × 1 (L0), 2 × 2 (L1) and 4 × 4 (L2) (see Xiao et al., 2010, for details). The similarity matrix for HOG 2 × 2 (L0–L2) was constructed as a cross correlation of these histogram features at different image regions and spatial resolutions.

SSIM (L0–L2)

Self-similarity descriptors (Shechtman and Irani, 2007) capture the internal geometric layout of edges (i.e., shape information) using recurring patterns in edge distributions. The descriptors are obtained by computing the correlation map of a 5 × 5 patch in a window with 40 pixels radius, followed by angular quantization. These SSIM descriptors are further quantized into 300 visual words using k-means (see Xiao et al., 2010, for details). The similarity matrix for SSIM was constructed as a cross correlation of these histograms of visual words at different spatial resolutions.

Finally, we included a variety of local image features based on image gradient/texture and color. Following the standard Bag-of-Words approach (vector quantization of features using k-means), we generated a fixed-length representation for each image. We used various dictionary sizes ($k = 50$, 250, 400, 1000) for each feature. For implementation, (van de Sande et al., 2011) was used for feature extraction and (Vedaldi and Fulkerson, 2010) for k-means quantization of features. As suggested by Vedaldi and Fulkerson (2010), we also L2 normalized each of the histograms. The similarity matrix for each of the local image features below was constructed as a cross correlation of these histograms of visual words for each local feature. The local features used were as follows:

- Hue histogram (50, 250, 400, 1000): A histogram based on the hue channel 1 of the image in the HSV color space representation. Roughly speaking, hue captures the redness/greenness/blueness etc. of the color.
- SIFT (50, 250, 400, 1000): Scale invariant feature transform (SIFT) (Lowe, 2004) characterizes each image based on local edge features. For each point in the image, it captures the gradient distribution around it, generally by computing histograms of edge feature in local neighborhood/patch and normalizing these histograms to make the descriptor rotationally invariant (even if the patch of pixels is rotated, the computed SIFT feature is the same). Standard SIFT works on grayscale images, and we use dense-SIFT (see Xiao et al., 2010; van de Sande et al., 2011).
- Hue-SIFT (50, 250, 400, 1000): SIFT computed only on the hue channel of the HSV representation of the input image.
- RGB-SIFT (50, 250, 400, 1000): SIFT computed on each color channel (R, G, and B) independently, and then concatenated.

CORRELATIONS ACROSS MEASURES

The similarity matrix arising from each method was converted into a vector using data from one side of the diagonal. This data were then fisher corrected for all analyses. First, a cross correlation analysis was performed to acquire the Pearson's r correspondence between each method. The p-values in this cross correlation are assumed to survive a Bonferroni correction correcting for 4950 pairwise correlations of scenes ($p < 0.00001$). For the regression analysis, p-values were corrected against 39 correlations (All ROIs, behavior, CV measures). To test the significance between model fits, a bootstrapping method was implemented. Testing across 1000 iterations of samples with replacement, a 95% confidence interval between model fits (r^2) was defined. The confidence interval reflected a $p < 0.05$ correcting for multiple comparisons. If the difference between the model correlations exceeded the confidence interval, the models were considered significantly different from each other (Wasserman, 2004).

RESULTS

We examined scene encoding in the human visual cortex by defining ROIs in the brain that preferred scene stimuli to weak-contextual objects and phase-scrambled scenes. This gives rise to three ROIs: the PPA, RSC, and TOS where the BOLD signal was found to be significantly greater when viewing scenes as compared to objects or phase-scrambled scenes. Additional two brain regions were defined, an early visual region and a region in the dorsolateral prefrontal cortex (DLPFC, see Materials and Methods). These regions were chosen as control regions to compare the scene ROIs (PPA, RSC, and TOS) to regions of the brain involved in visual processing or in a cognitive task involving visual stimuli, but that are not believed to be specific to scene processing. Data for each of the 100 scenes were then extracted on a voxel by voxel basis for each ROI. To examine the encoding of scenes each pairwise correlation of the scenes was computed to determine how similar the patterns of activity across the voxels of an ROI were from scene to scene. The resulting data were used to create a similarity matrix describing the scene space in each ROI, see Figure S3 for the similarity matrices of each ROI.

A separate behavioral study asking for an explicit judgment of scene similarity was performed to examine the perceived similarity between the 100 scenes. Using this data a similarity matrix was derived that was representative of scene space as defined by perceived similarity (see Figure S2). The data was split in half to test reliability of the scores, and similarity measures across the two halves correlated with an $r = 0.84$.

Finally, feature spaces defined through 30 different computer vision (CV) techniques were used to construct a scene space for each CV method. The features were cross-correlated for each pairwise correlation of the 100 different scenes to obtain a measure of similarity, which resulted in each similarity matrix or scene-space. Data were Fisher corrected, or log transformed (behavioral data), and correlated across the different scene-spaces to determine the similarity between these different scene representations (**Figure 1**; **Table 1**).

One of the clearest results within the similarity matrix across methods shown in **Figure 1** is how much more similar the scene-selective brain regions are to themselves as contrasted with any other measure, and how similar subsets of the computer vision methods are to themselves as contrasted with either the brain or behavioral methods. One of the implications of this pattern is that we still have a ways to go in accounting for the consistent patterns of neural encoding for visual stimuli. This work-to-be-done notwithstanding, the average correlation across scene ROIs not including hemisphere correlate (e.g., LH PPA × LH RSC; LH PPA × RH TOS) was $r = 0.34$, $SD = 0.07$. The greatest similarities resulted from comparing across hemisphere of the same region (e.g., LH PPA × RH PPA); mean $r = 0.58$, $SD = 0.11$. The correlations between brain regions was considerably lower when comparing a scene ROI with a control region, mean $r = 0.16$, $SD = 0.07$, demonstrating the similarity specific to scene selective regions. CV measures were similar with themselves, $r = 0.37$, $SD = 0.29$. And the least similarity was when comparing scene

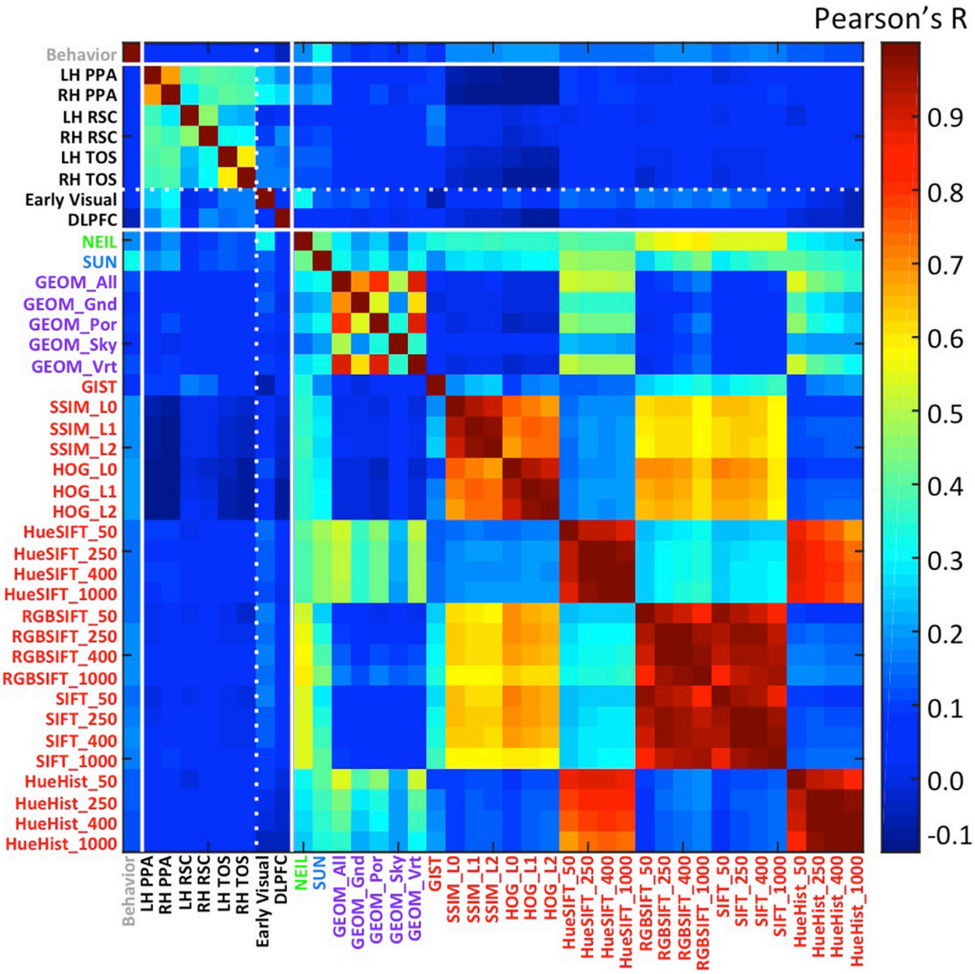


FIGURE 1 | Similarity matrix across different methods for constructing a scene space. Each cell is the *r*-value computing the correlation between the similarity of one scene space (e.g., voxel space in LH PPA) with another (e.g., attribute space in NEIL). Scene ROIs include the PPA, RSC, and TOS for each hemisphere, and two control brain regions—an early visual region as well the DLPFC. Computer vision methods are grouped according to their nominal level of representation—e.g., GEOM is mid-level (purple); and HOG is low-level (red).

brain ROIs with CV measures $r = 0.04$, $SD = 0.06$, however the correlations did get as high as $r = 0.22$, $p < 1.5 \times 10^{48}$ found between the RH PPA and the SUN measures. Similarity matrices derived from low-level features such as SSIM and HOG were either non-significant or negatively correlated with voxel space from scene regions, but found to be positively correlated with the early visual ROI. In general, the high-level CV methods (NEIL, SUN) significantly correlated with the scene ROIs, where, the low-level CV methods showed little correlation (although some did reach significance, see **Table 1**). Suffice it to say, there is a great deal of room for improvement in using CV measures to explain brain encoding of scenes. Critically, this is not due to noise in the signal—as already mentioned, there are strong correlations across the scene-selective ROIs, supporting the assumption that there is a meaningful code being used to process scenes, it just has yet to be cracked. However, that we observe significant correlations with some CV measures suggests we are making progress in explicating this code, and that the continued search for correspondences between computer vision models and patterns of brain activity may prove fruitful.

A more surprising result from our study is that correlations with brain regions was stronger with CV models (especially those with high- and mid- level features; average of SUN, NEIL, and GEOM All $r = 0.11$, $SD = 0.05$) than with behavioral similarity judgments (average $r = 0.05$, $SD = 0.01$). From these results we infer that perceived similarity between scenes is based on different visual and semantic parameters than those encoded in scene-selective ROIs. From an empirical point of view, the fact that our neurally-derived scene spaces do correlate more with some of the scene spaces derived from CV models suggests that methods drawn from computer vision offer a tool for isolating specific, and perhaps more subtle, aspects of scene representation as encoded in different regions of the human brain.

Beyond examining the general correspondence between CV metrics and the neural encoding of scenes, we were interested in the nature of the CV metrics offering the best correspondence and

Table 1 | Pearson's *r*-values for the correlations between similarity matrices.

	Behavior	LH+PPA	RH+PPA	LH+RSC	RH+RSC	LH+TOS	RH+TOS	Early+Visual	DLPFC
Behavior		**0.050**	0.030	**0.061**	0.044	**0.051**	**0.050**	**0.055**	−0.044
NEIL	**0.182**	**0.136**	**0.190**	**0.066**	**0.092**	**0.140**	**0.111**	**0.322**	**0.080**
SUN	**0.319**	**0.188**	**0.215**	**0.086**	**0.111**	**0.140**	**0.109**	**0.139**	**0.061**
GEOM_All	**0.107**	**0.067**	**0.085**	**0.055**	**0.070**	**0.089**	**0.077**	**0.111**	0.042
GEOM_Gnd	**0.078**	0.038	**0.048**	0.045	**0.059**	**0.053**	**0.054**	**0.133**	0.015
GEOM_Por	**0.094**	**0.079**	**0.115**	0.043	**0.067**	**0.058**	**0.060**	**0.100**	**0.054**
GEOM_Sky	**0.059**	0.040	0.018	**0.074**	**0.050**	**0.062**	0.038	**0.082**	0.019
GEOM_Vrt	**0.113**	**0.080**	**0.102**	0.041	**0.077**	**0.073**	**0.066**	**0.101**	**0.048**
GIST	**0.064**	**0.101**	**0.095**	**0.171**	**0.141**	**0.050**	**0.072**	**−0.057**	0.032
SSIM_L0	**0.182**	**−0.053**	**−0.071**	0.005	0.006	−0.008	−0.028	**0.088**	−0.014
SSIM_L1	**0.178**	**−0.070**	**−0.096**	0.008	0.012	−0.019	−0.039	**0.073**	−0.024
SSIM_L2	**0.175**	**−0.069**	**−0.093**	0.011	0.018	−0.017	−0.036	**0.087**	−0.021
HOG_L0	**0.201**	**−0.099**	**−0.131**	−0.001	−0.015	−0.041	**−0.074**	**0.108**	**−0.047**
HOG_L1	**0.193**	**−0.116**	**−0.151**	0.014	0.000	**−0.050**	**−0.075**	**0.085**	**−0.068**
HOG_L2	**0.198**	**−0.115**	**−0.157**	0.025	0.005	**−0.050**	**−0.072**	**0.072**	**−0.076**
HueSIFT_50	**0.151**	**0.105**	**0.105**	0.013	0.038	**0.083**	**0.066**	**0.154**	0.025
HueSIFT_250	**0.145**	**0.078**	**0.087**	0.023	0.043	**0.068**	**0.057**	**0.139**	0.024
HueSIFT_400	**0.151**	**0.096**	**0.100**	0.028	0.045	**0.076**	**0.064**	**0.128**	0.021
HueSIFT_1000	**0.153**	**0.087**	**0.090**	0.035	**0.049**	**0.066**	**0.057**	**0.116**	0.014
RGBSIFT_50	**0.156**	0.007	0.029	0.011	0.033	0.009	−0.017	**0.175**	**0.066**
RGBSIFT_250	**0.178**	0.014	0.028	**0.048**	**0.050**	0.015	0.011	**0.154**	**0.046**
RGBSIFT_400	**0.188**	0.046	**0.054**	**0.066**	**0.060**	0.025	0.025	**0.167**	0.039
RGBSIFT_1000	**0.188**	**0.077**	**0.087**	**0.083**	**0.076**	**0.046**	**0.052**	**0.153**	**0.048**
SIFT_50	**0.155**	−0.002	0.025	0.018	0.043	0.006	−0.012	**0.173**	**0.066**
SIFT_250	**0.173**	0.018	0.037	0.045	**0.048**	0.010	0.005	**0.136**	**0.058**
SIFT_400	**0.180**	0.030	0.044	**0.050**	**0.051**	0.016	0.014	**0.123**	**0.052**
SIFT_1000	**0.173**	**0.070**	**0.089**	**0.072**	**0.065**	0.036	0.043	**0.099**	**0.065**
HueHist_50	**0.137**	**0.061**	**0.055**	0.001	0.029	**0.068**	0.045	**0.093**	0.003
HueHist_250	**0.127**	**0.057**	0.041	0.043	**0.053**	**0.060**	0.044	0.012	−0.017
HueHist_400	**0.123**	**0.052**	0.032	**0.053**	**0.059**	**0.053**	0.042	−0.012	−0.024
HueHist_1000	**0.114**	0.044	0.022	**0.058**	**0.060**	0.044	0.038	−0.033	−0.032

Gray values indicate p >0.05; and bolded values indicate survived correction for multiple correlations.

what this might reveal about the kind of information encoded in scene-selective brain regions. Interestingly, we find that CV metrics that consider high-level visual attributes, that is, SUN and NEIL, have the strongest correlation with the scene-selective ROIs (**Figure 2A**). In general, the lower-level CV metrics performed the worst (e.g., SSIM and HOG) and the mid-level features as defined through the GEOM faired reasonably well and were significantly correlated with scene-selective ROIs. This latter result was not unexpected in that the GEOM feature space is designed to divide a scene into those visual properties that define major features of scenes (e.g., sky). Of particular note, GEOM Por, which emphasizes material properties was significantly correlated with the responses of the PPA, a result consistent with previous studies in which it was found that the PPA is sensitive to both textures and material information (Arnott et al., 2008; Cant and Goodale, 2011). GIST was the low-level CV model that most strongly correlated with scene ROIs—primarily RSC. This result in general was not unexpected as GIST has previously been shown to be correlated with scene ROIs (Watson et al., 2014) and with scene recognition (Oliva and Torralba, 2006).

Helping validate the significance of our results, we note that the hierarchy of correlations, with decreasing correlations progressing from high- to mid- to low-level visual features was observed only in the scene-selective ROIs, but not in the two control regions (early visual and DLPFC; **Figure 2A**). More specifically, although NEIL produced a strong correlation in both control regions, the other high-level model, SUN, and mid-level model, GEOM, were not the most significant correlations when compared to low-level feature models (e.g., SIFT). That low-level feature models resulted in higher-ranked correlations in early visual regions as compared to high- and mid-level feature models is consistent with the central role of these brain regions in early visual processing.

To examine the consistency of this hierarchy of feature sensitivity within the six scene-selective ROIs, we examined each ROI separately and plotted the six CV metrics that showed the best correlations (**Figure 2B**). Both SUN and NEIL (except for the LH RSC) consistently resulted in close to the strongest correlations with our neuroimaging data. To test the significance of the model fits, a bootstrapping method was used to test for a

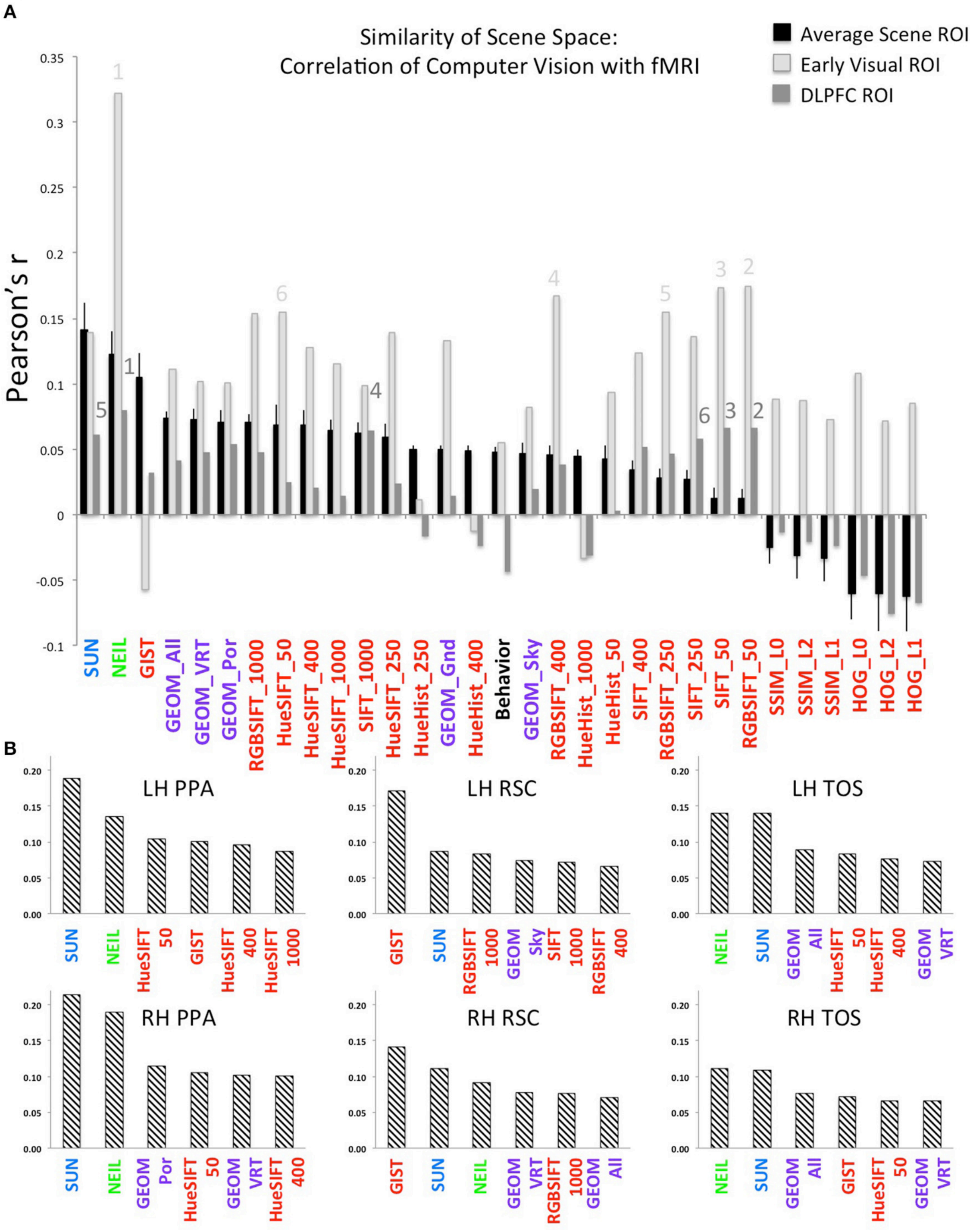


FIGURE 2 | Strength of correlation between the similarity matrix of computer vision (CV) metrics and the similarity matrix of patterns of brain activity across voxels in each ROI. (A) The average correlation across each CV metric and each of the scene ROIs is shown by the black bars. The X-axis is ordered by the strength of this correlation. By way of comparison, the correlations between the CV metrics and the two control regions are illustrated by the light gray bars (early visual region) and the dark gray bars (DLPFC). Error bars indicate standard error across the six scene ROIs (LH and RH of the PPA, RSC, TOS). Note that font color indicates the approximate level of featural analysis implemented in each specific CV metric: blue and green are high-level; purple is mid-level; and red is low-level. Numbers indicate the top 6 correlations in the early visual regions (light gray font, above light gray bars) and in the DLPFC (dark gray front, above dark gray bars). **(B)** The top-ranked 6 CV metrics that correlated with the neurally-derived similarity matrix in each of the 6 scene-selective ROIs. The Y-axis is Pearson's *r*-value.

$p < 0.05$ correcting for multiple correlations. For a full plot of all correlations, see Figure S4. Only in the LH PPA did the SUN features significantly account for more variance in brain data than NEIL features, in the other ROIs they were statistically equivalent. The PPA and the TOS both had SUN and NEIL fitting the data the best, performing significantly better than behavior, low level features such as HueHist, SIFT, RGBSIFT. In some cases the variances accounted for by HOG and SSIM, which was negatively correlated, did not significantly differ from SUN and NEIL (LH PPA, RH PPA, RH TOS). However, it is hard to interpret the significance of a negative correlation, so we provide this result with caution. Interestingly, color also seemed to be an important feature in encoding scene space. Hue SIFT, which takes into account scale invariant local features with respect to different hue maps, gave rise to scene spaces that were correlated with the neural responses measured in both TOS and demonstrated significance above a number of other models in the PPA. Although numerically midlevel features—GEOM—correlated better than low level features, significance was only reached for GEOM_por and GEOM_sky in the RH PPA, and GEOM_all in the LH TOS. On the other hand, the RSC had a different pattern of correlations. GIST showed the strongest correlation with our neuroimaging data within the RSC, fitting significantly better than all other models in the LH RSC, and all models except for the SUN features in the RH RSC. This is consistent with previous results demonstrating a correspondence for GIST with the responses of this region (Watson et al., 2014). In the LH RSC and RH RSC SUN features and RGB SIFT correlated at levels significantly over other models, and within the RH RSC NEIL also correlated significantly over and above other models. Overall, high-level feature models produced the scene spaces most consistently correlated with the scene spaces derived from scene-selective ROIs in the PPA and TOS, whereas GIST correlated the best, and the high-level SUN and NEIL features correlated next best in the RSC.

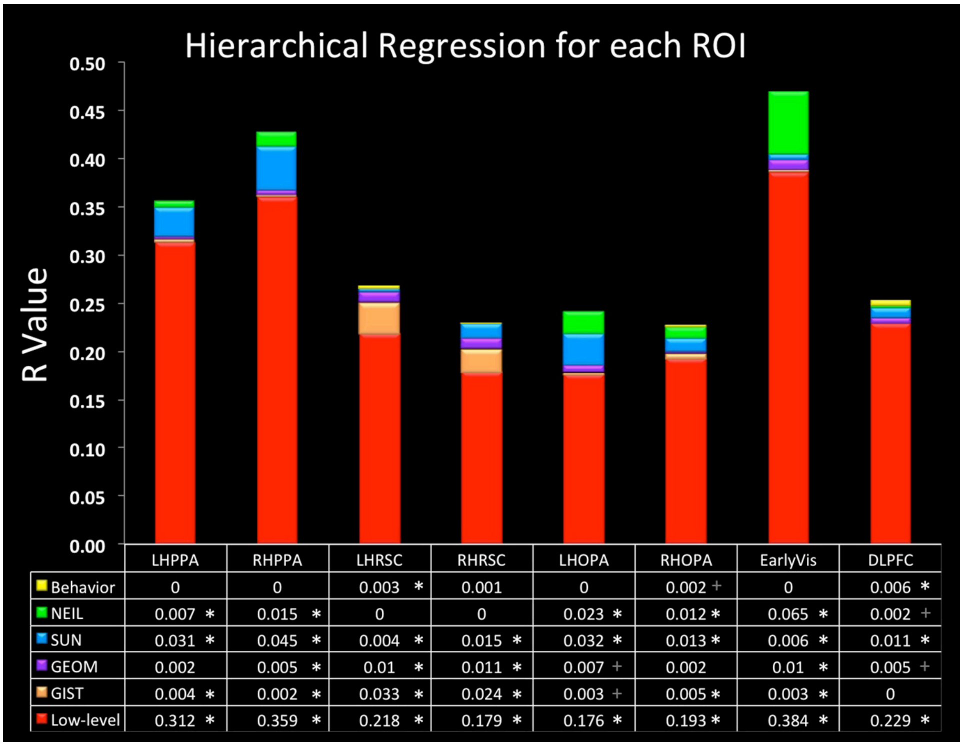
To investigate the reliability of this dataset we split the data in two (one for each session) and tested the consistency of the results. We found the correlations between the brain data with the CV measures and behavioral judgments were very consistent over the two sessions, resulting in an average $r = 0.76$, $SD = 0.19$; where the strongest consistency was in the PPA and early visual regions $r = 0.94$, $SD = 0.02$, and the lowest consistency was in the RH RSC ($r = 0.43$) and the DLPFC ($r = 0.63$). In addition, we examined the effect of including the trials that "jiggled" on the analysis, until this point all analyses include the rotated trials. We performed the analyses with and without the rotated trials, showing very little effect of including all trials in the average, the average r-value obtained across all ROIs with the CV measures and behavioral data across the two analyses was 0.97, $SD = 0.02$. The most notable difference in the analysis that did not include the rotated trials was an increase in the correlation with GIST. This result provides some insight into the nature of the correlation between GIST and scene ROIs, one that may be less stable than the others and therefore may not allow theoretical inference about the nature of scene representations in these brain regions.

Finally, we were especially interested in examining the similarity between NEIL-derived scene-space and our neuroimaging data. The web-scale nature of how NEIL learns about regularities across scene categories is appealing in that it seems to best capture both the evolutionary history of our visual systems and the kind of neural statistical learning that seems to emerge over a lifetime of experience. NEIL's features capture the visual regularities that give rise to semantic information, helping to define the visual features that give rise to scene understanding. **Table 1** shows that the scene space derived from NEIL's attributes is significantly correlated with our neurally-derived scene space within each scene-selective ROI. However, the question remains about how well does NEIL do over and above all the other CV measures. To address this, we ran a hierarchical regression for each ROI (**Figure 3**). In this regression the first input was the low-level CV metrics (Hue Histogram, SIFT, HOG, SSIM) and the second input was to separately add GIST, to see what variance was left over when the low-level visual features were removed. Next we entered the GEOM metrics, followed by the SUN attributes, followed by NEIL, and, finally, the last block being our behavioral data. This regression demonstrates that NEIL accounts for a significant amount of the variance in defining the neurally-derived scene space over and above any of the other CV metrics in both the PPA and the TOS, as well as in early visual regions. As such, it appears as if NEIL is capturing something unique about scene representation within the PPA, TOS, and early visual regions that is not captured by any of the other models. The behavioral data only accounted for unique variance above that already accounted for in the LH RSC and the DLPFC.

DISCUSSION

We started with the challenge of specifying the "language" of mid- and high-level features supporting object and scene recognition. Given the large space of possible answers to this question, we attempted to constrain the possible answers by applying a variety of computer vision models that make somewhat different assumptions regarding the nature of this language. To evaluate the effectiveness of these different assumptions, we explored the degree to which each model accounted for patterns of neural data arising from scene processing by scene-selective brain regions. We found that:

- The NEIL and SUN models—both of which rely on mid- and high-level visual features—were best at accounting for variation in the neural responses of both the PPA and the TOS. The fact that NEIL was equivalent to SUN indicates that statistically-derived features offer a viable model of scene representation that may, ultimately, reveal non-intuitive coding principles for scenes.
- The GIST model—a model which relies on global spatial properties of scenes—was best at accounting for variations in the neural responses of the RSC. Additional unique variance in the RSC was accounted for by our behaviorally-obtained similarity ratings.
- Given points (1) and (2), there is support for a model of scene processing in which PPA and TOS are coding scene information differently from RSC, with the former coding for the visual attributes within scenes and the latter coding for higher-order, scene categories.



	LHPPA	RHPPA	LHRSC	RHRSC	LHOPA	RHOPA	EarlyVis	DLPFC
🟨 Behavior	0	0	0.003 *	0.001	0	0.002 +	0	0.006 *
🟩 NEIL	0.007 *	0.015 *	0	0	0.023 *	0.012 *	0.065 *	0.002 +
🟦 SUN	0.031 *	0.045 *	0.004 *	0.015 *	0.032 *	0.013 *	0.006 *	0.011 *
🟪 GEOM	0.002	0.005 *	0.01 *	0.011 *	0.007 +	0.002	0.01 *	0.005 +
🟧 GIST	0.004 *	0.002 *	0.033 *	0.024 *	0.003 +	0.005 *	0.003 *	0
🟥 Low-level	0.312 *	0.359 *	0.218 *	0.179 *	0.176 *	0.193 *	0.384 *	0.229 *

FIGURE 3 | Hierarchical regression. Data in the bottom row of the table is the initial *R*-value yielded from the low-level CV measures. Each row above indicates the change in *R*-value when the variables listed were added. Order of blocks are (1) Low-level (HueHist, SIFT, HOG, SSIM, entered simultaneously), (2) GIST, (3) GEOM (All, Gnd, Pos, Sky, Vrt, entered simultaneously), (4) SUN, (5) NEIL, (6) Behavioral. *Denotes changes in R that reached significance $p < 0.05$ corrected for multiple correlations; + denotes changes in R that reached significance $p < 0.05$ uncorrected.

- The most effective computer vision models were better than behaviorally-obtained ratings of scene similarity at accounting for variance in our neural data.

Of note, we found that regions of the brain selective for scene processing respond similarly to the same scenes, and treating, similar scenes as defined in one ROI as similar in another ROI, and, different scenes as defined in one ROI as different in another ROI. This pattern of results suggests that there is a stable encoding pattern for scenes within scene-selective brain regions and that voxel-to-voxel variation carries meaningful information regarding commonalities and differences between scenes.

These results suggest that, as a first step, applying computer vision models to neural data may allow us to better understand how scene information is encoded in neural systems. In particular, we view the application of NEIL as having the most promise in that its "vocabulary" of scene attributes does not ultimately depend on intuition, but rather on those regularities that can be learned from scene data. By way of example, NEIL includes visual features such as textures, color/shape combinations, and geometric configurations that do not readily correspond to any typical part label, but that may help enable NEIL's ability to categorize scenes. More generally, models such as NEIL offer better-specified theories of visual representation: it is our contention that NEIL and other artificial vision models offer meaningful—and testable—constraints at multiple levels of visual processing. With respect to our present results, we can now iterate toward more fine-grained tests of the most promising models (NEIL, SUN, GIST).

Beyond the well-specified representational constraints inherent in any functional model of computer vision, adopting multiple models also allowed us to consider a range of feature representations. In particular, the computer vision methods employed here ranged from analyzing low-level features, such as orientation information and spatial frequency, to high-level features, such as semantic categories. As expected, the low-level feature spaces (e.g., SIFT) were best correlated with patterns of voxel activity found in early visual brain regions, but were not highly positively correlated with the patterns of activity arising from scene-selective cortex. In contrast, as discussed, NEIL, SUN, and GIST gave rise to feature spaces that were most strongly correlated with the patterns of activity arising from scene-selective brain regions. Moreover, we found that NEIL's feature space, in particular, accounted for unique variance that could not be accounted for by any of the other methods. Together, our results indicate that the PPA, RSC, and TOS are involved in the processing of mid- to high-level features of scenes. We should note also that one curious result is the fact that NEIL accounted for significant variance in early visual areas. However, without a map of retinotopy for these early visual areas, it is difficult to say much about what NEIL's features may reveal about these processing areas.

Finally, we also observed that two models relying primarily on low-level features were significantly correlated with certain scene-selective brain regions. First, GIST correlated quite strongly with the RSC, replicating previous findings demonstrating a connection between GIST and the RSC functional properties (Watson et al., 2014). This suggests that the RSC may contribute to processing an image's spatial envelope or global scene properties which are known to be involved in scene understanding (Oliva and Torralba, 2006; Greene and Oliva, 2009). Moreover the RSC has been shown to process a representation of the scene that is abstracted from what is seen in the environment, typically processing a broader environment that extends beyond the current saccade (Epstein and Higgins, 2007; Park et al., 2007; Park and Chun, 2009). One possibility is that the RSC may process the low spatial frequencies or global properties of a scene that are strongly indicative of scene category. In addition, RSC was found to correlate with behavioral ratings of similarity, which was not found in the PPA or the TOS. That the correlations with GIST and behavior were unique to the RSC may suggest that RSC may provide a more categorical, or high-order representation of scenes. The second low-level model proved to be important were SIFT features in color domains that correlated strongly with multiple scene-selective regions: Hue SIFT showed strong correlations with the PPA and TOS, while RGB SIFT showed strong correlations with the RSC. In earlier work, junctures within scenes, which may be similar to SIFT features, were found to be important for scene categorization (Walther and Shen, 2014). Our results add to this finding by suggesting that key features but specifically within different color domains also carry information regarding scene categories. That is, scene-selective brain regions may rely on color cues in scene understanding—a claim consistent with earlier behavioral research on scene processing (Oliva and Schyns, 2000). At the same time, the lower correlations observed for the Hue Histogram model as compared to the Hue SIFT and RGB SIFT models suggest that it is not color *per se* that carries this information, but rather information about scene categories arises from an interaction of SIFT features within color domains. In particular, the perirhinal cortex—a region of the parahippocampal gyrus adjacent to the PPA—has been shown to unitize properties across an object; for example, that stop signs are red (Staresina and Davachi, 2010). As such, this function may extend to the parahippocampal region more generally being seen as unitizing diagnostic features, with the PPA supporting this function within scene processing.

In sum, we explored the visual dimensions underlying the neural representation of scenes using an approach in which models derived from computer vision are used as proxies for any psychological theory. While this approach may seem somewhat indirect, we argue that it is a necessary precursor in that extant psychological models have typically been somewhat underspecified with respect to the potential space of visual features. Humans can identify scenes effortlessly under a wide variety of conditions. For example, we can name scenes with near-equivalent accuracy when shown both photographs and line drawings, and with color present or absent. There is, then, no single feature dimension that drives the organization of scene-selective cortex. However, some dimensions are likely to prove more effective than others. Color is just one example of the many diagnostic cues that are used to aid in scene perception. There are almost surely a range of visual attributes and their associations within scenes that are diagnostic as to their categories and to which we are sensitive (Bar et al., 2008; Aminoff et al., 2013). Computer vision models, to the extent that they make representational assumptions with respect to scene attributes and their associations (i.e., models with a less well-understood representational basis may not actually be particularly informative), are, therefore, useful for better explicating those featural dimensions involved in human visual scene processing.

ACKNOWLEDGMENTS

Supported by Office of Naval Research—MURI contract number N000141010934 and National Science Foundation 1439237. We thank Carol Jew for help in data collection; and Ying Yang and Tim Verstynen for technical assistance.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fncom.2015.00008/abstract

REFERENCES

Aminoff, E. M., Kveraga, K., and Bar, M. (2013). The role of the parahippocampal cortex in cognition. *Trends Cogn. Sci.* 17, 379–390. doi: 10.1016/j.tics.2013.06.009

Arnott, S. R., Cant, J. S., Dutton, G. N., and Goodale, M. A. (2008). Crinkling and crumpling: an auditory fMRI study of material properties. *Neuroimage* 43, 368–378. doi: 10.1016/j.neuroimage.2008.07.033

Baldassi, C., Alemi-Neissi, A., Pagan, M., DiCarlo, J. J., Zecchina, R., and Zoccolan, D. (2013). Shape similarity, better than semantic membership, accounts for the structure of visual object representations in a population of monkey inferotemporal neurons. *PLoS Comput. Biol.* 9:e1003167. doi: 10.1371/journal.pcbi.1003167.s004

Bar, M., and Aminoff, E. (2003). Cortical analysis of visual context. *Neuron* 38, 347–358. doi: 10.1016/S0896-6273(03)00167-3

Bar, M., Aminoff, E., and Schacter, D. L. (2008). Scenes unseen: the parahippocampal cortex intrinsically subserves contextual associations, not scenes or places *per se*. *J. Neurosci.* 28, 8539–8544. doi: 10.1523/JNEUROSCI.0987-08.2008

Barenholtz, E., and Tarr, M. J. (2007). "Reconsidering the role of structure in vision," in *Categories in Use*, Vol. 47, eds A. Markman and B. Ross (San Diego, CA: Academic Press), 157–180.

Cant, J. S., and Goodale, M. A. (2011). Scratching beneath the surface: new insights into the functional properties of the lateral occipital area and parahippocampal place area. *J. Neurosci.* 31, 8248–8258. doi: 10.1523/JNEUROSCI.6113-10.2011

Cant, J. S., and Xu, Y. (2012). Object ensemble processing in human anterior-medial ventral visual cortex. *J. Neurosci.* 32, 7685–7700. doi: 10.1523/JNEUROSCI.3325-11.2012

Chen, X., Shrivastava, A., and Gupta, A. (2013). "NEIL: extracting visual knowledge from web data," in *IEEE International Conference on Computer Vision (ICCV)* (Sydney), 1409–1416.

Dalal, N., and Triggs, B. (2005). Histograms of oriented gradients for human detection. *IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recogn.* 1, 886–893. doi: 10.1109/CVPR.2005.177

Diedrichsen, J., and Shadmehr, R. (2005). Detecting and adjusting for artifacts in fMRI time series data. *Neuroimage* 27, 624–634. doi: 10.1016/j.neuroimage.2005.04.039

Epstein, R. A., and Higgins, J. S. (2007). Differential parahippocampal and retrosplenial involvement in three types of visual scene recognition. *Cereb. Cortex* 17, 1680–1693. doi: 10.1093/cercor/bhl079

Farhadi, A., Endres, I., Hoiem, D., and Forsyth, D. (2009). "Describing objects by their attributes," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Miami Beach, FL).

Felzenszwalb, P. F., Girshick, R. B., McAllester, D., and Ramanan, D. (2010). Object detection with discriminatively trained part-based models. *Pattern Anal. Mach. Intell. IEEE Trans.* 32, 1627–1645. doi: 10.1109/TPAMI.2009.167

Greene, M. R., and Oliva, A. (2009). Recognition of natural scenes from global properties: seeing the forest without representing the trees. *Cogn. Psychol.* 58, 137–176. doi: 10.1016/j.cogpsych.2008.06.001

Harel, A., Kravitz, D. J., and Baker, C. I. (2013). Deconstructing visual scenes in cortex: gradients of object and spatial layout information. *Cereb. Cortex* 23, 947–957. doi: 10.1093/cercor/bhs091

Hoiem, D., Efros, A. A., and Hebert, M. (2007). Recovering surface layout from an image. *Int. J. Comput. Vis.* 75, 151–172. doi: 10.1007/s11263-006-0031-y

Khaligh-Razavi, S., and Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLOS Comput. Biol.* 10:e1003915. doi: 10.1371/journal.pcbi.1003915

Kravitz, D. J., Peng, C. S., and Baker, C. I. (2011). Real-world scene representations in high-level visual cortex: it's the spaces more than the places. *J. Neurosci.* 31, 7322–7333. doi: 10.1523/JNEUROSCI.4588-10.2011

Lampert, C., Nickisch, H., and Harmeling, S. (2013). Attribute-based classification for zero-shot visual object categorization. *IEEE Trans. Pattern Anal. Mach. Intell.* 36, 453–465. doi: 10.1109/TPAMI.2013.140

Leeds, D. D., Seibert, D. A., Pyles, J. A., and Tarr, M. J. (2013). Comparing visual representations across human fMRI and computational vision. *J. Vis.* 13, 25. doi: 10.1167/13.13.25

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* 60, 91–110. doi: 10.1023/B:VISI.0000029664.99615.94

Naphade, M., Smith, J., Tesic, J., Chang, S., Hsu, W., Kennedy, L., et al. (2006). Large-scale concept ontology for multimedia. *IEEE Multimedia Mag.* 13, 86–91. doi: 10.1109/MMUL.2006.63

Nasr, S., Echavarria, C. E., and Tootell, R. B. H. (2014). Thinking outside the box: rectilinear shapes selectively activate scene-selective cortex. *J. Neurosci.* 34, 6721–6735. doi: 10.1523/JNEUROSCI.4802-13.2014

Nestor, A., Vettel, J. M., and Tarr, M. J. (2008). Task-specific codes for face recognition: how they shape the neural representation of features for detection and individuation. *PLoS ONE* 3:e3978. doi: 10.1371/journal.pone.0003978

Oliva, A., and Schyns, P. G. (2000). Diagnostic colors mediate scene recognition. *Cogn. Psychol.* 41, 176–210. doi: 10.1006/cogp.1999.0728

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. doi: 10.1023/A:1011139631724

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. doi: 10.1016/S0079-6123(06)55002-2

Park, S., Brady, T. F., Greene, M. R., and Oliva, A. (2011). Disentangling scene content from spatial boundary: complementary roles for the parahippocampal place area and lateral occipital complex in representing real-world scenes. *J. Neurosci.* 31, 1333–1340. doi: 10.1523/JNEUROSCI.3885-10.2011

Park, S., and Chun, M. M. (2009). Different roles of the parahippocampal place area (PPA) and retrosplenial cortex (RSC) in panoramic scene perception. *Neuroimage* 47, 1747–1756. doi: 10.1016/j.neuroimage.2009.04.058

Park, S., Intraub, H., Yi, D.-J., Widders, D., and Chun, M. M. (2007). Beyond the edges of a view: boundary extension in human scene-selective visual cortex. *Neuron* 54, 335–342. doi: 10.1016/j.neuron.2007.04.006

Park, S., Konkle, T., and Oliva, A. (2014). Parametric coding of the size and clutter of natural scenes in the human brain. *Cereb. Cortex*. doi: 10.1093/cercor/bht418. [Epub ahead of print].

Patterson, G., and Hays, J. (2012). "Sun attribute database: discovering, annotating, and recognizing scene attributes," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Providence, RI), 2751–2758.

Shechtman, E., and Irani, M. (2007). "Matching local self-similarities across images and videos," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Minneapolis, MN), 1–8.

Shrivastava, A., Singh, S., and Gupta, A. (2012). "Constrained semi-supervised learning using attributes and comparative attributes," in *Proceedings of European Conference on Computer Vision (ECCV)* (Florence), 369–383.

Stansbury, D. E., Naselaris, T., and Gallant, J. L. (2013). Natural scene statistics account for the representation of scene categories in human visual cortex. *Neuron* 79, 1025–1034. doi: 10.1016/j.neuron.2013.06.034

Staresina, B. P., and Davachi, L. (2010). Object unitization and associative memory formation are supported by distinct brain regions. *J. Neurosci.* 30, 9890–9897. doi: 10.1523/JNEUROSCI.0826-10.2010

van de Sande, K. E., Gevers, T., and Snoek, C. G. (2011). Empowering visual categorization with the GPU. *Multimedia IEEE Trans.* 13, 60–70. doi: 10.1109/TMM.2010.2091400

Vedaldi, A., and Fulkerson, B. (2010). VLFeat: an open and portable library of computer vision algorithms. *Proc. Int. Conf. Multimedia* 1469–1472. doi: 10.1145/1873951.1874249

Vedaldi, A., Ling, H., and Soatto, S. (2010). Knowing a good feature when you see it: ground truth and methodology to evaluate local features for recognition. *Stud. Comput. Intell.* 285, 27–49. doi: 10.1007/978-3-642-12848-6_2

Walther, D. B., and Shen, D. (2014). Nonaccidental properties underlie human categorization of complex natural scenes. *Psychol. Sci.* 25, 851–860. doi: 10.1177/0956797613512662

Wasserman, L. (2004). "The bootstrap", in *All of Statistics: A Concise Course of Statistical Inference* (New York, NY: Springer Publishing Company), 107–118.

Watson, D. M., Hartley, T., and Andrews, T. J. (2014). Patterns of response to visual scenes are linked to the low-level properties of the image. *Neuroimage* 99, 402–410. doi: 10.1016/j.neuroimage.2014.05.045

Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., and Torralba, A. (2010). "Sun database: large-scale scene recognition from abbey to zoo," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (San Francisco, CA), 3485–3492.

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., and DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl. Acad. Sci. U.S.A.* 111, 8619–8624. doi: 10.1073/pnas.1403112111

Yu, F., Ji, R., Tsai, M., Ye, G., and Chang, S. (2012). "Weak attributes for large-scale image retrieval," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Providence, RI), 2949–2956.

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 September 2014; accepted: 15 January 2015; published online: 04 February 2015.

Citation: Aminoff EM, Toneva M, Shrivastava A, Chen X, Misra I, Gupta A and Tarr MJ (2015) Applying artificial vision models to human scene understanding. Front. Comput. Neurosci. ***9****:8. doi: 10.3389/fncom.2015.00008*

This article was submitted to the journal Frontiers in Computational Neuroscience.

ORIGINAL RESEARCH
published: 05 November 2015
doi: 10.3389/fncom.2015.00135

Fourier power, subjective distance, and object categories all provide plausible models of BOLD responses in scene-selective visual areas

Mark D. Lescroart[1], *Dustin E. Stansbury*[2] *and Jack L. Gallant*[1, 2, 3*]

[1] Helen Wills Neuroscience Institute, University of California, Berkeley, Berkeley, CA, USA, [2] Vision Science Program, University of California, Berkeley, Berkeley, CA, USA, [3] Department of Psychology, University of California, Berkeley, Berkeley, CA, USA

Perception of natural visual scenes activates several functional areas in the human brain, including the Parahippocampal Place Area (PPA), Retrosplenial Complex (RSC), and the Occipital Place Area (OPA). It is currently unclear what specific scene-related features are represented in these areas. Previous studies have suggested that PPA, RSC, and/or OPA might represent at least three qualitatively different classes of features: (1) 2D features related to Fourier power; (2) 3D spatial features such as the distance to objects in a scene; or (3) abstract features such as the categories of objects in a scene. To determine which of these hypotheses best describes the visual representation in scene-selective areas, we applied voxel-wise modeling (VM) to BOLD fMRI responses elicited by a set of 1386 images of natural scenes. VM provides an efficient method for testing competing hypotheses by comparing predictions of brain activity based on encoding models that instantiate each hypothesis. Here we evaluated three different encoding models that instantiate each of the three hypotheses listed above. We used linear regression to fit each encoding model to the fMRI data recorded from each voxel, and we evaluated each fit model by estimating the amount of variance it predicted in a withheld portion of the data set. We found that voxel-wise models based on Fourier power or the subjective distance to objects in each scene predicted much of the variance predicted by a model based on object categories. Furthermore, the response variance explained by these three models is largely shared, and the individual models explain little unique variance in responses. Based on an evaluation of previous studies and the data we present here, we conclude that there is currently no good basis to favor any one of the three alternative hypotheses about visual representation in scene-selective areas. We offer suggestions for further studies that may help resolve this issue.

