Four Blobs: "Y" or Face?

Lingyun Zhang & Garrison W. Cottrell
University of California, San Diego
Computer Science and Engineering
9500 Gilman Dr., La Jolla, CA 92093-0114 USA
{lingyun,gary}@cs.ucsd.edu

Abstract

What is the difference between processing faces and other objects such as letters? What makes humans face experts, and what makes this expertise different from other identification skills? It is well known that people are very sensitive to the configural information in faces. How does the sensitivity to face configuration compare to sensitivity to configurations of other stimuli? To investigate these issues, Nishimura et al. (2004) designed a test to contrast two types of processing using the same stimuli. They primed subjects to see four blobs as either a "Y" or as a face. Then they asked the subjects to discriminate pairs of these stimuli that differed only in small shifts in the blob locations. Although the stimuli were exactly the same, subjects were more accurate in the face condition than the "Y" condition. With Nishimura et al., we assumed that the subjects were relying on their letter recognition networks in the "Y" condition and their face recognition networks in the face condition to perform the task. We therefore trained two networks, a face recognition network and a letter recognition network that were otherwise identical in structure, and show here that the internal representations in the letter network for the blobs were less differentiated than the internal representations for the blobs in the face network. We argue that this is a natural consequence of the requirements of the two tasks.

1. Introduction

We have developed a simple neurocomputational model of face and object recognition that accounts for a number of important phenomena in facial expression processing, holistic processing and visual expertise [9, 7, 11, 19]. Here, we investigate the model's ability to account for a recent experiment that shows differential human sensitivity to configural information based on priming. Nishimura et al. (2004) constructed "blob" stimuli consisting of four gaussian blobs in the same spatial arrangement as the eyes, nose and mouth of human faces. The blob stimuli were also constructed to have about the same variability in location as those features in human faces. They then primed one group of subjects to see these blobs as a "Y" and another group to see them as parts of a face. The first group was less able to discriminate the blob stimuli than the second group. They suggest that this is because the face group is using their face recognition system to discriminate the blobs, and present this as further evidence that face processing utilizes a sensitivity to configuration that other tasks do not.

Why would subjects show these differential sensitivities? We first discuss what is known about face processing. Face processing has long been described as *holistic* or *configural*. Holistic is typically taken to mean that subjects use some kind of whole-face representation when processing faces. This is reflected at least two ways. First, subjects have difficulty recognizing parts of the face in isolation – there is a whole-face superiority effect. Second, subjects have difficulty ignoring parts of a face when making a decision about another part. For example, subjects are slower in making an expression judgement about the top half of a face if the bottom half is displaying an incongruent expression [3]. Our model of face processing is able to account for this kind of data because it uses representations that are global; that is, they are composed of whole-face templates we have called *holons* [7, 8]. Inputs that match part of one of these representational units cause it to fire. Units later in the processing stream take this to be a vote for the whole template, so that the system as a whole responds as if both halves of the face had been of the type matched.

Configural processing means that subjects are sensitive to the relationships between the parts, e.g., the distances between the eyes. Thus, small changes in the spacing of the eyes cause subjects to see the faces as different people. This is presumably due to long experience with many people, and the need to differentiate these faces. This sensitivity to configuration, however, takes a surprisingly long time to develop [23].

What kind of processing is required to recognize objects,

and how does it differ from faces? Diamond and Carey [14] were among the first to discriminate between the types of processing involved in face and object recognition. They proposed that first-order relational information, which consists of the coarse spatial relationships between the parts of an object (e.g. eyes are above the nose), is sufficient to recognize most objects at the basic level. By contrast, second-order relational information (e.g. the spacing between individual features such as the eyes and the mouth), is needed for face recognition. They found that inverting images severely disrupted subjects' ability to discriminate between faces, and that this effect was stronger for faces than for dogs and landscapes in naive subjects. However, dog experts also showed an inversion effect for dogs. These results suggest that face processing is a kind of expertise, and that experts become over-tuned to the typical orientation of the stimuli in their domain of expertise. Their ability to discriminate based upon subtle configural differences is overly disrupted by inversion. Diamond and Carey [14] 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 greater reliance on second-order relational information, and that perhaps this characteristic distinguishes face/expert-level processing from regular object recognition. Farah et al. [15] 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. Thus, regular object classification is thought to use a parts-based representation.

