Pre-publication draft of:
Cottrell, G.W. and Hsiao, J.H. (2011) Neurocomputational Models of Face Processing. In A.J. Calder, G. Rhodes, M. Johnson, and J. Haxby (Eds.) *The Oxford Handbook of Face Perception*. Oxford, UK: Oxford University Press.

Neurocomputational Models of Face Processing

Garrison W. Cottrell[1] and Janet H. Hsiao[2]

[1]University of California, San Diego

[2]University of Hong Kong

Until the day we can record from multiple neurons in undergraduates, understanding how humans process faces requires an interdisciplinary approach, including building computational models that mimic how the brain processes faces. Using machine learning techniques, we can often build models that perform the same tasks people do, in neurophysiologically plausible ways. These models can then be manipulated and analyzed in ways that people cannot, providing insights that are unavailable from behavioral experiments. For example, as we will see below, our model of perceptual expertise can be "raised" in an environment where its "parents" are cups or cans instead of faces, and the same kind of processing ensues. This demonstrates, at least from our point of view, that there is nothing special about faces as an object class *per se*; rather, it is what we have to do with them – fine level discrimination of a homogeneous class - that is special.

In this chapter, we will delineate two dimensions along which computational models of face (and object) processing may vary, and briefly review three such models (Dailey and Cottrell, 1999; O'Reilly and Munakata, 2000; Riesenhuber and Poggio, 1999). Subsequently, we will focus primarily on the model we are most familiar with (our own!) and how this model has been used to reveal potential mechanisms underlying the neural processing of faces and objects – the development of a specialized face processor, how it could be recruited for other domains, hemispheric lateralization of face processing, facial expression processing, and the development of face discrimination. At the end, we return to the Riesenhuber and Poggio model to describe the elegant way it has been used to predict fMRI data on face processing. The overall strategy of these modeling efforts is to sample problems that are constrained by neurophysiological and behavioral data, and to stress the ways in which models can generate novel hypotheses about the way humans process faces.

1. The Model Space

There are at least two dimensions along which neurocomputational models of face processing vary[1]. The first is the amount of fidelity to the known neural architecture and processing. The models created by O'Reilly and Munakata (2000) include realistic constraints on the neural units themselves, increasing large receptive fields in layers corresponding to V1, V2, V4, and the dorsal pathway, as well as inputs based upon known representations in Lateral Geniculate Nucleus (LGN) (Figure 1). Images are transformed by a center-surround transform to the input layer. The final layer corresponds to pre-frontal categorization units. While this model is perhaps the most neurophysiologically plausible, the (published) accounts of its use are on very simple stimuli.

[1] In this chapter we focus on neurocomputational models of face processing. Hence we will not be discussing, for example, models of face recognition from the field of computer vision. For a recent analysis of such models from a cognitive science point of view, see (O'Toole, this volume).

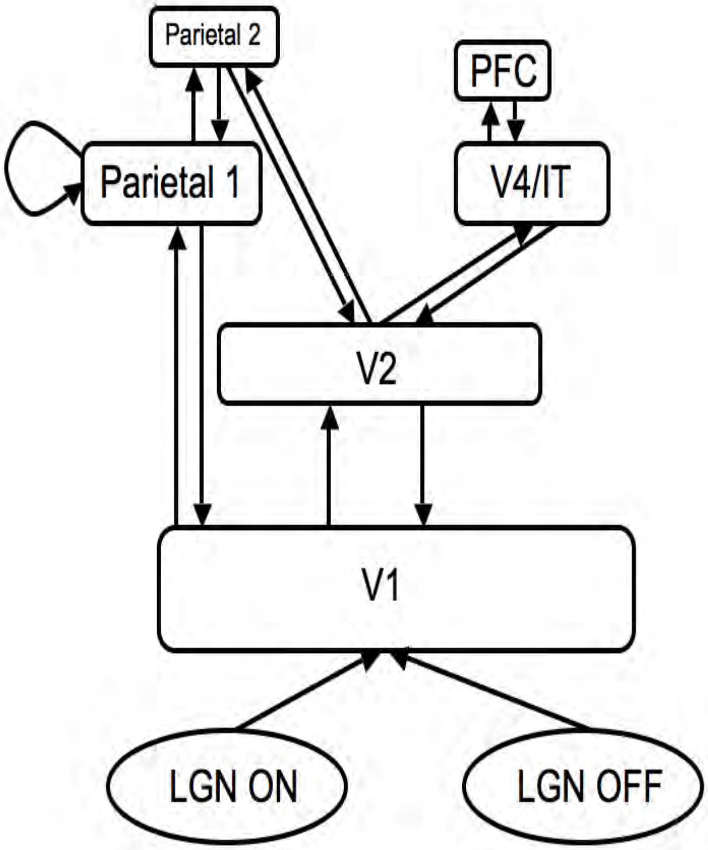


Figure 1. Architecture of O’Reilly and Munakata’s model.

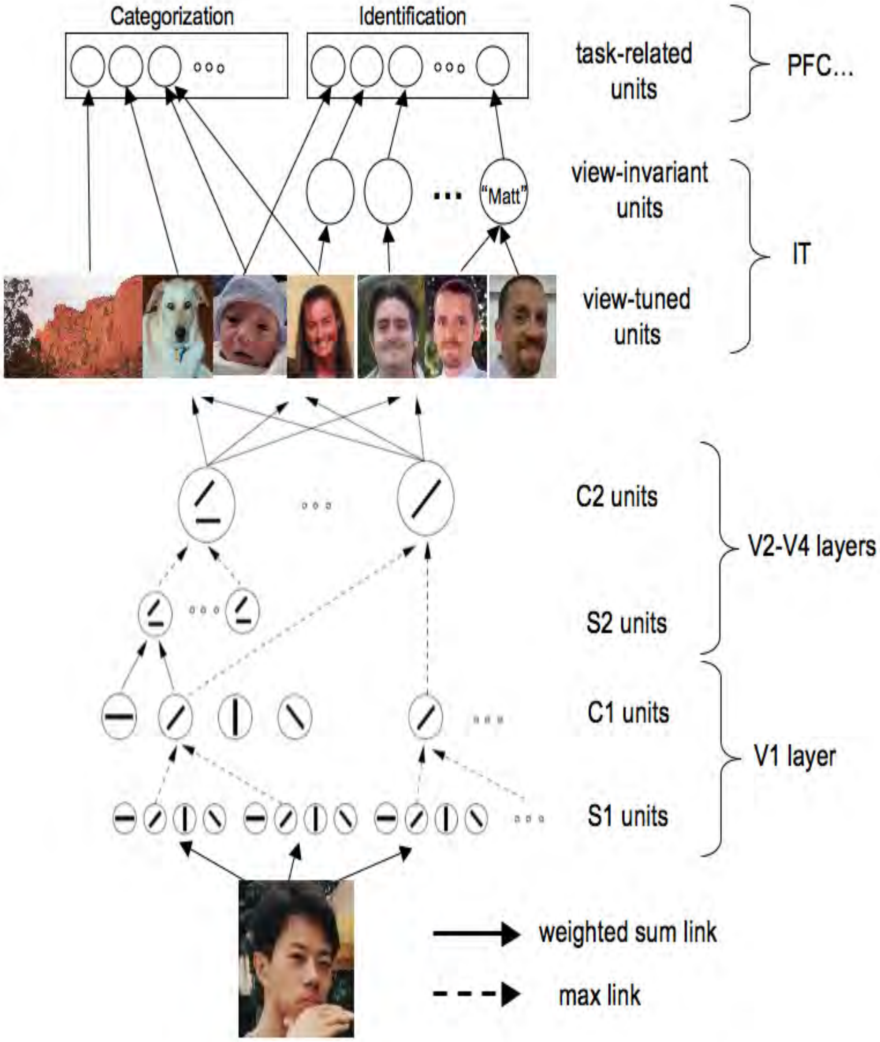
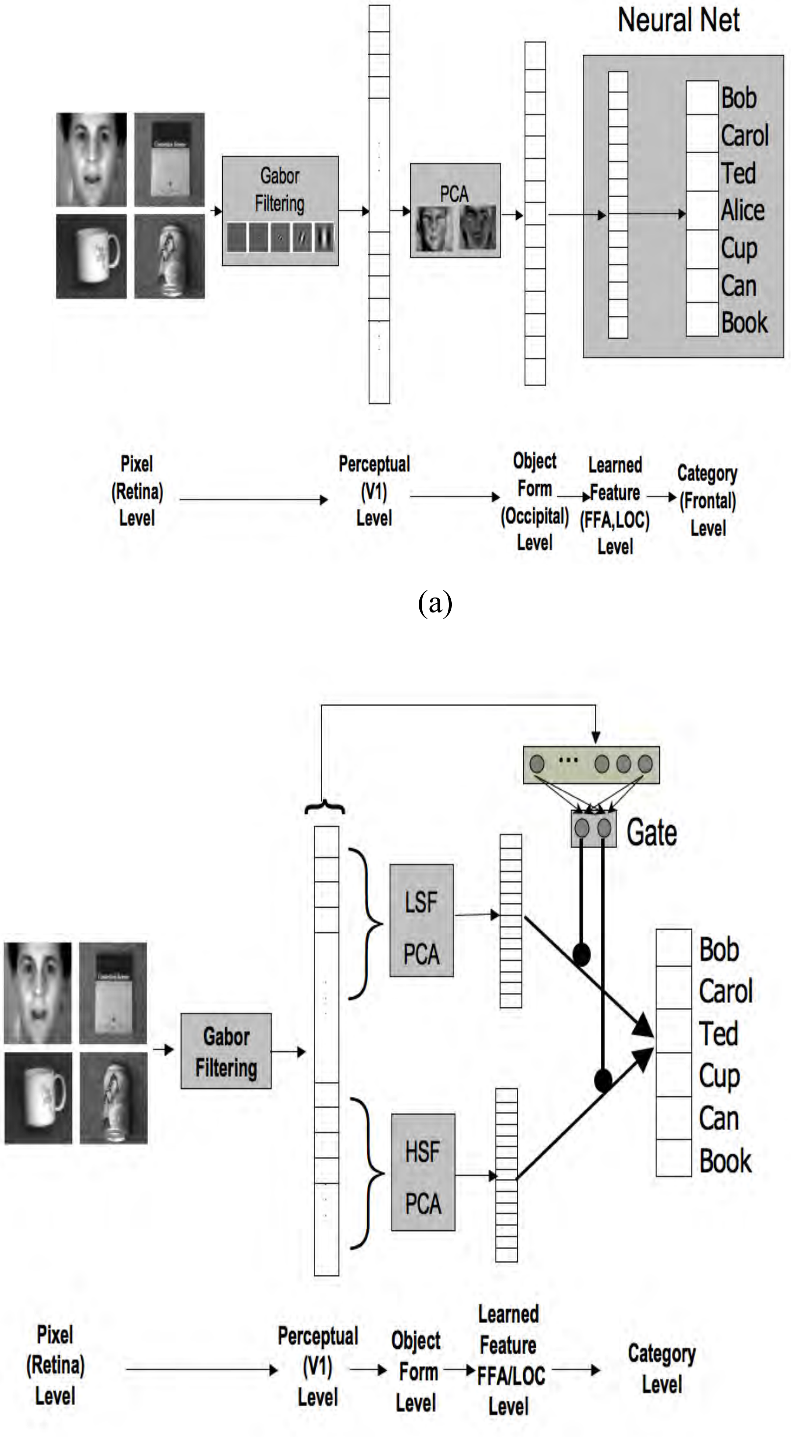


Figure 2. Architecture of Riesenhuber and Poggio's model (a.k.a., "The Standard Model.")

