Common object representations for visual recognition and production

Judith E. Fan
Department of Psychology
Princeton University
jefan@princeton.edu

Daniel L. K. Yamins
McGovern Institute for Brain Research
Massachusetts Institute of Technology
yamins@mit.edu

Nicholas B. Turk-Browne
Department of Psychology
Princeton University
ntb@princeton.edu

Abstract

What is the relationship between recognizing objects and drawing objects? We examine the possibility that both functions are supported by a common internal representation. First, we show that a model of ventral visual cortex only optimized to recognize objects in photographs generalizes to drawings of objects, suggesting that the capacity for visual abstraction is rooted in the functional architecture of the visual system. Next, we tested whether practice drawing objects might alter how those and other objects are represented. On each trial, participants sketched an object. The model then guessed the identity of the sketched object, providing real-time feedback. We found that repeatedly sketched objects were better recognized after training, while sketches of unpracticed but similar objects worsened. These results show that *visual production* can reshape the representational space for objects: by differentiating trained objects and merging other nearby objects in the space.

Keywords: communication; drawing; learning; perception and action; computer vision

Introduction

Although the retinal images cast by physical objects and line drawings differ dramatically, humans effortlessly recognize objects in either format. How does the brain accomplish this feat of visual abstraction? Moreover, with just a few well-placed strokes, humans are able to communicate abstract ideas (e.g., object identity) by drawing. What are the mechanisms that underlie the ability to produce a sketch that represents an object?

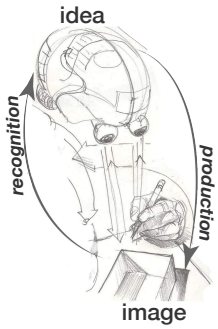

Figure 1: Visual recognition entails mapping an image onto a specific idea (e.g., object identity based on a photograph). Visual production entails expressing a specific idea in an image (e.g., identifiable sketch based on an object concept). (Art credit: Jeffery Thompson)

Here we examine the possibility that the ability to recognize objects and produce drawings of objects are linked by a common internal substrate — a *generalized object representation*. This premise is plausible, given the reciprocal functions of recognition and production (Fig. 1). Specifically, visual recognition entails mapping an image from the external world onto a specific idea in mind; visual production entails expressing a specific idea in mind as an image residing in the external world. We test this hypothesis by evaluating two predictions it makes: (1) that recognition of both objects and human drawings can be achieved by a common visual feature representation; and (2) that training on a novel drawing task can alter this representation, just as training on visual recognition tasks can alter object representations (Goldstone, 1998).

Part One: Recognizing Pictures of Objects

People can recognize objects in the face of enormous variation in pose, size, position, lighting, and other sources of noise, a fact which belies the computational difficulty of this feat (Pinto, Cox, & DiCarlo, 2008). This ability is supported by a set of hierarchically organized brain regions known as the ventral visual stream (Malach, Levy, & Hasson, 2002), by which simple visual features (e.g., orientation, spatial frequency) encoded in the lowest area, V1, are successively combined and transformed in such a way as to support read out of abstract object properties (e.g., category, identity) at the top level in the hierarchy, inferior temporal (IT) cortex (Hung, Kreiman, Poggio, & DiCarlo, 2005).

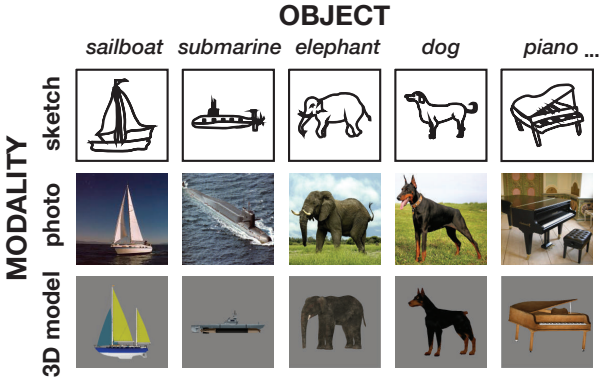

Figure 2: Multi-domain imageset containing sketches, photographs, and 3D-rendered images of 147 real-world objects.

Supplementing natural visual inputs from objects in the environment, humans have also devised a wide range of technologies for producing pictures that represent objects. The most ancient among these is drawing, whereby lines and marks are made on a surface by manipulating a stylus (Clottes, 2008). Despite large differences between drawings of objects and visual inputs from physical objects (or photorealistic images of objects), they are just as effective at evoking the real-world object (Biederman & Ju, 1988).

