Implementing the expert object recognition pathway

Bruce A. Draper, Kyungim Baek*, Jeff Boody

Department of Computer Science, Colorado State University, Fort Collins, CO, 80523, USA

Abstract. Brain imaging studies suggest that expert object recognition is a distinct visual skill, implemented by a dedicated anatomical 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 complex (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 (a.k.a. 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.

Keywords: Cognitive vision – Computer vision – Biological vision – Appearance-based vision

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 [23]. The primitive components of complex systems interact in nontrivial ways to produce phenomena that are not easily measured or explained at the level of the components. As an example, Marr cited the study of gases, which are composed of freely moving molecules. Nonetheless, the ideal gas law describes gases in terms of collective properties such as temperature and pressure that are not easily modeled at the molecular level. To understand these types of phenomena, complex systems need to be modeled at many levels of abstraction. Marr proposed three levels for modeling information processing systems: the functional level, the algorithm and representation level, and the implementation level [23].

Marr proposed these levels while studying human vision in the 1970s. His argument is even stronger today, with the advent of brain imaging technologies such as fMRI, PET, and rTMS. Unlike single-cell recordings, these sensors measure the responses of large collections of neurons. This is akin to measuring the pressure of a gas rather than the energy of individual molecules, and it suggests that brain imaging data should be used to build cognitive models of the human visual system at the functional and algorithmic levels, rather than the implementation level of neural models.

This paper models human vision at a functional level, based largely on data from brain imaging studies. It attempts to associate specific functions with gross anatomical structures and to implement these structures so that their behaviors and interactions can be studied in the context of real data.

It is important not to oversimplify the human visual system when modeling it. Human vision is not one mechanism but a collection of related subsystems. The best known division within the human visual system is the ventral/dorsal split [37], 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).

This paper presents an implementation of a functional model of the expert object recognition pathway. The model is divided into four stages: selective attention and Gabor filters in the early visual system, nonaccidental 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 Sect. 7 draws conclusions.

---

* *Current address*: Department of Biomedical Engineering, Columbia University, New York, NY, USA
Correspondence to: B.A. Draper (e-mail: draper@cs.colostate.edu)

2 Expert object recognition

The expert object recognition pathway was first identified in fMRI studies of human face recognition [6,15,31]. 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 imaged a larger portion of the brain and 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 [21] (see also [24]).

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 [35]. 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 [12]. 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 [33]. 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 become experts who can identify a greeble's identity, gender or family in equal time. When trained subjects view greebles, their fusiform gyrus is active [33]. Gauthier and Logothetis provide evidence that training produces similar results in monkeys [9]. 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

In order to model expert object recognition, it is important to know what functional properties it has. The most important properties are that expert object recognition is trainable and combines categorical with instance-level (a.k.a. subcategorical) recognition. People become expert at recognizing familiar objects, such as faces, animals, and chairs, and they can recognize these objects at both the instance and category level. Kosslyn uses pencils as an example [16]: we can all recognize pencils, but if we have one long enough we also recognize *our* pencil, from its dents and imperfections. Multiple-level categorization was used to define expert recognition in the greeble studies cited above.

Expert object recognition is viewpoint dependent. In fMRI studies, the response of the fusiform gyrus to images of upside-down faces is minimal [11]. When upright and inverted greebles are presented to experts, only the upright greebles activate the fusiform gyrus [10]. 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 poststimulus [32]. 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 140 ms or less 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 imagelike representations in the primary visual cortex [17]. These memories can therefore be viewed as a form of compressed image [16], implying that expert recognition is a form of image matching.
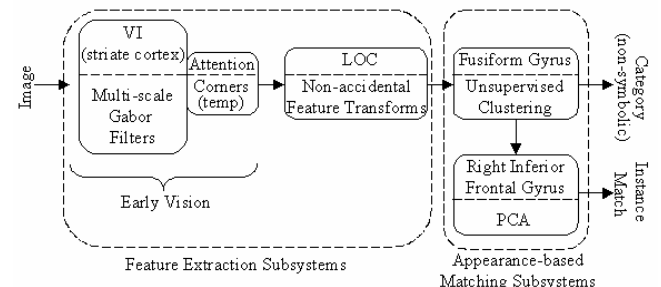
4 Modeling the expert object recognition pathway

