Implementing the Expert Object Recognition Pathway

Bruce A. Draper, Kyungim Baek and Jeff Boody

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

Brain imaging studies suggest that expert object recognition is a distinct visual skill, implemented by a dedicated anatomic pathway. Like all visual pathways, the expert recognition pathway begins with the early visual system (retina, LGN/SC, striate cortex). It is defined, however, by subsequent diffuse activation in the lateral occipital cortex (LOC), and sharp foci of activation in the fusiform gyrus and right inferior frontal gyrus. This pathway recognizes familiar objects from familiar viewpoints under familiar illumination. Significantly, it identifies objects at both the categorical and instance (subcategorical) levels, and these processes cannot be disassociated. This paper presents a four-stage functional model of the expert object recognition pathway, where each stage models one area of anatomic activation. It implements this model in an end-to-end computer vision system, and tests it on real images to provide feedback for the cognitive science and computer vision communities.

1 Introduction

In the introduction to his book, David Marr argued that complex systems are more than just the easily extrapolated properties of their primitive components, and need to be modeled at many levels of abstraction [22]. As an example, Marr sited gases in physics, which can be modeled either at the molecular level or at the level of the ideal gas law. The ideal gas law describes gases in terms of collective properties such as temperature and pressure that are not easily extracted from the molecular model. By analogy, Marr proposed three levels for modeling information processing systems: the functional level, an algorithm and representation level, and the implementation level.

Marr proposed these three levels while studying human vision in the 1970's. His argument is even stronger today, with the advent of brain imaging technologies such as fMRI, PET, and rTMS. These sensors measure responses not of individual neurons, but of large collections of neurons. This is more like measuring the pressure of a gas than the properties of individual molecules. We therefore model the human visual system at the functional level based on data from brain imaging studies.

The human visual system, however, is not one pathway but a collection of related subsystems. The best known division is the ventral/dorsal split [34], but brain imaging studies suggest that the ventral and dorsal streams are themselves divided into many subsystems. One of the ventral subsystems is the expert object recognition pathway, which recognizes familiar objects such as human faces, pets and chairs, when seen from familiar viewpoints. The expert recognition pathway begins with the early vision system. It is anatomically defined in brain imaging studies by additional centers of activation in the fusiform gyrus and right inferior frontal gyrus, and diffuse activation in the lateral occipital complex (LOC).

The goal of this paper is to present an implementation of a functional model of the expert object recognition pathway. The model is divided into four stages: Gabor-based edge detection in the early visual system, non-accidental feature transformations in the LOC, unsupervised clustering in the fusiform gyrus and PCA-based subspace matching in the right inferior frontal gyrus. Sections 2-4 of this paper provide background on expert object recognition and appearance-based models of human object recognition. Section 5 describes the four processing stages. Section 6 applies the system to real-world data, and Section 7 draws conclusions.

2 Expert Object Recognition

The expert object recognition pathway was first identified in fMRI studies of human face recognition [6, 14, 29]. In these studies, patients were shown images of 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 also noting activation in the right inferior frontal gyrus, an area previously associated through lesion studies with visual memory [20] (see also [23]).

More recent evidence suggests that this pathway is used for more than recognizing faces. Tong, et al. report that the fusiform gyrus is activated by animal faces and cartoon faces [33]. Chao, et al. report that the fusiform gyrus is activated by images of full-bodied animals with obscured faces [5]. Ishai et al. find that the fusiform gyrus responds to chairs [11]. Tarr and Gauthier considered the past experience of their subjects, and found fusiform gyrus activation in dog show judges when they view dogs, and in bird experts when they view birds [31]. Most important of all, Tarr and Gauthier show that the expert recognition pathway is trainable. They created a class of cartoon characters called greebles, which are grouped by gender and family. When novice subjects view greebles, fMRIs show no activity in the fusiform gyrus. The subjects are then trained to be experts who can identify a greeble’s identity, gender or family in equal time. When the experts view greebles, their fusiform gyrus is active [31]. Gauthier and Logothetis provide evidence that training produces similar results in monkeys [8]. We conclude that expert object recognition is a general mechanism that can be trained to recognize any class of familiar objects.