What commonalities across line drawings and photorealistic images (e.g., photographs, 3D-renderings) allow people to recognize the same object depicted in such different ways?

Discovering the computational principles that underlie such robust recognition is a challenge that lies at the heart of human visual abstraction. Here we present a computational approach to quantifying such commonalities (Fan, Yamins, DiCarlo, & Turk-Browne, 2014; Yamins et al., 2014).

Methods

Imageset We first assembled a multi-domain imageset containing sketches, photographs, and 3D-rendered images (Fig. 2). From an existing sketch corpus (Eitz, Hays, & Alexa, 2012), we obtained ∼12,000 sketches of 147 common, real-world objects. These sketches were produced by human participants on Amazon Mechanical Turk, who were prompted on each trial with a randomly chosen entry from a list of 250 basic-level object categories to sketch on a digital drawing canvas. From the annotated Imagenet database (Deng et al., 2009), we acquired ∼200K photographs of the same 147 objects, depicting diverse exemplars from each object class embedded in their natural backgrounds. Finally, using 3D mesh models, we rendered ∼200K synthetic images of these same objects in highly variable positions, sizes, and poses against randomly selected real-world backgrounds.

Neurally Predictive Model of Object Recognition We then applied a recently developed deep convolutional neural network model that was inspired by the functional architecture of the ventral visual stream in order to extract features from these images (Yamins et al., 2014; Fig. 3a). This model had been identified using hierarchical modular optimization (HMO), a procedure for efficiently searching among mixtures of convolutional neural networks for candidate hierarchical model architectures that achieve high performance on basic-level object recognition tasks. The HMO procedure was performed on an independent imageset containing photographs only, with no objects in common with the multi-domain imageset described above. In addition to achieving human-level performance on these tasks, the higher layers of the resulting model are also quantitatively predictive of neural population responses in high-level visual cortex (e.g., V4 and IT). As such, it was an attractive candidate for investigating the visual invariants that support recognition across image domains.

Results

The model uses a fixed, but large number of feature dimensions to represent all images. Each image elicits a pattern of feature values at every layer in the model, which may be expressed as a vector in this high-dimensional feature space. For a given image domain, we computed average feature vectors within an object class, then derived correlation matrices based on these feature vectors. This procedure was performed at each of the five layers of the model. Each matrix entry represents the proximity between the average feature vectors from the model for a pair of objects (Kriegeskorte et al., 2008). Higher values (cooler colors) reflect relatively proximal pairs of objects, whereas smaller values reflect more distant object pairs. Each 147x147 matrix provides a compact

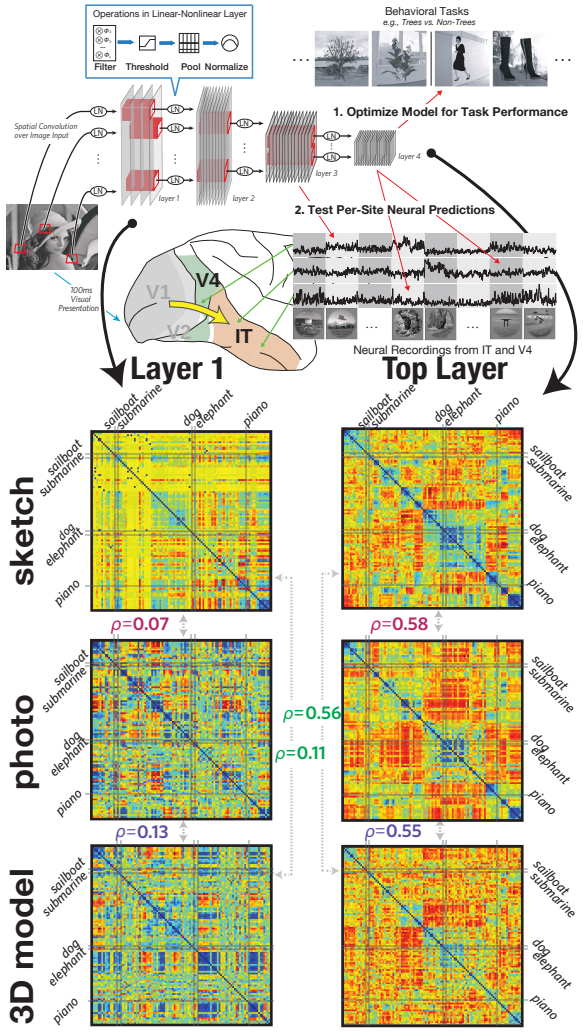

Figure 3: (a) Feature extraction using a neurally predictive, deep convolutional neural network model optimized for performance on challenging object recognition tasks. (b) Correlation matrices for each image domain, displaying the overall layout of objects in high-dimensional feature space. Each entry shows correlation distance (1-ρ) between feature vectors for a pair of objects.