The model of Riesenhuber and Poggio (1999) is based on the standard model of early visual processing in neuroscience, so their model is sometimes called "The Standard Model" (TSM) (Figure 2). They include alternating processing stages of pattern recognition, followed by spatial pooling of responses. These alternating stages are hypothesized to exist in V1, V2, and V4 (with increasingly large receptive fields). Further stages correspond to Inferior Temporal Cortex (IT) and prefrontal cortex (PFC). Images are presented to the network and processed by Gabor filters representing the simple cells of the V1 layer. These units compute a weighted sum of their inputs, which results in them "firing" in proportion to how well the input matches with the Gabor filter. The stages alternate between units that compute weighted sums of their inputs (S-units, like the Gabor filters) to detect particular patterns, and units that compute the maximum of their inputs (C-units) in order to pool the responses of neurons on the previous layers. For example, a unit in the C1 layer fires if any S1 unit connected to it fires. All of the S1 units connected to the same C1 unit have the same orientation within a small area. So, the C1 unit "spatially pools" the S1 units' responses over a small region of space, similar to complex cells in V1. After four such layers (2 S layers and 2 C layers), the next layer up contains units tuned to particular views of whole objects, which feed into view-invariant units. These, finally, provide input to a classification layer.

The models of Dailey and Cottrell (1999; and following models from our lab, which we dub here "The Model" (TM)), are the least biologically plausible, starting with a layer corresponding to V1 which directly feeds into a layer representing early IT, sometimes followed by a layer that represents category specific areas of IT, followed by the PFC (Figure 3a). The first layer units represent the magnitude response of Gabor filters, modeling complex cells in V1 over a very small region of space. The second layer projects the first layer inputs linearly onto cells whose receptive fields are the principal components of the first layer. That is, the second layer responds to *correlations* between the first layer units over a particular training set of images, resulting in global, holistic features, because they derive from inputs across the entire face. The next layer is either the category units or a layer of hidden units before the category units that are task-specific.

The second dimension along which these models vary is the way in which the connection strengths are set. The O'Reilly and Munakata model starts with priors on the architecture in the form of receptive fields being retinotopically mapped, receiving inputs from localized regions of the layer before them, thus giving larger receptive fields at higher layers. These are then trained by the Leabra algorithm, a biologically plausible learning system that combines contrastive Hebbian learning (error driven), standard Hebbian learning (for structure learning), and k-Winner Take All competition, ensuring sparse representations at each layer.



(a)

(b)

Figure 3. (a) Basic architecture of Cottrell and colleagues' model (a.k.a "The Model"). (b) Dailey and Cottrell's (1999) modular network.

TSM has connections set by hand for the first several layers, reflecting the same priors as the O'Reilly and Munakata model, with weights based roughly on physiological data. In some versions, later layers (e.g., the complex V1 to simple V2 connections (C1 to S2) in (Cadieu et al., 2007)) are "trained" by greedy search of parameters that lead to C2 responses that are within some error tolerance of physiologically-measured V4 responses. In other versions (e.g., Jiang, et al., 2006), parameters of later layers are set by brute force search through parameter space to find parameters that lead to behavior that is not statistically significantly different from the desired behavior (e.g., responses of humans in a behavioral experiment). View-tuned units can be set by presenting the network with a face and setting the weights to the observed outputs of the C2 layer, and then view-invariant units can be tuned to these for each person.

Finally, in TM, as noted above, the first layer past the Gabor filters is trained in an unsupervised way by principal components analysis, capturing the statistical regularities over the training set. The next layer is trained by either a simple delta rule if there are no hidden units, or by backpropagation when there are. The network is usually trained to do the same general task required by human subjects (e.g., recognize a facial expression, a particular person, or an object) on a separate set of stimuli, and then tested on the same stimuli as used with the subjects. While backpropagation is not biologically plausible, similar representations can be learned using the contrastive Hebbian algorithm as used in Leabra, which is biologically plausible, as it is completely based on propagation of activations rather than errors.

The three modes of setting the connection strengths have different motivations and consequences. For the O'Reilly and Munakata model and TM, the motivation is to not to impose any preconceptions on the weights, except for the basic architectural constraints. The training tasks are the same sort of tasks people have to perform, and therefore the analysis of the final network involves analyzing the representations formed as a prediction about the actual neural representations. In the O'Reilly and Munakata model, a further motivation is to use a biologically plausible learning rule, so that the model as a whole, which uses biologically plausible units and connectivity patterns (including feedback between layers, a feature neither of the other two models shares), becomes a theory of how the object recognition system comes to be in the first place.

In TM, when there is a wealth of data concerning the task (as in the model of facial expression recognition (Dailey, et al., 2002)), the idea is to create a working model of the task that is at least neurally plausible, and then apply it to novel stimuli used in experimental settings to see how well it fits human data, and make predictions concerning neural representations for the tasks. There is no fitting to human data before testing it in the experimental setting; the model is simply trained to do the same task people do. When there is not a wealth of neurophysiological data (e.g., single cell recordings in the Fusiform Face Area in humans are absent), the model can be used as an intuition pump to make predictions of what we would find were we able to (Tong et al., 2005; 2008).

In TSM, the motivation is quite different. Since the "training" is really a fit to human behavioral or monkey neural data, the idea is that, if the model architecture is neurally plausible, we can then analyze the resulting network's representations and behavior after the fitting process in ways that are unavailable in monkeys or humans. This analysis can provide a better understanding of the actual representations, and make novel predictions that can be tested.

In the following, while we mainly concentrate on TM and its variants, because it is the model we are most familiar with, we will also review an interesting result using TSM.

2. The development of the FFA and its right-hemisphere lateralization

What would turn a relatively generic object recognition system into a face recognition system? There are two main differences between face recognition and generic object recognition. First, faces are a type of visual stimuli that we are exposed to extensively from birth onwards. Hence they have both a primacy and a frequency advantage over objects in other categories. Second, faces must be categorized beyond the basic level – we must recognize individuals, even though in terms of basic features, we all look alike: We (almost) all have two eyes, two ears, a nose and a mouth, so fine-level discrimination is necessary to distinguish each other. It is clear that faces play a privileged role in our everyday lives, and our brains reflect this. fMRI studies have shown that an area within the fusiform gyrus (the *fusiform face area*, or FFA) responds more to faces than to other stimuli, especially in the right hemisphere (Kanwisher, et al., 1997). The domain-specificity of this area has been challenged by some on the grounds that the FFA becomes more active after subordinate-level training on a new visual category

(Gauthier, et al., 1999), and hence that it is a visual expertise area, rather than a face-specific area. Whatever side one comes down on in this debate, it is still the case that, no matter what the FFA's role is, it is more active for faces *first*, and an account of how this comes to be is in order. In this section, we are going to introduce models accounting for the development of FFA and its right-hemisphere lateralization, in order to address how a specialized face processor may emerge under developmental constraints.

2.1 The development of the FFA

The model we describe here has been under development since 1990 (e.g., Fleming and Cottrell, 1990; Cottrell and Metcalfe, 1991; Dailey and Cottrell, 1999). As described above, the basic architecture of the model (TM) incorporates several known observations about visual anatomy (Figure 3(a)). In the first layer, Gabor filters (Daugman, 1985) are used to simulate neural responses in early visual areas such as V1 (Lades et al., 1993). The second layer uses Principal Component Analysis (PCA) as a biologically plausible way (because it can be implemented using the Generalized Hebbian Algorithm, Sanger, 1989) to reduce the dimensionality of the information to simulate possible information extraction processes beyond V1, up to the level of lateral occipital regions; this layer can be thought of as the structural description layer from the classic Bruce and Young (1986) model. The next layer is an adaptive hidden layer that is trained by back propagation to learn features appropriate for a given task; when the task is face identification, it can be thought of as corresponding to the FFA. The fourth layer represents the output of the model, providing labels to the input; it may be analogous to frontal areas (Palmeri and Gauthier, 2004). Because the categories to be discriminated have a strong influence on the hidden layer, these are not just passive outputs; rather, they drive the kinds of representations developed through error feedback.

Dailey and Cottrell (1999) used TM to account for the development of the FFA. The model assumes two main developmental constraints: 1) infants' visual acuity is quite low in the high spatial frequencies (Teller et al., 1986); and 2) their goal is to differentiate their caregivers from each other and from the rest of the people they come in contact with. The model shows that these constraints are sufficient to drive the formation of a specialized face processing area. In order to implement the acuity constraint, they assumed separate spatial frequency input channels that were processed by a channel-specific PCA: there was a (relatively) low spatial frequency input channel and a (relatively) high spatial frequency input channel, and a separate PCA captured the covariances within each channel. By assuming separate spatial frequency input channels, they could model the developmentally appropriate input to the cortex as the low spatial frequency channel, and observe what effects that had on the model's performance. This representation can also be conceptualized as roughly modeling the left and right hemispheres: according to some theories, the right hemisphere has a relatively low spatial frequency bias and the left hemisphere has a relatively high spatial frequency bias (Sergent, 1982; Ivry and Robertson, 1998; cf. Hsiao, Shieh, and Cottrell, 2008; see also the next subsection). The two spatial frequency channels also had separate hidden layers before they reached the output layer (Figure 3(b)). The model assumed that the two spatial frequency networks competed for tasks through a gating network, which fed more error back to the module with the lower error (i.e., a "mixture of experts" model, Jordan and Jacobs, 1995). As shown in Figure 3(b), the gating network mixed the hidden layers of the two spatial frequency modules and gave feedback to each module during learning in proportion to the value of the gating units. The entire network was trained through back-propagation of error with the generalized delta rule; it implemented a simple form of competition in which the gating units settled on a division of labor that minimized the network's output error. If the task being solved was best performed by the low spatial frequency network, then the gating network would learn to weight the low spatial frequency inputs to the output more highly, and the hidden units on that channel would be trained more on that task.

To examine how different learning experiences influence the development of representations, in particular the difference between basic- and subordinate-level categorization, the model was trained to either categorize four classes of twelve objects each at a basic level (i.e., books, faces, cups, and cans), or to individuate one of the classes into twelve different identities while continuing to simply categorize the other three classes of stimuli at the basic level. Consistent with behavioral data, the results showed that a strong specialization of the low spatial frequency module emerged when the model was trained to individuate faces while categorizing the other three classes of stimuli at the basic level; no other combinations of tasks and input stimuli showed a similar level of specialization. Further experiments showed why: a monolithic network trained on the low spatial frequency inputs generalized much better to new faces than a similar network trained only on the high spatial frequencies – which means that learning would proceed much faster in the low spatial frequency network. Thus, the model supported the hypothesis that something resembling a face processing "module", i.e., the FFA, could arise as a natural consequence of infants' developmental environment – poor visual acuity coupled with the goal of individuating people's faces - without being innately specified (see the chapter by Mark Johnson for an alternate view).

