On the Role of Shape Prototypes in Hierarchical Models of Vision

Michael D. Thomure, Melanie Mitchell, and Garrett T. Kenyon
To appear in *Proceedings of the International Joint Conference on Neural Networks* (IJCNN), 2013.

***Abstract*— We investigate the role of learned shape-prototypes in an influential family of hierarchical neural-network models of vision. Central to these networks' design is a dictionary of learned shapes, which are meant to respond to discriminative visual patterns in the input. While higher-level features based on such learned prototypes have been cited as key for viewpoint-invariant object-recognition in these models [1], [2], we show that high performance on invariant object-recognition tasks can be obtained by using a simple set of unlearned, "shape-free" features. This behavior is robust to the size of the network. These results call into question the roles of learning and shape-specificity in the success of such models on difficult vision tasks, and suggest that randomly constructed prototypes may provide a useful "universal" dictionary.**

I. Introduction

In this paper we examine the role of shape prototypes in one well-known family of hierarchical object-recognition architectures—those with multiple layers that alternate between prototype matching and activation pooling. In the literature, this architecture has been argued to allow a trade-off between selectivity to specific patterns (encoded by learned prototypes) and invariance to irrelevant variations in object pose (via activation pooling) [1]. Here we term these as *alternating multilayer* architectures. Recent models in this family have been reported to achieve state-of-the-art performance on a variety of object-recognition tasks [3], [4], [5], and have been shown to out-perform many alternative approaches on image classification tasks that specifically emphasize viewpoint invariance [2]. In this work, we conducted a series of detailed experiments to investigate how a set of learned shape prototypes in alternating multilayer models mediate improved classification performance. Surprisingly, we found that the classification performance of networks using randomly generated prototypes—with no apparent spatial structure—perform nearly identically to networks using prototypes learned from natural images in a way so as to capture "useful" shape components.

In the next section we describe the general architecture of the hierarchical networks we are studying. In Section III we outline the experiments we performed to investigate the role of multiple layers and, in particular, of shape prototypes. In Section IV, we give the results of our experiments on several well-known image datasets. In Section V we discuss these results. Finally, in Section VI we give our conclusions and sketch plans for future work.

Michael Thomure and Melanie Mitchell are with the Computer Science Department, Portland State University, (email: {thomure, mm}@cs.pdx.edu); Melanie Mitchell is with the Santa Fe Institute; Garrett Kenyon is with Los Alamos National Laboratory (email: gkenyon@lanl.gov).

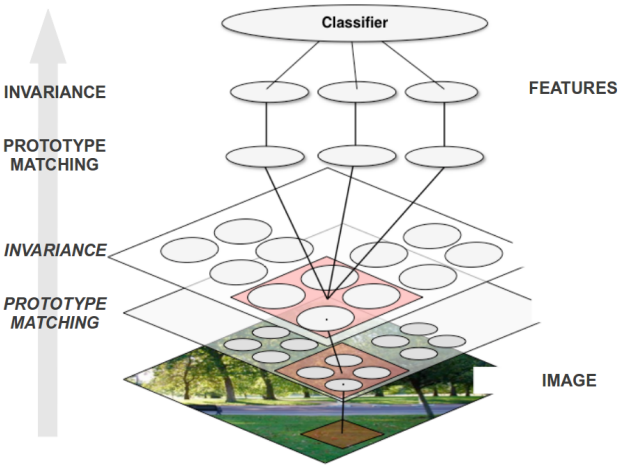


Fig. 1: Sketch of an alternating multilayer architecture, similar to the model described in [1]. Shaded regions correspond to the input of a single unit at the layer above. See text for explanation.

II. Hierarchical Models of Vision

This work considers an important family of hierarchical neural networks, which are used to classify an image based on its contents. These networks combine multiple layers of prototype matching, as shown in Figure 1, in which a visual pattern is matched to a local image neighborhood. These layers are interleaved with pooling operations that provide invariance to certain deformations of the prototype, such as scaling or translation. In these networks, prototype matching is achieved by a (dis)similarity measure, such as dot product or radial basis function, while invariance is achieved via summarization of input neighborhoods by their average or maximum value. At the top of the network, the activity of the nodes is fed to a classification algorithm, with logistic regression or support vector machine (SVM) being a common choice.