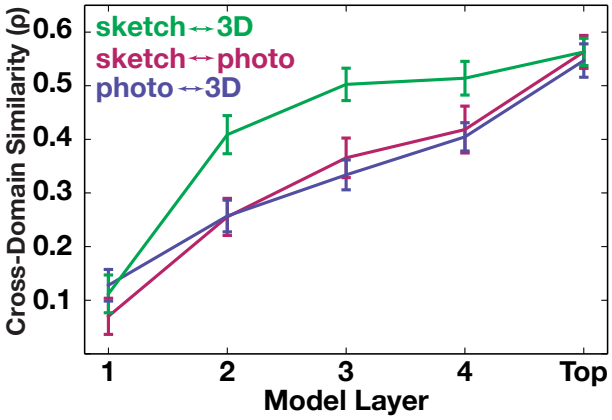

Figure 4: Cross-domain similarity (Spearman's ρ) between image domains increases as a function of model layer.

visualization of the layout of objects in the high-dimensional feature space inherent to each layer of the model, for each image domain (see Fig. 3b for first and top-layer matrices).

All matrices individually show clear block-diagonality, indicating the presence of higher-order structure due to clustering of objects with similar features.

The matrices computed based on top-layer output also show striking cross-domain similarities, both visually and as quantified by Spearman rank correlation comparisons (Fig. 3b). This indicates an underlying commonality in the feature representations for the three image modalities at the top layer in the model, the layer whose output has been previously shown to be highly predictive of neural population responses in IT cortex.

By contrast, cross-domain similarities are negligible at the lowest layer in the model, the layer approximating the local/simple features encoded in V1. This shows that low-level image statistics (e.g., edge fragments) are insufficient to explain robust recognition across image modalities, especially under conditions of high image variation.

We found that the strength of cross-domain similarities increased over successive layers in the model (Fig. 4), consistent with the understanding that the ventral visual stream computes progressively more abstract properties of objects over successive processing stages.

In sum, our results show that a hierarchical neural network model only optimized to recognize photorealistic images of objects generalized to abstract drawings of objects, having produced congruent object-similarity 'maps' across image domains based on the same visual feature representation. These results suggest that the capacity for visual abstraction may be rooted in the functional architecture of the visual system.

Part Two: Producing Drawings of Objects

How does learning refine object representations? For example, although people tend to label an object at the basic level (Mervis & Rosch, 1981), domain-specific expertise (e.g., knowledge of dogs) makes subordinate-level names (e.g., 'schnauzer') as accessible (if not more accessible) than basic-level names for objects in the domain of expertise (Tanaka & Taylor, 1991). This suggests that initially similar stimuli can become more differentiated with practice. Expertise can also lead to unitization of features that were initially processed separately. For instance, dog experts are worse at recognizing inverted images of dogs than non-experts (Diamond & Carey, 1986), suggesting that extensive experience with an object can lead to automatic binding of features into a viewpoint-specific functional unit.

Such findings suggest that training on recognition tasks can alter object representations. Here we ask whether training on *visual production* tasks can also alter representations of objects. In the previous section, we found that computations approximating those performed by the ventral visual stream produced congruent object similarity 'maps' for both photorealistic images and hand-drawn sketches of objects. Insofar as both the ability to recognize sketches and to produce recognizable sketches recruit a common internal representation, we hypothesized that practice drawing some objects (e.g., horse, cow) might affect the way that those and other, related objects (e.g., sheep) are subsequently represented.

Since recognizability is a key attribute of successful drawings, a natural starting point for examining learning is to identify objects for which untrained participants have trouble producing clearly recognizable drawings — that is, that are frequently confused with drawings of other objects. The most confusable objects are likely to be objects whose drawings share many visual features, even if the objects themselves are not semantically related, *per se* (e.g., bell and pear). To define groups of related objects (i.e., 'visual categories'), we exploit pre-existing object clusters revealed by the model in the original sketch corpus collected by Eitz et al. (2012).

