Visual Expertise is a General Skill

Maki Sugimoto (mxs@hnc.com)
HNC Software, Inc.
5935 Cornerstone Court West
San Diego, CA 92121-3728 USA

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

Abstract

The fusiform face area (FFA) in the ventral temporal lobe has been shown through fMRI studies to selectively respond with high activation to face stimuli, and has been identified as a face specific processing area. Studies of brain-lesioned subjects with face recognition or object recognition deficits also have often been cited as evidence for face specific processing. Recent studies, however, have shown evidence that the FFA also responds with high activation to a wide variety of non-face objects if the level of discrimination and the level of expertise are controlled. Based on these recent results, we hypothesized that the features of faces that the FFA respond to can be useful for discriminating other classes of visually homogeneous stimuli with some tuning through experience. To test our hypothesis, we trained two groups of feed-forward neural networks on visual classification tasks. The first group was pretrained on basic level classification of four stimulus classes, including faces. The second group was pretrained on subordinate level classification on one of the stimulus classes and basic level classification on the other three. In two experiments that used different criteria to stop pretraining, we show that networks that fully acquire the skill of subordinate level classification consistently show an advantage in learning the new task.

Introduction

The functional role of the so-called fusiform face area (FFA) located in the ventral temporal lobe is controversial. The FFA has been shown in fMRI studies to respond with high activation to face stimuli but not to other visual object stimuli, and has thus been identified as a face specific processing area (Kanwisher, 2000). Studies of patients with face recognition or object recognition deficits also have often been cited as evidence for face specific processing. Recent studies by Gauthier and colleagues have questioned whether the FFA is really a face-specific area (Gauthier, Behrmann & Tarr, 1999a). They have proposed an alternative theory that the FFA engages in expert level classification of visually similar stimuli from a wide variety of categories not limited to faces.

The current study is an attempt to shed light on the debate through simulations of computational models. We constructed our hypothesis based on the recent view that the FFA is a domain-general processing area, specializing in visual expertise of fine level discrimination of homogeneous stimuli. Our experimental results show strong support for the hypothesis, thus providing further evidence for the plausibility of the domain-general view of the FFA.

In the following sections, we will describe the evidence for and against the FFA's face specificity, and our refinement of the domain-general hypothesis. The experimental methods are then described in detail followed by the results. We conclude by summarizing our findings and suggesting future research.

Evidence for the Face Specificity of the FFA

Studies of brain-lesioned subjects provide the strongest evidence for localized face specific processing. Patients with *associative prosopagnosia* reportedly have deficits in individual face identification, but are normal in face detection and object recognition (Farah, Levinson & Klein, 1995). On the other hand, patients with *visual object agnosia* are normal in identifying individual faces but have deficits in recognizing non-face objects (Moscovitch, Winocur & Behrmann, 1997). The two groups of patients serve as evidence for a double dissociation of visual processing of faces and other objects.

Through fMRI studies of normal brains, the FFA has been identified as the area being most selective to faces (Kanwisher, McDermott & Chun, 1997). Prosopagnosia patients usually have a lesion in an area encompassing the FFA (De Renzi et al., 1994), providing consistent evidence for the face specificity of the FFA.

Evidence against the Face Specificity of the FFA

Gauthier and colleagues argued that the FFA showed high activity in response to various classes of visual stimuli when the levels of discrimination and expertise were properly controlled (Gauthier et al., 1999a). One study showed significantly high activity of the FFA for car and bird experts when stimuli from their respective expert class were presented (Gauthier et al., 2000). Another study that utilized 3-D artificially rendered models called "Greebles" (Gauthier & Tarr, 1997), showed the FFA increasing its activation in response to the rendered models as the subjects were trained to classify them at a fine level (Gauthier et al., 1999b). For the latter study, the use of the Greebles allowed the authors to develop human subject experts of non-face objects while fully controlling the subjects' experience with the stimuli.

These results showing high activity of the FFA for non-face objects including completely novel objects,

serve as strong evidence against the face specific view of the FFA.

Our Approach with Computational Models

Why does the FFA engage in expert classification of non-face objects? We hypothesized that the features of faces that the FFA responds to can be useful for discriminating any class of visually homogeneous stimuli with some tuning through experience. If our hypothesis is correct, possession of expertise with faces should facilitate the expert level learning of other classes. In this paper, we consider individuating members of a homogeneous class (subordinate classification) to be an expert level task.

To test our hypothesis, we trained two groups of neural networks with hidden layers to perform a subordinate level Greeble classification task. Prior to training on the Greebles, we pretrained the networks on one of the following two tasks:

1. Basic level classification of faces and objects
2. Subordinate level classification of one of the classes and basic level classification of the rest

Developing the first visual expertise for non-face objects is one of the conditions that cannot be ethically achieved in human experiments. Our computational model attempts to overcome this limitation by pretraining neural networks on subordinate classification of non-face objects. If the advantage can be observed for all groups of networks with various pretraining tasks, we would conclude that the features that are discriminative of homogeneous visual stimuli *in general* are robust features that translate well to any other class of stimuli.

