Visual Expertise Depends on How You Slice the Space

Brian A. Tran (b3tran@ucsd.edu)
UCSD Department of Computer Science and Engineering, 9500 Gilman Drive
La Jolla, CA 92093-0114 USA

Carrie A. Joyce (cjoyce@cs.ucsd.edu)
UCSD Department of Computer Science and Engineering, 9500 Gilman Drive
La Jolla, CA 92093-0114 USA

Garrison W. Cottrell (gary@cs.ucsd.edu)
UCSD Department of Computer Science and Engineering, 9500 Gilman Drive
La Jolla, CA 92093-0114 USA

Abstract

Previous studies using fMRI have found that the Fusiform Face Area (FFA) responds selectively to face stimuli. More recently however, studies have shown that FFA activation is not face-specific, but can also occur for other objects if the level of experience with the objects is controlled. Our neurocomputational models of visual expertise suggest that the FFA may perform fine-level discrimination by amplifying small differences in visually homogeneous categories. This is reflected in a large spread of the stimuli in the high-dimensional representational space. This view of the FFA as a general, fine-level discriminator has been disputed on a number of counts. It has been argued that the objects used in human and network expertise studies (e.g. cars, birds, Greebles) are too "face-like" to conclude that the FFA is a general-purpose processor. Further, in our previous models, novice networks had fewer output possibilities than expert networks, leaving open the possibility that learning more discriminations, rather than learning fine-level discriminations, may be responsible for the results. To challenge these criticisms, we trained networks to perform fine-level discrimination on fonts, an obviously non-face category, and showed that these font networks learn a new task faster than networks trained to identify letters. In addition, all networks had the same number of output options, illustrating that visual expertise does not rely on number of discriminations, but rather on how the representational space is partitioned.

Introduction

The Fusiform Face Area (FFA) in the ventral temporal lobe has recently received much attention. Initial work appeared to show that this area was selective for processing faces. Several fMRI studies showed high activation in the FFA only to face stimuli and not other objects (Kanwisher et al., 1997; Kanwisher, 2000). Further, studies involving patients with *associative prosopagnosia*, the inability to identify individual faces (Farah et al., 1995), and *visual object agnosia*, the inability to recognize non-face objects (Moscovitch et al., 1997), seemed to indicate a clear double dissociation between face and object processing. Prosopagnosic patients had lesions that encompassed either right hemisphere or bilateral FFA, while object agnosic patients' lesions did not (De Renzi et al., 1994).

Gauthier and colleagues have challenged the notion of the face specificity of the FFA by pointing out that the earlier studies failed to equate the level of experience subjects had with non-face objects, to the level of experience they had with faces (Gauthier et al., 1997; Gauthier et al., 1999a; Gauthier et al., 1999b). She showed that the FFA was activated when bird and dog experts were shown pictures of the animals in their area of expertise. Further, she illustrated that, if properly trained, individuals can develop expertise on novel, non-face objects (e.g. Greebles), and subsequently show increased FFA activation to them (Gauthier et al., 1999a). Expertise in these studies was operationally defined as the point in training when a subject's default response level (i.e. entry level) "shifts" from basic to the individual level. This is indexed by the subject's reaction time for verifying individual names becoming as fast as the time to verify category membership.

Neurocomputational models done first by Sugimoto and Cottrell (2001) and later extended by Joyce and Cottrell (2004) began to address the question of how and why the FFA gets recruited for these other tasks (Sugimoto & Cottrell, 2001, Joyce & Cottrell, 2004). Using four different types of stimulus classes (books, cans, cups and faces), Sugimoto and Cottrell found that the amount of expert-level experience on a previous task correlates with faster subordinate level learning relative to a system that processes the same stimuli, but not to a subordinate level. Thus, an area that is used for one expertise task will learn a second expertise task faster than an area used only for basic level discriminations. Joyce and Cottrell (2004) further found that an expert network's ability to separate individuals is reflected in highly variable responses at the representational layer (the hidden layer). This response variability extended to novel categories, permitting faster learning of these

categories. This suggests that the FFA is primed to win the competition for a new expertise task because of its ability to fine-tune its feature representations when given a novel fine-level discrimination task (Joyce & Cottrell, 2004).

