A Mixture of Experts Model Exhibiting Prosopagnosia

Matthew N. Dailey, Garrison W. Cottrell, and Curtis Padgett
Computer Science & Engineering 0114
University of California, San Diego
La Jolla, CA 92093
{mdailey,gary,cpadgett}@cs.ucsd.edu

Abstract

A considerable body of evidence from prosopagnosia, a deficit in face recognition dissociable from nonface object recognition, indicates that the visual system devotes a specialized functional area to mechanisms appropriate for face processing. We present a modular neural network composed of two "expert" networks and one mediating "gate" network with the task of learning to recognize the faces of 12 individuals and classifying 36 nonface objects as members of one of three classes. While learning the task, the network tends to divide labor between the two expert modules, with one expert specializing in face processing and the other specializing in nonface object processing. After training, we observe the network's performance on a test set as one of the experts is progressively damaged. The results roughly agree with data reported for prosopagnosic patients: as damage to the "face" expert increases, the network's face recognition performance decreases dramatically while its object classification performance drops slowly. We conclude that data-driven competitive learning between two unbiased functional units can give rise to localized face processing, and that selective damage in such a system could underlie prosopagnosia.

Introduction

For years, researchers attempting to deduce the functional architecture of the visual system have debated whether face recognition occurs in a specialized "module" not used for recognition of nonface objects. A considerable body of evidence from prosopagnosia seems to indicate that faces are processed by a more or less independent system. Prosopagnosia is a rare condition in which brain damage reduces a person's ability to recognize faces. Although the condition is almost always accompanied by other visual impairments, the deficit can be remarkably specific to faces.

One possible explanation is that face recognition is in some way more difficult than other types of recognition, so mild damage to a general-purpose recognition system could affect face recognition more than nonface object recognition (Damasio, Damasio, & Van Hoesen, 1982; Humphreys & Riddoch, 1987). However, recent experiments showing a double dissociation between face and nonface object recognition provide evidence that some separable mechanism serves face recognition better than object recognition and vice versa.

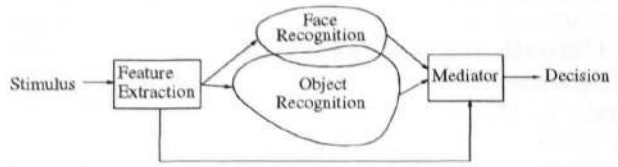
McNeil and Warrington (1993) report that W.J., a patient with severe prosopagnosia but apparently normal recognition of famous buildings, dog breeds, car makes, and flower species, had acquired a flock of sheep and learned to recognize the individuals from their markings. In a test with unfamiliar sheep of a breed unfamiliar to W.J., a control group performed significantly better on recognition of human faces than of the sheep faces, indicating the advantages humans normally have in identifying human faces. But W.J. performed significantly better on the sheep face task than on the human face task. The unfamiliar sheep face recognition task was in many ways as difficult in terms of complexity and confusability as face recognition, yet W.J. performed well.

Martha Farah and her colleagues have performed two important experiments providing further evidence of a specialization for face processing. In the first, they constructed a within-class discrimination task involving faces and visually similar eyeglasses (Farah, Levinson, & Klein, 1995a). Normal subjects were significantly better at discriminating the faces than the eyeglasses, but the prosopagnosic patient L.H. did not show this effect. His face discrimination performance was significantly lower than that of the control group, but his eyeglass discrimination performance was comparable to that of the controls. In the other experiment, the researchers compared L.H.'s performance in recognizing inverted faces to that of normals (Farah, Wilson, Drain, & Tanaka, 1995b). The surprising result was that whereas normal subjects were significantly better at recognizing upright faces than inverted ones, L.H. performed normally on the inverted faces but was actually worse at recognizing the upright faces than the inverted ones. This study indicates that normal face recognition not only utilizes some form of specialized processing, but also that the face processing pathway is mandatory, even after damage.