Keywords: scene perception, fMRI, voxel-wise modeling, encoding models, neuroscience, vision

OPEN ACCESS

Edited by:
Judith Peters,
The Netherlands Institute for Neuroscience, Netherlands

Reviewed by:
Chris I. Baker,
National Institutes of Health, USA
Ko Sakai,
University of Tsukuba, Japan

****Correspondence:***
Jack L. Gallant
gallant@berkeley.edu

Received: *12 July 2015*
Accepted: *19 October 2015*
Published: *05 November 2015*

Citation:
Lescroart MD, Stansbury DE and Gallant JL (2015) Fourier power, subjective distance, and object categories all provide plausible models of BOLD responses in scene-selective visual areas.
Front. Comput. Neurosci. 9:135.
doi: 10.3389/fncom.2015.00135

INTRODUCTION

fMRI experiments have shown that natural scene perception activates several distinct functional areas in the human cerebral cortex. These include the Parahippocampal Place Area (PPA), Retrosplenial Complex (RSC), and the Occipital Place Area (OPA, also known as the Temporal Occipital Sulcus or TOS) (Aguirre et al., 1998; Epstein and Kanwisher, 1998; Maguire, 2001;

Nasr et al., 2011; Dilks et al., 2013). Which specific scene-related features are represented in these areas has been the subject of substantial debate.

Several qualitatively different scene-related features have been proposed to be represented in scene-selective areas. Some studies have suggested that these areas represent simple 2D features related to the Fourier power spectrum (Rajimehr et al., 2011; Nasr and Tootell, 2012; Nasr et al., 2014; Watson et al., 2014). Others have argued that PPA, RSC, and OPA represent features related to 3D spatial structure, such as expanse or openness (Kravitz et al., 2011; Park et al., 2011), the distance from objects in a scene to an observer (Amit et al., 2012; Park et al., 2015), or the size of objects in a scene (Cate et al., 2011; Konkle and Oliva, 2012). A third position is that scene-selective areas represent information about the semantic categories of natural scenes or their constituent objects (Walther et al., 2009, 2011; Huth et al., 2012; Stansbury et al., 2013).

Previous studies have not resolved which of these hypotheses provides the best account of the representation of natural scenes in scene-selective areas. One reason that this has been a difficult issue to resolve is that almost every previous study of scene-selective cortical areas has used stimuli that were pre-selected or manipulated to maximize variation in specific stimulus features of interest. Consequently, different experiments use different stimuli, and thereby sample different ranges of variation in stimulus features. If the brain operated according to purely linear mechanisms, this would not cause any problems for scientific interpretation of the results. However, feature tuning in the human visual system is conferred by nonlinear mechanisms that operate at all levels of the visual hierarchy (Van Essen et al., 1992). In such a nonlinear system, responses to a limited range of stimulus variation cannot necessarily be used to infer responses to stimulus variation outside that range (Wu et al., 2006; Gallant et al., 2012). Thus, any experiment that constrains stimulus variation may fail to characterize nonlinear tuning properties for stimuli (or stimulus features) that fall outside the experiment's pre-selected stimulus set.

The most straightforward way to probe the visual system in an ecologically valid range is to use a broad distribution of natural images as stimuli. The human visual system is exquisitely tuned to the statistical variance and covariance of features in natural images (Field, 1987; Simoncelli and Olshausen, 2001). Thus, one efficient way to determine what features are represented in scene-selective areas is to record brain activity elicited by a wide range of natural scenes, extract features from the stimulus images that reflect the various hypotheses, and then determine which features best account for the measured brain activity (Naselaris et al., 2009, 2012; Nishimoto et al., 2011; Stansbury et al., 2013).

In this study, we analyzed BOLD fMRI responses to a large set of natural photographs to determine which features of natural scenes are represented in PPA, RSC, and OPA. We employed a voxel-wise modeling (VM) approach in which we directly compared predictive models based on three different classes of scene-related features: 2D features derived from the Fourier power spectrum of each scene, the distance to salient objects in each scene, and semantic categories of the constituent objects in each scene. For each class of features, we defined a feature space to formalize each alternative hypothesis in quantitative terms.

To estimate the relationship between each feature space and measured BOLD responses, we used linear regression to fit each feature space to the fMRI data recorded from each voxel in the posterior part of the brain (encompassing the visual cortex). Each feature space and its associated β weights constitute an encoding model that maps a stimulus onto brain responses. We evaluated each model based on how accurately it predicted BOLD responses in a separate validation data set. Finally, we applied a variance partitioning analysis to determine whether different models predict unique or shared variance in BOLD responses.

METHODS

The data used for this experiment came from previously published studies from our laboratory. The four subjects in this experiment are the same four subjects as in Stansbury et al. (2013). Data for two of these subjects (subjects 1 and 2) were originally collected for Naselaris et al. (2012). Here we provide a brief description of the stimuli, subjects, data collection, and image response estimation. For full details, see Stansbury et al. (2013).

fMRI Data Acquisition and Preprocessing

All fMRI data were collected at the UC Berkeley Brain Imaging Center using a 3 Tesla Siemens Tim Trio MR Scanner (Siemens, Germany). Data were collected from each of four human subjects (1 female) while they viewed 1386 natural images. The data were collected over six or seven scanning sessions for each subject, and the total scan time per subject was 4 h and 53 min. Voxels were approximately 2.25 × 2.25 × 2.99 mm, and the repetition time (TR) was approximately 2 s. The fMRI scan protocol used for subject one was slightly different from the protocol used for the others; see Stansbury et al. (2013) for full details. Anatomical scans were acquired for each subject using a T1-weighted magnetization-prepared rapid gradient echo (MP-RAGE) sequence. All subjects gave their written informed consent to participate, and the experimental protocol was approved by the UC Berkeley Committee for the Protection of Human Subjects.

Freesurfer was used to automatically extract cortical surfaces from the T1-weighted scans (Dale et al., 1999). These surfaces were manually edited to improve the match to the anatomical data. Surface flattening and visualization were performed with Freesurfer and custom python code (Gao et al., 2015; available at http://github.com/gallantlab/pycortex).

Functional MRI data were preprocessed using custom Matlab (R2014a, MathWorks) code and SPM8 (http://www.fil.ion.ucl.ac.uk/spm/software/spm8/). For each subject, data were motion corrected and coregistered to the first volume collected. The motion correction and coregistration transformations were concatenated, and the data were re-sliced only once. Data were divided into two separate subsets: one used for model estimation and one used for model validation. The preprocessed BOLD responses were de-convolved into a unique hemodynamic response per voxel and a unique response amplitude per image per voxel. Response amplitudes for the validation data were estimated slightly differently from the method in Stansbury et al. (2013) in order to obtain an estimate of the noise in each voxel (see Noise Ceiling Estimation section below).

Stimuli

The experimental stimuli consisted of 1386 photographs of natural scenes. Most of the photographs used in this study were selected first from a collection of 4000 labeled images curated by the Lotus Hill Institute (Wuhan, China). The labels provided by the Lotus Hill Institute were used to re-label all images as containing (or not containing) each of four non-mutually exclusive superordinate categories: *animal*, *human*, *manmade*, and *natural*. Animals and humans were prioritized because animacy was a principal feature of interest in Naselaris et al. (2012). An additional 242 images were downloaded from Google Images, in order to increase the number of scenes containing both animals and humans. Finally, 1386 images were randomly chosen from the full set of 4242 images, such that approximately the same number of images had the labels *animal* and *human*, and (independently) such that approximately the same number of images had the labels *natural* and *manmade*. Thus, images were not specifically selected based on the features of interest in this study. All four subjects saw the same 1386 stimulus images. Figure S02 shows all 126 validation images shown to all four subjects.

Images subtended 20° × 20° of visual angle (500 × 500 pixels). Each image presentation consisted of five brief flashes in 1 s, followed by 3 s of isoluminant gray screen. The 1260 images in the estimation data set were repeated twice each. The 126 images in the validation data set were repeated 12 times each. During the experiment subjects maintained steady fixation on a small (0.2° × 0.2°) square that changed colors at 3 Hz. Subjects were instructed to try to understand each scene as it was presented, but had no explicit task besides maintaining fixation. Stimulus presentation and all statistical analyses were conducted using custom Matlab (R2014a, MathWorks) and python code.

Feature Spaces Used for Voxel-wise Encoding Models

In voxel-wise modeling, a *feature space* is a quantification of the features of a stimulus that are hypothesized to be related to brain responses (Naselaris et al., 2011; Gallant et al., 2012). For this study we created three different feature spaces: a Fourier power feature space, subjective distance feature space, and an object category feature space. Each feature space embodies a different hypothesis about which features are represented in scene-selective areas.

Fourier Power Feature Space

To parameterize variation in spatial frequency energy at different orientations, we created a Fourier power feature space. First, the color images were converted to Commission Internationale de l'Éclairage L*A*B* color space, and the luminance layer was extracted. A 2D Fourier transform was computed for each luminance image. The amplitude spectrum for each image was divided into eight bins: one high-frequency and one low-frequency bin at each of four orientations (0, 45, 90, and 135°). The divide between high and low frequency bins was set at five cycles/degree, as in Rajimehr et al. (2011) and Nasr and Tootell (2012). A schematic of the Fourier domain bins is shown in **Figure 1C**. Fourier power was averaged over each bin for each image. To reduce correlations between Fourier power bins, each bin in each image was divided by the L2 norm of all bins for that image. The L2 norm itself was retained as a separate feature reflecting the overall spatial frequency energy in each image. Thus, the final Fourier power feature space consisted of nine feature channels: one total spatial frequency energy channel (i.e., the L2 norm), four low spatial frequency channels, and four high spatial frequency channels (One can think of each feature channel as a separate column in a regression design matrix). To match the range of variation in the Fourier power feature channels to the range of the z-scored BOLD responses, each feature channel was z-scored separately across all images.

Subjective Distance Feature Space

To parameterize distance in each scene, we created a subjective distance feature space based on human distance judgments. Human raters were instructed to estimate the distance to the main content (the most salient or subjectively important objects) in each of the 1386 stimulus images. The determination of the main content of each image was left to the discretion of each rater, so these distance ratings were inherently subjective. For each image, raters chose one of five roughly logarithmically spaced distance bins: (1) extreme closeup, ∼1–2 ft., (2) arm's length, ∼3–4 ft., (3) nearby/same room, <20 ft., (4) semi-distant, <100 ft., or (5) far away, >100 ft. Raters viewed each image for 300 ms before making each rating. These brief durations approximated the brief image presentation time used in the fMRI experiment. Raters had the option to repeat an image if they felt they had not adequately understood it, but repeated viewing was discouraged. Three different raters provided distance judgments. Two of them were also subjects in the fMRI experiment. The ratings produced by the three raters were consistent: the correlations between the three raters' distance ratings were 0.845, 0.857, and 0.861. The median distance rating for each image across all three raters was used to code the features. The final subjective distance feature space consisted of five mutually exclusive binary feature channels (one for each distance bin).

Object Category Feature Space

To parameterize semantic variation in our stimulus images, we used an object category feature space based on human-assigned labels indicating the presence of objects or other scene elements (such as land, water, and sky) in each image. This feature space was originally created for an earlier study (Naselaris et al., 2012). For a full description of the labeling process, see the original paper. Briefly, 15 human raters assigned natural language labels to each object in each image. These labels were binned into 19 categories: creepy animal (e.g., insects, snakes, and reptiles), bird, fish, water mammal, land mammal, many humans, few humans, vehicle, artifact, text, prepared food, fruit vegetable, other plants, furniture, sky, water, land,

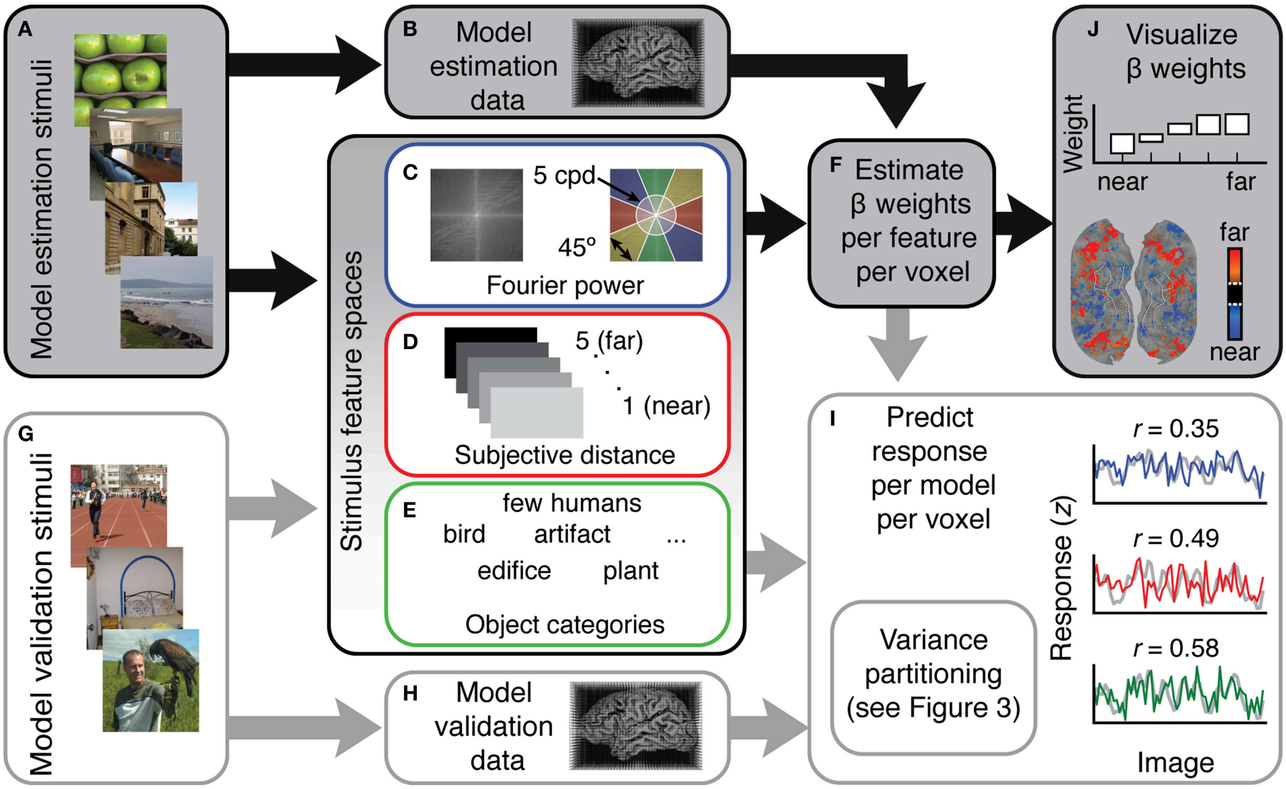


FIGURE 1 | Overview of the voxel-wise modeling (VM) procedure used in this study. (A) Human subjects were shown 1260 natural images while **(B)** fMRI data were recorded. **(C–E)** These data were modeled as a function of three different feature spaces. Each feature space reflects a different hypothesis about which features are represented in scene-selective areas. **(C)** For the Fourier power mode, the feature space was computed by taking the Fourier transform of each stimulus image and then averaging the amplitude spectrum over the orientation and spatial frequency bins shown at right. **(D)** For the subjective distance model, the feature space consisted of ratings from three humans who judged whether the main content of each stimulus scene was (1) <2 ft away, (2) <4 ft away, (3) <20 ft away, (4) <100 ft away, and (5) >100 ft away. **(E)** For the semantic category model the feature space consisted of labels from three human raters who labeled the objects in each stimulus image using 19 semantic labels. **(F)** Ordinary least squares regression was used to find a set of weights (β) that map the features in each model onto the BOLD responses in each voxel. Each feature space and its associated β weights constitute a different encoding model. **(G)** In order to validate the models in an independent data set, the same subjects were shown a different set of 126 images while **(H)** fMRI responses were collected. **(I)** To assess model accuracy, the β weights estimated from the training data were used to predict responses in this withheld model validation data set. **(J)** To reveal patterns of tuning in the features quantified by each different model, pre-specified *t* contrasts were computed between β weights in each model and projected onto the cortical surface, and β weights were averaged over voxels in different regions of interest and plotted.

part of building, and edifice. These categories span several superordinate categories known to be represented in higher-order visual areas (animate/inanimate, large/small, human/non-human). Thus, the full object category feature space consisted of 19 non-exclusive binary feature channels, each indicating the presence of a different object category in each stimulus image. In previous work models based on this feature space have been shown to provide accurate predictions of BOLD responses in several higher-order visual areas (Naselaris et al., 2012). This object category model also provides a simple approximation of the WordNet (Miller, 1995) feature space used to model BOLD data in Huth et al. (2012).

These three feature spaces were chosen as simple examples of three broader classes of hypotheses regarding the representation in scene-selective areas: that scene-selective areas represent low-level, image-based features, 3D spatial information, and categorical information about objects and scenes. Many other implementations of these broad hypotheses are possible, but an exhaustive comparison of all of the potential models is impractical at this time. Instead, here we focus on just three specific feature spaces that each capture qualitatively different information about visual scenes and that are simple to implement. We emphasize simplicity here for instructional purposes, for ease of interpretation, and to simplify the model fitting procedures and variance partitioning analysis presented below.

Model Fitting and Evaluation

We used ordinary least squares regression to find a set of weights (β) that map the feature channels onto the estimated BOLD responses for the model estimation data (**Figure 1H**). Separate β weights were estimated for each feature channel and for each voxel. Each β weight reflects the strength of the relationship between variance in a given feature channel and variance in the

BOLD data. Thus, each β weight also reflects the response that a particular feature is likely to elicit in a particular voxel. The model β weights as a whole demonstrate the *tuning* of a voxel or an area to specific features within the feature space for that model. The full set of β weights for all feature channels for a voxel constitute an encoding model for that voxel. Note that many previous fMRI studies from our laboratory (Nishimoto et al., 2011; Huth et al., 2012; Stansbury et al., 2013) have used ridge regression or another regularized regression procedure to produce voxel-wise encoding models that have the highest possible prediction accuracy. We did not use regularized regression in the current study because the use of regularization complicates interpretation of the variance partitioning analysis described below. Furthermore, the number of features in each model fit here was small relative to the amount of data collected, so regularization did not improve model performance.

Many studies describe the tuning of voxels across the visual cortex by computing *t* contrasts between estimated regression β weights for each voxel (Friston et al., 1994). To facilitate comparison of our results to the results of several such studies, we computed three *t* contrasts between β weights in each of our three models. Each contrast was computed for all cortical voxels. Using the β weights in the Fourier power model, we computed a contrast of cardinal vs. oblique high-frequency orientations (Nasr and Tootell, 2012). This contrast was specifically (+ high freq 0° + high freq 90° – high freq 45° – high freq 135°) (see **Figure 4** for feature naming scheme). Using the β weights in the subjective distance model, we computed a contrast of far vs. near distances (+ v. far + distant – near – closeup) (Amit et al., 2012; Park et al., 2015). Using the β weights in the object category model, we computed a contrast of people vs. buildings (+ few people –0.5 edifice –0.5 part of building) (Epstein and Kanwisher, 1998). Since these contrasts were computed for every voxel in the brain, the *p*-values for each *t* contrast were adjusted using False Discovery Rate (FDR) with an α level of 0.05 to correct for multiple comparisons (Benjamini and Yekutieli, 2001).

To evaluate the accuracy of each model, we used the model fit to each voxel to predict BOLD responses of the same voxel in the validation data set. Prediction accuracy was assessed by computing Pearson's product-moment correlation (r) between the predicted response and the validation response estimated for each voxel. To convert prediction accuracy to an estimate of the variance explained, we squared the prediction accuracy (r) for each model in each voxel value while maintaining its sign (David and Gallant, 2005).

Noise Ceiling Estimation

Noise in the validation data set will nearly always bias prediction accuracy downward, and the magnitude of this bias may differ across voxels. This makes raw prediction accuracy difficult to interpret: for any given voxel, imperfect predictions may be caused by a flawed model, measurement noise, or both. To correct this downward bias and to exclude noisy voxels from further analyses, we used the method of Hsu et al. (Hsu et al., 2004; Huth et al., 2012) to estimate a noise ceiling (γ) for each voxel in our data. The noise ceiling is the amount of response variance in the validation data that could theoretically be predicted by the perfect model.

Noise ceiling estimation requires repeated measurement of responses to the same stimulus (Hsu et al., 2004). Thus, we estimated 11 different responses to each of our validation stimuli for each voxel. We split the validation data into 11 partially overlapping blocks. Each block contained two presentations of each stimulus image. The first block contained the first and second presentations of each image, the second block contained the second and third presentations of each image, and so on. For each block, the BOLD data were de-convolved into a unique hemodynamic response per voxel and a unique response amplitude per image per voxel. This procedure resulted in 11 different estimates of the response to each of our validation images for each voxel. These 11 validation image response estimates were used to compute the noise ceiling (γ) for each voxel.

γ can be interpreted as a measure of signal repeatability. If the same stimuli reliably elicit similar responses, γ is high (near one); if not, it is low (near zero). To give a sense for this metric, **Figure 2** shows estimated responses for three voxels with noise ceilings (γ-values) that are relatively high, average, and just above chance. Estimated γ-values were used to select voxels for all analyses presented in this paper. Voxels with noise ceilings greater than $\gamma = 0.04$ [a value corresponding to bootstrapped $p(\gamma) < 0.01$ for a single voxel] were retained, and all others were discarded before further analysis. In auditory cortex, where the signal should not be strongly related to the stimuli in this experiment, this threshold retains approximately five percent of the voxels. Figure S01 shows the absolute number of voxels kept, the percent of voxels kept, and the mean γ-value for each region of interest for each subject.

The noise ceiling was also used to normalize prediction accuracy in order to estimate the proportion of potentially explainable response variance that is actually explained by the models. The square root of the noise ceiling ($\gamma^{1/2}$) gives the theoretical maximum correlation between predicted and observed responses for each voxel. Following Hsu et al. (2004), all estimates of prediction accuracy were divided by $\gamma^{1/2}$. Estimates of variance explained were divided by γ. Note that very low noise ceilings can result in divergent normalized correlation estimates. For example, for $\gamma = 0.0001$ and $r = 0.07$, the normalized value of r would be $0.07/0.0001^{1/2} = 7$. Our voxel selection criterion allows us to avoid such divergent estimates, since all voxels with low γ-values are discarded.

Model Comparison

To determine which features are most likely to be represented in each visual area, we compared the predictions of competing models on a separate validation data set reserved for this purpose. First, all voxels whose noise ceiling failed to reach significance [$\gamma > 0.04$, $p(\gamma) > 0.01$ uncorrected] were discarded. Next, the predictions of each model for each voxel were normalized by the estimated noise ceiling for that voxel. The resulting values were converted to z scores by the Fisher transformation (Fisher, 1915). Finally, the scores for each model were averaged separately across each ROI.

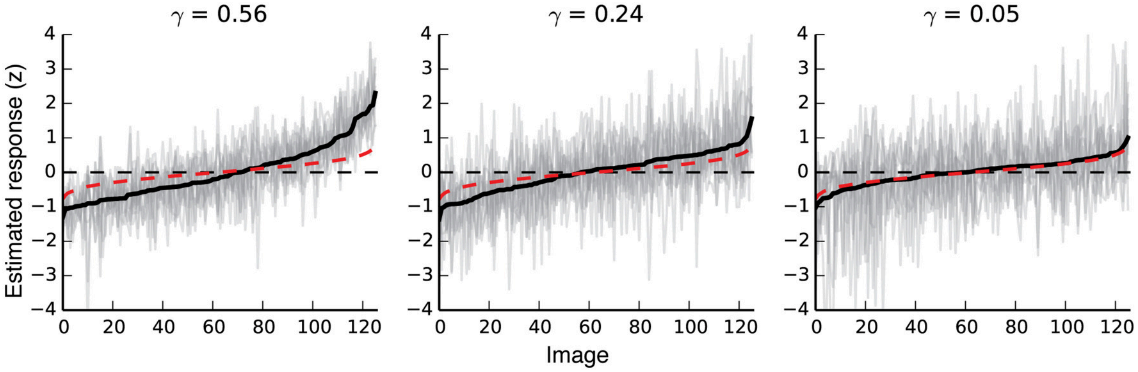


FIGURE 2 | Response variability in voxels with different noise ceilings. The three plots show responses to all validation images for three different voxels with noise ceilings that are relatively high, moderate, and just above chance. The far-right plot shows the response variability for a voxel that meets our minimum criterion for inclusion in further analyses. Black lines show the mean response to each validation image. For each plot, images are sorted left to right by the average estimated response for that voxel. The 11 gray lines in each plot show 11 separate estimates of response amplitude per image for each voxel. Red dotted lines show random responses (averages of 11 random Gaussian vectors sorted by the mean of the 11 random vectors). Note that even random responses will deviate slightly from zero at the high and low ends, due to the bias induced by sorting the responses by their mean.

For each ROI, a permutation analysis was used to determine the significance of model prediction accuracy (vs. chance), as well as the significance of *differences* between prediction accuracies for different models. For each feature space, the feature channels were shuffled across images. Then the entire analysis pipeline was repeated (including fitting β weights, predicting validation responses, normalizing voxel prediction correlations by the noise ceiling, Fisher *z* transforming normalized correlation estimates, averaging over ROIs, and computing the average difference in accuracy between each pair of models). This shuffling and re-analysis procedure was repeated 10,000 times. This yielded a distribution of 10,000 estimates of prediction accuracy for each model and for each ROI, under the null hypothesis that there is no systematic relationship between model predictions and fMRI responses. Statistical significance was defined as any prediction that exceeded 95% of all of the permuted predictions ($p = 0.05$), calculated separately for each model and ROI. Note that different numbers of voxels were included in each ROI, so different ROIs had slightly different significance cutoff values. Significance levels for differences in prediction accuracy between models were determined by taking the 95th percentile of the distribution of differences in prediction accuracy between randomly permuted models ($p = 0.05$).

Variance Partitioning

Estimates of prediction accuracy can determine which of several models best describes BOLD response variance in a voxel or area. However, further analysis is required to determine whether two models each explain unique or shared variance in BOLD responses. For example, consider two hypothetical models A and B. Suppose that model A makes slightly more accurate predictions than does model B for a given voxel. One possibility is that the variance explained by model B is a subset of the larger variance explained by model A. Another possibility is that model B explains a unique and complementary component of the response variance that is not explained by model A (For example, even if model B is worse overall it might make more accurate predictions than model A for a subset of images). **Figure 3B** shows two simulated examples in which competing models explain unique and shared response variance.

We performed a variance partitioning analysis (**Figure 3**) to determine the extent to which the three models in this study predict unique or shared components of the response variance in each scene-selective area. First, β weights were fit to each feature space independently (**Figure 1**). Then, feature spaces were concatenated in the features dimension (**Figure 3A**) for each possible pair or trio of feature spaces (Fourier power ∪ subjective distance, Fourier power ∪ semantic categories, subjective distance ∪ semantic categories, and Fourier power ∪ subjective distance ∪ semantic categories). For example, the feature space matrix resulting from the concatenation of all three models had 33 feature channels (nine from the Fourier power model, five from the subjective distance model, and 19 from the semantic category model). Each concatenated feature space was fit to the data for each voxel, and used to predict responses in the validation data for each voxel. Prediction accuracy was converted to variance explained by squaring the prediction correlation while maintaining its sign.

For pairwise variance partitioning, the unique and shared variance explained by each model or pair of models was computed according to the equations in **Figure 3C**. Similarly straightforward arithmetic was used to perform three-way variance partitioning to compute each element of the Venn diagram in **Figure 9**. For example, the unique variance explained by the semantic category model was estimated as the difference between variance explained by the full, 3-part concatenated model (Fourier power ∪ subjective distance ∪ semantic category) and the 2-part concatenation of the Fourier power and subjective distance models (Fourier power ∪ subjective distance).

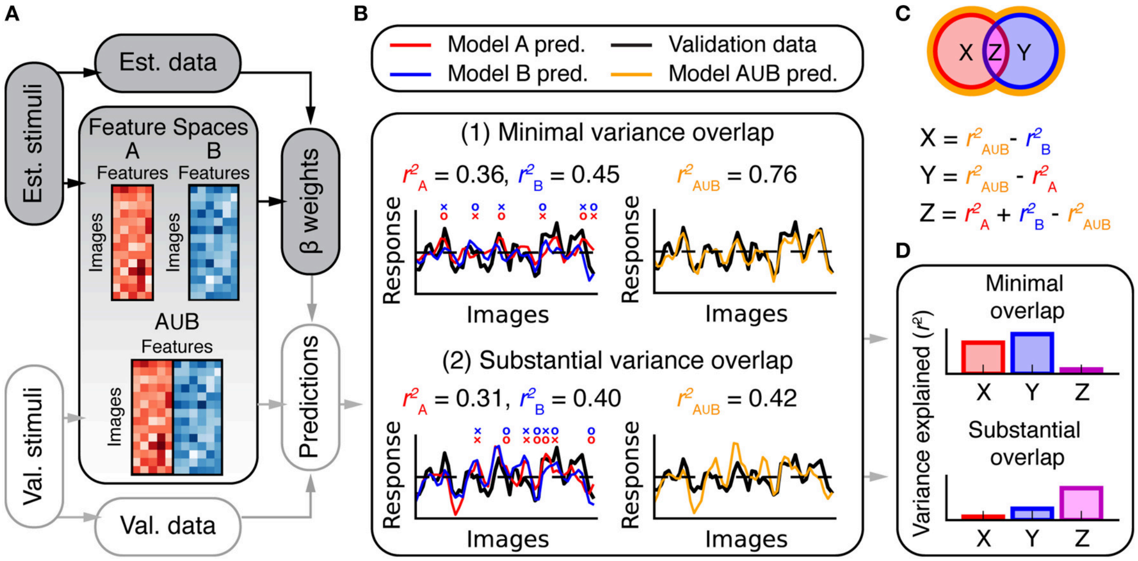


FIGURE 3 | Overview of variance partitioning analysis. Variance partitioning determines what fraction of variance in BOLD responses is shared between two models. **(A)** To estimate the amount of shared variance between each pair or trio of feature spaces, all pairs or trios of feature spaces were concatenated (in the features dimension) and the resulting combined feature spaces were fit to the data and used to compute predictions of the validation data. **(B)** Two simulated models that predict (1) independent variance and (2) shared variance. In (1), each model tends to make accurate predictions (o marks) where the other fails (× marks). Consequently, the combined model (A∪B) performs well. In (2), both models succeed and fail for the same images (that is, the predictions are correlated). Consequently, the combined model does not perform better than the individual models. The total variance explained by models A and B can be subdivided into the partitions shown in the Venn diagram in **(C).** Each partition corresponds to variance explained by: (X) only model A, (Y) only model B, and (Z) both A and B (shared variance). The variance explained by the combined model ($r^2_{A∪B}$) provides an estimate of the convex hull of the Venn diagram (shown by the orange border). Thus, X, Y, and Z can be computed as shown. **(D)** Bar graphs of the values for X, Y, and Z computed for the two cases in **(B)**.

Evaluation of Correlations between Stimulus Features

One risk associated with the use of natural images as stimuli is that features in different feature spaces may be correlated. If some of the features in different feature spaces are correlated, then models based on those feature spaces are more likely to generate correlated predictions. And if model predictions are correlated, the variance explained by the models will be shared (see **Figure 3**). To explore the consequences of correlated features, we computed the Pearson correlation (r) between all features in the Fourier power, subjective distance, and object category feature spaces. To determine whether the correlations between features that we measure in our stimulus set are general to many stimulus sets, we also explored feature correlations in two other stimulus sets (from Kravitz et al., 2011 and Park et al., 2015—see Supplementary Methods).

Non-zero correlations between a subset of the features in different feature spaces may or may not give rise to models that share variance. Two partially correlated feature spaces are most likely to lead to models that share variance if the feature channels that are correlated are also correlated with brain activity.

For example, imagine two simple feature spaces A and B, each consisting of three feature channels. A and B are used to model some brain activity, Y. Suppose that the first feature channel in A (A_1) is correlated with the first feature channel in B (B_1) at $r = 0.5$, and that the other feature channels (A_2, A_3, B_2, and B_3) are not correlated with each other or with Y at all. If A_1 and B_1 are both correlated with Y, then a linear regression that fits A and B to Y will assign relatively high β weights to A_1 and B_1 in the fit models (call the fit models M_A and M_B). This, in turn, will make the predictions of M_A and M_B more likely to be correlated. Thus, M_A and M_B will be more likely to share variance.

Now, imagine a second case. Suppose instead that A_1 and B_1 are correlated with one another but neither A_1 nor B_1 is correlated with Y. Suppose that the other feature channels in A and B are correlated with Y to varying degrees. In this case, A_1 and B_1 will be assigned small β weights when A and B are fit to Y. The small β weights on A_1 and B_1 will mean that those two channels (the correlated channels) will not substantially affect the predictions of M_A and M_B. Thus, in this case, the predictions of M_A and M_B will not be correlated, and M_A and M_B will each explain unique variance. These two simple thought experiments illustrate how the emergence of shared variance depends on correlations between feature channels and the β weights on those feature channels.

To illustrate how the correlations between features in this specific study interact with the voxel-wise β weights for each feature to produce shared variance across models, we conducted a simulation analysis. In brief, we simulated voxel responses based on the real feature values and two sets of β weights and performed variance partitioning on the resulting data. First, we used the concatenated stimulus feature spaces (X) and a set of semi-random weights (β) to generate simulated voxel data, according to the regression equation:

$$Y_{\text{sim}} = X\beta + \varepsilon \quad (1)$$

ε is Gaussian noise ∼N(0,1). To assure that the simulated data had approximately the same signal-to-noise ratio as the fMRI data in our experiment, we modified the basic regression equation to scale the noise according to a distribution of expected correlations (ρ), thus:

$$Y_{\text{sim}} = \rho X\beta + (1 - \rho^2)^{1/2}\varepsilon \quad (2)$$

We simulated the same number of voxels that we measured in all the scene-selective areas in all four subjects (761 voxels). We used the following procedure to assure that the simulation β weights were plausible given the covariance structure of the different feature spaces. First, we generated 761 different sequences of Gaussian random noise. Then we used ordinary least squares regression to fit β weights for each feature channel to the noise sequences. This resulted in 761 sets of β weights that map the feature spaces onto random data. Since ordinary least squares regression uses the feature covariance matrix to estimate β weights, the β weights generated by this procedure are guaranteed to be plausible given the covariance of the feature channels. Each set of semi-random β weights was then used to generate a simulated voxel timecourse according to Equation (2) above. We also created a second set of simulated data, based on the actual β weights we estimated for each of the 761 voxels in the experiment.

To illustrate how the specific β weights (the real β weights or the semi-random β weights) affected estimates of shared variance, we applied the same variance partitioning analysis that we applied to the fMRI data to both sets of simulated data. Note that the results of the variance partitioning of the simulated data based on the real β weights should match the results of the variance partitioning of the BOLD data. We include these results to show that our simulation procedure is operating as expected, and to demonstrate that any difference between the two simulations is a result of differences in the weights, and not anything to do with the simulation procedure.

Functional Area Localizers

Visual areas in retinotopic visual cortex as well as functionally defined category-selective visual areas were identified in separate scan sessions using conventional methods (Spiridon et al., 2006; Hansen et al., 2007). Scene-selective areas PPA, RSC, and OPA were all defined by a contrast of places vs. objects. The Fusiform Face Area (FFA) was defined by a contrast of faces vs. objects. The boundaries of each area were hand drawn on the cortical surface at the locations at which the *t* statistic for the contrast of places vs. objects changed most rapidly.

RESULTS

To investigate how natural scenes are represented in scene-selective areas in the human brain, we analyzed BOLD fMRI signals evoked by a large set of natural images (These data were collected for two studies from our laboratory that were published previously: Naselaris et al., 2012 and Stansbury et al., 2013). We tested three specific hypotheses about scene representation in these areas that have been proposed in previous studies: that scene selective areas represent Fourier power, subjective distance, and object categories. To formalize each of these hypotheses, we defined three feature spaces that quantified three classes of features: Fourier power at different frequencies and orientations, distance to the salient objects in each scene, and the semantic categories of objects and other components of each scene. To determine the relationship between each feature space and brain activity, we used ordinary least squares regression to estimate sets of β weights that map each feature space onto the BOLD fMRI responses in the model estimation data set.

We present our results in four sections. First, we examine the tuning revealed by the estimated model β weights in V1, the FFA, the PPA, RSC, and the OPA. Second, we estimate the importance of each feature space by predicting responses in a withheld data set. Third, we evaluate whether each of these feature spaces predicts unique or shared response variance in the fMRI data. Finally, we investigate the correlations between features in the Fourier power, subjective distance, and object category feature spaces.

Voxel-wise Model β Weights Replicate Tuning Patterns described in Previous Studies

The voxel-wise model β weights for the features in each model are shown in **Figures 4**, **5**. For each area, all voxels for each subject that met our voxel selection criterion [$\gamma > 0.04, p(\gamma) < 0.01$—see Methods] are shown. Overall, the tuning profiles revealed by the β weights in each area appear to be broadly consistent with tuning revealed by previous studies. We first describe the β weights in two comparably well-understood areas (V1 and FFA), and then describe the β weights for each model for all three scene-selective areas.

In V1, the β weights for the Fourier power model (**Figures 4A–C**) show that images containing high Fourier power tend to elicit responses above the mean. This is consistent with many studies showing that V1 responses increase with increasing image contrast (Albrecht and Hamilton, 1982; Gardner et al., 2005). The β weights for the subjective distance model show that very distant scenes elicit responses below the mean in most V1 voxels. This is likely because the most distant scenes (such as the image of the ocean in **Figure 1A**) have low overall Fourier power. The β weights for the object category model show that the images with labels for *fruit and vegetable*, *prepared food*, and *creepy animal* all elicit responses above the mean. These are also likely be related to different levels of Fourier power. We analyze the correlations between Fourier power and specific object categories, as well as other correlations between feature channels in different models, in detail below.

In FFA, the β weights for the Fourier power model (**Figures 4D–F**) show that images with high frequency energy at 135° tended to elicit BOLD responses above the mean, while high frequency energy at vertical and horizontal (90° and 0°) orientations elicit responses below the mean. Several previous studies have rigorously argued that FFA responds to faces rather than low-level image features (Kanwisher and

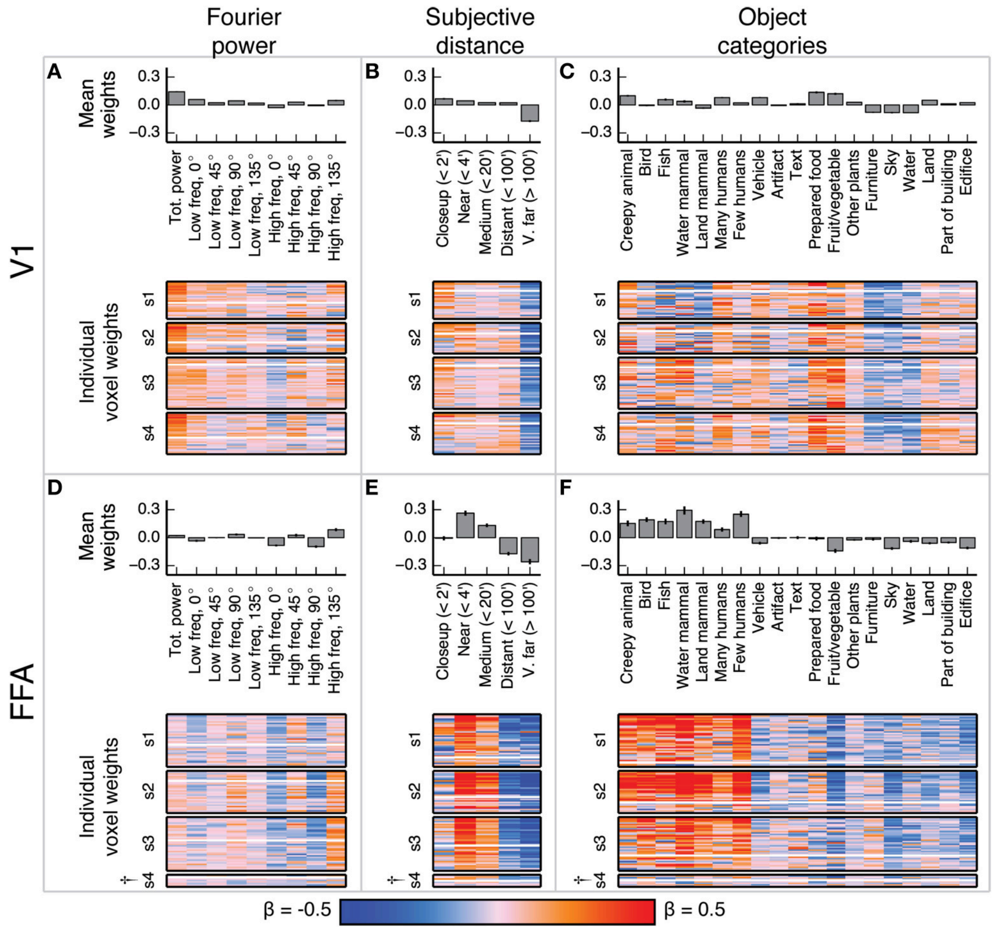


FIGURE 4 | Voxel-wise model β weights for all models for all voxels in V1 and FFA. (A) Model β weights for the Fourier power model for V1. The image in the lower part of the panel shows the weight for every voxel in V1 that met our selection criterion [$\gamma > 0.04$, $p(\gamma) < 0.01$, see Methods]. Voxels are separated by subject (s1–s4), and the relative size of each subject's section indicates the relative number of voxels selected in V1 for that subject. † marks indicate specific ROIs in specific subjects with low signal quality (and thus few voxels selected for analysis). See Figure S01 for evaluation of signal across subjects. Each horizontal stripe through the image shows the weights for a different voxel. Voxels are sorted within each subject by normalized prediction accuracy for the Fourier power model. Weights from the model that produced the most accurate predictions in V1 are at the top. The solid white line across the image for each subject shows the chance threshold for prediction accuracy ($p < 0.05$, FDR corrected). The bar graph at the top of the panel shows the mean β weights for all V1 voxels for all subjects. Each text label corresponds to both the bar above it and the column of weights below it. Error bars are 99% confidence intervals across all voxels. These tuning patterns are consistent with known response properties of V1, where voxel responses are related to the amount of Fourier power in each image. **(B)** Same plots as **(A)**, for the subjective distance model in V1. Voxels are sorted by normalized prediction accuracy for the subjective distance model. **(C)** Same plots as **(A)**, for the object category model in V1. Voxels are sorted by normalized prediction accuracy for the object category model. **(D–F)** Same plots as **(A–C)**, but for FFA. These tuning patterns are consistent with known response properties of FFA, where voxel responses are related to object categories associated with animate entities.

Yovel, 2006). Thus, the tuning for specific frequencies and orientations is likely to reflect natural correlations between the presence of humans or other animate entities and particular spatial frequency patterns. The β weights for the subjective distance model show that relatively nearby objects elicit BOLD responses above the mean in FFA, while distant objects elicit responses below the mean, and the nearest objects do not affect responses in either direction. This is consistent with at least one study that showed parametrically increasing responses in FFA to scenes with increasingly nearby objects (Park et al., 2015). Finally, the β weights for the object category model show that images containing object categories relating to humans and animals elicited BOLD responses above the mean, while images containing categories related to structural features of scenes (*water*, *land*, *edifice*, etc.) elicit BOLD responses below the mean. These results replicate well-established tuning properties of FFA (Kanwisher et al., 1997; Kanwisher and Yovel, 2006; Huth et al., 2012; Naselaris et al., 2012), and are consistent across subjects in voxels that have sufficient signal to model (See Figure S01 for assessment of signal quality by subject and ROI).

Figure 5 shows the model β weights for all models and all voxels in PPA, RSC, and OPA. Since the β weights in each of the three models show similar tuning in all three areas, we describe the tuning model by model in all three areas.

The β weights for the Fourier power model (**Figures 5A,D,G**) show a somewhat variable pattern across subjects. In general,

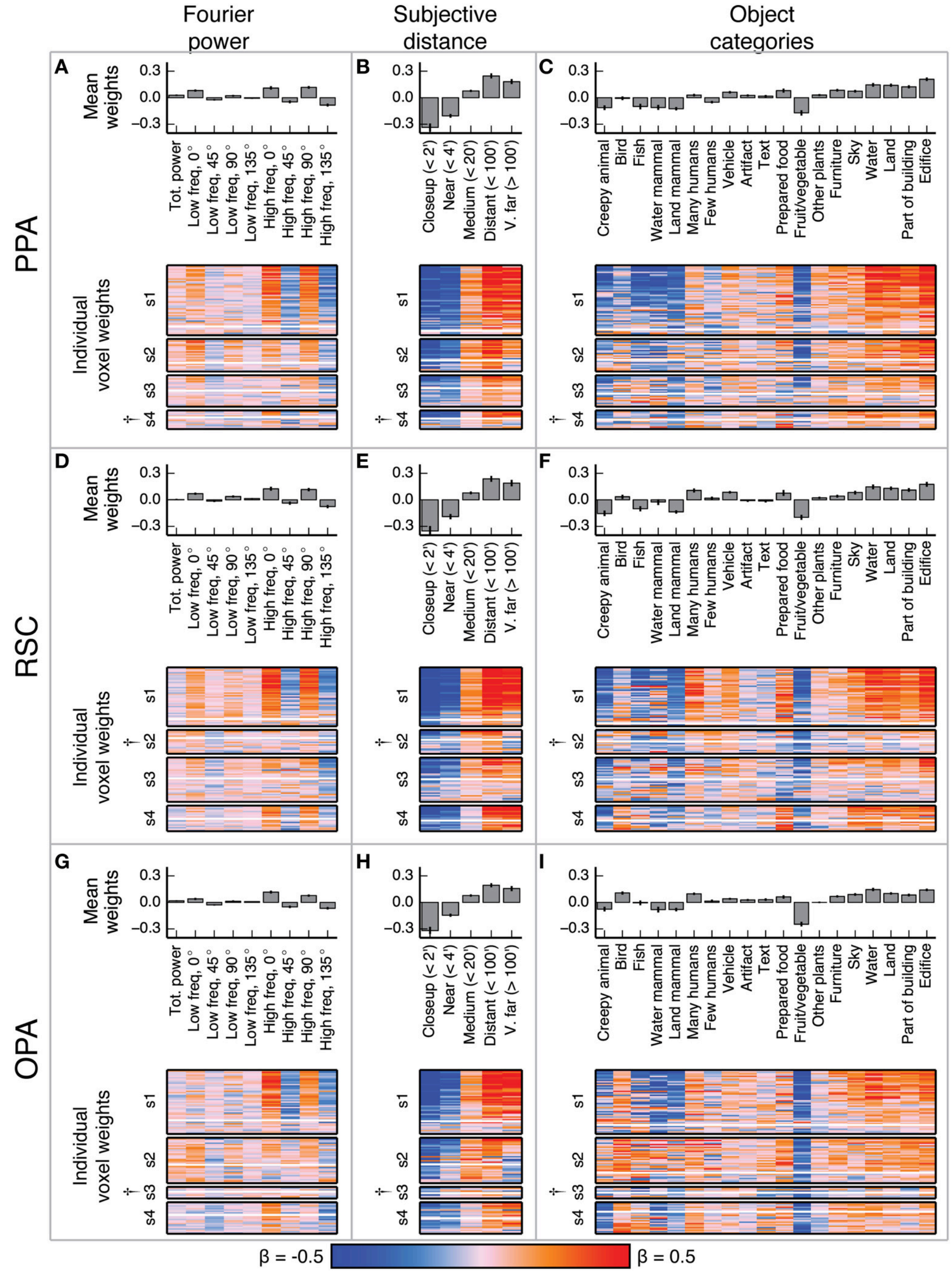


FIGURE 5 | Voxel-wise model β weights for all models for all voxels in PPA, RSC, and OPA. (A–C) Same plots as **Figures 4A–C** but for PPA, with conventions as in **Figure 4**. **(D–F)** Same plots as **Figures 4A–C** but for RSC. **(G–I)** Same plots as **Figures 4A–C** but for OPA. † marks indicate specific ROIs in specific subjects with low signal quality (and thus few voxels selected for analysis). See Figure S01 for evaluation of signal across subjects. For the Fourier power model, the voxel-wise β weights are generally large for high frequency cardinal (vertical and horizontal) orientations, though this varies across subjects. For the subjective distance model, voxel-wise β weights are large for distant objects and small for nearby objects across all subjects. For the object category model, voxel-wise β weights were large for object categories related to the scene structure (e.g., *edifice*, *land*, and *sky*) and small for object categories associated with animate entities (e.g., *few people, land mammal,* and *water mammal*). This pattern of β weights in the object category model was consistent across subjects and ROIs with good signal. All these results are generally consistent with previous reports.