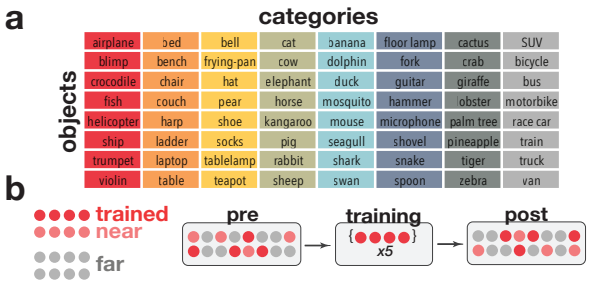

Figure 5: (a) Stimuli: Objects belonged to eight visual categories, each containing eight items. (b) Design: Each participant was randomly assigned two of these categories. During training, participants drew four randomly selected objects in one category (Trained) multiple times. Before and after training, participants drew the other four objects in that category (Near), as well as the objects in the second category (Far), once each.

Methods

Participants Six hundred and fifty-one participants were recruited via Amazon Mechanical Turk (AMT) for the drawing experiment, with sixty excluded for failing to complete the session. Participants were paid a base amount of $1.50 and up to $3.00 bonus for high task performance. Three hundred and twenty-seven additional participants were recruited (via AMT) to provide labels for the sketches from the drawing experiment, and were paid $0.85 for their participation. All provided informed consent in accordance with the Princeton IRB.

Stimuli and Design In order to identify groups of objects that are drawn similarly prior to training, we applied a clustering algorithm (affinity propagation with damping=0.9; Frey & Dueck, 2007) to the features extracted from the 147-object sketch corpus described above (Eitz et al., 2012). This yielded 16 clusters containing between 3 and 20 objects each. Among clusters containing at least 8 objects, we defined 8 visual categories containing 8 objects each (Fig. 5a). Each participant was randomly assigned two of these categories. During training, participants sketched 4 randomly selected

objects in one category (Trained) multiple times. Before and after training, participants sketched the other 4 objects in that category (Near), as well as the objects in the second category (Far), once each (Fig. 5b).

Task The sketching task was performed in the context of a game ('Guess My Sketch') in which participants teamed up with two avatars (red, blue) in order to earn points. At the start of each trial, only the red avatar was onscreen. This avatar cued participants with either an image (N=324) or word (N=267) that referred to a target object for them to sketch (Fig. 6). After cue offset, the blue avatar appeared, prompting the participant to begin sketching. Upon sketch submission, the blue avatar listed its top three guesses as to the identity of the drawn object, thus providing participants with immediate feedback about the quality of their sketch. These guesses were listed in order of confidence. Participants earned points if any of these guesses were correct, and more points the earlier the correct guess was in the list.

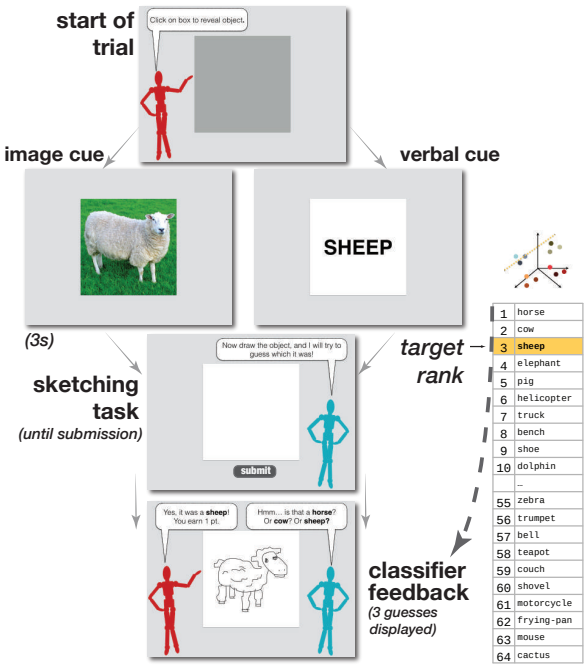

Figure 6: Task: On each trial, participants were prompted with an image (N=324) or word (N=267) that referred to a target object for them to sketch. The computer guessed the identity of the drawn object in real time, providing participants immediate feedback about the quality of their sketch. The rank of the target in the list of all 64 guesses (ordered by confidence) returned by the computer was used to track changes in performance across trials.