2.2 Hemispheric lateralization in face processing

In face perception, it has been shown that a chimeric face made from two left half faces from the viewer's perspective is usually judged more similar to the original face compared with that made from two right half faces (Gilbert and Bakan, 1973; Brady, Campbell, and Flaherty, 2005; Figure 4). This *left side bias* has been argued to be an indicator of RH involvement in face perception (Burt and Perrett, 1997), and may be related to visual expertise. For example, Hsiao and Cottrell (2009) found a RH bias in Chinese character experts. As noted above, fMRI studies of face processing usually find stronger FFA activation in the right hemisphere (e.g., Kanwisher et al., 1997). Similarly, electrophysiological studies of face processing usually show a stronger face-specific wave 170 ms after the stimulus onset over the right hemisphere (the so-called "N170," e.g., Rossion, Joyce, Cottrell, and Tarr, 2003). In an expertise training study with artificial objects, Gauthier and colleagues found that an increase in holistic processing for the trained objects was correlated with increased right fusiform area activity (Gauthier and Tarr, 2002; Gauthier et al., 1999). Neuropsychological data also suggest a link between RH damage and deficits in face recognition and perception (e.g., Meadows, 1974). In short, RH lateralization in face and face-like processing has been consistently reported.

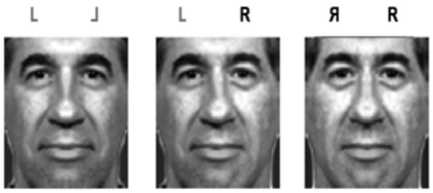


Figure 4. Left chimeric, original, and right chimeric faces. The original face image is taken from the FERET database (Phillips, Moon, Rauss, and Rizvi, 2000).

In order to account for this perceptual asymmetry phenomenon, Hsiao et al. (2008) developed a computational model that aimed to examine the fundamental processing differences between the two hemispheres, and the stage at which the information in the two hemispheres converges (cf. Dailey and Cottrell, 1999). It has been consistently reported that there is a hemispheric asymmetry in the perception of local and global features: an advantage for detecting global features when the stimuli are presented in the left visual field/RH, and an advantage for detecting local features when the stimuli are presented in the right visual field/LH (e.g., Sergent, 1982). In order to account for this hemispheric asymmetry, Ivry and Robertson (1998; cf. Sergent 1982) proposed the Double Filtering by Frequency (DFF) theory; the theory posits that information coming into the brain goes through two frequency filtering stages; stage one is an attentional selection of the task-relevant frequency range; at stage two, the LH amplifies high frequency information, whereas the RH amplifies low frequency information. Hsiao et al.'s (2008) hemispheric model implemented this DFF theory.

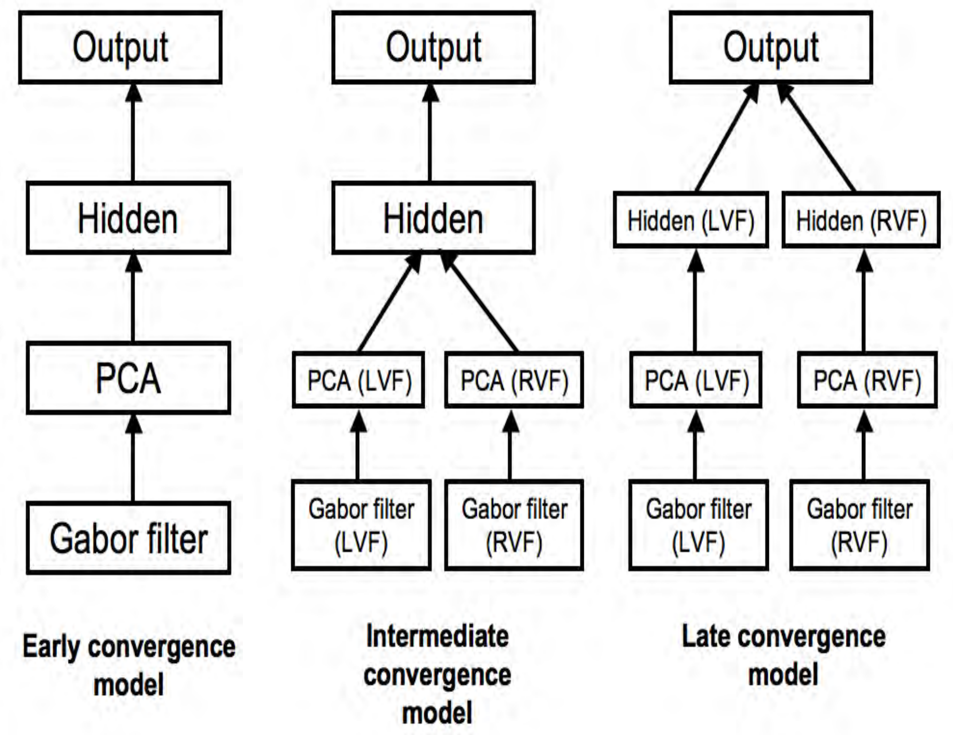


Figure 5. Models of different timings of convergence (adapted from Hsiao et al., 2008).

In order to examine at what stage the information in the two hemispheres starts to converge, Hsiao et al. (2008) compared three models with different timings of convergence (Figure 5). Following the basic architecture of TM (Figure 3(a)), the model incorporated several known observations about visual anatomy: Gabor filters simulated V1 neural responses; PCA simulated possible information extraction processes up to the level of lateral occipital regions; the hidden layers were by analogy with the fusiform area. With this level of abstraction, the timing of convergence may happen at three different stages: in the early convergence model, the convergence happened right after Gabor filters/V1; in the intermediate convergence model, it converged after the PCA/information extraction stage; in the late convergence model, it had two separate hidden layers, and the convergence happened at the output layer. If the DFF manipulation was applied, it was to the Gabor filters before the PCA stage in each model. This was done by attenuating the low frequency Gabor filters for the Left Hemisphere and attenuating the high frequency Gabors for the Right Hemisphere.

In the simulations, two conditions were created: in the baseline condition, no frequency bias (i.e. no DFF theory) was applied, and in the DFF condition, the information in the LH was biased to high spatial frequency ranges whereas that in the RH was biased to low spatial frequency ranges. The models' task was to map each face image to its identity; a localist representation was used in the output layer, with each output node corresponding to each face identity. To examine the models' fit to human data, the left side bias was defined as the size of the difference between the activation of the output node representing the original face when the left chimeric face was presented versus when the right chimeric face was presented; this activation reflected how much the model "thinks" the all-left or all-right stimulus "looks like" the original stimulus. In the unbiased condition (no DFF theory applied), no left side bias was found, suggesting there is no *a priori* bias in facial structure. In the DFF-biased condition, the early convergence model failed to produce the left side bias effect, whereas the intermediate and late convergence models did show the effect (Figure 6(a) and 6(b)). In other words, the model showed that the combination of spatial frequency bias *and* splitting of the information between left and right were sufficient to show the left side bias, but neither alone can show the effect. This result suggests that the visual pathways may converge at an intermediate or late stage, at least after some information extraction has been applied to each separately; Hsiao et al. (2008) speculated that this convergence may be in the lateral occipital region in the visual stream (cf. Tootell et al., 1998). In addition, they trained the model to perform a recognition task on Greebles, a novel type of object (Gauthier, Williams, Tarr, and Tanaka, 1998), and the results again showed that, when the DFF theory was applied, the early convergence model failed to produce the left side bias, whereas the intermediate and late convergence models did exhibit the bias (Figure 6(c)). This predicts that the left side bias will also be observed in expert object processing (cf. Hsiao and Cottrell, 2009).

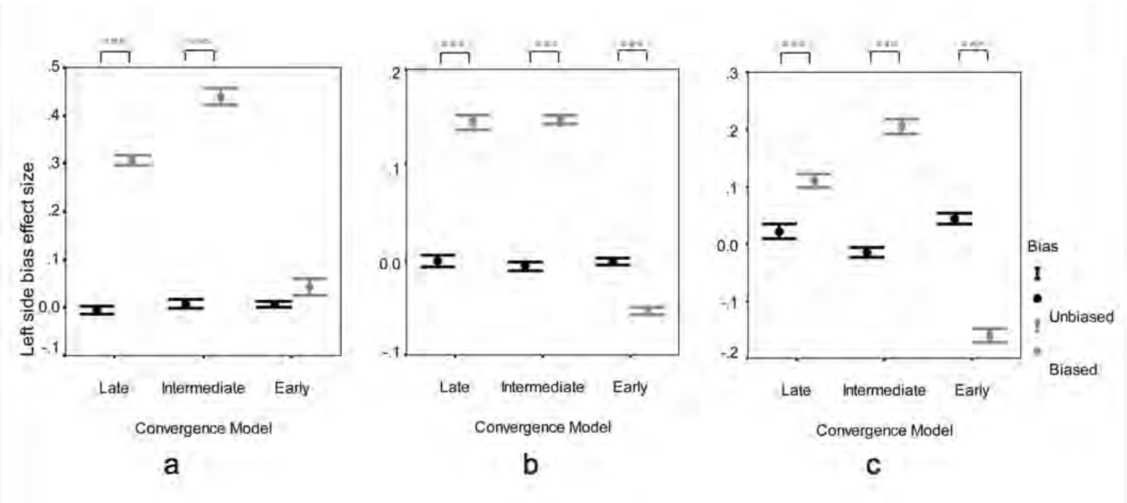


Figure 6. The left side bias effect size in different models in a face identification task with faces of different expressions (left), a face identification task with faces in different lighting conditions (middle), and a Greeble identification task with Greebles in different lighting conditions (right) (images are taken from Hsiao et al., 2008).

3. Behavioral data accounted for by the model

In this section, we review some of the behavioral data accounted for by TM. We begin with facial expression recognition, as historically this is one of the first applications of our model. Next we discuss how the development of face discrimination can be accounted for solely by the PCA level of the model. Finally, we summarize a number of other experiments in modeling behavioral studies.

3.1 Facial Expression Processing

There has been a controversy concerning whether facial expression processing is “discrete” or “continuous.” The proponents of the discrete view consider facial expressions as being perceived categorically, that is, facial expression recognition shows the operational definition of categorical perception: Sharp boundaries between the categories (when subjects judge morphs between expressions) and greater discrimination of two faces along the morph continuum when they are near or cross a category boundary (Calder et al., 1996). On this view, facial expressions are placed in a category, and there are no underlying dimensions of similarity once they are categorized (but see Calder et al., 2000, for a more nuanced view). Proponents of the continuous view point to data showing that when subjects make similarity judgments of facial expressions, and the data is subjected to a Multidimensional Scaling (MDS) procedure, a facial expression "circumplex" is revealed, where surprise is between happy and fear, and the other negative emotions are arranged opposite happy (Russell, 1980). Under this view, facial expression perception is graded, and relies on an underlying two- or three-dimensional space, with two of the dimensions being valence and intensity.

The puzzle was intensified by Young et al. (1997), who used multiple measures from the same subjects when viewing images from a 15-way morph between the six basic emotions (6-way alternative forced choice (6-AFC), ranked choice (a second and third button push for any other emotions perceived), reaction time, discrimination and similarity measures) to reveal both data consistent with categorical perception, and data consistent with the dimensional view. For example, the ranked choice results showed that subjects were above chance at selecting the mixed-in category even at a morph level (70/30) when they were consistently picking the 70% category in the 6-AFC.

