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

OPEN ACCESS

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).

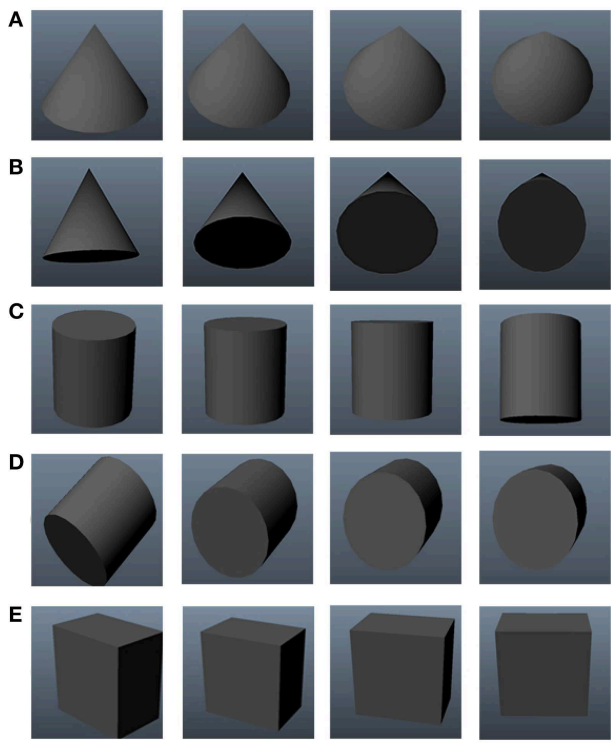

FIGURE 1 | Representative appearance changes undergone by objects during out-of-plane rotations. Variations of metric properties here include: **(A)** increasing angle at a point and **(B)** increasing size, shape and curvature of cross section of a cone **(C)** increasing size, shape and curvature of 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.

