A Biologically Plausible Approach to Cat and Dog Discrimination

Bruce A. Draper, Kyungim Baek, Jeff Boody

Department of Computer Science
Colorado State University
Fort Collins, CO 80523-1873 U.S.A.
draper,beak,boody@cs.colostate.edu

Abstract. The paper describes a computational model of human expert object recognition in terms of pattern recognition algorithms. In particular, we model the process by which people quickly recognize familiar objects seen from familiar viewpoints at both the instance and category level. We propose a sequence of unsupervised pattern recognition algorithms that is consistent with all known biological data. It combines the standard Gabor-filter model of early vision with a novel cluster-based local linear projection model of expert object recognition in the ventral visual stream. This model is shown to be better than standard algorithms at distinguishing between cats and dogs.

The Human Visual System

The basic anatomical stages of the human visual process are well known. Images form on the retina, and pass via the optic nerve to the lateral geniculate nucleus (LGN) and superior colliculus (SC), and on to the primary visual cortex (area V1). From here, the human visual system divides into two streams: the dorsal visual pathway and the ventral visual pathway. As described by Milner and Goodale [1], the dorsal stream is responsible for vision in support of immediate physical action. It models the world in egocentric coordinates and has virtually no memory. The ventral stream is responsible for vision in support of cognition. It is responsible for both object recognition and 3D allocentric (i.e. object-centered) object modeling, and maintains a complex visual memory.

Although the early visual process (up to and including V1) is fairly uniform, the dorsal/ventral dichotomy is just one of many divisions that can be drawn in later stages of the human visual system. The dorsal stream, for example, can be further divided into anatomically distinct components for specific egocentric coordinates, e.g. eye-centered, head-centered, and shoulder-centered subsystems ([1], p. 53-55). By analogy, the ventral stream should also be composed of multiple, anatomically distinct systems. This hypothesis is verified by brain imaging studies, which show activity in different anatomical locations depending on whether the subject is viewing, for example, faces or places [2, 3].

It is important, therefore, for claims of biologically plausible object recognition to be specific: *which* object recognition subsystem is being modeled? In this paper, we focus on the recognition of familiar objects from familiar viewpoints at both the

category and instance level, a process sometimes called “expert object recognition” [4]. We suggest that expert object recognition is a very fast process in humans, and uses an anatomical pathway from the primary visual cortex (V1) to the fusiform gyrus and the right inferior frontal gyrus. We also suggest that this pathway can be modeled as a sequence of statistical pattern recognition algorithms.

The Expert Object Recognition Pathway

The pathway associated with expert object recognition was first identified in fMRI studies of the more limited task of human face recognition [5-7]. In these studies, patients were shown images of human faces while in a scanner. The resulting fMRI images revealed activation not only in the primary visual cortex, but also in the fusiform gyrus. Subsequent PET studies (which imaged a larger portion of the brain) confirmed the activation in the fusiform gyrus, while adding another locus of activity in the right inferior frontal gyrus, an area previously associated through lesion studies with visual memory [8] (see also [2]). Moreover, in both the fMRI and PET studies, the activation was unique to the task of face recognition. Images of places triggered another distinct pathway with activation in the parahippocampal gyrus [2, 3, 8]. This led to speculation that evolution had created a special visual pathway for recognizing faces, and the locus of activation within the fusiform gyrus was dubbed the Fusiform Face Area (FFA; [3]).

More recent evidence suggests, however, that this pathway is used for more than just recognizing human faces. Tong, et al. report that the FFA is activated by animal faces and cartoon faces as well as human faces [9]. Chao et al. report that the FFA is activated by images of full-bodied animals, and animals with obscured faces [10]. Ishai et al. find an area in the fusiform gyrus that responds to chairs [11]. Tarr and Gauthier factored in the past experience of their subjects, and found FFA activation in dog show judges when they view dogs, and in bird experts when they view birds [12]. Most convincing of all, Tarr and Gauthier show that as people become expert at recognizing a class of objects, their recognition mechanism changes. Tarr & Gauthier created a class of cartoon characters called greebles, which in addition to individual identities can be grouped according to gender and family. When novice subjects view greebles, fMRIs show no activity in the FFA. Subjects are then trained to be greeble experts, where the definition of expert is that they can identify a greeble’s identity, gender or family with an equal response time. After training, the FFAs of experts become active when they view greebles [12]. It is therefore reasonable to conclude that the FFA is part of a general mechanism for recognizing familiar objects.