Fourier power at cardinal orientations tends to elicit BOLD responses above the mean in voxels in PPA, RSC, and OPA, while Fourier power at oblique orientations elicits BOLD responses that are small or below the mean. This result is obvious in subject 1, but weaker in the other subjects. In subject 1, the β weights are large for high frequency Fourier power and small for low frequency Fourier power, but this pattern also is weak in the other subjects. We note that subject 1 had substantially better signal (a higher average noise ceiling and more voxels retained) than the other subjects (Figure S01). Thus, the slightly inconsistent tuning across subjects may have been a result of differences in signal quality. The pattern of responses we observe in subject 1 and in the highest-signal voxels in the other subjects are qualitatively consistent with the results of Nasr and Tootell (2012), who found reliably larger responses to cardinal orientations vs. oblique orientations in PPA (Note that in the Nasr and Tootell study, some of the individual voxels within RSC and OPA also showed a cardinal > oblique orientation effect, even though the ROIs as a whole did not).

The β weights for the subjective distance model (**Figures 5B,E,H**) show that images with distant salient objects elicited BOLD responses above the mean in most voxels in PPA, RSC, and OPA. Images that contain nearby salient objects elicit BOLD responses below the mean in these same areas. These results were consistent across subjects. Several other studies have also found increased responses to distant scenes (vs. nearby scenes) in scene-selective areas (Amit et al., 2012; Park et al., 2015).

The β weights for the object category model (**Figures 5C,F,I**) show that images containing buildings or vistas (i.e., images with *edifice*, *water,* and/or *land* labels) elicit BOLD responses above the mean in PPA, RSC, and OPA. Some voxels also respond above the mean to images with *sky* and *furniture* labels. In contrast, images labeled with animate categories (e.g., *land mammal, water mammal,* and *few humans*) elicited BOLD responses below the mean. These results were consistent across subjects. The low weight for the *fruit and vegetable* category is likely due to a bias in stimulus sampling. The stimulus set contained numerous close-up images of fruits and vegetables, such as the top image in **Figure 1A**. The overall pattern of responses in all three areas is consistent with numerous previous studies that have demonstrated increased responses to landscapes, buildings, and other large, inanimate objects in scene-selective areas (Epstein and Kanwisher, 1998; Huth et al., 2012; Naselaris et al., 2012).

To visualize the cortical extent of each of these patterns of tuning independent of ROIs, we computed three different t contrasts between the β weights in each of the models for each voxel in the cortex. We used the β weights from the Fourier power model, the subjective distance model, and the object category model, respectively, to compute contrasts of cardinal vs. oblique, far vs. near, and humans vs. buildings. Each of these contrasts has been emphasized in previous work. Thus, we provide them here for purposes of comparison with other studies that have computed similar maps. However, note that these contrasts are simplifications of the full tuning profile revealed by the weights, particularly for the object category model, which contains many categories besides humans and buildings.

Figures 6A–C show each of these contrasts for one subject, projected onto that subject's cortical surface. Figures S04–S06 show the same maps for the other three subjects. For all three contrasts, many voxels with reliably large ($p < 0.05$, FDR corrected) positive t-values are located in PPA, RSC, and OPA. Relatively few voxels outside scene-selective areas have large positive t-values (Some voxels in the posterior medial parietal lobe also show large t-values in some subjects, particularly for the near vs. far contrast). These contrasts are broadly consistent with contrast maps reported in other studies (Rajimehr et al., 2011; Amit et al., 2012; Nasr and Tootell, 2012; Park et al., 2015). However, as in **Figure 5**, there is variability across subjects in the weights in the Fourier power model. Thus, our replication of tuning for cardinal orientations (as observed by Nasr and Tootell, 2012) is weaker than our replication of tuning for far distances and categories associated with scene structure.

In summary, the voxel-wise models of Fourier power, subjective distance, and object categories reveal three qualitatively different patterns of tuning that are common to all three scene-selective areas: (somewhat) stronger responses to cardinal than to oblique orientations, stronger responses to distant than to nearby objects, and stronger responses to object categories associated with buildings and landscapes than to categories associated with animate objects. However, the tuning revealed by the voxel-wise model β weights does not reveal which of the three models provides the best overall account of the responses in each area. Furthermore, some of the tuning results in V1 and FFA suggest that correlations between features in different models may have affected the estimated tuning for each model (For example, it seems unlikely that V1 truly represents fruits and vegetables, as **Figure 4** seems to indicate). We address both of these issues below.

The Object Category Model Makes the Best Predictions in Scene-selective Areas

To determine which model provides the best description of BOLD responses in each area, we used each fit model to predict responses in a separate validation data set (**Figure 1**). We then computed the correlation between the predictions of each model and the estimated BOLD responses in the validation data. Correlations were normalized by the estimated noise ceiling for each voxel.

Figures 6D–F show estimates of prediction accuracy for all three models for one subject projected onto that subject's cortical surface. Figures S04–S06 show similar maps for the other three subjects. All three models accurately predict brain activity in PPA, RSC, and OPA. The object category model also makes good predictions in the FFA, the Occipital Face Area (OFA), and the Extrastriate Body Area (EBA), as reported previously (Naselaris et al., 2012). This is likely because the object model contains labels for the presence of humans and other animate categories.

Figure 7 shows estimates of prediction accuracy for all three models, averaged across voxels in all four subjects within each of several different ROIs. Figure S07 shows the same result for each individual subject.

In area V1 the Fourier power model provides the best predictions of brain activity (bootstrap $p < 0.05$). This suggests that tuning in the Fourier power model (**Figure 4A**) is more important than tuning in the subjective distance and object category models in V1 (**Figures 4B,C**). In FFA the object category model provides the best predictions (all bootstrap $p < 0.05$). This suggests that tuning in the object category model (**Figure 4F**) is more important than tuning in the Fourier power or subjective

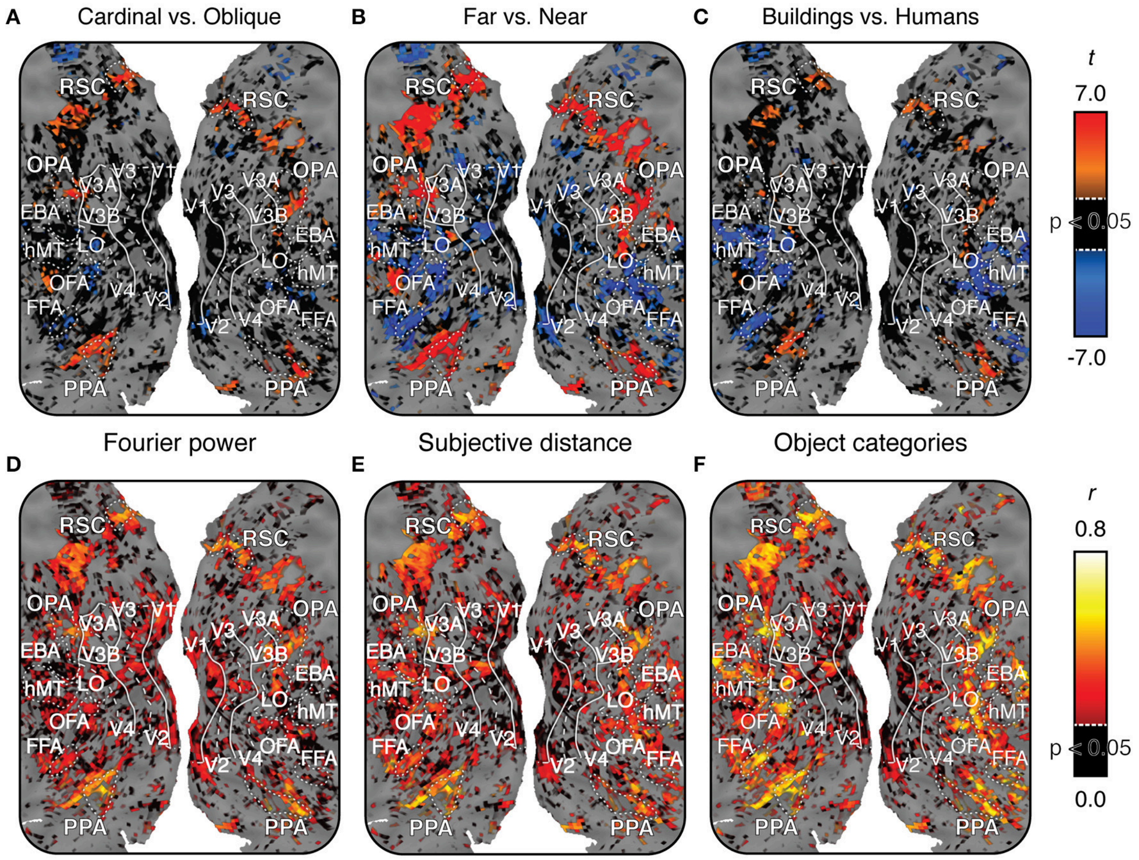


FIGURE 6 | Maps of voxel-wise *t* contrasts and normalized prediction accuracy for subject 1. Figures S04–S06 show the same maps for the other three subjects. For all maps, dashed lines indicate the horizontal meridian in the visual field, solid lines indicate the vertical meridian, and dotted lines indicate the boundaries of regions of interest defined by functional contrasts. **(A)** *t* contrast computed for β weights within the Fourier power model (cardinal vs. oblique). *t*-values are scaled from -7 to 7, black voxels indicate *t*-values below the chance threshold ($t < 3.36$, FDR-corrected $p > 0.05$) despite good signal [$\gamma > 0.04$, $p(\gamma) < 0.01$]. Gray voxels indicate poor signal [$\gamma < 0.04$, $p(\gamma) > 0.01$] and thus no basis for comparing models. **(B)** *t* contrast for β weights within the subjective distance model (far vs. near). **(C)** *t* contrast computed for β weights within the object category model (buildings vs. people). Voxels with significant *t* contrasts for each of the three models are located in the same regions of the cortex. **(D)** Prediction accuracy for the Fourier power model. Prediction accuracy has been normalized by the noise ceiling. Black voxels indicate correlations that are below the chance threshold ($r < 0.21$, FDR-corrected $p > 0.05$) despite good signal [$\gamma > 0.04$, $p(\gamma) < 0.01$]. Gray voxels indicate poor signal [$\gamma < 0.04$, $p(\gamma) > 0.01$], and thus no potential to test predictions. **(E)** Prediction accuracy for the subjective distance model. **(F)** Prediction accuracy for the object category model. All three models make accurate predictions in similar locations across the cortex, though the object category model makes more accurate predictions in FFA, OFA, and EBA. Combined with the *t* contrast maps, this suggests that the three different models may each describe the same response variance in scene-selective areas in a different way.

distance models in FFA (**Figures 4D,E**). Thus, in both V1 and FFA, choosing the best model based on prediction accuracy favors the models that are most consistent with previous results for these areas (Jones and Palmer, 1987; Kanwisher and Yovel, 2006; Kay et al., 2008; Naselaris et al., 2009). These examples demonstrate how assessing prediction accuracy can (and should) affect the interpretation of tuning revealed by β weights.

In PPA, the object category model provides the best predictions of brain activity (all bootstrap $p < 0.05$). This suggests that tuning in the object category model is more important than tuning in the Fourier power or subjective distance models in PPA. In RSC, the object category model provides more accurate predictions than those provided by the Fourier power model (bootstrap $p < 0.05$), but the predictions of the object category model are not significantly different from those of the subjective distance model (bootstrap $p = 0.14$). This suggests that tuning in the object category model is more important than tuning in the Fourier power model, but it is unclear whether the tuning in the subjective distance model or the tuning in the object category model is more important. In OPA, the object category model provides the best predictions of brain activity (all bootstrap $p < 0.05$). Thus, as in PPA, tuning in the object category model is more important than tuning in the Fourier power or subjective distance models in OPA.

Among the options tested here, the representation in two of three scene-selective areas (PPA and OPA) is best described in terms of tuning for object categories. In RSC, tuning for object categories is more important than tuning for Fourier power. Thus, the object category model seems to be a good model for all three areas. However, this conclusion is weakened by variability

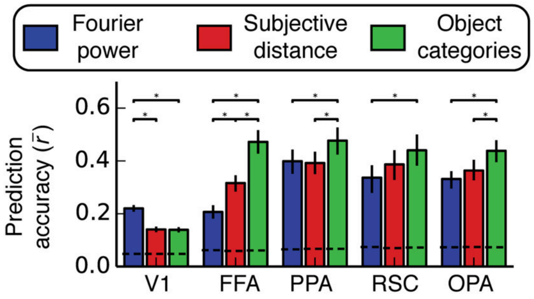


FIGURE 7 | Prediction accuracy (Pearson's *r*) averaged across all voxels and all subjects within several different regions of interest. Predictions are normalized by the noise ceiling and only voxels with reliable stimulus-evoked responses are included. Error bars are 99% confidence intervals, asterisks indicate significant differences between models (bootstrapped $p < 0.05$), and the dotted lines across the bottom indicate the chance threshold (bootstrapped $p = 0.05$) for the mean correlation for each ROI (Thresholds differ slightly across ROIs because of the differing number of voxels in each ROI). The Fourier power model makes the best predictions in V1, and the semantic category model makes the best predictions in all other ROIs (except in RSC, where the subjective distance model and the semantic category model are not reliably distinguishable). We note, however, that the object category model was not reliably better than the Fourier power and subjective distance models in all three scene-selective areas in all subjects (see Figure S07 for individual subject results).

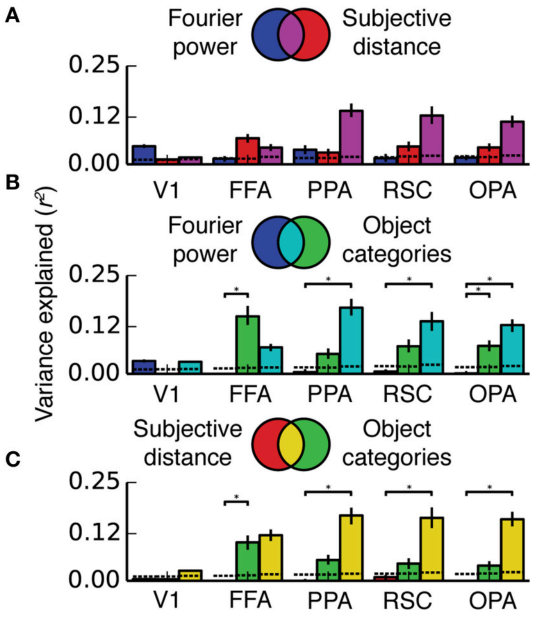


FIGURE 8 | Two-way variance partitioning analyses. All plots are based on concatenated data for all four subjects. **(A)** Independent and shared variance explained by Fourier power and subjective distance models. Dotted lines at the bottom of the graph indicate chance levels (bootstrapped $p = 0.05$) of variance explained, and asterisks indicate significant differences in variance explained (bootstrapped $p < 0.05$). Error bars are 99% confidence intervals across all voxels in a region. **(B)** Independent and shared variance explained by Fourier power and object category models. **(C)** Independent and shared variance explained by subjective distance and object category models. In PPA, RSC, and OPA, all pairs of models share a substantial amount of variance. Compared to the object category model, neither the Fourier power model nor the subjective distance model explains any unique variance.

in relative prediction accuracy across individual subjects (Figure S07). Furthermore, the fact that all three models make quite accurate predictions in all three areas (across all subjects with good signal) suggests that each model may each describe the same underlying representation in different ways.

The Fourier Power, Subjective Distance, and Object Category Models All Explain the Same Response Variance

The Fourier power, subjective distance and object category models all provide accurate predictions of BOLD responses in scene-selective visual areas. Given this result, an obvious question arises: do the Fourier power and subjective distance models explain the same BOLD response variance as is explained by the object category model? That is, can tuning for Fourier power and/or subjective distance almost fully account for category tuning? This question cannot be answered by merely examining prediction accuracy, because two models that make comparably accurate predictions could describe either unique or shared components of response variance (see example in **Figure 3B**). We performed a variance partitioning analysis to determine whether the three models explain unique or shared response variance in the ROIs of interest here. Variance partitioning allocates variance to each model based on whether two models can be combined for a gain in variance explained. If they can, then each model explains unique response variance; if not, the variance explained by the models is shared (see **Figure 3** and Methods for an overview).

Figures 8, **9** show the results of the variance partitioning analysis. In V1, only the Fourier power model explains any unique variance that cannot be explained by the other two models. All three models also share a small amount of variance in V1. The shared variance is likely due to natural correlations between specific features that affect responses in V1 and other features. For example, images with distant objects often have low overall contrast (and thus low Fourier power, as the image of the ocean in **Figure 1A**); thus distance and Fourier power are likely to be correlated (We analyze correlations between all features in detail below). Since total Fourier power affects responses in V1 (**Figure 4A**), this correlation could lead to the subjective distance model and the Fourier power model providing similar predictions (and thus explaining shared variance). Thus, it is likely that the subjective distance and object category models only explain any variance in V1 (**Figure 7**) because of the variance that they share with the Fourier power model.

In FFA, only the object category model explains any unique variance. All three models also share a significant amount of variance, and the subjective distance model and the object category model share a significant amount of variance that is independent of the Fourier power model. The unique variance explained by the object category model is in keeping with known response properties of FFA (Kanwisher and Yovel, 2006; Huth et al., 2012; Naselaris et al., 2012). As in V1, the shared variance between the object category model and the subjective distance model may be due to natural correlations between

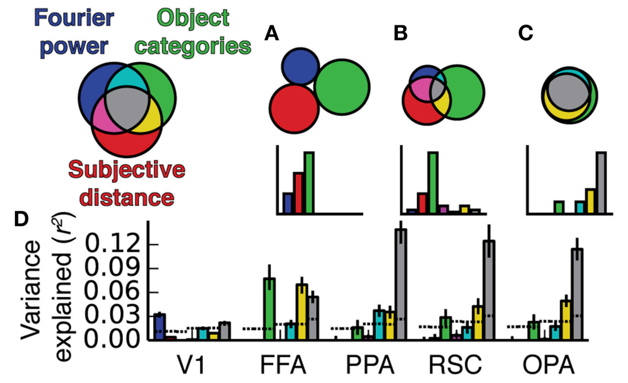


FIGURE 9 | Three-way variance partitioning analysis. (A) Venn diagram representing hypothetical relationships between the variance explained by the three models, such that each explains a unique component of the variance. **(B)** A different hypothetical relationship in which the semantic category model explains a large fraction amount of independent variance, but the three models all share small amounts of variance. **(C)** A third possible relationship in which all three models explain shared variance, and the Fourier power and subjective distance models account for most of the variance explained by the object category model. **(D)** Three-way variance partitioning results obtained in our experiment. This plot is based on concatenated data for all four subjects; results for individual subjects are shown in Figure S08. Dotted lines at the bottom of the graph indicate chance thresholds (bootstrapped $p = 0.05$) for the amount of variance explained. Chance thresholds differ depending on the number of voxels per ROI and the number of subtractions between fit models necessary to compute each component of the variance. Error bars are 99% confidence intervals across all voxels in a region. The pattern of results is most consistent with the Venn diagram in **(C)**.

features. For example, people and other animate categories are more likely to be present at specific distances (in this particular stimulus set, and also potentially in natural visual experience in general). Interestingly, at least one other study has found similar tuning for distance in FFA (Park et al., 2015). However, this study may be subject to the same stimulus feature correlations.

In scene-selective areas PPA, RSC and OPA, most of the variance explained by the Fourier power, subjective distance, and object category models is shared among all three models (**Figure 9**). That is, most of the variance explained by any one of the three models is explained by all three models. Only the object category model explains any unique variance in PPA, RSC, or OPA that cannot be explained by the other two models (**Figure 9**). Thus, the Fourier power and subjective distance models provide partial (but not complete) explanations of variance explained by the object category model in scene-selective areas.

The Fourier power and subjective distance models could be favored on grounds of parsimony, since both models have fewer feature channels than the object category model, and both Fourier power and distance are presumably less complex to compute than abstract category labels. However, neither simpler model provides a more accurate description of BOLD responses in scene-selective areas than that provided by the object category model, and neither model predicts any variance that is not already accounted for by the object category model.

Fourier Power, Subjective Distance, and Object Category Labels are Highly Correlated in Natural Images

The shared variance among the three models in PPA, RSC, and OPA is likely due to correlations between features in the feature spaces underlying the models. To investigate this possibility we computed the correlations between all features in the Fourier power, subjective distance, and object category feature spaces. **Figure 10** shows the resulting correlations. The highest correlations are between the features within the Fourier power feature space. This was expected, since correlations between different spatial frequency bands are a well-known property of natural images (Field, 1987). The average correlation magnitude for features in different feature spaces is $r = 0.11$.

We found reliable relationships between several Fourier power and subjective distance channels. For example, **Figure 10B** shows that horizontal high frequency Fourier power is positively correlated with far distances and negatively correlated with near distances. These correlations may be a result of thin horizontal horizon lines in distant images. Conversely, two low frequency Fourier power channels (*Low freq* 45° and *Low freq* 135°) are positively correlated with near and medium distances and negatively correlated with far distances. Vertical low frequency Fourier power is also positively correlated with intermediate distances and negatively correlated with far distances. The correlations between most low frequency channels and near distances could be a result of perspective projection: nearby objects will fill more of the visual field, and thereby increase low frequency Fourier power. Low frequency horizontal Fourier power may not follow the same trend as other low frequency orientations because the land/sky boundaries will increase both high and low horizontal Fourier power in distant scenes.

To determine whether the relationships between Fourier power and distance that we observe are general to other stimulus sets as well, we computed the same Fourier power feature space for the stimuli used in two previous fMRI studies of distance representation (Kravitz et al., 2011; Park et al., 2015). In both stimulus sets, we found the same relationships between Fourier power and distance as in our stimuli (**Figure 11**; See Figures S09, S10 for further analysis of these two data sets).

We also note that many of the features that elicit large responses in scene-selective areas (**Figure 5**) have relatively high correlations with each other. For example, the category label *sky* is correlated with the subjective distance label *Distant (<100′)* ($r = 0.39$), and horizontal high-frequency Fourier power (Fourier power channel *High freq*, 0°) is correlated with the semantic labels *vehicle*, *sky*, and *water* ($r = 0.19$, 0.27, and 0.27, respectively). Each of these labels is fairly common in the stimulus set (each occurs in at least 230/1326 images—see Figure S03 for frequencies of all object category and distance labels). Thus, the correlations between Fourier power feature channels and the category labels *vehicle*, *sky*, and *water* are reasonably likely to be representative of natural relationships between features in the real world.

Other correlations between less common labels may reflect sampling biases in this particular set of images. For example, the correlation between the nearest distance label [*Closeup (<2′)*]

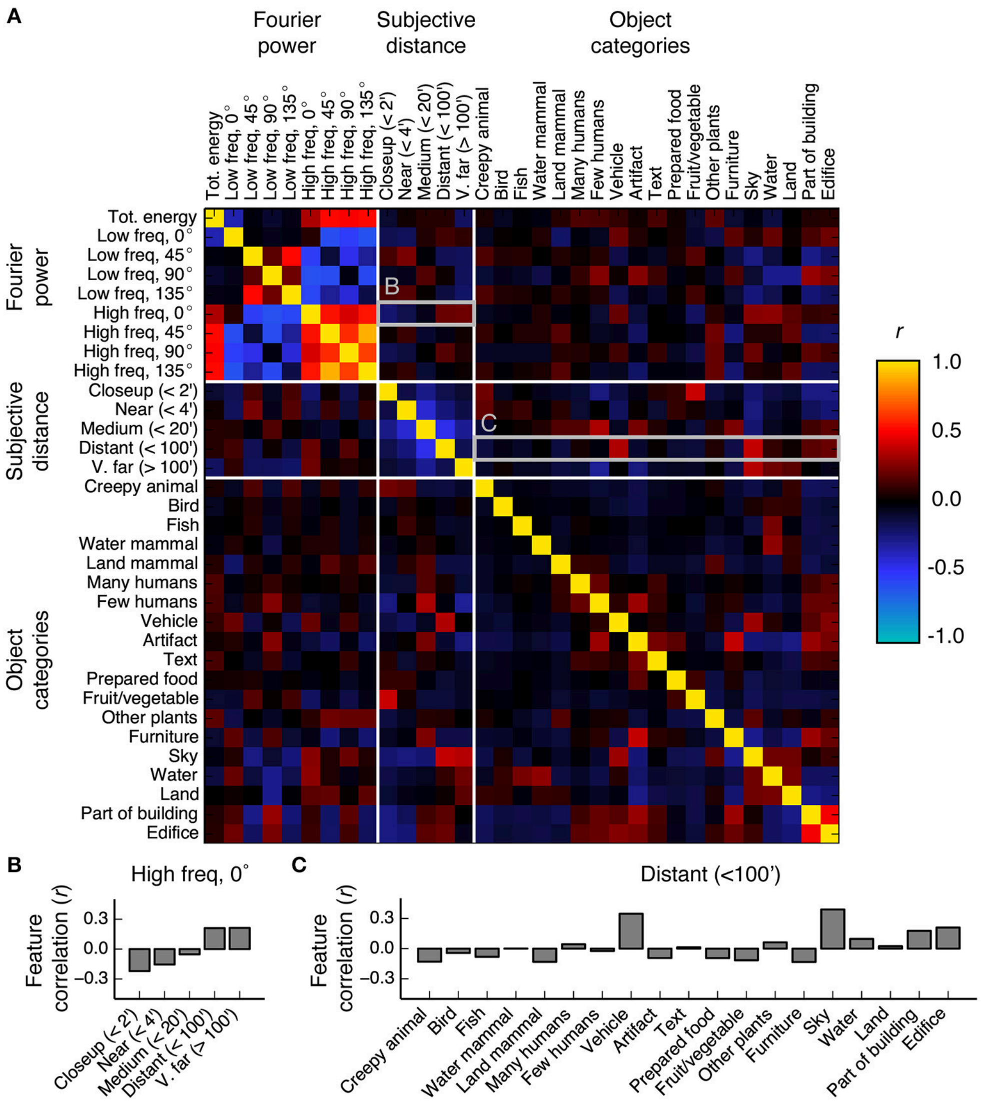


FIGURE 10 | Correlations between all features in the Fourier power, subjective distance, and object category feature spaces. (A) Full correlation matrix. White lines demarcate boundaries between feature spaces. Features that elicit responses above the mean in scene-selective areas [the Fourier power features labeled *High freq*, 0° and *High freq*, 90°; the subjective distance features labeled *Distant (<100′)* and *V. far (>100′)*; and the semantic category labels *Edifice, Part of building, Water, Land,* and *Sky*] tend to have high correlations between them. Panels **(B,C)** provide zoomed in views of the correlation values for the rows marked **(B,C)** in the correlation matrix. **(B)** Bar graph of the correlations between the Fourier power channel *High freq,* 0° and all subjective distance features. High frequency horizontal Fourier power is positively correlated with large subjective distances, potentially due to the presence of a thin horizon line and tiny objects in faraway scenes. **(C)** Bar graph of the correlations between the subjective distance channel *Distant (<100′)* and all object category features. Distant scenes are tend to have the labels *Vehicle*, *Sky*, *Part of building*, and *Edifice*. The high correlations between features with high β weights in scene-selective areas could be a consequence of all three models attempting to parameterize the space of scene features.

and the object label *Fruit/vegetable* is 0.39. *Fruit/vegetable* only occurs in 62 images, of which 32 are rated as *Closeup (<2′)*. The relative rarity of the *Fruit/vegetable* labels, combined with the observation that fruits do not usually appear less than two feet from one's face, suggest that this correlation is potentially spurious.

Whether feature correlations are due to natural statistics or sampling biases, there is a risk that they will lead to biases in estimation of weights, and thereby to models that spuriously share variance. However, it is unclear whether correlations of the magnitude that we observe will necessarily give rise to models that share variance.

A Combination of Correlations between Features and Voxel Tuning Produce Shared Variance

We performed a simulation to illustrate how the feature correlations and voxel-wise β weights in our experiment give rise to models that explain the same variance. We generated

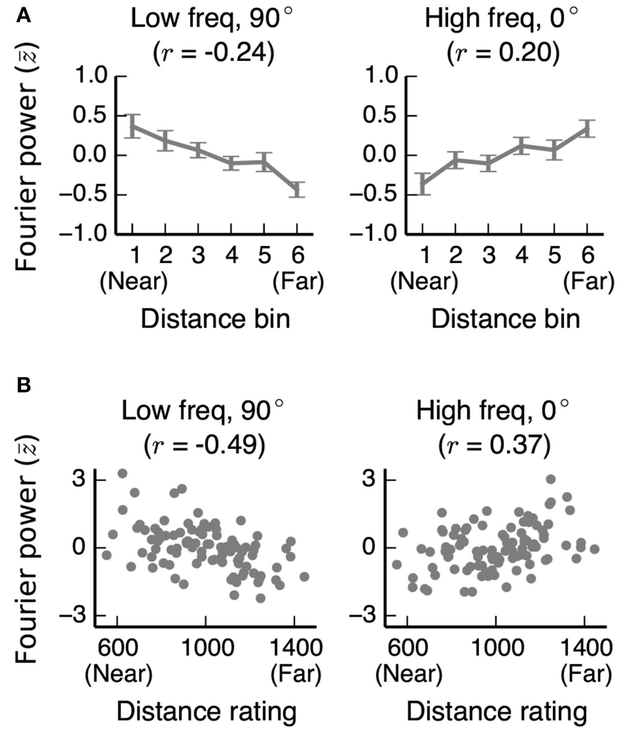


FIGURE 11 | Correlations between distance and two Fourier power channels in stimulus sets from other studies. (A) Mean low frequency vertical and high frequency horizontal Fourier power for each distance bin for images used in Experiment 2 of Park et al. (2015). Fourier power channels were *z*-scored across all images in the stimulus set before averaging across bins. Error bars are standard errors of the mean. **(B)** Low frequency vertical and high frequency horizontal Fourier power for each image in Kravitz et al. (2011), plotted against the behavioral distance ratings for each image obtained in that study. In both stimulus sets, as in our stimulus set, low frequency vertical (90°) Fourier power is reliably associated with nearer scenes, and high frequency horizontal (0°) Fourier power is reliably associated with far-away scenes.

two simulated data sets. The first was based on the stimulus feature spaces and the β weights estimated from the fMRI data for voxels in scene-selective areas, and the other was based on the same feature spaces and a set of semi-random β weights (see Methods for details). The two sets of β weights differed in whether the features that were correlated across feature spaces had relatively high β weights or not (the real weights did, but the random weights generally did not). We applied the same variance partitioning analysis that we previously applied to the fMRI data to both sets of simulated data.

Figure 12 shows the results of the simulation. When semi-random β weights were used to generate the simulated data, the variance partitioning still detected unique variance explained by each model despite the correlations between some of the features in the feature spaces. However, when the real β weights were used to generate the simulated data, the variance partitioning analysis found a large fraction of shared variance between all three models. Thus, the simulation makes it clear that correlated features in different feature spaces only lead to shared variance if the correlated features also have relatively high β weights.

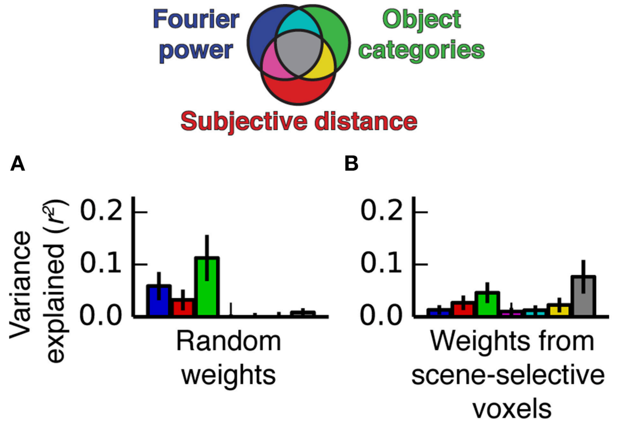


FIGURE 12 | Simulated variance partitioning. (A) Variance partitioning conducted on simulated data generated based on the feature spaces for all three models and a set of semi-random β weights (see Methods for details). This shows that, despite the correlations between feature spaces, there are many patterns of tuning that could result in estimates of unique variance explained for each model. **(B)** Variance partitioning conducted on simulated data generated based on the feature spaces for all three models and actual β weights from voxels in scene-selective areas. This shows that the specific pattern of tuning that we observed (with high weights on the most correlated features) is likely to result in shared variance across these three models.

The β weights, which reflect the specific response properties of PPA, RSC, and OPA, can selectively magnify correlations between particular correlated features when predictions are computed, which can lead to shared variance between the different models.

This suggests that new models of scene-selective areas are more likely to explain unique variance to the extent that the features they parameterize are *not* correlated with other features known to be associated with responses in scene-selective areas.

DISCUSSION

Several areas in the human brain respond to visual scenes, but which specific scene-related features are represented in these areas remains unclear. We investigated three hypotheses that have been proposed to account for responses in scene-selective areas such as PPA, RSC, and OPA. Specifically, we investigated whether these areas represent (1) information about the Fourier power of scenes, (2) the subjective distance to salient objects in scenes, or (3) semantic categories of scenes and their constituent objects. We evaluated these three hypotheses by applying voxel-wise modeling to a data set consisting of BOLD fMRI responses elicited by a large set of natural images. We created and compared the prediction performance of three voxel-wise encoding models, one reflecting each of these alternative hypotheses.

We found that a voxel-wise model based on semantic categories makes slightly more accurate predictions than a model based on Fourier power (in PPA, RSC, and OPA) or subjective distance (in PPA and OPA). However, a variance partitioning analysis revealed that, in all three areas, the variance predicted by these three models is mostly shared. The shared variance is likely a result of a combination of the response patterns of voxels in scene-selective areas and high natural correlations between the stimulus features in the feature spaces underlying each of the models. We therefore conclude that any or all of these models can provide a plausible account of visual representation in PPA, RSC, and OPA.

Previous Studies Have Not Resolved which Model Best describes Scene-selective Areas

Several previous studies of PPA, RSC, and/or OPA have argued in favor of each of the hypotheses tested here, or in favor of closely related hypotheses (Walther et al., 2009; Kravitz et al., 2011; Park et al., 2011, 2015; Rajimehr et al., 2011; Nasr and Tootell, 2012; Watson et al., 2014). However, none have completely resolved which features are most likely to be represented in scene-selective areas. We briefly review three representative and well-designed studies of scene-selective areas here, and assess their conclusions in light of our results.

Nasr and Tootell argued that PPA represents Fourier power (Nasr and Tootell, 2012). Specifically, they showed that filtered natural images with Fourier power at cardinal orientations elicit larger responses in PPA than do filtered images with Fourier power at oblique orientations. In two control experiments, they measured fMRI responses to stimuli consisting of only simple shapes, and found the same pattern of responses. Thus, their results suggest that Fourier power at cardinal orientations influences responses in PPA independent of subjective distance or semantic categories. This in turn suggests that the Fourier power model in our experiment should predict some unique response variance that is independent of the subjective distance and semantic category models. We did find that the Fourier power model gave accurate predictions in scene-selective areas. However, we did not find any unique variance explained by the Fourier power model. There are at least two possible explanations for this discrepancy. First, the Fourier power model may explain some unique variance, but we may have mischaracterized it as shared variance because of stimulus correlations. Second, the results of Nasr and Tootell's study, which relied on filtered and artificial stimuli, simply may not generalize to explain responses to natural images. This is a known pitfall of using artificial or manipulated stimuli (Talebi and Baker, 2012). In any case, the data from the Nasr and Tootell study provide no information about the strength of the relationship between Fourier power and BOLD responses in scene-selective areas relative to the effects of other features. Thus, their study cannot resolve the question of which model is best, nor the question of how Fourier power features are related to other features.

Park et al. (2015) argued that PPA and RSC represent scene size. Their metric for scene size was based on human judgments, and so is closely related to the subjective distance model that we tested here. They measured BOLD responses to a large and carefully chosen set of photographs of natural scenes, and found that responses in PPA and RSC increased parametrically with scene size. However, we found a strong relationship between scene size and Fourier power in the images used in the Park et al. study (**Figure 11A**, Figure S10). To try to avoid just such confounds, Park and colleagues created a control stimulus set in which high-frequency Fourier power was approximately equalized across different scene sizes. We did not test this control stimulus set directly, but since the differences in Fourier power that we observed were specific to particular orientations, it is unlikely that their control removed all Fourier power differences between scenes. This suggests that differences in particular Fourier power channels between different scene sizes might account for the results reported in the Park study, just as both the Fourier power and subjective distance models provide equivalent descriptions of scene-selective regions in our data. Finally, Park and colleagues did not assess whether the specific semantic categories of objects in each of their scenes might have affected BOLD responses. Without this comparison, it is unclear whether the presence of different object categories in their scenes may have also affected their results. For all these reasons, the results reported by Park and colleagues cannot provide a basis for choosing between the three models of scene-selective areas that we consider.

Kravitz et al. (2011) argue that PPA and OPA represent scene expanse (defined as the difference between open and closed scenes) and relative distance (defined as the difference between near and far scenes). They find that voxel patterns in PPA and OPA distinguish both open scenes from closed scenes and near scenes from far scenes better than the same voxels distinguish natural from manmade scenes. However, variation in Fourier power across their experimental conditions complicates the interpretation of their results. They acknowledge that the open and closed scenes in their stimulus set have visibly different Fourier power spectra. When we processed their stimuli with our Fourier power model, we found significant differences between their open and closed scenes in several Fourier power channels (Figure S09A). This suggests that the different patterns of responses they observed to open and closed scenes could be equally well explained by differences in Fourier power between open and closed scenes. Kravitz et al. do not report any differences between the Fourier spectra of the near and far scenes in their stimulus set. Our analysis of their stimuli also does not find any reliable difference in any Fourier power channel between their near and far scenes (Figure S09B). However, their Near and Far condition labels were based on relative distance within each scene category, which means that the scenes in the Near condition were not necessarily all the subjectively nearest scenes. For example, half their images of beaches were labeled as Near and half their images of hallways were labeled as Far, regardless of whether the beaches were subjectively nearer than the hallways. They did, however, obtain a measure of the relative subjective distance of each scene. When we compared the Fourier power features for each image to these distance ratings (instead of to the near/far condition labels), we found reliable correlations between Fourier power and relative subjective distance in their stimulus set (**Figure 11B** and Figure S09C), just as in our stimuli and in the Park et al. (2015) stimuli. Thus, the correlation between subjective distance ratings and fMRI-based distance scores reported in Kravitz et al. (2011) might be explained by variation in Fourier power—specifically, by the presence of high frequency horizontal Fourier power in distant scenes. In sum, our reanalysis of the stimuli from Kravitz et al. (2011) suggest that their results cannot provide a basis for choosing between the three models of scene-selective areas that we consider here.

Other Hypotheses Regarding Scene-selective Areas

The Fourier power, subjective distance, and object category feature spaces that we investigated broadly sample the space of hypotheses regarding the representation in scene-selective areas. However, three specific feature spaces obviously do not constitute a comprehensive test of every hypothesis in the literature.

Several other feature spaces have been proposed that parameterize variation in the same three broad domains that our models do (low-level image features, 3D spatial layout, and categorical or semantic information), but with different parameters. For example, low-level image variation can be parameterized using Gabor wavelets (Jones and Palmer, 1987; Kay et al., 2008), scene gist (Oliva and Torralba, 2001; Watson et al., 2014), or extended contours (Walther et al., 2011). 3D spatial variation can be parameterized according to scene expanse (Kravitz et al., 2011; Park et al., 2011) or local scene structure (Epstein and Kanwisher, 1998; Kornblith et al., 2013). And categorical information about scenes can be parameterized using hierarchical object labels (Huth et al., 2012) or labels for categories of scenes rather than objects, including distinctions between natural and man-made scenes (Naselaris et al., 2009; Walther et al., 2009; Stansbury et al., 2013).

Previous studies have also proposed that scene-selective areas may represent scene familiarity (Epstein et al., 2007), landmarks (Janzen and van Turennout, 2004; Auger et al., 2012), or other scene features relevant for navigation (Epstein, 2008; Morgan et al., 2011). None of these hypotheses are obviously related to the feature spaces we investigated.

Any of these feature spaces, if they were formalized and tested in the voxel-wise modeling framework, could potentially yield better or more unique models of BOLD responses than those we tested. However, all these other feature spaces—particularly those in the same broad categories of hypotheses as our models—may be strongly related to each other in the same way that the feature spaces we tested are. Our work provides a blueprint for how to address the correlations between feature spaces in a quantitative and principled way, and to assess which models explain unique or shared variance.

Suggestions for Further Studies on Representation in Scene-selective Areas

Our study suggests that the data available currently are not sufficient to discriminate between the alternative hypotheses that scene-selective areas represent information about Fourier power, subjective distance, or object categories. It could be the case that scene-selective areas represent all of these distinct feature classes. Alternatively, it could be the case that scene-selective areas represent only one of these three distinct classes of features, but that the presence of stimulus correlations in our study and missing controls and analyses in previous studies have precluded identification of the most appropriate feature space. Is there any way to resolve this issue?

The only way forward is to test the same models (and/or related models) on different stimulus sets, and to search for stimuli for which some models fail to make accurate predictions of brain responses and other models succeed. However, new stimuli must be chosen carefully to reduce the correlations between stimulus features in different alternative models. Simply removing problematic features (e.g., by Fourier bandpass filtering the stimuli) is not a good solution because the visual system is highly nonlinear (Carandini et al., 2005; Wu et al., 2006). Spatial frequencies that are filtered out of a stimulus may be reintroduced within the visual system by nonlinear processes operating at any level. An analogous process occurs in the missing fundamental phenomenon, which is well known in audition (Wightman, 1973a,b).

Restricting feature variation in experimental stimuli to avoid correlations between features is also not a good solution. This approach might produce satisfying results within the range of stimuli tested in an experiment, but the resulting model will be unlikely to generalize to the larger range of stimuli encountered in the natural world (Talebi and Baker, 2012). This is a lesson that has been well learned in the visual neurophysiology community over the past 20 years: if models are developed using filtered, constrained or highly artificial stimuli, they tend to perform poorly when tested on natural images (David et al., 2004; Talebi and Baker, 2012).

We suggest that one useful way forward would be to create natural stimulus sets that reduce the covariance of stimulus features while maintaining a natural range of variance in as many features as possible. It might be possible to generate stimuli that satisfy these constraints parametrically. Alternatively, it might be possible to develop an appropriate stimulus set by sampling images from an extremely large online database such as ImageNet (http://www.image-net.org/) or the Flickr image database (https://www.flickr.com/creativecommons/). A stimulus set that is designed specifically to minimize covariance between features while maintaining natural variability will reduce the amount of shared variance between models, and lead to clearer conclusions as to which model is best for each area.

Our suggestion that new stimulus sets should be developed is not completely novel. The imperative to include a reasonable amount of natural variation in a stimulus set seems to be an implicit guiding principle in many studies (e.g., Kravitz et al., 2011; Park et al., 2015). However, such implicit guiding principles are imprecise and likely to vary across experiments. Thus, we suggest that more effort should be devoted to defining stimulus features quantitatively rather than operationally. Quantitative definitions of features improve the ability to measure and control feature coverage and feature covariance. One substantial advantage of the voxel-wise modeling approach used here is that it provides a very clear and quantitative picture of what is known and what is not known. Stimulus properties can be quantified and modeled directly. Correlations between features within models and across models can also be quantified and assessed. This approach provides an unambiguous view of where the field is today, and it leads to clear recommendations for future studies.

ACKNOWLEDGMENTS

This work was supported by grants to JG from the National Eye Institute (EY019684) and the National Science Foundation Center for the Science of Information (CCF- 0939370), as well as NIH NEI F32EY021710 to ML. We thank Thomas Naselaris, Shinji Nishimoto, Tolga Cukur, Michael Oliver, Alex Huth, James Gao, Natalia Bilenko, Anwar Nuñez, and Storm Slivkoff for helpful discussions and comments, Talia Konkle for making the stimuli from Park, Konkle, and Oliva available on her web site, Chris Baker for sharing his stimuli and behavioral data (as well as for a constructive and helpful review), and doubly thank Thomas Naselaris for collecting half of the data.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fncom.2015.00135

REFERENCES

Aguirre, G. K., Zarahn, E., and D'Esposito, M. (1998). An area within human ventral cortex sensitive to "building" stimuli: evidence and implications. *Neuron* 21, 373–383.

Albrecht, D. G., and Hamilton, D. B. (1982). Striate cortex of monkey and cat: contrast response function. *J. Neurophysiol.* 48, 217–237.

Amit, E., Mehoudar, E., Trope, Y., and Yovel, G. (2012). Do object-category selective regions in the ventral visual stream represent perceived distance information? *Brain Cogn.* 80, 201–213. doi: 10.1016/j.bandc.2012.06.006

Auger, S. D., Mullally, S. L., and Maguire, E. A. (2012). Retrosplenial cortex codes for permanent landmarks. *PLoS ONE* 7:e43620. doi: 10.1371/journal.pone.0043620

Benjamini, Y., and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.* 29, 1165–1188. Available online at: http://www.jstor.org/stable/2674075

Carandini, M., Demb, J. B., Mante, V., Tolhurst, D. J., Dan, Y., Olshausen, B. A., et al. (2005). Do we know what the early visual system does? *J. Neurosci.* 25, 10577–10597. doi: 10.1523/JNEUROSCI.3726-05.2005

Cate, A. D., Goodale, M. A., and Köhler, S. (2011). The role of apparent size in building- and object-specific regions of ventral visual cortex. *Brain Res.* 1388, 109–122. doi: 10.1016/j.brainres.2011.02.022

Dale, A. M., Fischl, B., and Sereno, M. I. (1999). Cortical surface-based analysis. I. Segmentation and surface reconstruction. *Neuroimage* 9, 179–194. doi: 10.1006/nimg.1998.0395

David, S. V., and Gallant, J. L. (2005). Predicting neuronal responses during natural vision. *Network* 16, 239–260. doi: 10.1080/09548980500464030

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. doi: 10.1523/JNEUROSCI.1422-04.2004

Dilks, D. D., Julian, J. B., Paunov, A. M., and Kanwisher, N. (2013). The occipital place area is causally and selectively involved in scene perception. *J. Neurosci.* 33, 1331–1336a. doi: 10.1523/JNEUROSCI.4081-12.2013

Epstein, R. A. (2008). Parahippocampal and retrosplenial contributions to human spatial navigation. *Trends Cogn. Sci.* 12, 388–396. doi: 10.1016/j.tics.2008.07.004

Epstein, R., and Kanwisher, N. (1998). A cortical representation of the local visual environment. *Nature* 392, 598–601. doi: 10.1038/33402

Epstein, R. A., Parker, W. E., and Feiler, A. M. (2007). Where am I now? Distinct roles for parahippocampal and retrosplenial cortices in place recognition. *J. Neurosci.* 27, 6141–6149. doi: 10.1523/JNEUROSCI.0799-07.2007

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. doi: 10.1364/JOSAA.4.002379

Fisher, R. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika* 10, 507–521. doi: 10.2307/2331838

Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J.-P., Frith, C. D., and Frackowiak, R. S. J. (1994). Statistical parametric maps in functional imaging: a general linear approach. *Hum. Brain Mapp.* 2, 189–210. doi: 10.1002/hbm.460020402

Gallant, J. L., Nishimoto, S., Naselaris, T., and Wu, M. C.-K. (2012). "System identification, encoding models, and decoding models: a powerful new approach to fMRI research," in *Visual Population Codes: Toward a Common Multivariate Framework for Cell Recording and Functional Imaging,* eds N. Kriegeskorte and G. Kreiman (Cambridge, MA; London, UK: MIT Press), 163–188.