However, our model uses the same kind of representation for all stimuli. The only difference is the requirements of the task, between a version of our model that recognizes faces and one that recognizes objects, such as letters. In face identification, the model must take similar looking stimuli, and magnify small differences between them in its internal representation (see Figure 1, left). On the other hand, in order to recognize letters, the model must take similar looking things (the same letter in different fonts, for example) and represent them as the same thing (see Figure 1, right). This point has been made before [17]; here, we construct models that automatically implement those differences via learning the different tasks. Then we may analyze the models in ways that we cannot analyze human subjects.

Our use of separate networks for these two tasks is motivated by fMRI experiments that have shown that face-related tasks and letter-related tasks activate different brain regions [17, 20, 5]. The fusiform face area tends to be in right medial fusiform gyrus, whereas there appears to be a letter form or word form area in the left midfusiform gyrus [5, 6, 22]. We model this by having two networks, each trained to do one of the tasks. We hypothesize that the priming in Nishimura et al.'s experiment causes one of these networks to be primed, and therefore used for the task. Then we show how blobs are represented differentially in the two networks.

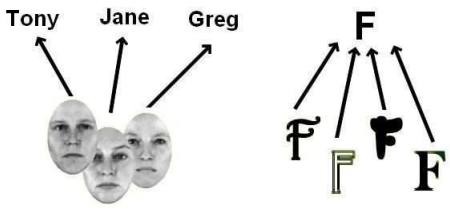

Figure 1. Faces are automatically perceived as different individuals despite the similarity, while the "F"s are perceived as the same letter. Adapted from [17].

In the following, we describe Nishimura et al.'s experiments and our account of their data. We found that different encoding due to different tasks in our model could account for their data, which suggests that the effect in human subjects may come from using the letter system verses the face system to encode the stimuli. Finally, we discuss plans for future work.

1.1. Nishimura et al's Stimuli and Experiments

To compare sensitivity to configural information for faces and non-faces, Nishimura et al. (2004) created a group of stimuli that contain 4 blobs located over the eyes, nose and mouth of configurally different faces (Figure 2). Two groups of subjects were then primed to see the blobs as either a face or as the letter "Y", respectively (Figure 3). The subjects then performed same/different tasks on different pairs of blobs that differed only in their relative locations. The results showed that when people take these blobs as faces, they discriminate them better than when they see them as the letter Y (Figure 4). This suggests that by different priming, ambiguous stimuli might be represented differently. In this work, we concentrate on modelling this using our neurocomputational model of visual object recognition.

2. A Computational Model of Classification

Our model is a three level neural network that has been used in previous work (Figure 5). The model takes manually aligned images as input. The images are first filtered by

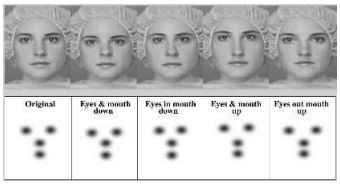

Figure 2. 4 blobs are located over the eyes, nose and mouth of 5 faces used in previous studies (from [24]).

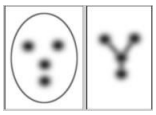

Figure 3. The blobs are primed as either face or letter Y. (from [24]).

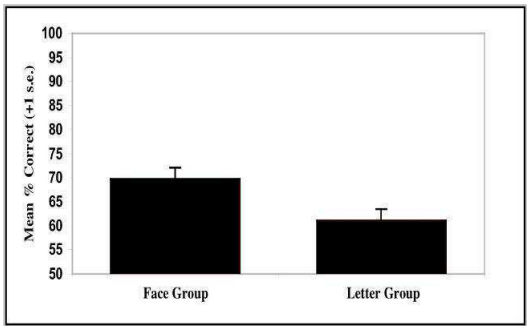

Figure 4. Mean accuracy of the subjects in the face group was higher than that of those in the letter group. (from [24]).

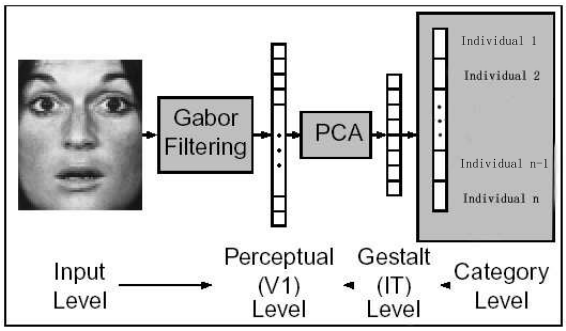

Figure 5. Object recognition model (from [11]).