Properties of Expert Object Recognition

What constitutes expert object recognition? People become expert at recognizing objects when they encounter them often and when instances look alike, as with faces, animals and chairs. Just as important, people become experts at recognizing objects when they have to do so at multiple levels. For example, people recognize faces they

have never seen before as being human faces. At the same time, people almost instantly recognize the identity of familiar faces. Gauthier and Tarr use this multiple-level categorization as the defining characteristic of expert object recognition in their greeble studies [13], and it is a critical property of expert object recognition: objects are identified at both the class and instance level.

Expert object recognition is also very fast. While fMRI and PET studies do not give timing information, face recognition in humans can also be detected in ERP studies through a negative N170 signal. This signal occurs, on average, 164 milliseconds post stimulus onset [14]. This implies that the unique stages of expert object recognition – which we equate with the activation of the FFA and right inferior frontal gyrus – must begin within 164 milliseconds of the presentation of the target. Since visual processing begins in V1, this implies that the early stages of visual processing must also be quick. In particular, the early responses of simple and complex cells in V1 mimic Gabor filters and their combinations, and appear within 40 milliseconds of stimulus onset -- quickly enough to serve as input to the FFA. Later responses in the same cells reflect textural and boundary completion properties, but appear as late as 200 milliseconds post onset [15], probably too late to influence the expert recognition process.

Finally, expert object recognition is viewpoint dependent. The response of the FFA, for example, to images of faces presented upside-down is minimal [16]. The FFA responds to faces viewed head-on or in profile, but not to images of the back of the head [9]. In [17], upright and inverted greebles are presented to both novices and expert subjects. Expert subjects only show activation in the FFA for upright greebles (novice subjects have no FFA activation at either orientation).

Kosslyn's Model of Object Recognition

Brain imaging studies delineate the path within the ventral visual stream for recognizing familiar objects from familiar viewpoints at both the category and instance level. We believe this process accesses visual memory (as opposed to symbolic memory) because of the activation of the right inferior frontal gyrus. We also believe that the input to this process can be modeled in terms of Gabor filters because of the temporal constraints on V1 responses.

Brain imaging studies do not, however, indicate how objects are recognized, or even what the output of the object recognition process might be. For this we turn to a psychological model of object recognition originally proposed by Kosslyn and shown in Figures 1 and 2 [18]. It should be noted that Kosslyn originally proposed this model for visual perception in general, but that we apply it in the more limited context of expert object recognition.

Kosslyn's model makes a strong distinction between the strictly visual system (shown in white in Figure 1) and other mixed-modality systems, including high-level reasoning (shown in gray). Although high-level reasoning can influence vision, particularly in terms of visual attention, vision is distinct from other parts of the brain. The goal of vision is to "see again", in the sense that object recognition retrieves a visual memory that closely matches the current image [19]. Semantics in the form of object labels or other facts are assigned later by an associative memory, which

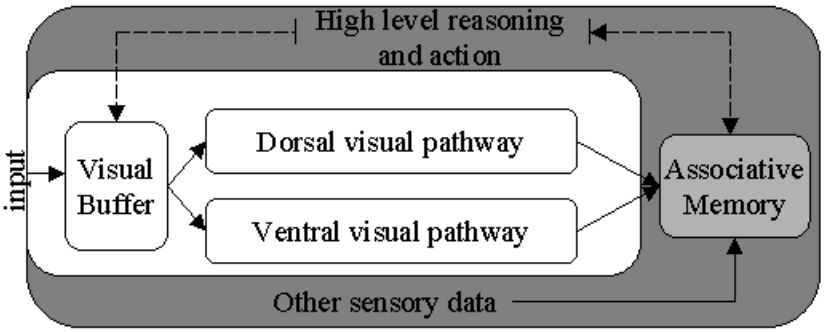

Fig. 1. Our interpretation of Kosslyn's model of the human visual system [18].

receives data from many systems, including other sensors and high-level reasoning systems.