Experimental Methods

As described briefly in the previous section, we trained neural networks on subordinate level classification with various pretraining tasks. In this section, we will describe further details on the input database, the preprocessing procedure, network configurations and the simulation procedures.

Image Database

The images were 64x64 8-bit grayscale images consisting of five basic classes: human faces, books, cans, cups, and Greebles. Each class included 5 different images of 12 individuals, resulting in a total of 60 images for each class. Example images are shown in Figure 1 and 2. The non-Greeble images are described elsewhere (Dailey & Cottrell, 1999). For the Greebles, the 12 individuals were selected exclusively from one of the five families. Five images were obtained for each individual by performing random operations of shifting up to 1 pixel vertically and horizontally, and rotating up to 3 degrees clockwise or counterclockwise in the image plane. A region from the background of the common object images was randomly extracted and applied to the background of the Greeble images.

Figure 1: Example of face and common object images (Dailey & Cottrell, 1999)

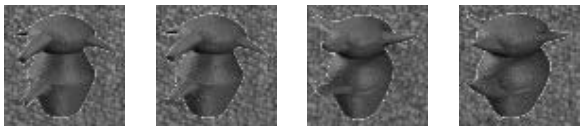

Figure 2: Example of Greeble images;The left two are images of the same Greeble

Preprocessing

To preprocess the images, we followed the procedures introduced by Dailey and Cottrell (1999), applying Gabor based wavelet filters and using principal component analysis (PCA) for dimensionality reduction.

2-D Gabor wavelet filters, which are relatively robust to variations in background, translation, distortion and size (Lades et al., 1993), have been used previously for face recognition tasks with neural networks. Each image was represented by the magnitude of the responses of 40 filters tuned to 8 orientations and 5 spatial frequencies, measured at 64 points subsampled in an 8x8 grid, resulting in a vector of 2560 elements (Buhman, Lange & von der Malsburg, 1990; Dailey & Cottrell, 1999).

PCA was done separately on each spatial frequency, extracting 8 components for each of the 5 scales to form 40-dimensional input vectors. Each element of the input vectors were normalized across all face/object images by z-scoring, i.e., a linear transformation to mean 0 and standard deviation 1. The Greeble patterns were not represented in the principal components to prevent any knowledge of the Greebles contaminating the model.

Network Configuration

Standard feed forward neural networks with a 40-unit hidden layer were used for all the experiments. The hidden layer units used the logistic sigmoid function while the output units were linear. The learning rate and momentum were .005 and .5, respectively. These parameters were tuned so that the networks reliably learned the most difficult task, which was the subordinate level classification on faces and Greebles with basic level classification on the common objects.

Training Set Variations

Each network was trained on a subset of the whole data set as follows. For the classes on which subordinate level

classification were performed, one image for each individual was randomly selected to test generalization. Another image was removed to be used as the holdout set (for early stopping) from the rest of the images, resulting in a reduced training set of 3 images per individual.

For the classes on which basic level classification were performed, images of one randomly selected individual were reserved for testing. Images of a different individual were used as the holdout set, resulting in a reduced training set of images of 10 individuals.

With the arrangements mentioned above, 3 images of 12 individuals were available for use as the training set for the subordinate level classification task and 5 images of 10 individuals were available for the basic level task. In order to control the number of images presented to the networks during the training, the training set was restricted to 3 images from 10 individuals for both levels of classification, for a total of 30 images for each class.

In the experiments reported below, we do not use the holdout and test sets, as we use RMSE thresholds and amount of training as conditions for stopping training phases. The holdout and test sets are used in preliminary experiments to find appropriate values of the RMSE thresholds.

Task Variations

Training of the neural networks was done in two phases. In the first phase, the pretraining phase, the networks were trained using only the face/common object data on one of the following two tasks:

1. Basic level classification on all 4 input categories
2. Subordinate level classification on 1 category and basic level on the rest.

The networks that were assigned the first task had 4 outputs, corresponding to book, can, cup, and face. We will refer to these networks as "Non-experts".

The networks that were assigned the second task had 13 outputs; 3 for the basic level categories and 10 for the individuals in the subordinate level. For example, if a network was assigned a subordinate level classification task for cans and basic level for the rest, the output units corresponded to book, cup, face, can 1, can 2, can 3, etc. We will refer to these networks as "Experts".

In the second phase, the pretrained networks were trained on a subordinate level classification task of individuating Greebles in addition to the pretrained task. Greebles were included in the input data set and 10 output units corresponding to each Greeble were added. Thus, the networks performed either a 14-way or a 23-way classification depending on their pretrained task.