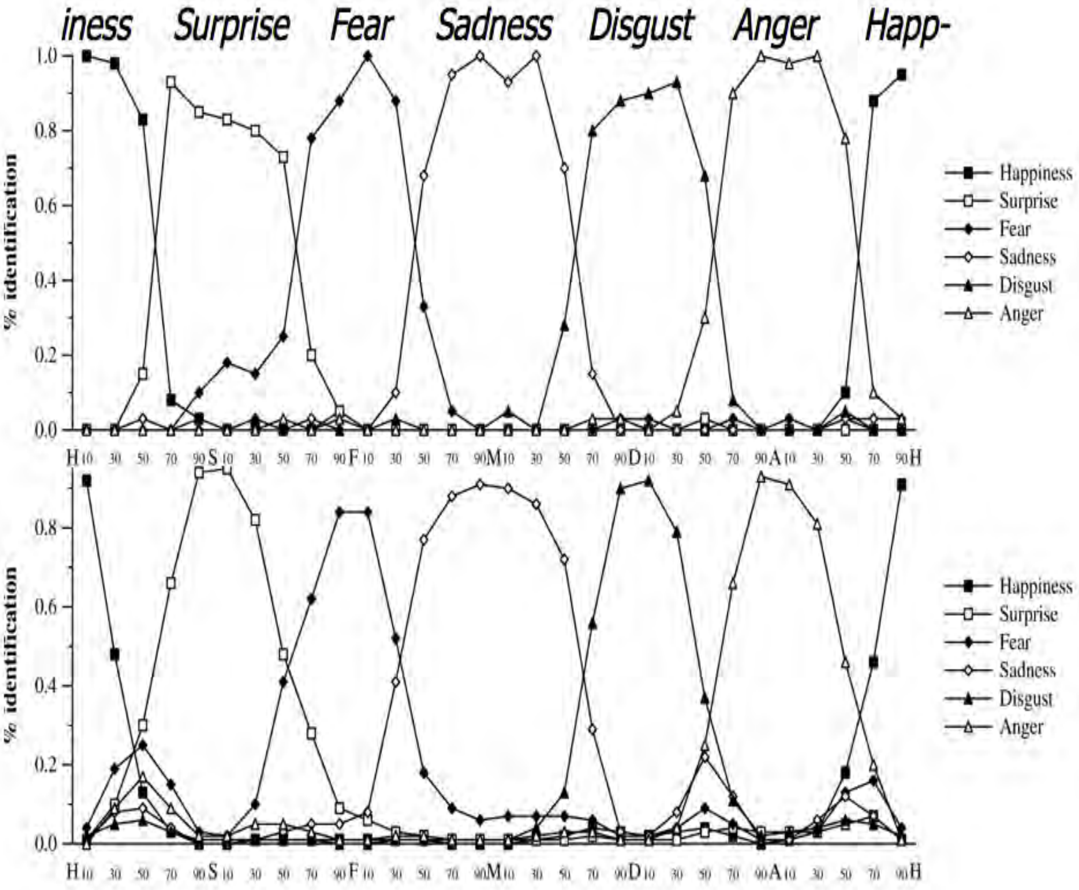
In (Dailey et al., 2002), we modeled these results using a simple version of TM (Figure 3(a) with no hidden layer), and were able to fit *both* the data supporting categorical perception and the data supporting the dimensional view. We concluded that categorical perception is really the result of a simple decision process over a continuous representation (see also Ellison and Massaro, 1997, for a similar view), and that the data could be accounted for by a simple neurocomputational model trained to recognize facial expressions.

We modeled this experiment using TM trained on the six basic emotions using the Pictures of Facial Affect (POFA, Ekman and Friesen, 1976) dataset, excluding the subject "JJ," as he was the one used in the Young et al. experiment (i.e., the network in Figure 3(a) is trained with six outputs, one for each emotion, and no hidden units were used between the PCA and the outputs). We constructed image-quality morphs of JJ and tested the trained TM on them (it should be pointed out that the model had never been exposed to facial morphs or JJ), deriving the same measures from TM as were used with the human subjects (Dailey et al., 2002). For 6-AFC, we chose the most activated output unit as the model's choice. For ranked choice, we simply chose the second and third most activated outputs, and for reaction time, we used the uncertainty of the most activated output, where uncertainty is defined as the difference from 1 (the maximum possible output). The idea is that reaction time reflects certainty – an output of .9 will cause a faster button press than an output of .8. For similarity judgments (in order to apply MDS) with a model that only processes one face at a time, we presented the model with each face, and then correlated the responses of the model. Finally, to model discrimination between two faces, we use one minus the correlation (similarity). An interesting aspect of these last two measures is that they can be performed at any layer of the network, from the image pixels up to the output layer. We can then compare the results to the human data to see which layer of the model best accounts for the data. This led to some surprising results, as will be seen shortly.

First, the model finds fear the hardest expression to recognize, and often confused it with surprise, just as human subjects do. This is an emergent property of the model, since it was trained upon the majority agreed-upon category for each face. I.e., if 90% of the human subjects rated a face as surprised, and 10% rated it as fear, the model was trained to label the face as surprised. The (human) confusion between fear and surprise could therefore simply be a consequence of the similarity of the facial patterns and the categories we place them in. On this view, the reason why fear is the hardest expression to recognize is that it is not sufficiently differentiated from surprise in its appearance, rather than due to some more complex psychological account.

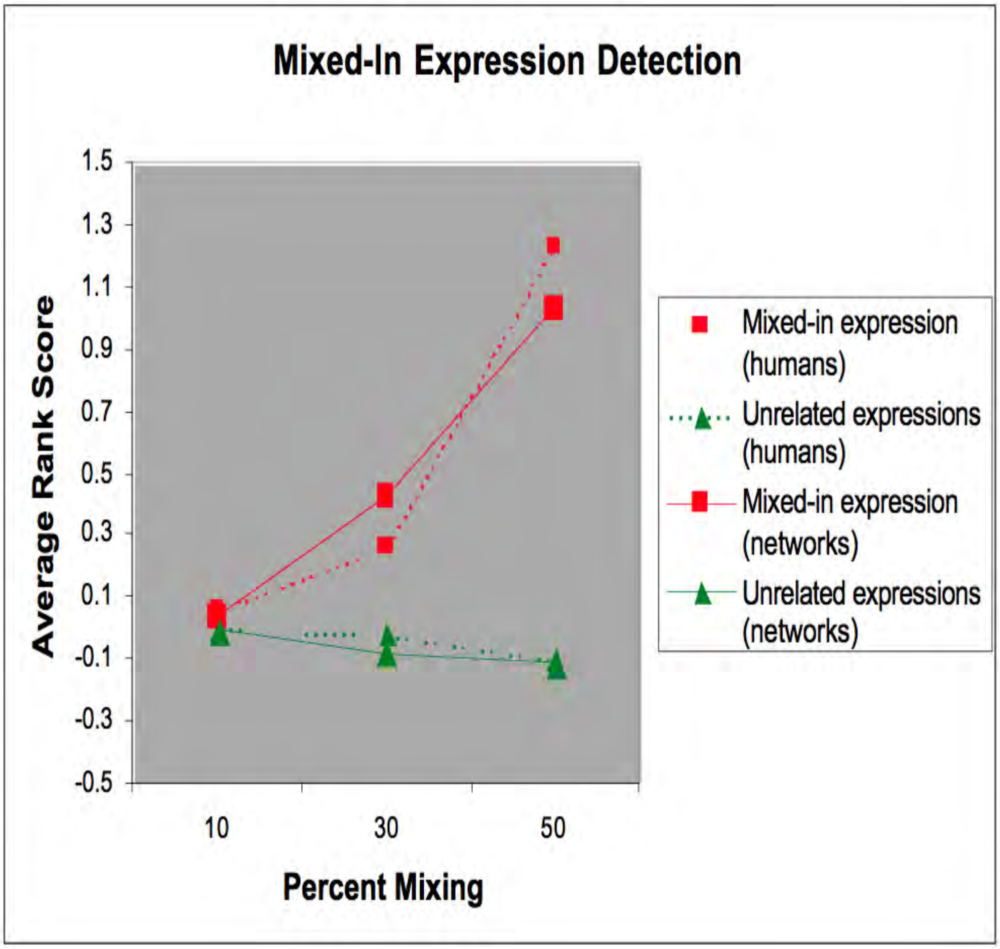
The model was able to account for the data consistent with categorical perception, showing high correlations (*r*=0.942) with the human categorization responses to the morph stimuli (Figure 7(a)), and good correlations with the discrimination scores (*r*=.65).

The model was also able to account for the data consistent with the dimensional account. The human reaction times showed a characteristic “scalloped” shape – slower

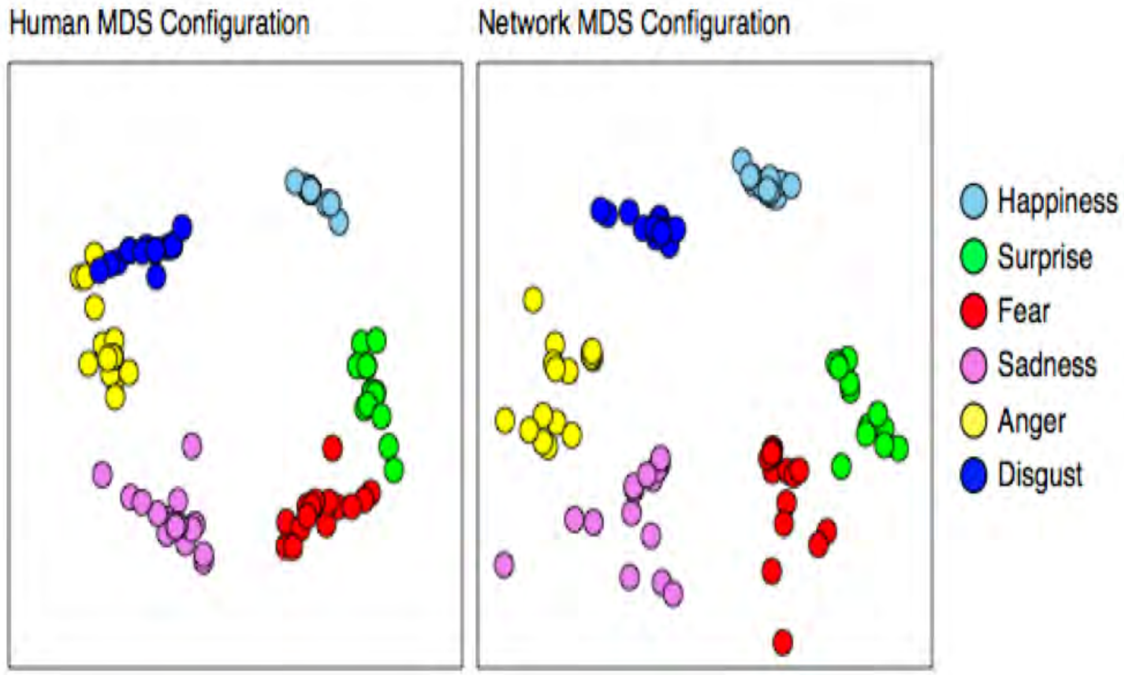


iness
Surprise
Fear
Sadness
Disgust
Anger
Happ-
% identification
1.0
0.8
0.6
0.4
0.2
0.0
H 10 30 50 70 90 S 10 30 50 70 90 F 10 30 50 70 90 M 10 30 50 70 90 D 10 30 50 70 90 A 10 30 50 70 90 H
Happiness
Surprise
Fear
Sadness
Disgust
Anger