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 Fig. 1. This model is similar to Kosslyn's model of the ventral visual stream, which also featured an early visual system, feature-based "preprocessing", and interacting categorization and exemplar matching subsystems [16]. This section describes each stage; the next section applies the model to real data.

4.1 Early vision

Computational models of the early visual system have a long history [27]. 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 [30]. Emulating Pollen, we model early vision as a bank of multiscale Gabor filters. Our system computes an image pyramid from the input, convolves it with nonsymmetric even and odd Gabor filters at every 15° of orientation, and computes the resulting energy.

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 stream. Instead, a

Fig. 1. Major components of the human expert object recognition pathway

portion of these data is selected by position (and possibly scale or frequency [25]) for further processing. Parkhurst et al. are able to show a positive correlation between human eye tracking data and a bottom-up model of selective attention based on color, intensity, and orientation [29]. Maki et al. present a model based on image flow, motion, and stereo [22]. Other computational models of selective attention are presented by Tsotsos et al. [36] and Park et al. [28]. We have recently developed a computational selective attention system based on the work of Itti et al. [13] and designed to be insensitive to similarity transforms of the source image [7]. 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.

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 [18]. Although their study cannot determine what the structural properties are, Kosslyn and others [16] have suggested they could be nonaccidental properties of the type proposed by Lowe [20] and Biederman [1]. Examples include edge collinearity, parallelism, symmetry, and antisymmetry. Psychological studies show that line drawings with nonaccidental features obscured are harder to recognize than obscured line drawings with nonaccidental features intact [1].

Another clue about LOC processing may come from the nonclassical, modulated responses of V1 cells. As discussed in the previous section, simple and complex cells in V1 initially respond as Gabor filters of the input stimulus. However, later responses of V1 cells (80–120 ms poststimulus) are modulated by portions of the stimulus outside their classically defined receptive fields [19,38]. These modulations appear to "fill in" gaps in lines or curves and to respond to axes of symmetry. This may be feedback from deeper in the visual stream, possibly the LOC. If so, it provides further evidence that collinear and parallel properties may be computed there.

This work models the LOC as computing fixed-length nonaccidental 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 nonaccidental features, such as parallelism, symmetry, and antisymmetry. The preliminary results in this paper, however, show the surprisingly powerful results of simplistically 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 selecting

fixation windows and computing Gabor features, and the LOC computing nonaccidental feature vectors, as shown in Fig. 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, 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 [21], and rTMS and PET data suggest that these memories are compressed images [17]. 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 we 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 appearance based. 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 [8,14]. We have implemented K-means and an EM algorithm for mixtures of PCA analyzers similar to [34]. Surprisingly, so far we get the best results by using K-means and overestimating the number of clusters K, possibly because nonsymmetric 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 human object recognition [3], and O'Toole showed that human memories for individual faces correlate to the quality of their PCA reconstruction [26]. Bülthoff in particular has focused on view interpolation for viewpoint-invariant appearance-based object recognition [4].

Fig. 2. Examples of images from the cat and dog database
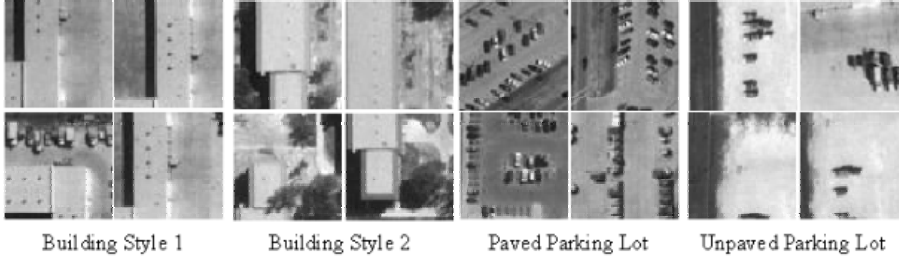
Fig. 3. Examples of building styles 1 and 2 and paved and unpaved parking lots in aerial images of Fort Hood, TX

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 nonaccidental features of the Gabor responses and then groups these features into localized subspaces prior to matching, where Bülthoff's model matches images in a single PCA space.