While the human and computational studies of expertise are compelling, they are not undisputed. For example, proponents of the view that the FFA is face-specific claim that the objects used in human expertise studies, such as cars, dogs, birds, or Greebles, are "face-like", meaning they possess properties similar to faces. Thus any response of FFA to these stimuli is due to their featural similarity with faces, not because the FFA is a general-purpose, fine-level discriminator. While the network simulations, which illustrate expertise across a wide variety of non-face objects, may seem to argue against this criticism, a methodological issue makes these results less compelling. In previous simulations, non-expert networks were trained on a lesser number of discriminations (4 category labels) than expert networks (10 individual labels plus the 4 category labels). It has been argued (Mike Tarr, personal communication) that if an object recognition network simply had to make as many discriminations as the expert one, then it would also be able to learn Greebles faster.

The current simulations were designed to address the criticisms cited above. First, we train the networks to perform fine-level discrimination on an obviously non-face category: fonts. In this case, the basic level networks learn to identify letters presented in a variety of different fonts (a task any human can do with ease) while the subordinate level networks learn to distinguish the particular font in which a letter is written (a task few humans can do). To address the second criticism, we present both basic and subordinate level networks with the same stimulus set and have them perform an equal number of discriminations (e.g. 6 letter vs. 6 font discriminations). Thus, any advantage to learning to distinguish Greebles by the font network over the letter network cannot be due to the number of discriminations learned.

Experiments

We ran two sets of experiments. In the first, we investigated the ability of our basic visual object processing architecture (Dailey & Cottrell, 1999; Dailey et al. 2002; Joyce & Cottrell, 2004) to recognize letters and fonts. This allowed us to discover which fonts were difficult and which letters gave good generalization once their font had been learned (by training on other letters). We then used these results in the second set of experiments to perform a very controlled version of our previous "basic versus expertise network" experiments, and investigate generalization to Greeble expertise.

Experiment 1: Stimuli and Methods

The images used were 300x300 pixel images of letters. For this experiment, 15 different fonts were used, and for each of those 15 fonts, we had images of all 26 letters. The fonts were chosen to be somewhat difficult. Image preprocessing

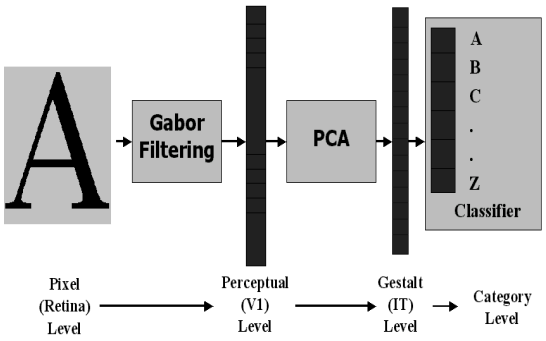

Figure 1: The expertise model.

of the different letters and fonts followed the procedures outlined in Dailey and Cottrell (1999). Each image was first processed using 2-D Gabor wavelet filters (5 spatial frequencies at 8 orientations each), a simple model of complex cell responses in visual cortex. The filters were applied at 64 points in an 8x8 grid, resulting in a vector of 2560 elements (Buhmann, Lades & von der Malsburg, 1990; Dailey & Cottrell, 1999). The vectors were then normalized via z-scoring (scaled and shifted so that they had zero mean and unit standard deviation) on a per-filter basis, a local operation. A principal components analysis (PCA) was then applied to the normalized vectors. The top 40 components were saved and renormalized. Projections of the stimuli onto these 40 dimensional vectors constituted the input to the networks. Figure 1 shows the expertise model, which includes the image preprocessing procedure.