Expert object recognition is performed within the ventral stream in a multi-step process, as shown in Figure 2. The visual buffer corresponds roughly to V1. The attention window is a mechanism that selects a small portion of the image for further processing. There is evidence that it can select scales as well as positions [19]. The data selected by the attention window is passed to a preprocessing subsystem, which according to Kosslyn computes non-accidental features and object-specific signal features. For the limited case of expert object recognition, we adopt a simpler model in which the preprocessing system simply computes edge magnitude responses from the complex Gabor filter responses.

After the attention window, the most significant subsystems are the category subsystem and the exemplar subsystem. As the name implies, the category subsystem assigns images to categories, although the categories are not defined in terms of symbolic object labels. Instead, the category system groups images in memory that look alike (as measured by their V1 responses). New images are then "categorized" in the sense of being assigned to one group or another. As a result, there is no one-to-one mapping between image clusters and object labels. If an object class has many variants (e.g. dogs or chairs) or is commonly seen from many viewpoints, its images may occur in many groups. Alternatively, if two distinct objects look similar to each other, they may fall within a single group, and the category subsystem will not

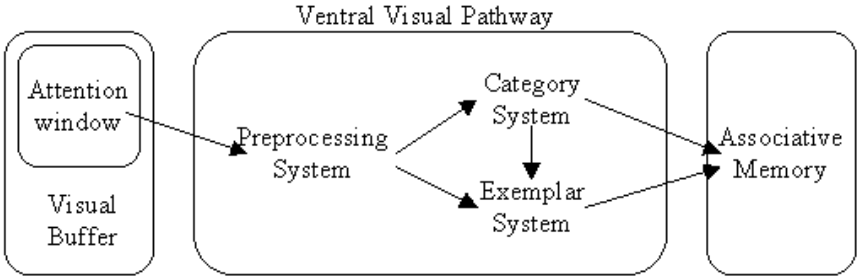

Fig. 2. The expert object recognition within the ventral visual stream.

distinguish between them.

The exemplar subsystem matches the current image to a single visual memory. Kosslyn describes visual memories as "compressed images", based in part on the anatomical structure of visual memory and in part on evidence that visual memories can be reconstructed in V1 as mental images [20]. We interpret this as implying that the exemplar subsystem performs subspace matching. The outputs of the category and exemplar subsystems are then passed to the associative memory, which can draw semantic conclusions based on both the best visual match in memory and a cluster of similar images, as well as non-visual inputs.

A Computational Model of Expert Object Recognition

We implement Kosslyn's model through the EOR (Expert Object Recognition) system shown in Figure 3. EOR is initially trained on a set of unlabeled training images as shown above the dotted line in Figure 3. The images are filtered through a pyramid of orientation-selective Gabor filters, using the Gabor parameters suggested in [21] for biological systems and the complex cell models suggested in [22], and then responses are combined into edges. We assume that the attention window can consistently select both the position and scale of the target object, and that it can compensate for small in-plane rotations. In effect, the attention window registers images of target objects. We do not know how the attention window works algorithmically, but we finesse the issue by giving the system small, registered images as input.

During training, the categorization system is modeled as an unsupervised clustering algorithm operating on edge data. We currently implement this using K-Means [23]. K-Means is simple and robust, and can be applied to high-dimensional data. Unfortunately, K-Means is also limited to modeling every data cluster as a symmetric Gaussian distribution. (We are experimenting with other clustering algorithms.)

The exemplar subsystem is implemented as a subspace projection and matching system. We have tested three different subspace projection algorithms: principal component analysis (PCA [24]); independent component analysis (ICA [25]), and factor analysis (FA [26]). So far, PCA has proven as effective as other techniques, although FA is useful as a pre-process for suppressing background pixels [27]. ICA can be applied so as to produce either (1) spatially independent basis vectors or (2) statistically independent compressed images ([28], 3.2-3.3). Although some have argued for the benefits of localized basis vectors in biological object recognition [29], we find they perform very poorly in practice [30]. Linear discriminant analysis (LDA [31]) has not been considered, since biological constraints dictate that it be possible to reconstruct an approximation of a source image from its compressed form. The experiments with EOR described below use PCA to model the exemplar subsystem.