3 Properties of Expert Object Recognition

People become expert at recognizing familiar objects, such as faces, animals and chairs. As experts, they can recognize these objects at both the instance and category level. Kosslyn uses pencils as an example [15]: we can all recognize pencils, but if we have one long enough we also recognize *our* pencil, from its dents and imperfections. This multiple-level categorization is used to define expert recognition in the greeble studies sited above.

Expert object recognition is also viewpoint dependent. In fMRI studies, the response of the fusiform gyrus to images of upside down faces is minimal [10]. When upright and inverted greebles are presented to experts, only the upright greebles activate the fusiform gyrus [9]. Expert recognition is also illumination dependent; our accuracy at face recognition, for example, drops if the faces are presented upside down or illuminated from below [2].

Expert object recognition is fast. In ERP studies, face recognition can be detected through a negative N170 signal that occurs 140-188 ms post stimulus [30]. Since the fusiform gyrus and right inferior frontal gyrus are the unique structures in expert recognition, we assume that one or both become active at this time, leaving only about 140ms for processing in the early vision system and LOC.

Finally, expert object recognition is probably appearance based. We know that expert recognition activates the right inferior frontal gyrus, an area associated with visual memories. We also know from PET and rTMS data that visual memories can reconstruct image-like representations in the primary visual cortex [16]. These memories can therefore be viewed as a form of compressed image [15], implying that expert recognition is a form of image matching.

4 Modeling the Expert Object Recognition Pathway

We interpret the activation data from fMRI and PET studies of expert object recognition as a four-stage pipeline. Processing begins in the early visual system, and then proceeds through the LOC to the fusiform gyrus and the right inferior frontal gyrus, as shown in Figure 1. This model is similar to Kosslyn’s model of the ventral visual stream, which also featured an early visual system, feature-based “pre-processing”, and interacting categorization and exemplar matching subsystems [15]. This section describes each stage; the next section applies the model to real data.

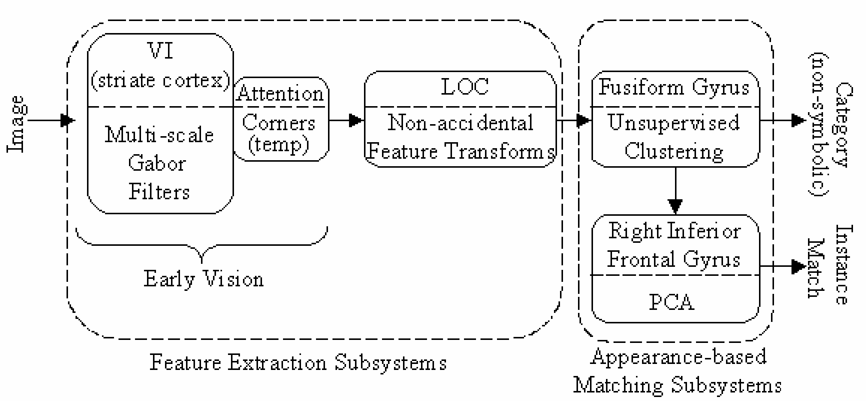

Figure 1. The major components of the human expert object recognition pathway.

4.1 Early Vision

Computational models of the early visual system have a long history [26]. Of particular interest are functional models of simple and complex cells in V1. Through single cell recordings, Pollen and others have shown that the outputs of cells in V1 can be directly modeled in terms of visual stimuli, combining the effects of retinal, LGN and V1 processing. Simple cell responses in V1 can be modeled as Gabor filters of the stimulus, parameterized by location, orientation, scale and phase. Complex cell responses combine the energy of Gabor filters across phases [28]. Following Pollen, this work models early vision as a bank of multi-scale Gabor filters. Our system computes an image pyramid from the input, convolves it with non-symmetric even and odd Gabor filters at every 15º of orientation, and computes the resulting energy.

It should be noted that the responses of V1 cells can be modulated by portions of the stimulus outside their classically defined receptive fields [18, 35]. This conflicts with the model of complex cells as Gabor filters, but the first modulation effects do not occur until 80-120ms post stimulus. From ERP studies, it seems unlikely that contextual modulation effects appear soon enough to influence expert recognition.