On the other hand, studies of several patients have shown that visual object recognition can be impaired while face recognition is spared. Feinberg et al. (1994), on the basis of neuroanatomical assessment of such patients, argue that complex object recognition largely depends on visual decomposition into parts, for which the left hemisphere is superior, whereas face stimuli do not require such decomposition. This double dissociation between face and object recognition provides a strong argument that the visual system contains elements specialized for face processing, although it does not necessarily imply a distinct face "module" (Plaut, 1995).

In light of the double dissociation, we propose a simple

visual system model:



Recognition of face-like stimuli and non-face-like stimuli is accomplished by specialized but possibly overlapping mechanisms. A mediator, on the basis of a representation of the stimulus itself, mixes the output of the two systems to generate a final decision on the identity or class of the stimulus.

For the current study, we implemented the model with the mixture of experts neural network architecture (Jordan & Jacobs, 1995), trained the network to perform a combined face identification and object classification task, and found that competitive learning between two identical "expert" modules can result in a division of labor in which one expert dominates in face pattern processing and the other dominates in nonface object pattern processing. Furthermore, damage to the "face" expert disproportionately ablates the model's face recognition performance, indicating that data-driven specialization of separate processors and the fact that faces require fine within-class discrimination might play an important role in the type of dissociation observed in prosopagnosia. After describing the experiment and its results, we discuss the possible implications of these findings and directions for further research.

Experimental Methods

Face and Object Data

This study utilized static images of 12 individuals' faces, 12 different cups, 12 different books, and 12 different soda cans. See Figure 1 for examples from each class.

For the faces, we collected 5 images of each of 12 individuals from the Cottrell and Metcalfe database (1991). In these images, the subjects attempt to display various emotions, while the lighting and camera viewpoint is held constant. We then captured 5 images of each of the 36 objects with a CCD camera and video frame grabber. For these images, we performed minor, pseudorandom perturbations of each object's position and orientation while lighting and camera viewpoint remained constant. After capturing the 640x480 grayscale images, we cropped and scaled them to 64x64, the same size as the face images.

Image Preprocessing

In order to transform raw 64x64 8-bit grayscale images into a representation more appropriate for a neural network classifier, we preprocessed the images with a Gabor wavelet-based feature detector and principal components analysis (PCA). These preprocessing steps qualitatively resemble some of the preprocessing done in early stages of the visual system.

The Gabor Jet Feature Detector We first transformed the input image set by extracting Gabor "jet" features. The wavelet resembles a sinusoid restricted by a Gaussian function, may be tuned to a particular orientation and frequency, and is similar to the observed receptive fields of simple cells in primary visual cortex (Jones & Palmer, 1987). A "jet" is formed by combining the response of several filters with different orientations. As an image feature detector, the jet exhibits some invariance to background, translation, distortion, and size (Buhmann, Lades, & von der Malsburg, 1990).

The basic wavelet is:

$$G(\vec{k},\vec{x}) = \exp(i\vec{k}\cdot\vec{x})\exp\left(-\frac{k^2\vec{x}\cdot\vec{x}}{2\sigma^2}\right),$$

where

$$\vec{k} = k(\cos\phi, sin\phi)$$

and $k \equiv |\vec{k}|$ controls the wavelength or "scale" of the filter function G, $\vec{x}$ is a point in the plane relative the wavelet's origin, ϕ is the angular orientation of the filter, and σ is a constant. As in Buhmann et al. (1990), we let $\sigma = \pi$, let ϕ range over $\{0, \frac{\pi}{8}, \frac{\pi}{4}, \frac{3\pi}{8}, \frac{\pi}{2}, \frac{5\pi}{8}, \frac{3\pi}{4}, \frac{7\pi}{8}\}$, and we let

$$k_i = \frac{2\pi}{N}2^i, \text{ with } i \in \{1,...,6\}.$$

Since the input image size is 64x64, $N = 64$.