At run-time (i.e. during testing) the process is very much simpler, as shown below the dotted line in Figure 3. Test images are Gabor filtered, and the edge responses are compared to the cluster centers learned during training using a nearest neighbor algorithm. Then images are compressed by projecting them into cluster-specific

subspaces, and nearest neighbors is applied again, this time to match the compressed images to compressed memories.

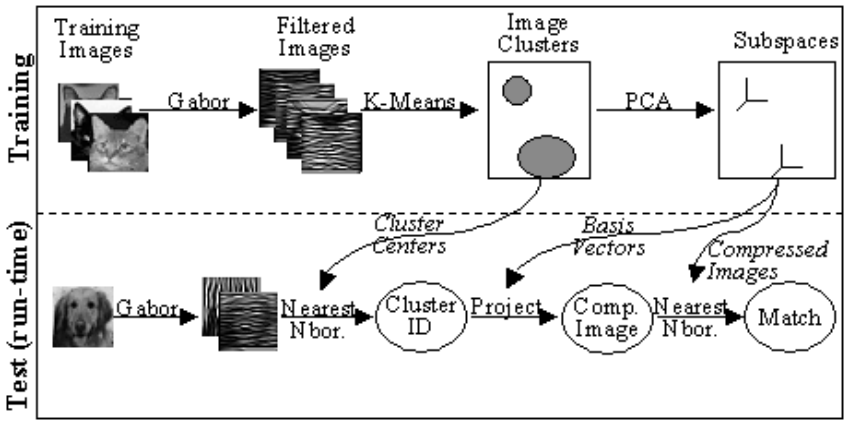

Fig. 3. EOR: Expert Object Recognition System.

There are several practical advantages to this biologically inspired design. First, the subspace matching algorithm is local, not global. We compute a unique subspace for every image cluster defined by the category system, and project only the images from that cluster into it. This creates a set of localized subspaces, rather than a single, global subspace as in most PCA-based systems. Localized subspaces have previously been used with face images [32, 33], but never in the context of multiple object classes. The argument for local subspaces is that while global PCA basis vectors are optimal for images drawn from a single, normal distribution, they are not optimal for images drawn from a mixture of Gaussian distributions. In the context of expert object recognition, people are expert are recognizing many types of objects, so the images are drawn from a mixture of distributions.

Another advantage of EOR's design is that the category and exemplar subsystems exploit different properties of the data. The vectors clustered by the category system are combinations of Gabor responses. As a result, they exploit (multi-scale) edge information in images. The exemplar subsystem, on the other hand, projects raw images into the subspace. As a result, the first stage groups according to boundary information, while the second phase includes information about intensities.

Experiment

To test EOR, we collected a dataset of 100 images of cat faces and 100 images of dog faces, some of which are shown in Figure 4. (Biological studies clearly show that cat and dog face activate the FFA [9, 10].) No subjects are repeated in the database, which contains images of 200 different animals. The images are 64x64 pixels, and

have been registered by hand so that the eyes are in approximately the same position in every image.

Fig. 4. Samples of images from Cat and Dog data set.

The system is trained on 160 images; the remaining 40 images are saved for testing. Test images are then presented to the system, which retrieves the closest matches from memory. If the retrieved image is of the same species (dog or cat) as the test image, the trial is a success, otherwise it is a failure. The system was trained 25 times (using randomly selected sets of 160 training images), yielding a total of 1,000 trials. Overall, EOR succeeded 89.9% of the time, as shown in Table 1.

Is this a good result? We compare our results to several standard techniques. The first baseline is global PCA followed by nearest neighbor image retrieval (labeled "PCA" in Table 1). The second is even simpler: we simply correlate the test image to the training images, and retrieve the training image with the highest correlation score. The third baseline correlates the Gabor edge responses of the training and test images. These approaches are labeled "Corr" and "Edge Corr." in Table 1. For completeness, we also clustered the edge responses of the training data, and then labeled test images as cat or dog according to the dominant label in the nearest cluster. To our surprise, this worked better if we gave it only the highest resolution Gabor responses, rather than a pyramid (starting at 32x32) of Gabor responses (see the last two columns in Table 1).