Examples of this approach include the early Neocognitron [6], the HMAX models [7], [1], Convolutional Neural Networks [5], and Sparse Localized Features [3]. Many of these approaches achieved state-of-the-art performance, and have garnered significant interest within the computer vision community.

Using a linear-kernel SVM as classifier, for example, the architecture described above has been reported to achieve state-of-the-art performance on a variety of object-recognition tasks [2], [4]. According to some [8], a key to the performance of this model is the inclusion of a second

(a) *Car v. Plane* (b) *Face1 v. Face2*

Fig. 2: Example images from the dataset of [2].

layer of learned prototypes that match specific discriminative visual patterns. However, to our knowledge, the role of the second-layer shape prototypes in the performance of such networks has never been carefully tested. The purpose of this paper is to investigate this claim.

III. Methods

In this work, we test the "shape dictionary" hypothesis discussed above, which suggests that using imprinted (or otherwise learned) prototypes succeeds by constructing representations that capture "shape-based" features of objects. To investigate this, we make use of unlearned, "shape-free" prototypes. These representations are constructed randomly, where each prototype component is drawn independently from a uniform distribution over activation values in the range $[0, 1]$. (This approach should not be confused with imprinting, in which randomness is used to choose the location of training patches.) As such, these prototypes lack the spatial structure expected of a "shape" prototype. Recent evidence suggests that various kinds of random features can be surprisingly useful in hierarchical networks [9], [10], [11], though the reasons for this behavior are still unclear.

A. Glimpse

We have developed a novel system for experimentation on hierarchical visual models, which hides low-level implementation details without sacrificing run-time efficiency. The system provides a simple interface for running experiments, is designed using only free and open-source components, and provides native support for parallel compute resources [12]. This system can be used to build networks with a range of different parameters, layer operations, and connectivity with a minimum of coding effort. We used this system to create Glimpse, our implementation of an alternating multilayer visual network.

B. Datasets

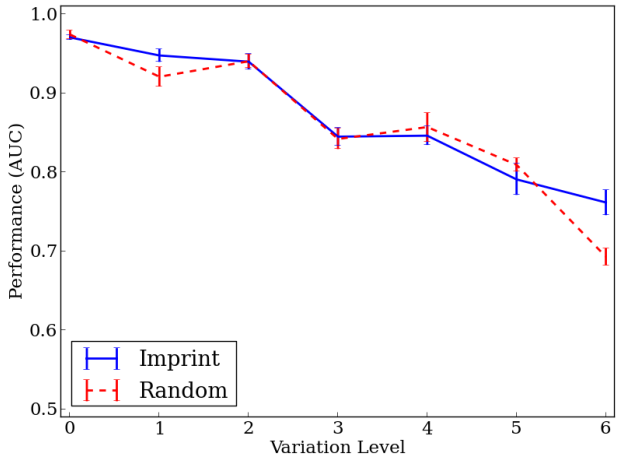
Recently, a number of authors have raised concerns that many common object recognition datasets contain significant confounds [13], [14]. To address these concerns, we consider two artificial tasks introduced previously by Pinto et al. [2]. These tasks were designed to probe a system's ability to demonstrate view-point invariant object recognition, without using visual cues from the surrounding environment. The dataset is constructed by rendering a 3D object model from various points of view, and then composing the object with a randomly-chosen natural image background. The difficulty of each task depends on the range of view-points from which an object is rendered. Following Pinto et al., the task is quantized into seven "variation levels", with difficulty increasing with each level. (See [2] for details.) The first dataset contains rendered examples of cars and airplanes (*Car v. Plane*), and measures category-level discrimination. The second dataset contains rendered examples of two different faces (*Face1 v. Face2*), and measures subordinate-level discrimination. Figure 2 gives a sample of the images for these two tasks.

