PERGAMON

Neural Networks 12 (1999) 1053–1073

Neural Networks

www.elsevier.com/locate/neunet

Organization of face and object recognition in modular neural network models

M.N. Dailey*, G.W. Cottrell[1]

Department of Computer Science and Engineering, University of California, San Diego, CA, USA

Received 3 December 1998; accepted 7 May 1999

Abstract

There is strong evidence that face processing in the brain is localized. The double dissociation between prosopagnosia, a face recognition deficit occurring after brain damage, and visual object agnosia, difficulty recognizing other kinds of complex objects, indicates that face and non-face object recognition may be served by partially independent neural mechanisms. In this paper, we use computational models to show how the face processing specialization apparently underlying prosopagnosia and visual object agnosia could be attributed to (1) a relatively simple competitive selection mechanism that, during development, devotes neural resources to the tasks they are best at performing, (2) the developing infant's need to perform subordinate classification (identification) of faces early on, and (3) the infant's low visual acuity at birth. Inspired by de Schonen, Mancini and Liegeois' arguments (1998) [de Schonen, S., Mancini, J., Liegeois, F. (1998). About functional cortical specialization: the development of face recognition. In: F. Simon & G. Butterworth, *The development of sensory, motor, and cognitive capacities in early infancy* (pp. 103–116). Hove, UK: Psychology Press] that factors like these could bias the visual system to develop a processing subsystem particularly useful for face recognition, and Jacobs and Kosslyn's experiments (1994) [Jacobs, R. A., & Kosslyn, S. M. (1994). Encoding shape and spatial relations—the role of receptive field size in coordination complementary representations. *Cognitive Science*, *18*(3), 361–368] in the mixtures of experts (ME) modeling paradigm, we provide a preliminary computational demonstration of how this theory accounts for the double dissociation between face and object processing. We present two feed-forward computational models of visual processing. In both models, the selection mechanism is a gating network that mediates a competition between modules attempting to classify input stimuli. In Model I, when the modules are simple unbiased classifiers, the competition is sufficient to achieve enough of a specialization that damaging one module impairs the model's face recognition more than its object recognition, and damaging the other module impairs the model's object recognition more than its face recognition. However, the model is not completely satisfactory because it requires a search of parameter space. With Model II, we explore biases that lead to more consistent specialization. We bias the modules by providing one with low spatial frequency information and the other with high spatial frequency information. In this case, when the model's task is subordinate classification of faces and superordinate classification of objects, the low spatial frequency network shows an even stronger specialization for faces. No other combination of tasks and inputs shows this strong specialization. We take these results as support for the idea that something resembling a face processing "module" could arise as a natural consequence of the infant's developmental environment without being innately specified. © 1999 Published by Elsevier Science Ltd. All rights reserved.

Keywords: Face recognition; Object recognition; Visual system development; Modular networks

1. Introduction

Two complementary deficits, prosopagnosia (impaired recognition of faces) and visual object agnosia (impaired recognition of common objects), together form a classic neuropsychological double dissociation. This might be taken as evidence for a domain-specific face processing mechanism in the brain that is distinct from the mechanisms serving general object recognition. However, two issues have led to a long-running debate on this view: (1) it is not entirely clear how specific or independent prosopagnosia and visual object agnosia are, and (2) double dissociations do not necessarily implicate separate, domain-specific mechanisms. In this section, we first briefly review the data on prosopagnosia and visual object agnosia; these data support the view that the mechanisms underlying facial identity recognition are at least somewhat different from those underlying most other object recognition tasks. We then review the theories attempting to explain the seemingly

* Corresponding author. UCSD Computer Science and Engineering, 9500 Gilman Dr, La Jolla, CA 92093-0114, USA. Tel.: + 1-619-453-4364; fax: + 1-619-534-7029.

E-mail addresses: mdailey@cs.ucsd.edu (M.N. Dailey), gary@cs.ucsd.edu (G.W. Cottrell)

[1] Tel.: + 1-619-534-6640; fax: + 1-619-534-7029.

PII: S0893-6080(99)00050-7

remarkable dissociation and motivate the current computational modeling studies.

1.1. Is prosopagnosia really specific to faces?

Prosopagnosia is almost always accompanied by other visual impairments, so it is difficult to determine the extent to which a prosopagnosic's deficit is limited to face processing. We limit discussion here to so-called "associative" prosopagnosics who may have normal ability in face detection tasks (is it a face?) but whose primary deficit is that they cannot recognize the *identity* of familiar faces (De Renzi, 1986). This condition is usually associated with either unilateral right hemisphere or bilateral lesions in the fusiform gyrus area. For reviews that include lesion locations, see De Renzi, Perani, Carlesimo, Silveri and Fazio, (1994); Farah (1990).

Although many prosopagnosics have difficulty performing difficult subordinate (within-class) classification tasks with objects other than faces, in some cases, the condition can be remarkably face-specific. De Renzi's (1986) "case 4" was profoundly prosopagnosic but claimed to have no trouble with day-to-day within-class discrimination of common objects such as keys and automobiles. However, there have been objections that perhaps this patient was not tested extensively enough to determine whether his deficit was truly face specific.

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 that face processing can be impaired with little impact on within-category discrimination of objects. 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. We must be cautious in interpreting these results, however. de Gelder, Bachoud-Levi and Degos (1998) report on an agnosic patient with a similar inversion effect for both faces *and shoes*, challenging the idea of an "inverted inversion effect" for faces as a face-specific phenomenon. Also, prosopagnosic patients may show no object processing deficit when the performance measure is classification accuracy, but other measures, such as response time and sensitivity, may actually reveal impairments in their performance (Gauthier, Behrmann & Tarr, 1999).

Thus with important caveats, it appears that brain damage can disproportionately impair processing of normal, upright faces. On the other hand, studies of several patients have shown that visual object recognition can be severely impaired while face recognition is spared. Associative visual agnosia sparing face recognition is normally associated with left occipital or bilateral occipital lesions and usually coincides with alexia, in which patients have difficulty reading because they cannot rapidly piece letters into words (Farah, 1990). It seems to reflect an impairment in decomposing complex objects into parts (Feinberg, Schindler, Ochoa, Kean & Farah, 1994). Although it is difficult to assess exactly what is impaired and what is preserved (researchers obviously cannot test patients on *all* objects), Farah's (1990) review cites many such cases. Perhaps the most dramatically impaired and well-known visual agnosic without prosopagnosia is C.K. (Behrmann, Moscovitch & Winocur, 1994). This patient has a striking deficit in part integration; he can identify the component parts of objects but cannot put them together to recognize the whole. His face processing abilities, however, are largely spared, to the point that he can see faces in "composite" images where a face is composed of or hidden amongst other objects, but cannot see the objects themselves. Moscovitch, Wincour & Behrmann (1997) show in a series of experiments that C.K.'s ability to recognize (upright) famous faces, family resemblances, caricatures, and cartoons is completely normal, as is his ability to match unfamiliar faces. On the other hand, he is impaired at tasks involving inverted faces, which presumably activate his damaged "object processing" mechanisms.

These complementary patterns of brain damage constitute a double dissociation between face and object recognition and provide evidence that the visual system contains elements specialized for (or merely very useful for) face processing. However, double dissociations certainly do not imply that two tasks are served by entirely separate and distinct "modules". As Plaut (1995) points out, two seemingly independent tasks might not be independent at all, but simply rely more heavily or less heavily on particular mechanisms. In the worst case, the apparent distinction between face and object processing could simply reflect the expected outliers in a random distribution of patterns of brain damage (Juola & Plunkett, 1998). However, there are independent reasons, other than the patterns of brain damage, to believe that prosopagnosia reflects damage to a system that is specialized for face processing (and possibly certain other types of stimuli): we next review the behavioral distinctions between face processing and general object processing.

1.2. How might face recognition differ from general object recognition?

Given that the neuropsychological data indicate there may be something special about faces, one issue of debate is whether there is an innate, mandatory, domain-specific module (Fodor, 1983) for face processing. Moscovitch et al. (1997), for instance, give a convincing argument for modularity based on their experiments with C.K., the object agnosic. At the same time, many other researchers have attempted to find a more parsimonious explanation for the face/object double dissociation that places face recognition at some extreme end of a continuum of mechanisms.

There are many ways in which prosopagnosia could reflect damage to a general-purpose object recognition system yet appear to be face specific. One early explanation was that face recognition is simply more difficult than other types of recognition, so mild damage to a general-purpose recognition system could affect face recognition more than non-face object recognition (Damasio, Damasio & Van Hoesen, 1982; Humphreys & Riddoch, 1987). Although most if not all prosopagnosic patients also have some level of impairment at subordinate-level classification, a strict interpretation of this hypothesis is ruled out by the fact that many visual object agnosic patients have impaired recognition of common objects but spared face recognition (see the previous section).

Currently, there are at least two related classes of plausible theories attempting to characterize the differences between face and object processing. The first posits that faces are perceived and represented in memory "holistically" with little or no decomposition into parts, whereas most other object recognition tasks require part-based representations. Farah, Wilson, Drain and Tanaka (1998) review the literature on this topic and provide new evidence for holistic perceptual representations. Biederman and Kalocsai (1997) propose a computational basis for such representations. They show that the outputs of an array of overlapping local spatial filters similar to some of the receptive fields in visual cortex, as used in Wiskott, Fellous, Kruger & von der Malsburg's (1997) face recognition system, can account for human performance in experiments using face stimuli but cannot account for human performance in other object recognition tasks. Clearly, such simple representations would be holistic at least in the sense that there is no explicit encoding of the parts of the face independent of the whole face.

Theories in the second, related class suggest that the main reason for the seemingly special status of face recognition is that it involves expert-level subordinate classification within a relatively homogeneous object class. In this view, faces are only special in that they are very similar to each other, and we must acquire a great deal of sensitivity to configural differences between them. Tanaka and Sengco (1997) have shown that subtle configuration information, such as the distance between the eyes in a face, plays a crucial role in face processing but not in processing other objects types. But face processing is not necessarily the only task that engages this type of processing. It appears that the acquisition of expertise in subordinate classification of a novel synthetic object class, "Greebles", leads to a similar sensitivity to configuration information (Gauthier & Tarr, 1997). Gauthier, Tarr, Anderson, Skuklarski and Gore (1998) have also observed in fMRI studies that expert-level Greeble classification activates an area in fusiform gyrus thought by some to be specialized for faces (McCarthy, Puce, Gore & Allison, 1997; Sergent, Ohta & MacDonald, 1992).

Thus the main observable differences between face processing and general object processing (in the most common cases) involve holistic representations and our level of expertise with subordinate-level classification of faces. In this paper, we propose a theoretical model that explains how such specialized representations and mechanisms might develop, and describe preliminary computational modeling experiments that support the theory. The next section outlines some of the important data on the development of face recognition in infants, which we use to inform the construction of our computational models.

1.3. Developmental data and a possible low spatial frequency bias

In the previous sections, we have outlined evidence from neuropsychology and adult behavior that faces (and possibly other similar classes of stimuli) are processed by specialized mechanisms. Experiments exploring the development of face recognition abilities in human infants have also provided important clues to the organization of the putative face processing system and how that organization arises.