In the image-cue version of the task, unique photographs were used as cues on every trial in order to discourage overly stereotyped sketches. Participants were instructed to "make a sketch in which someone else is likely to recognize the object depicted," but were informed that the sketch did not have to exactly depict what was in the photo. Other than the lack of an image cue, the verbal-cue version of the task was identically structured.

Feedback We trained a 64-way support vector machine (SVM) linear classifier on model responses to photographs of the objects used in this study, but no sketches. (Thus, sketch-classification during the experiment reflects pure generalization across image domains.) On each trial, top-layer model features were extracted from the submitted sketch in real time, which were passed to the 64-way classifier to determine feedback. The classifier returned a list of 64 margin values, corresponding to the level of confidence that the test image belonged to each object class. The three objects with the most positive margin values (highest confidence) were returned to the participant as guesses. In the verbal-cue version of the task, when none of the three top guesses were correct, the rank of the target in this ordered object list was also returned to the participant (e.g.,*"Too bad... 'giraffe' would have been my 9th guess."*). Because this *target rank* value provides a consistent measure of the 'goodness-of-fit' of the submitted sketch to the target object representation in the model, this value served as our primary measure of task performance. Since the criteria for recognition by the model were fixed, we interpret changes in task performance as reflecting changes to the participants' internal object representations.

Validating Model Representation Because the conditions used by Eitz et al. (2012) to collect sketches differ somewhat from our own (e.g., only verbal cues were used, each participant sketched an object only once, and could apply 'undo', 'redo', 'clear' on their sketches), we first sought to assess the similarity between their sketch corpus and the sketches collected for this experiment. To accomplish this, we extracted features of sketches from the verbal-cue and image-cue versions of the task using top-layer output from the model.

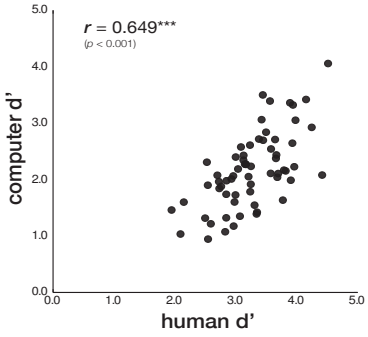

Figure 8: An independent cohort of human participants guessed the identity of objects depicted in drawings from the image-cue experiment. Human and computer recognition performance (d') was highly consistent across objects (r=0.649).

For each version of the task, we computed the average feature vectors for all sketches within an object class, then derived correlation matrices on these feature vectors. In both the image-cue and verbal-cue datasets, we found that sketches of objects within a category were highly similar, validating category assignments. The two matrices were also highly similar to each other (Spearman's ρ=0.890), suggesting that this feature representation successfully captured object iden-

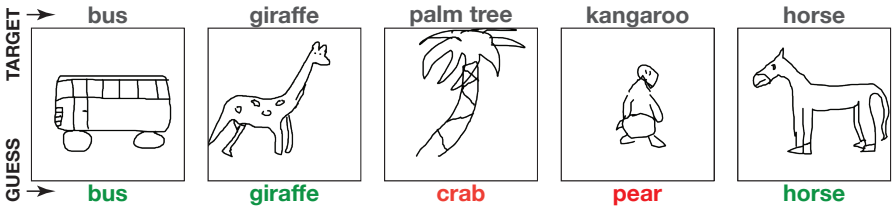

Figure 7: Sample sketches from the experiment, with target label and model's top guess.

tity despite low-level task differences. Moreover, both matrices were highly similar to the original sketch corpus (image-original: ρ=0.715; verbal-original: ρ=0.708). An independent cohort of human participants (N=327) provided three labels to each sketch from the image-cue experiment, in order of confidence, from the set of 64 object labels. We found that human and model recognition performance (d′) was highly consistent across objects (Spearman's ρ=0.649, Fig. 8).

Consequences of Drawing Practice Since the assignment of objects to condition was randomized across participants, no differences in performance (target rank) on Trained, Near, and Far objects were predicted during the pre-test. To test this, we computed the mean target rank in each condition for each participant (Trained=9.92, Near=9.24, Far=9.57), which we then analyzed using a 3-condition (Trained, Near, Far) x 2 cue-type (image, verbal) repeated-measures ANOVA. There was no main effect of condition on pre-test performance ($F_{2,1178}$=1.70, p=0.184). Cue type did have an effect ($F_{1,589}$=19.7, p<0.001), but did not interact with condition ($F_{2,1178}$=0.258, p=0.773).