Again as in Buhmann et al. (1990), for each of the eight orientations and six wavelengths, we convolve $G(\vec{k},\vec{x})$ with the input image $I(\vec{x})$:

$$(\mathcal{WI})(\vec{k},\vec{x_0}) = \int G_{\vec{k}}(\vec{x_0}-\vec{x})I(\vec{x})d^2x$$

then normalize the response values across scales:

$$(\mathcal{TI})(\vec{k},\vec{x_0}) = \frac{|(\mathcal{WI})(\vec{k},\vec{x_0})|}{\int|(\mathcal{WI})(\vec{k},\vec{x})|d^2xd\phi}.$$

With eight orientations and six scale factors, the process results in a vector of 48 complex values at each point of an image (see Figure 2 for an example). We subsampled an 8x8 grid of these vectors and computed the magnitude of the complex values to get a 3072-element vector representing the image.

Dimensionality Reduction with Principal Components Analysis The feature extraction method described above produced 240 input patterns of 3072 elements. To reduce the dimensionality of the input patterns, we first divided them into a training set composed of four examples for each individual face or object (192 patterns total) and a test set composed of one example of each individual (48 patterns total). Using the technique described by Turk and Pentland (1991), we projected each pattern onto the basis formed by the 192 most-significant eigenvectors of the training set's covariance matrix, resulting in 192 coefficients for each pattern. As a final step, we normalized each pattern by dividing each of

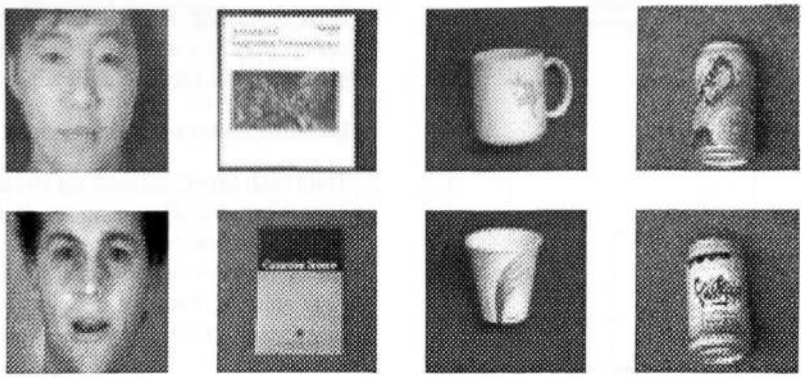


Figure 1: Example face, cup, book, and can images

Figure 2: Original image and Gabor jets at six scales. Each pixel's intensity in the processed images represents the log of the sum of the magnitudes of the filter responses in each of the eight directions.

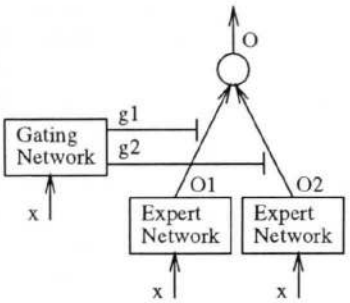


Figure 3: Modular mixture of experts network

its coefficients by its maximum coefficient magnitude so all coefficients fell in the range $[-1, 1]$.

With the resulting representation, our networks exhibited good training set accuracy and adequate generalization, so we did not further reduce the pattern dimensionality or normalize the variance of the coefficients. Note that with 192 patterns and 192 dimensions, the training set is almost certainly linearly separable.

Mixture of Experts Network Architecture

We modeled the face and object recognition task with the "mixture of experts" architecture (Jordan & Jacobs, 1995).

Figure 3 shows a simple modular network. Each expert network i is a single-layer linear network that computes an output vector $\mathbf{O}_i$ as a function of the input vector $\mathbf{x}$ and a set of parameters θ_i.