Although the early vision system processes the whole retinal image through a bank of Gabor filters, not all of this information is passed downstream to the ventral and dorsal systems. Instead, a portion of this data is selected by position (and possibly scale or frequency [24]) for further processing. Parkhurst, et al are able to show a positive correlation between human eye tracking and a bottom-up model of attention selection based on color, intensity and orientation. [27]. Maki et al present a model based on image flow, motion and stereo [21]. Unfortunately, the system described in this paper does not yet use a biological model of attention selection. Instead, it runs a corner detector over the image, and successively selects image patches around each corner. In the future, we hope to replace this with the attentional model in the Neuormorphic Vision Toolkit [12] (this is the system evaluated by Parkhurst, et al).

4.2 Modeling the Lateral Occipital Complex (LOC)

The lateral occipital complex is a large area of the brain that is diffusely active during object recognition. Using fMRI, Kourtzi and Kanwisher show object selective activation in the LOC, and demonstrate through fatigue effects that cells in the LOC respond to structural (edge-based) properties [17]. Although their study can't determine what the structural properties are, Kosslyn and others [15] have suggested they could be non-accidental properties of the type proposed by Lowe [19] and Biederman [1]. Examples include edge collinearity, parallelism, symmetry and anti-symmetry. Psychological studies show that line drawings with non-accidental features obscured are harder to recognize than obscured line drawings with non-accidental features intact [1].

This work models the LOC as computing fixed length non-accidental feature transforms. The first and simplest example is the Hough transform – it projects edge responses into the space of geometric lines, thereby making collinearity explicit. As long as the temptation to threshold the Hough space and produce symbolic lines is avoided, the Hough space is an appropriate feature representation for appearance-based recognition. We are currently developing new transforms to capture other non-accidental features, such as parallelism, symmetry and anti-symmetry. The preliminary results in this paper, however, show the surprisingly powerful results of modeling the LOC as a Hough transform.

4.3. Categorization: Modeling the Fusiform Gyrus

Together, the early vision system and the LOC form a feature extraction subsystem, with the early vision system computing Gabor features and the LOC transforming them into non-accidental feature vectors, as shown in Figure 1. Similarly, the fusiform gyrus and right inferior frontal gyrus combine to form a feature-based appearance matching subsystem.

The appearance-based matching system is divided into two components: an unsupervised clustering system and a subspace projection system. This is motivated by the psychological observation that categorical and instance level recognition cannot be disassociated, and the mathematical observation that subspace projection methods exploit the commonality among images to compress data. If the images are too diverse, for example pictures of faces, pets, and chairs, then there is no commonality for the subspaces to exploit.

To avoid this, we model the fusiform gyrus as an unsupervised clustering system, and the right inferior frontal gyrus as a subspace matching system. This anatomical mapping is partly for simplicity; the exact functional division between these structures is not clear. Lesion studies associate the right inferior frontal lobe with visual memory [20], and rTMS and PET data suggest that these memories are compressed images [16]. Since compressed memories are stored in the frontal gyrus, it is easy to imagine that they are matched there as well, perhaps using an associative network. At the same time, clustering is the first step that is unique to expert recognition and the fusiform gyrus is the first anatomically unique structure on the expert pathway, so it makes sense to associate clustering with the fusiform gyrus. Where images are projected into cluster-specific subspaces is not clear however; it could be in either location, or both.

It is important to note that the categories learned by the clustering mechanism in the fusiform gyrus are non-linguistic. The images in a cluster do not need to be of the same object type or viewpoint, nor do all images of one object need to appear in one cluster. Clustering simply divides the training data into small groups of similar samples, so that PCA can fit a unique subspace to each group. This is similar to the localized subspace projection models in [7, 13]. We have implemented K-Means and an EM algorithm for mixtures of PCA analyzers similar to [32]. Surprisingly, so far we get the best results by using K-Means and overestimating the number of clusters K, possibly because non-symmetric Gaussians can be estimated by collections of symmetric ones.

4.4 Appearance Matching in the Right Inferior Frontal Gyrus

The last stage applies subspace projection to every cluster of feature vectors, and stores the training sample in the compressed subspaces. Currently, PCA is used as the subspace projection mechanism. New images are assigned to a cluster and projected into that cluster's PCA subspace, where nearest neighbor retrieval selects the best available instance match.