A standard backpropagation network architecture was used for learning classifications. The network had 40 input units, each representing a principal component vector, a 30-unit hidden layer using the logistic sigmoid function, and 15 linear output units for the font network, and 26 linear outputs for the letter network. The learning rate and momentum were 0.01 and 0.5, respectively.

Letter training Letter networks were trained to identify letters across a subset of the 15 different fonts. Each network was given the letters from 13 different fonts as the training set and another font as the holdout set. It was then

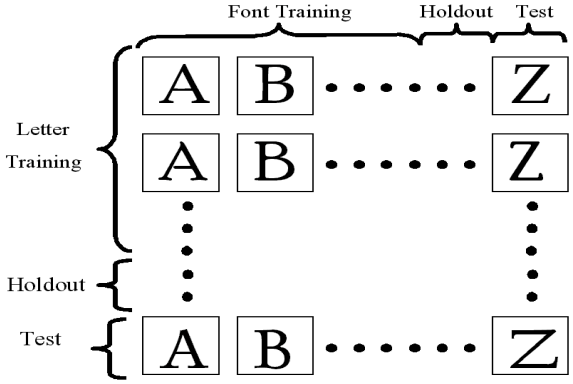

Figure 2: Training for Experiment 1. 26 letters and 15 fonts were used.

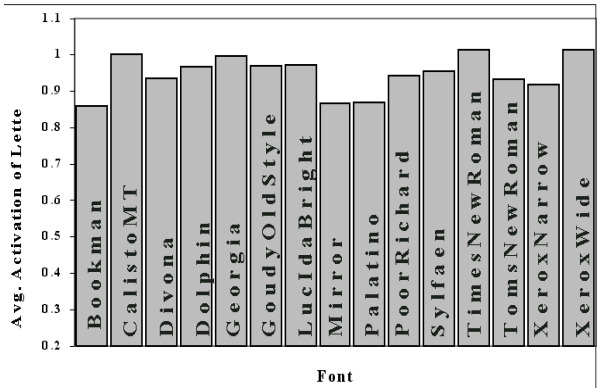

Figure 3: Average activation of letters for test fonts.

tested on the letters from the remaining font. Training was stopped at either an RMSE of 0.02 or when overtraining started to occur. The result was 15 letter networks, all trained and tested on different fonts. Figure 2 illustrates the training and test sets for Experiment 1.

Letter networks learned their task quickly. Figure 3 illustrates the average activation of letters for each font when that font was used as the test set. The amount of activation of an output unit can be thought of as the level of confidence that the letter unit activated corresponds to the correct letter. Although the letters in some fonts were harder to generalize to than others, the average activations were quite high across all fonts. Accuracy of the networks was also computed: if the activation of the unit corresponding to the correct letter is the highest among all other units, then the network was correct in naming the letter. As expected, all letter networks were able to name the correct letter with 100% accuracy.

Font training Training networks to be font experts (i.e. identify the font a letter is written in) for 15 different fonts proved to be quite difficult. Our networks never satisfactorily learned the problem. In order to determine which fonts were easy enough to learn, we performed multidimensional scaling on the distances between the fonts. Distances between fonts were defined as one minus the

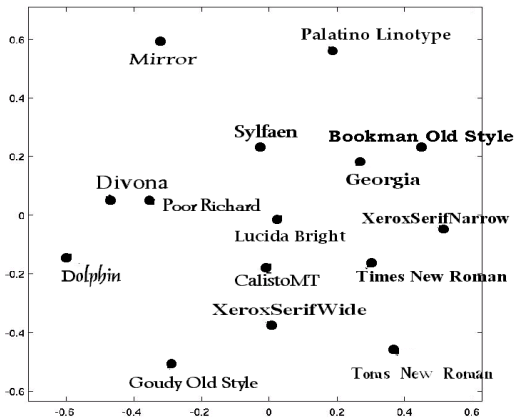

Figure 4: MDS of fonts.