We assume that each expert specializes in a different area of the input space. The gating network assigns a weight g_i to each of the experts' outputs $\mathbf{O}_i$. The gating network determines the g_i as a function of the input vector $\mathbf{x}$ and a set of parameters $\mathbf{w}$. The g_i can be interpreted as estimates of the prior probability that expert i can generate the desired output $\mathbf{y}$, or $P(i|\mathbf{x}, \mathbf{w})$. The gating network is a single-layer linear network with a softmax nonlinearity at its output. That is, the linear network computes

$$\xi_i = \sum_j x_j w_{ij}$$

then applies the softmax function to get

$$g_i = \frac{\exp(\xi_i)}{\sum_j \exp(\xi_j)}$$

Clearly, then, the g_i are nonnegative and sum to 1. The final, mixed output of the network is

$$\mathbf{O} = \sum_i g_i \mathbf{o}_i.$$

Adaptation by Maximum Likelihood Gradient Ascent For adapting the network's estimates of the parameters $\mathbf{w}$ and θ_i, we used the gradient ascent algorithm for maximizing the log likelihood described by Jordan & Jacobs. Assuming the probability density associated with each expert is Gaussian with identity covariance matrix, they obtain the online learning algorithms

$$\Delta\theta_i = \eta_e h_i (\mathbf{y} - \mathbf{o}_i)\mathbf{x}^T$$

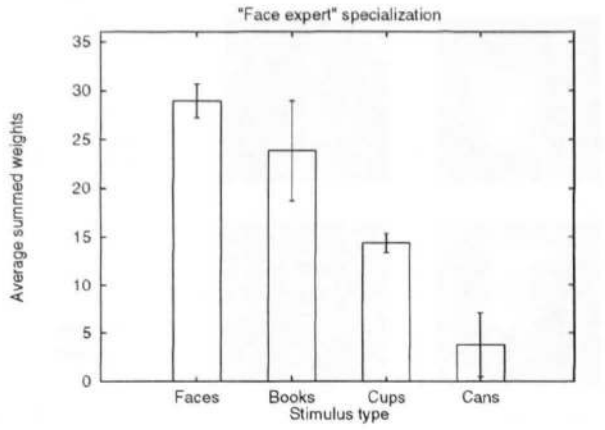


Figure 4: Weights assigned to the face-dominant expert network for each stimulus class. Error bars denote standard error.

and

$$\Delta \mathbf{w}_i = \eta_g (h_i - g_i)\mathbf{x}^T$$

where η_e and η_g are learning rates for the expert networks and the gating network, respectively, and h_i is an estimate of the posterior probability that expert i can generate the desired output $\mathbf{y}$:

$$h_i = \frac{g_i \exp\left(-\frac{1}{2}(\mathbf{y} - \mathbf{o}_i)^T(\mathbf{y} - \mathbf{o}_i)\right)}{\sum_j g_j \exp\left(-\frac{1}{2}(\mathbf{y} - \mathbf{o}_j)^T(\mathbf{y} - \mathbf{o}_j)\right)}$$

This is essentially a softmax function computed on the inverse of the sum squared error of each expert's output, smoothed by the gating network's current estimate of the prior probability that the input pattern was drawn from expert i's area of specialization.

As the network learns, the expert networks "compete" for each input pattern, while the gate network rewards the winner of each competition with stronger error feedback signals. Thus, over time, the gate partitions the input space in response to the experts' performance. We found that adding momentum terms to the update rules enabled the network to learn more quickly and the gate network to partition the input space more reliably. With this change, if c is a weight change computed as above, the update rule for an individual weight becomes $\Delta w_i(t) = c + \alpha \Delta w_i(t-1)$. The next section describes how we chose the learning parameters η_g, η_e, α_g, and α_e during the training process.

In these experiments, the network's task was to recognize the faces as individuals and the objects as members of their class. Thus the network had 15 outputs, corresponding to cup, book, can, face 1, face 2, etc. For example, the desired output vector for the cup patterns and the face 5 patterns were $[1,0,0,0,0,0,0,0,0,0,0,0,0,0,0]^T$ and $[0,0,0,0,0,0,0,1,0,0,0,0,0,0,0]^T$, respectively.