PCA is not a new model of human object recognition. Bülthoff and Edelman first used PCA to model of human object recognition [3], and O'Toole showed that human memories for individual faces correlate to the quality of their PCA reconstruction [25]. Bülthoff in particular has focused on view interpolation for viewpoint invariant appearance-based object recognition [4].

The computational model presented in this paper is more modest than Bülthoff's proposal, in the sense that it only models expert object recognition, not human object recognition in general. As a result, PCA is not used for view interpolation, since expert recognition is not viewpoint invariant. Moreover, our system first transforms the source image with non-accidental features of the Gabor responses, and then groups these features into localized subspaces prior to matching, where Bülthoff's model uses a single PCA space to match images.

5 Performance

We implemented the system described in Section 4 and tested it on two domains: aerial images of Fort Hood, TX, and facial images of cats and dogs. For the cat and dog data (shown in Figure 3), the images were already small (64x64 pixels) and hand registered, so the selective attention mechanism was disabled. For the Fort Hood data, each source image is 1000x1000 pixels and contains approximately 10,000 corners (i.e. possible attention points). We randomly selected 100 points on each of four object types for further processing. Similarly, we randomly selected 400 attention points on another, non-overlapping image for testing. Figure 4 shows example attention windows for each type of object (two building styles, paved parking lots, and unpaved parking areas).

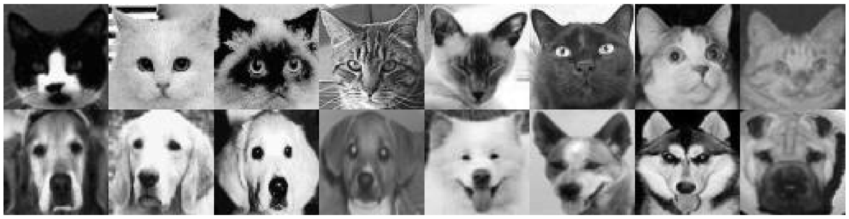

Figure 3. Examples of images from the cat and dog data base.

Our model of expert object recognition uses only unsupervised learning, so no object labels were provided during training. During testing, the system retrieves a cluster and stored image for every attention window. Since clusters do not correspond to semantic labels, the cluster response is not evaluated. A trial is a success if the retrieved instance match is of the same object type as the test window.

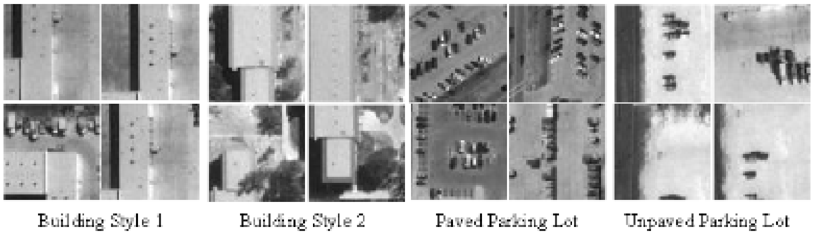

Figure 4: Examples of Building Styles 1 & 2 and paved and unpaved parking lots in aerial images of Fort Hood, TX.

In Figure 5, we compare the performance of the biomimetic system to a baseline system that applies PCA to the pixels in an image pyramid. The horizontal axis is the number of PCA dimensions retained; and the vertical axis is the instance-level recognition rate. The biomimetic model clearly outperforms PCA, which is reassuring, since it uses PCA as its final step. It would have been disappointing if all the additional mechanisms failed to improve performance!

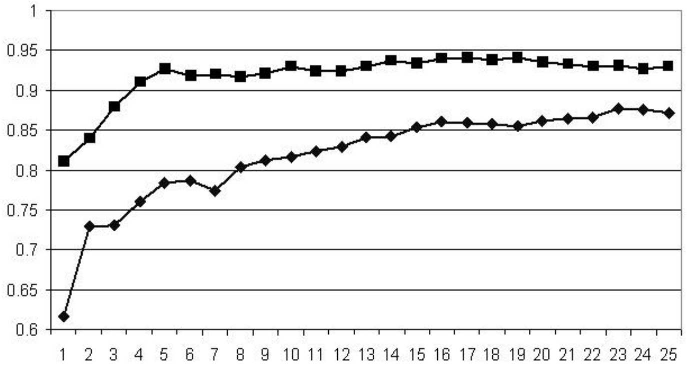

Figure 5: Recognition rates for the proposed biomimetic system (plotted with squares) versus a baseline PCA system (plotted with diamonds) on the cat and dog data. The horizontal axis is the number of PCA dimension retained.