	Avg. RMSE	Avg.Accuracy(%)
Easy Font Network	0.3419	86.19
Hard Font Network	0.4382	76.62

Table 1. Average RMSE and accuracy for font networks

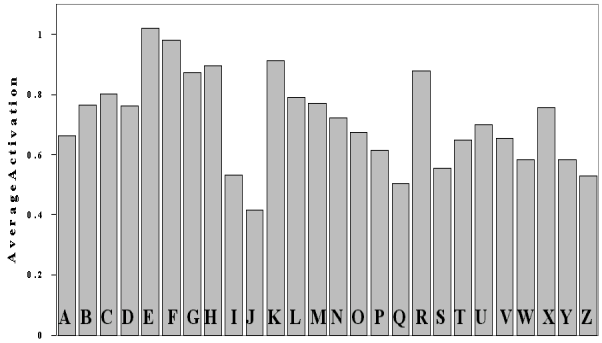

Figure 5: Average activation of fonts across a test letter.

average correlations between their corresponding letters, using their PCA representations. We then formed a 15 by 15 matrix of inter-font distances, and submitted this to a standard non-metric multidimensional scaling routine. The results for a two dimensional solution are shown in Figure 4. The plot shown had a stress of 1.9694.

We used this graph to find the three most separated and the three most correlated fonts. One group of networks was trained on the easier fonts (3 least correlated) and another group of networks was trained on the harder fonts (3 most correlated). Here, 24 letters from each font were used as the training set, 1 letter as the holdout set, and 1 letter as the test set. This was repeated so that each letter had a chance to be the test letter once. Training was stopped when overtraining started to occur. The result is 52 different networks, 26 from the easy font training and 26 from the hard font training.

With these reduced training sets, networks were successfully able to learn to discriminate fonts. Verifying our analysis, the hard font networks had a slightly harder time learning the task than the easy font networks (Table 1). Although the RMSE for both networks were high, they were still accurately able to name the correct font. In fact many of the networks had an accuracy of 100%.

We again computed average activations of output units, except this time for fonts across a test letter (Figure 5). The importance of this plot comes in the activation for particular letters. A high activation means that the network had an easier time generalizing to the font in that letter. This assisted us in choosing the highly generalizable letters as stimuli for Experiment 2.

Experiment 2: Stimuli and Methods

As discussed in the Introduction, Experiment 2 was carried out in order to provide a novel control for our computational model of the visual expertise hypothesis. We used the six

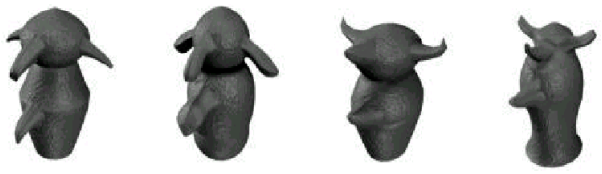

Figure 6: Examples of Greeble images.

most discriminable fonts, and the six letters that were the easiest to generalize to between fonts. Two sorts of networks, both with exactly the same training set, preprocessing, architecture, and number of outputs, were then trained to be either letter classifiers or font experts. While one might consider a letter network an "expert," for our purposes, we consider it a basic level classifier. The main characteristic of basic level categorization is that similar things are classified into the same category. The font expert, on the other hand, must take similar things (the same letter in different fonts) and differentiate between them. Our hypothesis is that such a network will learn the Greeble task faster than a letter network. Thus, training in Experiment 2 was divided into two separate phases. Phase 1 involved training the letter and font networks in a manner similar to that of Experiment 1. In Phase 2, the letter and font networks trained in Phase 1 were then trained to classify Greebles. Examples of Greebles are shown in Figure 6.