(a)


Mixed-In Expression Detection
Average Rank Score
1.5
1.3
1.1
0.9
0.7
0.5
0.3
0.1
-0.1
-0.3
-0.5
10
30
50
Percent Mixing
Mixed-in expression (humans)
Unrelated expressions (humans)
Mixed-in expression (networks)
Unrelated expressions (networks)

(b)



(c)

Figure 7 Results of the expression recognition network. (a) Model results (bottom) versus human proportion of button push (top) for six morph sequences. (b) Model and human results for mixed-in expression detection. The data show that humans and the model both can detect the mixed-in expression at above chance levels at the 30% morph level, when both the networks and the humans are consistently rating the image as the majority (70%) expression. (c) Circumplexes from the humans (left) and the model (right) based on similarity scores. The rotational orientation is meaningless. The order around the circumplex is matched by the model.

near the category boundaries and faster in the middle of a category, which Young et al. took to be inconsistent with the categorical account. Our model mirrored this (with correlation of *r*=.677), because of the way reaction time was modeled. The probability of a face being in a category drops smoothly as the morph moves farther away from the category, so the output goes down, and reaction time goes up. Furthermore, the ranked choice plots for the human subjects and the networks were very similar (Figure 7(b)).

A surprising result concerning the dimensional account is that the MDS of the model's *output* representation gave the same circumplex as the human data (Figure 7(c)).

Hence, human similarity judgments of facial expressions can be accounted for by a simple pattern recognizer's judgment of the probability of the category, given the face, *not* by the underlying perceptual space (the PCA layer representation did not match the human circumplex as well as the output layer). This is, of course, quite counter to the categorical perception view, because that view is based on the idea of the actual *selection* of the highest likelihood category, and ignores the actual probability distribution over categories.

This model has also been used to account for cultural differences in facial expression recognition (Dailey et al., 2010). We conducted a cross-cultural facial expression recognition experiment using Japanese and American subjects viewing Japanese and American facial expressions, the first to demonstrate directly an in-group advantage in facial expression (previous evidence was found via a meta-analysis of a number of experiments (Elfenbein and Ambady, 2002; ; see also Ambady and Weisbuch, this volume). We then used the model to account for the ingroup advantage by training it on Japanese versus American expressions of emotion, and showed that the differences could be accounted for by experience.

3.2 The development of face processing

Face processing is typically described as holistic or configural. Holistic processing is typically taken to mean that the context of the whole face has an important contribution to processing the parts: subjects have difficulty recognizing parts of the face in isolation, and subjects have difficulty ignoring parts of the face when making judgments about another part (Tanaka and Farah, 1993). Configural processing means that subjects are sensitive to the relationships between the parts, e.g., the spacing between the eyes. These two, related, forms of processing, are characteristic of face processing, but they are somewhat oversold in the literature. People are also, of course, highly sensitive to *featural* differences between faces. Recently there has been a great deal of interest in the development of these sensitivities. In this section, we describe one such study and how we can account for the development of sensitivity to features and configuration in our model. Holistic processing can easily be captured by a model that uses whole-face template-like representations as ours does (the principal components of the Gabor filters are global representations of the face). It is interesting that this representation also captures featural differences.

The ability to discriminate faces develops throughout childhood. It has been shown that even children of ten years of age do not achieve adult-level sensitivity to configural differences in faces (Mondloch, et al., 2002). What is the mechanism underlying this protracted developmental period? While there are likely to be a number of influences on development of face expertise, we used TM to investigate two that seemed potentially relevant. First, neuroplasticity studies have shown that changes in inputs to cortical areas lead to altered neural representations in several modalities, a result we assume carries over to changes in visual experience and hence higher level representations (Jenkins et al., 1987; Buonamono & Merzenich, 1998; Gauthier et al., 1999). Also, fMRI studies suggest that the FFA develops over many years in children (Scherf, et al., 2007). Hence, the first variable we investigate is increasing representational resources over time. Second, children meet an increasing number of people over time – first their immediate family, then classmates in elementary school, then more people in high school. It is possible that the social requirement of having to distinguish a greater number of individuals over the years drives the need to use configural information.

We investigated this question using only the first two layers of our model (Zhang and Cottrell, 2006), i.e., principal components analysis of Gabor filter representations. We simulated development qualitatively using a very simple manipulation of this preprocessing, i.e., how many people the model "knows," and how many principal components it uses to represent the data. These two variables were sufficient to account for the slower development of configural processing compared to featural processing, and how the child's performance improves with age. We trained the model on from 100 (19 individuals) to 500 faces (107 individuals) from the FERET face database (Phillips, et al., 2000), and evaluated them on the "Janes" images from Mondloch et al. (2002), which we describe shortly. Because human subjects had to discriminate the faces in that experiment, the evaluation of the model consisted of measuring the average discriminability of the face sets. We measured this at the PCA layer, using 1 minus the correlation between the representations of the faces as the discrimination measure, as in the expression recognition model. The rank order of the discriminability was then compared to the human subject data.

The Mondloch et al. (2002) images vary in several carefully controlled ways. Starting with "Jane," there are a number of variations of her face made by varying the distance between the eyes and the distance between the mouth and the nose. This is the configural set. Then, there are variations constructed by replacing her mouth or eyes with someone else's. This is the featural set. There is also a contour set, constructed by replacing the outside of her face with someone else's. Finally, there is the "cousin" set, which are simply pictures of different women of about the same age. Children ages 6, 8 and 10, as well as adults, were asked to discriminate pairs of faces, both upright and inverted. The results of children's performance on average, in order of best to worst, is: cousin > featural > contour > configural. The adult ranking is cousin > featural > configural > contour. For the inverted faces, the adult and child performance was ranked identically: featural > contour > configural (cousins were not tested in the inverted condition).

We found that in our model, when the number of the training images (and individuals) is small, the discriminability of the configural set is the lowest, which is also observed in children' performance. As the number of images increases, the discriminability of the configural set slowly catches up and exceeds that of the contour set, as observed in adults' performance. In parallel, we found that an increase of the number of components is able to account for the continuous improvement in the performance of all the Jane's sets, but does not cause a change in rank discriminability. Taken together, a "child" model exposed to a small number of people, and with a representation using a small number of resources will not discriminate faces as well as an adult, and the ranking of difficulty on the Jane's sets will mimic that of the children. An "adult" model exposed to a large number of people and with greater representational resources will have a better ability to discriminate the stimuli, and the rank order on the four sets matches that of the adults, both for upright and inverted faces. The careful reader will notice that these layers of model are *unsupervised*, so the increased representation of configural information is simply a consequence of the statistical structure of the face space as more individual faces are added, rather than the need to distinguish them, as hypothesized above.

3.3 Additional behavioral data accounted for

We do not have space to cover other behavioral effects that TM has been used to account for, hence we briefly summarize them. The model has been used to account for holistic processing effects in identity and expression, and the lack of interaction between them, as found by Calder and colleagues (Calder et al., 2000; Cottrell et al., 2002). Of main interest here is that the model did not need separate processing pathways in order to show a lack of interaction (see Calder, this volume, for more discussion). In addition to the developmental work described above, the model has been used to account for age of acquisition effects in face processing (Lake and Cottrell, 2005). Although models of the Other Race Effect (ORE) have been built before (O'Toole, et al., 1991), we have used our model to account for the Other Race Advantage in visual search (Haque and Cottrell, 2005; Levin, 2000). The basic idea of the model was that other race faces contain more information in the Shannon sense, and hence are more salient (see also (Zhang, et al., 2007) for a similar account). Finally, the model has been used to account for priming effects in the discrimination of nonsense characters (McCleery, et al., 2008). It had been shown that Chinese subjects primed to think of nonsense characters as a face discriminated them better than when they were primed to think of them as a Chinese character, when they differed configurally. We constructed a model virtually identical to our face processing model that recognized Chinese characters. We assumed that the priming caused the subjects to use either their face network or their Chinese character network. Since recognizing Chinese characters involves mapping similar characters (in different fonts) to the same category, it is a compressive mapping, lumping stimuli together and ignoring configural differences, while face processing is the opposite – it must differentiate between similar things (faces), and hence it involves an expansive mapping, spreading faces out in representational space, including on the basis of configural differences (see the next section for more discussion of this point). Hence when the Chinese character network is used, the nonsense characters were closer together in representational space, and more difficult to discriminate. When the face network is used, the nonsense characters are farther apart in representational space, and hence better discriminated, as in the behavioral data (McCleery, et al., 2008).

4. The recruitment of the FFA: Solving the "visual expertise mystery"

Although it remains highly controversial whether what makes faces "special" is that they are a domain of expertise or that they are faces *per se* (e.g. Bukach, Gauthier, and Tarr, 2006; McKone, Kanwisher, and Duchaine, 2007; McKone and Robbins, this volume; Scott, this volume), it is still of interest to account for the fMRI data on expertise learning. For example, Gauthier, Skudlarski, Gore, and Anderson (2000) showed that over training in a new domain of expertise ("Greebles", a novel type of object), activity in the FFA increased. They argued that the FFA's computational role may be fine level discrimination of homogeneous categories, such as faces, and as such, it becomes recruited for other expertise tasks (i.e., the expertise hypothesis). However, it is still incumbent upon modelers invested in this hypothesis to show *why* this happens mechanistically. We have called the question of why the FFA is recruited for other objects of expertise "the visual expertise mystery," and it has been a focus of research in our lab for many years.

The main result of the model, rather surprisingly to us, is that the features useful for discriminating members of a homogeneous class, whether it be faces, cups, cans or books, are useful in discriminating members of other homogeneous classes. That is, there is nothing special about faces *per se*, rather, it is the fine-level discrimination process that encourages the development of universally useful features (Joyce and Cottrell, 2004; Sugimoto and Cottrell, 2001; Tong, Joyce, and Cottrell, 2005; Tong, Joyce, and Cottrell, 2008).