EOR outperformed all five baseline techniques. As shown in Table 1, the performance improvement of EOR over PCA and clustering is statistically significant at the 95% confidence level, according to McNemar's significance test for paired binomial values. This is interesting, since these are the techniques combined inside EOR. The improvement over correlation is significant at only a 90% confidence level, and therefore needs to be verified in other studies.

	EOR	PCA	Corr	Edge Corr	Cluster (full res)	Cluster (multiscale)
% Correct	89.9%	88.3%	88.6%	89.2%	85.1%	73.7%
$P(H_0)$	--	4.44%	8.27%	9.79%	0.03%	0%

Table 1. Recognition rates for EOR, PCA, correlation, and K-Means, and McNemar's confidence values for EOR vs. other techniques.

Conclusions

People are experts at recognizing objects they see often, even when many instances look alike. Moreover, they recognize familiar objects very quickly, and

categorize them at both the class and instance level. Brain imaging studies identify this type of expert object recognition as a specific visual skill, and suggest an anatomical pathway involving the fusiform gyrus and right inferior frontal gyrus. This paper proposes a computational model of this pathway as unsupervised clustering followed by localized subspace projection, and shows that this model outperforms global PCA, correlation, and K-Means clustering on the task of discriminating between cats and dogs.

References

[1] A. D. Milner and M. A. Goodale, *The Visual Brain in Action*. Oxford: Oxford University Press, 1995.

[2] K. Nakamura, R. Kawashima, N. Sata, A. Nakamura, M. Sugiura, T. Kato, K. Hatano, K. Ito, H. Fukuda, T. Schormann, and K. Zilles, "Functional delineation of the human occipito-temporal areas related to face and scene processing: a PET study," *Brain*, vol. 123, pp. 1903-1912, 2000.

[3] K. M. O'Craven and N. Kanwisher, "Mental Imagery of Faces and Places Activates Corresponding Stimulus-Specific Brain Regions," *Journal of Cognitive Neuroscience*, vol. 12, pp. 1013-1023, 2000.

[4] I. Gauthier and M. J. Tarr, "Unraveling mechanisms for expert object recognition: Bridging Brain Activity and Behavior," *Journal of Experimental Psychology: Human Perception and Performance*, vol. in press, 2002.

[5] A. Puce, T. Allison, J. C. Gore, and G. McCarthy, "Face-sensitive regions in human extrastriate cortex studied by functional MRI," *Journal of Neurophysiology*, vol. 74, pp. 1192-1199, 1995.

[6] V. P. Clark, K. Keil, J. M. Maisog, S. Courtney, L. G. Ungeleider, and J. V. Haxby, "Functional Magnetic Resonance Imaging of Human Visual Cortex during Face Matching: A Comparison with Positron Emission Tomography," *NeuroImage*, vol. 4, pp. 1-15, 1996.

[7] N. Kanwisher, M. Chun, J. McDermott, and P. Ledden, "Functional Imaging of Human Visual Recognition," *Cognitive Brain Research*, vol. 5, pp. 55-67, 1996.

[8] E. Maguire, C. D. Frith, and L. Cipolotti, "Distinct Neural Systems for the Encoding and Recognition of Topography and Faces," *NeuroImage*, vol. 13, pp. 743-750, 2001.

[9] F. Tong, K. Nakayama, M. Moscovitch, O. Weinrib, and N. Kanwisher, "Response Properties of the Human Fusiform Face Area," *Cognitive Neuropsychology*, vol. 17, pp. 257-279, 2000.

[10] L. L. Chao, A. Martin, and J. V. Haxby, "Are face-responsive regions selective only for faces?," *NeuroReport*, vol. 10, pp. 2945-2950, 1999.

[11] A. Ishai, L. G. Ungerleider, A. Martin, J. L. Schouten, and J. V. Haxby, "Distributed representation of objects in the human ventral visual pathway," *Science*, vol. 96, pp. 9379-9384, 1999.

[12] M. J. Tarr and I. Gauthier, "FFA: a flexible fusiform area for subordinate-level visual processing automatized by expertise," *Neuroscience*, vol. 3, pp. 764-769, 2000.