Using the results from Experiment 1, the 6 most generalizable letters and the 6 most discriminable fonts were chosen as the stimuli. In addition to these 36 stimuli, 5 different images of 10 unique Greebles were introduced in phase 2. The five different images were produced by jittering the image of a specific Greeble a few pixels on the x, y or x and y axes. Preprocessing of the images was as described in Experiment 1. Greeble images were also preprocessed using Gabor filters and PCA, however they were not included in the generation of the PCA eigenvectors. Rather, the eigenvectors produced via the PCA on the letter/font stimuli were applied to the Greebles. Thus, the PCA representations given to the networks contained no *a priori* information about how Greebles fit into the representational space.

As in Experiment 1, the networks consisted of 40 input units. The hidden unit layer was increased to 40 units due to the increased difficulty of having to solve two tasks. Finally, there were 16 output units, where 6 represented the category (fonts or letters), and 10 the Greebles. Learning rate and momentum remained the same.

Training procedures in Phase 1 were similar to that of Experiment 1, except that only 6 letters and 6 fonts were used. Here, 10 letter networks were trained such that for each network, the letters for a randomly selected font were used as the test set, the letters from another font were the holdout set, and the remaining 4 were used for training. For font networks, each of the 10 networks was tested on the fonts for a randomly selected letter, another randomly selected letter was used as holdout, and the rest were for training. All networks were trained to 2560 epochs. At each log base 2 epoch of training in Phase 1, the weights of the

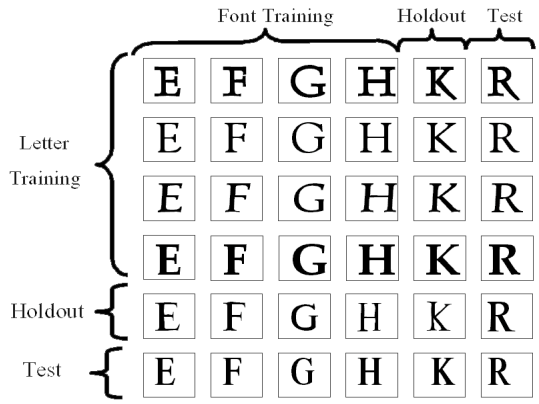

Figure 7: Phase 1 training for Experiment 2. 6 letters and 6 fonts were used.

letter and font network were saved. These weights were used as the starting points for networks in Phase 2 in order to show how varying levels of experience with a preliminary task affected learning of a secondary subordinate level task. For this phase, both font and letter networks ignored the 10 Greeble output units. This training procedure is shown in Figure 7.

In Phase 2 training, the networks trained in Phase 1 were trained to perform subordinate level classification on 10 Greebles. Training for this phase stopped when an RMSE of 0.05 was reached.

Experiment 2: Results

Phase 1 Training Based on the results from Experiment 1, we trained letter and font networks on stimuli that seemed the easiest to generalize to. Both networks were able to learn the task with extremely low error. As expected, the letter networks initially had an easier time learning the letters than the font networks did learning fonts. More importantly, accuracy on the fine-level discrimination task (classifying fonts) became just as good as basic level discrimination (classifying letters).

Phase 2 Training In the second phase of Experiment 2, the letter and font networks were trained to perform fine-level discrimination on Greebles. Again the results were as

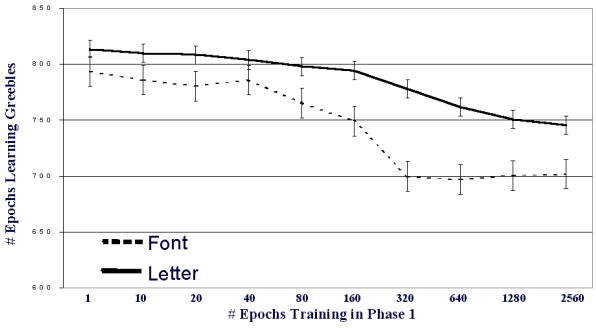

Figure 8: Number of epochs to learn the new task. Error bars denote standard error.