5 Performance

We implemented the system described in Sect. 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 Fig. 2), the images were already small (64×64 pixels) and hand registered, so the selective attention mechanism was disabled. For the Fort Hood data, each source image is 1000×1000 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, nonoverlapping image for testing. Figure 3 shows example attention windows for each type of object (two building styles, paved parking lots, and unpaved parking areas).

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 (i.e., category) and stored image (i.e., instance or subcategory) 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.

In Fig. 4, 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.
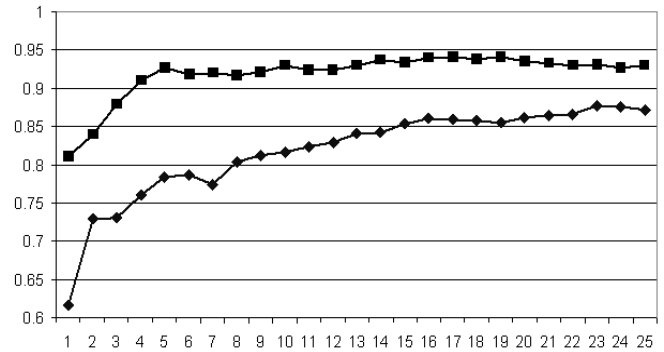

Fig. 4. Recognition rates for the proposed biomimetic system (*squares*) vs. a baseline PCA system (*diamonds*) on the cat and dog data. The horizontal axis is the number of PCA dimension retained
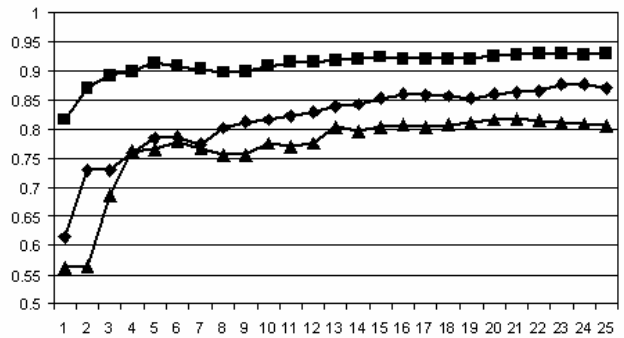
Fig. 5. Recognition rates vs. number of subspace dimensions for **a** PCA applied to image pyramid pixels (*diamonds*); **b** PCA applied to Gabor energy responses (*triangles*); and **c** PCA applied to the Hough transform (*squares*)

It would have been disappointing if all the additional mechanisms had failed to improve performance!

