Look Ma! No Network!: PCA of Gabor Filters Models the Development of Face Discrimination

Lingyun Zhang and **Garrison W. Cottrell**
lingyun,gary@cs.ucsd.edu
UCSD Computer Science and Engineering
9500 Gilman Dr., La Jolla, CA 92093-0114 USA

Abstract

In previous work, we modeled Mondloch's behavioral data [Mondloch et al., 2002] on adult face discrimination [Zhang and Cottrell, 2004]. We found that our standard model [Dailey and Cottrell, 1999, Dailey et al., 2002] was overly holistic and that by adding local feature processing we could qualitatively match the human data. However, further investigation has lead us to reconsider our conclusions. In particular, we have found that the form of the Gabor filters we used were not biologically realistic, in that the spatial frequency range was overly limited. In this paper, we show that biologically realistic Gabor filters allow us to match the adult data using only the first two processing steps of our model. No neural network is necessary. In addition, we find that we can also model the developmental data 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 are sufficient to account for the slower development of configural processing compared to featural processing, and how the child's performance improves with age.

Introduction

We have developed a model of visual categorization that accounts for a number of important phenomena in face and object processing and visual expertise [Dailey and Cottrell, 1999, Cottrell et al., 2002, Dailey et al., 2002, Joyce and Cottrell, 2004, Tran et al., 2004, Tong et al., 2005]. Here, we investigate the model's ability to account for human sensitivity to variations in faces that are considered theoretically important for face identification. 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. Configural processing means that subjects are sensitive to the relationships between the parts, e.g., the spacing between the eyes. Holistic processing can easily be captured by a model that uses whole-face template-like representations as ours does: interference from incongruent halves of a face occurs when making judgments about a different part (e.g, expression on top when a different expression is on bottom [Cottrell et al., 2002]). However, configural effects related to spacing information are attenuated by the alignment procedure that we typically use, which warps the image so the eyes and mouth are always in the same three positions.

Diamond and Carey [Diamond and Carey, 1986] were among the first to discriminate between the types of processing involved in face/object perception and recognition. Based on studies looking at the inversion effect to faces, landscapes and dogs in both dog novices and dog experts, they proposed that first-order relational information, which consists of the coarse spatial relationships between the parts of an object (i.e. eyes are above the nose), is sufficient to recognize most objects. By contrast, second-order relational information, which is needed for face recognition and recognition of individuals within categories of expertise, is reserved for visually homogeneous categories where slight differences in configuration must be used to distinguish between individuals (e.g. a slight change in the distance between the eyes and the nose). Diamond and Carey [Diamond and Carey, 1986] suggest that experience allows people to develop a fine-tuned prototype and to become sensitive to second-order differences between that prototype and new members of that category (e.g. new faces).

One implication of the Diamond and Carey study is that the inversion effect (a large reduction in same/different performance on inverted faces, compared to inverted objects) is based on a relative reliance on second-order relational information, and that perhaps this characteristic distinguishes face/expert-level processing from regular object recognition. Farah et al. [Farah et al., 1995] found that encouraging part-based processing eliminated the inversion effect, whereas allowing/encouraging non-part-based processing resulted in a robust inversion effect. Thus Farah et al. conclude that the inversion effect, in faces and other types of stimuli, is associated with holistic pattern perception.

However, this emphasis on holistic and configural processing has led to less consideration of the obvious fact that subjects are also quite sensitive to changes in the features themselves – substitutions of different eyes or mouths can make the face look quite different. The Thatcher illusion [Thompson, 1980] suggests that parts are processed somewhat independently, and only loosely connected to the representation of the whole face. A study by Mondloch et al. (2002) that varied these different aspects of a face (configuration, feature changes, and changes to contour of the face) found differing levels of sensitivity to the type of manipulation in a same/different paradigm (the stimuli are shown in Figure 1). Importantly, they investigated sensitivity to these manipulations in adults and children of three different age levels in order to investigate how face processing changes over development. While the manipulations were not performed parametrically (no equating of the difficulty of discrimination was performed), but in a rather ad hoc manner, the results are consistent across subjects. Hence this is a crucial set of data to account for with our model.