Experiments have shown that at birth, an infant's visual attention is spontaneously directed toward face-like stimuli (Johnson, Dziurawiec, Ellis & Morton, 1991). An infant can visually discriminate between his or her mother's face and a stranger's face, but only external features such as hairline and head contours are salient to the task at an early age (Pascalis, de Schonen, Morton, Deruelle & Fabre-Grenet, 1995). Later, around the age of 6–8 weeks, infants begin to use the face's internal features to discriminate their mothers from strangers (de Schonen & Mancini, 1998). A possibly related developmental factor is the fact that the newborn infant's acuity and contrast sensitivity are such that they can only detect large, high contrast stimuli; at one month of age, infants are typically insensitive to spatial

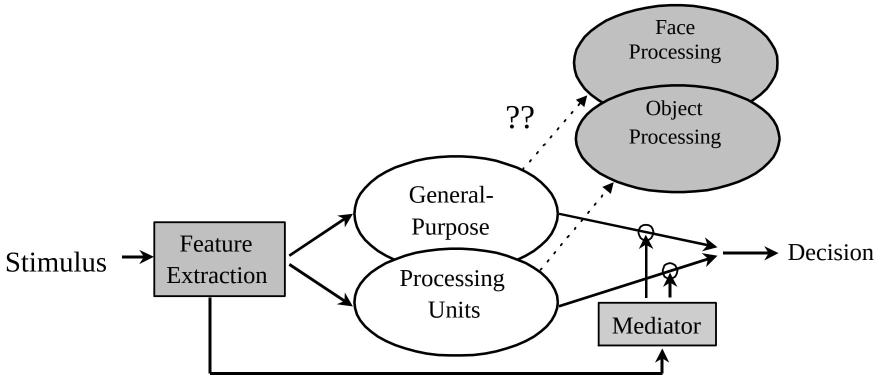


Fig. 1. Theoretical visual recognition model. We assume that recognition of face-like stimuli and non-face-like stimuli is accomplished by interdependent but possibly specialized processing units. Some mechanism, on the basis of a representation of a stimulus, mediates a competition between the units to generate a final decision on the identity or class of the stimulus. We explore the conditions under which one system specializes for face processing.

frequencies greater than 2 cycles/degree (Teller, McDonald, Preston, Sebris & Dobson, 1986).

It is not clear whether the infant's shift to the use of internal features for distinguishing his or her mother from strangers represents the use of an entirely new system or a gradual refinement of the old system (Johnson, 1997). But the experimental data on the development of face recognition capabilities make it seem likely that the infant visual system begins training a cortical "face processor" utilizing external facial features very early on. At the same time, these capabilities must develop on the basis of extremely low resolution stimuli. de Schonen and Mancini (1998) propose a scenario accounting for some of the known data. The scenario holds that several factors, including different rates of maturation in different areas of cortex, the infant's tendency to track faces, and the infant's initially low acuity, all conspire to force an early specialization for face recognition in right hemisphere. This specialized mechanism would necessarily be based on a "configurational" as opposed to a "componential" approach, due to the low resolution involved. Later in life, the adult visual system might then be biased to favor the same holistic subsystem when faced with tasks involving new stimuli having similar computational requirements.

de Schonen and Mancini's scenario resonates with some of the recent experimental data showing a low spatial frequency bias in adult face processing. Costen, Parker and Craw (1996) showed that although both high-pass and low-pass image filtering decrease face recognition accuracy, high-pass filtering degrades identification accuracy more quickly than low-pass filtering. Also, Schyns and Oliva (1999) have shown that learning the identity of a set of faces later biases subjects' perception toward low spatial frequency information. The bias is tested by briefly presenting a hybrid image combining a low-pass filtered image of one individual's face and a high-pass filtered image of another person's face and requiring them to categorize the face they perceive according to its gender, expressiveness, or type of expression. The identity learning stage shifted the subjects' perception toward the low-pass filtered face compared to subjects without prior exposure. These studies provide evidence that low spatial frequency information may be relatively more important for face identification than high spatial frequency information.

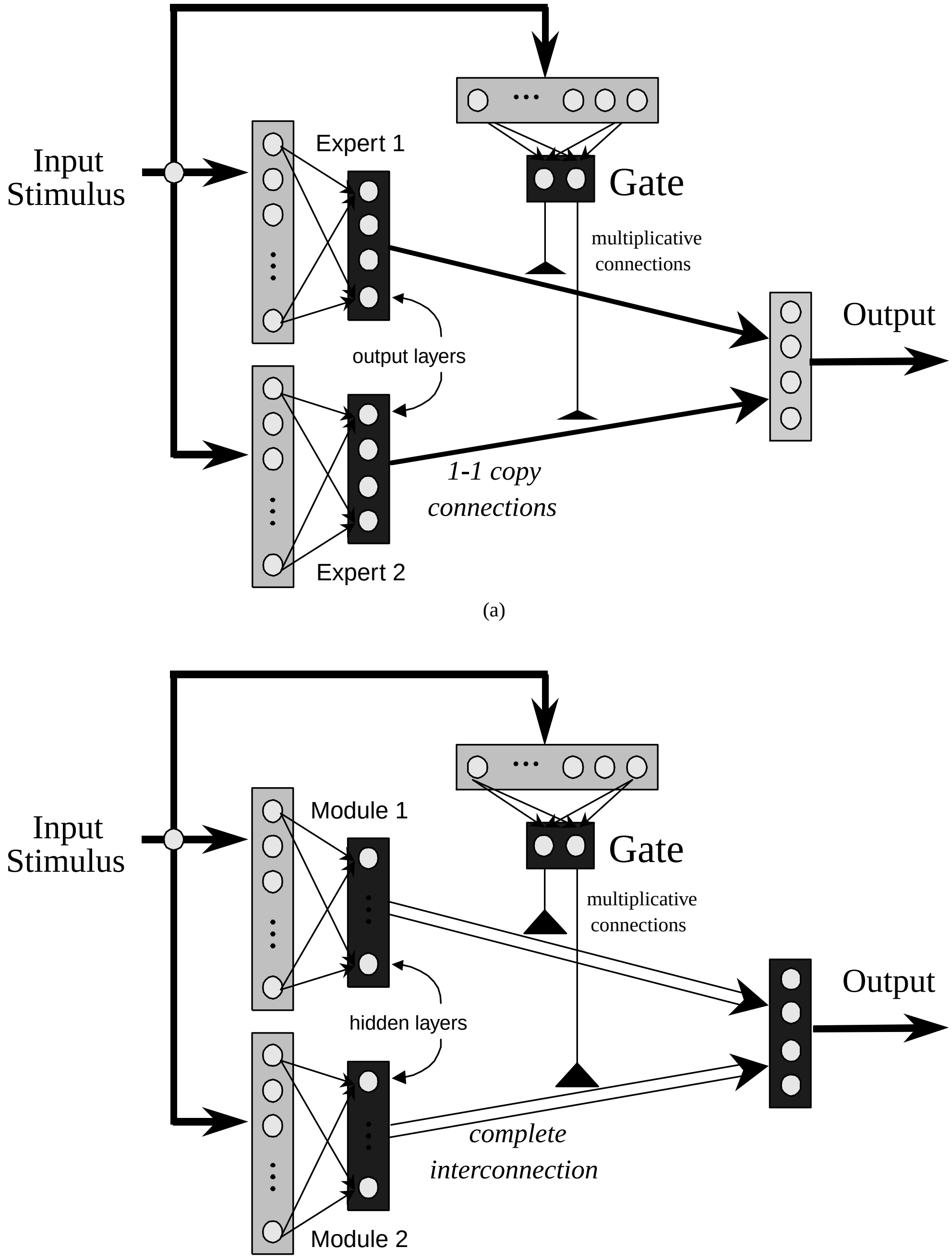
1.4. Outline

In a series of computational modeling studies, we have begun to provide a computational account of the face specialization data. We propose that a neural mechanism allocating resources according to their ability to perform a given task could begin to explain the apparent specialization for face recognition evidenced by prosopagnosia. We have found that a model based on the mixture of experts architecture, in which a gating network implements competitive selection between two simple homogeneous modules, can develop a specialization such that damage to one module disproportionately impairs face recognition compared to non-face object recognition.

We then consider how the availability of spatial frequency information and the task to be performed affects face recognition specialization given this hypothesis of neural resource allocation by competitive selection. We find that when high and low spatial frequency information is "split" between two modules in our system, and the task is to identify the faces while simply classifying the objects, the low-frequency module consistently specializes for face recognition. After describing the models in more detail, we present our experimental results, and discuss their implications.

2. The modeling paradigm

We have performed two computational modeling experiments designed to explore the ways in which a general-purpose learning mechanism might specialize for face recognition vs. object recognition, such that localized



(b)

Fig. 2. Modular network model architectures. (a) The standard mixture of experts (ME). See Appendix A for details. (b) The modular network for Model II, described in Appendix B. In this network, the gate mixes hidden layer representations rather than expert network output layers. In the ME network, the experts are self-contained linear networks with their own output layers. The entire network's output vector is a linear combination of the expert output vectors. In contrast, the second network's modules are not self-contained—each module's "output" is a standard network's hidden layer. Each of the overall network's output units is a function of *all* of the nonlinear hidden units in *both* of the modules, modulated by the gate network's outputs.

random "damage" to the model results in decreased face recognition performance or decreased object recognition performance. Both of the models are feed-forward neural networks with special competitive "modular" architectures that allow us to conveniently study the conditions under which specialization arises. In this section, we describe the computational models then describe how we acquired and preprocessed the object/face image data used in both experiments.

2.1. The theoretical model

Our basic theoretical model for face and object recognition is displayed in Fig. 1. We generally assume

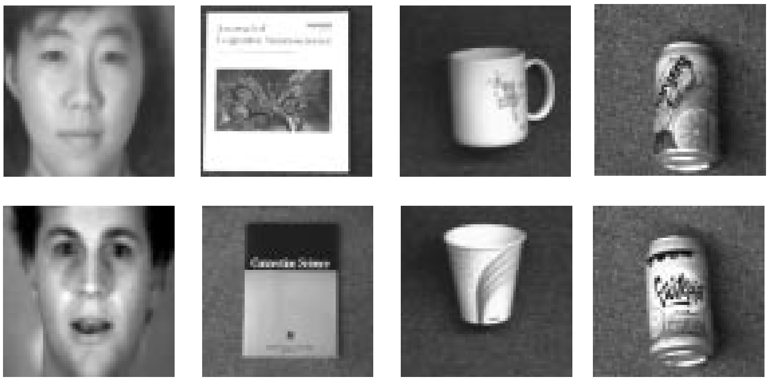


Fig. 3. Example face, book, cup, and can images.

that prosopagnosia (sparing object recognition) and visual object agnosia (sparing face recognition) are symptoms of damage to subsystems that are more or less specialized for face recognition or object recognition. We imagine an array of general-purpose "processing units" that compete to perform tasks and a "mediator" that selects processing units for tasks. This mediator could be intrinsic to the processing unit architecture itself, as in the self-organizing map (Kohonen, 1995) or a more external, explicit mechanism, as in the mixture of experts (ME) (Jordan & Jacobs, 1995). We instantiate this theoretical model with modular neural networks by presenting modular networks with various face/object classification tasks. We then study the conditions under which, through competition, one expert or module specializes for faces to the extent that "damaging" that model by removing connections results in a "prosopagnosic" network. By allowing the networks to learn and discover potentially domain-specific representations on their own, we can gain some insight into the processes that might lead to such specializations in the brain. Although we make no claims that our models are biologically plausible in any significant way, the experts or modules in a given network could be interpreted as representing, for instance, analogous regions in the left and right hemispheres, or two nearby relatively independent processing units in the same region of the brain.