IV. Results

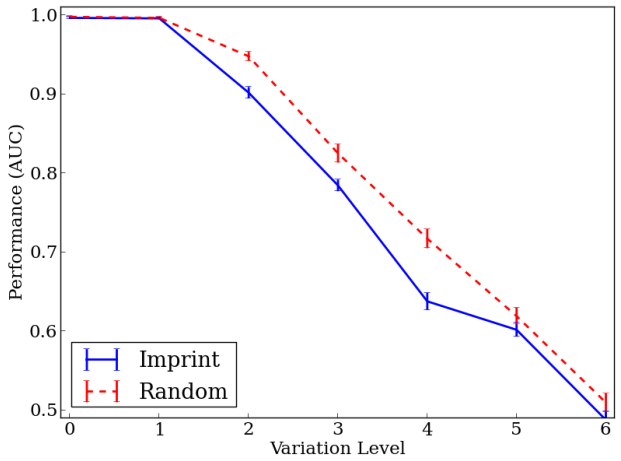
Figure 3 compares performance given as the mean Area Under the ROC Curve (AUC) over five independent training/testing splits using two different image representations: features based on 4,075 *imprinted* prototypes, and features based on 4,075 *random* prototypes. Figure 3a shows this comparison for the *Car v. Plane* task, and Figure 3b shows the same comparison for the *Face1 v. Face2* task. The data provided by Pinto et al. [2], [15] is split into seven different *variation levels*, i.e., levels of variation in rotation, position, and scale of the objects of interest. Each level of variation defines a separate object-recognition task. Following [2], we plot performance (mean AUC, with error bars giving standard error) as the variation level is increased. We found that results were similar to [2], but that this did not depend on the technique for choosing prototypes; behavior for random and imprinted prototypes was nearly identical. This result seems to contradict the "shape dictionary" hypothesis. Here we consider a number of possible explanations.

We first consider the possibility that a sufficiently large network is simply robust to a bad choice of prototypes. That is, perhaps any sufficiently large set of prototypes would lead to the behavior seen in Figure 3. To investigate this, we compare the performance of these representations using different numbers of prototypes. Figure 4 shows that the performance of Glimpse using imprinted and random prototypes is quite similar even when using as few as 10 prototypes. Regardless of the size of the network, we were unable to find a significant difference in performance between the two representations.

Alternatively, we considered the possibility that random prototypes provide a kind of "weak" feature that, when used alone, is non-discriminative, but in combination with others provides a "strong ensemble". In contrast, we expect imprinting to generate at least some prototypes that provide

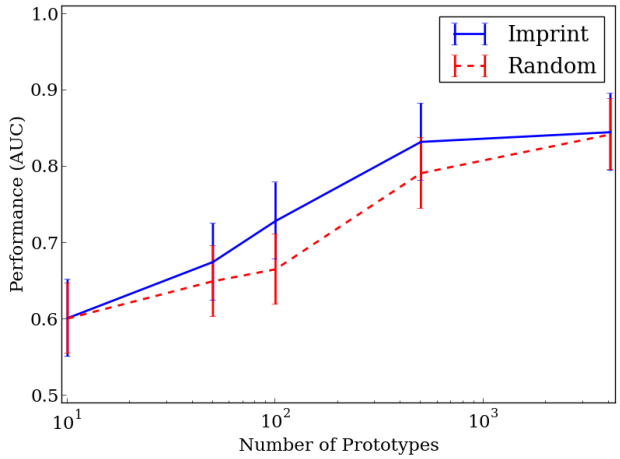


(a) *Car v. Plane*

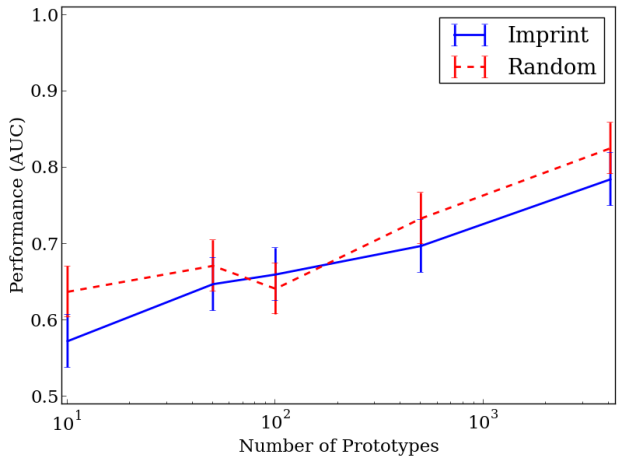


(b) *Face1 v. Face2*

Fig. 3: Comparison of Glimpse's performance on two tasks, using (1) 4,075 imprinted prototypes; and (2) 4,075 random prototypes. The horizontal axis shows the variation level (over rotation, position, and scale) of the object of interest, and the vertical axis shows the mean AUC over five independent training/testing splits at each variation level. Error bars show the standard error.