The more interesting question is why the system performs better. Figure 5 shows the results from a credit assignment experiment on the cat and dog data where system components are isolated. In the baseline system, an image pyramid is com-
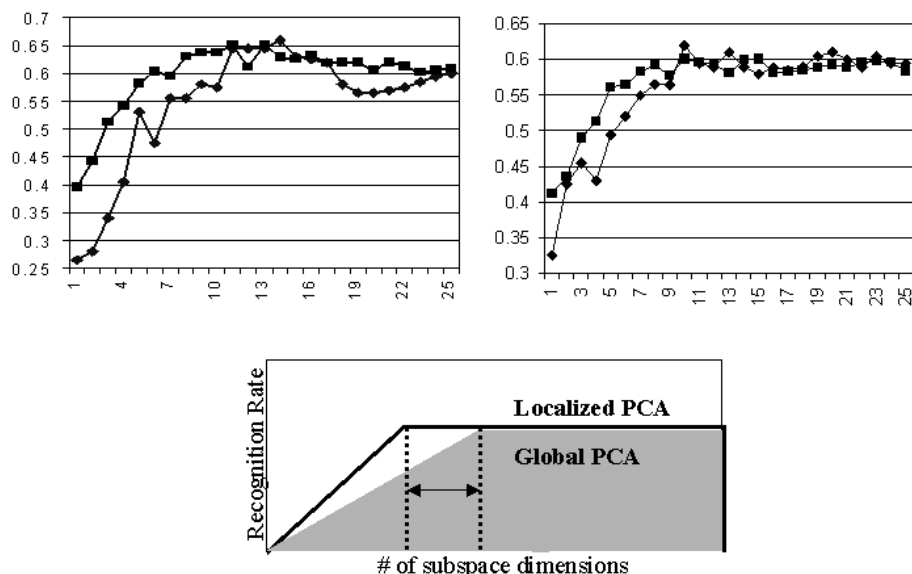
Fig. 6. Number of subspace dimensions (horizontal) vs. recognition rate (vertical) with and without localized clustering for the Fort Hood data. The plot on the *left* is for localized PCA applied to the Hough transform; the plot on the *right* is for localized PCA applied to complex cell responses. The *bottom* figure summarizes these and other plots, showing how clustering improves compression by achieving the maximum recognition rate with fewer subspace dimensions

puted for each image, and a single PCA subspace is computed from the pixels in the pyramid. In other words, with the Gabor filters, nonaccidental transforms and clustering are disabled. (This is also the baseline for Fig. 4.) 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 Fig. 5. 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 (Figs. 4 and 5).

Further experiments confirm that clustering only marginally improves recognition rates when the number of subspace dimensions is large (Fig. 6). What clustering does is group images by appearance, allowing for more image compression. As a result, peak recognition performance is reached with fewer subspace dimensions, as shown iconically at the bottom of Fig. 6. Clustering therefore improves the system's ability to compress visual memories, presumably allowing more memories to be retained.

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 one of the domains is cat and dog faces, which 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 nonaccidental 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 represents 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.

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 [18] and Biederman [1] apply in the special case of expert recognition. More biological studies are needed.

References

1. Biederman I (1987) Recognition-by-components: a theory of human image understanding. Psychol Rev 94(2):115–147
2. Bruce V, Young A (1998) In the eye of the beholder: the science of face perception. Oxford University Press, New York, p 280
3. Bülthoff HH, Edelman S (1992) Psychophysical support for a 2-D view interpolation theory of object recognition. Proc Natl Acad Sci USA 89:60–64
4. Bülthoff HH, Wallraven C, Graf A (2002) View-based dynamic object recognition based on human perception. In: Proceedings of the international conference on pattern recognition, Quebec City
5. Chao LL, Martin A, Haxby JV (1999) Are face-responsive regions selective only for faces? NeuroReport 10(14): 2945–2950
6. Clark VP et al (1996) Functional magnetic resonance imaging of human visual cortex during face matching: a comparison with positron emission tomography. NeuroImage 4:1–15
7. Draper BA, Lionelle A (2003) Evaluation of selective attention under similarity transforms. In: Proceedings of the workshop on performance and attention in computer vision, Graz, Austria
8. Frey BJ, Colmenarez A, Huang TS (1998) Mixtures of local linear subspaces for face recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. IEEE Press, Santa Barbara, CA