2.2. The network architectures

The first model's network architecture is the well-known "mixture of experts" (ME) network (Jacobs, Jordon, Nowlan & Hinton, 1991). The ME network contains a population of simple linear classifiers (the "experts") whose outputs are mixed by a "gating" network. During learning, the experts compete to classify each input training pattern, and the gating network directs more error information (feedback) to the expert that performs best. Eventually, the gating network learns to partition the input space such that expert 1 "specializes" in one area of the space, expert 2 specializes in another area of the space, and so on.

The second network architecture we use is inspired by the ME network but is slightly more complicated. The main difference between it and ME is that it contains separate hidden layers for each module and a gating network that essentially learns to decide which hidden layer representation to "trust" in classifying a given input stimulus. Fig. 2 summarizes the differences between the two architectures; Appendices A and B describe their operation and learning rules in detail.

Modular networks like the mixture of experts can be useful in a variety of engineering applications, but as Jacobs (1997) argues, they have also been very useful tools for exploring hypotheses about brain function. Jacobs and Kosslyn (1994), for instance, showed that if one expert in a two-expert network was endowed with large "receptive fields" and the other was given smaller receptive fields, one expert specialized for a "what" task whereas the other specialized for a "where" task. As another example, Erickson and Kruschke (1998) have successfully used the mixture of experts paradigm to model aspects of human categorization of visual stimuli. Thus the mixture of experts approach is a potentially powerful computational tool for studying functional specialization in the brain.

2.3. Measuring specialization and effects of local "brain damage"

Since these modular networks naturally decompose given problems in a data-driven way, we can explore hypotheses about the modularity of face and object recognition by training the models to perform combined face/object classification tasks. In both of the network models we have described, the gating network assigns a weight to each expert or module given an input pattern; this weight is the gate network's estimate of the probability the given pattern was drawn from the expert's area of expertise. To determine whether expert or module *n* is a "face specialist", we can present the face patterns in the test set to the network, record gate unit *n*'s activation for each of the patterns, and average them. If that average is high, we can say that expert or module *n* is indeed a face specialist.

We can model localized brain damage by randomly eliminating some or all of the connections in one of the experts or modules. If one expert or module is specialized for a task, such as book classification, but not other tasks, eliminating its connections will degrade the overall model's performance on that task, with less impact on performance of other tasks.

2.4. Face/object stimuli

Our studies utilized static images of 12 individuals' faces, 12 different cups, 12 different books, and 12 different soda cans. See Fig. 3 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. For the 36 objects, we captured 5 images of each with a CCD camera and video frame grabber. We performed minor, pseudorandom perturbations of each object's position and orientation while lighting and camera viewpoint remained constant. After capturing the 640 × 480 grayscale images, we cropped and scaled them to 64 × 64, the same size as the face images.

2.5. Preprocessing with Gabor wavelet filters

In order to transform raw 64 × 64 8-bit gray-scale images into a representation more appropriate for a neural network classifier, we preprocessed the images with a Gabor wavelet-based feature detector. von der Malsburg and colleagues have been using grids of these wavelet filters to extract good representations for face recognition for several years. The 2-D Gabor filter (Daugman, 1985) is a two-dimensional sinusoid localized by a Gaussian envelope; it can be tuned to a particular orientation and spatial frequency. The filter is biologically motivated—it is a good model of observed receptive fields of simple cells in cat striate cortex (Jones & Palmer, 1987). von der Malsburg and colleagues form a "jet" by concatenating the response of several filters with different orientation and spatial frequency tunings. As an image feature detector, the jet exhibits some invariance to background, translation, distortion, and size (Buhmann, Lades & von der Malsburg, 1990).

Early versions of their face recognition system (Lades et al., 1993) stored square meshes of these jets at training time and used them as deformable templates at recognition (test) time to match a test face. More recent versions (Wiskott et al., 1997) place the jets over particular facial features (fiducial points) for greater accuracy. Biederman and Kalocsai (1997) show how Wiskott et al.'s representation can account for psychological phenomena in face recognition, and the system was recently the top performer in the U.S. Army's FERET Phase III face recognition competition (Okada, et al., 1998). Thus the Gabor wavelet jet is a good representation for face recognition. We use a simple version of the square mesh (Buhmann et al., 1990) as described below. Since we use prealigned images and phase-invariant filter responses, the more complicated fiducial point techniques are unnecessary.

The basic kernel function 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

$$\vec{k} = [k\cos\phi, k\sin\phi]^{\mathrm{T}}$$

and $k \equiv |\vec{k}|$ controls the spatial frequency (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 $\{0, (\pi/8), (\pi/4), (3\pi/8), (\pi/2), (5\pi/8), (3\pi/4), (7\pi/8)\}$, and we let

$$k_i = \frac{2\pi}{N}2^i$$

where N is the image width and i an integer. In the first series of experiments, we used 6 scales ($i \in \{1,\ldots,6\}$), and in the second series we used 5 scales ($i \in \{1,\ldots,5\}$). See Fig. 4 for examples of the filters at three particular orientation/scale combinations.

Again as in Buhmann et al. (1990), for each of the orientation/spatial frequency pairs, we convolve $G(\vec{k},\vec{x})$ with the input image $I(\vec{x})$:

$$(\mathscr{W}\mathscr{I})(\vec{k},\vec{x}_0) = \int G_{\vec{k}}(\vec{x}_0 - \vec{x})I(\vec{x})\,\mathrm{d}^2\vec{x}$$

then normalize the response values across orientations:

$$(\mathscr{T}\mathscr{I})(\vec{k},\vec{x}_0) = \frac{|(\mathscr{W}\mathscr{I})(\vec{k},\vec{x}_0)|}{\int\int|(\mathscr{W}\mathscr{I})(\vec{k},\vec{x})|\,\mathrm{d}^2\vec{x}\,\mathrm{d}\phi}.$$

With eight orientations and six scale factors, this process results in a vector of 48 complex values at each point of an image (see Fig. 5 for example filter responses). We subsampled an 8 × 8 grid of these vectors and computed the magnitude of the complex values, resulting in a large vector (3072 elements for the 6-scale representation in Model I or 2560 for the 5-scale representation in Model II) representing the image.

3. Model I: mixture of experts network

Our first model, reported in Dailey, Cottrell and Padgett (1997), was designed to explore the extent to which specialization could arise in a simple competitive modular system

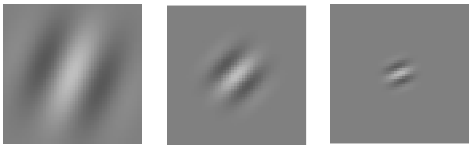


Fig. 4. Real components of Gabor wavelet filters at three different orientations and scales.

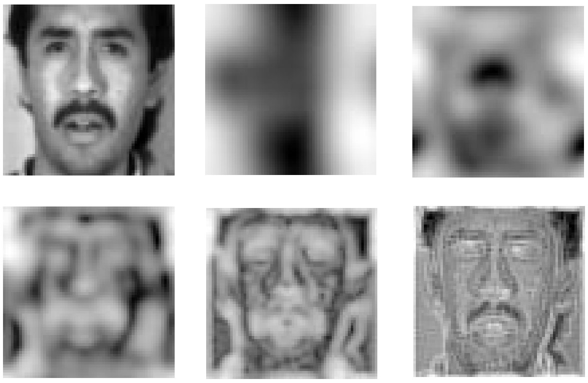


Fig. 5. Original image and Gabor jets at five scales. Each pixel's intensity in the processed images represents the log of the sum of the magnitudes of the filter responses in each of the eight directions.

in which the expert networks' inputs were not biased in any way. The network model was a mixture of experts (Jacobs et al., 1991). Fig. 2(a) shows our two-expert network schematically, and Appendix A describes the network and its learning rules in detail. In short, the "experts" are simple single-layer linear networks, and the gating network learns an input space partition and "trusts" one expert in each of these partitions. The gate network's learning rules attempt to maximize the likelihood of the training set assuming a Gaussian mixture model in which each expert is responsible for one component of the mixture.

We trained the ME model with a simple face/object classification task, observed the extent to which each expert specialized in face, book, cup, and can classification, and finally observed how random damage localized in one expert affected the model's generalization performance. As described in the Face/Object Stimuli section (Section 2.4), we preprocessed each image to generate a 3072-element vector representing the image. The rest of this section describes the training procedure and specialization/damage results.

3.1. Dimensionality reduction with principal components analysis

The feature extraction method described above produced 240 input patterns of 3072 elements. Since neural networks generalize better when they have a small number of independent inputs, it is desirable to reduce the input pattern dimensionality. To accomplish this, 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 efficient technique for PCA described by Turk and Pentland (1991), we projected each pattern onto the basis formed by the 192 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 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.

3.2. Network training

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 was $[1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]^T$, and the pattern for "face 5" was $[0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0]^T$.

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 Fig. 7 results for 0% damage), but we found that the a priori estimates (g_1 and g_2 in Appendix A) learned by the gate network were extremely sensitive to the learning rate parameters (η_g and η_e in Appendix A) and momentum parameters (α_g and α_e in Appendix A). 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 $\alpha_g = 0.6$.

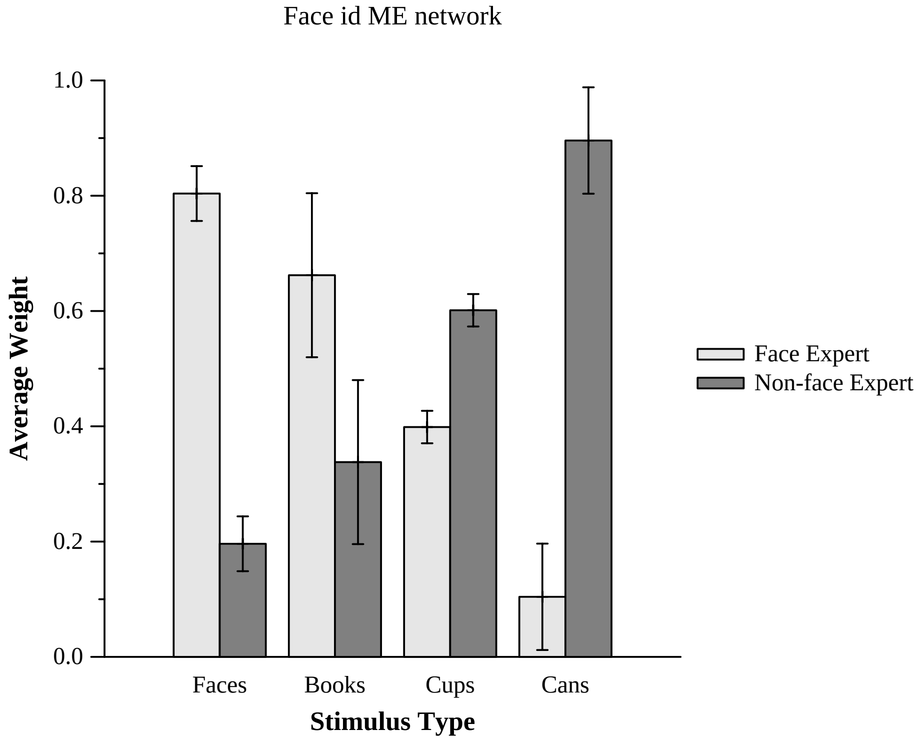


Fig. 6. Weights assigned to the face-dominant expert network for each stimulus class. Error bars denote standard error.