In our previous work [Zhang and Cottrell, 2004], we made a first attempt at modeling the adult data. We followed the model in [Dailey et al., 2002] and found that our model was overly holistic. I.e. the human adults found the featural set more discriminable than the configural set while our model found the opposite. In that work, we introduced a representation of the important parts of the face (eyes and mouth) to the model and found that only a relatively small amount of holistic representation, compared to parts representations, was necessary to account for the data. However, further investigation has lead us to reconsider our conclusions. In particular, we have found that the form of the Gabor filters we used was not biologically realistic, in that the spatial frequency range was overly limited.

Here we will show that with biologically realistic Gabor filters, our model can match the adult data using only the first two processing steps, i.e. no neural network classifier is necessary. Furthermore, we found that our model can account for the developmental data as well, using two very simple manipulations of the preprocessing. We hypothesized that sensitivity to configuration might just be a consequence of how many people one has to distinguish. When you only know a few people, featural differences may be sufficient, but as you get to know more people, you may need to be sensitive to configural differences as well. Secondly, we hypothesized that as subjects mature, they may allocate more processing resources to the task of representing faces. These two variables are indeed sufficient for accounting broadly for changes over development.

Mondloch's Stimuli and Experiments

Mondloch et al. [Mondloch et al., 2002] began with a single face (called Jane) and modified it to create twelve new versions (called Jane's Sisters). These were divided to three sets of stimuli: a configural set, a featural set, and a contour set. The four faces in the configural set were created by moving the eyes and/or the mouth. The four faces in the featural set were created by replacing Jane's eyes, nose and mouth with those of four different females. The four faces in the contour set were created by pasting the internal portion of Jane's face within the outer contour of four different females. The control stimuli were called "cousins" and consisted of three different female faces (Figure 1).

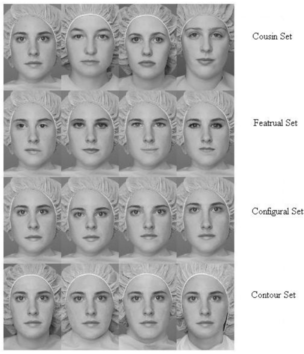

Figure 1: The four sets of Jane stimuli generated by Mondloch et al. Jane is the leftmost face on the top row. The four rows are the cousin set, featural set, configural set and contour set respectively from top to bottom. (Adapted from [Mondloch et al., 2002])

These stimuli were presented to 6, 8 and 10-year-old children as well as adults in a series of same-different trials. One face appeared for 200ms. After a 300ms interval, the second one appeared until the participant responded. There were also trials in which upside down versions of these faces were presented. Figure 2 shows the performance results.

In earlier work [Zhang and Cottrell, 2004], we concentrated on modeling the adult data, and hence focused on the black bars in Figure 2. The results showed that when stimuli were presented upright, the relative accuracy for adults in each set of stimuli was $cousin > featural > configural > contour$. This is interesting because it suggests that, at least for this stimulus set, subjects were more sensitive to individual feature differences than to configural changes.

In this work, we extended our focus to also model the developmental data in the upright situation, i.e. the left panel in Figure 2. The human data showed that for children, the relative accuracy is $cousin > featural > contour > configural$. I.e. the rank among the cousin, featural and contour sets do not change with the age, but the relative accuracy of the configural increases from the worst in children to the third in adults. Mondloch et al. concluded that configural face processing develops more slowly than featural face processing.

A Computational Model of Face Recognition

Our model is a three level neural network that has been used in previous work (Figure 3). The model takes manually aligned face images as input. The images are first filtered by 2D Gabor wavelet filters. PCA (principal component analysis) is then used to extract a set of features from the high dimensional data. In the last stage, a simple back propagation network is used to assign a name to each face. We now describe each of the components of the model in more detail.