Network Training

After removing one example of each face and object (48 patterns) from the training set for use as a validation set to stop training, we used the following training procedure:

1. Initialize network weights to small random values.
2. Train each expert network on 10 randomly-chosen patterns from the (reduced) training set. Without this step, both networks would perform equally well on every pattern and the gating network would not learn to differentiate between their abilities, because the gate weight update rule is insensitive to small differences between the experts' performance.
3. Repeat 10 times:
   (a) Randomize the training set's presentation order.
   (b) Train the network for one epoch.
4. Test the network's performance on the validation set.
5. If mean squared error over the validation set has not increased two consecutive times, go to 3.
6. Test the network's performance on the test set.

The training regimen was sufficient to achieve near-perfect performance on the test set (see Figure 5 results for 0% damage), but we found that the *a priori* estimates (g_1 and g_2) learned by the gate network were extremely sensitive to the learning parameters η_g, η_e, α_g, and α_e. If the gate network learns too slowly relative to the experts, they generally receive the same amount of error feedback and the g_i never deviate far from 0.5. If the gate network learns too quickly relative to the experts, it tends to assign all of the input patterns to one of the experts. To address this problem, we performed a search for parameter settings that partition the training set effectively. For 270 points in the four-dimensional parameter space, we computed the variance of one of the gate network outputs over the training set, averaged over ten runs. This variance measure was maximal when $\eta_e = 0.05, \eta_g = 0.15, \alpha_e = 0.4$, and $\alpha_g = 0.6$.

Maximizing the gate output variance is a reasonable strategy for selecting the model's learning parameters. It encourages a fairly sharp partition between the experts' areas of specialization without favoring one partition over another. On the other hand, it would have been preferable to include a term penalizing low gate value variance in the network's objective function, since this would eliminate the need for a parameter search.

Results

Figure 4 summarizes the division of labor performed by the gate network over 10 runs with $\eta_e = 0.05, \eta_g = 0.15, \alpha_e = 0.4$, and $\alpha_g = 0.6$. The bars denote the weights the gate network assigned to whichever expert emerged as face-dominant, broken down by stimulus class, and the error bars denote standard error. Figure 5 illustrates the performance

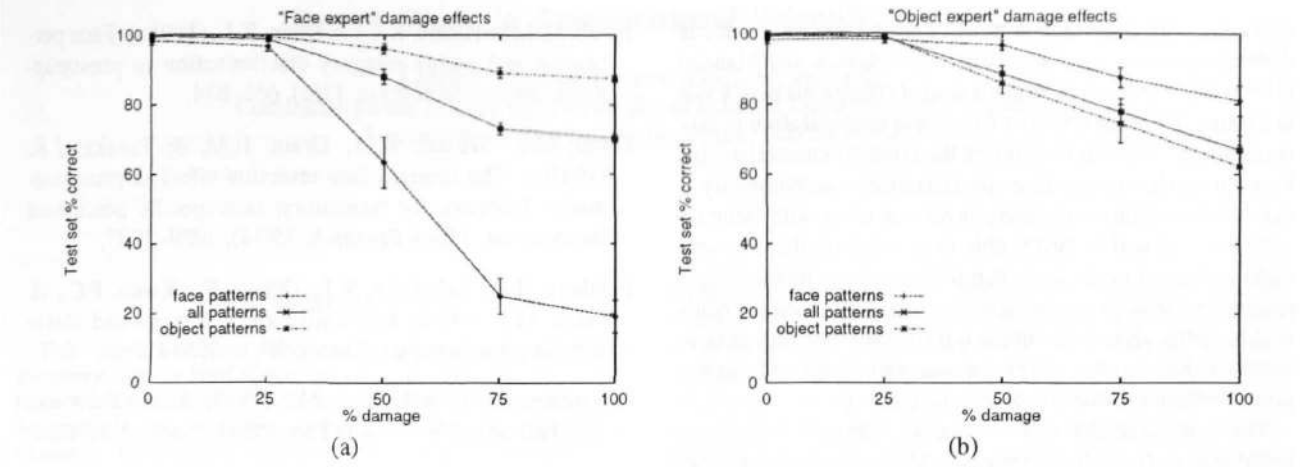


Figure 5: (a) Face identification classification errors increase as damage to the face-dominating expert module increases. (b) Object categorization classification errors increase as damage to the non-face-dominating expert module increases.