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 may 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; we experimented with this technique and found that the results (as reported in the next section) were robust to this change in the training procedure.

3.3. Model I results

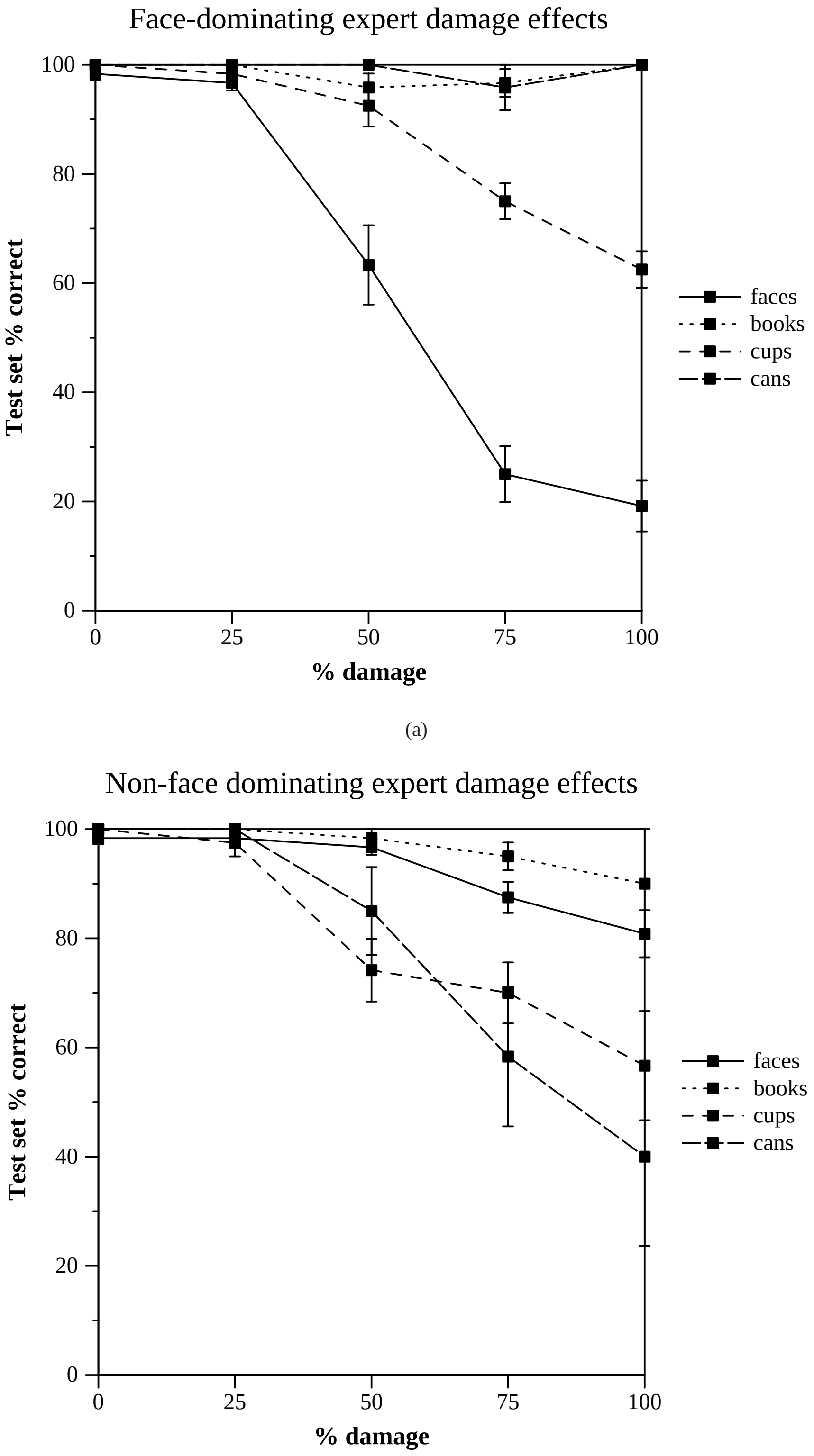
Fig. 6 summarizes the division of labor performed by the gate network over 10 runs with $\eta_e = 0.05$, $\eta_g = 0.15$, $\alpha_e = 0.4$, and $\alpha_g = 0.6$. The bars denote the weights the gate network assigned to whichever expert emerged as face-dominant, broken down by stimulus class, and the error bars denote standard error. Fig. 7 illustrates the performance 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.

3.4. Discussion of model I results

The results show that localized damage in a trained ME network can model prosopagnosia: as damage to the "face" module increases, the network's ability to recognize faces decreases dramatically. From this we conclude that it is plausible for competition between 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 system good at identifying one face will also be good at identifying others. However, since the degree of separation between face and non-face patterns in the model is not clean and is sensitive to training parameters, additional constraints would be necessary to achieve a face/non-face division reliably. Indeed, as discussed earlier, 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/non-face separation in the network, damaging the "face expert" affects face recognition accuracy disproportionately, compared with how damage to the non-face expert affects object recognition accuracy. This is most likely due to the fact that the network is required to perform subordinate classification between members of a homogeneous class (the faces) but gross superordinate classification of the members of the other classes.

This experiment shows how a functional specialization



(a)

(b)

Fig. 7. (a) Face identification classification errors increase as damage to the face-dominating expert module increases, with less impact on object classification. (b) Object categorization classification errors increase as damage to the non-face-dominating expert module increases, with less impact (on average) on face identification.

for face processing could arise in a system composed of unbiased "expert" modules. The next modeling experiment shows that adding simple biologically motivated biases to a similar competitive modular system can make the effect even more reliable.

4. Model II: modular hidden layer network

In the mixture of experts model just described, the experts were very simple linear classifiers and the system was not biased in any way to produce a face expert, although the specialization was sensitive to parameter settings and was not always strong. Our second model, reported in Dailey and Cottrell (1998), was designed to explore the extent to which the learning task and structural differences between modules might strengthen the specializations we observed in the earlier model. In order to allow the expert networks to develop more sophisticated representations of the input stimuli than a simple linear decision boundary, we added hidden layers to the model. In order to make the gating network more sensitive to the task at hand (and less sensitive to the a priori structure of the input space), we trained it by backpropagation of error instead of the ME's Gaussian mixture model. The connections to the modular network's output units come from two separate input/hidden layer pairs; these connections are gated multiplicatively by a simple linear network with softmax outputs. Fig. 2(b) illustrates the model's architecture, and Appendix B describes its operation and learning rules in detail. The model is very similar to the ME in that it implements a form of competitive selection in which the gating network learns which module is better able to process a given pattern and rewards the "winner" with more error feedback.

The purpose of the experiments described in this section was to explore how two biases might affect specialization: (1) the discrimination level (subordinate vs. superordinate) of the task being learned, and (2) the range of spatial frequency information available in the input. We used the same stimuli as in the mixture of experts experiments and trained the model with several different face/object classification tasks while varying the range of spatial frequencies available to the modules. In each case, we observed the extent to which each module specialized in face, book, cup, and can classification. We found that when the system's task was subordinate classification of faces and superordinate classification of books, cups, and cans, the module receiving only low spatial frequency information developed a strong, reliable specialization for face processing. After describing this experiment and its results, we repeat Model I's damage experiments with the specialized networks and analyze the contribution of the input representation to the results.

4.1. Preprocessing with principal components analysis

The Gabor wavelet filtering procedure we used produced a 2560-element vector for each stimulus. As in the mixture of experts model, it is desirable to reduce the input's dimensionality. In this experiment, however, we wanted to maintain a segregation of the responses from each Gabor wavelet filter scale, so we performed a separate principal components analysis on each spatial frequency component of the pattern vectors. For each of the 5 filter scales in the jet, we extracted the subvectors corresponding to that scale from each pattern in the training set, computed the eigenvectors of their covariance matrix, projected the subvectors from each of the patterns onto these eigenvectors, and retained the eight most significant coefficients. Reassembling the pattern set resulted in 240 40-dimensional vectors.

4.2. Network training

Of the 240 40-dimensional vectors, we used four examples of each face and object to form a 192-pattern training set, and one example of each face and object to form a 48-pattern test set. We held out one example of each individual in the training set for use in determining when to stop network training. We set the learning rate for all network weights to 0.1 and their momentum to 0.5. Both of the hidden layers contained 15 units in all experiments. We used the network's performance on the hold out set to determine appropriate criteria for stopping training. For the identification tasks, we determined that a mean squared error (MSE) threshold of 0.02 provided adequate classification performance on the hold out set without overtraining and allowed the gate network to settle to stable values. For the four-way classification task, we found that an MSE threshold of 0.002 was necessary to give the gate network time to stabilize and did not result in overtraining. For all runs reported in the results section, we simply trained the network until it reached the relevant MSE threshold.

We trained networks to perform three tasks:

1. Four-way superordinate classification (4 outputs).
2. Subordinate book classification; superordinate face, cup, and can classification (15 outputs).
3. Subordinate face classification; superordinate book, cup, and can classification (15 outputs).

For each of these tasks, we trained networks under two conditions. In the first, as a control, both modules and the gating network were trained and tested with the full 40-dimensional pattern vector. In the second, the gating network received the full 40-dimensional vector, but module 1 received a vector in which the elements corresponding to the largest two Gabor filter scales were set to 0, and the elements corresponding to the middle filter scale were reduced by 0.5. Module 2, on the other hand, received a vector in which the elements corresponding to the smallest two filter scales were set to 0 and the elements corresponding to the middle filter were reduced by 0.5. Thus module 1 received mostly high-frequency information, whereas module 2 received mostly low-frequency information,

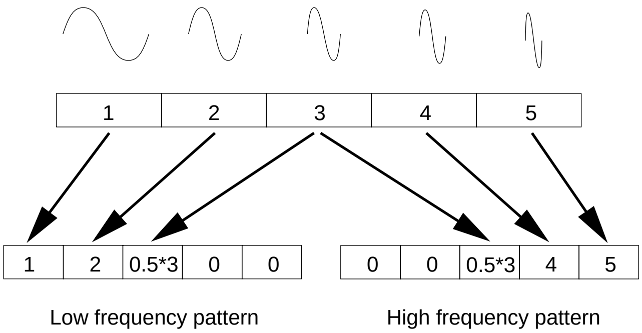


Fig. 8. Splitting input patterns into high spatial frequency and low spatial frequency components.

with de-emphasized overlap in the middle range, as shown in Fig. 8.

For each of the 3 × 2 experimental conditions, we trained networks using 20 different initial random weight sets and recorded the softmax outputs learned by the gating network on each training pattern. As in the ME model, this indicates the extent to which a module is functionally specialized for a class of stimuli. To test performance under localized random damage conditions, we randomly removed connections from a module's hidden layer to the output layer.

4.3. Model II results

Fig. 9 displays the resulting degree of specialization of each module on each stimulus class. Each chart plots the average weight the gating network assigns to each module for the training patterns from each stimulus class, averaged over 20 training runs with different initial random weights. The error bars denote standard error. For each of the three reported tasks (four-way classification, book identification, and face identification), one chart shows division of labor between the two modules in the control condition, in which both modules receive the same patterns, and the other chart shows division of labor between the two modules when one module receives low spatial frequency (LSF) information and the other receives high spatial frequency (HSF) information.