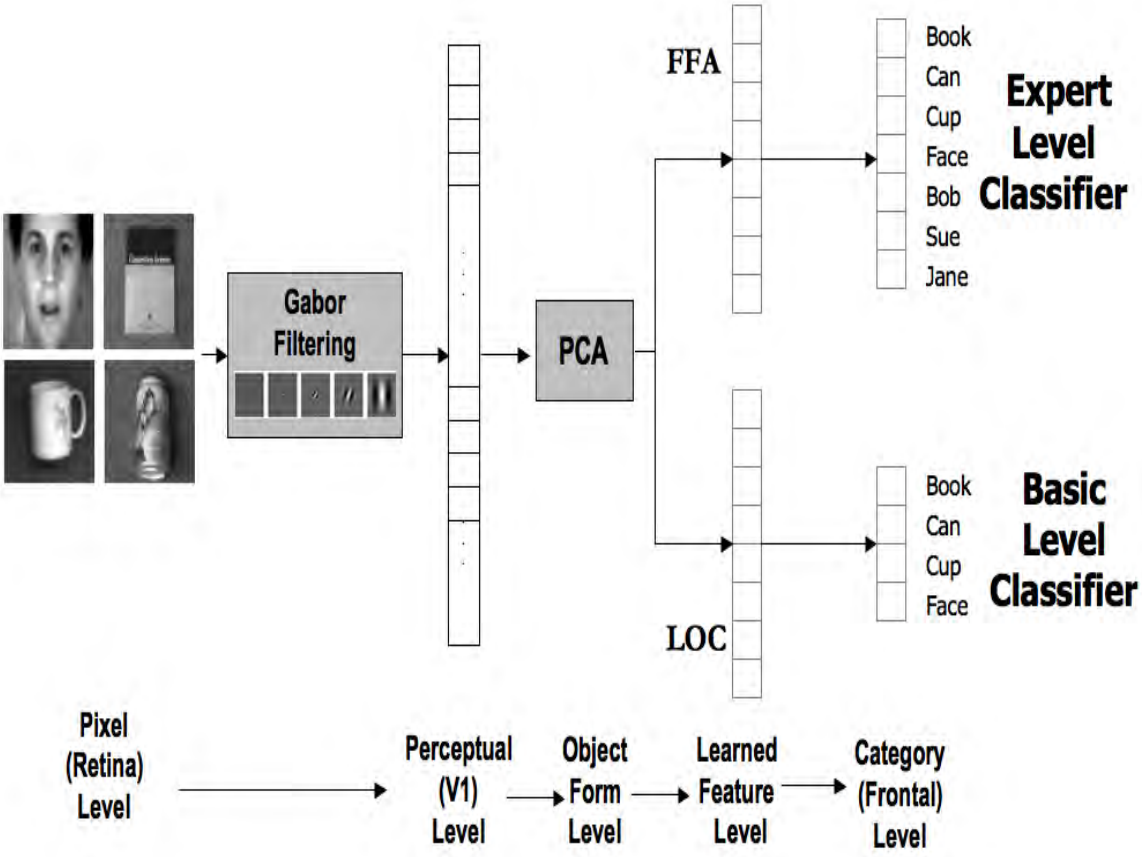
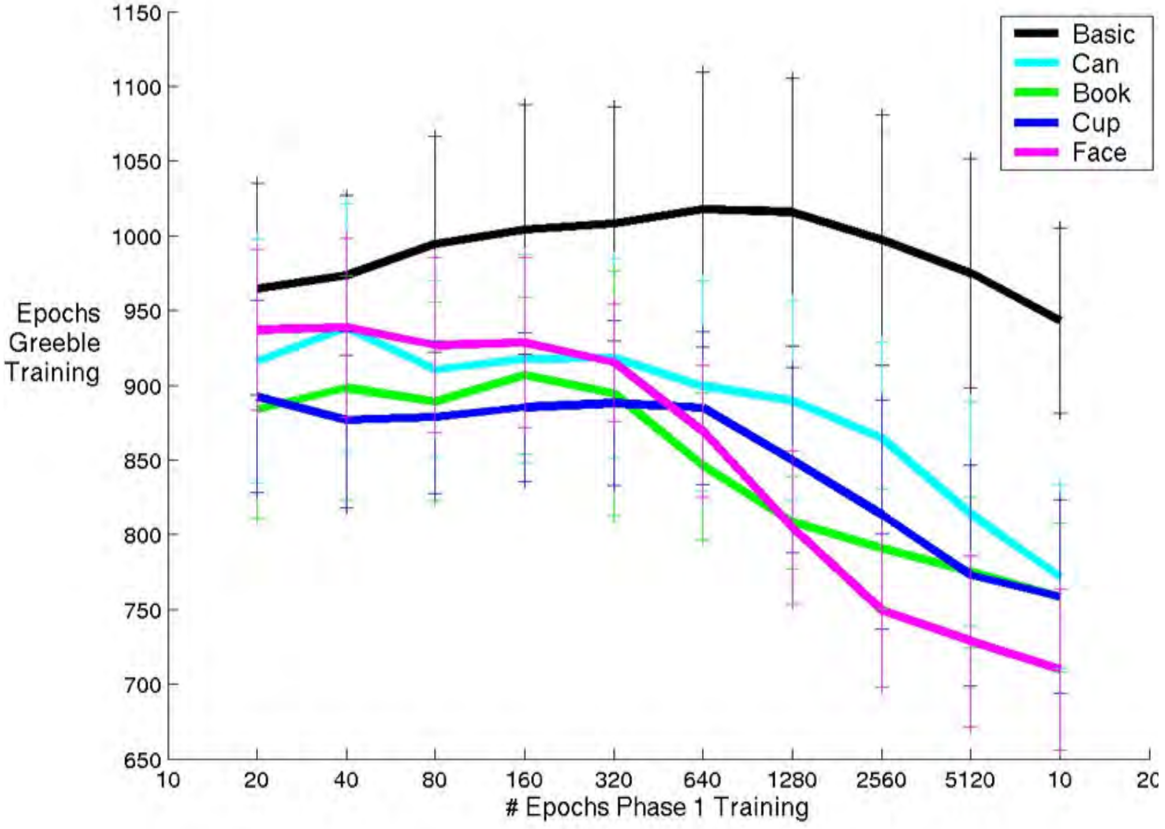
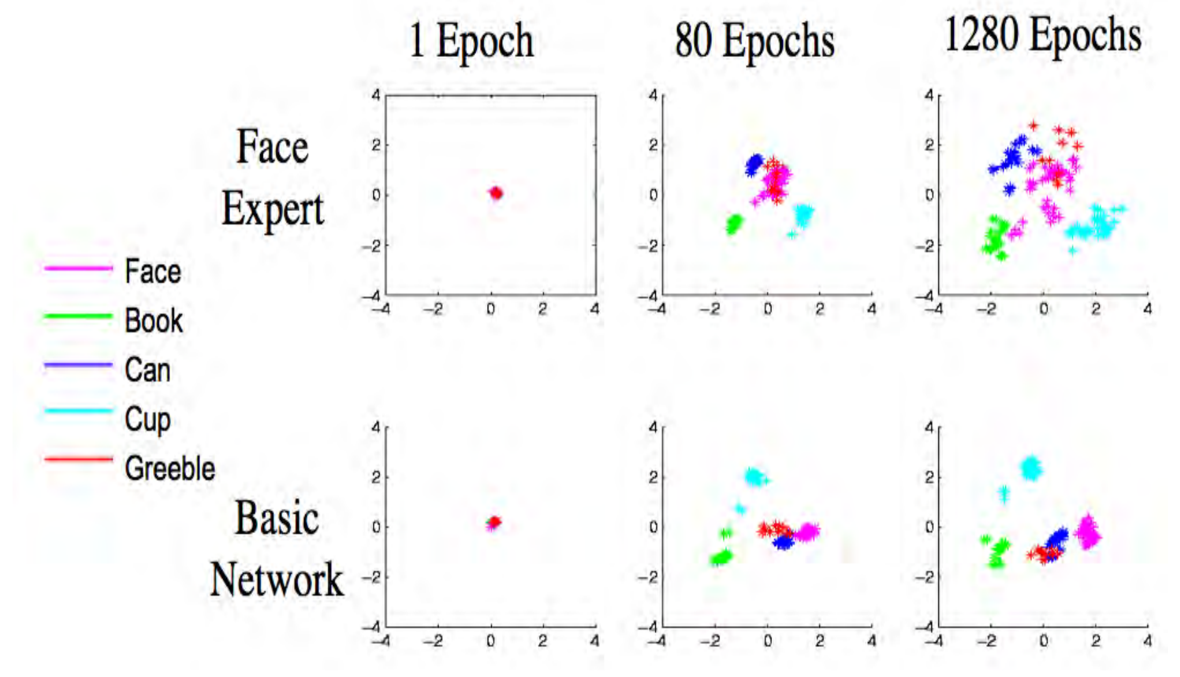


Figure 8. The expertise and basic networks. The two networks share pre-processing up to the level of the hidden units.

The basic assumption of the model is that (like the developmental model in Figure 3(b)), areas of cortex *compete* to solve tasks. The model assumes that there are two networks that share preprocessing up to the point of features specialized for a task (Figure 8). These correspond to two cortical areas that compete so that one becomes specialized for basic-level categorization (i.e. lateral occipital complex or other visual processing areas), and the other is specialized for subordinate-level (expert) categorization (i.e., the FFA). In our first simulation, similar to the model of the development of the FFA (Dailey and Cottrell, 1999; section one of this chapter), one network was trained to categorize four classes of twelve objects each (i.e., books, faces, cups, and cans) at a basic level (i.e., the basic network); the other network was trained to individuate one of the classes into its twelve identities, while continuing to simply categorize the other three classes of stimuli at the basic level (i.e., the expert network). Except for the difference in their tasks, the two networks were given identical processing resources in their hidden layers and received the same input during training. The results showed that, consistent with human data (Gauthier and Tarr, 1997), the model demonstrated the entry-level shift in reaction times, measured as the uncertainty of the output: it was just as fast for subordinate-level/expert responses as for basic-level ones (Joyce and Cottrell, 2004).



(a)



(b)

After learning the first task, the two networks were then placed in a competition to learn a new category of objects (e.g., Greebles) at the subordinate/expert level. As shown in Figure 9(a), the expert networks performed better in learning the new task than the basic network regardless of the kind of expertise that was acquired on the first task; in addition, the more training they received on the first task, the faster they learned the new task. This result suggests that in the expert networks, the features learned on the first task facilitated learning the second task. This hypothesis is confirmed by an analysis of the hidden layer representations before training on the new task. We performed PCA of the hidden layer activations over learning of the first task, and plotted the projections of the stimuli on the second and third component in Figure 9(b) (the first component just represents the growth of the connections over time). What this shows is the position in representational space of each stimulus; the coordinates can be thought of as the firing rates of two "über-units" in response to each stimulus. As can be seen from the Figure, the representations developed in the expert networks spread out the individuals in the representational space more than the basic networks did. Thus the expert networks were more sensitive to within-class variability, while the basic networks were only sensitive to between-class variability. This is because the expert networks had to differentiate among stimuli in the same class, and this lead to a representation that magnified small differences between the individuals. In contrast, the basic network needed to *ignore* differences between individuals. The red crosses shown in the Figure correspond to the Greeble stimuli projected into the space without training. It is clear that the reason that the Greebles are learned faster is that the "spreading" transform by the expert network generalizes to the Greeble stimuli. That is, the individual Greebles are already differentiated by the representation even before training on them.

A common confusion concerning this result is that it seems to suggest that becoming an expert in one domain should facilitate the development of expertise in another domain. The prediction is somewhat more subtle than this. The model concerns competitions between multiple cortical areas, and what it says is that if a cortical area is used for visual expertise, that same area will be recruited for other domains of visual expertise. At some point, the amount of representational resources within the expertise area will be exhausted, and increased expertise in one domain may actually lead to *reduced* expertise in another. Indeed, it has been reported that the FFA is difficult to locate in bird experts (Kung, et al. 2007). Of course, our results *do* suggest that being an expert in one domain facilitates acquiring expertise in a second domain. The problem is, it is difficult to test this prediction, since nearly every subject is a face expert to begin with. For the acquisition of expertise in a *third* domain, the predictions are less clear, and would depend on factors such as the amount of representational resources in the expertise area, the frequency of exposure in each domain, etc.