Perceptual Layer

Research suggests that the receptive fields of the striate neurons are restricted to small regions of space, responding to narrow ranges of stimulus orientation and spatial

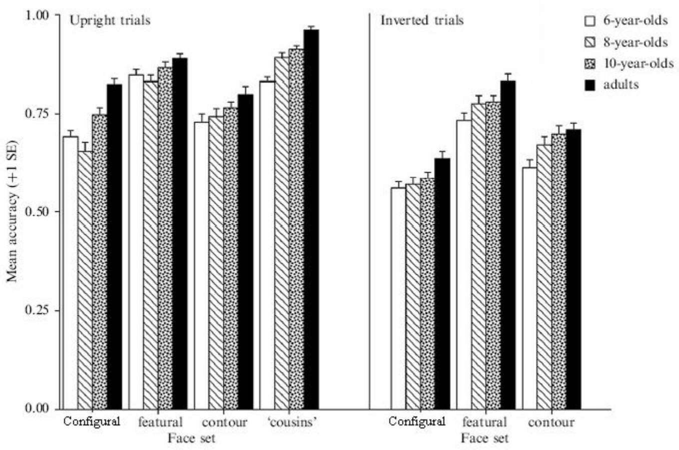

Figure 2: Mean accuracy for each face set and each age group when stimuli were presented upright (left panel) and inverted (right panel). (Adapted from [Mondloch et al., 2002])

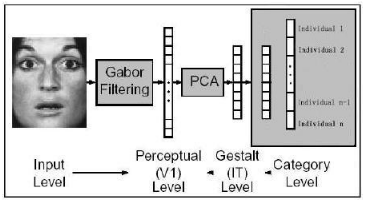

Figure 3: Our standard model. From left to right are the input level (aligned face images), perceptual level (Gabor filtering), gestalt level (PCA) and the category level (two layer neural network). (Adapted from [Dailey et al., 2002])

frequency [Jones and Palmer, 1987]. Two-D Gabor filters [Daugman, 1985] have been found to fit the 2D spatial response profile of simple cells quite well. In this processing step the image was filtered with overlapping 2-D Gabor filters in quadrature pairs at five scales and eight orientations.

In the earlier work, we used gabor filters following [Dailey et al., 2002] (Figure 4, upper row). In this work, we used more biologically realistic filters following [Dailey and Cottrell, 1999] and [Hofmann et al., 1998] (Figure 4, lower row), where the parameters are based on those reported in [Jones and Palmer, 1987], to be representative of real cortical cell receptive fields. The basic kernel function is:

$$G(\vec{k},\vec{x}) = \exp i\vec{k}\cdot\vec{x}\exp\left(-\frac{k^2\vec{x}\cdot\vec{x}}{2\sigma^2}\right) \tag{1}$$

where

$$\vec{k} = [k\cos\phi, k\sin\phi] \tag{2}$$

and $k = |\vec{k}|$ controls the spatial frequency of the filter function G. $\vec{x}$ is a point in the plane relative ot the wavelet's origin. ϕ is the angular orientation of the filter, and σ is a constant. Here, $\sigma = \pi$, ϕ ranges over $\{0, \frac{\pi}{8}, \frac{\pi}{4}, \frac{3\pi}{8}, \frac{\pi}{2}, \frac{5\pi}{8}, \frac{3\pi}{4}, \frac{7\pi}{8}\}$, and

$$k_i = \frac{2\pi}{N}2^i \tag{3}$$

where N is the image width and i an integer. We used 5 scales here with $i \in \{1, 2, 3, 4, 5\}$.

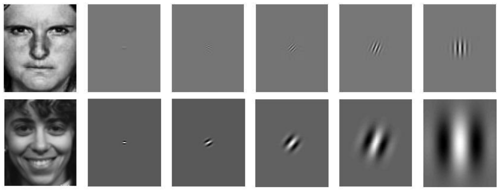

Figure 4: Examples of Gabor filters used in the earlier work [Zhang and Cottrell, 2004] (upper row) and current work (lower row) which is biologically realistic.