In the control condition, both modules receive the same input, so averaged over many runs, each module wins the competition for any given pattern about half the time. So although on any single run, the modules would show a pattern of specialization similar to the results from Model I, on average, there is no reason for one module to specialize for a pattern class consistently, and the gate weights are thus symmetric.

When required to identify faces on the basis of HSF or LSF information, compared with the four-way-classification and same-pattern controls, the LSF module wins the competition for face patterns extremely consistently (lower right graph). Book identification specialization, however, shows considerably less sensitivity to spatial frequency. We have performed the equivalent experiments with a cup discrimination and a can discrimination task. Both of these tasks show a LSF sensitivity lowe than that for face identification but higher than that for book identification.[2] We have also performed the same experiments providing different patterns of spatial frequency information to the two modules, and the pattern of face specialization is robust. The full pattern contains filters with five different spatial frequency ranges; if the LSF module receives the lowest two ranges (range 1–2) and the HSF module receives either range 3–4 or range 4–5, the face identification specialization is essentially the same.

As shown in Fig. 10, damaging the specialized face identification networks provides a good model of prosopagnosia and visual object agnosia: when the face-specialized (LSF) module's ouput is "damaged" by removing connections from its hidden layer to the output layer, the overall network's generalization performance on face identification drops dramatically, while its generalization performance on object recognition drops much more slowly. When the non-face-specialized (HSF) module's outputs are damaged, the opposite effect occurs: the overall network's performance on each of the object recognition tasks drops, whereas its performance on face identification remains high.

4.4. Why does the low spatial frequency network specialize for face processing?

In order to attain a better understanding of why the LSF module specializes for face processing in model II, we have performed two additional experiments involving the same stimuli but simpler classification models. In the first, we train simple "monolithic" (as opposed to modular) backpropagation neural networks on the same face identification and book identification tasks under three conditions. In the first condition, we train the networks to perform their task given the full range of PCA'ed Gabor filter responses. In the second and third conditions, we present the networks with either the "low pass" patterns or the "high pass" patterns. The monolithic network only has difficulty when the task is face identification and the input stimuli are high pass, i.e. when we attenuate the LSF information in the representation.

In a second experiment, we used an even simpler classification algorithm, nearest neighbors. In this case, the classifier's face accuracy is not impaired by attenuation of the HSF information, but it is impaired by attenuation of the LSF information. In this section, we present the methods and results of the experiments, and in Section 4.5, we

[2] In cup identification, the LSF module receives and average weight of 0.90 for the cup patterns, and in can identification, the LSF module receives an average weight of 0.85 for the can patterns. This is the pattern one might expect given the nearest neighbor analysis below.

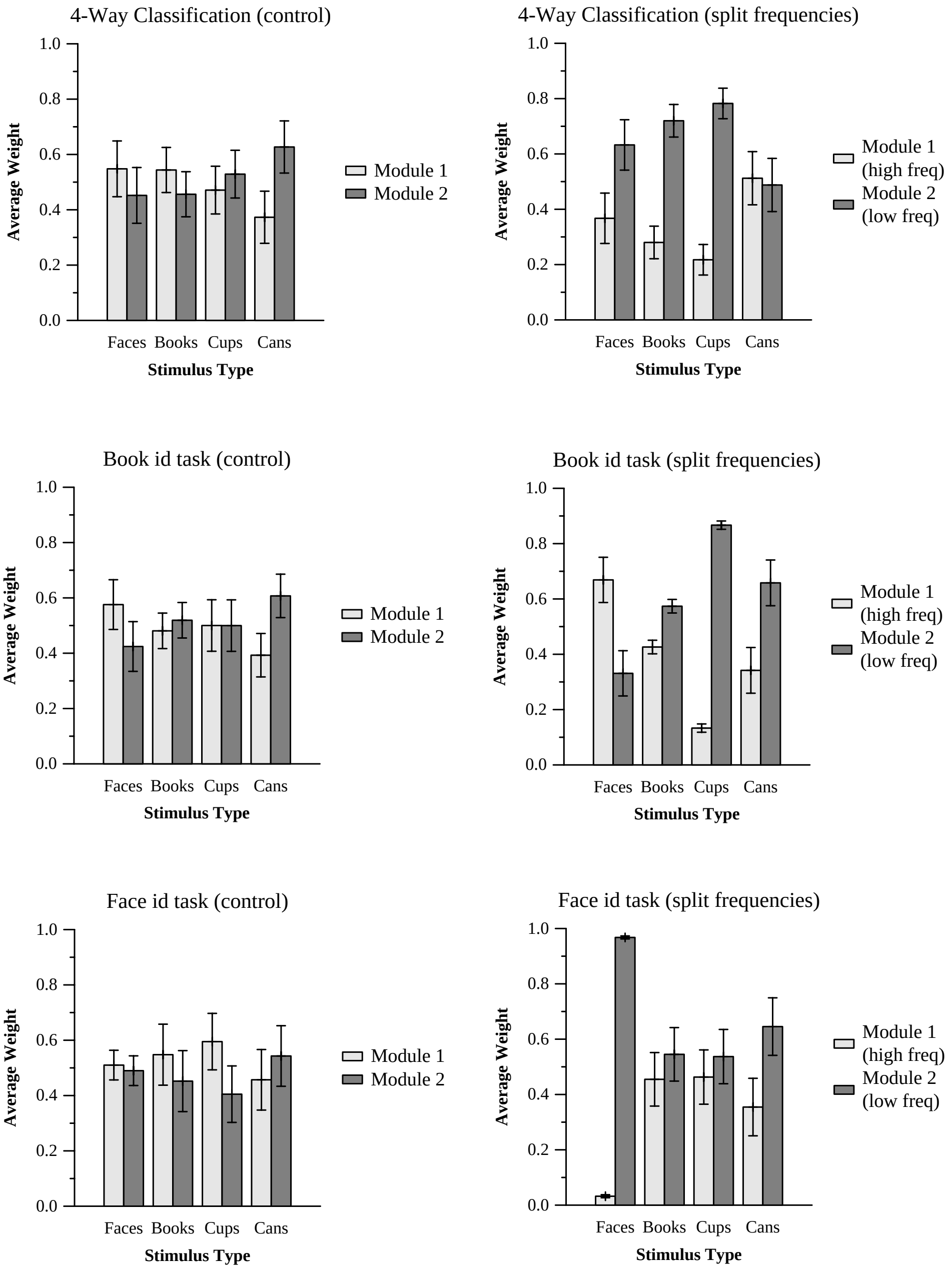


Fig. 9. Average weight assigned to each module broken down by stimulus class. For each task, in the “control” experiment (left column), each module receives the same pattern. The averaged control experiment weights are symmetric because the modules are not biased to prefer one class of stimuli over another. The split-frequency charts summarize the specialization resulting when module 1 receives high-frequency Gabor filter information and module 2 receives low-frequency Gabor filter information.

discuss their implications together with those of the model II results.

4.4.1. *Monolithic network experiment*

The goal of this experiment is to examine the contribution of the low spatial frequency and high spatial frequency information in our stimulus representation to classification by a simple backpropagation neural network.

4.4.1.1. Methods. We used a 2 × 2 design, manipulating the classification task and input stimulus representations, and trained 10 networks under each of the 4 conditions. The two tasks were the same identification tasks used in the model II experiment previously described:

1. *Face identification*: Train networks to perform identification of the faces of 12 individuals (Bob, Carol, Ted, Alice, …) while performing basic-level classification of the books, cups and cans.
2. *Book identification*: Train networks to perform identification of the images of 12 books (Book1, Book2,

Book3, …) while performing basic-level classification of the faces, cups and cans.

We manipulated the stimulus representations in the same manner as in Model II:

1. *Low Pass*: Train networks with high spatial frequency information in their input patterns attenuated.
2. *High Pass*: Train networks with low spatial frequency information in their input patterns attenuated.

Recall that there is overlap between the information present in the low pass and high pass input conditions, i.e. the response of the middle-scale Gabor filters is attenuated by 50% in each condition.

Each of the 40 networks were standard backpropagation neural networks. We constructed the monolithic networks to have approximately the same number of weights as the model II networks, so each network had a 40-unit input layer, a 31-unit hidden layer, and a 15-unit output layer.

As in the previous experiments, we have five examples of each individual face, cup, book, and can. One example of each class was reserved for testing after training was completed (a 48-pattern test set), one example of each class was placed in a hold out set to determine when to stop training (48 patterns), and the remaining three examples of each class formed a 144-pattern training set.

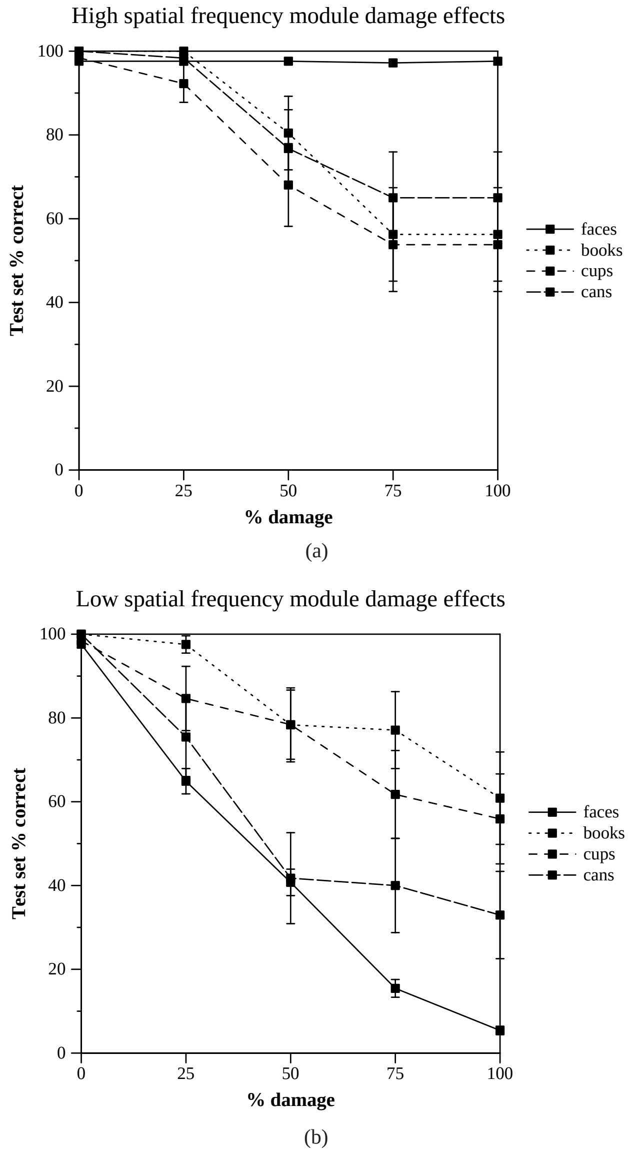
Network training was standard. We initialized each network with small random weights. Each training epoch consisted of one learning pass through the training set in a random order with learning rate 0.1 and momentum 0.5. After every 10 epochs, we tested the network's performance (measured with mean squared error) on the hold out set. We continued this process until hold out set performance failed to increase for 20 subsequent tests (200 training epochs).