One criticism of this simulation was that the expert networks made more discriminations than the basic network, and this may have lead to more primitive representations in the basic network that were unable to do the expertise task. To respond to this criticism, Tong et al. (2005) conducted another simulation in which the number of output classes in the two networks was controlled instead of the input training set. They trained one network to categorize ten categories at the basic level (i.e. the basic network) and another network to categorize ten individual's faces at the individual level. Each network was trained on five images of each category or individual, but the networks' training sets were mutually exclusive. Then a novel category was used as a new expertise category that both networks had to learn to individuate. The results showed that the basic and expert networks took about the same length of time to train, showing that the tasks were now equally difficult. Nevertheless, the expert network still learned a new expert task more quickly, whether it was ships, swords, or lamps. Analysis of the network representations showed that this was for the same reason as in the previous simulation – representations were more spread out in the expertise network.

It is difficult to come up with a situation in which both the input images and the number of output categories for both basic and expert level categorization are controlled. However, Tran et al. (2004) conducted a simulation that used images of six letters in six different fonts, with letter categorization as the basic level task, while font categorization is the expert-level task. This is clearly an expertise discrimination, as similar-looking inputs (e.g., the letter "A" in six different fonts) must be separated into six different categories. After training two networks on these tasks, the networks were trained on Greebles, and the font network was faster at learning the new expertise task. This is because the expert (font) network spread out the representation of the exemplars in the new category prior to training on them, while the letter network did not, just as in the face/object network. The conclusion is that it is the *type* of distinction that matters, not the number.[2]

In summary, these simulations consistently showed that the expert network always learned the new expertise category first, even when the first category of expertise was not faces. Further analyses showed that, compared with the basic network, the expert network spread the stimuli into broader regions of representational space; this is because subordinate-level categorization requires amplification of within-category variability. Consequently, this spreading transformation generalized to new objects and thus facilitated learning within-category identification of the new objects. In other words, the expert network had a head start in differentiating new objects from one another. Tong et al.'s study (2008) thus provides a computational explanation of why the FFA may be the location of processes for fine level discrimination of homogeneous categories that can be recruited for other expertise tasks (e.g., Gauthier et al., 2000).

The expertise hypothesis itself is, of course, controversial, and several research efforts have claimed to disprove it. For example, Rhodes et al. (2004) showed that butterfly experts showed more activation outside of FFA than within FFA. However, they

[2] Or to put it another way, it is the distinction of type that matters, not the letters!

also showed that the overlap with FFA increased when subjects were performing an individuation task, whether it was for objects of expertise or not. It is not surprising that there is activation outside of FFA for non-faces – after all, these objects have different shapes than faces, and activation outside of FFA could simply reflect differences in categorical, rather than individuating, representations. The increase in overlap with FFA when performing individuation tasks, however, does suggest that the FFA is playing a role in this task. The expertise hypothesis does not claim that it is only FFA that is activated by objects of expertise, nor does it claim that the FFA will be equally activated by faces and other objects of expertise. Hence we do not see a conflict between our model and these results.

5. The Standard Model, fMRI, and face discrimination

Although we have focused (perhaps overly so) on our own models, as noted in the introduction, there are models that are more neurophysiologically plausible, such as the Riesenhuber and Poggio (1999) HMAX model (Figure 2), often called The Standard Model (TSM). Again, the reason this is called The Standard Model is that it is based on the standard interpretation in neuroscience of what neurons in V1 (simple and complex) and V2 are thought to be doing. While TSM has for the most part been applied to modeling object recognition, it has also been applied to face processing, where similar points to ours have been made. For example, in (Jiang et al., 2006), they show that the model does not require any special mechanism for face processing – it is just another form of object recognition. The difference between faces and objects in TSM is a quantitative one: Units tuned to faces have more narrow tuning profiles than object-tuned units, due to the requirement to differentiate them from one another, leading to a sparse representation of faces (in this context, *sparse* means that few units are activated by any face).

Jiang et al. use TSM to model a face discrimination experiment in which faces vary either by configural or featural changes in the “different” condition. Complementary to our model, where holistic representations were shown to be sufficient to discriminate featural changes (Cottrell, et al., 2002), Jiang et al. were concerned with whether a model based on features could show sensitivity to configural differences in faces. The model is then used to make predictions concerning an fMRI experiment with facial morphs, which were confirmed by human experiments.

Jiang et al. model the Fusiform Face Area with 180 View Tuned Units (VTUs, see Figure 2). These units are created by presenting the model with 180 training faces, and a radial basis function[3] unit is created that is tuned to each face by setting its weight vector to the pattern of activation on the C2 units. The set of faces used in the behavioral experiment is separate from the ones used to create the VTUs. These faces produce a pattern of activation across the 180 VTUs, depending on how similar the faces are to the training faces. As in TM, in order to model a discrimination experiment in a model that processes one face at a time, they assume that subjects store the activity pattern of the faces engendered on the VTUs, and compute the Euclidean distance between them. Unlike our model, one of the parameters of TSM is the number of elements of the activity pattern that are stored. Using only the most active units is seen as an efficient way to use memory. Also, since the model would never make errors when shown the same face image twice, noise is added to the units' activities, so that the model is capable of making mistakes on "same" trials. The standard deviation of this noise is one of the parameters of the model.

In order to fit the model to the human discrimination data, Jiang et al. use a brute force search through parameter space. The parameters are 1) the number of afferents to the face units from the C2 units; 2) the standard deviation of the face units; 3) the number of unit activations stored in memory for face discrimination tasks; 4) the size of the noise variance used in storing the first face responses; and 5) the threshold for determining "same." The first two parameters constrain the tuning specificity of the face units – a small standard deviation results in tighter tuning, whereas a smaller number of afferents results in broader tuning. They find 35 sets of parameters that produce data that are not significantly different from the subject data, and the 35 networks' data are within the error bars of the human data for featural changes, configural changes, and inverted faces.

[3] A radial basis function unit is a unit with a circular Gaussian response function; hence any deviation from its mean in any direction reduces the response only based on the distance from the mean. The mean for each unit is set to the responses of units at the previous layer (C2 units) for that unit's face. Tuning is set by the standard deviation of the Gaussian and by setting some components of the mean to 0.

This procedure achieves their goal of showing that their feature-based model is capable of producing data consistent with subject behavior with configurally-manipulated faces. The results are telling: the model predicts that face units will be relatively tightly tuned, with standard deviations of about 0.1, with connections to less than half the C2 units. Of note, the number of stored activations for comparing two faces comes out to be less than 10, suggesting only a very sparse representation of a face needs to be stored in memory in a discrimination task. To show this is a face-specific result, a model fit to an additional experiment using cars as stimuli for the human subjects required more broadly tuned units. Note that this can be seen as a similar story to the one we told with our model above: a small number of highly specific units fire for each face, which means that they will be spread out in representational space, whereas for a category like cars, broadly-tuned units will lead to a similar set of units being activated for most cars.

Since the model was directly fit to the data, it is important to show that the model "as is" can predict new data. Hence, to validate the model, Jiang et al. show that the same 35 networks with no further tuning can predict fMRI Rapid Adaptation (fMRI-RA) data on a completely different set of faces. The model and the human subjects are presented with morphed faces of varying distances from a set of prototypes. In a rapid adaptation paradigm, the BOLD response to one image is compared to the BOLD response to a different image presented immediately after the first. If the BOLD response adapts, that is, it decreases from the presentation of the first image, this suggests that the two images are represented by a similar population of neurons. If the BOLD response recovers, then this suggests that they use a different population of neurons. The model predicted at what point the difference in BOLD adaptation response would asymptote. In the model, at some point, a completely separate population of units are activated by the two faces, so no further adaptation can occur. This prediction was confirmed. Importantly, this prediction did *not* follow from parameter sets that did not fit the behavioral data. Finally, they also performed a face discrimination task on the faces used in the scanner, and showed that the model's correlation with the human discrimination performance was significantly better when only the maximally activated units (less than 10) were used in the comparisons, as opposed to using all of the 180 unit activations. This finding supports the sparse representation prediction of the model.

6. Summary and conclusions

As noted at the outset, computational modeling is an important tool in trying to understand the nature of face representation and processing. As we have shown in this chapter, models can be manipulated in ways people cannot, in order to shed light on such questions as whether faces are special. We have shown here that, according to our model, at least, faces are only special to the extent that they are a class of stimuli for which we must discriminate individual members. This fine-level discrimination task requires transformations that separate representations of similar objects, and at least in our model, this transformation generalizes to novel objects. Hence when someone learns a new domain of expertise, this region of cortex is used to bootstrap the process of discriminating the new class.

Models can also be analyzed in ways that (currently) human brains cannot – down to the single unit level. This kind of analysis can lead to predictions that are difficult to conceive of in the absence of a concrete model. An example from this chapter is the way in which Jiang et al. (2006) used their model to compare the behavioral predictions of sparse versus dense representations to suggest that representations in FFA are sparse. Another example is the prediction from the Dailey et al. (2002) EMPATH model that the similarity structure of facial expressions can be captured by the similarity of categorical representations.

We also discussed two ways in which models differ, in their degree of realism and their methods of parameter setting. The models we focused on the most here, TM and TSM, differ considerably along these dimensions, with TSM being more neurally realistic, while requiring a search for parameters to fit the data, and TM being less realistic, but using learning on a realistic task to set the parameters. While the use of backpropagation generally leads to the criticism of biological implausibility, more biologically plausible techniques for learning, such as those used in Leabra (O'Reilly and Munakata, 2000) should eventually lead to models that are biologically plausible in both their learning methods and their neural processing mechanisms.

One future direction, then, might be to first use the Leabra model to replicate many of the results in this chapter, extending the modeling effort to include interactions between the dorsal pathway and the ventral pathway. Since the dorsal pathway is believed by many to include a salience map, this approach could therefore add new dimensions to the modeling effort by explicating the role of attention in object and face recognition. This leads to a second observation and direction for future research: the models described here are all wrong in a fundamental way, in that they do not sample the environment with the eye movements that are required by a foveated retina. Indeed, eye movements have been shown to be functional in face recognition (Hsiao & Cottrell, 2009). Since we make about three eye movements per second, or about 173,000 per day, it seems important to take this factor into consideration when modeling human perception. We look forward to new research that attempts to understand object and face perception in the light of such constraints.

References

Ambady, N. and Weisbuch, M. (this volume).

Brady, N., Campbell, M., and Flaherty, M. (2005). Perceptual asymmetries are preserved in memory for highly familiar faces of self and friend. *Brain and Cognition*, *58*, 334-342.

Bruce, V., and Young, A. (1986). Understanding face recognition. *British Journal of Psychology*, **77**, 305-327.

Bukach, C. M., Gauthier, I., and Tarr, M. J. (2006). Beyond faces and modularity: the power of an expertise framework. *Trends in Cognitive Sciences*, **10**:159-166.