The more interesting question is why the system performs better. Figure 6 shows the results from a credit assignment experiment on the cat and dog data where system components isolated. In the baseline system, an image pyramid is computed for each image, and a single PCA is computed for pixels in the pyramid. In other words, the Gabor filters, non-accidental transforms and clustering have been disabled. (This is also the baseline for Figure 5.) We then reintroduced the Gabor filters, applying PCA to the energy values produced by the complex cell models. Performance does not improve, in fact it degrades, as shown in Figure 6. Next we reintroduced the Hough transform, so that PCA is applied to the Hough space. Performance improves markedly, approaching the best recognition rates for the system as a whole. This suggests that the LOC model is critical to overall system performance. It also calls into question the need for clustering, since recognition performance is essentially the same with or without it (see Figures 5 & 6).

Further experiments confirm that clustering only marginally improves recognition rates when the number of subspace dimensions is large (see Figure 7). What clustering does is force the images stored in a subspace to be similar, allowing for more compression. As a result, peak recognition performance is reached with fewer subspace dimensions, as shown iconically at the bottom of Figure 7. Clustering therefore improves the system's ability to compress visual memories.

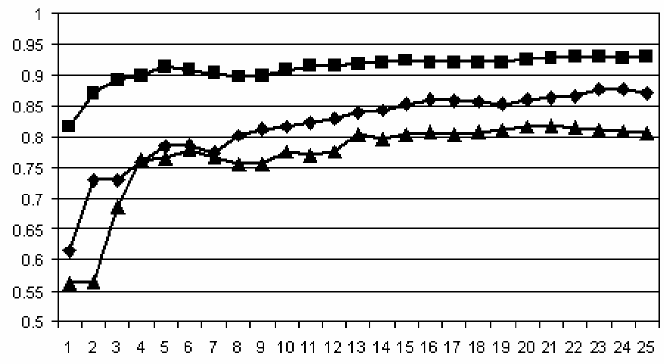

Figure 6: Recognition rates vs. number of subspace dimensions for (a) PCA applied to image pyramid pixels, plotted with diamonds; (b) PCA applied to Gabor energy responses, plotted with triangles; and (c) PCA applied to the Hough transform, plotted with squares.

6 Conclusion

The most surprising result so far from our model of expert object recognition is the performance of the Hough transform with PCA. Most appearance-based methods apply PCA to raw images or to the results of simple image operations (e.g. image differences). We observe a significant benefit, however, from applying PCA to the output of a Hough transform in two domains, even though cat and dog faces have few straight lines. We do not observe the same benefit when PCA is applied to the outputs of the Gabor filters. We hypothesize that the recognition rate increases because the Hough transform makes collinearity (a non-accidental property) explicit.

We also observe that clustering to create localized PCA projections improves compression more than recognition. This may only be true for instance matching tasks; in classification tasks the PCA subspace represent an underlying class probability distribution, and Mahalanobis distances are meaningful. Localized PCA subspaces may therefore improve the recognition rate. In instance matching, however, clustering improves compression but not recognition.

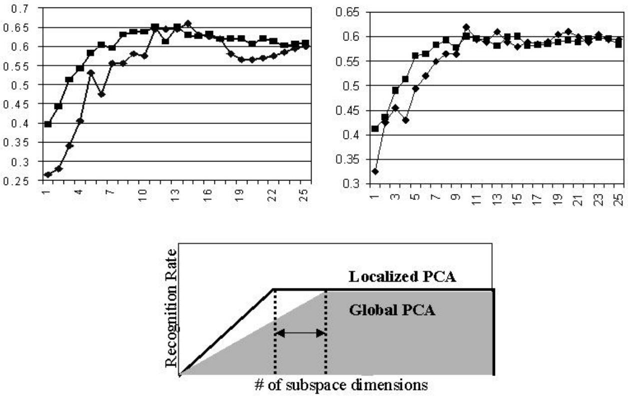

Figure 7: The number of subspace dimensions (horizontal) vs. recognition rate (vertical) with and without localized clustering for the Fort Hood data. The plot on the left is for localized PCA applied to the Hough transform; the plot on the right is for localized PCA applied to complex cell responses. The bottom figure summarizes these and other plots, showing how clustering improves compression by achieving the maximum recognition rate with fewer subspace dimensions.