Our main hypotheses concerned changes in performance due to focused drawing practice on the Trained objects. Specifically, we predicted that repeatedly sketching a subset of the objects in one category would affect how other similar objects belonging to the same category were drawn, but that such practice would not affect how unrelated objects were drawn.

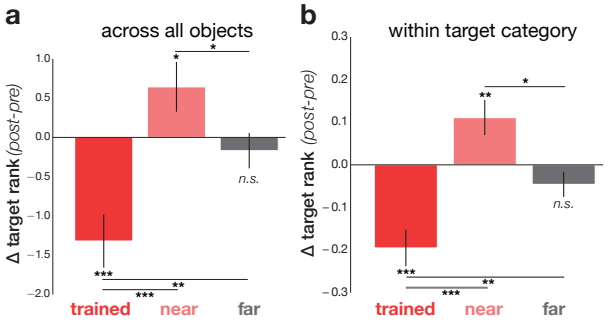

Figure 9: Changes in performance between post-test and pre-test for each condition: (*left*) when computing target rank among all 63 distractors and (*right*) when computing target rank relative to remaining 7 objects within the target category only. Error bars represent 1 s.e.m. *p<0.05, **p<0.01,***p<0.001

To evaluate this prediction, we calculated the change in target rank for each item ($\Delta_{rank} = rank_{post} - rank_{pre}$), then averaged these Δ_{rank} values for each condition within-participant.

We then performed the same type of ANOVA as was used for the pre-test analysis, which revealed a highly significant difference among conditions ($F_{2,1178}$=12.7, p<0.001). There was no main effect of cue-type ($F_{1,589}$=0.213, p=0.644), and no interaction with condition ($F_{2,1178}$=0.365, p=0.695), so in subsequent analyses we collapsed across cue-type. A follow-up t-test revealed that sketches of Trained objects were better recognized by the model following training ($\Delta_{rank} <$ 0:t_{590}=3.89, p=0.0001), and this improvement was also statistically reliable when compared with performance on Far objects ($\Delta_{rank,trained} < \Delta_{rank,far}$: t_{590}=3.05, p=0.002). By contrast, model performance for sketches of Near objects worsened after training relative to baseline ($\Delta_{rank} < 0$: t_{590}=2.03, p=0.04) and relative to control Far objects ($\Delta_{rank,near} < \Delta_{rank,far}$: t_{590}=2.15, p=0.03). Recognition of Far objects did not change significantly relative to baseline ($\Delta_{rank} < 0$: t_{590}=0.751, p=0.453). This was true when computing the target rank among all distractors (Fig. 9a), as well as when classification was restricted to objects within the target category (Fig. 9b).

These results show that visual production reshaped participants' representational space for objects: by differentiating trained objects and merging other objects nearby in the space (Fig. 10). More broadly, these findings suggest that the outward expression of visual concepts can itself bring about changes to their internal representation.

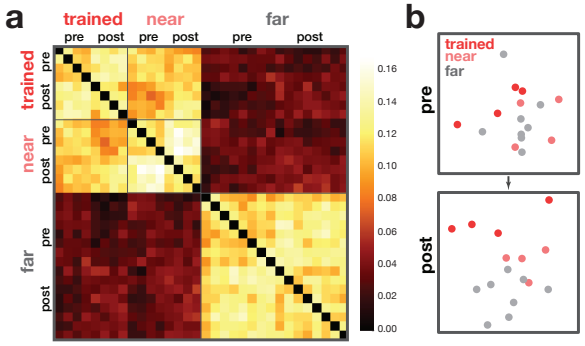

Figure 10: (a) Representational similarity between conditions across phases (averaged over object identity): low values in the large off-diagonal blocks correspond to larger average correlation distances between objects in different categories. More subtle changes underlying the effects shown in Fig. 9 are reflected in the smaller diagonal and off-diagonal blocks. (b) Visualization of changes to the representation using multidimensional scaling on the correlation distances between objects in each condition.

Discussion

Humans draw for many reasons: to depict, to record, to plan, to explain, to create (Tversky, 2011). Drawn images predate the historical record (Clottes, 2008), are pervasive in human culture (Gombrich, 1989), and are often produced prolifically in childhood (Kellogg, 1969). Moreover, drawing is a powerful tool for communication — with just a few strokes it is possible to convey the identity of a face (Bergmann, Dale, & Lupyan, 2013) or express an intention (Galantucci, 2005). Just as investigations of both verbal comprehension and production are indispensable to theories about linguistic communication, a more complete understanding of visual communication will entail examining how visual recognition and production interact to achieve our goals.