Buonomano, Dean V. and Merzenich, Michael M. (1998). Cortical plasticity: From synapses to maps. *Annual Review of Neuroscience* **21**:149–186.

Burt, D. M. and Perrett, D. I. (1997). Perceptual asymmetries in judgments of facial attractiveness, age, gender, speech and expression. *Neuropsychologia*, **35**:685-693.

Cadieu, C., Kouh, M., Pasupathy, A., Conner, C., Riesenhuber, M., and Poggio, T.A. (2007). A Model of V4 Shape Selectivity and Invariance. *J Neurophysiol* **98**: 1733-1750.

Calder, A.J. (this volume) Interpreting facial expressions. In A. Calder, G. Rhodes, J. V. Haxby, and M. Johnson (Eds.), *Handbook of Face Perception.* Oxford University Press, Oxford: UK.

Calder, A.J., Rowland, D., Young, A.W., Nimmo-Smith, I., Keane, J. and Perrett, D.I.

(2000a) Caricaturing facial expressions. *Cognition*, **76**, 105-46.

Calder, A.J., Young, A., Keane, J., and Dean, M. (2000b). Configural information in facial perception. *Journal of Experimental Psychology: Human Perception and Performance*, **26**(2):527–551.

Calder, A., Young, A., Perrett, D., Etcoff, N., and Rowland, D. (1996). Categorical perception of morphed facial expressions. *Visual Cognition*, **3**:81–117.

Cottrell, G.W., and Metcalfe, J. (1991). EMPATH: Face, gender and emotion recognition using holons. In R.P. Lippman, J. Moody, and D.S. Touretzky (Eds.), *Advances in Neural Information Processing Systems Vol. 3* (pp. 564-571), Morgan Kaufmann: San Mateo.

Cottrell, G.W., Branson, K. and Calder, A.J. (2002) Do expression and identity need separate representations? In *Proceedings of the 24th Annual Cognitive Science Conference*, Fairfax, Virginia, Mahwah: Lawrence Erlbaum, pp: 238-243.

Dailey, M.N., and Cottrell, G.W. (1999). Organization of face and object recognition in modular neural network models. *Neural Networks*, *12*, 1053-1073.

Dailey, M. N., Cottrell, G. W., Padgett, C., and Adolphs, R. (2002). EMPATH: A neural network that categorizes facial expressions. *Journal of Cognitive Neuroscience*, *14*, 1158–1173.

Dailey, M.N., Joyce, C.A., Lyons, M.J., Kamachi, M., Ishi, H., Gyoba, J. and Cottrell, G.W. (2010). Evidence and a computational explanation of cultural differences in facial expression recognition. *Emotion.*(in press).

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.

Ekman, P. and Friesen, W. (1976). Pictures of Facial Affect. Consulting Psychologists, Palo Alto, CA.

Elfenbein, H., and Ambady, N. (2002). On the universality and cultural specificity of emotion recognition: A meta-analysis. Psychological Bulletin, 128(2), 203–235.

Fleming, M., and Cottrell, G.W. (1990). Categorization of faces using unsupervised feature extraction. *Proceedings of the International Joint Conference on Neural Networks*, *2*, 65-70.

Gauthier, I., Skudlarski, P., Gore, J.C., and Anderson, A.W. (2000). Expertise for cars and birds recruits brain areas involved in face recognition. *Nature Neuroscience*, *3(2)*, 191–197.

Gauthier, I., and Tarr, M. J. (1997). Becoming a "greeble" expert: Exploring mechanisms for face recognition. *Vision Research*, *37*, 1673-1682.

Gauthier, I. and Tarr, M. J. (2002) Unraveling mechanisms for expert object recognition: bridging brain activity and behavior. *Journal of Experimental Psychology: Human Perception and Performance*, *28*, 431-446.

Gauthier, I., Tarr, M. J., Anderson, A. W., Skudlarski, P., and Gore, JC. (1999). Activation of the middle fusiform "face area" increases with expertise in recognizing novel objects. *Nature Neuroscience*, *2*, 568-573.

Gauthier, I., Williams, P., Tarr, M. J., and Tanaka, J. (1998). Training "Greeble" experts: A framework for studying expert object recognition processes. *Vision Research*, *38*, 2401-2428.

Gilbert, C. and Bakan, P. (1973). Visual asymmetry in perception of faces. *Neuropsychologia*, *11*, 355-362.

Haque, AfmZakaria, and Cottrell, G.W. (2005). Modeling the Other Race Advantage with PCA. In *Proceedings of the 27th Annual Cognitive Science Conference,* La Stresa, Italy. Mahwah: Lawrence Erlbaum, pp. 899-904

Hsiao, J. H. and Cottrell, G. W. (2009). Not all visual expertise is holistic, but it may be leftist: The case of Chinese character recognition. *Psychological Science* **20**(4):455-463.

Hsiao, J. H., Shieh, D. X., and Cottrell, G. W. (2008). Convergence of the visual field split: Hemispheric modeling of face and object recognition. *Journal of Cognitive Neuroscience*, **22**(12):2298-2307.

Jiang, X., Rosen, E., Zeffiro, T., VanMeter, J., Blanz, V., and Riesenhuber, M. (2006) Evaluation of a shape-based model of human face discrimination using fMRI and behavioral techniques. *Neuron* **50**:159–172.

Jenkins, W.M.; Merzenich, M. M. and M.T. Ochs (1987), Behaviorally controlled differential use of restricted hand surfaces induce changes in the cortical representation of the hand in area 3b of adult owl monkeys., *Soc. Neurosci.* Abstract

10:665

Johnson (this volume) Face Perception: A Developmental Perspective.

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

Joyce, C., and Cottrell, G. W. (2004). Solving the visual expertise mystery In H. Bowman and Labiouse, C. (Eds.), *Connectionist models of cognition and perception II: Proceedings of the eighth neural computation and psychology workshop*. World Scientific.

Kanwisher, N., McDermott, J. and Chun, M. M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience*, *17*, 4302-4311.

Kung, C-C., Ellis, C., and Tarr, M. J. (2007) Dynamic Reorganization of fusiform gyrus: long-term bird expertise reduces face selectivity. Poster presented at the *Cognitive Neuroscience Society (CNS) Annual Meeting*, New York, NY.

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

Lake, Brenden M., and Cottrell, G.W. (2005). Age of Acquisition in Facial Identification: A connectionist approach. In *Proceedings of the 27th Annual Cognitive Science Conference,* La Stresa, Italy. Mahwah: Lawrence Erlbaum, pp. 1236-1241.

Levin, D. T. (2000). Race as a visual feature: using visual search and perceptual discrimination asks to understand face categories and the cross-race recognition deficit, Journal of Experimental Psychology: General 2000 vol. 129, No. 4.

McCleery, J.P., Zhang, L., Ge, L., Wang, Z., Christiansen, E.M., Lee, K. and Cottrell, G.W. (2008) The roles of visual expertise and visual input in the face inversion effect: Behavioral and neurocomputational evidence. *Vision Research* 48:703-715.

McKone, E., Kanwisher, N., and Duchaine, B. C. (2007). Can generic expertise explain special processing for faces? *Trends in Cognitive Sciences*, *11(1)*, 8-15.

McKone, E., and Robbins, R. (this volume) Are faces special?

Meadows, J.C. (1974). The anatomical basis of prosopagnosia. *J Neurol Neurosurg*

Psychiatry **37**:489-501.

O'Reilly, R. and Munakata, Y. (2000) *Computational Explorations in Cognitive Neuroscience*. MIT Press: Cambridge, MA.

O'Toole, A. J. (this volume). Cognitive and Computational Approaches to Face Perception. In A. Calder, G. Rhodes, J. V. Haxby, and M. Johnson (Eds.), *Handbook of Face Perception.* Oxford University Press: Oxford, UK.

Palmeri, T. J., and Gauthier, I. (2004). Visual object understanding. *Nature Reviews Neuroscience*, *5*, 291-303.

Phillips, P. J., Moon, H., Rauss, P. J., and Rizvi, S. (2000). The FERET evaluation methodology for face recognition algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*, 1090-1104.

Rhodes, Gillian, Byatt, Graham, Michie, Patricia T., and Puce, Aina (2004) Is the Fusiform Face Area Specialized for Faces, Individuation, or Expert Individuation? *Journal of Cognitive Neuroscience* 16(2):189-203.

Riesenhuber, M. and Poggio, T. (1999). Hierarchical Models of Object Recognition in Cortex. *Nature Neuroscience*, *2*, 1019-1025.

Rossion, B., Joyce, C. A., Cottrell, G. W., and Tarr, M. J. (2003). Early lateralization and orientation tuning for face, word, and object processing in the visual cortex. *Neuroimage*, *20*, 1609-1624.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, **39**:1161–1178.

Scherf, K. S., Behrmann, M., Humphreys, K. and Luna, B. (2007). Visual category-selectivity for faces, places, and objects emerges along different developmental trajectories. *Developmental Science*, **10**(4):F15-F30.

Scott, L.S. (this volume). Face perception and perceptual expertise in adult and developmental populations.

Sergent, J. (1982) The Cerebral Balance of Power: Confrontation or Cooperation? *J. Exp. Psychol. Hum. Percept. Perform*. **8**, 253-272.

Sugimoto, M., and Cottrell, G.W. (2001). Visual Expertise is a General Skill. In *Proceedings of the 23rd Annual Cognitive Science Conference*, pp. 994-999. Mahwah: Lawrence Erlbaum.

Tanaka,J.W. and Farah, M.J. (1993). Parts and wholes in face recognition. *Quarterly Journal of Experimental Psychology,* 46A, 225-245.

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

Tong, M. H., Joyce, C. A. and W. Cottrell, G. W. (2005). Are Greebles special? Or, why the Fusiform Fish Area (if we had one) would be recruited for sword expertise. *Proceedings of the 27th Annual Cognitive Science Conference*. Mahwah: Lawrence Erlbaum.

Tong, M. H., Joyce, C. A., and Cottrell, G. W. (2008). Why is the fusiform face area recruited for novel categories of expertise? A neurocomputational investigation. *Brain Research*, *1202*, 14-24.

Tootell, R. B. H., Mendola, J. D., Hadjikhani, N. K., Liu, A. K., and Dale, A. M. (1998). The representation of the ipsilateral visual field in human cerebral cortex. *The Proceedings of the National Academy of Sciences USA*, *95*, 818–824.

Tran, B., Joyce, C. A., and Cottrell, G. W. (2004). Visual expertise depends on how you slice the space. *Proceedings of the 26th Annual Conference of the Cognitive Science Conference.* Mahwah, NJ: Lawrence Erlbaum.

Zhang, Lingyun and Cottrell, Garrison (2006) Look Ma! No network: PCA of Gabor filters models the development of face discrimination. In *Proceedings of the 28th Annual Cognitive Science Conference*, Vancouver, BC, Canada. Mahwah: Lawrence Erlbaum.

Zhang, Lingyun, Tong, Matthew H., and Cottrell, Garrison W. (2007) Information attracts attention: A probabilistic account of the cross-race advantage in visual search. In *Proceedings of the 29th Annual Cognitive Science Society Meeting*, Nashville, TN. Mahwah: Lawrence Erlbaum.