Gao, J. S., Huth, A. G., Lescroart, M. D., and Gallant, J. L. (2015). Pycortex: an interactive surface visualizer for fMRI. *Front. Neuroinform.* 9:23. doi: 10.3389/fninf.2015.00023

Gardner, J. L., Sun, P., Waggoner, R. A., Ueno, K., Tanaka, K., and Cheng, K. (2005). Contrast adaptation and representation in human early visual cortex. *Neuron* 47, 607–620. doi: 10.1016/j.neuron.2005.07.016

Hansen, K. A., Kay, K. N., and Gallant, J. L. (2007). Topographic organization in and near human visual area V4. *J. Neurosci.* 27, 11896–11911. doi: 10.1523/JNEUROSCI.2991-07.2007

Hsu, A., Borst, A., and Theunissen, F. E. (2004). Quantifying variability in neural responses and its application for the validation of model predictions. *Network* 15, 91–109. doi: 10.1088/0954-898X_15_2_002

Huth, A. G., Nishimoto, S., Vu, A. T., and Gallant, J. L. (2012). A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron* 76, 1210–1224. doi: 10.1016/j.neuron.2012.10.014

Janzen, G., and van Turennout, M. (2004). Selective neural representation of objects relevant for navigation. *Nat. Neurosci.* 7, 673–677. doi: 10.1038/nn1257

Jones, J. P., and Palmer, L. A. (1987). An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. *J. Neurophysiol.* 58, 1233–1258.

Kanwisher, N., McDermott, J., and Chun, M. M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. *J. Neurosci.* 17, 4302–4311.

Kanwisher, N., and Yovel, G. (2006). The fusiform face area: a cortical region specialized for the perception of faces. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 361, 2109–2128. doi: 10.1098/rstb.2006.1934

Kay, K. N., Naselaris, T., Prenger, R. J., and Gallant, J. L. (2008). Identifying natural images from human brain activity. *Nature* 452, 352–355. doi: 10.1038/nature06713

Konkle, T., and Oliva, A. (2012). A real-world size organization of object responses in occipitotemporal cortex. *Neuron* 74, 1114–1124. doi: 10.1016/j.neuron.2012.04.036

Kornblith, S., Cheng, X., Ohayon, S., and Tsao, D. Y. (2013). A network for scene processing in the macaque temporal lobe. *Neuron* 79, 766–781. doi: 10.1016/j.neuron.2013.06.015

Kravitz, D. J., Peng, C. S., and Baker, C. I. (2011). Real-world scene representations in high-level visual cortex: it's the spaces more than the places. *J. Neurosci.* 31, 7322–7333. doi: 10.1523/JNEUROSCI.4588-10.2011

Maguire, E. A. (2001). The retrosplenial contribution to human navigation: a review of lesion and neuroimaging findings. *Scand. J. Psychol.* 42, 225–238. doi: 10.1111/1467-9450.00233

Miller, G. A. (1995). WordNet: a lexical database for English. *Commun. ACM* 38, 39–41. doi: 10.1145/219717.219748

Morgan, L. K., Macevoy, S. P., Aguirre, G. K., and Epstein, R. A. (2011). Distances between real-world locations are represented in the human hippocampus. *J. Neurosci.* 31, 1238–1245. doi: 10.1523/JNEUROSCI.4667-10.2011

Naselaris, T., Kay, K. N., Nishimoto, S., and Gallant, J. L. (2011). Encoding and decoding in fMRI. *Neuroimage* 56, 400–410. doi: 10.1016/j.neuroimage.2010.07.073

Naselaris, T., Prenger, R. J., Kay, K. N., Oliver, M., and Gallant, J. L. (2009). Bayesian reconstruction of natural images from human brain activity. *Neuron* 63, 902–915. doi: 10.1016/j.neuron.2009.09.006

Naselaris, T., Stansbury, D. E., and Gallant, J. L. (2012). Cortical representation of animate and inanimate objects in complex natural scenes. *J. Physiol. Paris* 106, 239–249. doi: 10.1016/j.jphysparis.2012.02.001

Nasr, S., Echavarria, C. E., and Tootell, R. B. H. (2014). Thinking outside the box: rectilinear shapes selectively activate scene-selective cortex. *J. Neurosci.* 34, 6721–6735. doi: 10.1523/JNEUROSCI.4802-13.2014

Nasr, S., Liu, N., Devaney, K. J., Yue, X., Rajimehr, R., Ungerleider, L. G., et al. (2011). Scene-selective cortical regions in human and nonhuman primates. *J. Neurosci.* 31, 13771–13785. doi: 10.1523/JNEUROSCI.2792-11.2011

Nasr, S., and Tootell, R. B. H. (2012). A cardinal orientation bias in scene-selective visual cortex. *J. Neurosci.* 32, 14921–14926. doi: 10.1523/JNEUROSCI.2036-12.2012

Nishimoto, S., Vu, A. T., Naselaris, T., Benjamini, Y., Yu, B., and Gallant, J. L. (2011). Reconstructing visual experiences from brain activity evoked by natural movies. *Curr. Biol.* 21, 1641–1646. doi: 10.1016/j.cub.2011.08.031

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. doi: 10.1023/A:1011139631724

Park, S., Brady, T. F., Greene, M. R., and Oliva, A. (2011). Disentangling scene content from spatial boundary: complementary roles for the parahippocampal place area and lateral occipital complex in representing real-world scenes. *J. Neurosci.* 31, 1333–1340. doi: 10.1523/JNEUROSCI.3885-10.2011

Park, S., Konkle, T., and Oliva, A. (2015). Parametric coding of the size and clutter of natural scenes in the human brain. *Cereb. Cortex* 25, 1792–1805. doi: 10.1093/cercor/bht418

Rajimehr, R., Devaney, K. J., Bilenko, N. Y., Young, J. C., and Tootell, R. B. H. (2011). The "parahippocampal place area" responds preferentially to high spatial frequencies in humans and monkeys. *PLoS Biol.* 9:e1000608. doi: 10.1371/journal.pbio.1000608

Simoncelli, E. P., and Olshausen, B. A. (2001). Natural image statistics and neural representation. *Annu. Rev. Neurosci.* 24, 1193–1216. doi: 10.1146/annurev.neuro.24.1.1193

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. doi: 10.1002/hbm.20169

Stansbury, D. E., Naselaris, T., and Gallant, J. L. (2013). Natural scene statistics account for the representation of scene categories in human visual cortex. *Neuron* 79, 1025–1034. doi: 10.1016/j.neuron.2013.06.034

Talebi, V., and Baker, C. L. (2012). Natural versus synthetic stimuli for estimating receptive field models: a comparison of predictive robustness. *J. Neurosci.* 32, 1560–1576. doi: 10.1523/JNEUROSCI.4661-12.2012

Van Essen, D. C., Anderson, C. H., and Felleman, D. J. (1992). Information processing in the primate visual system: an integrated systems perspective. *Science* 255, 419–423. doi: 10.1126/science.1734518

Walther, D. B., Caddigan, E., Fei-Fei, L., and Beck, D. M. (2009). Natural scene categories revealed in distributed patterns of activity in the human brain. *J. Neurosci.* 29, 10573–10581. doi: 10.1523/JNEUROSCI.0559-09.2009

Walther, D. B., Chai, B., Caddigan, E., Beck, D. M., and Fei-Fei, L. (2011). Simple line drawings suffice for functional MRI decoding of natural scene categories. *Proc. Natl. Acad. Sci. U.S.A.* 108, 9661–9666. doi: 10.1073/pnas.1015666108

Watson, D. M., Hartley, T., and Andrews, T. J. (2014). Patterns of response to visual scenes are linked to the low-level properties of the image. *Neuroimage* 99, 402–410. doi: 10.1016/j.neuroimage.2014.05.045

Wightman, F. L. (1973a). Pitch and stimulus fine structure. *J. Acoust. Soc. Am.* 54, 397–406. doi: 10.1121/1.1913591

Wightman, F. L. (1973b). The pattern-transformation model of pitch. *J. Acoust. Soc. Am.* 54, 407–416. doi: 10.1121/1.1913592

Wu, M. C.-K., David, S. V., and Gallant, J. L. (2006). Complete functional characterization of sensory neurons by system identification. *Annu. Rev. Neurosci.* 29, 477–505. doi: 10.1146/annurev.neuro.29.051605.113024

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.

frontiers in
COMPUTATIONAL NEUROSCIENCE

ORIGINAL RESEARCH ARTICLE
published: 26 March 2014
doi: 10.3389/fncom.2014.00034

Optimal attentional modulation of a neural population

Ali Borji[1]* and Laurent Itti[1,2,3]

[1] Department of Computer Science, University of Southern California, Los Angeles, CA, USA
[2] Neuroscience Graduate Program, University of Southern California, Los Angeles, CA, USA
[3] Department of Psychology, University of Southern California, Los Angeles, CA, USA

Edited by:
Hans P. Op De Beeck, University of Leuven, Belgium

Reviewed by:
Sam Ling, Boston University, USA
Janneke Jehee, Donders Center for Cognitive Neuroimaging, Netherlands

***Correspondence:**
Ali Borji, University of Southern California - Hedco Neuroscience Building - 3641 Watt Way, Los Angeles, CA 90089-2520, USA
e-mail: borji@usc.edu

Top-down attention has often been separately studied in the contexts of either optimal population coding or biasing of visual search. Yet, both are intimately linked, as they entail optimally modulating sensory variables in neural populations according to top-down goals. Designing experiments to probe top-down attentional modulation is difficult because non-linear population dynamics are hard to predict in the absence of a concise theoretical framework. Here, we describe a unified framework that encompasses both contexts. Our work sheds light onto the ongoing debate on whether attention modulates neural response gain, tuning width, and/or preferred feature. We evaluate the framework by conducting simulations for two tasks: (1) classification (discrimination) of two stimuli s_a and s_b and (2) searching for a target T among distractors D. Results demonstrate that all of gain, tuning, and preferred feature modulation happen to different extents, depending on stimulus conditions and task demands. The theoretical analysis shows that task difficulty (linked to difference Δ between s_a and s_b, or T, and D) is a crucial factor in optimal modulation, with different effects in discrimination vs. search. Further, our framework allows us to quantify the relative utility of neural parameters. In easy tasks (when Δ is large compared to the density of the neural population), modulating gains and preferred features is sufficient to yield nearly optimal performance; however, in difficult tasks (smaller Δ), modulating tuning width becomes necessary to improve performance. This suggests that the conflicting reports from different experimental studies may be due to differences in tasks and in their difficulties. We further propose future electrophysiology experiments to observe different types of attentional modulation in a same neuron.

Keywords: top-down attention, neural modulation, neural coding, gain, tuning width, feature selectivity

1. INTRODUCTION

Optimal neural coding, or efficient coding, suggests that sensory systems have evolved to optimize the representation of the world around us. Two seemingly different fields of study, neural coding and visual search, have addressed neural modulation. The former has mainly investigated the optimal tuning width for a population of neurons (often one value for all neurons) in stimulus reconstruction and discrimination tasks (e.g., Zhang and Sejnowski, 1999; Jazayeri and Movshon, 2006; Berens et al., 2011; Wang et al., 2012). For example the question of whether sharpening or broadening a neuron's tuning might improve performance has attracted significant interest (e.g., Pouget et al., 1999; Zhang and Sejnowski, 1999). Computational studies of top-down biasing of visual search, on the other hand, have primarily addressed optimal gain modulation (e.g., Navalpakkam and Itti, 2007; Scolari and Serences, 2009, 2010; Scolari et al., 2012). Optimal neural modulation, in general, is a complex optimization problem since several variables such as statistics of stimuli, task variability, limitations of neural systems (e.g., number of neurons and parameters, metabolic cost, noise), and coupled nonlinear dynamics are involved. Here, we present a reconciled and abstract account of optimal neural modulation by solving for the best set of gain, tuning width and preferred feature of individual neurons to maximize classification and visual search performance. We use terms *attention* and *optimal neural modulation* interchangeably since the term "attention," as currently used in the literature, refers to a highly heterogeneous class of phenomena. Characteristics of these phenomena vary significantly depending on the specific context in which the nervous system is operating (e.g., different time scales, tasks, environments, etc.).

1.1. OVERVIEW OF ATTENTIONAL MODULATION

Finding a friend amidst several hundred passengers at an airport can be a nightmare. Yet, our brain handles the explosion of information efficiently by filtering out irrelevant or distracting stimuli, and by drawing our gaze to salient and relevant visual stimuli, through a process known as visual attention (Treisman and Gelade, 1980; Tsotsos, 1992; Desimone and Duncan, 1995; James, 2011). Specifically, visual attention is believed to help in at least two ways: *goal-driven top-down attention* (Yarbus, 1967; Corbetta and Shulman, 2002; Borji and Itti, 2014) might help in focusing on relevant image regions that resemble our friend's appearance, thereby accelerating our search, and *stimulus-driven bottom-up attention* (Koch and Ullman, 1985) might alert us to salient image regions like moving cars, pedestrians or dollies in our way, thereby avoiding accidents (Itti and Koch, 2001). Together, top-down and bottom-up attention help us select a few relevant and salient image regions for further processing, including recognition, representation, awareness and action (Desimone and Duncan, 1995; Crick and Koch, 1998). Please see Itti and Koch (2001), Hayhoe and Ballard (2005), Macknik et al. (2008), Eckstein et al. (2009), Baluch and Itti (2011), Carrasco (2011), Eckstein (2011), Kowler (2011), Nakayama and Martini (2011), Schütz et al. (2011), Tatler et al. (2011), and Borji and Itti (2013) for recent reviews of attentional mechanisms at behavioral, computational, and neural levels.

There exists at least three types of attention – *spatial* (Posner et al., 1980; Moran and Desimone, 1985; Kastner et al., 1999; Womelsdorf et al., 2006; Talsma et al., 2007), *feature-based* (Treue and Trujillo, 1999; Saenz et al., 2003; Sohn et al., 2005; Maunsell and Treue, 2006; Serences and Boynton, 2007; Jehee et al., 2011) and *object-based attention* (Duncan, 1984, 1996; Roelfsema et al., 1998; Kanwisher and Wojciulik, 2000; Reynolds et al., 2003; Chen, 2012; Cohen and Tong, 2013), depending on whether the basic unit of attentional deployment is a spatial location/region (e.g., the attentional "spotlight" Treisman and Gelade, 1980; Crick, 1984; Brefczynski and DeYoe, 1999), visual feature (e.g., color, orientation), or an object.

Attention offers several behavioral advantages. It is known to:

- Improve processing of stimuli at the attended location (Posner et al., 1980),
- Improve detection of faint stimuli and to lower contrast thresholds (Carrasco et al., 2000; Baldassi and Verghese, 2005),
- Improve feature discrimination (Lee et al., 1999),
- Increase spatial resolution (He et al., 1996; Yeshurun and Carrasco, 1998),
- Reject unwanted stimulus noise (Lu and Dosher, 1998; Ling et al., 2009),
- Increase the rate of visual processing (Carrasco and McElree, 2001),
- Affect appearance (Liu et al., 2006).

In effect, attention filters out irrelevant stimuli from the visual input and enables neural resources to be focused on the relevant locations, features and objects (Zhang et al., 2011).

Attentional modulation is widespread in the brain and has been observed in multiple areas along the cortical hierarchy including:

- V1 (Motter, 1993; Watanabe et al., 1998; Martinez et al., 1999; Huk and Heeger, 2000; Saenz et al., 2002; Verghese et al., 2012),
- V2 (Motter, 1993; Luck et al., 1997),
- V4 (Haenny and Schiller, 1988; Spitzer et al., 1988; Motter, 1993; Connor et al., 1997; Luck et al., 1997; McAdams and Maunsell, 1999; Williford and Maunsell, 2006; David et al., 2008; Ipata et al., 2012),
- MT (Treue and Maunsell, 1996; O'Craven et al., 1997; Treue and Trujillo, 1999; Saenz et al., 2002; Sohn et al., 2005),
- Lateral Intra-Parietal cortex (LIP) (Bushnell et al., 1981; Colby et al., 1996; Gottlieb et al., 1998; Bisley and Goldberg, 2003),
- Frontal Eye Fields (FEF) (Bichot and Schall, 2002; Moore and Fallah, 2004; Bichot et al., 2005),
- Subcortical structures like Lateral Geniculate Nucleus (LGN) (O'Connor et al., 2002) and Superior Colliculus (SC) (Munoz et al., 1991; Fecteau and Munoz, 2006).

Attentional effects are task-dependent. In separate studies, attention to color/shape has been shown to enhance BOLD activity in V4, while attention in a speed discrimination task increases activity in MT, and attention in a contrast discrimination task increases activity in V1 (Corbetta et al., 1990; Beauchamp et al., 1997; O'Craven et al., 1997; Huk and Heeger, 2000; Verghese et al., 2012). In fact, simply instructing observers to pay attention to different aspects of a same stimulus on different blocks of trials triggers different observable attentional modulation effects, in distinct anatomical and functional cortical areas. For example, Watanabe et al. (1998) showed, using one stimulus with superimposed translating and expanding fields of dots, differential attentional modulation of BOLD activation, depending on whether the task was to attend to the translating or the expanding feature of the stimulus.

Although different neural mechanisms for attention have been reported, the physiology literature presently appears to be divided. Attention to a neuron's preferred location or feature could:

- Cause a leftward shift in the neuron's contrast response function thus increasing the effective contrast of the stimulus (Reynolds et al., 2000; Martinez-Trujillo and Treue, 2002),
- Increase the response gain of the neuron a.k.a multiplicative scaling (McAdams and Maunsell, 1999; Treue and Trujillo, 1999; Womelsdorf et al., 2008; Boynton, 2009; Reynolds and Heeger, 2009; Saproo and Serences, 2010; Scolari and Serences, 2010; Scolari et al., 2012),
- Decrease the neuron's tuning width a.k.a bandwidth scaling (Moran and Desimone, 1985; Haenny and Schiller, 1988; Spitzer et al., 1988),
- Increase neuron's baseline or spontaneous activity a.k.a additive scaling (Luck et al., 1997; Chelazzi et al., 1998; Chawla et al., 1999; Kastner et al., 1999),
- Shift neurons tuned to nearby locations toward the attended location (Connor et al., 1996; Womelsdorf et al., 2006; David et al., 2008; Ipata et al., 2012),
- Modulate neuronal interactions through neuronal synchronization (Fries et al., 2001; Womelsdorf and Fries, 2007; Womelsdorf et al., 2007).

Note that the underlying mechanisms responsible for these observed effects at the single-unit level may be more complex, for example involving biasing or winner-take-all (WTA) competitions among neurons in a local population (Desimone and Duncan, 1995; Lee et al., 1999), or through gain modulation of upstream neurons (McAdams and Maunsell, 1999). **Figure 1** illustrates four possible types of attentional modulation of a neural population. Here, we discard the additive scaling since it has been argued that uniform translation of a tuning function does not affect the coding precision of that tuning function (Cover and Thomas, 1991) (but see Saproo and Serences, 2010), Paragraph 4 in the Discussion section and hence information content of a

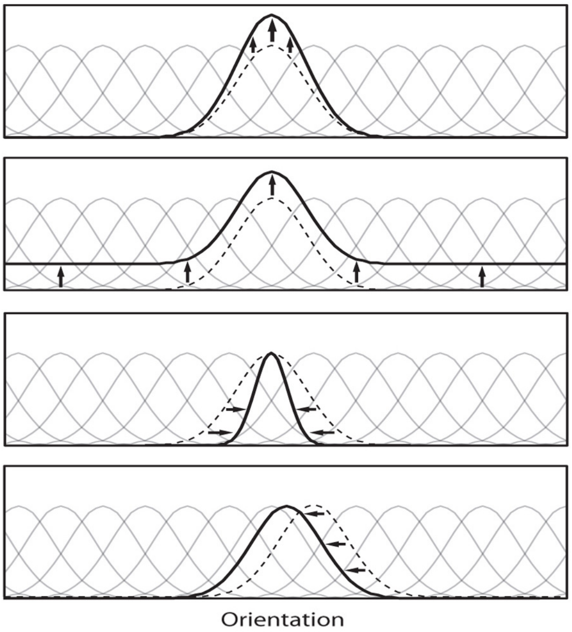


FIGURE 1 | This illustration depicts four possible attention-induced modulations of a neural population to a given visual task (here classification and visual search). Attention theoretically can: **(A)** Increase the gain of some important neurons a.k.a multiplicative scaling. This modulation selectivity increases the gain of the neurons that are more useful to find the target in visual search (or two classes in the classification and discrimination tasks). **(B)** Enhance response amplitudes in a feature-nonspecific manner a.k.a additive scaling. **(C)** Increase the selectivity of a neuron by modulating its tuning width (here sharpening) a.k.a bandwidth scaling, and **(D)** Shift tuning curves of neurons around to concentrate on important regions of the feature space (or shifting physical spatial receptive field of a neuron). Faint tuning curves correspond to the neural population before modulation, dotted black curve is the neuron under investigation, and the solid black curve is the modulated tuning curve. Here, we discard case **(B)** to make our simulations easier and tractable. Further, it has been argued that this case does not affect information decoding much.

neural population. Further, this simplification makes our analysis easier and tractable.

1.2. OPTIMAL ATTENTIONAL MODULATION

To gain better insight into above-mentioned discrepancies, we propose a unified account for optimal modulation of neural activity over two tasks: (1) *stimulus classification* (which of two stimuli was presented on the basis of the neural response pattern) and (2) *visual search* (i.e., enhancing the representation of the target stimulus, thus making search easier). Target selection often comes up in the context of a real world task such as visual search where the observer may be looking for a particular target, or for an unknown target that is the odd-ball. Our proposed framework can extend to additional tasks, including match-to-sample (as a neuron's response to the matching stimulus is enhanced while response to any non-matching stimulus is suppressed), discrimination, and stimulus reconstruction.

Let $p(\mathbf{r}|s_a)$ and $p(\mathbf{r}|s_b)$ be probability distributions of population activity $\mathbf{r}$ to two stimuli s_a and s_b. The goal of optimal population modulation is to find the best set of parameters for each of n sensory neurons (i.e., $\theta_i = [g_i, \sigma_i, \mu_i]$ including gain, tuning width, and feature selectivity) such that:

$$\boldsymbol{\phi}^* = \arg\max_{\boldsymbol{\phi}} f(p(\mathbf{r}(\boldsymbol{\phi})|s_a), p(\mathbf{r}(\boldsymbol{\phi})|s_b)), \quad \boldsymbol{\phi} = [\boldsymbol{\theta}_{i=1\ldots n}] \tag{1}$$

where f denotes the task objective function. For *classification* and *discrimination* tasks, f can be the mutual information between neural activity and behavioral response, or classification accuracy (e.g., linear discrimination error). Here we choose to maximize the inverse of minimum discrimination error (MDE) as the optimality criterion for the classification task. It has been shown that MDE has several advantages over other criteria such as Fisher Information (Berens et al., 2011). For *visual search* tasks, we choose to maximize signal to noise ratio (SNR). The concept of SNR has been suggested by psychophysicists as measured by the amount of overlap between target (="signal") and distractor (="noise") response distributions. If the purpose is *reconstruction* (i.e., estimate the true value of the presented stimulus on the basis of the noisy neural response $\mathbf{r}$: $\hat{s} = \arg\max_s p(s|\mathbf{r}) \propto \arg\max_s p(\mathbf{r}|s)p(s)$), then f can be the inverse of the mean squared error (MSE) between estimated stimulus (by means of a decoding method such as maximum-likelihood or population vector) and the actual input stimulus.

Optimizing above objective functions is a complex and time consuming process. For the brain this would be an optimization across many (usually thousands of) neurons, involving many different parameters which seems to be very daunting. Note that this does not happen instantly, rather it is a slow process of an organism learning to perform a task. Further, the stimulus distribution is also not available at once and demands the organism to interact with the environment and observe sensory data over time. Indeed, previous work by Baluch and Itti (2010) has shown that human observers become increasingly more efficient at biasing their visual system toward search targets in a triple conjunction search task. This suggests that humans learn over time how to bias the setting of their neural parameters so as to maximize task performance. Navalpakkam and Itti (2007) proposed a three-phase mechanism for learning top-down attentional modulation. In the first phase, bottom-up and top-down cues (learned previously) are applied to render some visual items salient. In the second phase, distributions of target and distractor features are learned through past trials, preview of picture cues, verbal instructions, etc. and in the third phase, optimal top-down gains (as well as other parameters) are computed (see Figure 2 in Navalpakkam and Itti, 2007). These gains will be later recalled and applied during future search trials.

2. THEORETICAL PERSPECTIVE

We formalize, in the Bayesian sense, how attention may modulate neural activity to optimize task performance. In classification tasks, the goal is to distinguish between a stimulus from class $C = 1$ [defined by a distribution of features $P(s|C = 1)$ in some dimension such as orientation] from a stimulus from class $C = -1$ [defined by a distribution of features $P(s|C = -1)$]. In visual search, class $C = 1$ is considered the target T that is to be found among distractors D ($C = -1$).

We assume that the incoming visual display is processed by a population of n neurons tuned to different features. We further assume that all neurons have idealized and homogeneous tuning functions. Let $\mathbf{r}(s) = [r_1(s), r_2(s), \dots, r_n(s)]$ denote the population vector of responses to input stimulus s. Assuming independent neurons, the probability distribution of response to a single stimulus s is:

$$L_r(s) = p(\mathbf{r}|s) = \prod_{j=1}^{n} p(r_j|s) \tag{2}$$

2.1. CLASSIFICATION

In classification tasks, a Bayesian ideal observer needs to estimate $\hat{C} = \arg\max_C P(C|\mathbf{r}) = \arg\max_C P(\mathbf{r}|C)P(C)/P(\mathbf{r})$ where $\hat{C}$ represents the estimated class (out of m classes). This equation means that the classifier chooses the class that was most likely to have caused the observed response pattern $\mathbf{r}$ on the basis of the stimulus conditional response distributions. For a two-class problem, the optimal neural decision variable depends on distributions of neural response to classes $P(\mathbf{r}|C = 1)$ and $P(\mathbf{r}|C = -1)$, each defined as:

$$p(\mathbf{r}|C) = \int p(\mathbf{r}|s)p(s|C)ds = \int L_r(s)p(s|C)ds \tag{3}$$

Thus, to maximize classification performance, the MDE objective function (the error of the ideal observer model) tries to minimize the overlap between neural response distributions to the two classes:

$$MDE(C = 1, C = -1) = \frac{1}{2}\int \min\big(p(\mathbf{r}|C = 1), p(\mathbf{r}|C = -1)\big)d\mathbf{r} \tag{4}$$

Discrimination is a special case of classification, with $p(s|C = 1) = d(s - s_a)$ and $p(s|C = -1) = d(s - s_b)$, where d denotes the Dirac delta function. In Berens et al. (2011), authors have used MDE to solve for the optimal tuning width of a neural population in reconstruction and discrimination tasks.

2.2. VISUAL SEARCH

Assuming that attention during visual search is guided to locations of high neural activity, search performance can be optimized by maximizing the strength of the signal (expected total neural response to the target $C = 1$) relative to the noise (expected total neural response to the distractors $C = -1$). Thus, using the above formulas, SNR can be written as:

$$\begin{aligned} SNR(C = 1, C = -1) &= \frac{\sum_i E(r_i|C = 1)}{\sum_i E(r_i|C = -1)} \\ &= \frac{\sum_i \int r_i p(r_i|C = 1)dr_i}{\sum_i \int r_i p(r_i|C = -1)dr_i} \\ &= \frac{\sum_i \int\int r_i p(r_i|s)p(s|C = 1)dsdr_i}{\sum_i \int\int r_i p(r_i|s)p(s|C = -1)dsdr_i} \end{aligned} \tag{5}$$

A closed-form solution for optimal gain modulation using SNR has been previously proposed in Navalpakkam and Itti (2007). Please note that here we attempt to solve visual search in feature space, irrespective of spatial organization of items in the search array. The SNR formulation has been shown to be capable of explaining a large number of psychophysics findings in the visual search literature (Verghese, 2001; Navalpakkam and Itti, 2007; Scolari and Serences, 2009, 2010; Jehee et al., 2011; Scolari et al., 2012). In addition, it has been shown that feature-based attention occurs independently of spatial attention (David et al., 2008), and feature-based attention changes activity globally throughout the visual-field representation (McAdams and Maunsell, 1999; Treue and Trujillo, 1999; Saenz et al., 2002; Maunsell and Treue, 2006; Serences and Boynton, 2007). In other words, attentding to a spatial location all features in that location are enhanced (McAdams and Maunsell, 1999; Boynton, 2009; Ling et al., 2009; Reynolds and Heeger, 2009). Conversely, attention to a specific feature results in global biases to that feature across the entire visual field (Treue and Maunsell, 1996; Treue and Trujillo, 1999; Saenz et al., 2002; Serences and Boynton, 2007).

3. SIMULATION RESULTS

We run two numerical simulations to investigate the optimal coding quality of a population of neurons under a range of stimulus conditions. The goal of this analysis is to reveal patterns or profiles of modulations depending on tasks and stimuli. Understanding how different patterns arise in different conditions can help design future experiments to pinpoint the neural basis of attentional modulation. In the first simulation, for simplicity and tractability, we choose a neural population of size 6 and we exhaustively search the parameter space for optimal solutions. We then run a second, larger simulation with 60 neurons on the most interesting cases. To illustrate our simulations, we consider the feature dimension of stimulus orientation, although our results apply interchangeably to other features such as color, spatial location, or direction of motion.

3.1. SMALL-SCALE SIMULATION

We assume a conventional model of neural response, where the i-th neuron ($i \in [1\ \ n]$, in a population of $n = 6$ equi-spaced uncorrelated neurons in $[0\ \ 180]$) has a bell-shaped tuning function:

$$\begin{aligned} f_i(s) &= g_i \times \left(\lambda_1 + \lambda_2\left(\frac{1}{2} + \frac{1}{2}cos(s - \mu_i)\right)^{20\sigma_i}\right); \\ p(r|s) &= \frac{1}{\sqrt{2\pi\upsilon_i^2}}e^{-\frac{(r - f_i(s))^2}{2\upsilon_i^2}} \end{aligned} \tag{6}$$

where s is the scalar stimulus feature (here orientation) and μ_i is the preferred feature of neuron i. The parameter g_i is the multiplicative gain. The parameter σ_i controls the width of the tuning curve. Large σ corresponds to steep tuning curves with small width. The parameters λ_1 and λ_2 set the baseline rate to 5 Hz and the maximal rate (amplitude) to 50 Hz. The firing activity of each neuron is assumed to follow a Gaussian distribution with Poisson-like noise, where variance is identical to mean spike count [i.e., $\upsilon_i^2 = \bar{r}_i(s) = 10f_i(s)$]. We estimate MDE and SNR (Equations 4, 5) using Monte Carlo techniques, by iteratively sampling from $p(s|C)$, and, for each s, many times from $p(r|s)$ to finally estimate $p(r|C)$ (similar approach as in Berens et al., 2011).

We consider two types of constraint regimens on neural parameters. The *first regimen* constrains each free parameter to change only within a restricted window, to adhere to biophysical constraints. Note that, otherwise, in visual search, a trivial solution to optimize SNR would be for every neuron to shift its preference to the target feature, change its tuning to infinitely narrow, and enhance its gain infinitely. However, such unbounded changes would likely consume enormous energy (every spike is costly), would prevent neurons from adapting to dynamically changing environments, and are implausible given the electrophysiological observations described in the Introduction. Thus, to prevent indiscriminate changes leading to this mathematical singularity, we constrain each free parameter to change only within a restricted window. We set bounds for g_i to [0.5 2], for σ_i to [0.5 3], and for μ_i to $[-0.2\ 0.2]$ (in radian, $\sim 11.46°$). A default value of 1 for g_i and σ_i, and 0 for μ_i means no change.

Constraint regimen one imposes constraints at the single cell level. Another possibility is to consider constraints at the population level as suggested by Navalpakkam and Itti (2007) where the sum of each parameter over the neural population is constrained (Our *second regimen*, $\sum g_i = 2,\ \sum \sigma_i = 3$, and $\sum \mu_i = 2$). This type of constraint needs more complex mechanisms to impose than constraint type one, for example by means of another neural network or a low-level molecular process. Similar to regimen one, regimen two leads to efficient spending of resources and energy but has more selective pressure as several solutions in regimen one may have equal objective function but in regime two optimization favors most informative neurons. Eventually, our treatment here is theoretical and further biological research is needed to discover which constraint is really implemented in the brain.

We also set the minimum value of g_i and σ_i to be 0.1 to preserve baseline activity. We employ real-valued Genetic Algorithms to exhaustively search the parameter space, in each individual dimension (i.e., g alone), for $g + \sigma$, as well as all three 3 parameters, to maximize SNR and MDE^{-1}. It is worth noting that the qualitative conclusions derived from our simulations do not depend on the exact values of bounds.

Figure 2 shows simulation results obtained by modulating g_i, σ_i, and μ_i in the above manner for two arrangements of stimulus classes: (1) an easy task where two classes are far apart ($C = 1$ at 45° and $C = -1$ at 135°), and (2) a difficult task where two classes are close to each other and thus more similar ($C = 1$ at 80° and $C = -1$ at 100°). We investigate two levels of uncertainty (low $\sigma_s = 5°$ and high $\sigma_s = 20°$) on stimulus distributions. For some cases in which solutions are not unique, we also show other good answers in insets. To further study the influence of stimulus distributions and initial parameterization, in **Figure 3** we illustrate solutions to some additional cases: (1) when only knowledge about one class is known, (2) three classes of stimuli (two targets and one distractor; See Supplementary materials for heterogeneous search, i.e., one target among two distractors), and (3) narrow default tuning curves ($\sigma_i = 5$). In each test case, we first describe results for classification, then search.

3.1.1. Response gain

In *classification*, under constraint regimen one, all neurons attain the maximum allowed gain, in both easy and difficult tasks. In regimen 2, all gains are concentrated around one of two classes, since both classes are equally important. Interestingly, and possibly counter-intuitively, if we were to distribute the gains equally around both stimulus classes, or equally among all neurons, the MDE would rise (i.e., worse classification). In *visual search*, SNR optimization shows that neurons tuned near the target feature undergo gain enhancement, while neurons tuned near the distractor feature undergo gain suppression (aligned with Treue and Trujillo, 1999 and Navalpakkam and Itti, 2007). While in regimen 2, only neurons at the target feature show gain enhancement, in regimen 1 neurons around the target are also enhanced. Interestingly in regimen 2, when target and distractor are very close and overlap is high (**Figure 2F**, $T = 80°,\ D = 100°,\ \sigma_s = 20$), in accordance with Navalpakkam and Itti (2007) and Scolari and Serences (2009), we also observe higher gain for the exaggerated neuron (at 45°) than for the neuron best tuned to the target (at 75°). However, unlike Navalpakkam and Itti (2007), baseline activity is sustained in our simulation, which agrees with electrophysiology findings (Chelazzi et al., 1998; Chawla et al., 1999; Kastner et al., 1999; David et al., 2008). Supporting single-unit evidence comes from feature-based attention tasks (McAdams and Maunsell, 1999; Treue and Trujillo, 1999; Martinez-Trujillo and Treue, 2004; David et al., 2008; Jehee et al., 2011).

3.1.2. Tuning width

Maximum *classification* accuracy, in the easy task and in regimen 1, is obtained when all neurons widen their tuning as much as possible. In other cases (difficult task, regimen 1, and both tasks in regimen 2), optimization leads to sharpening near both stimuli and widening elsewhere (see also **Figure 3**). In *visual search*, our results suggest that attention causes both narrowing and widening of tuning width, and the choice depends on the difficulty of the task. In regimen 1, in the easy task, neurons at and near the target feature are maximally widened while neurons near the distractor feature are maximally sharpened. In regimen 2, in the easy task, we observe widening of neurons both at target and distractor, which was unexpected. Since neurons tuned near the distractor feature already respond strongly to the distractor (due to our bounds), sharpening would indeed only boost the distractor and lower SNR; however, widening for these neurons represents a "better worst-case scenario," as it will make them respond to both distractor and target, resulting in slightly higher SNR compared to sharpening. When we made the task even easier (**Figure** 3*), we then observed that neurons at distractor sharpened. Over the difficult task in both regimens, we observe a sharpening at the target and widening near the distractor, which is the opposite of the easy task in regimen 1. When $p(s|T)$ and $p(s|D)$ do not overlap much (i.e., low uncertainty), and/or tuning curves are narrow and far apart, neural tuning widens near the target and sharpens near the distractor. The opposite happens when $p(s|T)$ and $p(s|D)$ highly overlap or the population is very dense. Note that parameter setting is important in the optimal answers. While exact values might

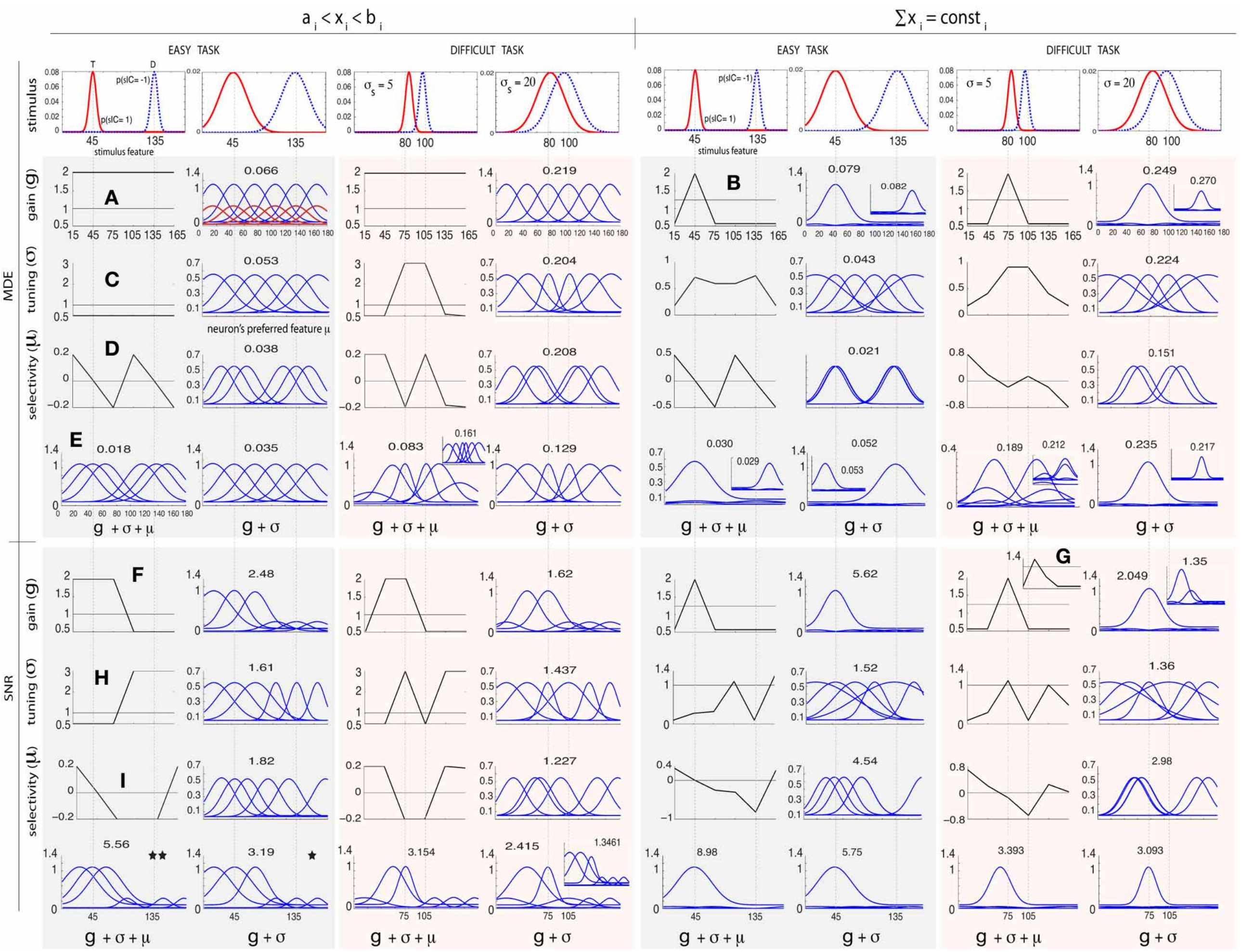


FIGURE 2 | Optimal attentional modulation for classification (MDE; top row) and visual search tasks (SNR; bottom row). The lower the MDE, the better (opposite is true for SNR). Left and right columns correspond to two parameter regimens for 2 classes of stimuli: (1) coarse classification (or easy search) $C = 1$ (target; solid line) at 45° and $C = -1$ (distractor; dashed line) at 135°, 2) fine classification (or hard search) $C = 1$ at 80°, and $C = -1$ at 100°, each at two uncertainty levels ($\sigma_S = 5°$ and $\sigma_S = 20°$). **(A)** MDE, g, regimen 1: gains should be maximized for all neurons in both tasks. The red tuning curves represent the default tunings ($\sigma = 1$). **(B)** MDE, g, regimen 2: All gain is allocated to one of the two classes. **(C)** MDE, σ, regimen 1: all neurons in easy task should be widened. In other cases, neurons at 2 classes should be sharpened while the rest should be widened. **(D)** MDE, μ: neurons should be moved to locations of classes in all cases. **(E)** MDE, $g + \sigma$ and $g + \sigma + \mu$ has the superposition of individual effects. **(F)** SNR, g, regimen 1: gains of neurons nearby target should be enhanced. In regimen 2, gain at the target should be amplified in easy task. **(G)** In difficult search ($\sigma_S = 20°$), the gain of the exaggerated neuron should be amplified more than the neuron at the target. **(H)** SNR, σ, easy task: neurons nearby target should be widened while neurons near distractor should be sharpened (see text). In difficult search task, neurons near target should be sharpened while neurons near distractor should be widened. **(I)** SNR, μ: neurons should be moved toward the target and away from the distractor.

differ for different parameter settings, we believe that patterns will stay the same (e.g., dependency of results to task difficulty). For experimental works, when biophysical properties of a neural population are known, it is easy to run a simulation (with our shared code) and verify a hypothesis. Supporting evidence for sharpening at the target comes from single-unit studies of orientation (Spitzer et al., 1988) and spatial tuning (Moran and Desimone, 1985).

3.1.3. Preferred feature

In *classification*, optimization moves neurons toward either of the two classes as much as possible, in both regimens over both tasks. The optimal answer in *visual search* is to move neurons toward the target and away from the distractor. Supporting evidence for tuning shifts comes from single-unit studies in feature-based (David et al., 2008; Ipata et al., 2012) and spatial attention (Connor et al., 1996; Womelsdorf et al., 2006).

3.1.4. All parameters

Comparing results obtained for the joint optimization of all parameters and the separate optimization of g, σ, and μ, we empirically find that the superposition of optimal answers to each individual parameter is always a good answer (although we do not have a theoretical guarantee on the optimality or uniqueness of such answer). For example, optimizing gain and tuning width jointly in easy visual search, regimen 1 (See **Figure** 2*),

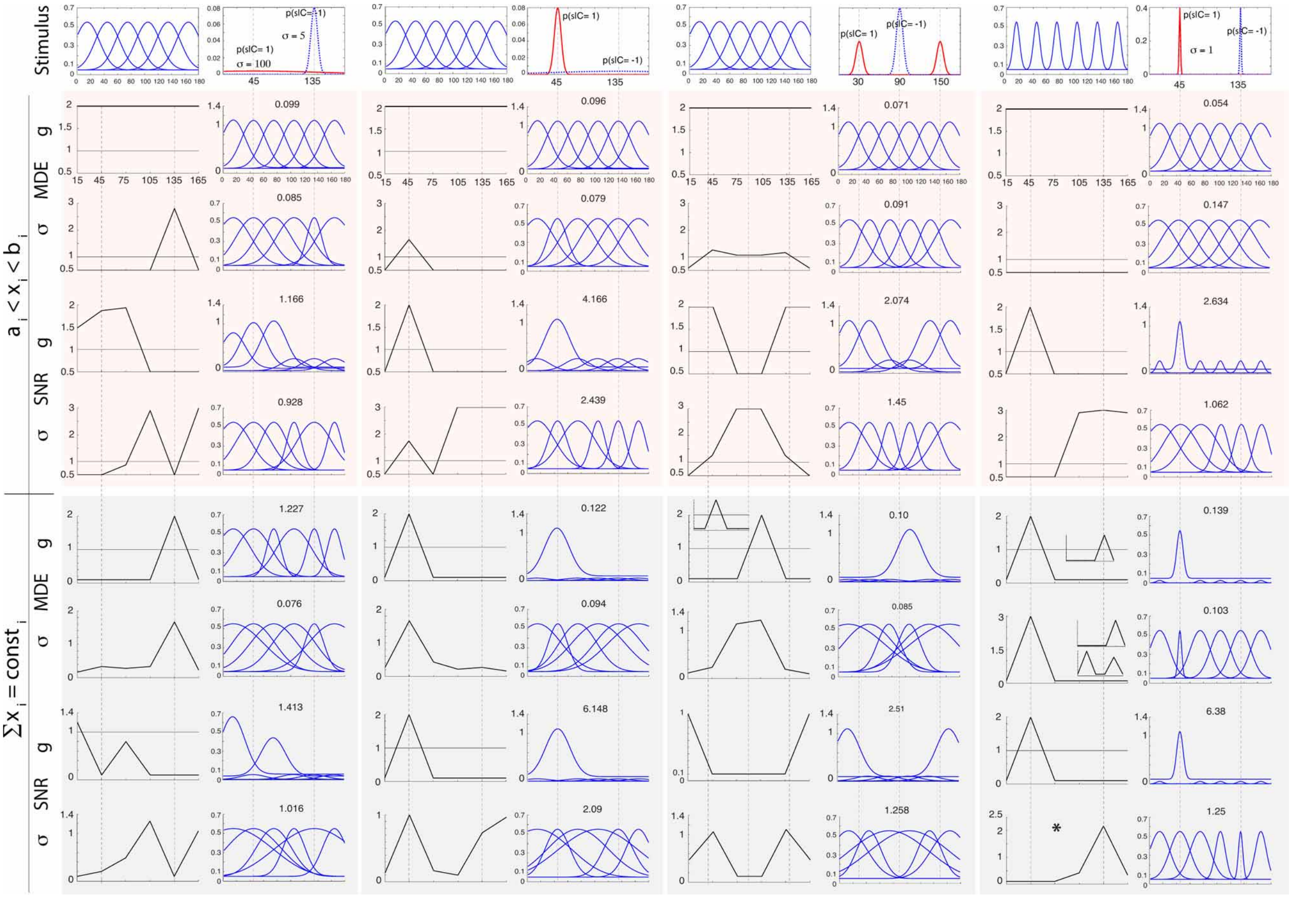


FIGURE 3 | Optimal neural modulation of g, σ and μ for additional cases mentioned in the text (small-scale simulation). Columns from left to right: unknown target (here modeled as a very wide distribution with $\sigma_S = 100$ shown with the dotted blue curve) and known distractor at 135° with $\sigma_S = 5$ (solid red curve), known target at 45° and unknown distractor, visual search for two targets at 30° and 150° and a single distractor at 90° with $\sigma_S = 5$, easy search for a target at 45° with a narrow distribution ($\sigma_S = 1$) and a distractor at 135°. In each column/setting, the left side shows the original neural population with 6 neurons and below that are the optimal parameters. The right side in each column shows the neural population after modulation. The top rows shows results for regimen 1 while the bottom one corresponds to regimen 2, for both classification (MDE) and visual search tasks (SNR). The panel with ***** shows the optimal σ in an easy visual search task.

leads to maximal gain amplification and widening of neurons around the target, while minimizing gains of neurons selective to the distractor. Note that tuning width modulation of neurons near the distractor is not important here since their gain has already been minimized. When optimizing all three parameters, in addition to the joint answer of gain and tuning width, neurons are also shifted toward the target and away from the distractor (See **Figure** 2**). Our results also show that modulation of multiple parameters always yields better performance than optimizing only one or two parameters. This suggests that biological top-down attention may also affect multiple parameters, although most previous reports have focused on one parameter at a time.

Optimal neural modulation in heterogeneous visual search (i.e., one target among two distractors and vice versa) and optimizing g, σ, and μ with 12 neurons shows the same patterns as in **Figure 2**. These results are shown in Supplementary materials.