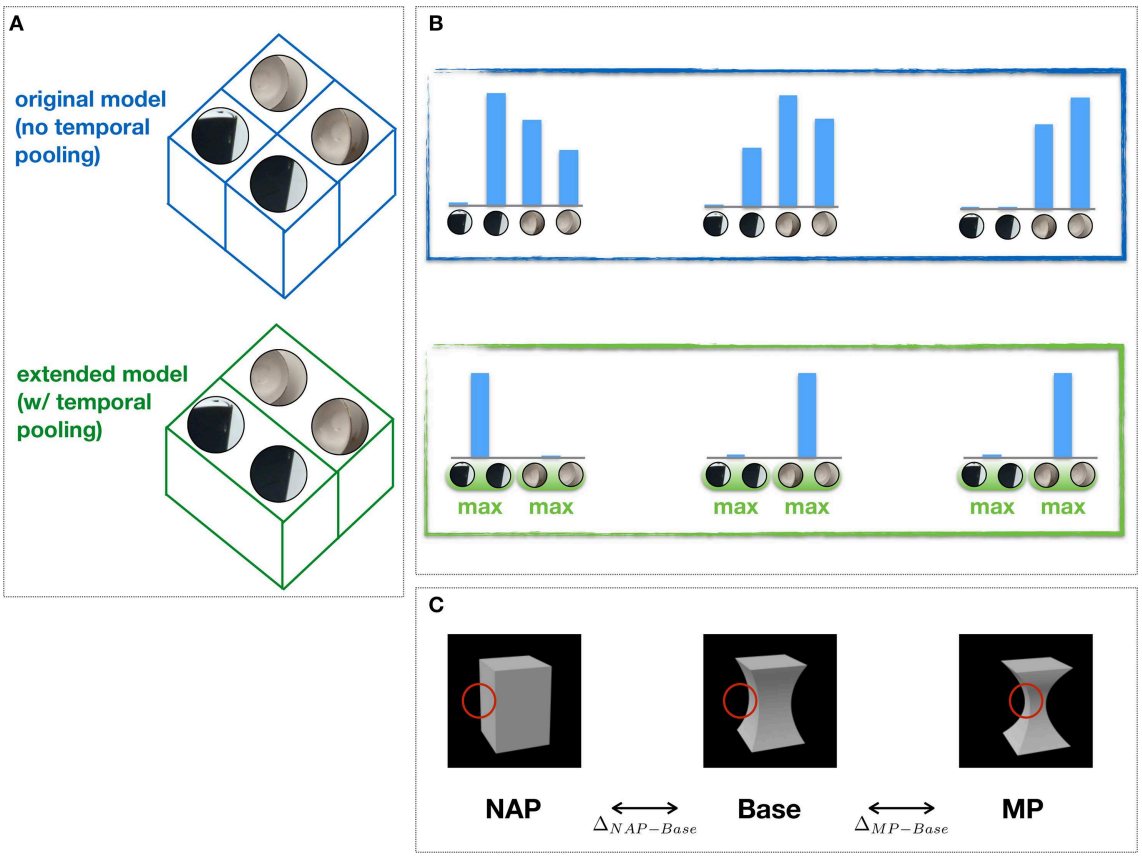

FIGURE 2 | Feature columns, invariance to object transformations and NAP sensitivity. (A) Feature columns in the extended (bottom) vs. the original (top) model (i.e., w/ and w/o temporal pooling). One of the key computational mechanisms in the HMAX builds on the proposal by Hubel and Wiesel (1962) to 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

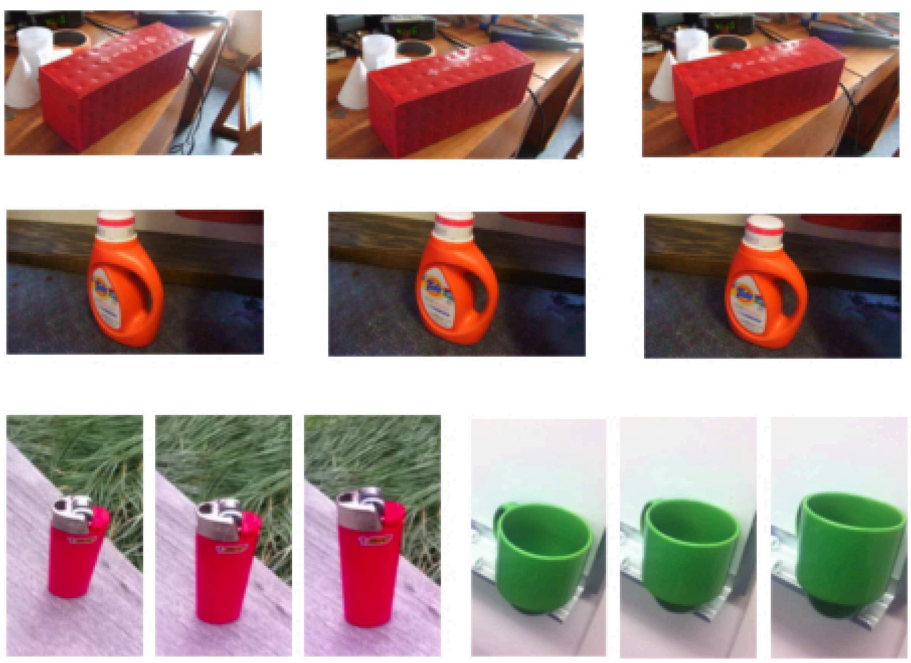

FIGURE 3 | Representative frames sampled from a collected video database of everyday objects undergoing 3D transformations (i.e., combination of translation, scaling, and in-depth rotation).