2D Gabor wavelet filters, which are a good model of simple cell receptive fields in cat striate cortex [18]. 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 class to each image. We now describe each of the components of the model in more detail.

2.1. Perceptual Level

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 frequency[18]. DeValois et al [13] mapped the receptive fields of V1 cells and found evidence for multiple lobes of excitation and inhibition. 2D Gabor filters [12](Figure 6) have been found to fit the 2D spatial response profile of simple cells quite well[18]. In this processing step the image was filtered with a rigid 23 by 15 grid of overlapping 2-D Gabor filters[12] in quadrature pairs at five scales and eight orientations [11](Figure 7). We thus obtained $23 \times 15 \times 5 \times 8 = 13,800$ filter responses in this level, which is termed the *perceptual* level [11].

2.2. Gestalt Level

In this stage we perform a PCA of the Gabor filter responses. This is a biologically plausible means of dimensionality reduction[11], since it can be learned in a Hebbian manner. PCA learns features that encode correlations between features at the previous level. Thus, for example, if the Gabor filter responses to the left eye are highly correlated with the Gabor filter responses to the right eye, there will be a principal component that corresponds to both of these, capturing the redundancy in the Gabor filter responses. 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 13,800 Gabor vectors.

2.3. Categorization Level

The classification portion of the model is a two-layer back-propagation neural network. 50 hidden units are used. A scaled tanh [21] activation function is used at the hidden layer and the softmax activation function $y_i = e^{a_i} / \sum_k e^{a_k}$ was used at the output level. The network is trained with the cross entropy error function [1] to identify the images using localist outputs. Networks trained in this way learn to produce the conditional probability of the output class given the input.

Figure 6. A Gabor function is constructed by multiplying a Gaussian function by sinusoidal function[12]. We use five scales and eight orientations.

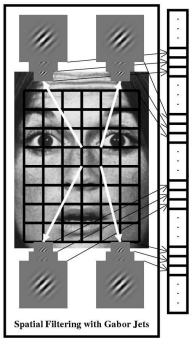

Figure 7. An image filtered with a rigid 23 by 15 grid of overlapping 2-D Gabor filters in quadrature pairs at five scales and eight orientations (from [10]).

3. Modelling Nishimura et al.

We set up two networks, one engaged in face identity classification and the other in letter classification. After training, blobs were fed to these two networks. The mean discriminabilities were computed respectively and then compared between the two classifiers. The results showed that the face network considers the blob stimuli to be more different than the letter network does.

3.1. The Image Sets

We had 182 face images of 26 individuals (7 images for each individual). We also used 182 letter images of the 26 upper case letters (7 images for each letter in 7 fonts). 27 blob stimuli were created by manipulating the eye blob and/or mouth blob's position.

The FERET database is a large database of facial images, which is now standard for face recognition from still images[25]. We used 182 face images of 26 individuals, 7 images each. In [11], where the task was to learn facial expressions, images were aligned so that eyes and mouth went to designated coordinates. This alignment removes the configural information which is crucial for our work, because we are trying to understand how configural processing is better learned in the face recognition task than in the letter recognition task. To avoid this negative effect, we required that the relative spacing between the parts of the face remain the same. We formed a triangle from the eyes and mouth of the original face, and then translated, rotated and scaled this triangle to be as close as possible to a target triangle in terms of the sum squared differences between the final eye and mouth coordinates and the target coordinates.[1] This manipulation preserves the relative distance between the features of the face (Figure 8). Thus, a triangle represented by the eyes and mouth is scaled and moved to fit closely to a reference location, but the relative sizes of the sides of the triangle are not changed. The aligned images were 192 pixels by 128 pixels.

Figure 8. Two examples of face image normalization. The faces were cropped with the eyes and the mouth as close as possible to the target position while keeping the shape of the triangle among these features the same.

We also used 182 192 by 128 pixel letter images of the 26 upper case letters, 7 images each (Figure 9). The letter images were aligned so that the ends of the letter Y were approximately where the eyes and mouth were in the face stimuli.

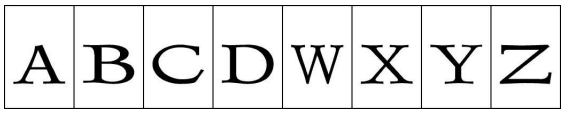

Figure 9. Some letter images.