effects of damaging one expert by randomly removing connections between its input and output units. Damaging the face-specializing network resulted in a dramatic decrease in performance on the face patterns. When the network not specializing in faces was damaged, however, the opposite effect was present but less severe. Clearly, the face specialist learned enough about the object classes during early stages of training (when the gating network estimates all prior probabilities at about 0.5) to correctly classify some of the object patterns.

Discussion

The simulation results show that the network is a good model of the prosopagnosic effect: as damage to the "face" module increases, the network's ability to recognize faces decreases dramatically. From this we conclude that it is possible for competitive learning between two unbiased functional units to give rise to a specialized face processor. Since faces form a fairly homogeneous class, it is reasonable to expect that a unit good at identifying one face will also be good at processing others. However, since the degree of separation between face and nonface patterns in the model is not clean and is sensitive to training parameters, additional constraints would be necessary to achieve a face/nonface division reliably. Indeed, such constraints, such as the prevalence of face stimuli in the newborn's environment, different maturation rates in different areas of the brain, and a possibly innate preference for tracking faces, may well be at work during infant development (Johnson & Morton, 1991).

Despite the lack of a strong face/nonface separation in the network, damaging the "face expert" affects face recognition accuracy disproportionately, compared with how damage to the nonface expert affects object recognition accuracy. Although we have not yet run the appropriate control experiments, we hypothesize that requiring the network to perform fine discrimination between members of a homogeneous class and gross classification of the other classes leads to the difference in damage effects.

Although we were not directly attempting to model visual object agnosia, it is interesting to consider how object classification performance degrades in Figure 5 (b). Even with 100% damage to the "object expert," the "face expert" alone is able to correctly classify 62% of the object patterns in the test set, compared with the object expert correctly classifying 19% of the face patterns when the face expert no longer contributes. This effect is most likely due to the fact that the face expert receives some error feedback on the object patterns early in training, when the gate network's prior probability estimates are close to 0.5 for all patterns, and that the classification task is relatively simple, involving only three classes. The face expert's performance would most likely decrease markedly on objects if the object classification task was more realistic, involving more classes or within-class discrimination. But in an interesting way, these results concur with the neuropsychological data on prosopagnosia. On the basis of a review of the literature on agnosia, Farah (1991) observes that visual object agnosia without prosopagnosia nearly always coincides with alexia (an inability to recognize words), and concludes that face recognition depends strongly on processing complex objects as a whole, word recognition depends strongly on breaking complex objects into parts, and nonface object recognition depends more on a mixture of the two mechanisms. Thus selective damage to a "part decomposition" mechanism can affect the processing of some object types more than others. For our model, this hypothesis predicts that objects amenable to processing as wholes will be more easily recognized by a face specialist that objects in which discrimination requires decomposition into parts.

Future Work

Building on this experiment, we will investigate several avenues of further research. The network's behavior indicates

that competitive learning is most likely not the sole factor in development of a face specialist. de Schonen and Mancini (1995) argue that innate organizational constraints play a role in biasing the brain toward a functional specialization in face recognition. We will investigate the types of constraints that bias our model toward face specialization; one possibility is that a low-resolution pathway involving units with large receptive fields will be better able to accomplish the discrimination required in the face recognition task, whereas a high-resolution pathway involving units with small receptive fields will be better able to accomplish the object recognition task. Jacobs and Kosslyn (1994) have successfully applied this approach within the mixture of experts paradigm.