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

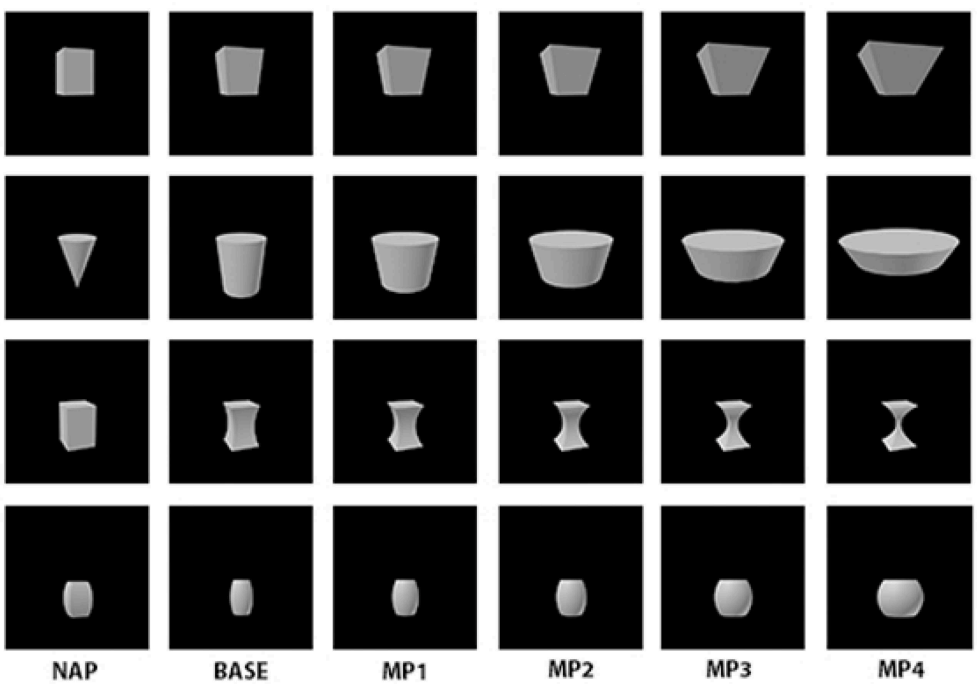

FIGURE 4 | Sample stimuli from the study by Kayaert et al. (2003). The column labeled BASE corresponds to a reference image. The column labeled NAP corresponds to a transformation of the base image where an NAP was changed. MP1, MP2, MP3, and MP4 correspond to a transformation of the base image with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

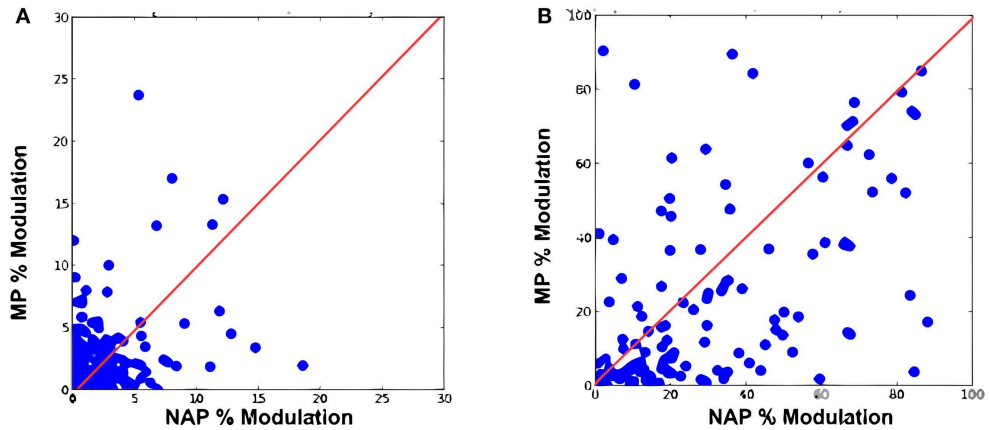

FIGURE 5 | Percent modulation of C_2 units from the base image to MP2 vs. percent modulation of C_2 units from the base image to NAP for (A) original HMAX (B) the extended model.

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 Körner, 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.