Figure 4 shows the optimal MDE and SNR values (in regimen 1) as a function of target-distractor dissimilarity for g, σ, and $g + \sigma$ (averaged over $T \in \{30°, 40°, 50°, 60°\}$ and $D = T + \{10°, 20°, 30°, 40°, 50°, 60°\}$). Increasing the distance between the two classes leads to decrease in MDE and a ramp up in SNR. This qualitatively matches with human performance as a function of task difficulty (Duncan and Humphreys, 1989). Over both MDE and SNR, modulating both g and σ wins over single parameters. The tuning width is more effective than gain in classification, as seen by lower MDE values of σ than MDE values using g. The opposite occurs in visual search using SNR. One reason why SNR values for σ are small might be because neurons in this simulation are not allowed to sharpen beyond a certain limit.

3.1.5. Note on noise correlation

In our simulations so far, we considered optimal modulation of an uncorrelated neural population for the sake of simplicity

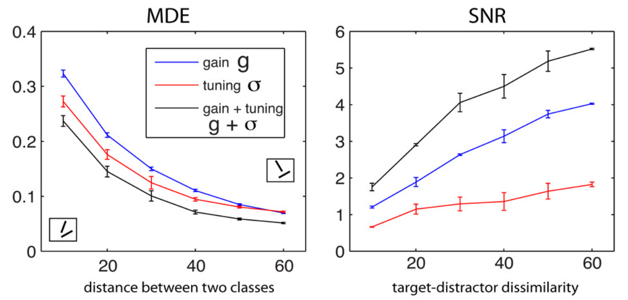


FIGURE 4 | Dependency of objective functions to dissimilarity between two classes for the small-scale simulation with 6 neurons for *g*, σ, and $g + \sigma$ (averaged over $T \in \{30°, 40°, 50°, 60°\}$ and $D = T + \{10°, 20°, 30°, 40°, 50°, 60°\}$). Left: MDE for classification and Right: SNR for visual search. MDE decreases as two classes become more separate from each other while SNR raises which means that in both cases task becomes progressively easier.

(i.e., uncorrelated noise). But, noise in the brain is correlated and this might influence the amount of information a neural population conveys (Averbeck et al., 2006) (See also Seriès et al., 2004 and Bejjanki et al., 2011). Here, we analyze the role of correlations (correlated noise) in optimal modulation of parameters for visual search (i.e., maximizing SNR) on our small scale neural population with 6 neurons.

Following Berens et al. (2011), we model the stimulus-conditional response distribution as a multivariate Gaussian:

$$p(\mathbf{r}|s) = \mathcal{N}(\bar{\mathbf{r}}(s), \Sigma(s)) \tag{7}$$

In above equation, $\bar{\mathbf{r}}(s) = (\bar{r}_1(s), \bar{r}_2(s), \ldots, \bar{r}_6(s))$ and $\Sigma(s)$ represent average spike counts and covariance matrix, respectively. This allows us to inject Poisson-like noise correlations into our simulation (See Berens et al., 2011 and their supplement for more details on adding correlated noise). Results are shown in supplementary materials for optimal answers of searching a target at 80° and distractor at 100° with $\sigma_s = 5°$ (see **Figure 2**). We consider 10% noise correlation in our simulations. As it can be seen patterns of results are similar to those shown in **Figure 2** for both constraint regimens and all three neural parameters. This could be because the effect of noise is vanished when averaging the neural activity, to targets and to distractors in SNR computation. For future research we encourage a more detailed look at noise correlations (e.g., non-uniform correlations) and how they may affect optimal solutions on larger neural populations.

3.2. LARGE-SCALE SIMULATION

The previous analysis revealed different patterns of modulation depending on task and stimulus conditions. Importantly, it revealed that joint optimization of all parameters always yields better performance than optimizing only one parameter. This prompts us to study the relative utility or contribution of modulating each parameter as part of a joint optimization. To further investigate this, we focus on visual search in a larger-scale, more detailed simulation. We simulated a population of $n = 60$ equi-spaced, broad, overlapping Gaussian neurons with preferred stimulus feature μ_i, tuning width σ_i, amplitude λ_2, gain factor g_i, and baseline firing rate λ_1:

$$f_i(s) = g_i \times \left(\lambda_1 + \lambda_2 e^{-(s-\mu_i)^2/2\sigma_i^2}\right), \quad i = 1, \ldots, n;$$
$$p(r|s) = \frac{e^{-f_i(s)} f_i(s)^r}{r!} \tag{8}$$

with default tuning width of 10°, default gains at unity, spacing between preferred orientations of adjacent neurons 3° spanning 0–180° in orientation space (**Figure 5**). In addition, we consider the noise in neural response (to repeated presentations of a same stimulus) to have Poisson variability (used to numerically compute the expectations in the Equation 5). Here, we set $\lambda_1 = 0$, for simplicity.

We jointly maximized SNR wrt. g_i, μ_i, and σ_i using a multi-start Nelder-Mead simplex algorithm (Nelder and Mead, 1965) (genetic algorithm was too slow in this larger-scale test). We used multiple initial conditions to avoid converging into local optima (20 different initial conditions, each with a random jitter in g_i, μ_i, and σ_i of up to 50% from default values), and considered the solution with maximum SNR. Here, attention can modulate g_i by up to ±50% of its default unity value, and σ_i and μ_i by up to ±50% of the default tuning width (corresponding to **regimen 1** and to avoid numerical instability).

Figure 5 shows how neural parameters may be optimally modulated in an easy search (with an orientation difference between target and distractors of $5\sigma_0 = 50°$), and a difficult search task (smaller orientation difference of $\sigma_0 = 10°$). After modulation, the expected neural response to the target is much higher than the distractor (**Figure 5C**) compared to before modulation (**Figure 5B**). This effect is more clearly seen in the difficult task, where the initial population response to the target and distractor are similar (**Figure 5B**, 2nd column, hence a low SNR), but different after modulation (**Figure 5C**, 2nd column), leading to an improvement in SNR. Optimization results here are aligned with our smaller-scale simulation (**Figure 2**). Interestingly, since here target and distractor are well separated in the easy task, neurons around the target widen while those tuned near the distractor sharpen. In contrast, neurons sharpen near the target and widen near the distractor in the difficult task.

3.2.1. Analysis of tuning curve overlap

How much is SNR dependent on the degree of neural overlap? Over our population of 60 neurons, we change σ from 6° to 35° and task difficulty from 10° to 100° and then find the optimal solutions for g, σ, and μ. **Figure 6** shows that increasing the overlap between neurons reduces SNR for all parameters regardless of task difficulty. This impairment is more profound in difficult tasks than in easy tasks. In easy tasks, irrespective of the degree of overlap, SNR values using gain are higher than SNR due to σ and μ. SNR using gain increases as the difference between target and distractor increases. Interestingly, there is an interaction between overlap and task difficulty when optimizing for σ and μ (non-monotonic curve shapes in **Figure 6**).

The analysis of SNR changes as a function of tuning overlap suggests explicit qualitative predictions that could be made

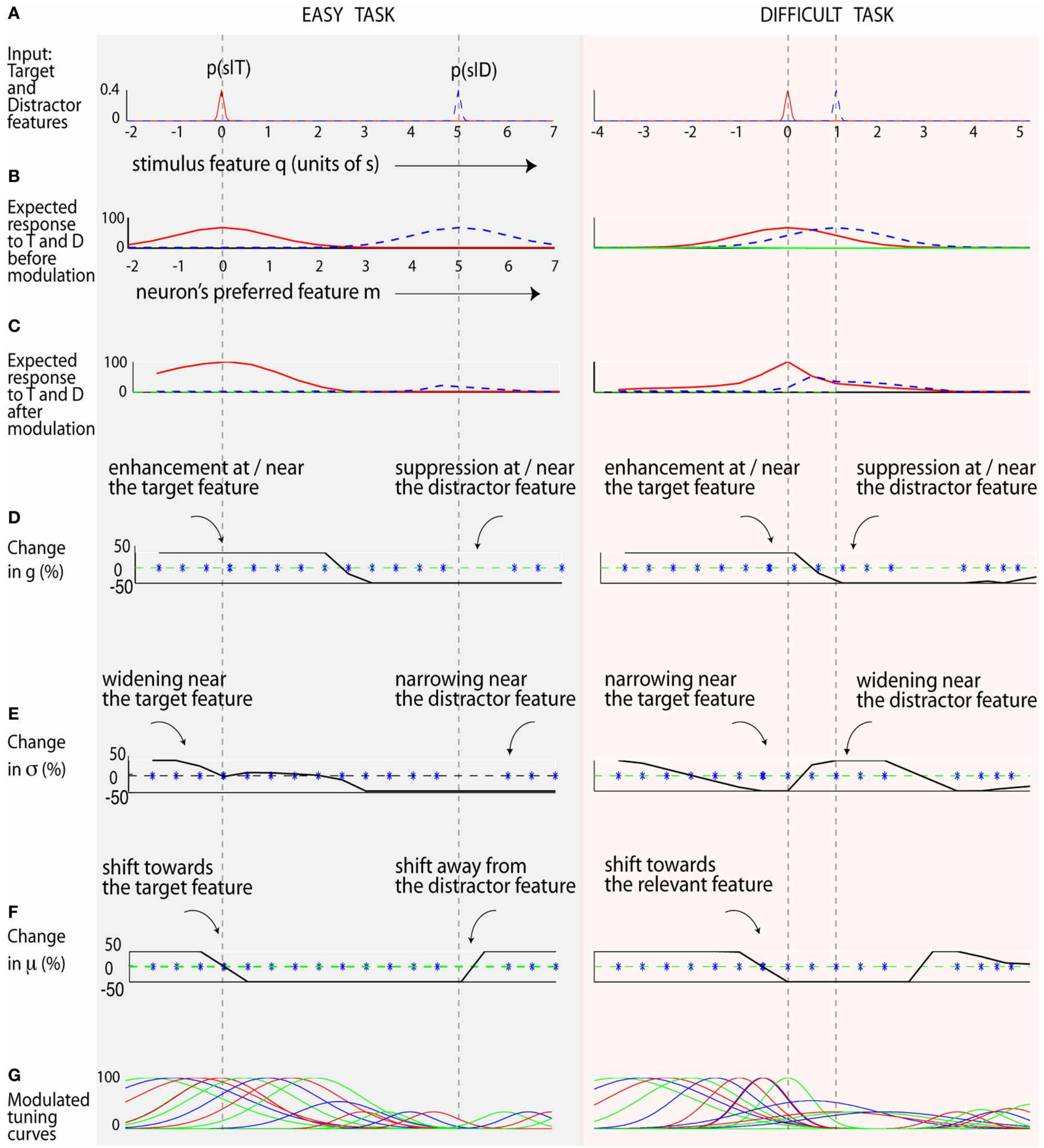


FIGURE 5 | Attentional modulation in easy and difficult visual search. **(A)** The input stimuli. Rows **(B,C)** show the expected response of neurons (tuned to different features) before and after modulation. The solid red line is the expected response to the target, while the dotted blue line represents the expected response to the distractor. **(D)** The optimal shift in response gain is shown by the solid black line. Neurons tuned near the target increase their gain, while others tuned near the distractor undergo suppression. **(E)** The optimal shift in neuron's tuning width (σ) is shown here in the solid black line. In the difficult task, neurons tuned to the target feature decrease their tuning width, while nearby neurons widen their tuning width. **(F)** The optimal shift in preferred features μ is shown by the solid black line. A positive shift ($\Delta\mu_i > 0$) indicates neurons shifting to the right, and vice versa. The blue star shows the neuron's preferred feature after the modulation. Neurons shift toward the target feature and away from the distractor feature (as seen by the lack of blue stars near the distractor). **(G)** The optimal tuning curves.

when looking across cortical areas (given that orientation tuning inherently broadens as one ascends the visuocortical hierarchy). Moving along the hierarchy, neurons become broader (thus higher overlap among neurons) which eventually causes lower SNR. Also note that the peak of the curves in **Figure 6** shifts to the right suggesting that maximum separability happens for more dissimilar stimuli.

3.2.2. Behavioral utility of neural modulation

How useful is the modulation of each neural parameter? To answer this question, we computed a utility statistic $u(p)$ for a parameter $p \in \{g, \sigma, \mu\}$ as the ratio of benefit to SNR obtained by modulating p alone vs. modulating everything. Higher utility values indicate that more performance is achieved by modulating p compared to other parameters, i.e., p is a high-yield parameter to modulate in the particular task and stimulus studied. As seen in **Figure 7**, $u(g)$ and $u(\mu)$ both decrease with increasing task difficulty, but $u(\sigma)$ does not. Thus, in easy tasks (where the target and distractor differ by $\Delta \geq 40°$) modulating g or μ is more useful, but becomes less useful in difficult tasks. On the other hand, while modulating σ is not very beneficial in easy tasks, it becomes necessary in difficult tasks ($\Delta \leq 25°$). Furthermore, in easy tasks, simulation predicts that the combined modulation of μ and g is sufficient to yield close to best behavioral performance, but their combined utility decreases with increasing task difficulty.

4. DISCUSSION AND CONCLUSION

Results of two consistent simulations reveal that:

1. In classification, when two classes are well separated, all neurons should be widened and gains should be boosted,
2. In classification, when two classes are close in feature space, neurons selective to both should be sharpened and their gains should be increased,
3. In easy search, the optimal solution is to widen and boost gain at the target, and sharpen and reduce gain around the distractor (the opposite is seen for tuning width in difficult search),
4. Only in constraint regimen 2 and in difficult search, maximum gain is allocated to the exaggerated neuron as predicted by Navalpakkam and Itti (2007) and seen by Scolari and Serences (2009),
5. Feature selectivity of neurons should be biased toward target features (the two classes in classification) and away from distractors,

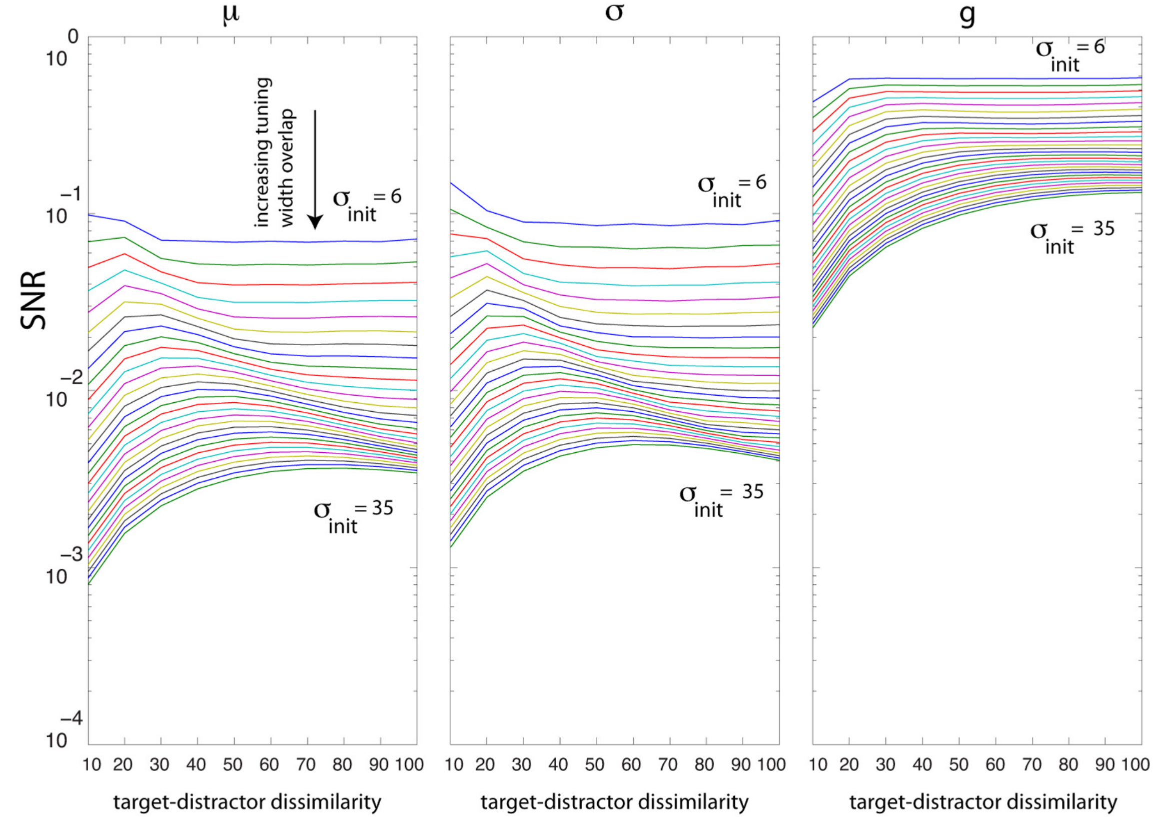


FIGURE 6 | Analysis of tuning curve overlap (σ from 6 to 35°; spacing between neurons is 3°). The x axis shows task difficulty due to target-distractor dissimilarity (measured by increasing orientation difference between the target and distractor: for $j = 1 : 10$, $T = 60° - j \times 5$, $D = 60° + j \times 5$). The y axis shows the best SNR achieved by optimizing each parameter. Curves from top to bottom indicate higher overlap between neurons. Increasing the neural overlap impairs the SNR due to optimal σ and μ more than SNR by g.

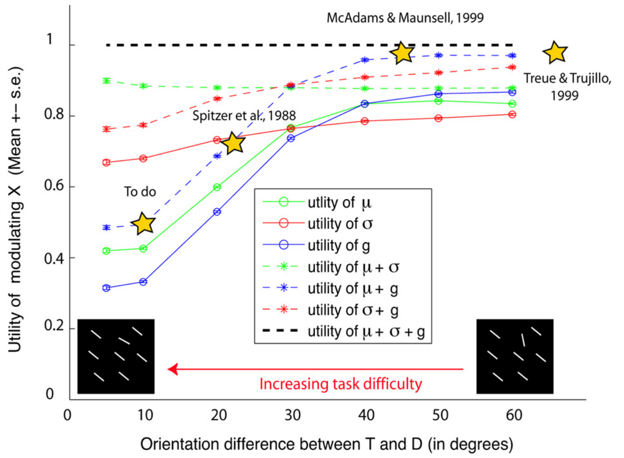


FIGURE 7 | Utility of attentional modulation. The *x* axis shows task difficulty due to target distractor similarity. The *y* axis shows simulation predictions of utility of modulating preferred features (μ), tuning width (σ), response gains (g), or any combination of these parameters. For easy tasks, we predict that modulating preferred features and gains are useful and sufficient (yielding 0.97 × the best performance). But their combined utility decreases with decreasing orientation difference between the target and distractors ($u = 0.49$), rendering them less useful in difficult tasks. On the other hand, modulating tuning width is more useful and necessary in difficult tasks. A similar trend is observed in separately modulating gains or preferred feature vs. tuning width.

6. Optimizing multiple parameters is better than optimizing a single one and joint solutions seem to be combinations of constituent ones,
7. Increasing overlap among neurons worsens SNR, which is more harmful in difficult than in easy search,
8. Uniform noise correlation did not affect our conclusions but more detailed analysis of different noise conditions is encouraged,
9. Task difficulty is a key factor in determining the utility of a neural parameter.

Our theoretical investigation sheds new light on the ongoing controversy of attentional modulation, by indicating that the reported discrepancies in the literature may be due to differences in task difficulty (**Figure 7**). For instance, previous physiological studies that reported gain modulation (McAdams and Maunsell, 1999; Treue and Trujillo, 1999) used easy tasks: McAdams and Maunsell used an angular difference of 45° or 90° between target and distractor, while Treue and Martinez-Trujillo used either no distractor or one 180° from the target. Previous studies that found preferred feature modulation also used easy tasks: (Womelsdorf et al., 2008) used a spatial attention task where monkeys attended to a target location in the absence of distractors. In such easy tasks, as predicted by our theoretical analysis, modulation of gains and preferred features (which is most useful) is observed, while tuning width modulation (not useful) is not observed. One of the few previous studies (Spitzer et al., 1988) that reported tuning width modulation, observed it in more difficult discrimination tasks (smaller angular difference of 22.5°). Nevertheless, as tuning width modulation remains a controversial issue (e.g., Treue and Trujillo, 1999), our main goal here it to show how tuning width modulation is an optimal strategy when the task is difficult.

It is difficult to disentangle the effect of gain and tuning width modulation behaviorally (see Ling et al., 2009). We suggest neurophysiology experiments for this purpose by systematically controlling for task difficulty. An ideal task for testing tuning width modulation would be when the monkey attends to a target feature in the presence of flanking distractor (e.g., attend to a 45° oriented moving random dot pattern (RDT) among 50 and 40° oriented RDTs). In such a task, modulating preferred features or gains will not suffice as neurons responding to the target will also respond to similar distractors. Instead, sharpening the tuning curve will help the target-sensitive neurons by decreasing interference from distractors, hence better resolving the difference between target and distractor. In contrast, when the target and flanking distractor are very different (e.g., more than 45° apart), modulating tuning widths is not useful, and thus modulation of preferred features and gains should be observed.

Our model generalizes over previous gain-only models: guided search theory (Wolfe et al., 1989), feature-similarity gain principle (Treue and Trujillo, 1999; Martinez-Trujillo and Treue, 2004), and optimal gain theory (Navalpakkam and Itti, 2007). The guided search theory revises the feature integration theory (FIT) and suggests that top-down attention acts as a linear weighted combination of multiple features which in effect makes an object of interest more salient among distractors and decreases the search time. However, similar to FIT, this theory only attempts to explain the behavior of the organism. In the the feature similarity gain model, gain modulation is a function of similarity between the neuron's preferred feature and the target feature. This theory does not consider target-distractor similarity. The optimal gain theory, combines information from both the target and distracting clutter to maximize the relative salience of the target. Interestingly, this model predicts that it is sometimes optimal to enhance the non-target features (e.g., **Figure 2G**). Here, we considered three neural parameters and showed how distribution of target and distractors can be used to optimally tune all these parameters and make the target salient.

In addition to gain, our model offers testable predictions for tuning width modulation and shifts in selectivity (seen by David et al., 2008 and Ipata et al., 2012 in area V4). Our model differs from the well-established normalization model of attention (Reynolds and Heeger, 2009) in one main aspect: the normalization model commits to explain low-level attentional mechanisms, while our model offers a high-level theoretical account for optimal attention over a population of neurons, considering task difficulty, and stimulus statistics. Obviously, our model has limited prediction power. It may need to be further expanded to account for optimal spatial attention, when deployed jointly with feature-based attention in hybrid spatial/feature tasks. We encourage future neurophysiology studies, with our theoretical framework in hand, to further explore such tasks, which will give new insights for developing unified models of spatial and feature-based attention.

In summary, we investigated three attentional mechanisms, namely attentional modulation of neural response gain, tuning width and preferred feature. Reports from different laboratories differ on whether attention modulates tuning width or gain or preferred feature. We have proposed a simple computational model that reconciles the above differences by predicting that task-difficulty (due to target-distractor similarity) plays a critical role in determining attentional modulation. Our model predicts that gain and preferred feature modulation is useful in easy tasks, while tuning width modulation is useful in difficult tasks – a prediction that is in good qualitative agreement with reported data. This unified model illuminates the similarities and differences in reported data from various laboratories, and provides guidelines for future experiments.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation (grant number CMMI-1235539), the Army Research Office (W911NF-11-1-0046 and W911NF-12-1-0433), and US Army (W81XWH-10-2-0076). The authors affirm that the views expressed herein are solely their own, and do not represent the views of the United States government or any agency thereof. We thank Vidhya Navalpakkam for very useful discussion and insights on this manuscript. Also thanks to reviewers for their valuable comments.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/journal/10.3389/fncom.2014.00034/abstract

REFERENCES

Averbeck, B. B., Latham, P. E., and Pouget, A. (2006). Neural correlations, population coding and computation. *Nat. Rev. Neurosci.* 7, 358–366. doi: 10.1038/nrn1888

Baldassi, S., and Verghese, P. (2005). Attention to locations and features: different top-down modulation of detector weights. *J. Vis.* 5, 556–570. doi: 10.1167/5.6.7

Baluch, F., and Itti, L. (2010). Training top-down attention improves performance on a triple-conjunction search task. *PLoS ONE* 5:e9127. doi: 10.1371/journal.pone.0009127

Baluch, F., and Itti, L. (2011). Mechanisms of top-down attention. *Trends Neurosci.* 34, 210–224. doi: 10.1016/j.tins.2011.02.003

Beauchamp, M. S., Cox, R. W., and DeYoe, E. A. (1997). Graded effects of spatial and featural attention on human area MT and associated motion processing areas. *J. Neurophysiol.* 78, 516–520.

Bejjanki, V. R., Beck, J. M., Lu, Z.-L. L., and Pouget, A. (2011). Perceptual learning as improved probabilistic inference in early sensory areas. *Nat. Neurosci.* 14, 642–648. doi: 10.1038/nn.2796

Berens, P., Ecker, A. S., Gerwinn, S., Tolias, A. S., and Bethge, M. (2011). Reassessing optimal neural population codes with neurometric functions. *Proc. Natl. Acad. Sci. U.S.A.* 108, 4423–4428. doi: 10.1073/pnas.1015904108

Bichot, N. P., Rossi, A. F., and Desimone, R. (2005). Parallel and serial neural mechanisms for visual search in macaque area v4. *Science* 308, 529–534. doi: 10.1126/science.1109676

Bichot, N. P., and Schall, J. D. (2002). Priming in macaque frontal cortex during popout visual search: feature-based facilitation and location-based inhibition of return. *J. Neurosci.* 22, 4675–4685.

Bisley, J. W., and Goldberg, M. E. (2003). The role of the parietal cortex in the neural processing of saccadic eye movements. *Adv. Neurol.* 93, 141–157.

Borji, A., and Itti, L. (2013). State-of-the-art in visual attention modeling. *Patt. Anal. Mach. Intell. IEEE Trans.* 35, 185–207. doi: 10.1109/TPAMI.2012.89

Borji, A., and Itti, L. (2014). Defending yarbus: eye movements reveal observers' task. *J. Vis.* 14, 1–22.

Boynton, G. M. (2009). A framework for describing the effects of attention on visual responses. *Vision Res.* 49, 1129–1143. doi: 10.1016/j.visres.2008.11.001

Brefczynski, J. A., and DeYoe, E. A. (1999). A physiological correlate of the 'spotlight' of visual attention. *Nat. Neurosci.* 2, 370–374. doi: 10.1038/7280

Bushnell, M. C., Goldberg, M. E., and Robinson, D. L. (1981). Behavioral enhancement of visual responses in monkey cerebral cortex. i. modulation in posterior parietal cortex related to selective visual attention. *J. Neurophysiol.* 46, 755–772.

Carrasco, M. (2011). Visual attention: the past 25 years. *Vision Res.* 51, 1484–1525. doi: 10.1016/j.visres.2011.04.012

Carrasco, M., and McElree, B. (2001). Covert attention accelerates the rate of visual information processing. *Proc. Natl. Acad. Sci. U.S.A.* 98, 5363–5367. doi: 10.1073/pnas.081074098

Carrasco, M., Penpeci-Talgar, C., and Eckstein, M. (2000). Spatial covert attention increases contrast sensitivity across the csf: support for signal enhancement. *Vision Res.* 40, 1203–1215. doi: 10.1016/S0042-6989(00)00024-9

Chawla, D., Rees, G., and Friston, K. J. (1999). The physiological basis of attentional modulation in extrastriate visual areas. *Nat. Neurosci.* 2, 671–676.

Chelazzi, L., Duncan, J., Miller, E. K., and Desimone, R. (1998). Responses of neurons in inferior temporal cortex during memory-guided visual search. *J. Neurophysiol.* 80, 2918–2940.

Chen, Z. (2012). Object-based attention: a tutorial review. *Attent. Percept. Psychophys.* 74, 784–802. doi: 10.3758/s13414-012-0322-z

Cohen, E. H., and Tong, F. (2013). Neural mechanisms of object-based attention. *Cereb. Cortex*. doi: 10.1093/cercor/bht303. [Epub ahead of print].

Colby, C. L., Duhamel, J. R., and Goldberg, M. E. (1996). Visual, presaccadic, and cognitive activation of single neurons in monkey lateral intraparietal area. *J. Neurophysiol.* 76, 2841–2852.

Connor, C. E., Gallant, J. L., Preddie, D. C., and Essen, D. C. V. (1996). Responses in area v4 depend on the spatial relationship between stimulus and attention. *J. Neurophysiol.* 75, 1306–1308.

Connor, C. E., Preddie, D. C., Gallant, J. L., and Essen, D. C. V. (1997). Spatial attention effects in macaque area v4. *J. Neurosci.* 17, 3201–3214.

Corbetta, M., Miezin, F. M., Dobmeyer, S., Shulman, G. L., and Petersen, S. E. (1990). Attentional modulation of neural processing of shape, color, and velocity in humans. *Science* 248, 1556–1559.

Corbetta, M., and Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nat. Rev. Neurosci.* 3, 201–215. doi: 10.1038/nrn755

Cover, T. M., and Thomas, J. A. (1991). *Elements of Information Theory*. New York, NY: Wiley-Interscience. doi: 10.1002/0471200611

Crick, F. (1984). Function of the thalamic reticular complex: the searchlight hypothesis. *Proc. Natl. Acad. Sci. U.S.A.* 81, 4586–4590. doi: 10.1073/pnas.81.14.4586

Crick, F., and Koch, C. (1998). Constraints on cortical and thalamic projections: the no-strong-loops hypothesis. *Nature* 391, 245–250.

David, S., Hayden, B., Mazer, J., and Gallant, J. (2008). Attention to stimulus features shifts spectral tuning of V4 neurons during natural vision. *Neuron* 59, 509–521. doi: 10.1016/j.neuron.2008.07.001

Desimone, R., and Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annu. Rev. Neurosci.* 18, 193–222. doi: 10.1146/annurev.ne.18.030195.001205

Duncan, J. (1984). Selective attention and the organization of visual information. *J. Exp. Psychol. Gen.* 113, 501–517. doi: 10.1037/0096-3445.113.4.501

Duncan, J. (1996). *Cooperating Brain Systems in Selective Perception and Action.* Cambridge, MA: Attention and Performance XVI.

Duncan, J., and Humphreys, G. W. (1989). Visual search and stimulus similarity. *Psychol. Rev.* 96:433. doi: 10.1037/0033-295X.96.3.433

Eckstein, M. P. (2011). Visual search: a retrospective. *J. Vis.* 11:14. doi: 10.1167/11.5.14

Eckstein, M. P., Peterson, M. F., Pham, B. T., and Droll, J. A. (2009). Statistical decision theory to relate neurons to behavior in the study of covert visual attention. *Vis. Res.* 49, 1097–1128. doi: 10.1016/j.visres.2008.12.008

Fecteau, J. H., and Munoz, D. P. (2006). Salience, relevance, and firing: a priority map for target selection. *Trends Cogn. Sci.* 10, 382–390. doi: 10.1016/j.tics.2006.06.011

Fries, P., Reynolds, J. H., Rorie, A. E., and Desimone, R. (2001). Modulation of oscillatory neuronal synchronization by selective visual attention. *Science* 291, 1560–1563. doi: 10.1126/science.1055465

Gottlieb, J. P., Kusunoki, M., and Goldberg, M. E. (1998). The representation of visual salience in monkey parietal cortex. *Nature* 391, 481–484. doi: 10.1038/35135

Haenny, P., and Schiller, P. (1988). State dependent activity in monkey visual cortex. single cell activity in v1 and v4 on visual tasks. *Exp. Brain Res.* 69, 245–259. doi: 10.1007/BF00247570

Hayhoe, M., and Ballard, D. (2005). Eye movements in natural behavior. *Trends Cogn. Sci.* 9, 188–194. doi: 10.1016/j.tics.2005.02.009

He, S., Cavanagh, P., and Intriligator, J. (1996). Attentional resolution and the locus of visual awareness. *Nature* 383, 334–337. doi: 10.1038/383334a0

Huk, A. C., and Heeger, D. J. (2000). Task-related modulation of visual cortex. *J. Neurophysiol.* 83, 3525–3536.

Ipata, A. E., Gee, A. L., and Goldberg, M. E. (2012). Feature attention evokes task-specific pattern selectivity in V4 neurons. *Proc. Natl. Acad. Sci. U.S.A.* 109, 16778–16785. doi: 10.1073/pnas.1215402109

Itti, L., and Koch, C. (2001). Computational modeling of visual attention. *Nat. Rev. Neurosci.* 2, 194–203. doi: 10.1038/35058500

Jehee, J. F., Brady, D. K., and Tong, F. (2011). Attention improves encoding of task-relevant features in the human visual cortex. *J. Neurosci.* 31, 8210–8219. doi: 10.1523/JNEUROSCI.6153-09.2011

James, W. (2011). *The Principles of Psychology*. New York, NY: Digireads.com Publishing.

Jazayeri, M., and Movshon, J. A. (2006). Optimal representation of sensory information by neural populations. *Nat. Neurosci.* 9, 690–696. doi: 10.1038/nn1691

Kanwisher, N., and Wojciulik, E. (2000). Visual attention: insights from brain imaging. *Nat. Rev. Neurosci.* 1, 91–100. doi: 10.1038/35039043

Kastner, S., Pinsk, M. A., De Weerd, P., Desimone, R., and Ungerleider, L. G. (1999). Increased activity in human visual cortex during directed attention in the absence of visual stimulation. *Neuron* 22, 751–761. doi: 10.1016/S0896-6273(00)80734-5

Koch, C., and Ullman, S. (1985). Shifts in selective visual attention: towards the underlying neural circuitry. *Hum. Neurobiol.* 4, 219–227.

Kowler, E. (2011). Eye movements: the past 25 years. *Vision Res.* 51, 1457–1483. doi: 10.1016/j.visres.2010.12.014

Lee, D. K., Itti, L., Koch, C., and Braun, J. (1999). Attention activates winner-take-all competition among visual filters. *Nat. Neurosci.* 2, 375–381. doi: 10.1038/7286

Ling, S., Liu, T., and Carrasco, M. (2009). How spatial and feature-based attention affect the gain and tuning of population responses. *Vision Res.* 49, 1194–1204. doi: 10.1016/j.visres.2008.05.025

Liu, T., Fuller, S., and Carrasco, M. (2006). Attention alters the appearance of motion coherence. *Psychon. Bull. Rev.* 13, 1091–1096. doi: 10.3758/BF03213931

Lu, Z.-L., and Dosher, B. A. (1998). External noise distinguishes attention mechanisms. *Vision Res.* 38, 1183–1198. doi: 10.1016/S0042-6989(97)00273-3

Luck, S. J., Chelazzi, L., Hillyard, S. A., and Desimone, R. (1997). Neural mechanisms of spatial selective attention in areas v1, v2, and v4 of macaque visual cortex. *J. Neurophysiol.* 77, 24–42.

Macknik, S. L., King, M., Randi, J., Robbins, A., Teller, Thompson, J., et al. (2008). Attention and awareness in stage magic: turning tricks into research. *Nat. Rev. Neurosci.* 9, 871–879. doi: 10.1038/nrn2473

Martinez, A., Anllo-Vento, L., Sereno, M. I., Frank, L. R., Buxton, R. B., Dubowitz, D. J., et al. (1999). Involvement of striate and extrastriate visual cortical areas in spatial attention. *Nat. Neurosci.* 2, 364–369. doi: 10.1038/7274

Martinez-Trujillo, J., and Treue, S. (2002). Attentional modulation strength in cortical area mt depends on stimulus contrast. *Neuron* 35, 365–370. doi: 10.1016/S0896-6273(02)00778-X

Martinez-Trujillo, J. C., and Treue, S. (2004). Feature-based attention increases the selectivity of population responses in primate visual cortex. *Curr. Biol.* 14, 744–751. doi: 10.1016/j.cub.2004.04.028

Maunsell, J., and Treue, S. (2006). Feature-based attention in visual cortex. *Trends Neurosci.* 29, 317–322. doi: 10.1016/j.tins.2006.04.001

McAdams, C. J., and Maunsell, J. H. (1999). Effects of attention on orientation-tuning functions of single neurons in macaque cortical area v4. *J. Neurosci.* 19, 431–441.

Moore, T., and Fallah, M. (2004). Microstimulation of the frontal eye field and its effects on covert spatial attention. *J. Neurophysiol.* 91, 152–162. doi: 10.1152/jn.00741.2002

Moran, J., and Desimone, R. (1985). Selective attention gates visual processing in the extrastriate cortex. *Science* 229, 782–784. doi: 10.1126/science.4023713

Motter, B. C. (1993). Focal attention produces spatially selective processing in visual cortical areas v1, v2, and v4 in the presence of competing stimuli. *J. Neurophysiol.* 70, 909–919.

Munoz, D. P., Pelisson, D., and Guitton, D. (1991). Movement of neural activity on the superior colliculus motor map during gaze shifts. *Science* 251, 1358–1360. doi: 10.1126/science.2003221

Nakayama, K., and Martini, P. (2011). Situating visual search. *Vision Res.* 51, 1526–1537. doi: 10.1016/j.visres.2010.09.003

Navalpakkam, V., and Itti, L. (2007). Search goal tunes visual features optimally. *Neuron* 53, 605–617. doi: 10.1016/j.neuron.2007.01.018

Nelder, J. A., and Mead, R. (1965). A simplex method for function minimization. *Comp. J.* 7, 308–313.

O'Connor, D., Fukui, M., Pinsk, M., and Kastner, S. (2002). Attention modulates responses in the human lateral geniculate nucleus. *Nat. Neurosci.* 5, 1203–1209. doi: 10.1038/nn957

O'Craven, K. M., Rosen, B. R., Kwong, K. K., Treisman, A., and Savoy, R. L. (1997). Voluntary attention modulates fMRI activity in human MT-MST. *Neuron* 18, 591–598.

Posner, M. I., Snyder, C. R., and Davidson, B. J. (1980). Attention and the detection of signals. *J. Exp. Psychol.* 109, 160–174. doi: 10.1037/0096-3445.109.2.160

Pouget, A., Denève, S., Ducom, J.-C., and Latham, P. E. (1999). Narrow versus wide tuning curves: what's best for a population code? *Neural Comput.* 11, 85–90. doi: 10.1162/089976699300016818

Reynolds, J. H., Alborzian, S., and Stoner, G. R. (2003). Exogenously cued attention triggers competitive selection of surfaces. *Vision Res.* 43, 59–66. doi: 10.1016/S0042-6989(02)00403-0

Reynolds, J. H., Pasternak, T., and Desimone, R. (2000). Attention increases sensitivity of v4 neurons. *Neuron* 26, 703–714. doi: 10.1016/S0896-6273(00)81206-4

Reynolds, J. H., and Heeger, D. J. (2009). The normalization model of attention. *Neuron* 61, 168–185. doi: 10.1016/j.neuron.2009.01.002

Roelfsema, P. R., Lamme, V. A., and Spekreijse, H. (1998). Object-based attention in the primary visual cortex of the macaque monkey. *Nature* 395, 376–381. doi: 10.1038/26475

Saenz, M., Buracas, G. T., and Boynton, G. M. (2002). Global effects of feature-based attention in human visual cortex. *Nat. Neurosci.* 5, 631–632. doi: 10.1038/nn876

Saenz, M., Buracas, G. T., and Boynton, G. M. (2003). Global feature-based attention for motion and color. *Vision Res.* 43, 629–637. doi: 10.1016/S0042-6989(02)00595-3

Saproo, S., and Serences, J. T. (2010). Spatial attention improves the quality of population codes in human visual cortex. *J. Neurophysiol.* 104, 885–895. doi: 10.1152/jn.00369.2010

Schütz, A., Braun, D., and Gegenfurtner, K. (2011). Eye movements and perception: a selective review. *J. Vis.* 11, 1–30. doi: 10.1167/11.5.9

Scolari, M., Byers, A., and Serences, J. T. (2012). Optimal deployment of attentional gain during fine discriminations. *J. Neurosci.* 32, 7723–7733. doi: 10.1523/JNEUROSCI.5558-11.2012

Scolari, M., and Serences, J. T. (2009). Adaptive allocation of attentional gain. *J. Neurosci.* 29, 11933–11942. doi: 10.1523/JNEUROSCI.5642-08.2009

Scolari, M., and Serences, J. T. (2010). Basing perceptual decisions on the most informative sensory neurons. *J. Neurophysiol.* 104, 2266–2273. doi: 10.1152/jn.00273.2010

Serences, J. T., and Boynton, G. M. (2007). Feature-based attentional modulations in the absence of direct visual stimulation. *Neuron* 55, 301–312. doi: 10.1016/j.neuron.2007.06.015

Seriès, P., Latham, P. E., and Pouget, A. (2004). Tuning curve sharpening for orientation selectivity: coding efficiency and the impact of correlations. *Nat. Neurosci.* 7, 1129–1135. doi: 10.1038/nn1321

Sohn, W., Chong, S. C., Papathomas, T. V., and Vidnyánszky, Z. (2005). Cross-feature spread of global attentional modulation in human area mt+. *Neuroreport* 16, 1389–1393. doi: 10.1097/01.wnr.0000174059.57144.62

Spitzer, H., Desimone, R., and Moran, J. (1988). Increased attention enhances both behavioral and neuronal performance. *Science* 240, 338–340. doi: 10.1126/science.3353728

Talsma, D., Doty, T. J., and Woldorff, M. G. (2007). Selective attention and audiovisual integration: is attending to both modalities a prerequisite for early integration? *Cereb. Cortex* 17, 679–690. doi: 10.1093/cercor/bhk016

Tatler, B., Hayhoe, M., Land, M., and Ballard, D. (2011). Eye guidance in natural vision: reinterpreting salience. *J. Vis.* 11, 1–23. doi: 10.1167/11.5.5

Treisman, A., and Gelade, G. (1980). A feature integration theory of attention. *Cogn. Psychol.* 12, 97–136. doi: 10.1016/0010-0285(80)90005-5

Treue, S., and Maunsell, J. H. (1996). Attentional modulation of visual motion processing in cortical areas mt and mst. *Nature* 382, 539–541. doi: 10.1038/382539a0

Treue, S., and Trujillo, J. C. M. (1999). Feature-based attention influences motion processing gain in macaque visual cortex. *Nature* 399, 575–579. doi: 10.1038/21176

Tsotsos, J. K. (1992). On the relative complexity of active vs passive visual-search. *Int. J. Comp. Vis.* 7, 127–141. doi: 10.1007/BF00128132

Verghese, P. (2001). Visual search and attention: a signal detection theory approach. *Neuron* 31, 523–535. doi: 10.1016/S0896-6273(01)00392-0

Verghese, P., Kim, Y.-J., and Wade, A. R. (2012). Attention selects informative neural populations in human v1. *J. Neurosci.* 32, 16379–16390. doi: 10.1523/JNEUROSCI.1174-12.2012

Wang, Z., Stocker, A. A., and Lee, A. D. (2012). "Optimal neural tuning curves for arbitrary stimulus distributions: discrimax, infomax and minimum L_p Loss," in *Advances in Neural Information Processing Systems 25*, 2177–2185. Available online at: http://books.nips.cc/papers/files/nips25/NIPS2012_1077.pdf

Watanabe, T., Sasaki, Y., Miyauchi, S., Putz, B., Fujimaki, N., Nielsen, M., et al. (1998). Attention-regulated activity in human primary visual cortex. *J. Neurophysiol.* 79, 2218–2221.

Williford, T., and Maunsell, J. (2006). Effects of spatial attention on contrast response functions in macaque area V4. *J. Neurophysiol.* 96, 40–54. doi: 10.1152/jn.01207.2005

Wolfe, J. M., Cave, K. R., and Franzel, S. L. (1989). Guided search: an alternative to the feature integration model of visual search. *J. Exp. Psychol. Hum. Percept. Perform.* 15, 4–433. doi: 10.1037/0096-1523.15.3.419

Womelsdorf, T., Anton-Erxleben, K., Pieper, F., and Treue, S. (2006). Dynamic shifts of visual receptive fields in cortical area MT by spatial attention. *Nat. Neurosci.* 9, 1156–1160. doi: 10.1038/nn1748

Womelsdorf, T., Anton-Erxleben, K., and Treue, S. (2008). Receptive field shift and shrinkage in macaque middle temporal area through attentional gain modulation. *J. Neurosci.* 28, 8934–8944. doi: 10.1523/JNEUROSCI.4030-07.2008

Womelsdorf, T., and Fries, P. (2007). The role of neuronal synchronization in selective attention. *Curr. Opin. Neurobiol.* 17, 154–160. doi: 10.1016/j.conb.2007.02.002

Womelsdorf, T., Schoffelen, J.-M., Oostenveld, R., Singer, W., Desimone, R., Engel, A. K., et al. (2007). Modulation of neuronal interactions through neuronal synchronization. *Science* 316, 1609–1612. doi: 10.1126/science.1139597

Yarbus, A. (1967). *Eye Movements and Vision*. New York, NY: Plenum Press. doi: 10.1007/978-1-4899-5379-7

Yeshurun, Y., and Carrasco, M. (1998). Attention improves or impairs visual performance by enhancing spatial resolution. *Nature* 396, 72–75. doi: 10.1038/23936

Zhang, K., and Sejnowski, T. J. (1999). Neuronal tuning: to sharpen or broaden? *Neural Comput.* 11, 75–84. doi: 10.1162/089976699300016809

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. doi: 10.1073/pnas.1100999108

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: 26 September 2013; accepted: 01 March 2014; published online: 26 March 2014.

*Citation: Borji A and Itti L (2014) Optimal attentional modulation of a neural population. Front. Comput. Neurosci. **8**:34. doi: 10.3389/fncom.2014.00034*

This article was submitted to the journal Frontiers in Computational Neuroscience.

frontiers in Computational Neuroscience

HYPOTHESIS AND THEORY
published: 21 April 2015
doi: 10.3389/fncom.2015.00045

On the role of spatial phase and phase correlation in vision, illusion, and cognition

Evgeny Gladilin[1]* *and Roland Eils*[1, 2]

[1] *Division of Theoretical Bioinformatics, German Cancer Research Center, Heidelberg, Germany,* [2] *BioQuant and IPMB, University Heidelberg, Heidelberg, Germany*

Numerous findings indicate that spatial phase bears an important cognitive information. Distortion of phase affects topology of edge structures and makes images unrecognizable. In turn, appropriately phase-structured patterns give rise to various illusions of virtual image content and apparent motion. Despite a large body of phenomenological evidence not much is known yet about the role of phase information in neural mechanisms of visual perception and cognition. Here, we are concerned with analysis of the role of spatial phase in computational and biological vision, emergence of visual illusions and pattern recognition. We hypothesize that fundamental importance of phase information for invariant retrieval of structural image features and motion detection promoted development of phase-based mechanisms of neural image processing in course of evolution of biological vision. Using an extension of Fourier phase correlation technique, we show that the core functions of visual system such as motion detection and pattern recognition can be facilitated by the same basic mechanism. Our analysis suggests that emergence of visual illusions can be attributed to presence of coherently phase-shifted repetitive patterns as well as the effects of acuity compensation by saccadic eye movements. We speculate that biological vision relies on perceptual mechanisms effectively similar to phase correlation, and predict neural features of visual pattern (dis)similarity that can be used for experimental validation of our hypothesis of "cognition by phase correlation."

Keywords: vision research, visual illusions, motion detection, pattern recognition, saccades, acuity, phase correlation, association cortex

Edited by:
Judith Peters,
The Netherlands Institute for Neuroscience, Netherlands

Reviewed by:
Christianne Jacobs,
University of Westminster, UK
Benedikt Zoefel,
Centre National de la Recherche Scientifique, France

Correspondence:
Evgeny Gladilin,
German Cancer Research Center,
Division of Theoretical Bioinformatics,
Im NeuenheimerFeld 580,
69120 Heidelberg, Germany
evgeny.gladilin@gmail.com

Received: *31 October 2014*
Accepted: *30 March 2015*
Published: *21 April 2015*

Citation:
Gladilin E and Eils R (2015) On the role of spatial phase and phase correlation in vision, illusion, and cognition.
Front. Comput. Neurosci. 9:45.
doi: 10.3389/fncom.2015.00045

1. Introduction

Continuous evolution of biological systems implicates a common origin of different functions and mechanisms that emerged as a result of successive modification of one particularly advantageous basic principle. Electrophysiological findings (Hubel and Wiesel, 1968) and psychophysical experiments (Campbell and Robson, 1968) indicate that visual system relies on the basic principle of frequency domain transformation of the retinal image in visual cortex which was initially believed to resemble a crude Fourier transformation (Graham, 1981). Even though, more recent mathematical models of sparse image coding revised the assumption of global Fourier transformation in favor of locally supported Gabor- (Marcelja, 1980), Wavelet- Mallat, 1989, Wedge-, Ridge- or Curvelet-functions (Donoho and Flesia, 2001), the concept of neural image representation in the frequency domain by phase and amplitude remained valid.