9. Gauthier I, Logothetis NK (2000) Is face recognition not so unique after all? Cogn Neuropsychol 17(1/2/3):125–142
10. Gauthier I et al (1997) Behavioral and neural changes following expertise training. In: Proceedings of the annual meeting of the Psychonomic Society, Philadelphia
11. Haxby JV et al (199♣?) The effect of face inversion on activity in human neural systems for face and object recognition. Neuron 22:189–199
12. Ishai A et al (1999) Distributed representation of objects in the human ventral visual pathway. Science 96:9379–9384
13. Itti L, Koch C, Neibur E (1998) A model of saliency-based visual attention for rapid scene analysis. IEEE Trans Patt Anal Mach Intell 20(11):1254–1259
14. Kambhatla N, Leen TK (1997) Dimension reduction by local PCA. Neural Comput 9:1493–1516
15. Kanwisher N et al (1996) Functional imaging of human visual recognition. Cogn Brain Res 5:55–67
16. Kosslyn SM (1994) Image and brain: the resolution of the imagery debate. MIT Press, Cambridge, MA, p 516
17. Kosslyn SM et al (1999) The role of area 17 in visual imagery: convergent evidence from PET and rTMS. Science 284:167–170
18. Kourtzi Z, Kanwisher N (2000) Cortical regions involved in perceiving object shape. J Neurosci 20(9):3310–3318
19. Lee TS et al (1998) The role of the primary visual cortex in higher level vision. Vision Res 38:2429–2454
20. Lowe DG (1985) Perceptual organization and visual recognition. Kluwer, Boston
21. Maguire E, Frith CD, Cipolotti L (2001) Distinct neural systems for the encoding and recognition of topography and faces. NeuroImage 13(4):743–750
22. Maki A, Nordlund P, Eklundh J-O (2000) Attentional scene segmentation: integrating depth and motion from phase. Comput Vision Image Understand 78(3):351–373
23. Marr D (1982) Vision. Freeman, Cambridge, MA
24. Nakamura K et al (2000) Functional delineation of the human occipito-temporal areas related to face and scene processing: a PET study. Brain 123:1903–1912
25. Oliva A, Schyns PG (1997) Coarse blobs or fine edges? Evidence that information diagnoticity changes the perception of complex visual stimuli. Cogn Psychol 34:72–107
26. O'Toole AJ et al (1994) Structural aspects of face recognition and the other race effect. Mem Cogn 22:208–224
27. Palmer SE (1999) Vision science: photons to phenomenology. MIT Press, Cambridge, MA, p 810
28. Park S-J, Shin J-K, Lee M (2002) Biologically inspired saliency map model for bottom-up visual attention. In: Proceedings of the international workshop on biologically motivated computer vision. Springer, Berlin Heidelberg New York
29. Parkhurst D, Law K, Neibur E (2002) Modeling the role of salience in the allocation of overt visual attention. Vision Res 42(1):107–123
30. Pollen DA, Gaska JP, Jacobson LD (1989) Physiological constraints on models of visual cortical function. In: Rodney M, Cotterill J (eds) Models of brain functions. Cambridge University Press, New York, pp 115–135
31. Puce A et al (1995) Face-sensitive regions in human extrastriate cortex studied by functional MRI. J Neurophysiol 74:1192–1199
32. Tanaka JW, Curran T (2001) A neural basis for expert object recognition. Psychol Sci 12(1):43–47
33. Tarr MJ, Gauthier I (2000) FFA: a flexible fusiform area for subordinate-level visual processing automatized by expertise. Neuroscience 3(8):764–769
34. Tipping ME, Bishop CM (1999) Mixtures of probabilistic principal component analysers. Neural Comput 11(2):443–482
35. Tong F et al (2000) Response properties of the human fusiform face area. Cogn Neuropsychol 17(1/2/3):257–279
36. Tsotsos JK et al (1995) Modeling visual attention via selective tuning. Artif Intell 78(1–2):507–545
37. Ungeleider LG, Mishkin M (1982) Two cortical visual systems. In: Ingle DJ, Goodale MA, Mansfield RJW (eds) Analysis of visual behavior. MIT Press, Cambridge, MA, pp 549–586
38. Zipser K, Lamme VAF, Schiller PH Contextual modulation in primary visual cortex. Neuroscience 16(22):7376–7389