Blob images were generated by setting gaussian blobs (of width $\sigma = 5$ pixels) at left eye $(80(\pm 3), 36(\pm 3))$, right eye $(80(\pm 3), 92(\pm 3))$, nose $(115, 65)$ and mouth $(150(\pm 3), 65)$ positions. Note the two eye blobs were always symmetric. Thus $3 * 3 * 3 = 27$ blob images were generated.

3.2. Training and Learning

A learning rate of 0.05 and a momentum of 0.5 were used in the results reported here. Two networks were set

[1]The objective function for the minimization was ($||Eye_{right} - targetEye_{right}||^2 + ||Eye_{left} - targetEye_{left}||^2 + ||Mouth - targetMouth||^2$).

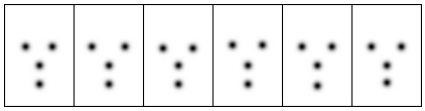

Figure 10. Some "blob" images.

up for faces and letters respectively. In a pilot experiment, 10 percent of the images were selected randomly as a test set and another 10 percent as a validation set [10]. Both networks achieve 80-90 percent accuracy within 50 epochs. This classification rate was good enough to show that our model represented the images well.

For the following experiments, we simply trained both networks on all 182 images, since we are only interested in obtaining a good representation at the hidden layer. Training was stopped at the 50th epoch based on the above pilot experiment. After training, the blob images were presented to the network. Note for the face network, the blob images were projected according to the face image eigenvectors in the PCA level while for the letter network, they were projected onto those of letters.

3.3. Modelling Discrimination

Hidden unit activations were recorded as the network's representation of images. In order to model discriminability between two images, we present an image to the network, and record the hidden unit response 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 [11]. The pairwise average within the blob image set was taken as the measure of the network's ability to discriminate the blob images. For both the face network and the letter network, the average of the discriminabilities was computed over 50 networks which were all trained in the same way, but used different initial random weights.

The results (Figure 11) showed that the face network better discriminates the difference between blob images than the letter network ($F = 24.72$, $p < 0.001$). I.e. the representations of blob images in the face network were more differentiated (further apart) from one other than those in the letter network.

To visualize this difference, we extracted the principal components of the hidden layer representations and then projected their activations onto the first three principal components (the ones that represent most of the variance of the activations) (Figure 12). Notice the hidden layer representations of the blobs in the face network are better separated, which suggests that the face network is especially sensitive to configural differences.

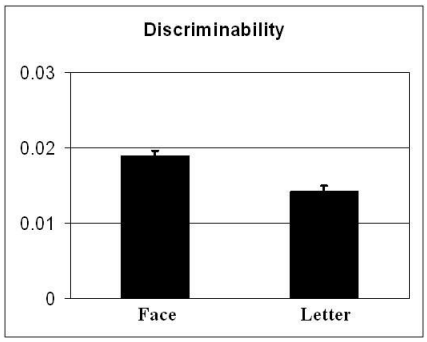

Figure 11. Mean discriminability of the blobs in the face network was higher than that in the letter network ($F = 24.72$,$p < 0.001$).

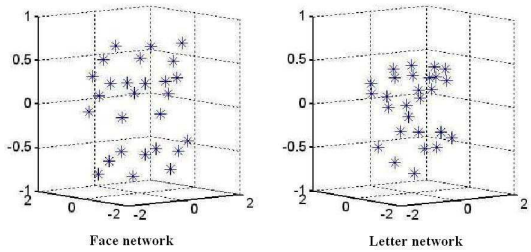

Figure 12. The projection of the hidden activities onto the first 3 principal components.

3.4. Inverted blobs

While Nishimura et al. did not present inverted blobs to their subjects, it is theoretically interesting to see what subjects would do with inverted stimuli if they are perceiving them as letters or as faces. If they see them as faces, then one expects that there should be an "inverted blob effect." This is easy to do with the networks. The same set of the blobs were inverted and then presented to the two types of networks. We expect that the face network should show a greater loss of discrimination than the letter network. The results, plotted along with the original data from Figure 11 for comparison, are shown in Figure 13. As expected, there is an inverted blob effect for the face network, while the letter network shows a slight increase in discriminability for the blobs. As there is no human data for this case, this is a prediction of the model.