(a) Car v. Plane



(b) Face1 v. Face2

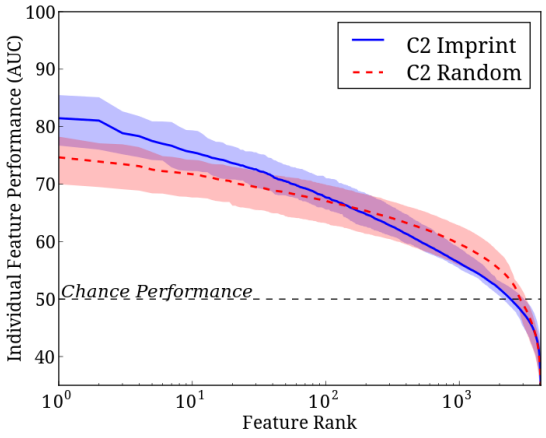
Fig. 4: Comparison of Glimpse's performance on (a) the *Car v. Plane* task and (b) the *Face1 v. Face2* task, for variation level 3 in each case. The curves give the performance (mean AUC over five independent training/testing splits) of the imprinted (solid blue) and random (dashed red) prototypes. Error bars give standard error.

highly-discriminative representations, even when considered in isolation. To investigate this, we measured performance based on individual features. For each prototype-generation method (imprinting or random), we generated 4,075 prototypes as before, except here we used them one at a time to create a single value to represent each image in order to train and test the SVM. As before, we performed five independent training/testing splits using each prototype. Figure 5a shows the performance (mean AUC) for single imprinted prototypes (solid blue curve) and single random prototypes (solid red curve) on the *Car v. Plane* task, where the prototypes are ranked by performance. The shaded areas give the range of performance for each case. Figure 5b shows the same values for the *Face1 v. Face2* task. We found no significant difference between the two representations in terms of the occurrence of individually discriminative features. In fact, it is striking how well the best random features perform when operating in isolation. In short, it appears that random prototypes are not limited to operating in ensembles.

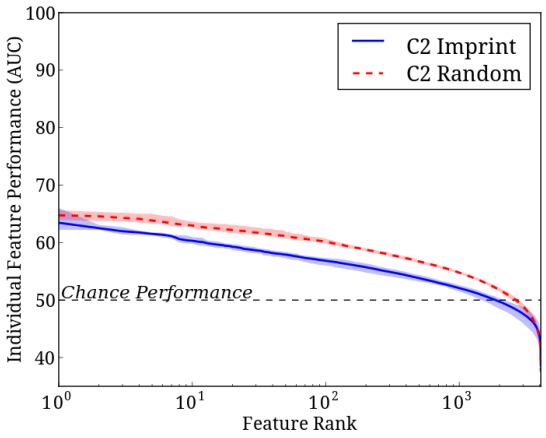
A. *Best matches between prototypes and image crops*

Finally, we investigated the hypothesis that the imprinted and random prototype representations behave similarly because they code for similar visual features. It is possible, in theory, that our process of random prototype generation occasionally creates the kind of useful shape selectivity that we expect under imprinting. In this case, we would expect these "lucky" random features to be among the most discriminative when used in isolation.

Due to the nature of these networks, it is difficult to directly interpret the contents of a prototype. (For example,



(a) Car v. Plane



(b) Face1 v. Face2

Fig. 5: Performance (mean AUC and range) using individual features from either imprinted (solid blue) or random (dashed red) prototypes for (a) the *Car v. Plane* task, and (b) the *Face1 v. Face2* task. In both cases, the tasks use variation level 3.