Gestalt layer

In this stage we perform a PCA of the Gabor filter responses. This is a biologically plausible means of dimensionality reduction[Dailey et al., 2002], since it can be learned in a Hebbian manner. PCA extracts a small set of informative features from the high dimensional output of the last perceptual stage. The eigenvectors of the covariance matrix of the patterns are computed, and the patterns are then projected onto the eigenvectors associated with the largest eigenvalues. At this stage, we produce a 50-element PCA representation from the Gabor vectors. Before being fed to the final classifier, the principal component projections are shifted and scaled so that they have 0 mean and unit standard deviation, known as z-scoring (or whitening).

Categorization layer

The classification portion of the model is a two-layer back-propagation neural network. We will show in this work that we do not need this layer to account for the human data. The representation at the Gestalt layer already matches the adult data qualitatively and that of the developmental data.

Modeling Mondloch et al.

The Training Set

The FERET database is a large database of facial images, which is now standard for face recognition from still images[Phillips et al., 1998]. We used 653 face images (539 upright images of 117 individuals and 114 inverted images of 20 individuals (that were also included in the upright images)) as the training set. The inverted faces were used in order to give a reasonable representation of upside down faces in the PCA layer of the network. The PC components are learned over Gabor filter outputs of the training set. These components are then frozen for use on the Jane images. In [Dailey et al., 2002], where the task was to learn facial expressions, images were aligned so that eyes and mouth went to designated coordinates. This alignment removed the configural information which is crucial for our work because we are trying to understand how configural processing and featural processing interact with each other in the face recognition task. To avoid this negative effect, we required that the relative spacing between the parts of the face remain the same. Thus, we *triangularly aligned* the face images. The face images were rotated, scaled and translated so that the sum of square distance between the target coordinates and those of the transferred features (eyes and mouth locations) was minimized. I.e. if the coordinates of the eyes and mouth are $re; le; m$ and the target coordinates are $(tre; tle; tm$, we minimize $||re - tre||^2 + ||le - tle||^2 + ||m - tm||^2$. Thus, a triangle represented by the eyes and mouth is scaled and moved to fit closely to a reference triangle, but the triangle is not warped. This way of alignment keeps configural information without affecting holistic processing.

Modeling Discrimination

In our earlier work, we regarded the hidden layer representation as internal representation and examined the discriminability at this level. In this work, we simply look at the PCA level, which simplifies the structure and does not vary from the random initialization of the classifier.

To model discriminability between two images, we present an image to the network, and record the PCA vector. We do the same with a second image. We model similarity as the correlation between the two representations, and discriminability as one minus similarity.

$$discrimination = 1 - correlation(image1, image2) \quad (4)$$

The PCA responses are whitened over images, i.e. every PC component is of zero mean and unit variance over images so that their contributions to the discriminability are normalized. Note that this measure may be computed at *any* layer of the network. We computed the average discriminability between images in each of the stimuli sets (featural, configural, etc., both upright and inverted). The average within each set was taken as the measure of the model's ability to discriminate each set.

Modeling the Adults' Data

We first revisited what had been the focus of the earlier work, i.e. modeling the adult data. In the earlier work we found that our model was overly holistic, i.e. it found the configural set more discriminable than the featural set. We compensated for this by introducing a representation of the face features (eyes and mouth). Now, with the biologically realistic Gabor filters, our model fits the adult data well. Figure 5 shows the rank of discriminability of the Jane's 4 sets at the PCA level. In the upright situation, the rank is $cousin > featural > contour > configural$. In the inverse situation, the rank is $cousin > featural > contour > configural$ where the contour is only slightly more discriminable than the configural set. Every set suffers a decrease in the discriminability when inverted, but the configural set suffers most and becomes the least discriminable of all. This fits the adult data (the black bars in Figure 2) qualitatively. Also note that the rank does not change when the number of PC changes, except for when the number of PC is very small. I.e. the discriminability rank of the Jane's sets is robust to the number of PC components used in our model.