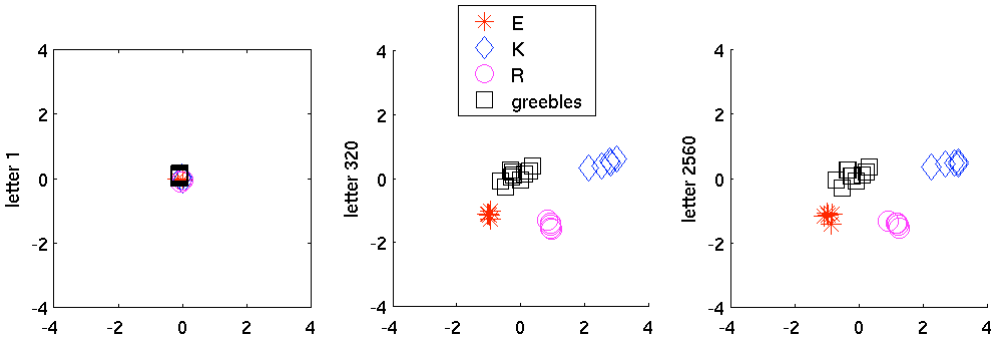

Figure 9: PCA of hidden units of letter network. Grouped by letter.

expected. Figure 8 shows the time in epochs needed for both the letter and font networks to learn the Greeble task as training on the initial task (either classifying letters or fonts) increased. All font networks, regardless of amount of training, learned the Greeble task faster than the letter networks. In addition, more experience with the font task resulted in improvement on learning the Greeble task while more experience with the letter task yielded little improvement (although there is some indication that the letter networks may catch up eventually). Further work will be necessary to evaluate this trend. However, the point remains that expertise in fonts is better than expertise in letters for Greeble training.

To further understand the behavior of the networks, PCA was done on the hidden unit representations prior to Greeble training. Figures 9 and 10 illustrate the spread of the stimuli in representational space based on the 2nd and 3rd principal components (the first PC just codes the overall magnitude change in the weights). In Figure 9 the six points in each symbol represent a given letter in 6 different fonts for a letter network, with one additional symbol representing how Greebles are represented prior to any training on Greebles. In Figure 10, each symbol represents a given font for a font network, and each individual point in that symbol a different letter.

Notice that for the letter network (Figure 9), the letters are grouped together by letter identity regardless of font. Similar inputs (the letters) are made *more similar* by this mapping. In the font network (Figure 10), over training, the fonts spread farther apart over time. Hence, in order to classify the font of each letter, the network must *amplify* small differences between similar items -- all the stimuli representing the same letter must be classified differently. This generalizes to the Greebles; in the font network, the Greebles are more spread out, making it easier for the font network to learn the distinctions between them. Figure 10 also shows that the fonts appear less spread out than the Greebles. This is because the network has learned to see all of the letters in the same font as "the same," whereas it has not learned anything about Greebles yet. It should be noted that each Greeble point is a different Greeble, so the network is already individuating them to some extent. These results are similar to those gathered in our previous network simulations using faces, cups, cans, books, and Greebles (Sugimoto & Cottrell, 2001; Joyce & Cottrell, 2004) illustrating that expertise in the font networks is due to the same mechanism as expertise in face and non-face object networks.

Conclusion

The current studies illustrate that: 1) expertise can be obtained with decidedly non-face-like stimuli and that font expertise exhibits similar properties to that of face and non-face objects seen in previous simulations, and 2) the expertise in previous simulations cannot be explained by a