Since pioneering works of Hubel and Wiesel (1962, 1968), Campbell and Robson (1968), Blakemore and Campbell (1969), Blakemore et al. (1969), and Thomas et al. (1969) it is known that different groups of neurons in the visual cortex show selective response to spatial-temporal characteristics of visual stimuli and operate as spatially organized filters (receptive fields) that extract particular image features (i.e., spatial frequency, orientation) within a certain range (bandwidth) of their sensitivity. Numerous subsequent studies dealt with experimental investigation and theoretical modeling of visual receptive fields and analysis of their amplitude-transfer (ATF) and phase-transfer functions (PTF). The existing body of evidence resulting from four decades of research on this field includes

- existence of frequency-selective V1 neurons operating as bandpass filters (Graham, 1989; De Valois and De Valois, 1990),
- coding of phase information using quadrature pairs of bandpass filters (Pollen and Ronner, 1983),
- odd-/even-symmetric filters in visual cortex (Morrone and Owens, 1987),
- linear ATF and PTF of simple striatic neurons (Hamilton et al., 1989),
- computation of complex-valued products in V1 neurons (Ohzawa et al., 1990),
- computation of magnitudes (energies) in complex V1 cells as a sum of squared responses of simple V1 cells (Adelson and Bergen, 1985),
- divisive normalization of neuronal filter responses (Heeger, 1992; Schwartz and Simoncelli, 2001),
- motion detection (Fleet and Jepson, 1990; Nishida, 2011),
- edge detection (Kovesi, 2000; Henriksson et al., 2009),
- stereoscopic vision (Fleet, 1994; Fleet et al., 1996; Ohzawa et al., 1997),
- 3D shape perception (Thaler et al., 2007),
- assessment of pattern similarity (Sampat et al., 2009; Zhang et al., 2014),
- triggering of diverse visual illusions (Popple and Levi, 2000; Backus and Oru, 2005).

Altogether, these findings support the concept of neural transformation of retinal images into frequency domain characteristics (i.e., phase and amplitude) that, in turn, serve as an input for subsequent higher-order mechanisms and functions of visual perception and cognition.

Despite recent advances in understanding of the overall topology and hierarchy of visual cortex (Riesenhuber, 2005; Poggio and Ullman, 2013), little is known yet about the underlying wiring schemes of phase/amplitude information processing in visual cortex. In particular, the observation that small cells of V1 show phase-sensitivity (Pollen and Ronner, 1981) while complex cells do not (De Valois et al., 1982) lead to controversial discussion about the role of spatial phase in visual information processing (Morgan et al., 1991; Bex and Makous, 2002; Shams and Malsburg, 2002; Hietanen et al., 2013).

In what follows we aim to address the following basic questions:

- What are the driving forces behind the evolutionary development of biological vision?
- What properties of spatial phase (further in this manuscript denoted as phase) make it an important feature for visual information processing?
- What is the origin of various phase-related visual phenomena including illusions of apparent motion, stereograms and virtual image context?
- How can phase information be used for motion detection and (dis)similarity cognition, and how can theoretical models be evaluated experimentally?

Our manuscript is organized as follows. First, we recapitulate the role of environmental constraints in development of biological vision in course of evolution. We review theoretical properties of phase using an extension of the Fourier phase correlation technique and demonstrate how phase information can be used for edge enhancement, motion detection, and pattern recognition. We show that saccadic strategy of image sampling naturally emerges within this concept as an algorithmic solution which improves the confidence of visual pattern discrimination and recognition. Further, we apply the concept of phase shift and correlation to analysis of different visual illusions and hypothesize about involvement of phase-based mechanisms in perception of motion and visual pattern (dis)similarity. In conclusion, we make suggestions for experimental evaluation of our theoretical predictions.

2. Invariants of Ecological Environment and Evolution of Vision

The evolutionary principle implies that remarkable abilities of biological vision result from adaptation of species to the environmental constraints that ancestors had to cope with in the past. It is generally recognized that progressive sophistication of vision is driven toward more efficient representation, processing and, probably, also modeling of the physical reality which stands behind the retinal images (Walls, 1962; Marr, 1982; Hyvärinen and Hoyer, 2001; Graham and Field, 2006). In addition to the basic optosensory function, the core tasks of visual perception in macroscopic organisms include orientation in the physical environment, which premises ability to detect obstacles and relative motion, as well as recognition of essential patterns related to food, threat and communication. Further, we recollect that biological organisms are composed of condensed matter and have to mainly take care about the objects of the physical world that also have rigid constitution and conservative shape. In contrast, highly deformable media such as gasses and liquids are biologically neutral which implicates that perception of non-rigid transformations did not fall under the early evolutionary pressure. Important is the notion that visual perception of rigid bodies with a preserved shape has to be independent on relative spatial position and orientation which means that it has to rely on some invariants (Ito et al., 1995; Booth and Rolls, 1998; Palmeri and Gauthier, 2004; Lindeberg, 2013) that are not given *per se* but have to be derived by subsequent processing of the raw retinal image. As a dimensionless quantity, phase bears topological information independently on the level of illuminance and contrast. Affine transformations in the image domain do not change the relative phase structure, but merely shift it as a whole. These properties of phase are of advantage for survival of the fittest and can be assumed to be "discovered" in course of the evolution of biological vision. Different features of visual perception emerge at evolutionarily distant time points and, thus, rely on different intrinsic invariances. Early forms of life are originated in the marine environment where movements are slowed down by viscosity of water, effects of gravitation are diminished and changes in the relative spatial position and orientation are more probable as it is the case in terrestrial environment with its stable gravitational axis and unresisting atmosphere. The ability to recognize abstract shapes (i.e., animal silhouettes) independently on their relative motion, orientation, and distance was essential to survival of species and probably originated already with the first marine animals. However, the translation-, rotation-, scaling-independent (i.e., TRS-invariant) perception of abstract shapes (Gladilin, 2004) does not apply to all kinds of visual stimuli. A prominent example of dependency of visual perception on changing environmental constraints is the Thatcher-Illusion, which consists in poor recognition of upside-down faces (Psalta et al., 2014). Comparative experiments with different primates demonstrate that perception of facial expression is a relatively new feature in biological vision (Weldon et al., 2013). Sensitivity of human face perception to rotations has obviously to do with the fact that the neuronal machinery of face recognition is relatively new cognitive feature which emerged in the terrestrial environment where primates encountered each other predominantly in the upright posture. In general, visual illusions can be attributed to optical stimuli that mislead evolutionarily conserved mechanisms of visual information processing based on a built-in knowledge of properties of the physical world (Ramachandran and Anstis, 1986). The ability to irritate or escape common cognitive schemes is, in turn, of evolutionary advantage. The fact that many animals use camouflage patterning, swarm motion or body morphing as a reliable survival strategy indicates that repetitive patterns and non-TRS transformations represent a principle challenge for biological vision which is evolutionarily predetermined to rely on TRS-invariants of the condensed matter world, see **Figure 1**.

3. The Role of Phase from the Viewpoint of Computer Vision

In this section, we elucidate the role of phase information for detection of image motion and pattern recognition from the viewpoint of computer vision. Readers who are not familiar with Fourier analysis may skip over math-intensive parts that will be concluded subsequently.

3.1. Image Representation in Spatial and Frequency Domains

In spatial domain, 2D images are represented by a matrix $A_{x,y}$ of $N \times M$ scalar intensity values on an Euclidian image raster ($x \in [0, N-1],\ y \in [0, M-1]$). Complex Fourier transformation maps an image $A_{x,y}$ onto the complex frequency domain $\alpha_{u,v}$:

$$\alpha_{u,v} = \mathcal{F}(A_{x,y}) = Re(\alpha_{u,v}) + i\, Im(\alpha_{u,v}) \tag{1}$$

or in a more explicit form for a discrete 2D case:

$$\alpha_{u,v} = \frac{1}{\sqrt{MN}} \sum_{x=0}^{N-1} \sum_{y=0}^{M-1} A_{x,y}\, \mathbf{e}^{-2\pi i (\frac{ux}{N} + \frac{vy}{M})}. \tag{2}$$

The inverse Fourier transformation mapping $\alpha_{u,v}$ onto the spatial domain is given by

$$A_{x,y} = \mathcal{F}^{-1}(\alpha_{u,v}) = \frac{1}{\sqrt{MN}} \sum_{u=0}^{N-1} \sum_{v=0}^{M-1} \alpha_{u,v}\, \mathbf{e}^{2\pi i (\frac{xu}{N} + \frac{yv}{M})}. \tag{3}$$

Further, we recollect that the complex conjugate of $\alpha_{u,v}$ is defined as $\alpha^*_{u,v} = Re(\alpha_{u,v}) - i\, Im(\alpha_{u,v})$.

3.2. Importance of Phase and Amplitude: Theoretical Perspective

The relative importance of Fourier phase and amplitude for retrieval of structural image features has been debated in several previous works (Oppenheim and Lim, 1981; Lohmann et al., 1997; Ni and Huo, 2007). The basic notion is that the phase bears topological information about image edges whereas amplitude encodes image intensity. To demonstrate the effect of amplitude and phase distortion, we perform reconstruction of the original image from amplitude-only and phase-only of its Fourier transform, see **Figure 2**. Here, the amplitude-only reconstruction (**Figure 2** (middle)) is computed as the Fourier inverse of the following amplitude-preserving and phase-eliminating transformation:

$$\begin{aligned} Re(\alpha_{u,v}) &\rightarrow \left(Re(\alpha_{u,v})^2 + Im(\alpha_{u,v})^2\right)^{1/2}, \\ Im(\alpha_{u,v}) &\rightarrow 0\,, \end{aligned} \tag{4}$$

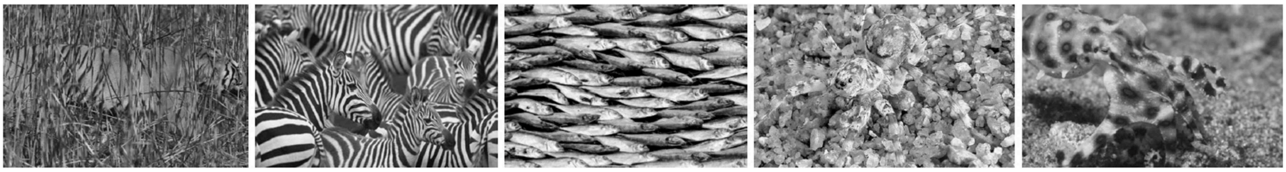


FIGURE 1 | Repetitive patterns, swarm motion, and body morphing disrupt detection of unique invariant features (i.e., rigid animal silhouettes). Examples of natural images are acquired from public Creative Commons sources (http://search.creativecommons.org/).

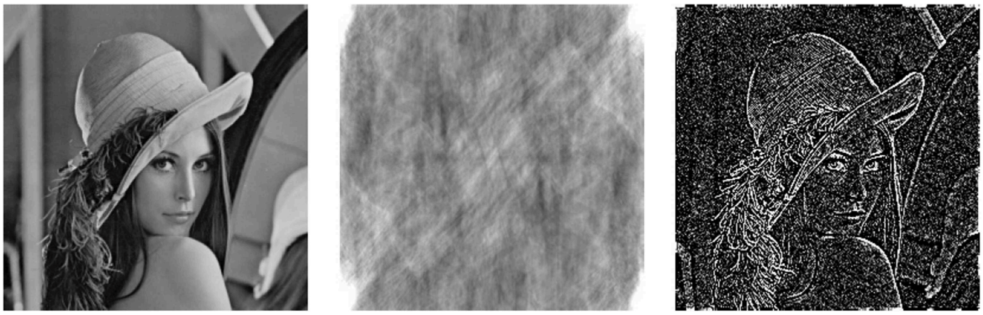


FIGURE 2 | Comparison of the effects of amplitude and phase distortion on image reconstruction. From left to right: the original Lenna image vs. amplitude-only and phase-only image transforms. The phase-only transformation works as an edge-enhancing filter resembling the Marr's Primal Sketch (Marr, 1982).

and the phase-only reconstruction (**Figure 2** (right)) is calculated as the Fourier inverse of the following phase-preserving and amplitude-normalizing transformation:

$$\begin{aligned} Re(\alpha_{u,v}) &\rightarrow \frac{Re(\alpha_{u,v})}{\left(Re(\alpha_{u,v})^2 + Im(\alpha_{u,v})^2\right)^{1/2}}\,, \\ Im(\alpha_{u,v}) &\rightarrow \frac{Im(\alpha_{u,v})}{\left(Re(\alpha_{u,v})^2 + Im(\alpha_{u,v})^2\right)^{1/2}}\,. \end{aligned} \tag{5}$$

This example demonstrates that the relative phase appears to be more significant for retrieval of cognitive image features (i.e., edges) that get completely lost in the amplitude-only transformation. Remarkably, the amplitude-normalizing phase-only reconstruction seem to effectively work as an edge-enhancing filter which generates a feature-preserving image sketch resembling the Marr's concept of the Primal Sketch generation in visual cortex (Marr, 1982).

3.3. Detection of Uniform Image Motion using Phase Correlation

The Fourier phase correlation (*PC*) is a powerful technique which has been originally developed for detection of affine image transformations such as uniform translational motion, rotation and/or scaling (De Castro and Morandi, 1987; Reddy and Chatterji, 1996). Phase correlation between two images $A_{x,y}$ and $B_{x,y}$, is computed as a Fourier inverse of the normalized cross-power spectrum (*CPS*):

$$PC_{x,y} = \mathcal{F}^{-1}(CPS_{u,v})\,, \tag{6}$$

where

$$CPS_{u,v} = \frac{\alpha_{u,v}\,\beta^*_{u,v}}{|\alpha_{u,v}\,\beta^*_{u,v}|} \tag{7}$$

and

$$\begin{aligned} \alpha_{u,v} &= \mathcal{F}(A_{x,y}) \\ \beta_{u,v} &= \mathcal{F}(B_{x,y}) \end{aligned} \tag{8}$$

are the complex Fourier transforms of the images $A_{x,y}$ and $B_{x,y}$, respectively. According to the Fourier shift theorem, relative displacement (Δx, Δy) between two identical images, i.e.,

$$B_{x,y} = A_{x-\Delta x, y-\Delta y}\,, \tag{9}$$

corresponds to phase-shift in the frequency domain

$$\beta_{u,v} = \mathbf{e}^{-2\pi i\varphi}\,\alpha_{u,v}\,, \tag{10}$$

where $\varphi = (\frac{u\Delta x}{N} + \frac{v\Delta y}{N})$. Consequently, the cross power spectrum between two identical images shifted with respect to each other in the spatial domain describes the phase-shifts of the entire Fourier spectrum in the frequency domain:

$$CPS_{u,v} = \frac{\alpha_{u,v}\,\mathbf{e}^{2\pi i\varphi}\alpha^*_{u,v}}{|\alpha_{u,v}\,\mathbf{e}^{2\pi i\varphi}\alpha^*_{u,v}|} = \mathbf{e}^{2\pi i\varphi}. \tag{11}$$

For two identical images with the relative spatial shift (Δx, Δy), the inverse Fourier integral of Equation (11), i.e., the phase correlation Equation (6), exhibits a single singularity at the point ($x = \Delta x$, $y = \Delta y$) and is given by

$$PC_{x,y} = \delta(x - \Delta x, y - \Delta y)\,. \tag{12}$$

Thus, phase correlation of two identical images has a single maximum-peak which coordinates in the spatial domain yield the relative image translation[1] ($x = \Delta x, y = \Delta y$), see **Figure 3A**.

3.4. Phase Correlation in the Presence of Noise

In the presence of additive statistical or structural noise, the cross power spectrum between two non-identical images takes the form:

$$CPS_{u,v} = \mathbf{e}^{2\pi i\varphi} + \varepsilon_{u,v}\,, \tag{13}$$

where $\varepsilon_{u,v}$ is a frequency-dependent perturbation-term whose properties depend on particular type of image differences. Consequently, the inverse Fourier integral of Equation (13), i.e., the phase correlation between two non-identical images, becomes different from the Dirac delta peak of the identical image shift Equation (12):

$$PC_{x,y} = \mathcal{F}^{-1}\left(\mathbf{e}^{2\pi i\varphi} + \varepsilon_{u,v}\right) \neq \delta(x - \Delta x, y - \Delta y)\,, \tag{14}$$

[1] Reformulation of phase correlation in polar coordinates results in detection of the image scaling and rotation (Reddy and Chatterji, 1996).

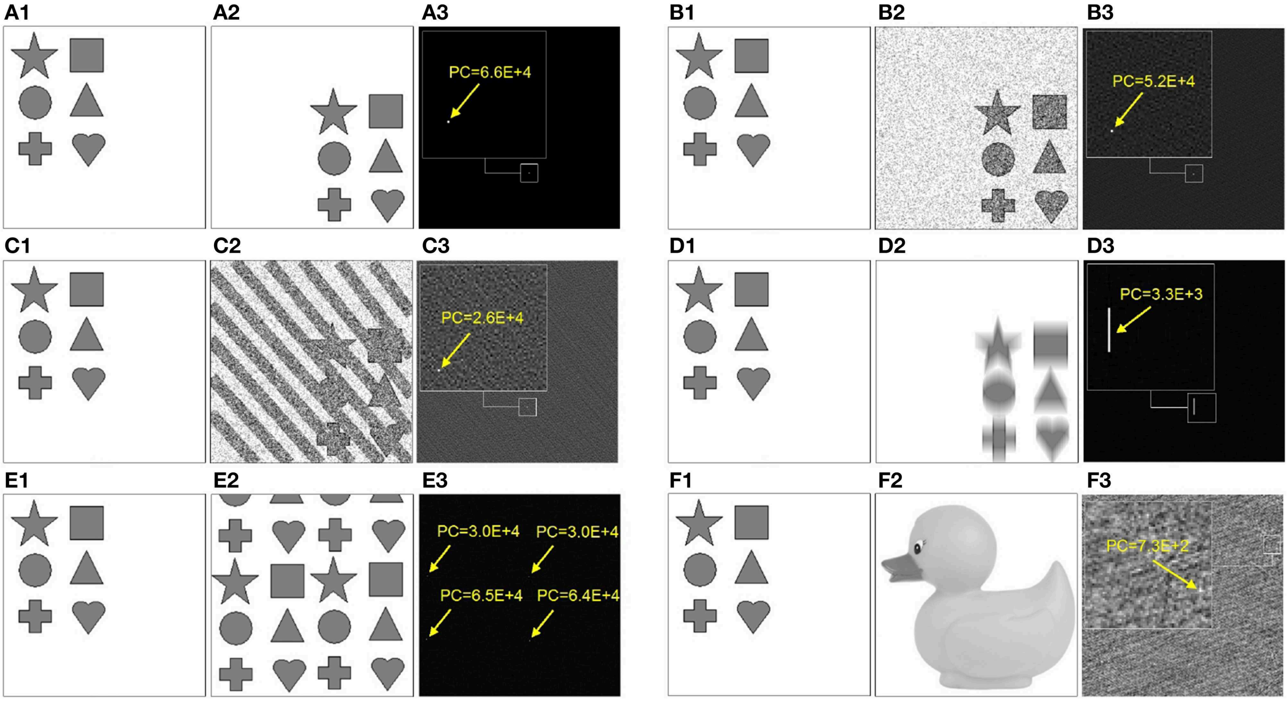


FIGURE 3 | Examples of phase correlation (right column) between the source (left column) and the target image (middle column). Target images **(A2-E.2)** represent the following transformations of the source image: **(A2)** uniform displacement, **(B2)** uniform displacement superimposed with 70% statistical noise, **(C2)** uniform displacement superimposed with 70% statistical and structural noise, **(D2)** uniform displacement superimposed with 20-pixel Y-motion-blur, **(E2)** superposition of four different uniform displacements (i.e., 4× fold repetition). **(F)** shows phase correlation between two significantly different images. Arrows point to the location of the absolute maximum peak of the *PC*. Visualization of the entire *PC* is performed using the following grayscale mapping: $PC_{x,y} \rightarrow 255(PC_{x,y} - \mathrm{MIN}(PC_{x,y}))/(\mathrm{MAX}(PC_{x,y}) - \mathrm{MIN}(PC_{x,y}))$.

which manifests in flattening of the maximum peak and overall more noisy *PC*, see **Figures 3B,C**. However, as long as the target pattern do not exhibit similarities with the background structures, phase correlation between two images remains a single-peak distribution. Remarkably, even a significant structural distortion does not affect the detection of the target pattern within the noisy visual scene, see **Figure 3C**. This example demonstrates that the height of maxima and the overall shape of the *PC* distribution can serve as quantitative characteristics of image (dis)similarity, i.e., the more sharp (Dirac-like) is the *PC* distribution, the more similar are the structures in the underlying images. An increasingly dispersed *PC* distribution indicates lower image similarity.

In the case of non-affine image transformations, phase correlation loses its exceptional properties and becomes a multi-peak distribution. **Figure 3D** shows the phase correlation of the original image with its blurred and displaced copy. Uncertainty of the 20-pixel Y-motion-blur applied in this example reflects in the horizontal line of peaks in *PC* that correspond to possible alignments between the original image with its transformed copy.

If the target pattern is multi-present or exhibits structural similarity with the surrounding structures, multiple peaks occur in *PC*. **Figure 3E** shows phase correlation between the target pattern and the image containing its four displaced copies. Finding the right correspondence in such visual scene becomes difficult or impossible. Camouflage textures and behavioral strategies of swarm animals generate repetitive patterns that irritate cognitive mechanisms of predators based on detection of unique target features, see **Figure 1**.

With increasing structural differences between each two images, *PC* becomes a random distribution with the significantly lower maximum peaks, see **Figure 3F**.

3.5. Phase Correlation in the Case of Non-Uniform Image Motion

Non-uniform motion means that displacements of image pixels differ in directions and/or magnitude. Consider time-series of images $A_{x,y}(t)$ that are composed of two non-uniformly moving regions:

$$A_{x,y}(t) = P_{x,y}(t) + B_{x,y}(t)\,, \tag{15}$$

where $P_{x,y}$ stands for a particular image pattern which has to be tracked in consecutive time steps, and $B_{x,y}$ is the background region. Let $P_{x,y}$ and $B_{x,y}$ in the subsequent time step $A_{x,y}(t+1)$ undergo different translations:

$$A_{x,y}(t+1) = P_{x,y}(t+1) + B_{x,y}(t+1)\,, \tag{16}$$

where

$$P_{x,y}(t+1) = P_{x+\Delta x_p, y+\Delta y_p}(t)\,,$$
$$B_{x,y}(t+1) = B_{x+\Delta x_b, y+\Delta y_b}(t)\,. \tag{17}$$

Considering the linearity of Fourier transformation, one obtains for $\mathcal{F}\left(A_{x,y}(t)\right)$ and $\mathcal{F}\left(A_{x,y}(t+1)\right)$

$$\begin{aligned}\alpha_{u,v}(t) &= \rho_{u,v} + \beta_{u,v}\\ \alpha_{u,v}(t+1) &= \mathbf{e}^{-2\pi i\varphi}\,\rho_{u,v} + \mathbf{e}^{-2\pi i\psi}\,\beta_{u,v}\,,\end{aligned} \tag{18}$$

where $\varphi = (\frac{u\Delta x_p}{N} + \frac{v\Delta y_p}{N})$ and $\psi = (\frac{u\Delta x_b}{N} + \frac{v\Delta y_b}{N})$, respectively. Consequently, the cross power spectrum between $A_{x,y}(t)$ and $A_{x,y}(t+1)$ takes the form

$$\begin{aligned}CPS_{u,v} = &\frac{\alpha_{u,v}(t)\,\alpha^*_{u,v}(t+1)}{|\alpha_{u,v}(t)\,\alpha^*_{u,v}(t+1)|} = \frac{1}{|\alpha_{u,v}(t)\,\alpha^*_{u,v}(t+1)|}\\ &(\rho_{u,v}\,\mathbf{e}^{2\pi i\varphi}\,\rho^*_{u,v} + \rho_{u,v}\,\mathbf{e}^{2\pi i\psi}\,\beta^*_{u,v} +\\ &\beta_{u,v}\,\mathbf{e}^{2\pi i\varphi}\,\rho^*_{u,v} + \beta_{u,v}\,\mathbf{e}^{2\pi i\psi}\,\beta^*_{u,v})\end{aligned} \tag{19}$$

or in a more compact form

$$CPS = CPS^p_{p'} + CPS^p_{b'} + CPS^b_{p'} + CPS^b_{b'}\,, \tag{20}$$

where CPS^*_* denote self- and cross-correlations between the Fourier transforms of the pattern and background regions in two consecutive time steps, respectively. Primed indexes are introduced to distinguish Fourier transforms of previous $(t:\ p, b)$ and subsequent $(t+1:\ p', b')$ time steps. By applying the inverse Fourier transformation to Equation (20), one obtains the phase correlation between $A(t)$ and $A(t+1)$:

$$PC = \mathcal{F}^{-1}(CPS) = PC^p_{p'} + PC^p_{b'} + PC^b_{p'} + PC^b_{b'}\,. \tag{21}$$

3.6. Saccades-Enhanced Phase Correlation

Phase correlation between two non-uniformly shifted image regions Equation (21) contains four terms:

- self-correlation of the target pattern ($PC^p_{p'}$),
- self-correlation of the background region ($PC^b_{b'}$) and
- two cross-correlation terms ($PC^p_{b'}$, $PC^b_{p'}$).

In order to detect the shift of the target pattern P, $PC^p_{p'}$ has to become the most dominant term of the total PC. Obviously, this condition is not automatically fulfilled,—other terms may have stronger weight in Equation (21). If the pattern and background regions do not exhibit similarities, i.e., if the pattern P is uniquely present in the image, cross-correlation terms ($PC^p_{b'}$ and $PC^b_{p'}$) should be smaller in comparison to self-correlation terms ($PC^p_{p'}$ and $PC^b_{b'}$). Thus, the major difficulty for detection of the target image pattern is caused by self-correlation of the background region ($PC^b_{b'}$) which properties are *a priori* unknown. Obviously, a single-step phase correlation between two images is not sufficient for detection of a particular image region. In order to maximize the weight of $PC^p_{p'}$ and, correspondingly, to minimize the weight of other terms in Equation (21), one can construct a cumulative phase correlation by iteratively composing PC between the (fixed) target pattern with differently shifted background. Due to formal similarity of such strategy with back-and-forth image sampling by saccadic eye movements (see **Figure 4**), we termed this procedure saccades-enhanced phase correlation (Gladilin and Eils, 2009). To show why this strategy appears to be promising, we write the average phase correlation of N recombinations between the target pattern and non-uniformly shifted background images:

$$\overline{PC} = \frac{1}{N}\sum_{i=1}^{N} PC_i = PC^p_{p'} + PC^p_{b'} + \frac{1}{N}\sum_{i=1}^{N} PC^{b_i}_{p'} + \frac{1}{N}\sum_{i=1}^{N} PC^{b_i}_{b'}\,. \tag{22}$$

Since first two terms in Equation (22) are independent on background variations (b_i), their absolute values remain unchanged. Further, it can be shown that the last two terms decrease with increasing N, and, thus, their weight in the average phase correlation can be arbitrarily decreased after sufficiently high number of saccadic iterations $N >> 1$. Without providing a precise

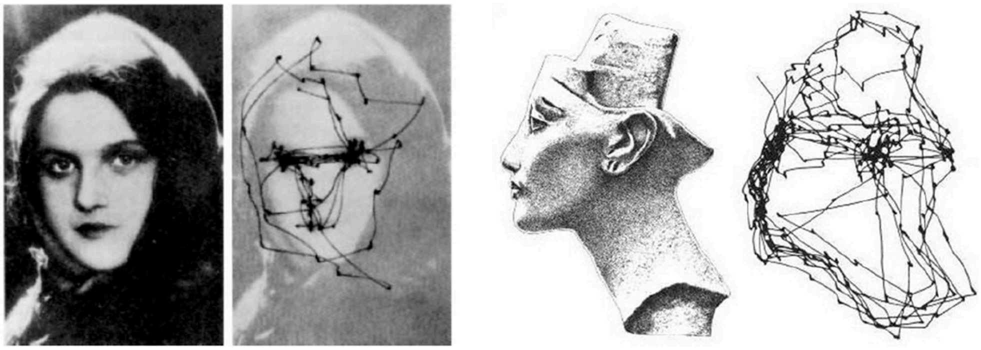


FIGURE 4 | Examples of saccadic eye movements from Yarbus (1967). Left the eyes of the observer exhibit remarkable back-and-forth movements between different regions of interest (i.e., eyes, mouth) and the image background. **Right** saccadic trajectories seem to follow the shape contours and edges.

proof, we can give the following plausible comment: for different shifts of the background region, positions of maxima in cumulative phase correlation differ as well. Consequently, the sum over different b_i remains bounded, and the average value of the last two terms in Equation (22) decreases as N^{-1}, i.e., $\lim_{N\to\infty}\left(\frac{1}{N}\sum_{i=1}^{N} PC^{b_i}_{b'}\right) \rightarrow 0$. As a result of saccadic image composition, self-correlation of the target pattern $PC^{p}_{p'}$ becomes the most dominant term and the shift of P can be determined from the coordinate of the absolute maximum of Equation (22).

The less structured is the target pattern and the more similar it is to the image background, the more difficult becomes the virtual separation of target and background regions using saccades-enhanced phase correlation. Consequently, analysis of poorly structured visual scenes requires more saccadic iterations for detection and recognition of the target pattern. Remarkably, experimental findings seem to confirm this theoretical prediction: the strategy of saccades by observation of unstructured textural images exhibits increasing frequency of target-background eye movements (He and Kowler, 1992).

3.7. Consideration of Visual Acuity

The foveal and peripheral areas of the retinal image are known to exhibit significant differences in acuity that have to be considered by construction of Fourier transforms and phase correlations of target and surrounding images. With approximately 3° of high-acuity foveal cone-projection (Osterberg, 1935), the observer's eye can sharply resolve only an area with the cross-section dimension of $D \approx 0.1\,L$, where L denotes the distance from observer to the focus plane. For a $L = 50$ cm far computer screen, it makes a $D = 5$ cm wide spot. The remaining peripheral area is progressively blurred with the distance from the focus. Consequently, a more natural representation of the retinal and higher-lever neural images is the composition of the central pattern surrounded by the low-pass smoothed periphery. For calculation of saccades-enhanced phase correlation this, in turn, means that not only the position of the focus but also spectral characteristics of the central and peripheral areas have to be appropriately filtered anew for each saccadic fixation image. Repetitive target-background sampling by saccades will, obviously, lead to enhancement of small details (i.e., high-frequent components) of more frequently focused regions and low-pass smoothing of less frequently sampled, peripheral areas. As a consequence, one can expect saccadic analysis to better discriminate images that show distinctive spectral differences between central and peripheral areas. Visual examination of images with similar spectral characteristics of pattern and background regions can be, in turn, associated with intensification of back-and-forth saccadic eye movements.

4. Psychophysical Evidence of Phase Involvement in Visual Information Processing

In this section, we review some psychophysical findings indicating the involvement of phase in visual information processing and analyze them from the perspective of theoretical concepts of phase-based motion and pattern detection.

4.1. Importance of Phase and Amplitude: Psychophysical Perspective

From theoretical considerations in Section 3.2, phase appears to be more essential for retrieval of structural information than amplitude. Psychophysical findings in Freeman and Simoncelli (2011) and Zhang et al. (2014) suggest, however, a combined phase-amplitude mechanism of pattern perception with higher weight of phase information near the fixation point and increasing importance of amplitude on the periphery of the visual field. On the other hand, one should consider that conscious fixations inhibit saccades which results in progressive low-pass blurring of peripheral image. Unconstrained image observation is always associated with saccadic eye movements that acquire high-frequency phase information from different image areas and, thus, substantially increase the real weight of phase information in image perception and (re)cognition.

4.2. On the Role of Phase and Saccades in Visual Illusions

Seemingly different visual illusions have a common feature to be triggered by coherently phase-shifted repetitive patterns. Below we briefly review three groups of visual illusions[2] that generate effects of (i) virtual depth (Tyler and Clarke, 1990), (ii) apparent motion (Kitaoka and Ashida, 2003), and (iii) non-local image tilt (Popple and Levi, 2000). Tight resemblance in stimulus configuration of different visual illusions has been supposed in previous works (Kitaoka, 2006). Though, a unified concept of underlying neural mechanisms that drive different perceptual illusions is still missing.

4.2.1. Virtual Depth Illusions

Stereogram images such as shown in **Figure 5** cause perceptual illusions of virtual depth and hidden 3D content. Stereograms are composed of repetitive patterns which retinal projections in the left and right eyes exhibit a relative spatial shift in the image domain and a corresponding phase-shift in the frequency domain. Accordingly, two basic models of binocular disparity based on position- and phase-shift receptive fields have been discussed in the literature in the last two decades (Arndt et al., 1995; Fleet et al., 1996; Ohzawa et al., 1997; Parker and Cumming, 2001; Chen and Qian, 2004; Goutcher and Hibbard, 2014). Anzai et al. (1997) conclude that "binocular disparity is mainly encoded through phase disparity." Fleet (1994) suggests a model of binocular disparity computation using the Local Weighted Phase Correlation which combines the features of phase-shift and phase correlation approaches. If phase correlation is, in fact, involved in binocular disparity calculation, the underlying neural mechanisms of virtual depth detection can be expected to depend on a certain threshold of neuronal activity, i.e., the strength of phase correlation, which, in turn, should be dependent on structural image properties. In particular, as we have seen above one can expect that structured (i.e., edge-rich, phase-congruent) patterns

[2] All examples of visual stimuli were taken from the "Illusion Pages" of A. Kitaoka http://www.psy.ritsumei.ac.jp/akitaoka/cataloge.html.

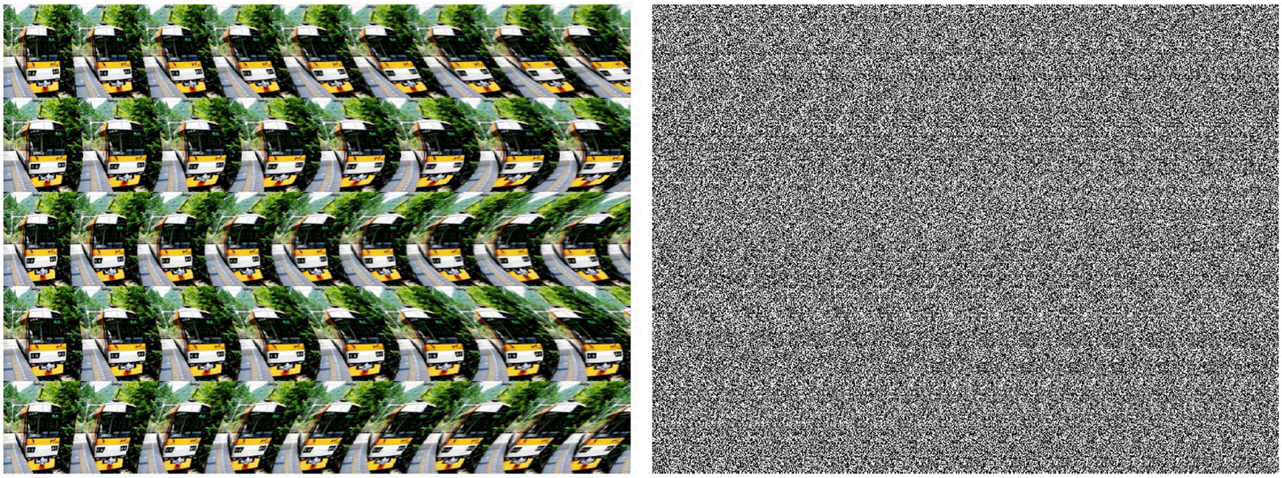


FIGURE 5 | Examples of virtual depth illusions (stereograms) based on structured (left) and diffuse textural (right) patterns (courtesy A. Kitaoka).

such as shown in **Figure 5** (left) produce stronger phase correlation signals and, thus, trigger virtual depth illusions easier re. faster than diffuse textural pattern such as **Figure 5** (right). Further experimental investigations are required to test this pure theoretical prediction.

4.2.2. Apparent Motion Illusions

Apparent motion illusions induce perception of dynamic image changes while observing static visual stimuli. Notably, the intensity of apparent motion illusions depends on spectral characteristics (i.e., low/high frequent image content) and the relative phase-shift of repetitive patterns.

4.2.2.1 The Rotating Snake

patterns from Kitaoka and Ashida (2003) induce a remarkably strong illusion of apparent rotational motion, see **Figures 6A,B**. The low-pass smoothed Rotating Snake in **Figures 6C,D** exhibit a reduced intensity of apparent rotational motion. Backus and Oru (2005) explain emergence of illusory motion of the Rotating Snakes by the difference in the temporal response of visual neurons to low- and high-contrast. This difference leads to misinterpretation of the temporal phase-shift as a spatial phase-shift (“phase advance”) at high contrast. The effect of low-pass smoothing, authors attribute to reduction of differences between high- and low-contrast regions. Recent findings indicate that signals of illusory motion in V1 and MT cortical areas can be also triggered by update of the retinal image as a result of saccadic eye movements or blinkers (Conway et al., 2005; Troncoso et al., 2008; Otero-Millan et al., 2012; Martinez-Conde et al., 2013). Consequently, conscious suppression of saccades inhibits illusions of apparent motion that are based on phase-advancing contrast patterns. To dissect the structural principle of the Rotating Snake in more detail, we performed its polar-to-rectangle transformation into the Translating Snake, see **Figures 6E–H**. This transformation changes the relative spatial orientation of repetitive patterns while preserving their local contrast structure. We observe that a pair of parallel Translating Snake patterns does not induce any significant perceptual effects, see **Figures 6E,F**. In contrast, antiparallel Translating Snakes patterns generate a weak illusion of translational motion, see **Figures 6G,H**. From this observation, we conclude that phase advancement due local contrast gradient is required but not sufficient for generation of apparent motion illusion. The sufficient condition consists in different spatial orientation of repetitive motion patterns: equally oriented motion patterns of the Translating Snake do not induce any illusory motion, while non-uniformly organized contrast gradients of the Rotating Snake do, see **Figures 6I,J**. Thus, we conclude that apparent motion signals are triggered not only by phase advancement at high contrast alone but by the difference in phase advancement between each two image regions subsequently fixated by saccades.

4.2.2.2 The Anomalous Motion

from Kitaoka (2006) is another example of apparent motion illusion which is induced by contrarily oriented contrast-gradient patterns, see **Figure 7** (left). In **Figure 7** (right), central and peripheral contrast-gradient patterns were aligned in the same direction. As a result, the illusion of apparent motion disappears. Only the combination of patterns with contrarily oriented contrast-gradients (i.e., the relative phase shift) is capable to generate a stable illusion of apparent relative motion, see **Figure 7** (left). Similar to the Rotation Snake, the Anomalous Motion illusion requires saccadic eye movements. Suppression of saccades by conscious point fixation stops the illusion of apparent motion.

4.2.3. Non-Local Tilt Illusion.

Figure 8 shows the virtual tilt illusion from Popple and Levi (2000) and Popple and Sagi (2000) which seems to be triggered without local cues. The particularity of this stimulus consists in a way it is constructed by horizontal lines of patterns that exhibit a relative vertical phase-shift. Consequently, the horizontal lines appear to have a vertical tilt which direction depends on the sign of the phase-shift. Based on our previous analysis of motion

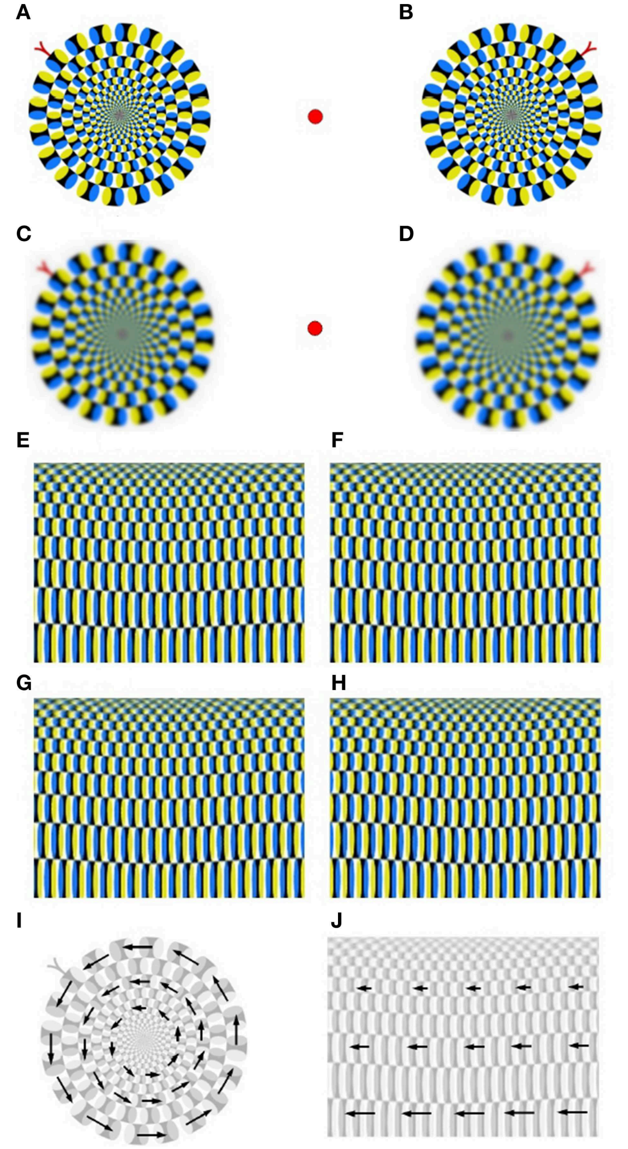


FIGURE 6 | **Apparent motion illusions.** **(A,B)** A pair of the Rotating Snake patterns from Kitaoka and Ashida (2003). **(C,D)** The low-pass filtered Rotating Snakes exhibit slower rotation. **(E,F)** Parallel patterns of the polar-to-rectangle transformation of the Rotating Snake, i.e., the Translating Snake, does not produce any motion illusions. **(G,H)** Antiparallel patterns of the Translating Snake generates a weak illusion of apparent translational motion. **(I,J)** Visualization of the Rotating and Translating Snake pattern shows that motion elements of the Rotating Snake exhibit a relative phase-shift to each other, while the Translating Snake elements are parallel and do not have any relative phase shift.

illusions, we presume that also the virtual tilt illusion is driven by saccadic eye motions along the horizontal lines of patterns. Consequently, the virtual tilt illusion is, nevertheless, based on local cues that are established by successive saccadic fixations.

Another puzzling property of this stimulus is the dependency of the tilt intensity on spectral image characteristics. Remarkably, the low-pass smoothed stimulus seems to exhibit stronger tilt as the unsmoothed version with high-frequent components. One possible explanation for this observation is that phase correlation of low-pass smoothed patterns results in a wide and blurry shift signal, cf. **Figure 3**. Another hypothetic assumption is that the strategy of saccadic eye movements differs for low-pass smoothed and unsmoothed stimuli. If, for instance, saccadic sampling of blurry images turns out to be associated with faster and/or more distant jumps,—this can effectively lead to stronger shift perception in comparison to unsmoothed stimuli.

5. Pattern Recognition using Phase Correlation

As we have seen above, pattern recognition and motion detection are closely related tasks in the frequency domain. In fact, detection of pattern motion using phase correlation premises the knowledge of complete spectral characteristics of a pattern, i.e., pattern recognition. The tight relationship between pattern's cognitive characteristics and motion can be seen as an exclusive feature of frequency domain techniques such as phase correlation, which differs them, for example, from gradient-based optical flow methods (Barron et al., 1994). The existing body of neurophysiological and psychophysical evidence do not allow to make a conclusion about the nature of neural mechanisms of pattern recognition. However, from the literature it is known that (i) the retinal images are frequency-coded, filtered and processed in visual cortex by several layers of specialized cells in a hierarchically organized manner (Mesulam, 1998; Kruger et al., 2013), (ii) recognition takes place in higher levels of this hierarchy, i.e., the association cortex, where high confidence pattern recognition has been related to activity of single cells (Quiroga et al., 2005), and (iii) saccades are involved in acquisition of the information for rapid scene recognition (Kirchner and Thorpe, 2006). By putting these findings together with our theoretical and experimental investigations, we hypothesize here that phase correlation (or an effectively similar mechanism) is involved in neural machinery of pattern recognition. The basic statements of this hypothesis are as follows:

- Images are coded in the neural network by their frequency domain features (i.e., phases and amplitudes).
- Phase correlation between neural images is performed by a special layer of cells [further termed as association layer neurons (ALN)].
- Similarity between each two visual stimuli is sensed by the spatial-temporal pattern of ALN activity in analogy to *PC* of two images, cf. **Figure 3**.

Figure 9 depicts the principle scheme of this hypothetic mechanism which postulates integration (phase correlation) of source and target images in association cortex and predicts the neural activity patterns related to perception of image (dis)similarity. According to this hypothesis, the physiological expression of high-confidence recognition of a visual stimulus is a coherent and persistent activity of a relatively small number of ALN (theoretically, even one single neuron as it has been observed in Quiroga et al. (2005)). In contrast, low similarity between visual stimuli would result in a diffuse and uncorrelated pattern of ALN activity.

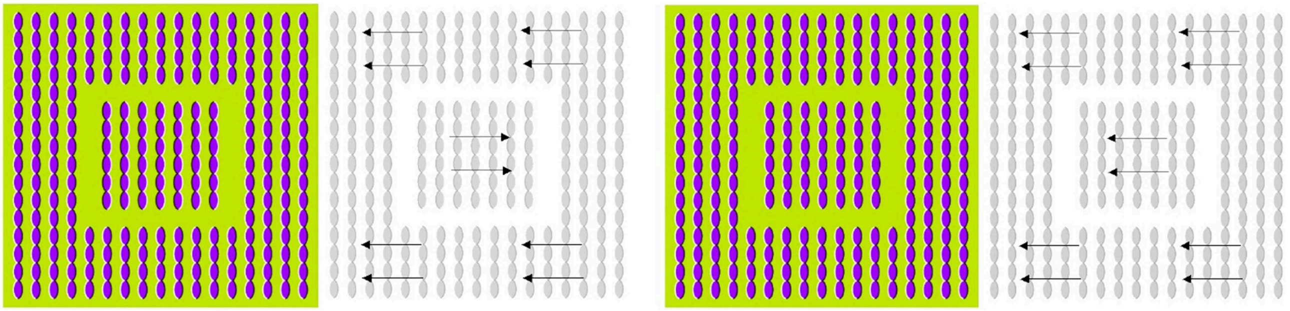


FIGURE 7 | The Anomalous Motion (courtesy A. Kitaoka) induces an illusion of apparent translational motion (left). Manipulated equidirectional stimulus **(right)** do not trigger any significant motion illusions.

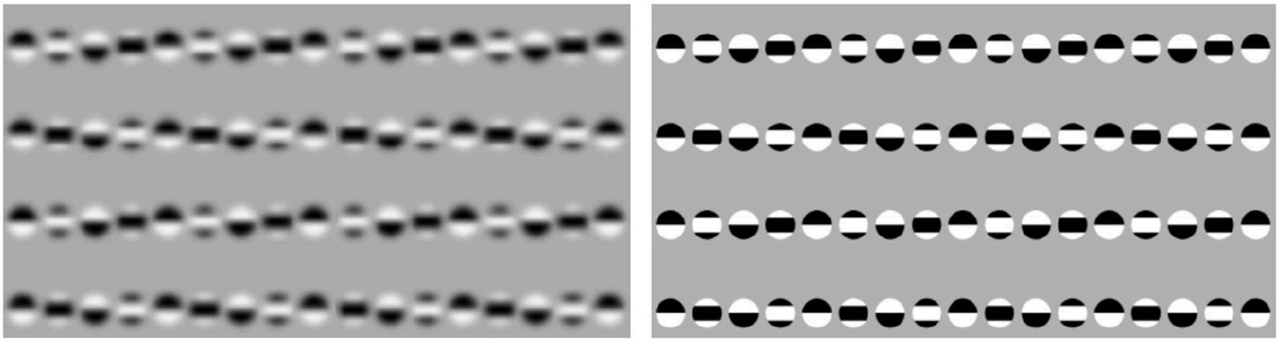


FIGURE 8 | Dependence of the non-local tilt illusion on low/high-frequent image content. From left to right: the low-pass filtered vs. unfiltered Popple illusion (courtesy A. Kitaoka).