4. Discussion

Our results qualitatively match those from human subjects. In the two networks, the difference in the discriminability of the same stimuli comes from the different en-

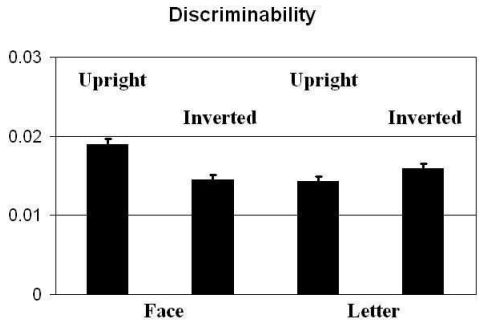

Figure 13. Discriminability results for the letter and face networks when the blobs are presented upright and inverted. There is no difference between the networks on the inverted stimuli ($F = 3.11$,$p = 0.0808$), but there is a significant difference between the face network's ability to discriminate the inverted blobs over the upright blobs ($F = 21.9$,$p < 0.001$). There is a small but significant increased discriminability for the letter networks ($F = 4.3$,$p = 0.0407$).

coding in the two tasks. Note that because all the faces share the same first-order relational features, the categorization need to be carried out at finer level. The face network needs to spread out similar face images to categorize individuals, while the letter network needs to squeeze different fonts of the same letter to a letter prototype. So the face network tends to magnify small differences in face images while the letter network tends to ignore such variability. This is consistent with our previous results with faces, objects and letters [19, 26]. Thus the configural differences between the blob stimuli are better noticed by the face network. Considering our separate face and letter networks to be analogous to the separate face and letter processing systems in the brain, we can apply the reasoning derived from our networks to the brain, as follows. The face processing system learns to pay attention to small differences such as configural changes in faces while the letter processing system ignores them. Thus, when the blobs are perceived as faces, the differences are more present in the inner representation by the face system, while when the blobs are perceived as letters, the differences are less present in the inner representation by the letter system.

Our model makes a prediction concerning an inverted blob effect. We have previously shown that our face networks show an inverted face effect – they are poorer at discriminating upside down faces [27]. We also showed that configural differences take the biggest hit in discrimination when compared with featural changes between the stimuli to be discriminated. The configural effect generalizes to the blobs. Interestingly, the letter network does not show this effect – its discriminability scores increase slightly when the stimuli are inverted. Possibly this is because it is poor at discriminating the upright stimuli that all match a particular letter "Y". In the inverted case, we speculate that the blobs fall in regions where it may have to discriminate slight differences between letters (e.g., "R" and "A").

How might these effects vary with development? As the participants were undergraduates [24], we would expect a lower strength of this effect in young children since sensitivity to configuration is lower in children [16, 23]. Furthermore, if the inverted blob effect is found in adults, as predicted by our model, we would expect that children would not show this effect. This is because developmentally, children show either no inversion effect, or less of an inversion effect, depending on their age [4, 2]. We are currently exploring adding a developmental component to our model in order to account for these developmental changes.

5. Acknowledgements

We thank Carrie Joyce for previous suggestions, Mayu Nishimura, Daphne Maurer and Catherine J. Mondloch for useful comments, and for inspiring this modelling idea. This research was supported by NIMH grant MH57075 to GWC.

References

[1] BISHOP, C. M. *Neural networks for pattern recognition*. Oxford University Press, 1995.

[2] BRACE, N. A., HOLE, G. J., KEMP, R., PIKE, G., DUUREN, M. V., AND NORGATE, L. Developmental changes in the effect of inversion: Using a picture book to investigate face recognition. *Perception 30* (2001), 85–94.

[3] CALDER, A. J., YOUNG, A. W., KEANE, J., AND DEAN, M. Configural information in facial expression perception. *Journal of Experimental Psychology: Human Perception and Performance 26*, 2 (2000), 526–551.

[4] CAREY, S., AND DIAMOND, R. From piecemeal to configurational representation of faces. *Science 195* (1977), 213–313.

[5] COHEN, L., DEHAENE, S., NACCACHE, L., LEHERICY, S., DEHAENE-LAMBERTZ, G., HENAFF, M.-A., AND MICHEL, F. The visual word form area. Spatial and temporal characterization of an initial stage of reading in normal subjects and posterior split-brain patients. *Brain 123* (2000), 291–307.

[6] COHEN, L., LEHERICY, S., CHOCHON, F., LEMER, C., RIVARD, S., AND DEHAENE, S. Language-specific tuning of visual cortex? Functional properties of the visual word form area. *Brain 125* (2002), 1054–1069.