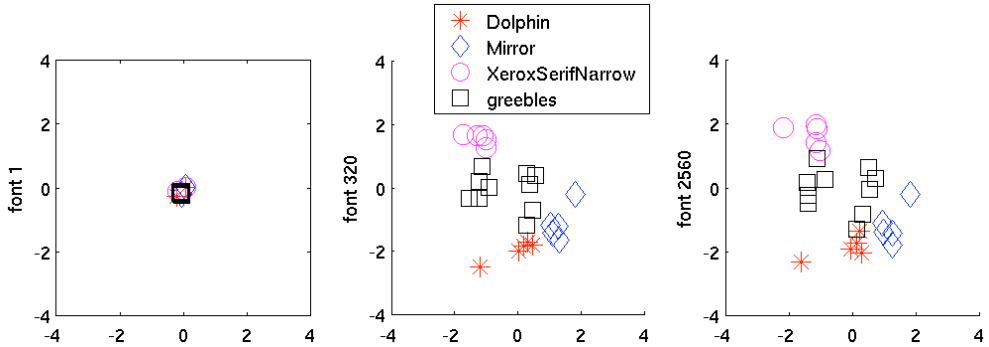

Figure 10: PCA of hidden units on font network. Points grouped by font.

greater number of subordinate level discriminations than basic level discriminations: in the current work these were equated and the results were qualitatively similar to those we have obtained previously.

Our first experiment gave us useful preliminary data for training font experts; it showed that the task of classifying fonts was possible, and revealed which letters and fonts were the easiest to generalize to and train on. The behavior of the networks in the second experiment was similar to previous studies, although our stimuli were different fonts, not "face-like" objects. When training the networks on a new task, the font expert networks learned Greeble classification faster than the letter networks, suggesting that previous visual expertise, whether it be on object or non-object, leads to relatively faster learning in a novel discrimination task. In addition, an equal number of discriminations were required of both letter and font networks. Thus, the expertise advantage could not be due to the sheer *number* of partitions the representational space was divided into, but instead is due to *how* the space was divided. We conclude that visual expertise does not depend on the type of stimuli, nor on the number of stimuli used for training, but on how you slice the space.

Future Work

We plan to train face networks to become font experts, thus generalizing the Greeble expertise work. We expect face expert networks will learn font expertise faster than basic level categorizers. We then plan to train human subjects to become font experts, using fMRI to image both prior to and after training to ascertain if font expertise training engages the FFA. We expect that the letter areas found in the left hemisphere will not become more highly activated by font training. Although it should be obvious from the way that letters are grouped together by the letter network, a future simulation should show that letter networks are difficult to train in font expertise.

Acknowledgements

We would like to thank Gary's Unbelievable Research Unit, the Perceptual Expertise Network, and the anonymous reviewers for comments. This work was supported by NIMH grant MH57075 to GWC and McDonnell Foundation grant #15573-S6 to the Perceptual Expertise Network, Isabel Gauthier, PI.

References

Buhmann, J., Lades, M., and von der Malsburg, C. (1990). Size and distortion invariant object recognition by hierarchical graph matching. *Proceedings of the IJCCN San Diego*, pp. II-411-416.

Dailey, M.N. and Cottrell, G. W. (1999). Organization of face and object recognition in modular neural network models. *Neural Networks*, 12(7-8):1053-1074.

Dailey, M.N., Cottrell, G. W., Padgett, Curtis, and Adolphs, R. (2002). EMPATH: A neural network that categorizes facial expressions. *J. Cog. Neuro.* 14(8):1158-1173.

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

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

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

Gauthier, I., Behrmann, M., Tarr, M. J., (1999a). Can face recognition really be dissociated from recognition? *Journal of Cognitive Neuroscience,* 11:349-370.

Gauthier, I., Tarr, M. J., Anderson, A. W., Skudlarski, P., and Gore, J. C. (1999b). Activation of the middle fusiform "face area" increases with expertise in recognizing novel objects. *Nature Neuroscience,* 2(6):568-573.

Joyce, C. A., Cottrell, G. W. (2004). Solving the visual expertise mystery. To appear in *Proceedings of the Neural Computation and Psychology Workshop 8*.

Kanwisher, N. (2000). Domain specificity in face perception. *Nature Neuroscience,* 3(8):759-762

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.

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

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