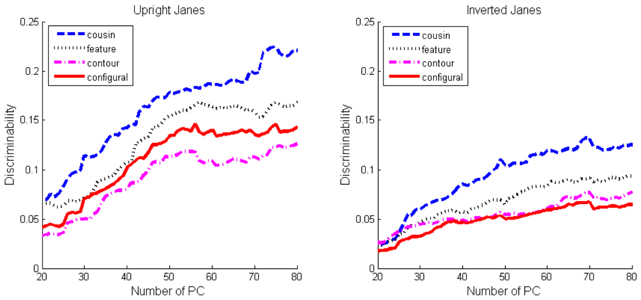

Figure 5: The discriminability of the Jane's sets at the PCA level when presented upright (left panel) and inverted (right panel).

Modeling the Developmental Data

In this section we will further explore the developmental data which was not considered in our earlier work. In this section, we examine the effects of the number of training images and the number of PC components on the discriminability respectively. The increase of the number of images along with the increase of the number of individuals models the fact that people get to know more and more people and get to see more and more faces. The increase of the number of PC components models the idea that more neural resources are allocated for face processing over development. We also discuss the effects of the frequencies of the Gabor filters and their possible contribution to development.

Manipulating the Number of Training Images A typical human will start to see faces soon after he is born, and throughout his life, he will know more and more people and see more and more faces. We think this ever growing experience with faces is one of the important factors playing a role in development. To understand how this might affect face processing, we examine how the increase of the number of individuals that the model "sees" changes its behavior. Figure 6 shows the discriminability of the Jane's sets in the upright situation when using from 100 images (19 individuals) to 500 images (107 individuals) at step of 100 images, all upright. As the figure shows, when only 100 training images are used, the rank of the discriminability of the Jane's sets is $cousin > featural > contour > configural$. As the number of images increase, the relative discriminability of the configural set increases. When more than 300 images are used, the rank starts to match that of the adults. This qualitatively matches the human data that configural processing develops more slowly than featural face processing.

Manipulating the Number of PC Components Besides seeing more and more faces, we believe that another important factor in child's maturation is brain development and consequently, more neural resources could be used for face processing. To understand this effect, we look at how the increase of number of PC components changes the model's behavior. As can been seen from both Figure 5 and Figure 6, the discriminabilities

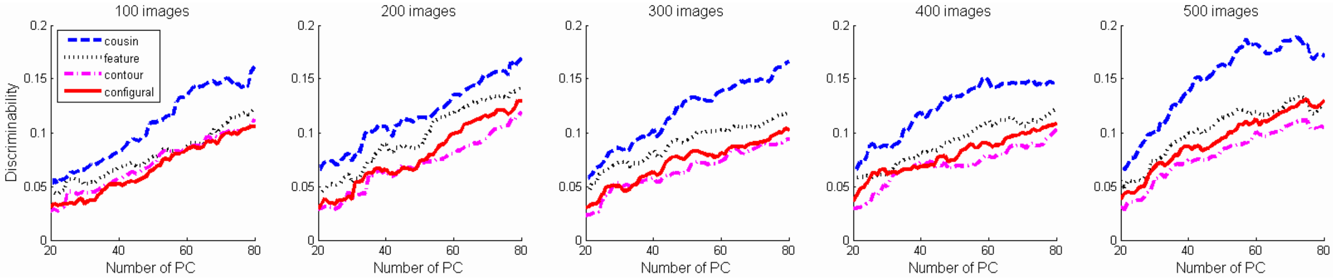

Figure 6: The discriminability of Jane's sets at PCA with the number of images taking values from 100 (upper row left most) to 500 (lower row right most) with a step of 100 images, all upright.

of all sets increases as the number of PC components increases. This qualitatively matches the human data (Figure 2, upright portion) that performance increases with age. Perhaps surprisingly, changing the number of PC's does not appear to affect the rank.