Finally, our work suggests that the LOC needs to be studied more closely. The LOC determines the overall recognition rate for our computational model, yet we have less information about it than any other anatomical component of the system. We cannot even be sure that the results reported by Kourtzi and Kanwisher [17] and Biederman [1] apply in the special case of expert recognition. More studies are needed.

References

[1] I. Biederman, "Recognition-by-Components: A Theory of Human Image Understanding," *Psychological Review*, vol. 94, pp. 115-147, 1987.

[2] V. Bruce and A. Young, *In the Eye of the Beholder: The Science of Face Perception*. New York: Oxford University Press, 1998.

[3] H. H. Bülthoff and S. Edelman, "Psychophysical Support for a 2-D View Interpolation Theory of Object Recognition," *Proceedings of the National Academy of Science*, vol. 89, pp. 60-64, 1992.

[4] H. H. Bülthoff, C. Wallraven, and A. Graf, "View-based Dynamic Object Recognition based on Human Perception," International Conference on Pattern Recognition, Quebec City, 2002.

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

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

[8] I. Gauthier and N. K. Logothetis, "Is Face Recognition Not So Unique After All?," *Cognitive Neuropsychology*, vol. 17, pp. 125-142, 2000.

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

[10] 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.

[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] L. Itti, "Modeling Primate Visual Attention," in *Computational Neuroscience: A Comprehensive Approach*, J. Feng, Ed. Boca Raton, FL: CRC Press, 2002.

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

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

[15] S. M. Kosslyn, *Image and Brain*. Cambridge, MA: MIT Press, 1994.

[16] 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.

[17] Z. Kourtzi and N. Kanwisher, "Cortical Regions Involved in Perceiving Object Shape," *The Journal of Neuroscience*, vol. 20, pp. 3310-3318, 2000.

[18] 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.

[19] D. G. Lowe, *Perceptual Organization And Visual Recognition*. Boston: Kluwer, 1985.

[20] 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.

[21] A. Maki, P. Nordlund, and J.-O. Eklundh, "Attentional Scene Segmentation: Integrating Depth and Motion from Phase," *Computer Vision and Image Understanding*, vol. 78, pp. 351-373, 2000.

[22] D. Marr, *Vision*. Cambridge, MA: Freeman, 1982.

[23] 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.

[24] A. Oliva and P. G. Schyns, "Coarse Blobs or Fine Edges? Evidence That Information Diagnoticity Changes the Perception of Complex Visual Stimuli," *Cognitive Psychology*, vol. 34, pp. 72-107, 1997.

[25] A. J. O'Toole, K. A. Deffenbacher, D. Valentin, and H. Abdi, "Structural Aspects of Face Recognition and the Other Race Effect," *Memory and Cognition*, vol. 22, pp. 208-224, 1994.

[26] S. E. Palmer, *Vision Science: Photons to Phenomenology*. Cambridge, MA: MIT Press, 1999.

[27] D. Parkhurst, K. Law, and E. Neibur, "Modeling the role of salience in the allocation of overt visual attention," *Vision Research*, vol. 42, pp. 107-123, 2002.

[28] 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.

[29] 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.

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

[31] 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.

[32] M. E. Tipping and C. M. Bishop, "Mixtures of Probabilistic Principal Component Analysers," *Neural Computation*, vol. 11, pp. 443-482, 1999.

[33] 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.

[34] L. G. Ungeleider and M. Mishkin, "Two cortical visual systems," in *Analysis of visual behavior*, D. J. Ingle, M. A. Goodale, and R. J. W. Mansfield, Eds. Cambridge, MA: MIT Press, 1982, pp. 549-586.

[35] K. Zipser, V. A. F. Lamme, and P. H. Schiller, "Contextual Modulation in Primary Visual Cortex," *Neuroscience*, vol. 16, pp. 7376-7389, 1996.