4.4.1.2. Results. The graphs in Fig. 11 show how training and hold out error decreases during training. The networks in the high pass condition generally had a more difficult time learning, though the impact was greater for the face identification task than for the book identification task.

Once the networks reached criterion (of no-longer-decreasing hold out set error), we tested their performance on the test set. The graphs in Fig. 12 show the networks' performance in terms of classification accuracy (% correct on the test set). Clearly, the monolithic network finds it impossible to generalize to unseen faces on the basis of a high spatial frequency representation.

4.4.2. Nearest neighbor classification

The fact that the monolithic network shows the same sensitivity to the presence of low spatial frequency information in its input representation provides evidence that the strong specialization for face processing in the modular network of model II is not simply an artifact of the particular architecture we chose for the competitive modular system.


High spatial frequency module damage effects

(a)

Low spatial frequency module damage effects

(b)

Fig. 10. Effect of damaging the specialized face identification networks from Fig. 9. Training on face specialization and splitting the spatial frequency information between the two modules leads to a strong specialization for faces in the low spatial frequency module.

To explore the issue of representation further, we examine the impact of spatial frequency attenuation on an even simpler system, a nearest-neighbor classifier. The goal of this experiment is to evaluate the extent to which the distance between test probes and learned exemplars in the input space can account for the results in the modular and monolithic network models.

4.4.2.1. Methods. In this experiment, we evaluated the ability of a nearest neighbor classifier to correctly place the 48 face, book, cup, and can stimuli in their *subordinate-level* categories by finding, given a test

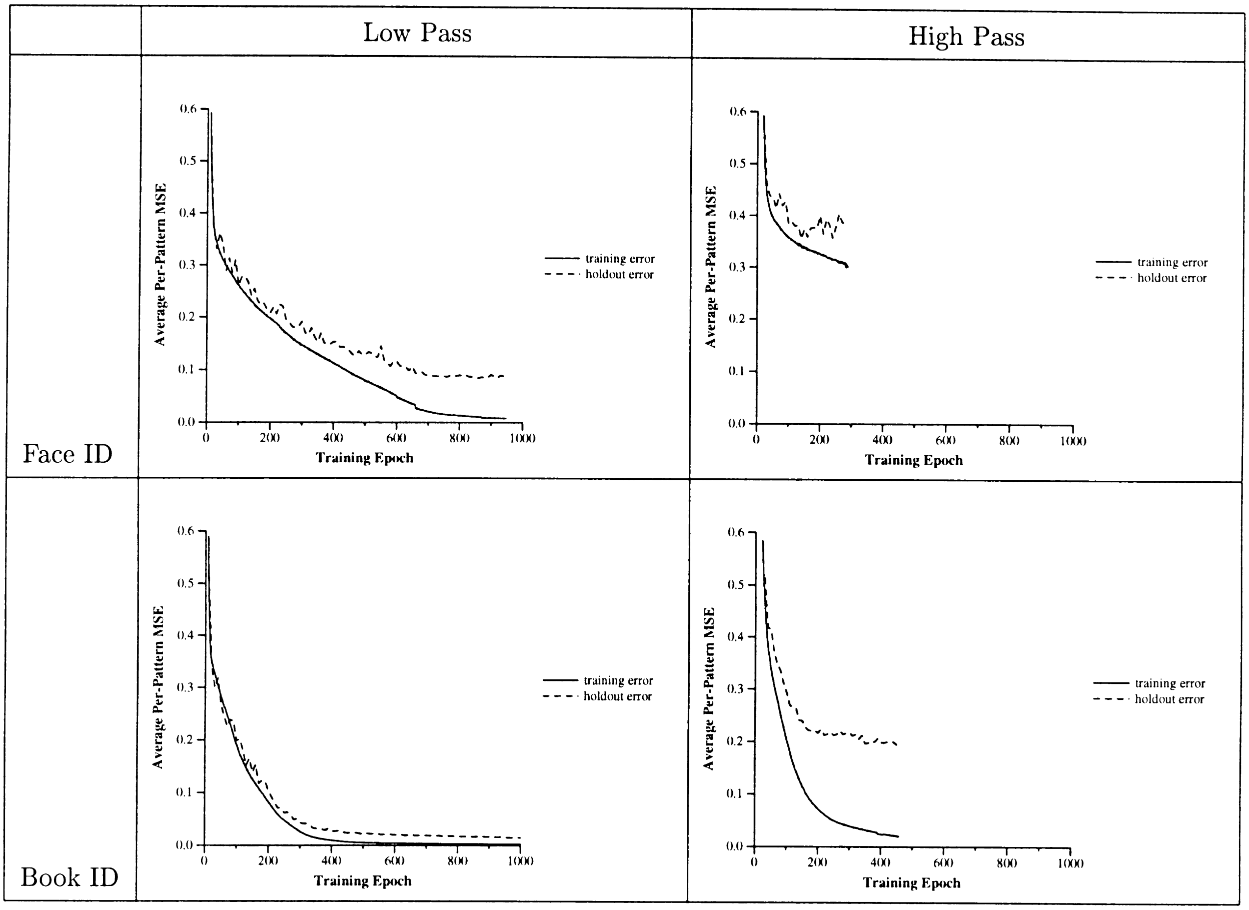


Fig. 11. Monolithic network training/hold out set error over time for the four conditions. Per-pattern mean squared error at epoch t is an average over the networks that actually trained that long. (Networks stop training when they reach criterion.) Each graph is cut off at the point where fewer than three networks were still training or at 1000 epochs.

pattern, the training example closest to it assuming a Euclidean metric. We used three stimulus conditions:

1. *Full pattern*: the same representation used in previous experiments, with the responses of all Gabor filter scales intact.
2. *LSF*: the same set of patterns with attenuated high spatial frequency information.
3. *HSF*: the same set of patterns with attenuated low spatial frequency information.

4.4.2.2. Results. The simplified classifier performs better than we expected; for all three object classes (books, cups, and cans), under all three input representation conditions, its accuracy is well above chance. As shown in Fig. 13, the LSF classifier outperforms the HSF classifier on face identification, cup identification, and can identification, although for book identification, all input representations lead to perfect accuracy. This shows that LSF information is usually more diagnostic of subordinate-level class than is HSF information, and that the face identification specialization results for Model II are largely due to the properties of the input representation. It also shows that our model is able to select the proper network for a given task.

4.5. Discussion of model II results

The results in Fig. 9 show a strong preference for LSF information in the face identification task, empirically demonstrating that, given a choice, a competition mediation mechanism will choose a module receiving low spatial frequency, large receptive field information for this task. The experiments with the monolithic network and the nearest neighbor classifier demonstrate that the large-scale Gabor filters carry the most information relevant to face identification given *this particular set of stimuli.* One problem is that we have only trained our networks on faces and objects at one distance from the camera, so the concept of “low spatial frequency information” is relative to the face or object, not to the viewer. Nevertheless, the resulting specialization in the network is remarkably strong. It demonstrates dramatically how effective a bias in the

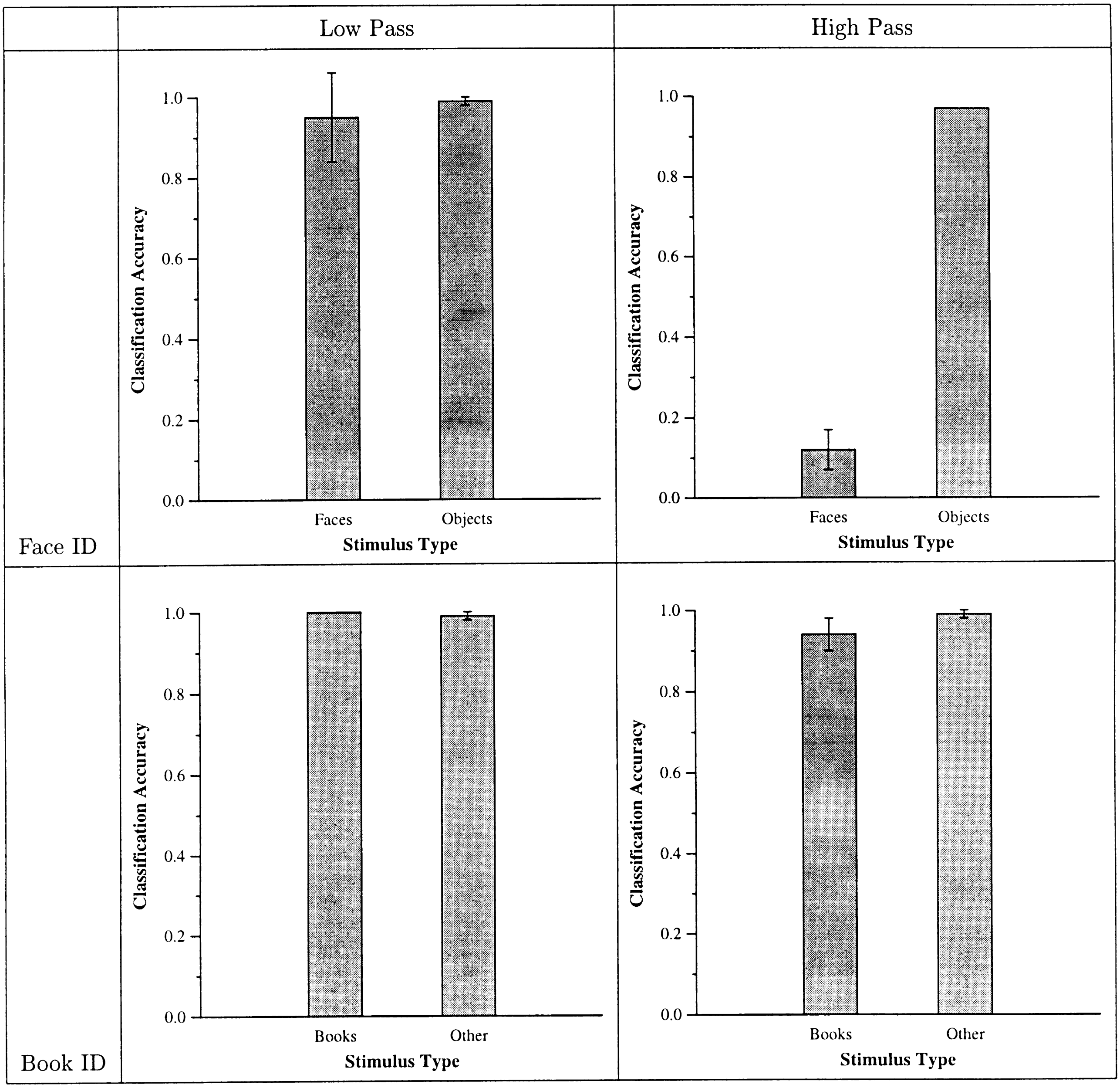


Fig. 12. Test set classification accuracy. Test set accuracy is averaged over ten networks. Only in the Face identification × High Pass condition did networks fail to achieve good test set generalization. Errors bars denote and $\alpha = 0.05$ confidence interval for the mean. Where error bars are not present, all 10 networks had the same classification accuracy for the given set of test patterns.

relative usefulness of a particular range of spatial frequencies can be. The result concurs with the psychological evidence for configural face representations based upon low spatial frequency information, and suggests how the developing brain could be biased toward a specialization for face recognition by the infant's initially low visual acuity.