[7] COTTRELL, G. W., BRANSON, K. M., AND CALDER, A. J. Do expression and identity need separate representations? In *Proceedings of the 24th Annual Conference of the Cognitive Science Society* (Mahwah, New Jersey, 2002), The Cognitive Science Society.

[8] COTTRELL, G. W., AND METCALFE, J. Empath: Face, gender and emotion recognition using holons. In *Advances in Neural Information Processing Systems 3* (San Mateo, 1991), R. P. Lippman, J. Moody, and D. S. Touretzky, Eds., Morgan Kaufmann, pp. 564–571.

[9] DAILEY, M. N., AND COTTRELL, G. W. Organization of face and object recognition in modular neural network models. *Neural Networks 12* (1999), 1053–1073.

[10] DAILEY, M. N., COTTRELL, G. W., AND ADOLPHS, R. A six-unit network is all you need to discover happiness. In *TwentySecond Annual Conference of the Cognitive Science Society* (2000).

[11] DAILEY, M. N., COTTRELL, G. W., PADGETT, C., AND ADOLPHS, R. Empath: A neural network that categorizes facial expressions. *Journal of Cognitive Neuroscience 14*, 8 (2002), 1158–1173.

[12] DAUGMAN, J. G. 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* (1985), 1160–1169.

[13] DEVALOIS, R. L., AND DEVALOIS, K. K. *Spatial Vision*. Oxford University Press, 1988.

[14] DIAMOND, R., AND CAREY, S. Why faces are and are not special: an effect of expertise. *Journal of Experimantal Psychology: General 115*, 2 (1986), 107–117.

[15] FARAH, M., LEVINSON, K., AND KLEIN, K. Face perception and within-category discrimination in prosopagnosia. *Neuropsychologia 33* (1995), 661–674.

[16] FREIRE, A., AND LEE, K. Face recognition in 4- to 7-year-olds: Processing of configural, featural, and paraphernalia information. *Journal of Experimental Child Psychology 80* (2001), 347–371.

[17] GAUTHIER, I., MOYLAN, J., TARR, M. J., ANDREW, A. W., SKUDLARSKI, P., AND GORE, J. C. Automatic subordinate-level processing for faces and letters, 1998. Presentation at *Human Brain Mapping*.

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

[19] JOYCE, C., AND COTTRELL, G. W. Solving the visual expertise mystery. In *Proceedings of the Neural Computation and Psychology Workshop 8*, Progress in Neural Processing. World Scientific, London, UK, 2004.

[20] KANWISHER, N., MCDERMOTT, J., AND CHUN, M. M. The fusiform face area: A module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience 17* (1997), 4302–4311.

[21] LECUN, Y., BOTTOU, L., ORR, G. B., AND MÜLLER, K.-R. Efficient backprop. In *Neural Networks—Tricks of the Trade, Springer Lecture Notes in Computer Sciences* (1998), vol. 1524, pp. 5–50.

[22] MCCANDLISS, B. D., COHEN, L., AND DEHAENE, S. The visual word form area: Expertise for reading in the fusiform gyrus. *Trends in Cognitive Sciences 7*, 7 (2003), 293–299.

[23] MONDLOCH, C. J., GRAND, R. L., AND MAURER, D. Configural face processing develops more slowly than featural face processing. *Perception 31* (2002), 553–566.

[24] NISHIMURA, M., MAURER, D., AND MONDLOCH, C. J. Perceiving changes in spacing among four 'blobs': Are adults more sensitive when primed to see them as facial features?, 2004. Poster presented at *The 2004 Cognitive Neuroscience Society Annual Meeting*.

[25] PHILLIPS, J., WECHSLER, H., HUANG, J., AND RAUSS, P. J. The feret database and evaluation procedure for face-recognition algorithms. *Image and Vision Computing 16*, 5 (1998), 295–306.

[26] TRAN, B. A., JOYCE, C. A., AND COTTRELL, G. W. Visual expertise depends on how you slice the space. In *Proceedings of the 26th Annual Conference of the Cognitive Science Society* (Mahwah, New Jersey, 2004), The Cognitive Science Society.

[27] ZHANG, L., AND COTTRELL, G. W. When holistic processing is not enough: Local features save the day. In *Proceedings of the 26th Annual Conference of the Cognitive Science Society* (Mahwah, New Jersey, 2004), The Cognitive Science Society.