(a) Imprinted prototypes

(b) Random prototypes

Fig. 6: Characterization of best-performing prototypes for the *Face1 v. Face2* task (cf. Figure 5b) based on the input patches to which they respond most strongly. (a): Each row corresponds to one of the top five imprinted prototypes (those ranked 1–5 in the imprinted set in Figure 5b). The 10 images in each row are the 10 image patches in the *Face1 v. Face2* dataset to which the prototype matched most closely. All patches in a row are drawn from different images. (b): Same as part (a), but here the five top prototypes are those ranked 1–5 in the random-prototype set in Figure 5b. In contrast to part (a), there is no obvious preferred "shape" along each row.

the invariance stage may cause the same feature values to be produced for multiple images.) Instead, we attempt to characterize a given prototype by examining those input patches that provide the best match. Figure 6 shows this data for the most discriminative prototypes on the *Face1 v. Face2* task (variation level 3). Each row in the figure corresponds to one of the five most discriminative prototypes (those ranked 1–5 in Figure 5b for (a) imprinted prototypes and (b) random prototypes. Each row gives the image 10 patches in the *Face1 v. Face2* dataset to which the corresponding prototype matched most closely, where each image is allowed at most one match. Although it may appear that patches in, say, the top row of Figure 6a are from slightly different positions of the same image, these patches are all from different images. As expected, it appears that the five imprinted prototypes are responding preferentially to specific "shape-based" patterns relevant to faces, and are relatively robust to rotation and translation of those patterns. However, the five random prototypes display no obvious "shape" preference or relevance to faces along each row.

These results show that, while imprinted features are highly selective to shape and somewhat invariant to background clutter, random prototypes are not easily interpretable as shape templates. Although the patches in Figure 6 came from one particular set of imprinted and random prototypes, we found that this behavior was qualitatively similar for other, independently generated, sets of imprinted and random prototypes.

V. Discussion

In this work, we investigated the hypothesis that shape-based prototypes are central to the ability of alternating multilayer networks to perform invariant object-recognition. To summarize our results:

- We applied our network to a pair of challenging benchmarks for invariant object recognition, and find that learned "shape" prototypes are not necessary to achieve

the performance seen in the literature. These benchmarks specifically emphasize viewpoint-invariance by carefully controlling for confounding factors. As such, our "shape-free" features seem to provide an unlearned, unbiased (i.e., universal) dictionary.

- Upon analysis, we find evidence that (1) our randomly-generated prototypes support performance that is on par with a learned shape dictionary (Figure 3), even in small networks (Figures 4 and 5). Critically, we also find evidence that (2) those prototypes lack shape specificity (Figure 6), a characteristic thought to be central to the success of these networks.

Taken together, these results argue that our understanding of successful hierarchical visual models is far from complete, and that further analysis is warranted. Furthermore, our work suggests that—when used properly—random projections may have an important role to play in these hierarchical networks.

We are left with several questions, yet to be answered. Chief among them are: (1) In what types of object-recognition tasks would a set of learned shape-based prototypes provide an advantage over randomly generated prototypes? Equivalently, for what sorts of tasks can we simply rely on random prototypes and thus avoid the costs of learning? (2) What are the mechanisms underlying the success of random prototypes in our experiments? For example, can this success be explained by mechanisms related to the methods of random projections or compressive sensing [16], [17]? These are questions our group hopes to address in future work.

VI. Conclusions

The family of alternating multilayer network models has been described as a state-of-the-art solution to the task of image classification, and has been shown to be superior to a variety of alternative approaches on tasks that specifically emphasize viewpoint invariant object detection. Central to the success of this model is claimed to be the unsupervised learning of discriminative "shape dictionaries". We performed a systematic evaluation of this claim, and showed that qualitatively identical behavior can be produced using only a simple, randomly constructed dictionary that displays little shape selectivity. This directly challenges the existing hypothesis, and suggests that a new explanation is required to understand the qualitative behavior of this important class of models. Our future work will include the search for this explanation.

Acknowledgment

This material is based upon work supported by the National Science Foundation under Grant Nos. 1018967 and 0749348. Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agency. We are grateful to Will Landecker and Max Quinn for many helpful discussions about this work, and to Nicolas Pinto for providing data.

VII. Appendix

In the following, we first give the detailed parameters and operations used in our visual network. We then verify that our model captures the qualitative behavior of other published implementations.