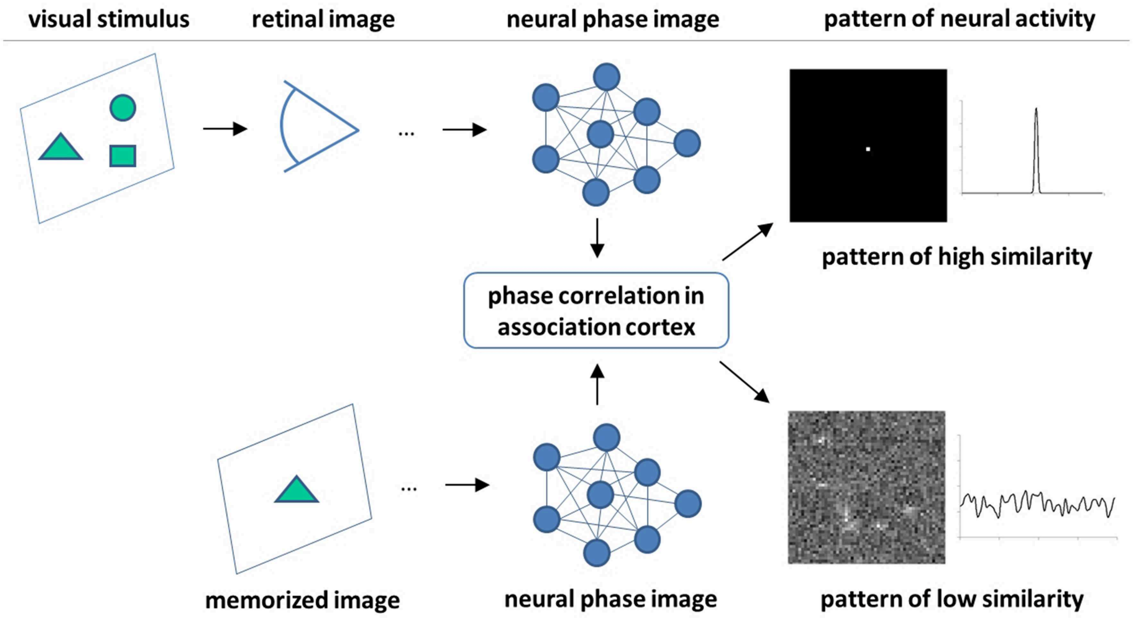

visual stimulus
retinal image
neural phase image
pattern of neural activity
...
pattern of high similarity
phase correlation in association cortex
...
memorized image
neural phase image
pattern of low similarity

FIGURE 9 | Scheme of the hypothetic mechanisms of visual pattern recognition. Persistent activity of a small number of neurons in association cortex is a feature of high image similarity. In the ideal case, similarity is detected by a single neuron. In contrast, a more disperse and stochastic pattern of neural activity indicates a low degree of image similarity.

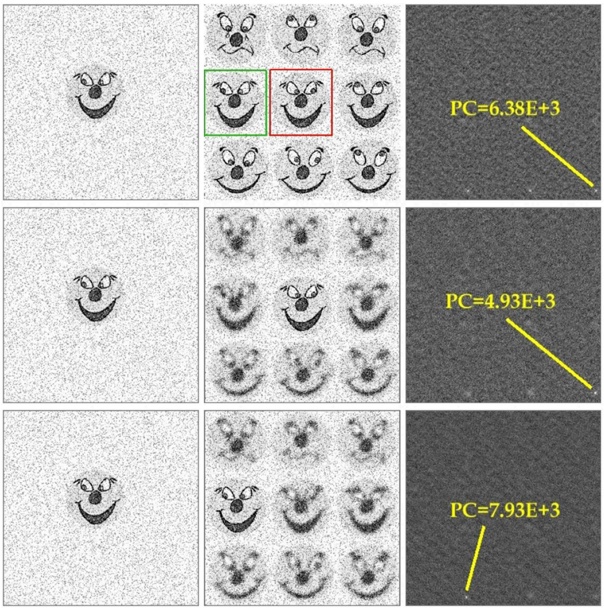


FIGURE 10 | Example of pattern recognition using phase correlation. From left to right: (i) the target smiley, (ii) multi-smiley image, phase correlation between (i) and (ii). The green frame indicates the correct location of the target pattern in the image, the red frame shows the wrong match which corresponds to the absolute maximum of the noisy phase correlation. Consideration of visual acuity improves the recognition score. Phase correlation between the target smiley and the images with three different acuity foci peaks out the right pattern location with the maximum height of PC = 7.93E+3.

Furthermore, missing similarity between images can be expected to provoke intensification of saccadic eye movements.

An example of repetitive pattern discrimination/recognition using phase correlation is shown in **Figure 10**. The task consists in finding a particular smiley within a group of similar patterns. Since phase correlation of noise-free images will immediately match the right location of the target smiley, the search is complicated by adding a large amount of high-frequency noise which substantially corrupts small image features (such as smiley's eyes). Single-step phase correlation between substantially noised images results in selection of the wrong pattern location (see yellow framed smiley in **Figure 10**). Due to high-level of noise, the peak of phase correlation corresponding to the correct pattern (green framed smiley) has the lower height. Remarkably, consideration of visual acuity (i.e., peripheral blurring) helps to improve the recognition score. Phase correlation between the target smiley and three images with different visual foci manages to peak out the right pattern location which corresponds to the highest peak of $PC = 7.93E + 3$.

Another example of remarkable features of phase correlation as a pattern recognition tool is detection of the virtual image content in visual completion illusions. **Figure 11** demonstrates detection of virtual geometrical patterns (i.e., triangle, circle) in the completion illusions from Idesawa (1991) and Kanizsa (1995). The correct location of the virtual figures corresponds to the absolute maximum of phase correlation. This examples demonstrate that phase correlation is capable to retrieve even extremely subtle pattern correspondences.

6. Discussion

Here, we merge existing phenomenological findings, computational analysis and theoretical hypotheses to dissect the role of image phase in diverse phenomena of visual information processing, illusion and cognition. We argue that fundamental importance of phase for detection of structural image features and transformations is of clear evolutionary advantage for survival of species and can be assumed to promote the development of phase-based mechanisms of neural image processing. A large body of neurophysiological and psychophysical evidence seems to confirm the assumption that biological vision relies on frequency domain transformation, filtering and higher-order processing of retinal images in the visual cortex. Hence, the emergence of efficient phase-based neural mechanisms in course of evolution appears to be plausible. We show that the concepts of phase shift, amplitude-normalizing phase-only transformation and phase correlation provide a qualitative description for a number of puzzling visual phenomena including

- preservation of cognitive features in the image sketch (in the sense of the Marr's Primal Sketch),
- robustness of pattern detection with respect to substantial level of noise and structural distortion,
- "eye exhaustion" by observation of repetitive and blurry scenes,
- advantages of saccadic strategy of iterative target-background sampling for pattern discrimination,
- dependency of saccadic eye movements on structural image properties (i.e., target-background similarity and spectral characteristics),
- advantages of differences in foveal and peripheral acuity for visual pattern recognition,
- dependency of the delay time by perception of virtual depth illusions on phase properties of stimuli,
- coherent phase shifts in contrast-gradient patterns of apparent motion illusions,
- driving role of saccades in apparent motion and tilt illusions,
- recognition of virtual patterns in completion illusions using phase correlation.
- singular pattern of neural activity in the association cortex by recognition of similar visual stimuli.

Although, straightforward projections of theoretical concepts onto biological systems can, in general, lead to too far-reaching extrapolations, some of our hypothetic predictions, such as dependency of saccades strategy on structural image properties and singular response of association cortex to structurally similar visual stimuli, can be, on principle, tested in experiment.

There is a tight resemblance between the concepts of amplitude-normalizing phase-only transformation and phase correlation we used in our work and energy models (Morrone and Owens, 1987; Morrone and Burr, 1988; Fleet et al., 1996) re. phase congruency detectors (Morrone et al., 1986; Kovesi, 2000). Both concepts take advantage of two basic principles:

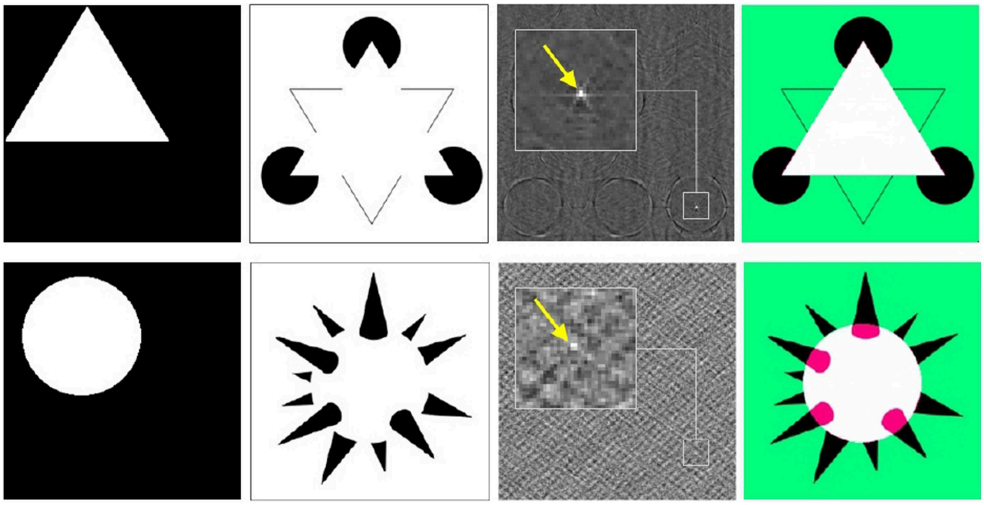


FIGURE 11 | Detection of the virtual image content using phase correlation. From left to right: (i) hidden patterns of illusion stimuli (i.e., triangle, circle), (ii) visual completion illusion from Kanizsa (1995) (top row) and Idesawa (1991) (bottom row), (iii) phase correlations between (i) and (ii) (maximum is indicated by the arrow), registration of (i) onto (ii) according to the maximum of (iii).

(i) amplitude-normalization, which effectively performs edge enhancement (i.e., image sketchification) and makes scene analysis independent of the level of illuminance and contrast, and (ii) calculation of the cognitive checksum by building an integral over the entire frequency spectrum, which, on one hand, makes the cognition extremely robust with respect to noise and, on the other hand, allows distributed storage of information in neural networks. Otherwise, there is a basic difference between these two concepts: phase congruency can be seen as an extended amplitude-normalizing, edge-enhancing filter, while phase correlation is constructed to detect the relative transformation and/or structural (dis)similarity between each two images. Furthermore, phase congruency is presumably performed by V1 neurons, while phase correlation can be expected to take place in a higher level of visual cortex hierarchy, i.e., association cortex. Finally, taking into consideration potential redeployment of the brain areas (Anderson, 2007), one can expect that the suggested principle of pattern recognition by phase correlation is not restricted to the visual system and could also play a role in other cognitive functions.

Within the general framework of recent hierarchical bottom-up top-down models of visual cortex (Lee and Mumford, 2003; Epshtein et al., 2008; Poggio and Ullman, 2013), our findings provide a theoretical explanation for what Marr called "early non-attentive vision" (Marr, 1976, 1982). In particular, our above results suggest that phase-only transformation in V1 with subsequent phase correlation in association cortex represent bottom-up neural mechanisms of Primal Sketch generation and perception, respectively. However, differently from the canonical edge operators that are based on derivatives (i.e., edge-mask convolution) of the image intensity function, edge information in the frequency domain is given implicitly by the relative phase structure and can be assessed for the entire image in a non-iterative and non-local manner. The ability of phase correlation to capture global structural information "on-the-fly" makes it to an ultimate tool for rapid bottom-up processing of the focused image content. The temporal focus of the observer is, in turn, controlled by higher-order cortical centers that integrate bottom-up streams and define conscious and unconscious strategies of visual scene sampling.

While the focus of our present work is on the role of image phase in visual information processing, it should be stated that phase does not exclusively bear cognitive features of visual stimuli. Findings in Freeman and Simoncelli (2011) and Zhang et al. (2014) suggest that amplitude information is also involved in visual (re)cognition and can be even overweight in peripheral vision or by perception of textural images. It is a subject of future research to reveal how phase and amplitude are weighted and merged to an integrated whole in association cortex upon structural properties of visual stimuli.

References

Adelson, E., and Bergen, J. (1985). Spatiotemporal energy models for the perception of motion. *J. Opt. Soc. A* 2, 284–299.

Anderson, M. (2007). Evolution of cognitive function via redeployment of brain areas. *Neuroscientics* 13, 1–9. doi: 10.1177/1073858406294706

Anzai, A., Ohzawa, I., and Freeman, R. (1997). Neural mechanisms underlying binocular fusion and stereopsis: position vs. phase. *Proc. Natl. Acad. Sci. U.S.A.* 94, 5438–5443.

Arndt, P., Mallot, H., and Biilthoff, H. (1995). Human stereovision without localized image features. *Biol. Cybern.* 72, 279–293.

Backus, B., and Oru, I. (2005). Illusory motion from change over time in the response to contrast and luminance. *J. Vis.* 5, 1055–1069. doi: 10.1167/5.11.10

Barron, J., Fleet, D., and Beauchemin, S. (1994). Performance of optical flow techniques. *Int. J. Comp. Vis.* 12, 43–77.

Bex, P., and Makous, W. (2002). Spatial frequency, phase, and the contrast of natural images. *J. Opt. Soc. Am. A* 19, 1096–1106. doi: 10.1364/JOSAA.19.001096

Blakemore, C., and Campbell, F. (1969). On the existence of neurones in the human visual system selectively sensitive to the orientation and size of retinal images. *J. Physiol.* 213, 237–260.

Blakemore, C., Nachmias, J., and Sutton, P. (1969). The perceived spatial frequency shift: evidence for frequency-selective neurones in the human brain. *J. Physiol.* 210, 727–750.

Booth, M., and Rolls, E. (1998). View-invariant representations of familiar objects by neurons in the inferior temporal visual cortex. *Cereb. Cortex* 8, 510–523.

Campbell, F., and Robson, J. (1968). Applciation of fourier analysis to the visibility of gratings. *J. Physiol.* 197, 551–566.

Chen, Y., and Qian, N. (2004). A coarse-to-fine disparity energy model with both phase-shift and position-shift receptive field mechanisms. *Neural Comput.* 16, 1545–1577. doi: 10.1162/089976604774201596

Conway, B., Kitaoka, A., Yazdanbakhsh, A., Pack, C., and Livingstone, M. (2005). Neural basis for a powerful static motion illusion. *J. Neurosci.* 25, 5651–5656. doi: 10.1523/JNEUROSCI.1084-05.2005

De Castro, E., and Morandi, C. (1987). Registration of translated and rotated images using finite fourier transforms. *IEEE Trans. Pattern Anal. Mach. Intell.* 9, 700–703.

De Valois, R., and De Valois, K. (1990). *Spatial Vision*. New York, NY: Oxford University Press.

De Valois, R., Albrecht, D., and Thorell, L. (1982). Spatial frequency selectivity of cells in macaque visual cortex. *Vis. Res.* 22, 545–559.

Donoho, D., and Flesia, A. (2001). Can recent innovations in harmonic analysis 'explain' key findings in natural image statistics. *Network* 12, 391–412. doi: 10.1080/net.12.3.371.393

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. doi: 10.1073/pnas.0800968105

Fleet, D., and Jepson, A. (1990). Computation of component image velocity from local phase information. *Int. J. Comp. Vis.* 5, 77–104.

Fleet, D., Wagner, H., and Heeger, D. (1996). Neural encoding of binocular disparity: energy models, position shifts and phase shifts. *Vis. Res.* 36, 1839–1857.

Fleet, D. (1994). "Disparity from local weighted phase correlation," in *Proceedings IEEE International Conference on Systems, Man and Cybernetics* (San Antonio, TX), 48–56.

Freeman, J., and Simoncelli, E. (2011). Metamers of the ventral stream. *Nat. Neurosci.* 14, 1195–1201. doi: 10.1038/nn.2889

Gladilin, E., and Eils, R. (2009). "Detection of non-uniform multi-body motion in image time-series using saccades-enhanced phase correlation," in *Proceedings of SPIE Medical Imaging 2009: Image Processing*, eds J. P. W. Pluim; B. M. Dawant, (San Diego, CA). doi: 10.1117/12.811120

Gladilin, E. (2004). "A contour based approach for invariant shape description," In *Proceedings of SPIE, Medical Imaging 2004: Image Processing* (San Diego, CA), 5370, 1282–1291.

Goutcher, R., and Hibbard, P. (2014). Mechanisms for similarity matching in disparity measurement. *Front. Psych.* 4:1014. doi: 10.3389/fpsyg.2013.01014

Graham, D., and Field, D. (2006). *Evolution of the Nervous Systems Chapter Sparse Coding in the Neocortex*. Ithaca, NY: Academic Press.

Graham, N. (1981). "The visual system does a crude Fourier analysis of patterns," in *Mathematical Psychology and Psychophysiology, SIAM-AMS Proceedings Vol. 13.,* ed S. Grossberg, (Providence, Rhode Island, American Mathematical Society), 1–16.

Graham, N. (1989). *Visual Pattern Analyzers*. New York, NY: Oxford University Press.

Hamilton, D., Albrecht, D., and Geisler, W. (1989). Visual cortical receptive fields in monkey and cat: spatial and temporal phase. *Vis. Res.* 29, 1285–1308.

He, P., and Kowler, E. (1992). The role of saccades in the perception of texture patterns. *Vis. Res.* 32, 2151–2163.

Heeger, D. (1992). Normalization of cell responses in cat striate cortex. *Vis. Neurosci.* 9, 181–197.

Henriksson, L., Hyvaerinen, A., and Vanni, S. (2009). Representation of cross-frequency spatial phase relationships in human visual cortex. *J. Neurosci.* 29, 14342–14351. doi: 10.1523/JNEUROSCI.3136-09.2009

Hietanen, M., Cloherty, S., van Kleef, J., Wang, C., Dreher, B., and Ibbotson, M. (2013). Phase sensitivity of complex cells in primary visual cortex. *J. Neurosci.* 237, 19–28. doi: 10.1016/j.neuroscience.2013.01.030

Hubel, D., and Wiesel, T. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol.* 160, 106–154.

Hubel, D., and Wiesel, T. (1968). Receptive fields and functional architecture of monkey striate cortex. *J. Physiol.* 195, 215–243.

Hyvärinen, A., and Hoyer, P. (2001). A two-layer sparse coding model learns simple and complex cell receptive fields and topography from natural images. *Vis. Res.* 41, 2413–2423. doi: 10.1016/S0042-6989(01)00114-6

Idesawa, M. (1991). "Perception of illusory solid object with binocular viewing," in *Proceedings IJCNN-91 Seattle International Joint Conference of Neural Networks* (Seattle, WA), Vol. II, A-943.

Ito, M., Tamura, H., Fujita, I., and Tanaka, K. (1995). Size and position invariance of neuronal responses in monkey inferotemporal cortex. *J. Neurophysiol.* 73, 218–226.

Kanizsa, G. (1995). Margini quasi-percettivi in campi con stimolazione omogenea. *Riv. Psycol.* 49, 7–30.

Kirchner, H., and Thorpe, S. (2006). Ultra-rapid object detection with saccadic eye movements: visual processing speed revisited. *Vis. Res.* 46, 1762–1776. doi: 10.1016/j.visres.2005.10.002

Kitaoka, A., and Ashida, H. (2003). Phenomenal characteristics of the peripheral drift illusion. *VISION* 15, 261–262. Available online at: http://www.psy.ritsumei.ac.jp/∼akitaoka/PDrift.pdf

Kitaoka, A. (2006). "Anomalous motion illusion and stereopsis," in *Journal Three Dimensional Images* (Tokyo), 9–14.

Kovesi, P. (2000). Phase congruency: a low-level image invariant. *Psych. Res.* 64, 136–148. doi: 10.1007/s004260000024

Kruger, N., Janssen, P., Kalkan, S., Lappe, M., Leonardis, A., Piater, J., et al. (2013). Deep hierarchies in the primate visual cortex: what can we learn for computer vision? *IEEE Trans. Pattern Anal. Mach. Intell.* 35, 1847–1871. doi: 10.1109/TPAMI.2012.272

Lee, T., and Mumford, D. (2003). Hierarchical bayesian infer-ence in the visual cortex. *J. Opt. Soc. Am. A* 20, 1434–1448. doi: 10.1364/JOSAA.20.001434

Lindeberg, T. (2013). Invariance of visual operations at the level of receptive fields. *PLoS ONE* 8:e66990. doi: 10.1371/journal.pone.0066990

Lohmann, A., Mendlovic, D., and Gal, S. (1997). Signicance of phase and amplitude in the fourier domain. *J. Opt. Soc. Am. A* 14, 2901–2904.

Mallat, S. (1989). A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Trans. Pattern Anal. Mach. Intell.* 11, 674–693.

Marcelja, S. (1980). Mathematical description of the responses of simple cortical cells. *J. Opt. Soc. Am.* 70, 1297–1300.

Marr, D. (1976). Early processing of visual information. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 275, 483–519.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. San Francisco, CA: W. H. Freeman and Company.

Martinez-Conde, S., Otero-Millan, J., and MacKnik, S. (2013). The impact of microsaccades on vision: towards a unified theory of saccadic function. *Nat. Rev. Neurosci.* 14, 83–96. doi: 10.1038/nrn3405

Mesulam, M. (1998). From sensation to cognition. *Brain* 121, 1013–1052.

Morgan, M., Ross, J., and Hayes, A. (1991). The relative importance of local phase and local amplitude in patchwise image recognition. *Biol. Cybern.* 65, 113–119.

Morrone, M., and Burr, D. (1988). Feature detection in human vision: a phase-dependent energy model. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 235, 221–245.

Morrone, M., and Owens, R. (1987). Feature detection from local energy. *Pattern Recogn. Lett.* 6, 303–313.

Morrone, M., Ross, J., Burr, D., and Owens, R. (1986). Mach bands are phase dependent. *Nature* 324, 250–253.

Ni, X., and Huo, X. (2007). Statistical interpretation of the importance of phase information in signal and image reconstruction. *Stat. Probab. Lett.* 77, 447–454. doi: 10.1016/j.spl.2006.08.025

Nishida, S. (2011). Advancement of motion psychophysics: review 2001-2010. *J. Vis.* 11, 1–53. doi: 10.1167/11.5.11

Ohzawa, I., DeAngelis, G., and Freeman, R. (1990). Stereoscopic depth descrimination in the visual cortex: neurons ideally suited as disparity detectors. *Science* 249, 1037–1041.

Ohzawa, I., DeAngelis, G., and Freeman, R. (1997). Encoding of binocular disparity by complex cells in the cat's visual cortex. *J. Neurophysiol.* 77, 2879–2909.

Oppenheim, A., and Lim, J. (1981). The importance of phase in signals. *Proc. IEEE* 69, 529–541. doi: 10.1109/PROC.1981.12022

Osterberg, G. (1935). *Topography of the Layer of Rods and Cones in the Human Retina* Vol. 13 of *Acta Ophthalmologica*. Copenhagen: A. Busck.

Otero-Millan, J., MacKnik, S., and Martinez-Conde, S. (2012). Microsaccades and blinks trigger illusory rotation in the rotating snakes illusion. *J. Neurosci.* 32, 6043–6051. doi: 10.1523/JNEUROSCI.5823-11.2012

Palmeri, T., and Gauthier, I. (2004). Visual object understanding. *Nat. Rev. Neurosci.* 5, 291–304. doi: 10.1038/nrn1364

Parker, A., and Cumming, B. (2001). Cortical mechanisms of binocular stereoscopic vision. *Prog. Brain Res.* 134, 205–216.

Poggio, T., and Ullman, S. (2013). Vision: are models of object recognition catching up with the brain? *Ann. N. Y. Acad. Sci.* 1305, 72–82. doi: 10.1111/nyas.12148

Pollen, D., and Ronner, S. (1981). Phase relationship between adjacent simple cells in the visual cortex. *Science* 212, 1409–1411.

Pollen, D., and Ronner, S. (1983). Visual cortical neurons as localized spatial frequency filters. *IEEE Trans. Sys. Man Cybern.* 5, 907–916.

Popple, A., and Levi, D. (2000). A new illusion demonstrates long-range processing. *Vis. Res.* 40, 2545–2549. doi: 10.1016/S0042-6989(00)00127-9

Popple, A., and Sagi, D. (2000). A fraser illusion without local cues? *Vis. Res.* 40, 873–878. doi: 10.1016/S0042-6989(00)00010-9

Psalta, L., Young, A., Thompson, P., and Andrews, T. (2014). The thatcher illusion reveals orientation dependence in brain regions involved in processing facial expressions. *Psychol. Sci.* 25, 128–136. doi: 10.1177/0956797613501521

Quiroga, R., Reddy, L., Kreiman, G., Koch, C., and Fried, I. (2005). Invariant visual representation by single neurons in the human brain. *Nature* 435, 1102–1107. doi: 10.1038/nature03687

Ramachandran, V., and Anstis, S. (1986). The perception of apparent motion. *Sci. Am.* 254, 102–109.

Reddy, B., and Chatterji, B. (1996). An fft-based technique for translation, rotation, and scale-invariant image registration. *IEEE Trans. Image Process.* 5, 1266–1271.

Riesenhuber, M. (2005). *Neurobiology of Attention Chapter Object Recognition in Cortex: Neural Mechanisms, and Possible Roles for Attention.* Philadelphia, PA: Elsevier.

Sampat, M., Wang, Z., Gupta, S., Bovik, A., and Markey, M. (2009). Complex wavelet structural similarity: a new image similarity index. *IEEE Trans. Image Process.* 18, 2385–2401. doi: 10.1109/TIP.2009.2025923

Schwartz, O., and Simoncelli, E. (2001). Natural signal statistics and sensory gain control. *Nat. Neurosci.* 4, 819–825. doi: 10.1038/90526

Shams, L., and Malsburg, C. (2002). The role of complex cells in object recognition. *Vis. Res.* 42, 2547–2554. doi: 10.1016/S0042-6989(02)00202-X

Thaler, L., Todd, J., and Dijkstra, T. (2007). The effects of phase on the perception of 3d shape from texture: psychophysics and modeling. *Vis. Res.* 47, 411–427. doi: 10.1016/j.visres.2006.10.007

Thomas, J., Bagrash, F., and Kerr, L. (1969). Selective stimulation of two form sensitive mechanisms. *Vis. Res.* 9, 625–627.

Troncoso, X., MacKnik, S., Otero-Millan, J., and Martinez-Conde, S. (2008). Microsaccades drive illusory motion in the enigma illusion. *Proc. Natl. Acad. Sci. U.S.A.* 105, 16033–16038. doi: 10.1073/pnas.0709389105

Tyler, C., and Clarke, M. (1990). "The autostereogram," In *Proceedings of SPIE, Stereoscopic Displays and Applications* (Santa Clara, CA), 182–196.

Walls, G. (1962). The evolutionry history of eye movements. *Vis. Res.* 2, 69–80.

Weldon, K., Taubert, J., Smith, C., and Parr, L. (2013). How the thatcher illusion reveals evolutionary differences in the face processing of primates. *Anim. Cogn.* 16, 691–700. doi: 10.1007/s10071-013-0604-4

Yarbus, A. (1967). *Eye Movements and Vision*. New York, NY: Plenum Press.

Zhang, F., Jiang, W., Autrusseau, F., and Lin, W. (2014). Exploring v1 by modeling the perceptual quality of images. *J. Vis.* 14, 1–14. doi: 10.1167/14.1.26

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.

ORIGINAL RESEARCH
published: 04 November 2015
doi: 10.3389/fncom.2015.00134

Aesthetic perception of visual textures: a holistic exploration using texture analysis, psychological experiment, and perception modeling

Jianli Liu[1]*, *Edwin Lughofer*[2] *and Xianyi Zeng*[3]

[1] College of Textile and Clothing, Jiangnan University, Wuxi, China, [2] Department of Knowledge-Based Mathematical Systems/Fuzzy Logic Laboratorium Linz-Hagenberg, Johannes Kepler University Linz, Linz, Austria, [3] The ENSAIT Textile Institute, University of Lille Nord de France, Roubaix, France

Modeling human aesthetic perception of visual textures is important and valuable in numerous industrial domains, such as product design, architectural design, and decoration. Based on results from a semantic differential rating experiment, we modeled the relationship between low-level basic texture features and aesthetic properties involved in human aesthetic texture perception. First, we compute basic texture features from textural images using four classical methods. These features are neutral, objective, and independent of the socio-cultural context of the visual textures. Then, we conduct a semantic differential rating experiment to collect from evaluators their aesthetic perceptions of selected textural stimuli. In semantic differential rating experiment, eights pairs of aesthetic properties are chosen, which are strongly related to the socio-cultural context of the selected textures and to human emotions. They are easily understood and connected to everyday life. We propose a hierarchical feed-forward layer model of aesthetic texture perception and assign 8 pairs of aesthetic properties to different layers. Finally, we describe the generation of multiple linear and non-linear regression models for aesthetic prediction by taking dimensionality-reduced texture features and aesthetic properties of visual textures as dependent and independent variables, respectively. Our experimental results indicate that the relationships between each layer and its neighbors in the hierarchical feed-forward layer model of aesthetic texture perception can be fitted well by linear functions, and the models thus generated can successfully bridge the gap between computational texture features and aesthetic texture properties.

Keywords: visual texture, aesthetic emotion, texture analysis, psychological experiment, dimension reduction, perception modeling, layered model architecture

Edited by:
Judith Peters,
The Netherlands Institute for Neuroscience, Netherlands

Reviewed by:
David R. Simmons,
University of Glasgow, UK
Petia D. Koprinkova-Hristova,
Bulgarian Academy of Sciences, Bulgaria

***Correspondence:**
Jianli Liu
jian-li.liu@hotmail.com

Received: *07 October 2014*
Accepted: *19 October 2015*
Published: *04 November 2015*

Citation:
Liu J, Lughofer E and Zeng X (2015) Aesthetic perception of visual textures: a holistic exploration using texture analysis, psychological experiment, and perception modeling. Front. Comput. Neurosci. 9:134. doi: 10.3389/fncom.2015.00134

INTRODUCTION

Texture is ubiquitous. It contains important visual information about an object and allows us to distinguish between animals, plants, foods, and fabrics. This makes texture a significant part of the sensory input that we receive every day. In the visual arts, texture is the perceived surface quality of a work of art. It is an element of two- and three-dimensional designs and is distinguished by its perceived visual and physical properties (Graham and Meng, 2011). From the research point of view, textures are classified into tactile and visual textures. The former, also known as actual textures or physical textures, are actual surface variations (Elkharraz et al., 2014), including, but not limited to, fur, wood grain, sand, and the smooth surfaces of canvas, metal, glass, and leather (Skedung et al., 2013). Physical texture is distinguished from visual texture by a physical quality that can be felt by touch (Manfredi et al., 2014). Visual texture is the illusion of physical texture. Every material has its own visual texture. Photographs, drawings, and paintings use visual texture to portray their participant matter both realistically and with interpretation (Guo et al., 2012). Above all, visual scientists have realized that the rich resource they are provided with by artists in the form of textures is worthy of scientific study (Zeki, 2002).

The challenge in aesthetic perception of visual textures and art is to understand the aesthetic emotion and judgment that are evoked when we experience beauty. To evaluate and explain beauty in science, models of aesthetic perception and judgment have been proposed in cognitive psychology and information science. According to the information-processing stage model of aesthetic processing, five stages-perception, explicit classification, implicit classification, cognitive mastering, and evaluation are involved in aesthetic experiences (Leder et al., 2004).

To discriminate between aesthetically pleasing and displeasing images, Datta et al., employed support vector machines and classification trees to perform explicit classification, and applied linear regression to polynomial terms of features to infer numerical ratings of aesthetics (Datta et al., 2006). Additionally, Datta et al., developed multi-category classifiers to recognize coarse-grained aesthetic categories and used support vector machines to predict fine-grained aesthetic scores (Datta et al., 2006). Jiang et al. used two model built algorithms to study automatic assessment of the aesthetic value in consumer photographic images (Jiang et al., 2010). Cela-Conde et al. pointed out that investigating the cognitive and neural underpinnings of aesthetic appreciation by means of neuro-imaging has yielded a wealth of fascinating information (Cela-Conde et al., 2011). Toet et al. explored the effects of various spatiotemporal dynamic texture characteristics on human emotions (Toet et al., 2011). Using structural equation modeling, Leder et al. explored aesthetic perception by analyzing expertise-related differences in the aesthetic appreciation of classical, abstract, and modern artworks (Leder et al., 2012). Simmons explored the relationship between color information and the emotions they induced by measuring along two affective dimensions, namely pleasant-unpleasant, and arousing-calming (Simmons, 2012).

In their research, Cela-Conde et al. discussed adaptive and evolutionary explanations for the relationships between the default mode network and aesthetic networks, and offered unique input to debates on the interaction between mind and brain (Cela-Conde et al., 2013). Reviewing from definitional, methodological, empirical, and theoretical perspectives of human aesthetic preferences, Palmer et al. concluded that visual aesthetic response can be studied rigorously and meaningfully within the framework of scientific psychology (Palmer et al., 2013). The research of Bundgaard addressed the phenomenology of aesthetic experience, which showed why and how aesthetic experience should be defined relative to its object and the tools for meaning-making specific to that object and not relative to the feeling (Bundgaard, 2014). Chatterjee and Vartanian reviewed recent evidence that approves aesthetic experiences emerge from the interaction between sensory–motor, emotion–valuation, and meaning–knowledge neural systems (Chatterjee and Vartanian, 2014). In experiment, Elkharraz et al. designed and manufactured 3D tactile textures with predefined affective properties, and used mixing algorithms to synthesize 48 new tactile textures that were likely to score highly against the predefined affective properties (Elkharraz et al., 2014).

However, surprisingly little funded research has been conducted on the emotional qualities and expectations associated with specific textures. In 2007, the project named "Measuring Feelings and Expectations Associated with Texture" (SynTex) was supported by the European Commission within the sixth framework program. SynTex was coordinated by Profactor GmbH and conducted in collaboration with six other research institutes in the European Union. In fact, SynTex is the only project to have ever attempted to measure, model and predict the psychological effects of texture. Thumfart et al. summarized the outcomes of this project (Thumfart et al., 2011). A further outcome is in the work of Groissboeck, which focused on synthesizing textures for predefined, desired emotions described by a numerical vector in aesthetic space (Groissboeck et al., 2010). We build upon this research, but go a step further in terms of significantly enhanced texture analysis, feature selection, and layered model-building for better interpretability, while achieving improved accuracy in the prediction of several core adjectives that define the aesthetic space. Expanding the aesthetic space used in Thumfart et al. (2011), we introduced two new adjectives in our experiments.

After reviewing related work in Section Introduction, we present the four different categories of low-level features that were extracted to objectively represent the visual textures in Section Materials and Methods. Further, we describe feature selection using Laplacian Score to reduce the complexity of the aesthetic perception model. Section Results and Discussion summarizes the semantic differential rating experiment, in which we collected aesthetic perceptions from participants with selected textural stimuli. We describe the modeling approaches in Section Results and Discussion; Section Conclusions conclude the paper.

MATERIALS AND METHODS

Selected Textural Stimuli

The visual texture database of stimuli used in our experiment consists of 151 selected high-resolution textural images, which are also the experiment materials used in SynTex project. This database is the Supplementary Material of the paper published by Thumfart et al., in the proceedings of the 13th international conference on Computer Analysis of Images and Patterns (CAIP 2009) (Thumfart et al., 2009). The project SynTex is by far outdated and the link that provides the visual texture database has been closed. The used visual textures for our study can be sent to readers upon request via email or dropbox exchange. Readers can contact us by using the email addresses given in the affiliations.

It includes natural, artificial, regular and stochastic textures in the textural stimuli, which were selected from various texture databases. In detail, 73 textures were chosen from the Brodatz texture album, 69 from the Outex texture database, 25 from the UIUCTex database, 12 from the USC-SIPI image database, and 64 from the VisTex database. Since the original sizes of the textures selected from different database varied, they were resized to a resolution of 480 × 480 pixels. Some examples of visual textures from the SynTex database are shown in **Figure 1**.

In the SynTex database, some textures are artificial and synthetic, some others are natural. So a nice diversity of different sorts and types of textures is given.

Texture Analysis

Texture analysis refers to the characterization of image regions by their textural content (Karu et al., 1996). Texture analysis attempts to quantify intuitive qualities described by terms such as rough, smooth, silky, and bumpy as functions of the spatial variations in pixel intensities (Guo et al., 2012). Texture analysis is used in a variety of applications, and can be helpful when objects in an image are better characterized by their textures than by intensity or traditional thresholding techniques (Bharati et al., 2004).

In our experiment, four different texture analysis methods are employed to extract statistical characteristics from visual textures, which were then categorized into color and statistical features, and perceptual and frequency-domain energy-based features. In total, we initially derived a set of 106 features for each texture image.

Color Characteristics

Colors play an important role in deciding what we like or dislike, because they evoke complex psychological reactions and give rise to relevant feelings (Ou et al., 2004a,b). In addition to the studies of Simmons (2012) mentioned in the introduction, there is growing interest in the understanding of human feelings in response to seeing colors and colored objects, which are also called “color emotions” in psychology (Lucassen et al., 2011). Experimental results show that the emotional responses to warm/cool, heavy/light, and active/passive are consistent across cultures, but that the like/dislike scale exhibits some differences (Ou et al., 2012). Visual perception of some emotions can be linked to different colors (Augello et al., 2013). Regression analysis is usually applied before product color design to reveal the relationships between human responses on these scales and the underlying color appearance attributes, such as lightness, chroma, and hue (Hanada, 2013; Man et al., 2013).

Six color features were computed from HSV (hue-saturation-value) space to describe each visual texture as the ones used in the work of Romani et al. (2012). In detail, average, and standard deviation of the HSV color matrix elements were calculated after conversion of each texture image from RGB to HSV color space.

Gray Level Co-occurrence Matrix Characteristics

If texture is the dominant information in a small area, then this area has statistically a wide variety of discrete textural features (Baraldi and Parmiggiani, 1995). The simplest texture analysis method uses statistical features computed from histograms. Haralick et al., went a step further and proposed a gray-level co-occurrence matrix (GLCM) in which the relative positions of pixels with respect to each other are considered as well (Haralick et al., 1973; Roberti et al., 2013). Given a spatial relationship between pixels in a texture, such a matrix represents the joint distribution of gray-level pairs of neighboring pixels (Davis et al., 1979). Thus, a considerable amount of information can be obtained by modifying the orientation θ or distance d between pixels, where d specifies the distance between the pixel of interest and its neighbor, and θ gives the direction from the pixel of interest to its neighbor. If either θ or d is set, one GLCM is generated. From each GLCM, four statistical characteristics called contrast, correlation, energy, and homogeneity can be calculated.

To research the effect of distance and orientation on statistical features, we extracted 16 GLCMs, choosing the distance from the set $d = \{2, 4, 6, 8\}$ and the orientation from $\theta = \{0^\circ, 45^\circ, 90^\circ, 135^\circ\}$. Use of these orientation angles, restriction to 135° is inspired by Haralick et al. They have been employed in many published statistical representations of textures and are deemed to provide sufficient information for building gray-level co-occurrence matrices. In total, we extracted 64 statistical features for each computed GLCM.

Tamura Texture Features

In color emotion research, an object usually has a uniform color. However, this is rarely the case for real-life objects. Therefore, the effect of texture on color emotion should be extended. Tamura, Mori, and Yamawaki found in psychological studies that humans respond best to coarseness, contrast, and directionality, and to lesser degrees to line-likeness, regularity, and roughness (Tamura et al., 1978). In most cases, only the first three Tamura features capture the high-level perceptual attributes of a texture well and are useful in visual art appreciation (Castelli and Bergman, 2002). Thus, in contrast to statistical data measures, Tamura texture features seem well suited to capture the emotional perception of visual textures. In our experiment, coarseness, contrast, and directionality were calculated as characteristics that represent the psychological responses to visual perception.

Wavelet-based Energy Texture Features

Wavelets have been successfully used as an effective tool to analyze texture information, as they provide a natural partitioning of the image spectrum into multi-scale and oriented sub-bands via efficient transforms (Brooks et al., 2001). Furthermore, wavelets are used in major image compression standards and are prominent in texture analysis (Dong and Ma, 2011). Wavelet-based energy features can be extracted as frequency features in conjunction with other spatial features to capture visual texture information. The basic idea underlying the wavelet energy signature is to generate textural features from wavelet sub-band coefficients or sub-images at each scale after wavelet transformation (Liu et al., 2011). Assuming that the energy distribution in the frequency domain identifies texture, we used L^1 and L^2 norms as measures in our work. More specifically, we calculated L^1 and L^2 norms from the high-frequency sub-bands of the first four levels that were proposed by Do and Vetterli (2002). To describe the quality of

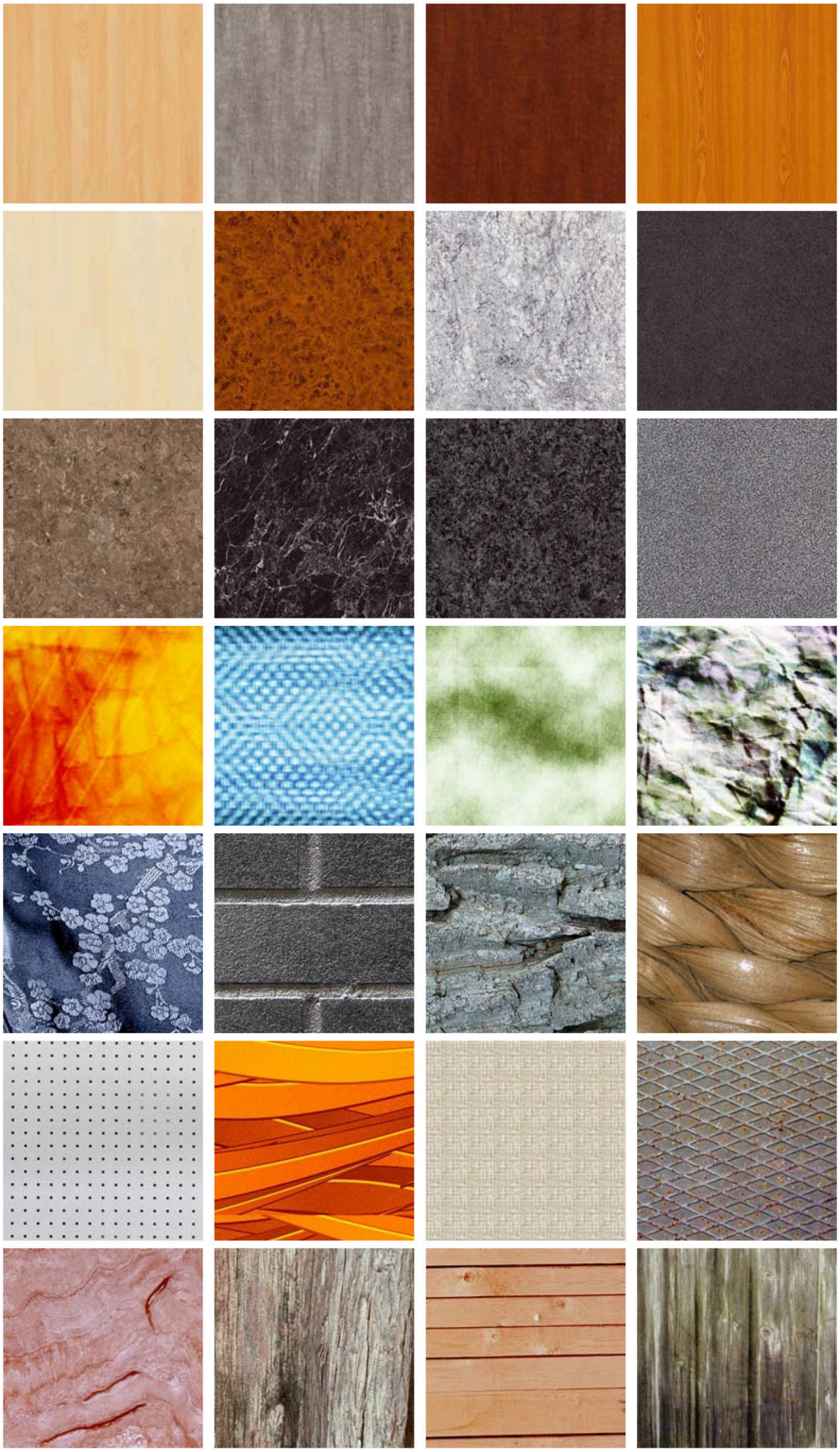


FIGURE 1 | Some examples of visual textures from the SynTex collection.

the information included in each sub-image to be reconstructed with the corresponding wavelet coefficients, we also calculated the Shannon entropy of each high-frequency sub-band.

When a texture image is decomposed at level *j* using a 2D discrete wavelet base, 3 *j* sub-bands are generated. Then 6 *j* energy signatures and 3 *j* entropy signatures are extracted. Since we decomposed each texture image into 4 levels, we extracted 36 wavelet signatures from each texture image.

Feature Selection Using the Laplacian Score

The goal of feature selection is to select the best features from a set of features that not only achieve the maximum prediction rate but can also reduce the complexity of model building (Vapnik, 1998). All feature selection approaches can be applied in either supervised or unsupervised mode (Chandrashekar and Sahin, 2014). In supervised mode, each training sample is described by a vector that consists of feature values with a class label. The class labels are used to guide the search process toward the optimal feature subset. In unsupervised mode, the training samples are not labeled, and thus feature selection is more difficult (Tabakhi et al., 2014). However, this mode provides more general information which can be used by an arbitrary model architecture.

Predicting aesthetic emotions linked to visual textures is a typical example of data mining, where the inputs are low-level texture features and the outputs are aesthetic properties of visual textures. The aesthetics properties used as outputs for modeling are not labeled by strings or 1-0 codes (class labels), as is usual in classification problems, but by discrete real decimal numbers. Ideally, as discussed above, the feature selection method should be independent of the chosen model architecture and also of the hierarchical layered structure (Breiman et al., 1993). Furthermore, we sought to optimize the information content of the feature space while reducing its complexity, with the aim of obtaining one unique reduced set with good interpretation capability.

In a kind of filter selection stage, we thus focused on an unsupervised feature selection scheme called Laplacian Score (LS). LS is a relatively recent unsupervised method for selecting top features (He et al., 2005). It is able to reduce truly redundant and correlated information content of the extracted features—note that only truly redundant features can be discarded without significant information loss (see Guyon and Elisseeff, 2003). In detail, firstly a nearest-neighbor graph was constructed for the original feature set. Secondly, the Laplacian scores for all features in the original feature set were computed using the LS algorithm. Thirdly, the features were ranked according to their Laplacian scores in ascending order. Finally, the last *d* features were discarded, and the feature set was updated with only the remaining features.

Psychological Experiments and Perception Modeling

Aesthetic experiences are very common in modern life, even we don't deliberately care about them. There is yet no scientifically comprehensive theory that explains what constitutes such experiences. As mentioned in the Introduction section, several scientific methods have been used to explore the complex systems that involve in aesthetic experiences. Except for measurements of physiological signals using bio-sensors, psychological experiments are also important tools in exploring cognitive challenges of aesthetic experience and judgments. This section describes the semantic differential rating experiment that was conducted to collect their aesthetic perceptions of visual textures from 10 male and 10 female subjects. The aesthetic properties were assigned to three different layers of the proposed aesthetic perception model.

Definitions of the Aesthetic Properties

Before the semantic differential experiment, we had to select and define the aesthetic properties. Which types of aesthetic properties should be defined and how many pairs of aesthetic antonyms should be selected are hot research topics in semantic analysis. The definition of the eight core adjectives as shown in **Table 1** has been derived from the findings in Levinson (2006) which emphasized that six of these define the aesthetic core space. The two additional ones (dark-light and disordered-harmonious) were considered because of the contents of the textures selected for experiment and the suggestions coming from the 20 subjects.