The hypothesis that face processing primarily depends on holistic or configural information, whereas processing other object types depends more on analyzing an object's subparts, is also testable in our model. We plan to explore the hypothesis by constructing more realistic object recognition tasks using a broader variety of objects and involving both within-class discrimination and gross classification of objects (such as words) requiring some level of "parts analysis" for recognition. We predict that these changes to the object recognition task will cause a clearer dissociation of object recognition from face recognition when damaging so-called object experts.

Work on "covert" face recognition in prosopagnosics measuring skin conductance during face recognition tasks (e.g. Tranel & Damasio, 1988) and evidence for the mandatory nature of the face processing system (Farah et al., 1995b) seem to argue that the process of mediating between the face and nonface object systems actually occurs *before* recognition. We plan to investigate ways to account for this data in future models.

References

Buhmann, J., Lades, M., & von der Malsburg, C. (1990). Size and distortion invariant object recognition by hierarchical graph matching. In *IJCNN International Joint Conference on Neural Networks* vol. 2 (pp. 411–416).

Cottrell, G.W., & Metcalfe, J. (1991). Empath: Face, gender, and emotion recognition using holons. In *Advances in Neural Information Processing Systems 3*, (pp. 564–571).

Damasio, A.R., Damasio, H., & Van Hoesen, G.W. (1982). Prosopagnosia: Anatomic basis and behavioral mechanisms. *Neurology*, 32, 331–341.

de Schonen, S., & Mancini, J. (1995). About functional brain specialization: The development of face recognition. Developmental Cognitive Neuroscience Technical Report 95.1, MRC Cognitive Development Unit, London, UK.

Farah, M.J. (1991). Patterns of co-occurrence among the associative agnosias: Implications for visual object representation. *Cognitive Neuropsychology*, 8, 1–19.

Farah, M.J., Levinson, K.L., & Klein, K.L. (1995a). Face perception and within-category discrimination in prosopagnosia. *Neuropsychologia*, 33(6), 661–674.

Farah, M.J., Wilson, K.D., Drain, H.M, & Tanaka, J.R. (1995b). The inverted face inversion effect in prosopagnosia: Evidence for mandatory, face-specific perceptual mechanisms. *Vision Research*, 35(14), 2089–2093.

Feinberg, T.E., Schindler, R.J., Ochoa, E., Kwan, P.C., & Farah, M.J. (1994). Associative visual agnosia and alexia without prosopagnosia. *Cortex*, 30(3), 395–412.

Humphreys, G.W., & Riddoch, M.J. (1987). *To See But Not to See*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Jacobs, R.A., & Kosslyn, S.M. (1994). Encoding shape and spatial relations: The role of receptive field size in coordinating complementary representations. *Cognitive Science*, 18(3), 361–386.

Johnson, M.H., & Morton, J. (1991). *Biology and Cognitive Development: The Case of Face Recognition*. Oxford, UK: Basil Blackwell Ltd.

Jones, J.P., & Palmer, L.A. (1987). An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. *Journal of Neurophysiology*, 58(6), 1233–1258.

Jordan, M.I., & Jacobs, R.A. (1995). Modular and hierarchical learning systems. In Arbib, M.A. (Ed.), *The Handbook of Brain Theory and Neural Networks*. Cambridge, Massachusetts: MIT Press.

McNeil, J.E., & Warrington, E.K. (1993). Prosopagnosia: A face-specific disorder. *The Quarterly Journal of Experimental Psychology*, 46A(1), 1–10.

Plaut, D.C. (1995). Double dissociation without modularity: Evidence from connectionist neuropsychology. *Journal of Clinical and Experimental Neuropsychology*, 17(2), 294–321.

Tranel, D., & Damasio, A.R. (1988). Non-conscious face recognition in patients with face agnosia. *Behavioural Brain Research*, 30, 235–249.

Turk, M., & Pentland, A. (1991). Eigenfaces for recognition. *The Journal of Cognitive Neuroscience*, 3, 71–86.