The input to Glimpse is a grayscale image, which is rescaled to have a shorter side of 220 pixels. A scale pyramid with nine bands is then generated from the image, with a constant ratio of $2^{1/4}$ between bands. Next, the first prototype-matching layer applies a battery of Gabor filters of size 11x11 pixels to each band of the scale pyramid. Given an input patch $\mathbf{x}$, the activation of each unit is calculated as

$$act_1(\mathbf{x}, \mathbf{g}) = \frac{|(\mathbf{x}, \mathbf{g})|}{\|\mathbf{x}\| \, \|\mathbf{g}\|} .$$

where $(\cdot)$ denotes the inner product, $|\cdot|$ the absolute value, and $\|\cdot\|$ the Euclidean norm. The filter $\mathbf{g}$ is given by the Gabor function

$$\begin{aligned} \mathbf{g} &= \exp\left(-\frac{\left(x_0^2 + \gamma^2 y_0^2\right)}{2\sigma^2}\right) \times \sin\left(\frac{2\pi x_0}{\lambda} + \phi\right) x_0 \\ &= x\cos\theta + y\sin\theta y_0 \\ &= -x\sin\theta + y\cos\theta \,, \end{aligned}$$

where x and y range over $[-w/2, w/2]$ for a unit with receptive field width of w. In our experiments, we used orientations of $\theta = (0°, 45°, 90°, 135°)$, an aspect ratio of $\gamma = 0.6$, a wavelength of $\lambda = 2.75$, a scale of $\sigma = \lambda/2$, a phase of $\phi = 0$, and a receptive field width of $w = 11$ pixels.

Each unit in the first invariance layer then applies maximum-value pooling over a local 10x10 neighborhood of outputs from the layer below, and the result is sub-sampled by a factor of two. The next prototype-matching layer then applies a set of stored prototypes, which receive input from a 7x7 spatial neighborhood of inputs. Given a patch $\mathbf{x}$ of inputs, the response of a prototype-matching unit at the second layer is given by

$$act_2(\mathbf{x}, \mathbf{p}) = \exp\left(-\beta \left\|\mathbf{x} - \mathbf{p}\right\|^2\right) \tag{1}$$

where $\mathbf{p}$ is the unit's prototype. In our experiments, we use a value of $\beta = 5.0$. Finally, the last invariance layer applies maximum-value pooling for each prototype, giving the best response over all scales and locations. The results are used as input to a linear kernel SVM.

In our work, we use random prototypes that are sparse and gain-invariant. Random prototypes use sparse input, in which high activation for one orientation suppresses activation at other orientations. We implement this by scaling the activation of each input unit x_i at location ℓ as $x_i' = \frac{x_i}{a_\ell}$. The total activation a_ℓ at location ℓ is measured as $a_\ell = \sqrt{\sum x_j^2}$, where the sum ranges over the set of inputs at location ℓ. Furthermore, gain invariance is achieved by constraining the input and prototype vectors to have fixed norm, where we compute activation as $act_2\left(\frac{\mathbf{x}'}{\|\mathbf{x}'\|}, \frac{\mathbf{p}'}{\|\mathbf{p}'\|}\right)$. (See Equation 1.)

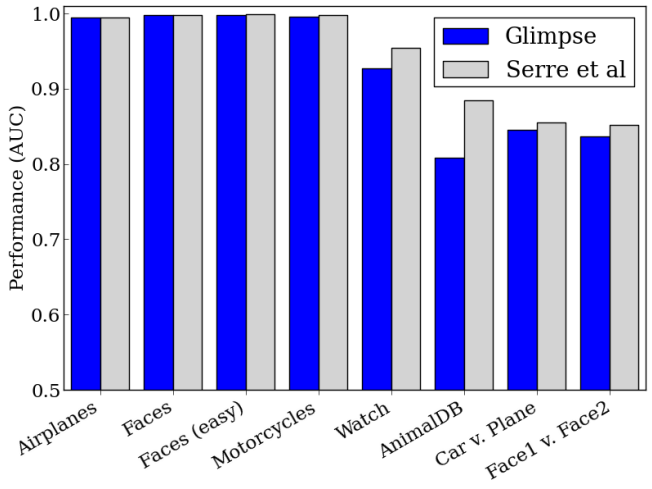


Fig. 7: Comparison of Glimpse with reference system of [1], given as mean AUC with standard errors shown.