[13] I. Gauthier, M. J. Tarr, J. Moylan, A. W. Anderson, P. Skudlarski, and J. C. Gore, "Does Visual Subordinate-level Categorization Engage the Functionally Defined Fusiform Face Area?," *Cognitive Neuropsychology*, vol. 17, pp. 143-163, 2000.

[14] J. W. Tanaka and T. Curran, "A Neural Basis for Expert Object Recognition," *Psychological Science*, vol. 12, pp. 43-47, 2001.

[15] T. S. Lee, D. Mumford, R. Romero, and V. A. F. Lamme, "The role of the primary visual cortex in higher level vision," *Vision Research*, vol. 38, pp. 2429-2454, 1998.

[16] J. V. Haxby, L. G. Ungerleider, V. P. Clark, J. L. Schouten, E. A. Hoffman, and A. Martin, "The Effect of Face Inversion on Activity in Human Neural Systems for Face and Object Recognition," *Neuron*, vol. 22, pp. 189-199, 199.

[17] I. Gauthier, M. J. Tarr, A. W. Anderson, P. Skudlarski, and J. C. Gore, "Behavioral and Neural Changes Following Expertise Training," presented at Annual Meeting of the Psychonomic Society, Philadelphia, PA, 1997.

[18] S. M. Kosslyn, *Image and Brain: The Resolution of the Imagery Debate.* Cambridge, MA: MIT Press, 1994.

[19] S. M. Kosslyn, "Visual Mental Images and Re-Presentations of the World: A Cognitive Neuroscience Approach," presented at Visual and Spatial Reasoning in Design, Cambridge, MA, 1999.

[20] S. M. Kosslyn, A. Pascual-Leone, O. Felician, S. Camposano, J. P. Keenan, W. L. Thompson, G. Ganis, K. E. Sukel, and N. M. Alpert, "The Role of Area 17 in Visual Imagery: Convergent Evidence from PET and rTMS," *Science*, vol. 284, pp. 167-170, 1999.

[21] N. Petkov and P. Kruizinga, "Computational models of visual neurons specialised in the detection of periodic and aperiodic oriented stimuli: bar and grating cells," *Biological cybernetics*, vol. 76, pp. 83-96, 1997.

[22] D. A. Pollen, J. P. Gaska, and L. D. Jacobson, "Physiological Constraints on Models of Visual Cortical Function," in *Models of Brain Functions*, M. Rodney and J. Cotterill, Eds. New York: Cambridge University Press, 1989, pp. 115-135.

[23] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*. New York: John Wiley and Sons, 1973.

[24] M. Turk and A. Pentland, "Eigenfaces for Recognition," *Journal of Cognitive Neuroscience*, vol. 3, pp. 71-86, 1991.

[25] A. Hyvärinen and E. Oja, "Independent Component Analysis: Algorithms and Applications," *Neural Networks*, vol. 13, pp. 411-430, 2000.

[26] B. G. Tabachnick and L. S. Fidell, *Using Multivariate Statistics.* Boston: Allyn & Bacon, Inc., 2000.

[27] K. Baek and B. A. Draper, "Factor Analysis for Background Suppression," presented at International Conference on Pattern Recognition, Quebec City, 2002.

[28] M. S. Bartlett, *Face Image Analysis by Unsupervised Learning*: Kluwer Academic, 2001.

[29] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788-791, 1999.

[30] K. Baek, B. A. Draper, J. R. Beveridge, and K. She, "PCA vs ICA: A comparison on the FERET data set," presented at Joint Conference on Information Sciences, Durham, N.C., 2002.

[31] P. Belhumeur, J. Hespanha, and D. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 19, pp. 711-720, 1997.

[32] B. J. Frey, A. Colmenarez, and T. S. Huang, "Mixtures of Local Linear Subspaces for Face Recognition," presented at IEEE Conference on Computer Vision and Pattern Recognition, Santa Barbara, CA, 1998.

[33] N. Kambhatla and T. K. Leen, "Dimension Reduction by Local PCA," *Neural Computation*, vol. 9, pp. 1493-1516, 1997.