Discriminability Using Bandpass PCA Here we examine how the frequency of the Gabor filters affects the discriminability at the PCA level. We did "bandpass" PCA [Dailey and Cottrell, 1999] on the Gabor filter responses, i.e. we performed PCA on the Gabor filter responses of each frequency individually and looked at the discriminability of the Jane's sets for each to observe the model's performance at different frequencies. Figure 7 shows the discriminability of the Jane's sets when the model is exposed to the 5 individual frequencies respectively. An interesting observation is that at the lowest frequency, the rank of the discriminability is $cousin > featural > contour > configural$, which matches that of the children, i.e. the configural set is the least discriminable. At the second lowest and the medium frequency, the rank of the discriminability is $cousin > featural > configural > contour$, i.e. the configural set catches up a little bit and its discriminability exceeds that of the contour set. At this moment, the rank is the same as the global PCA and that of the adults'. As the frequency goes higher to the two highest frequencies, the discriminability of the configural set exceeds those of the contour set and the featural set, contrary to the human data.

Overall, the relative discriminability of the configural set increases when the frequency of the Gabor filter goes higher. This is a very interesting but counter-intuitive result. Intuitively, featural processing is about local features and should use relatively higher frequencies while configural processing is sensitive to the distance between the local features and should use relatively lower frequencies. I.e. by our intuitions, we should have observed the discriminability of the featural set increase relative to that of the configural set, but we observed the opposite in our model. From our computational model's view, the low frequency is more shift invariant, i.e. small movements of the features on the face do not disturb the representation, while the high frequencies will detect this change because their receptive fields are so small that only the Gabor filters located right at the features will be activated. On the other hand, neither of the low nor high frequencies have an obvious advantage or disadvantage over the changes of the features themselves. Note that our data is suggesting that the discriminability of the configural set *relative* to that of the feature set will increase when the higher frequencies are available over time, NOT that the the configural processing is mainly using the high frequency.

The result is also interesting because this might also be contributing to development. First of all, people are born with only low acuity and gain high acuity over time. At the same time, configural processing kicks in more slowly than featural processing. Is there a connection? We do not know yet. Since the acuity of vision develops to adult-like levels around two years of age, while the human data we are considering is of children from 6 to 10 years of age, we hesitate to make a connection without more data. Second, the increase of the relative discriminability of the configural set as the frequency goes up coincides with the increase of the relative discriminability of the configural set as the number of training images goes up. Is there a connection? There might be one. I.e. it might be possible that when the number of training images is small, the low frequency dominates in the PCA representation and that when the number of training images increases, more of the higher frequencies kick in. We are currently working on qualitatively estimating the portion of each frequency at the PCA level to further investigate this hypothesis.

Also, this result explains the difference between the results we obtained in the earlier project and those we obtained now. The Gabor filters we used in the earlier project were much higher frequencies compared to the current ones (Figure 4.) So, not surprisingly, we found our model too sensitive to the configural set, which is exactly what we observe now in the PCA representation of Gabor filters with high frequencies.

Discussion

In our earlier work, we found that our standard model was overly holistic. We modified our model by adding a parts-based representation, implemented as a local feature PCA. In this work, we used biologically realistic Gabor filters and found that our standard model does not need the parts-based representation any more. It fits the adult data using only the PCA level.

Furthermore, we investigated how the developmental

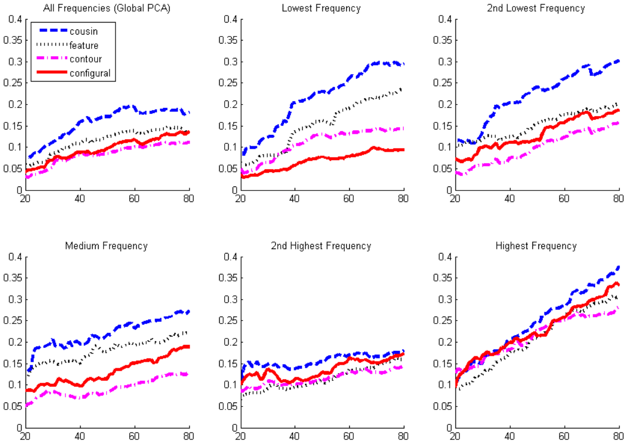

Figure 7: The discriminability of the Jane's sets at different Gabor filter frequencies. The leftmost image on the upper row is the global PCA for comparison. The rest five images are bandpass PCA of the 5 Gabor filter frequencies.