Inspired by this result, we predicted that human subjects performing face and object identification tasks would show more degradation of performance in high-pass filtered images of faces than in high-pass filtered images of other objects. Costen et al. (1996) have investigated the effect of high-pass and low-pass filtering on face images in isolation, and Parker, Lishman and Huges (1996) have investigated the effect of high-pass and low-pass filtering of face and object images used as 100 ms cues for a same/different task. Their results indicate that relevant high-pass filtered images cue object processing better than low-pass filtered images, but the two types of filtering cue face processing equally well. Similarly, Schyns and Oliva's (1999) results, described earlier, suggest that the human face identification system preferentially responds to low spatial frequency inputs. Finally, Gauthier, Epstein and Gore (1999) have recently provided some preliminary evidence more directly

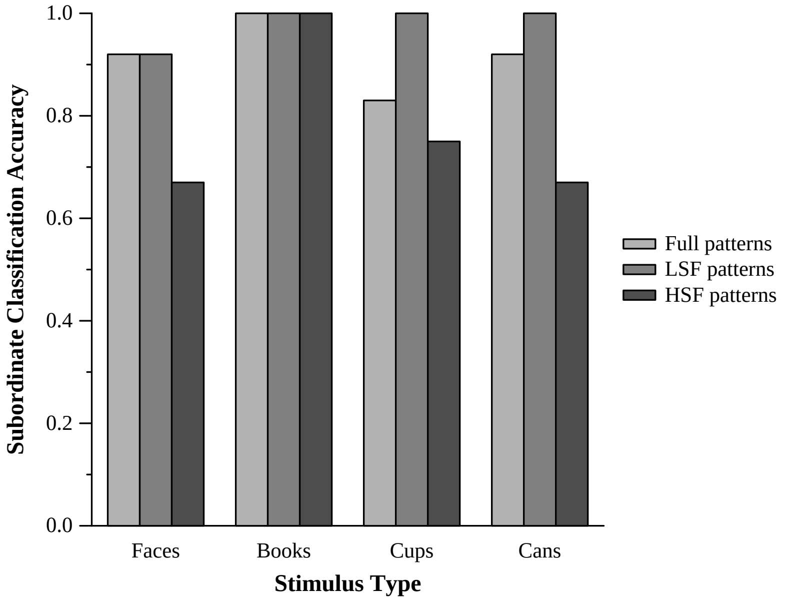


Fig. 13. Subordinate-level classification accuracy for a nearest-neighbor classifier for the face/object dataset. The full patterns were generated as described in Section 4.1 by PCA of the responses of a bank of Gabor filters. The LSF patterns had attenuated high spatial frequency information, whereas the HSF patterns had attenuated low spatial frequency information, as described in Section 4.2 and Fig. 8.

supporting our hypothesis. Their fMRI experiment involved detection of repetitions in a stream of faces, objects, and scenes that were low-pass or high-pass filtered. Subjects responded equally quickly to LSF versions of the stimuli, but were slower at responding to HSF faces than the HSF objects or scenes.

5. General discussion

Both of the models we have described show that localized damage in modular systems can model brain damage resulting in face or object agnosia. In the mixture of experts model, one expert tends to specialize for face recognition because the face patterns are generally near each other, and the gate module's Gaussian mixture model assumption encourages an a priori division of the input space.

Model II's gate network, which we train by the general technique of backpropagating error rather than imposing restrictive constraints on its solution space, is more sensitive to the requirements of the task at hand. The system adapts its division of labor to the given task—this is evident in Fig. 9, where the network's division of labor varies dramatically with the task we train it to perform. Our analysis of the low and high spatial frequency stimulus representations in the monolithic backpropagation network and the nearest neighbor classifier shows in a more direct empirical manner that *for this set of stimuli*, the low spatial frequency information in the representation is critical for subordinate-level face classification but no more useful for subordinate-level object classification than the high spatial frequency information.

The empirical analysis leads to two natural theoretical questions: why should low spatial frequency information be critical for face identification, and why should it be less important for face detection (basic-level classification) or subordinate-level identification of other objects? On the one hand, the task of face detection is very simple for the stimulus set we used in these experiments, so it should be no surprise that high spatial frequency Gabor filter responses can capture the difference between faces and cups, books, or cans. The high spatial frequency components of a jet characterize the local texture in an image, and PCA finds the axes along which the filter responses vary the most. For subordinate-level classification of the objects in our stimulus set, local texture (as encoded by the high spatial frequency filters) should be at least as diagnostic of identity as are the low spatial frequency filter responses, which can only encode the gross shape of features and relationships between them. On the other hand, local texture (again, as encoded in our representation) clearly does not provide enough information for reliable face identification with our stimulus set. We claim that in general, faces do not vary much in local texture—we all have eyes, a nose, a mouth, and a hairline with similar local contrast. *Only* the low spatial frequency components of our representation, then, can possibly encode the kinds of subtle configural differences that Tanaka and Sengco (1997) found were important in their subjects' face perception and recognition.

Taken together, the models empirically demonstrate that prosopagnosia could simply reflect random localized damage in a system trained by competition. This competition could easily be biased by structural and environmental constraints such as:

- infants appear to have an innate tendency to track faces at birth;
- faces are the only class of visual stimuli for which subordinate classification is important at birth;
- learning under these conditions would necessarily be based on gross features and low spatial frequency information due to the infant's low acuity and contrast sensitivity;
- low spatial frequency information is a suitable basis for the holistic representations apparently at work in adult face processing.

In contrast to prosopagnosia, however, localized damage in our networks does not model visual object agnosia (sparing face recognition) especially well. A lesioned network with object processing as badly impaired as C.K.'s, with intact face processing, would be an extremely rare occurrence. Of course, Juola and Plunkett (1998) might argue that C.K.'s brain damage is nothing more than an outlier. But as Moscovitch et al. (1997) point out, C.K. can perceive and recognize the component parts of complex objects but cannot put them together into a whole. The psychological evidence seems to indicate that faces are different from most other objects in that they are perceived and recognized holistically, but our networks do not have much opportunity to form part-based representations of the objects. Although the hidden units in networks like those of Model II could presumably discover parts for their intermediate representations of objects, that would probably require (at least) a much larger training set and more difficult classification tasks. Thus it seems that our models do not possess the part-based representations presumably destroyed in severe object agnosics without prosopagnosia.

As we stated in the introduction to this paper, one theoretical alternative to the "holistic hypothesis" is that faces are merely the most important class of stimuli for which expert-level subordinate classification within a homogeneous basic class is important.[3] Gauthier and colleagues have amassed a great deal of evidence for this hypothesis (Gauthier & Tarr, 1997; Gauthier, Anderson, Tarr, Skudlarski & Gore, 1997; Gauthier, Tarr, Moylan, Anderson & Gore, 1998; Gauthier, Tarr, Anderson, Skudlarski & Gore, 1999). Our experiments do not address the issue directly, but the nearest neighbor subordinate classification results for cups and cans (and, in fact, cup identification and can identification experiments we have not reported in detail) suggest that in many cases, subordinate classification may be biased toward low spatial frequency information. Perhaps this bias is more general than suggested by our preliminary experiments; we plan to explore it more directly in future work.

Finally, we anticipate a few possible criticisms of our model. First and perhaps foremost, the models and stimuli are largely static. We have given our model both high and low spatial frequency channels "from birth". The developing infant's environment is clearly dynamic, and its visual acuity gradually improves. However, we are simply demonstrating that given a choice, the low spatial frequencies are preferable for face discrimination. It is interesting that Nature has arranged for the neonates' visual capabilities to be matched to the task at a time when it is a salient distinction to make. This could be another variant of Elman's observation that a reduced capacity system may be necessary component of development in order to solve certain problems (Elman, 1991a,b). In future work, we plan to explore how the dynamics of these changing task requirements and an increasingly accurate sensory system might interact with functional specialization in our models.

Another potential criticism is that children do not appear to recognize faces holistically (as operationalized by Farah and Tanaka's part–whole paradigm) until at least the age of 6 (Tanaka, Kay, Grinnell, Stansfield & Szechter, 1998). Although the part–whole test is not the most direct way to assess the existence of configural or holistic processing (the advantage for wholes over parts alone has been observed for both novice and expert-level recognition of non-face objects including Greebles, cars, and cells (Gauthier & Tarr, 1997; Gauthier, personal communication), and a more direct test is Tanaka and Sengco's (1997) second-order configuration manipulation, the data could reflect a qualitative shift in the way children process faces. Perhaps there is contention between part-based processing and configural processing that is not resolved until a few years after birth. This is another topic for further research.

6. Conclusion

We have shown in two computational modeling studies that simple data-driven competitive selection combined with constraints and biases known or thought to exist during visual system development can account for some of the effects observed in normal and brain-damaged humans. Our studies lend support to the claim that there is no need for an innately specified face processing module—face recognition is only "special" insofar as faces form a remarkably homogeneous category of stimuli for which within-category discrimination is ecologically beneficial early in life.

Note that we are not arguing that face recognition is modular! In both models, the expert or module specializing for face recognition also plays a role in classifying other types of stimuli. Given that, it may be somewhat surprising that localized but random damage in the networks cause reliable double dissociations between face and object recognition. But this is simply another demonstration of Plaut's (1995) observation that double dissociations do not necessarily imply modularity.

Using competitive computational models to study

[3] Of course, the "holistic hypothesis" and the "subordinate hypothesis" are by no means mutually exclusive.

functional specialization in face processing appears to be a promising avenue for future research. In future work, we plan to explore mechanisms that lead to functional specialization and localization in unsupervised computational models that are more biologically plausible. As another route to increasing our models' plausibility and predictiveness, we will make efforts to realistically incorporate the time course of infant development. We also plan to study other neuropsychological double dissociations, such as that between facial expression and facial identity recognition, with similar techniques.

Acknowledgements

We thank Chris Vogt, Isabel Gauthier, Gary's Unbelievable Research Unit (GURU), and an anonymous reviewer for discussion and comments on previous drafts of this paper. The research was supported in part by NIMH grant MH57075 to GWC.

Appendix A. A mixture of experts learning rules

In this model, the output layers of an array of linear classifiers is combined by a gating network, as shown in Fig. 2(a). We trained this network with the maximum likelihood gradient ascent learning rules described by Jordan and Jacobs (1995).

A.1. Feed-forward phase

In the feed-forward stage, 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 y, or $P(i|\mathbf{x},\mathbf{w})$. The gating network is a single-layer linear network with 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

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

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

$$\mathbf{O} = \sum_i g_i o_i.$$

A.2. Adaptation by maximum likelihood gradient ascent

We adapted the network's estimates of the parameters $\mathbf{w}$ and θ_i, using Jordan and Jacobs' (1995) gradient ascent algorithm for maximizing the log likelihood of the training data given the parameters. Assuming the probability density associated with each expert is Gaussian with identity covariance matrix, they obtain the online learning rules

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

and

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

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 y:

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

This can be thought of as 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 expert's 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 $\Delta w_i(t) = c + \alpha\Delta w_i(t-1)$. We found that setting the learning parameters η_{g}, η_{e}, α_{g}, and α_{e} was not a simple task, as described in the text.

Appendix B. Mixed hidden layer network learning rules

This model is a simple modular feed-forward network. The connections to the output units come from two separate input/hidden layer pairs; these connections are mixed multiplicatively by a gating network similar to that of the mixture of experts. The architecture is shown in Fig. 2(b). We used standard backpropagation of error to adjust the network's weights, but since the multiplicative gating connections add some complexity, we give the detailed learning rules here.