Although very similar to the models described in [1], [3], Glimpse differs by using a larger input image (220 instead of 140 pixels), fixed-size first-layer prototypes (a scale pyramid of the image is used instead), fewer scales (9 instead of 16), and a single size for second-layer prototypes (instead of four sizes). These parameter choices allowed for increased computation speed without sacrificing performance, and are similar to those used in the SLF model [3].

For completeness, we compare our network against the reference systemof Serre et al. [18] for a range of benchmark datasets. (We obtained very similar results using the SLF model of Mutch and Lowe [3].) Figure 7 shows the results of this comparison, using 4,075 imprinted prototypes for each system. Performance is reported as mean AUC over five independent trials, where error bars give standard error. The datasets used here include subsets of the Caltech101 dataset [19], the *AnimalDB* benchmark of Serre et al. [4], and the synthetic tasks discussed in Section III. The Caltech101 subsets (*Airplanes, Faces, Faces (easy), Motorcycles, and Watch*) used all available examples of the given foreground class, and an equal number of examples from the background class. For all experiments, half of the available data was used for training, and the other half for testing.

In all cases, the variation among trials is low and performance is quite similar between the two systems. We note somewhat inferior performance for the Glimpse network on the *AnimalDB* dataset. We suspect that this is the result of a difference in parameters rather than reflecting a qualitative difference in network behavior.

References

[1] T. Serre, L. Wolf, and T. Poggio, "Object Recognition with Features Inspired by Visual Cortex," *CVPR*, 2005.

[2] N. Pinto, Y. Barhomi, D. D. Cox, and J. J. DiCarlo, "Comparing State-of-the-Art Visual Features on Invariant Object Recognition Tasks," in *Proceedings of the IEEE Workshop on Applications of Computer Vision (WACV 2011)*, 2011.

[3] J. Mutch and D. G. Lowe, "Object Class Recognition and Localization Using Sparse Features with Limited Receptive Fields," *International Journal of Computer Vision*, vol. 80, pp. 45–57, Oct. 2008.

[4] T. Serre, A. Oliva, and T. Poggio, "A feedforward architecture accounts for rapid categorization," *Proceedings of the National Academy of Sciences*, vol. 104, pp. 6424–6429, Apr. 2007.

[5] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Learning Hierarchical Features for Scene Labeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013 (in press).

[6] K. Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Biological Cybernetics*, vol. 36, pp. 193–202, Apr. 1980.

[7] M. Riesenhuber and T. Poggio, "Hierarchical models of object recognition in cortex," *Nature Neuroscience*, vol. 2, pp. 1019–1025, Nov. 1999.

[8] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio, "Robust object recognition with cortex-like mechanisms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, pp. 411–426, Mar. 2007.

[9] K. Jarrett, K. Kavukcuoglu, M. A. Ranzato, and Y. LeCun, "What is the best multi-stage architecture for object recognition?," in *Proceedings of the 2009 IEEE 12th International Conference on Computer Vision*, pp. 2146–2153, IEEE, Sept. 2009.

[10] N. Pinto, Z. Stone, T. Zickler, and D. Cox, "Scaling up biologically-inspired computer vision: A case study in unconstrained face recognition on facebook," in *CVPR 2011 Workshop on Biologically-Consistent Vision*, pp. 35–42, IEEE, June 2011.

[11] A. Saxe, P. W. Koh, Z. Chen, M. Bhand, B. Suresh, and A. Y. Ng, "On random weights and unsupervised feature learning," in *NIPS 2010 Workshop on Deep Learning and Unsupervised Feature Learning*, 2010.

[12] Source code for the Glimpse model. URL: `http://web.cecs.pdx.edu/~thomure/glimpse/`.

[13] N. Pinto, D. Cox, and J. J. Dicarlo, "Why is Real-World Visual Object Recognition Hard?," *PLoS Computational Biology*, vol. 4, Jan. 2008.

[14] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *CVPR 2011*, pp. 1521–1528, IEEE, June 2011.

[15] N. Pinto, "Personal communication."

[16] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face representation via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, 2009.

[17] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T. Huang, and S. Yan, "Sparse representation for computer vision and pattern recognition," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 1031–1044, 2010.

[18] "Source code for the HMAX model.." `http://cbcl.mit.edu/software-datasets/pnas07/`.

[19] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories," in *CVPR 2004, Workshop on Generative-Model Based Vision*, 2004.