data could be modeled. We used a very simple and intuitive manipulation, i.e. increasing the number of training images and increasing the number of the principal components, both of which have a straightforward correspondence in human development: children get to see more and more faces over time and they allocate more resources to face processing respectively. We found that in our model, when the number of the training images is small, the discriminability of the configural set is the lowest, which is also observed in children' performance. As the number of images increase, the discriminability of the configural set slowly catches up and exceeds that of the contour set, which is observed in adults' performance. In parallel, we found that the increase of the number of components is able to account for the continuous improvement in the performance of all the Jane's sets. Taken together, a "child" model with a small number of images to train on and with a representation of small number of PC components will not perform well on discrimination and the ranking of difficulty on the Jane's sets will mimic that of the children. A "adult" model trained with a large number of images and with more PC components will have a better ability to discriminate the stimuli and the rank on the four sets matches that of the adults.

We also investigated the effects of the Gabor filter frequencies on the relative discriminability of the Jane sets and its possible connection to development. However, we think this direction needs more careful treatment before we can draw any firm conclusions. We leave this for future work.

Acknowledgement

We thank Gary's Unbelievable Research Unit (GURU) for valuable comments, Daphne Maurer for Jane's data sets and anonymous reviewers for suggestions. This research was supported by NIMH grant MH57075 to GWC.

References

[Cottrell et al., 2002] Cottrell, G. W., Branson, K. M., and Calder, A. J. (2002). Do expression and identity need separate representations? In *Proc. 24th Ann. Cog. Sci. Soc. Conf.*, Mahwah, New Jersey. Cognitive Science Society.

[Dailey and Cottrell, 1999] 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 et al., 2002] 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(8):1158–1173.

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

[Diamond and Carey, 1986] Diamond, R. and Carey, S. (1986). Why faces are and are not special: an effect of expertise. *JEP: General*, 115(2):107–117.

[Farah et al., 1995] Farah, M., Levinson, K., and Klein, K. (1995). Face perception and within-category discrimination in prosopagnosia. *Neuropsychologia*, 33:661–674.

[Hofmann et al., 1998] Hofmann, T., Puzicha, J., and Buhmann, J. M. (1998). Unsupervised texture segmentation in a deterministic annealing framework. *IEEE PAMI*, 20(8):803–818.

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

[Joyce and Cottrell, 2004] Joyce, C. and Cottrell, G. W. (2004). Solving the visual expertise mystery. In *Proc. Neural Comp. and Psych. Workshop 8*, Progress in Neural Processing. World Scientific, London, UK.

[Mondloch et al., 2002] Mondloch, C. J., Grand, R. L., and Maurer, D. (2002). Configural face processing develops more slowly than featural face processing. *Perception*, 31:553–566.

[Phillips et al., 1998] Phillips, J., Wechsler, H., Huang, J., and Rauss, P. J. (1998). The feret database and evaluation procedure for face-recognition algorithms. *Image and Vision Computing*, 16(5):295–306.

[Thompson, 1980] Thompson, P. (1980). A new illusion. *Perception*, 9:483–484.

[Tong et al., 2005] Tong, M., Joyce, C., and Cottrell, G. W. (2005). Are greebles special? or, why the fusiform fish area would be recruited for sword expertise (if we had one). In *Proc. 27th Ann. Cog. Sci. Soc. Conf.*, La Stresa, Italy. Cognitive Science Society.

[Tran et al., 2004] Tran, B., Joyce, C. A., and Cottrell, G. W. (2004). Visual expertise depends on how you slice the space. In *Proc. 26th Ann. Cog. Sci. Soc. Conf.*, Chicago, Illinois. Cognitive Science Society.

[Zhang and Cottrell, 2004] Zhang, L. and Cottrell, G. W. (2004). When holistic processing is not enough: Local features save the day. In *Proc. 26th Ann. Cog. Sci. Soc. Conf.*, Chicago, Illinois. Cognitive Science Society.