B.1. Feed-forward phase

In the feed-forward stage, the hidden layer units u_{ij} (i is the module number and j is the unit number in the layer) compute the weighted sum of their inputs:

$$I_{ij} = \sum_k w_{ijk} x_k$$

then apply the sigmoid function to the sum:

$$z_{ij} = s(I_{ij}) = \frac{1}{1 + \mathrm{e}^{-I_{ij}}}.$$

Softmax unit i in the gate network computes the weighted sum of its inputs:

$$\xi_i = \sum_k \theta_{ik} x_k$$

then applies the softmax function to that weighted sum:

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

The g_i are positive and sum to 1. The final output layer then computes the weighted sum of the hidden layers of the modules, weighted by the gating values g_i:

$$o_i = \sum_m \left(g_m \sum_j w_{imj} z_{mj} \right).$$

B.2. Adaptation by backpropagation (generalized delta rule)

The network is trained by on-line back-propagation of error with the generalized delta rule. Each of the network's weights w_{ij} for a connection leaving unit i and feeding unit j is updated in proportion to δ_j, the error due to unit j, and x_i, the activation of unit i, with the addition of a momentum term.

For output unit i,

$$\delta_{o_i} = -2(y_i - o_i),$$

where y_i is the ith component of the desired output and o_i is unit i's actual output.

For hidden node u_{ij}, the jth unit in module i's hidden layer,

$$\delta_{u_{ij}} = s'(I_{ij}) \sum_p \delta_{o_p} g_i w_{pij},$$

where s' is the derivative of the sigmoid function, I_{ij} is the weighted sum of u_{ij}'s inputs, g_i is the ith softmax output unit of the gating module, and w_{pij} is the weight on the connection from u_{ij} to output unit o_p.

Finally, the error due to the softmax unit that gates module i is

$$\delta_{g_i} = (g_i - g_i^2) \sum_p \left(\delta_{o_p} \sum_j z_{ij} w_{pij} \right)$$

where z_{ij} is the output activation of hidden node u_{ij} and w_{pij} is the weight from u_{ij} to output node o_p.

Thus the gating units both mix the outputs of each module's hidden layer and give each module feedback during learning in proportion to its gating value (via $\delta_{u_{ij}}$). The architecture implements a simple form of competition in which the gate units settle on a division of labor between the modules that minimizes the entire network's output error.

References

Behrmann, M., Moscovitch, M., & Winocur, G. (1994). Intact visual imagery and impaired visual perception in a patient with visual agnosia. *Journal of Experimental Psychology: Human Perception and Performance*, *20* (5), 1068–1087.

Biederman, I., & Kalocsai, P. (1997). Neurocomputational bases of object and face recognition. *Philosophical Transactions of the Royal Society of London: Biological Sciences*, *352*, 1203–1219.

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

Costen, N. P., Parker, D. M., & Craw, I. (1996). Effects of high-pass and low-pass spatial filtering on face identification. *Perception & Psychophysics*, *38* (4), 602–612.

Cottrell, G. W., & Metcalfe, J. (1991). In R. P. Lippman & J. Moody & D. S. Touretzky (Eds.), *Empath: face, gender and emotion recognition using holons*, (pp. 564–571). *Advances in neural information processing systems 3* San Mateo: Morgan Kaufmann.

Dailey, M. N., & Cottrell, G. W. (1998). In M. I. Jordan & M. J. Kearns & S. A. Solla, *Task and spatial frequency effects on face specialization*, (pp. 17–23). *Advances in neural information processing systems 10* Cambridge, MA: MIT Press.

Dailey, M. N., Cottrell, G. W., & Padgett, C. (1997). A mixture of experts model exhibiting prosopagnosia. In M. G. Shafto, & P. Langley, *Proceedings of the Nineteenth Annual Conference of the Cogntive Science Society* (pp. 155–160) Hillsdale, NJ: Erlbaum.

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

Daugman, J. G. (1985). Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *Journal of the Optical Society of America A*, *2*, 1160–1169.

de Gelder, B., Bachoud-Levi, A. C., & Degos, J. D. (1998). Inversion superiority in visual agnosia may be common to a variety of orientation polarised objects besides faces. *Vision Research*, *38* (18), 2855–2861.

De Renzi, E. (1986). In H. Ellis & M. Jeeves & F. Newcombe & A. Young, *Current issues on prosopagnosia*, (pp. 243–252). *Aspects of face processing* Dordrecht: Martinus Nijhoff Publishers.

De Renzi, E., Perani, D., Carlesimo, G., Silveri, M., & Fazio, F. (1994). Prosopagnosia can be associated with damage confined to the right hemisphere—an MRI and PET study and a review of the literature. *Psychologia*, *32* (8), 893–902.

de Schonen, S., Mancini, J., & Liegeois, F. (1998). About functional cortical specialization: the development of face recognition. In F. Simon & G. Butterworth, *The development of sensory, motor, and cognitive capacities in early infancy*, (pp. 103–116). Hove, UK: Psychology Press.

Elman, J. L. (1991a). Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning*, *7* (2/3), 195–226.

Elman, J. L. (1991b). The importance of starting small. In: *Proceedings of the 13th Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Erickson, M. A., & Kruschke, J. K. (1998). Rules and exemplars in category learning. *Journal of Experimental Psychology: General*, *127*, 107–140.

Farah, M. J. (1990). *Visual agnosia: disorders of object recognition and what they tell us about normal vision*, Cambridge, MA: MIT Press.

Farah, M. J., Levinson, K. L., & Klein, K. L. (1995). 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. (1995). The inverted face inversion effect in prosopagnosia: evidence for mandatory, face-specific perceptual mechanisms. *Vision Research*, *35* (14), 2089–2093.

Farah, M. J., Wilson, K. D., Drain, M., & Tanaka, J. N. (1998). What is special about face perception? *Psychological Review*, *105* (3), 482–498.

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), 395–411.

Fodor, J. (1983). *Modularity of mind*, Cambridge, MA: MIT Press.

Gauthier, I., & Tarr, M. (1997). Becoming a greeble expert: exploring mechanisms for face recognition. *Vision Research*, *37* (12), 1673–1682.

Gauthier, I., Anderson, A. W., Tarr, M. J., Skudlarski, P., & Gore, J. C. (1997). Levels of categorization in visual recognition studied with functional MRI. *Current Biology*, *7*, 645–651.

Gauthier, I., Tarr, M. J., Moylan, J., Anderson, A. W., & Gore, J.C. (1998). The functionally defined face area is engaged by subordinate-level recognition. Poster presentation at *The Fifth Annual Meeting of the Cognitive Neuroscience Society*.

Gauthier, I., Behrmann, M., & Tarr, M. J. (1999). Can face recognition really be dissociated from object recognition? *Journal of Cognitive Neuroscience*, in press.

Gauthier, I., Tarr, M. J., Anderson, A. W., Skudlarski, P., & Gore, J. C. (1999). Activation of the middle fusiform face area increases with expertise in recognizing novel objects. *Nature Neuroscience*, in press.

Gauthier, I., Epstein, R., & Gore, J. C. (1999). The contribution of high and low spatial frequencies to the processing of objects, faces, and scenes. Poster presented at the *6th Annual Meeting of the Cognitive Neuroscience Society*.

Humphreys, G. W., & Riddoch, M. J. (1987). *To see but not to see*, Hillsdale, NJ: Erlbaum.

Jacobs, R. A. (1997). Nature, nurture, and the development of functional specializations: a computational approach. *Psychonomic Bulletin & Review*, *4* (3), 299–309.

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.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., & Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural Computation*, *3*, 79–87.

Johnson, M. H. (1997). *Developmental cognitive neuroscience*, Cambridge, MA: Blackwell.

Johnson, M. H., & Morton, J. (1991). *Biology and cognitive development: the case of face recognition*, Oxford, UK: Blackwell.

Johnson, M., Dziurawiec, S., Ellis, H., & Morton, J. (1991). Newborns' preferential tracking of face-like stimuli and its subsequent decline. *Cognition*, *40*, 1–19.

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

Jordan, M., & Jacobs, R. (1995). Modular and hierarchical learning systems. In M. Arbib, *The Handbook of brain theory and neural networks*, Cambridge, MA: MIT Press.

Juola, P., & Plunkett, K. (1998). Why double dissociations don't mean much. In *Proceedings of the 20th Annual Conference of the Cognitive Science Society* (pp. 561–566) Hillsdale, NJ: Erlbaum.

Kohonen, T. (1995). *Self-organizing maps*, Berlin: Springer.

Lades, M., Vorbrüggen, J. C., Buhmann, J., Lange, J., von der Malsburg, C., Würtz, R. P., & Konen, W. (1993). Distortion invariant object recognition in the dynamic link architecture. *IEEE Transactions on Computers*, *42* (3), 300–311.

McCarthy, G., Puce, A., Gore, J. C., & Allison, T. (1997). Face-specific processing in the human fusiform gyrus. *Journal of Cognitive Neuroscience*, *9* (5), 605–610.

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

Moscovitch, M., Winocur, G., & Behrmann, M. (1997). What is special about face recognition? Nineteen experiments on a person with visual object agnosia and dyslexia but normal face recognition. *Journal of Cognitive Neuroscience*, *9* (5), 555–604.

Okada, K., Steffens, J., Maurer, T., Hong, H., Elagin, E., Neven, H., & von der Malsburg, C. (1998). The Bochum/USC face recognition system and how it fared in the FERET phase III test. In H. Wechsler & P. J. Phillips & V. Bruce & F. F. Soulie & T. Huang, *Face recognition: from theory to applications*, *NATO ASI series F*Berlin: Springer.

Parker, D. M., Lishman, J. R., & Hughes, J. (1996). Role of coarse and fine spatial information in face and object processing. *Journal of Experimental Psychology: Human Perception and Performance*, *22* (6), 1445–1466.

Pascalis, O., de Schonen, S., Morton, J., Deruelle, C., & Fabre-Grenet, M. (1995). Mother's face recognition by neonates: a replication and an extension. *Infant Behavior and Development*, *18*, 79–85.

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

Schyns, P. G., & Oliva, A. (1999). Dr. Angry and Mr. Smile: when categorization flexibly modifies the perception of faces in rapid visual presentations. *Cognition*, *69* (3), 243–265.

Sergent, J., Ohta, S., & MacDonald, B. (1992). Functional neuroanatomy of face and object processing. A positron emission tomography study. *Brain*, *115*, 15–36.

Tanaka, J., & Sengco, J. (1997). Features and their configuration in face recognition. *Memory and Cognition*, *25* (5), 583–592.

Tanaka, J. W., Kay, J. B., Grinnell, E., Stansfield, B., & Szechter, L. (1998). Face recognition in young children: when the whole is greater than the sum of its parts. *Visual Cognition*,, 479–496.

Teller, D., McDonald, M., Preston, K., Sebris, S., & Dobson, V. (1986). Assesment of visual acuity in infants and children: the acuity card procedure. *Developmental Medicine & Child Neurology*, *28* (6), 779–789.

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

Wiskott, L., Fellous, J. M., Krüger, N., & von der Malsburg, C. (1997). Face recognition by elastic bunch graph matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *19* (7), 775–779.