Before the semantic differential experiments, we explained the meaning of each pair of aesthetic antonyms to the 20 participants and showed them some typical samples. In experiment, we emphasized that these samples were not their only references. We further suggested that knowledge about and preference for—or even prejudice against—some types of visual texture they would encounter should also be considered.

As shown in **Table 1**, the 8 pairs of aesthetic antonyms are also assigned to three emotion layers defined in Thumfart's work. In fact, the 8 pairs of aesthetic antonyms are assigned to effective, judgment or emotional layer by the 20 subjects after surveying 100 persons in 3 days. The logic relationships between these emotion layers are explained in Section Aesthetic Perception Model of Visual Textures.

Semantic Differential Experiment

Semantic differential experiments are commonly used to explore perceptual and emotional dimensions of visual art and music. In our case, a semantic differential experiment was carried out—with the approval of the ethical committee of Jiangnan University for experiments with human participants—to collect participant ratings for the eight aesthetic properties defined in **Table 1**.

TABLE 1 | Eight pairs of aesthetic properties are divided into three layers.

Aesthetic property	Emotion layer	Aesthetic property	Emotion layer
Warm-cold	Affective layer	Inelegant-elegant	Judgment layer
Rough-smooth	Affective layer	Simple-complex	Judgment layer
Dark-light	Affective layer	Artificial-natural	Judgment layer
Disordered-harmonious	Judgment layer	Like-dislike	Emotional layer

In the semantic differential experiment, 20 highly motivated Jiangnan University undergraduate students (aged 19–23) served as participants to rate 151 visual textures in terms of eight aesthetic antonyms. Before experiment, we introduced our research purpose, experimental procedures, and how long it takes to participate to all participants, and provided a written informed consent form to each participant.

After signing a written informed consent form, each participant enrolled for at least 5 daily sessions of 2 h and received payment. The purpose of the experiments was concealed from all participants, and they were trained to use a program we developed called Texture Aesthetic Annotation Assistant to rate the defined aesthetic properties. In each test, participants briefly (300 ms) viewed one visual texture, which was followed immediately by a perceptual mask (200 ms) presented at the same location. The viewing distance was 75 cm (screen to participant). After training, the 20 participants participated in the semantic differential experiment in our lab at their own leisure.

The participants operated the Texture Aesthetic Annotation Assistant which automatically displayed the texture and stored the ratings in a e. A visual texture and a rating bar were shown in the center and at the bottom of the display. The subject could drag the scrollbar to rate the texture according to the labeled aesthetic antonyms (placed at opposite ends of the scrollbar), and the eight pairs appeared sequentially as listed in **Table 1**. Rather than the seven point rating scale, we used a continuous rating scale within the interval [−100, 100] (Chuang and Chen, 2008). This kind of rating method is useful to build a continuous regression model with sufficiently fine granularity.

In the semantic differential experiment, each participant randomly evaluated each texture five times, and the ratings for each texture were stored in a text file. After completion of the semantic differential experiment, the ratings for each visual texture evaluated by the 20 participants (i.e., 100 ratings per texture) were averaged and used as final ratings to build a prediction model for aesthetic emotions (see below).

As the aim of this research was to gain general insights and explore potential relationships between human texture perception and low-level features of visual textures, we did not use individual experimental data to build an individual model for each subject, but created a general model that may be valid for a wider range of applications and purposes and reduces development costs.

Aesthetic Perception Model of Visual Textures

Axelsson summarized five theoretical models that are most important for the development of psychological aesthetics: (1) Berlyne's Collative-Motivational Model, (2) the Preference-for-Prototypes Model, (3) the Preference-for-Fluency Model, (4) Silvia's Appraisal-of-Interest Model, and (5) Eckblad's Cognitive-Motivational Model (Axelsson, 2007). However, these five models were developed in theoretical psychology and can hardly be explained in information-processing and mathematical terms because the input factors are specific human emotions that cannot be quantified. The hierarchical layer structure of these models, however, provides a reference for our work, and some aesthetic properties involved there are also helpful to us.

Achievements in neuroaesthetics are the most important basis for building a hierarchical structure of aesthetic perception, especially the research of Ishizu and Zeki provides powerful support (Ishizu and Zeki, 2013). Also, Thumfart et al. applied a similar hierarchical layer structure, in which we extended with two additional properties, "dark-light" and "disordered-harmonious." The structure of the hierarchical feed-forward model of aesthetic texture perception is shown in **Figure 2**.

In the hierarchical feed-forward model, the function of the affective layer is to complete the descriptions of the general and primary physical properties of the visual texture. Thus, the aesthetic antonyms selected for the affective layer are used to capture the primary emotions when we first skim the visual textures. In the judgment layer, the selected aesthetic antonyms should describe higher-level and more aggregate properties that are in part anchored in the subconscious, especially those induced after statistical and logical judgment. The emotions we feel when interacting with the textures are described in the emotional layer. The aesthetic antonyms selected for the emotional layer should describe the overall feelings people have and wish to express.

Building an Aesthetic Perception Model

Traditional machine learning techniques such as neural networks and support vector regression are useful prediction tools. However, they become completely impractical when interpretability of the implicit relations between low-level features and core adjectives is desired, because they are black boxes and cannot provide any meaningful and understandable insights. Hence, we propose a hierarchical feed-forward layer model of aesthetic texture perception with high interpretability that combines neuroaesthetics and information processing theory. In the layered structure model, each layer has a set of interpretable aesthetic antonyms.

As illustrated in **Figure 2**, there are three perception channels similar to neural circuits in neuroscience. In the first channel, the low-level texture features are used to model the aesthetic properties of the affective layer. In the second channel, the properties of the judgment layer are modeled using low-level features and aesthetic properties of the affective layer. Finally, the properties of the emotional layer are built accordingly by inputs from all previous layers and low-level features.

We set $M_p = \left\{A_i^p, B_j^p, C_k^p, \ldots\right\}$ as the low-level feature set of the p^{th} visual texture, where $i = 1, 2 \ldots n$, $j = 1, 2 \ldots s$, $k = 1, 2 \ldots t$ represents the number of different texture feature subsets A, B, C, etc. The aesthetics values of the affective layer, the judgment layer and the emotional layer for the p^{th} visual texture are represented by $G_p = \left\{g_1^p; g_2^p; g_3^p\right\}$, $T_P = \left\{t_1^p; t_2^p; t_3^p; t_4^p\right\}$, and $Q_p = \left\{q^p\right\}$, respectively. Considering the ideas conveyed in **Figure 2**, we employ six activation functions to construct the three perception channels.

The perception model of the affective layer is given by:

$$G = F_1(M) + R_0 \tag{8}$$

that of the judgment layer by:

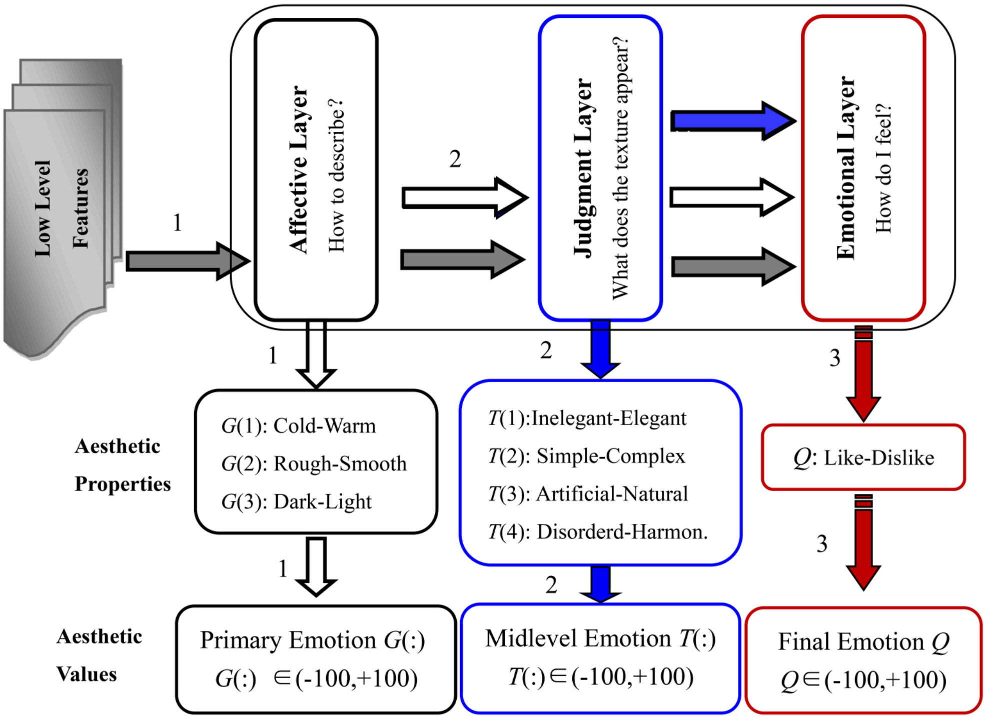


FIGURE 2 | Structure of the hierarchical feed-forward model of aesthetic perception of visual texture, the horizontal arrows indicate input flows to the different layers: e.g., the inputs to the judgment layer are low-level features (gray arrow) plus properties of the affective layer (white arrow).

$$T = F_2(M) + F_3(G) + R_1 \tag{9}$$

and that of the emotional layer by:

$$Q = F_4(M) + F_5(G) + F_6(T) + R_2 \tag{10}$$

where F_1, F_2, F_3, F_4, F_5, and F_6 are the six activation functions that can be linear or non-linear, and R_0, R_1, and R_2 refer to the emotion thresholds. The symbol "+" indicates emotions accumulated through different perception stages. Note that in our model-building cycles (as explained in the Results section), a particular set of activation functions best suited to the problem at hand is automatically applied. A standard procedure consists of a weighted linear combination of these activation functions where the weights are derived by least-squares optimization to obtain an optimal solution within a closed analytical formula, see Ljung (1999) or Lughofer (2011).

When aesthetic emotions are predicted for new incoming textures, the adjectives in the affective layer G(1), G(2), and G(3) are predicted using the low-level feature set stored in M and applying the activation function F_1. Next, the adjectives in the judgment layer T(1), T(2), T(3), and T(4) are predicted using the low-level feature set M and the predicted adjective values G(1) to G(3) by applying activation functions F_2 and F_3. Finally, the emotional layer adjective ("like-dislike") is predicted using the low-level feature set M, the predicted adjective values G(1) to G(3) and the predicted values T(1), T(2), T(3), and T4) by applying activation functions F_4, F_5, and F_6. Alternatively, if adjective values for G(1) to G(3) and/or T(1) to T(4) are already given by humans, these can be used in place of the predictions.

RESULTS AND DISCUSSION

The Selected Top Features

After feature selection, the original 106-D features were ranked according to their Laplacian scores. The first and most important 15 features are listed in **Table 2**.

The first 15 texture features are listed in **Table 2** according to their Laplacian score in ascending order. The first feature is the mean saturation, which is extracted from the HSV space. The number of contrast and homogeneity calculated using GLCMs is eight, which accounts for 53%. The ranks of coarseness and directionality are the seventh and the eighth, respectively. The ranks of the wavelet-based energy texture features (L^1 norm and L^2 norm) calculated from the horizontal and vertical sub-band at the first level are the ninth, the tenth and the fifth.

Visualization of the Selected Features

The magnitudes of the features extracted using the algorithms mentioned in Section Materials and Methods are different. In **Figure 3**, the feature set after normalizing to the interval $[-1, 1]$ using feature scaling method is visualized. In our experiment, 151 selected visual textures are used and 106 features are calculated for each visual texture. So, the size of the feature database is 151×106.

TABLE 2 | Feature list after selection using the Laplacian score.

ID	Laplacian score	Category	Parameters	Name
f_1	0.9742	Color characteristics	Mean of saturation	Mean of saturation
f_2	0.9101	GLCMs	$d = 8, \theta = 45°$	Contrast
f_3	0.8995	GLCMs	$d = 6, \theta = 45°$	Contrast
f_4	0.8855	GLCMs	$d = 8, \theta = 135°$	Contrast
f_5	0.8785	GLCMs	$d = 8, \theta = 90°$	Contrast
f_6	0.8778	GLCMs	$d = 4, \theta = 45°$	Contrast
f_7	0.8690	Tamura texture		Coarseness
f_8	0.8656	Tamura texture		Directionality
f_9	0.8551	Wavelet-based energy	horizontal sub-band at level 1	L^2 norm
f_{10}	0.8434	Wavelet-based energy	vertical sub-band at level 1	L^2 norm
f_{11}	0.8434	Tamura texture		Contrast
f_{12}	0.8427	GLCMs	$d = 8, \theta = 45°$	Homogeneity
f_{13}	0.8367	GLCMs	$d = 8, \theta = 135°$	Homogeneity
f_{14}	0.8306	GLCMs	$d = 6, \theta = 45°$	Homogeneity
f_{15}	0.8282	Wavelet-based energy	horizontal sub-band at level 1	L^1 norm

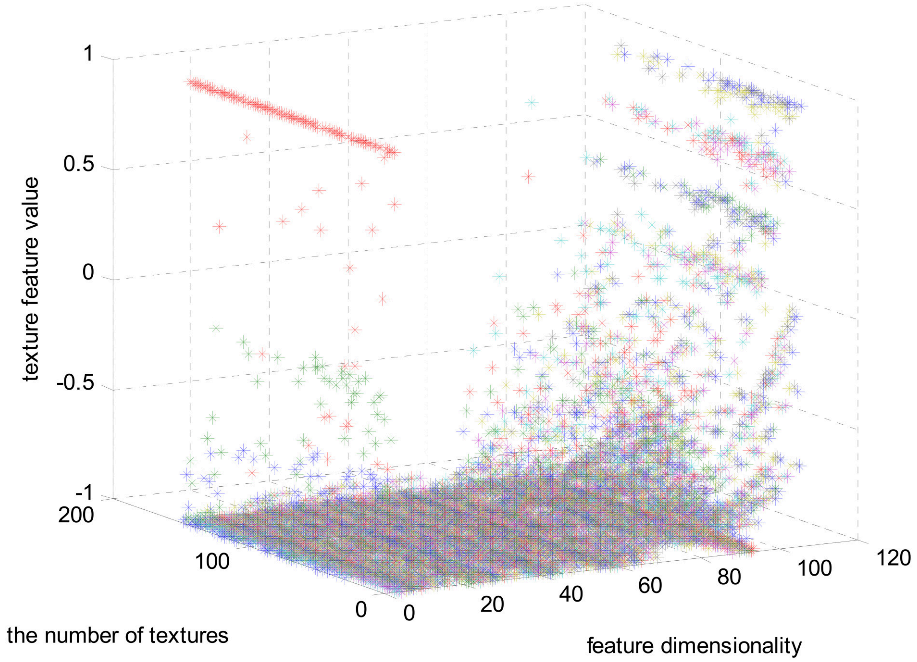


FIGURE 3 | The full feature matrix comprising 106 features and 151 textures.

In **Figure 3**, one color represents each type of features that locate in each dimensionality. We can find that the majority of the feature values compactly locates at the bottom of the space and only a few sparsely scatter among the concentrated feature stripes. One possible conclusion is that the features extracted using the algorithms mentioned in Section Materials and Methods are highly redundant, correlative and there is a quite low diversity of the features. In order to further examine this issue, the cross correlation coefficients of the 106-D features are calculated and illustrated in **Figure 4**. There are 1370 correlation coefficients that are larger than 0.75 in their absolute values, which accounts for 12.19% in total.

The first 10 features are visualized in **Figure 5**.

In **Figure 5**, the normalized 10 features regularly locate in the feature space. The features in the first feature vector, are much larger than the left ones. We also found that the first 10 selected feature vectors can be divided into two clusters, which locate on two poplars of the feature space. The structural risk and the computation complexity of the model will be under constraints through controlling the number of features that are used as inputs. Thus, in the model building process, the features with a Laplacian score lower than 0.85 are not used, which means that we used the first 10 features to build the aesthetic perception model.

Building a Model of Aesthetic Perception

Below, we discuss model building by means of Eureqa Desktop. Eureqa is a tool that uses a recent breakthrough in machine learning to unpick intrinsic relationships within complex data and explains them as simple mathematical formulas (Schmidt and Lipson, 2009). When the target expressions are defined by Equations (8–10), the basic, trigonometric and exponential functions are selected in the formula building blocks of Eureqa Desktop. In detail, the basic functions include addition, subtraction, multiplication, division and the constant operation. The trigonometric functions include sine, cosine and tangent functions. The exponential functions include exponential, natural logarithmic, factorial, power and square-root functions.

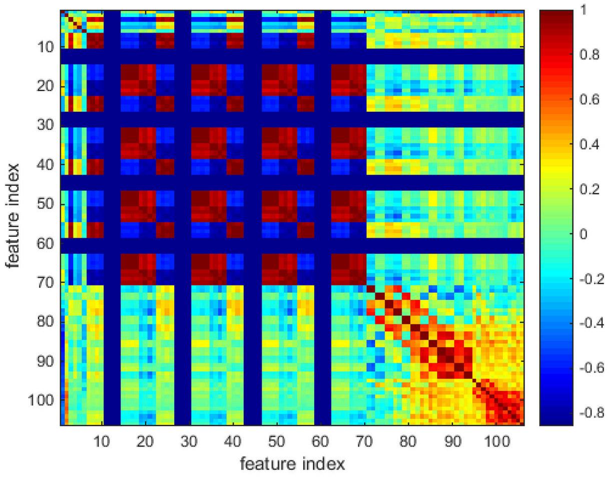


FIGURE 4 | The colored cross correlation coefficients matrix.

Before model building, the 10 selected features and emotion values were smoothed, outliers removed and normalized with the default algorithms embedded in Eureqa Desktop. The 151 visual textures were divided into two sets. One set is for model building and the other is for model test. The training set included 90% of the total number of textures, and was used for model building. The test set was used to evaluate the performance of the models built on the training set, to measure the expected quality on new textures.

We used a parameter called R^2 goodness of fit to evaluate the quantitative goodness of fit between each model and the used data. The model with the greatest R^2 is considered to be the best. The models thus selected for the eight pairs of aesthetics properties distributed in the hierarchical feed-forward model are:

$$\begin{aligned} G(1) &= 0.03 \cdot f_1 + 598.16 \cdot f_3 - 234.19 \cdot f_5 - 348.67 \cdot f_7 \cdots \\ &\cdots + 189.82 \cdot f_9 - 304.29 \cdot f_{10} + 22.17 \end{aligned} \tag{11}$$

$$G(2) = -1.31 \times 10^{-13} \cdot f_{10} - 1.73 \tag{12}$$

$$\begin{aligned} G(3) &= 0.29 \cdot f_1 + 83.14 \cdot f_7 - 53.82 \cdot f_5 - 214.60 \cdot f_9 \cdots \\ &\cdots - 0.34 \cdot f_1 \cdot f_7 + 216.71 \cdot f_9^{\wedge 2} + 19.22 \end{aligned} \tag{13}$$

$$\begin{aligned} T(1) &= 0.03 \cdot f_1 + 119.55 \cdot f_6 - 106.46 \cdot f_7 + 18.21 \cdot f_9 \\ &- 51.51 \cdot f_{10} + 0.76 \cdot G(3) + 2.86 \end{aligned} \tag{14}$$

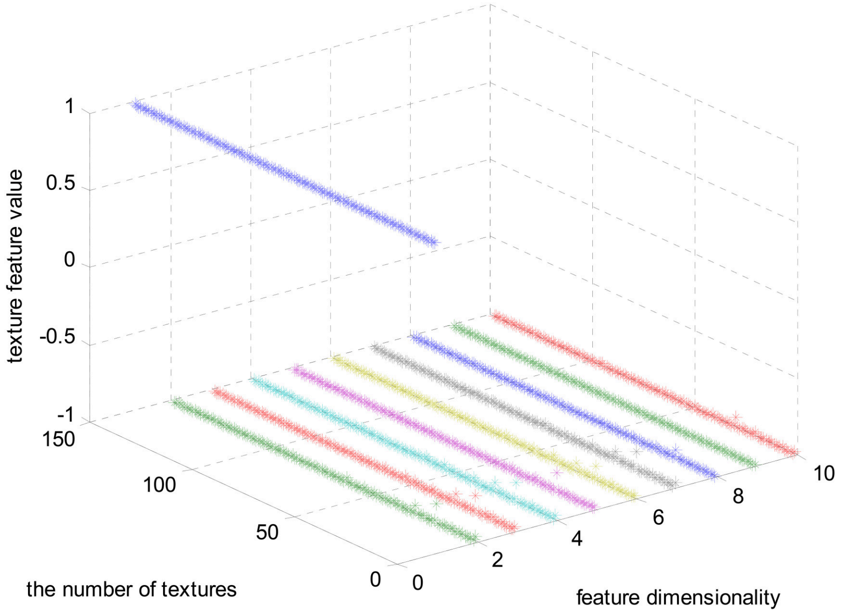


FIGURE 5 | The first 10 selected features.

$$T(2) = 0.09 \cdot f_1 + 691.81 \cdot f_3 - 737.98 \cdot f_4 - 609.32 \cdot f_6 + 683.64 \cdot f_7 \cdots - 66.11 \cdot f_8 + 0.33G(3) - 9.60 \quad (15)$$

$$T(3) = 1135.64 \cdot f_2 - 1123.31 \cdot f_3 - 582.81 \cdot f_4 + 571.40 \cdot f_7 \cdots \cdots - 364.15 \cdot f_8 + 376.91 \cdot f_9 - 14.60 \quad (16)$$

$$T(4) = 185.17 \cdot f_2 + 68.26 \cdot f_3 - 215.02 \cdot f_4 - 99.79 \cdot f_8 \cdots \cdots + 129.36 \cdot f_9 - 0.15 \cdot G(1) + 24.37 \quad (17)$$

$$Q = 123.09 \cdot f_2 - 144.83 \cdot f_3 - 113.49 \cdot f_8 - 148.05 \cdot f_9 \cdots \cdots + 0.64 \cdot T(1) + 0.12 \cdot T(3) + 1.14 \quad (18)$$

where $f_i, i = 1, 2 \cdots 10$ represents the 10 features selected using the Laplacian score algorithm.

In fact, 13 different non-linear terms are chosen for model building in Eureqa, which automatically selects those terms which are most feasible for establishing a high quality model (within a cross-validation procedure). During cross-validation, the training set is split into different folds, and always a separate test fold is used to elicit the error for each training fold combination. According to Hastie et al. (2009), CV is a good method to estimate the expected prediction error on future samples well. Furthermore, in order to overcome over-fitting, we studied how the models listed above performed on a separate test set. We should note the number of variables used to build each model is different. In detail, an input dimensionality of 10 in case of G(1) to G(3), of 13 in case of T(1) to T(4) and of 17 in case of Q.

Surprisingly, we found out that for all models linear terms were sufficient to reach the highest possible quality in terms of R^2 goodness of fit for explaining the targets. The exception was for the model for G(3), which uses two quadratic terms. However, these do not boost the quality of this model (cf. **Table 3**). This is the most noteworthy results of our experiment, as it keeps the model complexity low and thus emphasizes high interpretability capability. Even though the low-level texture features were integrated using non-linear models, the models bridging the gap between computational texture features and aesthetic texture properties turned out to be linear. Additionally, Equations (14–18) indicate that the higher level aesthetic properties in the judgment layer and emotional layers cover—with the exception of the texture features—the aesthetic properties in the lower-level layer. Interestingly, G(3) is an important adjective in the models for T(1) and T(2), whereas T(1) and T(3) have a direct influence on the "like-dislike" feeling.

The R^2 goodness of fit values of the eight models [shown in Equations (11–18)], are listed in **Table 3**. Complexity, correlation coefficient and the root mean squared error are also provided to fully evaluate goodness of fit and predictive power. Complexity is important to measure the model's capability in terms of interpretability because of higher complex models are always suffering from interpretability. The root mean squared error shows the expectation deviation between observed and predicted aesthetic property values. Correlation coefficient denotes the correlation between predicted and observed values. Thus, a value close to 1 indicates a nearly perfect prediction; usually, a value of 0.5 and below denotes a useless model. Eureqa's complexity metric (or size) is measured by the number of variables and the relative weights of each of the building blocks used in the solution. This is referred to as "enhanced complexity" in **Table 3**. Additionally, we report the basic complexity, which is simply the number of input terms in each model. These values are directly comparable with the values in Thumfart et al. (2011) and are directly related to the transparency and understandability of the model (a model with 100 terms can be hard to read and understood, for instance).

In **Table 3**, the R^2 (goodness of fit) for $G(2)$, $T(1)$, $T(2)$, $T(3)$, and Q are greater than 0.8. In other words, the models for $G(2)$, $T(1)$, $T(2)$, $T(3)$, and Q are instantiated that provide suitable representations of the aesthetic perceptions. However, it can be seen that the R^2 goodness of fit values for $G(1)$, $T(4)$ and particularly $G(3)$ are obviously lower than those of the other models. And, the MSEs of $G(1)$ and $G(3)$ are significantly greater when compared with the others. The models for T(1) and Q are fully useable and highly precise in case when real G(3) values are available for new textures. Another finding is that our new models based on specifically selected features can significantly outperform the models proposed by Thumfart in terms of prediction error (much lower MSE values).

Note that in model training and evaluation cycles, we always used the original data gotten from the semantic differential. In particular, for establishing a model for T(1) and Q, which both use G(3) as input, the original G(3) values from the interview data were used and not the predictions of the G(3) model (which were particularly poor as can be seen in **Table 3**). Model building and the final models for T(1) and Q were therefore not affected.

The aesthetics properties predicted using these models [according to Equations (11–18)] and the values from the interview-based test set that comprises 14 textures are plotted in **Figures 6–8** for the three most interesting and challenging properties "artificial-natural," "disordered-harmonious," and "like-dislike." The statistical measures of the predicted and real aesthetic property values from interviews for the test set are given in **Table 4**.

TABLE 3 | Statistical measures and qualities of models on the training data set (CV-based), the results *after the slashes* correspond to the results reported in (Thumfart et al., 2011) (if available), we offer two additional models for disordered-harmonious (T(4)) and dark-light (G(3)).

Aesthetic property	R^2	Basic complexity	Enhanced complexity	Correlation coefficient	RMSE
G(1)	0.57	6/3	23	0.80	07.5/8.87
G(2)	1.00	1/6	5	1.00	00.00/7.83
G(3)	0.28	6	17	0.44	10.80
T(1)	0.92	6/12	19	0.97	02.42/05.02
T(2)	0.84	7/5	25	0.93	04.31/04.81
T(3)	0.82	6/6	23	0.91	03.39/04.56
T(4)	0.47	6	29	0.93	1.53
Q	0.95	6/6	23	0.98	01.55/03.35

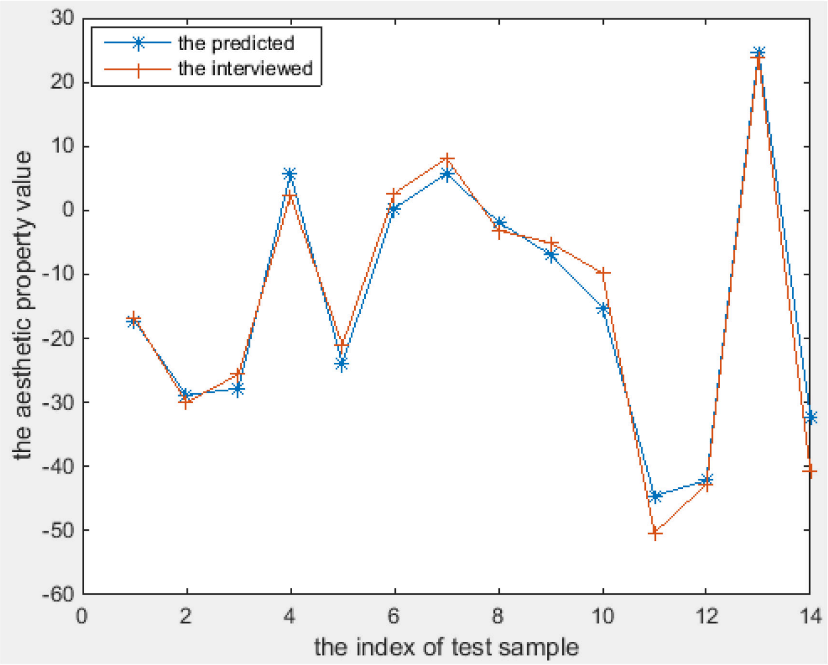


FIGURE 6 | The predicted and the interviewed test sample values for T(3).

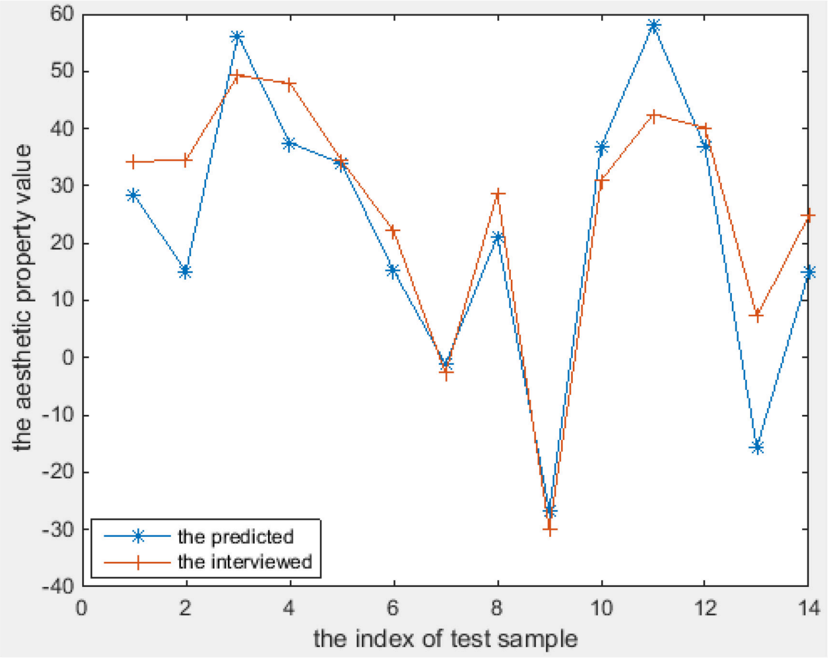


FIGURE 7 | The predicted and the interviewed test sample values for T(4).

As shown in **Figures 6–8**, the prediction power of G(1), G(2), T(1), T(2), and T(3) is better than that of G(3), T(4), and Q. However, the predictive power of G(3) and T(4) is much better on the test set than on the training set, at least for T(4). We therefore conclude that the models can be used to calculate the properties of textures. In fact, the maximum correlation coefficient of the training set is greater than that of the test set in all cases, as can be seen in **Table 3, 4**. In other words, the multiple linear regression

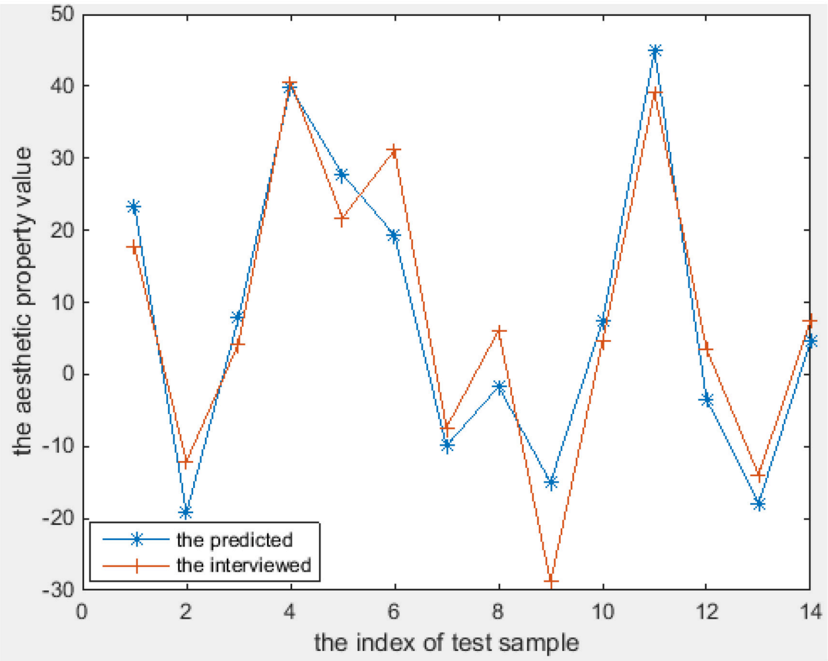


FIGURE 8 | The predicted and the interviewed test sample values for Q.

TABLE 4 | Statistical measures for the test set.

Aesthetic property	Correlation coefficient	RMSE
G(1)	0.99	1.74
G(2)	0.99	1.13
G(3)	0.76	9.28
T(1)	0.99	1.79
T(2)	0.99	2.51
T(3)	0.98	3.77
T(4)	0.94	5.68
Q	0.97	4.73

models can predict the aesthetic property well for some new visual textures that were not used in the training stage, even though the correlation coefficient on the overall training set is unsatisfactory. Another indication is that the bias is higher than the variance error, and thus over-fitting does not take place. On the other hand, we could find out that also the models for G(3) and T(4) can perform well on a subset of the whole texture set, which makes them promising for other textures collected in the future.

CONCLUSIONS

In this paper, we have proposed a hierarchical feed-forward layer structure built by multiple linear regression to investigate the relationship between human aesthetic texture perception and computational low-level texture features. Rather than black-box models, we sought to build nearly white-box models that can be interpreted both in terms of structure and interrelations between aesthetic properties and texture features according to feature weights.

First, we carried out a texture analysis and calculated 106 color and texture features for each visual texture. To achieve the best possible prediction rate and reduce the complexity of model building, feature selection using the Laplacian Score was employed to choose the best feature subset (finally comprising 10 features). Then, the aesthetic properties of a set of 151 visual textures were collected in a semantic differential experiment with 20 subjects. Eight pairs of antonyms were selected to describe aesthetic properties for emotion perception in different affective layers. Finally, we utilized multiple regression techniques employing a variety of functional terms to bridge the gap between computational texture features and aesthetic emotions in form of mappings within a hierarchical layered structure model.

The best model for each of the 8 aesthetic properties (except for the "dark-light" pair) is a linear function, even though non-linear terms were selected in Eureqa Desktop when models are initialized. Furthermore, these built models are in low dimensionality. In other words, the models only use a quite low number of terms, namely 7 maximal, and in most cases 6. This is helpful to the readability, interpretability and understandability for psychologists. The 8 models have lower errors than the models designed in Thumfart et al. (2011) for all aesthetic properties, which confirms the feasibility and applicability of our models in future works. Additionally, the experiment indicates that—with the exception of texture features—the higher level aesthetic properties in the judgment and emotional layers cover the aesthetic properties in the lower-level layer.

As part of future work, we will select more visual texture samples and include more subjects in the semantic differential experiment, especially to investigate the influences of the types of features and functions selected for model building. This should help to improve the lower quality models, especially that built for G(3). Additional future work will include:

1. Considering more complex non-linear regression modeling architectures (rather than plain transformations), especially regression trees and/or fuzzy systems, which both offer interpretability from another viewpoint. Their structures are readable as IF-THEN rules and provide better insights into the relations between input features and targets.
2. Perception modeling that considers different groups of people, e.g., a gender study or a study with respect to age, education etc.: the interview data is split into different groups and a model is created for each group. This could provide interesting answers to questions such as "Do women or men rate textures more consistently?" or "Do women or men trigger creation of different models?"

ACKNOWLEDGMENTS

We would like to thank the anonymous reviewers and the handling associate editor for their insightful comments. This project was supported by the "National Natural Science Foundation of China" [Grant number 61203364]. The second author also acknowledges support of the Austrian COMET-K2 programme of the Linz Center of Mechatronics (LCM), which is funded by the Austrian federal government and the federal state of Upper Austria. We thank Dr. Honglin Xu and Dr. Xueyan Wei for their help with the algorithm design and the psychological experiment. This publication reflects only the authors' views.

REFERENCES

Augello, A., Infantino, I., Pilato, G., Rizzo, R., and Vella, F. (2013). Binding representational spaces of colors and emotions for creativity. *Biol. Inspired Cogn. Archit.* 5, 64–71. doi: 10.1016/j.bica.2013.05.005

Axelsson, Ö. (2007). Individual differences in preferences to photographs.pdf. *Psychol. Aesthetics Creat. Arts* 1, 61–72. doi: 10.1037/1931-3896.1.2.61

Baraldi, A., and Parmiggiani, F. (1995). An investigation of the textural characteristics associated with gray level cooccurrence matrix statistical parameters. *IEEE Trans. Geosci. Remote Sens.* 33, 293–304. doi: 10.1109/36.377929

Bharati, M. H., Liu, J. J., and MacGregor, J. F. (2004). Image texture analysis: methods and comparisons. *Chemom. Intell. Lab. Syst.* 72, 57–71. doi: 10.1016/j.chemolab.2004.02.005

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1993). *Classification and Regression Trees.* Boca Raton, FL: Chapman and Hall.

Brooks, R. R., Grewe, L., and Iyengar, S. S. (2001). Recognition in the wavelet domain: a survey. *J. Electron. Imaging* 10, 757–784. doi: 10.1117/1.1381560

Bundgaard, P. F. (2014). Feeling, meaning, and intentionality—a critique of the neuroaesthetics of beauty. *Phenomenol. Cogn. Sci.* doi: 10.1007/s11097-014-9351-5. Available online at: http://link.springer.com/article/10.1007/s11097-014-9351-5

Castelli, V., and Bergman, L. D. (2002). *Image Databases: Search and Retrieval of Digital Imagery*. *The Second.* New York, NY: JOHN Wiley & SONS, INC.

Cela-Conde, C. J., García-Prieto, J., Ramasco, J. J., Mirasso, C. R., Bajo, R., Munar, E., et al. (2013). Dynamics of brain networks in the aesthetic appreciation. *Proc. Natl. Acad. Sci. U.S.A.* 110, 10454–10461. doi: 10.1073/pnas.1302855110

Cela-Conde, C. J., Agnati, L., Huston, J. P., Mora, F., and Nadal, M. (2011). The neural foundations of aesthetic appreciation. *Prog. Neurobiol.* 94, 39–48. doi: 10.1016/j.pneurobio.2011.03.003

Chandrashekar, G., and Sahin, F. (2014). A survey on feature selection methods. *Comput. Electr. Eng.* 40, 16–28. doi: 10.1016/j.compeleceng.2013.11.024

Chatterjee, A., and Vartanian, O. (2014). Neuroaesthetics. *Trends Cogn. Sci.* 18, 370–375. doi: 10.1016/j.tics.2014.03.003

Chuang, Y., and Chen, L. (2008). How to rate 100 visual stimuli efficiently. *Int. J. Des.* 2, 31–43.

Datta, R., Joshi, D., Li, J., and Wang, J. (2006). Studying aesthetics in photographic images using a computational approach. *Lect. Notes Comput. Sci.* 3953, 288–301. doi: 10.1007/11744078_23

Davis, L. S., Johns, S. A., and Aggarwal, J. K. (1979). Texture analysis using generalized co-occurrence matrices. *IEEE Trans. Pattern Anal. Mach. Intell.* 1, 251–259. doi: 10.1109/TPAMI.1979.4766921

Do, M. N., and Vetterli, M. (2002). Wavelet-based texture retrieval using generalized Gaussian density and Kullback-Leibler distance. *IEEE Trans. Image Process.* 11, 146–158. doi: 10.1109/83.982822

Dong, Y., and Ma, J. (2011). Wavelet-based image texture classification using local energy histograms. *Signal Process. Lett. IEEE* 18, 247–250. doi: 10.1109/LSP.2011.2111369

Elkharraz, G., Thumfart, S., Akay, D., Eitzinger, C., and Henson, B. (2014). Making tactile textures with predefined affective properties. *IEEE Trans. Affect. Comput.* 5, 57–70. doi: 10.1109/T-AFFC.2013.21

Graham, D. J., and Meng, M. (2011). Artistic representations: clues to efficient coding in human vision. *Vis. Neurosci.* 28, 371–379. doi: 10.1017/S0952523811000162

Groissboeck, W., Lughofer, E., and Thumfart, S. (2010). Associating visual textures with human perceptions using genetic algorithms. *Inf. Sci.* 180, 2065–2084. doi: 10.1016/j.ins.2010.01.035

Guo, X., Asano, C., Asano, A., Kurita, T., and Li, L. (2012). Analysis of texture characteristics associated with visual complexity perception. *Opt. Rev.* 19, 306–314. doi: 10.1007/s10043-012-0047-1

Guyon, I., and Elisseeff, A. (2003). An introduction to variable and feature selection. *J. Mach. Learn. Res.* 3, 1157–1182.

Hanada, M. (2013). Analyses of color emotion for color pairs with independent component analysis and factor analysis. *Color Res. Appl.* 38, 297–308. doi: 10.1002/col.20750

Haralick, R., Shanmugam, K., and Dinstein, I. (1973). Textural features for image classification. *IEEE Trans. Syst. Man Cybern.* 3, 610–621. doi: 10.1109/TSMC.1973.4309314

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference and Prediction, 2nd Edn*. New York, NY; Berlin; Heidelberg: Springer.

He, X., Cai, D., and Niyogi, P. (2005). "Laplacian score for feature selection," in *Advances in Neural Information Processing Systems, Vol. 18 (NIPS'05)* (Cambridge, MA; London: MIT Press), 1–8.

Ishizu, T., and Zeki, S. (2013). The brain's specialized systems for aesthetic and perceptual judgment. *Eur. J. Neurosci.* 37, 1413–1420. doi: 10.1111/ejn.12135

Jiang, W., Loui, A. C., and Cerosaletti, C. D. (2010). "Automatic aesthetic value assessment in photographic images," in *2010 IEEE International Conference on Multimedia and Expo* (Suntec City: IEEE), 920–925.

Karu, K., Jain, A., and Bolle, R. (1996). Is there any texture in the image? *Pattern Recognit.* 29, 1437–1446. doi: 10.1016/0031-3203(96)00004-0

Leder, H., Belke, B., Oeberst, A., and Augustin, D. (2004). A model of aesthetic appreciation and aesthetic judgments. *Br. J. Psychol.* 95, 489–508. doi: 10.1348/0007126042369811

Leder, H., Gerger, G., Dressler, S. G., and Schabmann, A. (2012). How art is appreciated. *Psychol. Aesthetics Creat. Arts* 6, 2–10. doi: 10.1037/a0026396

Levinson, J. (2006). *Contemplating Art*. Oxford: Oxford Scholarship Online Monographs.

Liu, J., Zuo, B., Zeng, X., Vroman, P., and Rabenasolo, B. (2011). Expert Systems with Applications Wavelet energy signatures and robust Bayesian neural network for visual quality recognition of nonwovens. *Expert Syst. Appl.* 38, 8497–8508. doi: 10.1016/j.eswa.2011.01.049

Ljung, L. (1999). *System Identification: Theory for the User*. New Jersey, NJ: Prentice Hall PTR; Prentic Hall Inc.; Upper Saddle River.

Lucassen, M. P., Gevers, T., and Gijsenij, A. (2011). Texture affects color emotion. *Color Res. Appl.* 36, 426–436. doi: 10.1002/col.20647

Lughofer, E. (2011). *Evolving Fuzzy Systems—Methodologies, Advanced Concepts and Applications*. Berlin; Heidelberg: Springer.

Man, D., Wei, D., and Chih-Chieh, Y. (2013). Product color design based on multi-emotion. *J. Mech. Sci. Technol.* 27, 2079–2084. doi: 10.1007/s12206-013-0518-8

Manfredi, L. R., Saal, H. P., Brown, K. J., Zielinski, M. C., Dammann, J. F. III., Polashock, V. S., et al. (2014). Natural scenes in tactile texture. *J. Neurophysiol.* 111, 1792–1802. doi: 10.1152/jn.00680.2013

Ou, L.-C., Luo, M. R., Woodcock, A., and Wright, A. (2004a). A study of colour emotion and colour preference. Part I: colour emotions for single colours. *Color Res. Appl.* 29, 232–240. doi: 10.1002/col.20010

Ou, L.-C., Luo, M. R., Woodcock, A., and Wright, A. (2004b). A study of colour emotion and colour preference. Part II: colour emotions for two-colour combinations. *Color Res. Appl.* 29, 292–298. doi: 10.1002/col.20024

Ou, L.-C., Ronnier Luo, M., Sun, P.-L., Hu, N.-C., Chen, H.-S., Guan, S.-S., et al. (2012). A cross-cultural comparison of colour emotion for two-colour combinations. *Color Res. Appl.* 37, 23–43. doi: 10.1002/col.20648

Palmer, S. E., Schloss, K. B., and Sammartino, J. (2013). Visual aesthetics and human preference. *Annu. Rev. Psychol.* 64, 77–107. doi: 10.1146/annurev-psych-120710-100504

Roberti, F., Siqueira, D., Robson, W., and Pedrini, H. (2013). Multi-scale gray level co-occurrence matrices for texture description. *Neurocomputing* 120, 336–345. doi: 10.1016/j.neucom.2012.09.042

Romani, S., Sobrevilla, P., and Montseny, E. (2012). Variability estimation of hue and saturation components in the HSV space. *Color Res. Appl.* 37, 261–271. doi: 10.1002/col.20699

Schmidt, M., and Lipson, H. (2009). Distilling free-form natural laws from experimental data. *Science* 324, 81–85. doi: 10.1126/science.1165893

Simmons, D. R. (2012). “Colour and emotion,” in *New Directions in Colour Studies*, eds C. P. Biggam, C. A. Hough, C. J. Kay, and D. R. Simmons (Amsterdam: John Benjamins), 395–414.

Skedung, L., Arvidsson, M., Chung, J. Y., Stafford, C. M., Berglund, B., and Rutland, M. W. (2013). Feeling small: exploring the tactile perception limits. *Sci. Rep.* 3, 1–6. doi: 10.1038/srep02617

Tabakhi, S., Moradi, P., and Akhlaghian, F. (2014). An unsupervised feature selection algorithm based on ant colony optimization. *Eng. Appl. Artif. Intell.* 32, 112–123. doi: 10.1016/j.engappai.2014.03.007

Tamura, H., Mori, S., and Yamawaki, T. (1978). Textural features corresponding to visual perception. *Syst. Man Cybern. IEEE Trans.* 75, 460–473. doi: 10.1109/TSMC.1978.4309999

Thumfart, S., Heidl, W., Scharinger, J., and Eitzinger, C. (2009). “A quantitative evaluation of texture feature robustness and interpolation behaviour,” in *Proceedings of The 13th International Conference on Computer Analysis of Images and Patterns* (Münster), 1154–1161.

Thumfart, S., Jacobs, R., Lughofer, E., Eitzinger, C., Cornelissen, F. W., Groissboeck, W., et al. (2011). Modeling human aesthetic perception of visual textures. *ACM Trans. Appl. Percept.* 8, 1–29. doi: 10.1145/2043603.2043609

Toet, A., Henselmans, M., Lucassen, M. P., and Gevers, T. (2011). Emotional effects of dynamic textures. *Iperception* 2, 969–991. doi: 10.1068/i0477

Vapnik, V. (1998). *Statistical Learning Theory*. New York, NY: John Wiley & Sons.

Zeki, S. (2002). Trying to make sense of art. *Nature* 418, 918–919. doi: 10.1038/418918a

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.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read, for greatest visibility

COLLABORATIVE PEER-REVIEW
Designed to be rigorous – yet also collaborative, fair and constructive

FAST PUBLICATION
Average 85 days from submission to publication (across all journals)

COPYRIGHT TO AUTHORS
No limit to article distribution and re-use

TRANSPARENT
Editors and reviewers acknowledged by name on published articles

SUPPORT
By our Swiss-based editorial team

IMPACT METRICS
Advanced metrics track your article's impact

GLOBAL SPREAD
5'100'000+ monthly article views and downloads

LOOP RESEARCH NETWORK
Our network increases readership for your article

Frontiers
EPFL Innovation Park, Building I • 1015 Lausanne • Switzerland
Tel +41 21 510 17 00 • Fax +41 21 510 17 01 • info@frontiersin.org
www.frontiersin.org

Find us on