Although we interpret our results as supporting the idea that training had reshaped participants' internal representation of objects, another possibility is that they had only adapted their responses based on classifier feedback. Examining the generality of these effects across different tasks in subsequent studies will be helpful for teasing apart these two accounts. Specifically, future experiments will examine how learning to draw objects affects how these objects are later perceived, to further evaluate the idea that visual production alters a generalized object representation that supports both recognition and production. In addition, we plan to investigate how visual learning achieved via active production differs from that achieved through passive observation (Gureckis & Markant, 2012), involving close examination of how sensory feedback (e.g., visual, tactile) and social interaction (Fay, Garrod, Roberts, & Swoboda, 2010) influence learning. Ultimately, inquiries into the psychological basis of visual production may shed new light upon the origins of symbolic writing systems for communication, and the very nature of our ability to apprehend abstract meanings from visual artifacts.

Acknowledgments

This work was supported by NSF GRFP DGE-0646086 (J.E.F), NIH R01 EY021755 (N.B.T.-B.), and the David A. Gardner '69 Magic Project at Princeton University.

References

Bergmann, T., Dale, R., & Lupyan, G. (2013). The impact of communicative constraints on the emergence of a graphical communication system. In *Proceedings of the 35th Annual Conference of the Cognitive Science Society* (pp. 1887–1992).

Biederman, I., & Ju, G. (1988). Surface versus edge-based determinants of visual recognition. *Cognitive Psychology*, *20*(1), 38–64.

Clottes, J. (2008). *Cave art*. Phaidon London.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009.* (pp. 248–255).

Diamond, R., & Carey, S. (1986). Why faces are and are not special: an effect of expertise. *Journal of Experimental Psychology: General*, *115*(2), 107.

Eitz, M., Hays, J., & Alexa, M. (2012). How Do Humans Sketch Objects? *ACM Transactions on Graphics (TOG)*, *31*(4), 44.

Fan, J. E., Yamins, D., DiCarlo, J., & Turk-Browne, N. B. (2014). Mapping core similarity among visual objects across image modalities. In *ACM Siggraph 2014 Posters* (p. 67).

Fay, N., Garrod, S., Roberts, L., & Swoboda, N. (2010). The interactive evolution of human communication systems. *Cognitive Science*, *34*(3), 351–386.

Frey, B. J., & Dueck, D. (2007). Clustering by passing messages between data points. *Science*, *315*(5814), 972–976.

Galantucci, B. (2005). An experimental study of the emergence of human communication systems. *Cognitive Science*, *29*(5), 737–767.

Goldstone, R. L. (1998). Perceptual learning. *Annual Review of Psychology*, *49*(1), 585–612.

Gombrich, E. (1989). *The story of art*. Phaidon Press, Ltd.

Gureckis, T. M., & Markant, D. B. (2012, September). Self-directed learning: A cognitive and computational perspective. *Perspectives on Psychological Science*, *7*(5), 464–481.

Hung, C. P., Kreiman, G., Poggio, T., & DiCarlo, J. J. (2005). Fast readout of object identity from macaque inferior temporal cortex. *Science*, *310*(5749), 863–866.

Kellogg, R. (1969). *Analyzing children's art*. National Press Books Palo Alto, CA.

Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., ... Bandettini, P. A. (2008). Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron*, *60*(6), 1126–1141.

Malach, R., Levy, I., & Hasson, U. (2002). The topography of high-order human object areas. *Trends in Cognitive Sciences*, *6*(4), 176–184.

Mervis, C. B., & Rosch, E. (1981). Categorization of natural objects. *Annual review of Psychology*, *32*(1), 89–115.

Pinto, N., Cox, D. D., & DiCarlo, J. J. (2008). Why is real-world visual object recognition hard? *PLoS Computational Biology*, *4*(1), e27.

Tanaka, J. W., & Taylor, M. (1991). Object categories and expertise: Is the basic level in the eye of the beholder? *Cognitive Psychology*, *23*(3), 457–482.

Tversky, B. (2011). Visualizing thought. *Topics in Cognitive Science*, *3*(3), 499–535.

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 201403112.