We ran two sets of experiments using different criteria to determine when to stop pretraining:

Experiment 1 The networks were trained until the training set RMSE dropped below a fixed threshold.

Experiment 2 The networks were trained for a fixed number of epochs.

The first criterion controls the networks' familiarity with the input data with respect to their given tasks. This criterion is partly motivated by Gauthier et al.'s definition of experts that takes into account not only the classification accuracy but also the response time which reflects the subjects' degree of certainty. Response time is often modeled in neural networks by the RMSE on a pattern. The second criterion controls the number of opportunities the networks can learn from the input. Employing this criterion corresponds to the idea of controlling the subjects' experience with their tasks, which is often difficult to control in human subject experiments.

For the Greeble phase, the networks were trained to a fixed RMSE threshold for both experiments.

Provided that the networks adequately learned the pretraining task in the pretraining phase, any difference in the learning process of the new task (in the second phase) between the Non-experts and the Experts must be due to the differences in the pretraining task. For the first experiment, we set the pretraining RMSE threshold to 0.0806. This threshold value was determined through preliminary experiments by estimating the training set RMSE for the face expert task to be learned without overfitting. For the second experiment, the epoch limits ranged over $5 * 2^n$ with $n \in 0, 1, .., 10$ to fully analyze the effect of the pretraining task differences. We set the RMSE threshold for the second (Greeble) phase to 0.158. This was determined from similar preliminary experiments based on the estimated optimal RMSE on the most difficult task, subordinate level classification on faces and Greebles.

Evaluation

For the two experiments, we compared the number of epochs required to learn the new task for the Non-experts and the Experts. For experiment 1, we trained 20 networks with different initial random weights for all 5 pretraining tasks, for a total of 100 networks. For experiment 2, we trained 10 networks with different initial random weights for all 5 pretraining tasks for 5120 epochs. We stored the intermediate weights of each network at 10 different intervals ranging over 5 to 2560 epochs, training a total of 550 networks in the second phase.

Results

Experiment 1: Fixed RMSE Criterion

Pretraining

Figure 3 shows the number of training epochs for the two phases averaged across the 20 networks for each pretraining condition. The Non-experts required a much shorter training period than all the expert networks for the pretraining phase, reflecting the ease of basic level classification. For the second phase, the Non-experts were significantly slower than all the Experts in learning the new task ($p < 0.001$, pairwise comparison between Non-experts and the face, can, cup, book experts with $t(38) = 7.03, 5.74, 14.69, 10.76$, respectively). The difference between the can experts and the face experts was insignificant ($t(38) = 1.20, p > 0.2$), the face experts

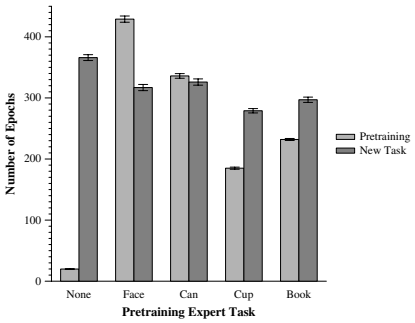

Figure 3: Number of epochs to reach the RMSE threshold. Error bars denote standard error.

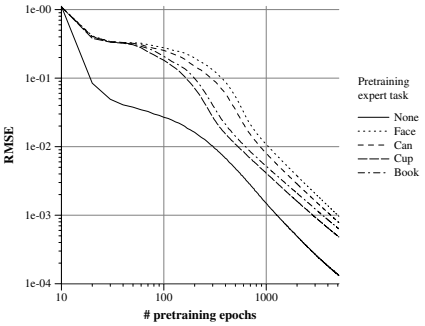

Figure 4: Learning curve for the pretraining phase

Table 1: Training set accuracy for just the Greebles. Figures in parentheses denote standard error.

Expert task	Greebles training set accuracy(%)
Non-expert	71.2 (2.00)
Face	93.5 (1.17)
Can	95.2 (1.04)
Cup	89.8 (2.00)
Book	88.8 (1.46)

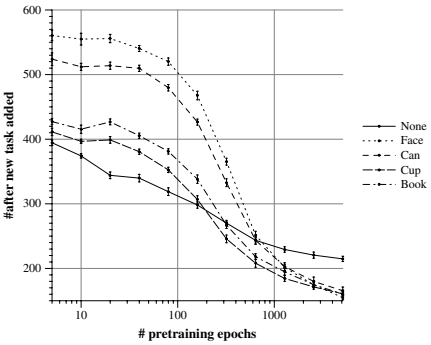

Figure 5: Number of epochs to learn the new task. Error bars denote standard error.

were slower than the book experts ($t(38) = 3.08, p < 0.005$), and the book experts were slower than the cup experts ($t(38) = 3.22, p < 0.005$).

Table 1 shows that despite the overall RMSE having been controlled, the Non-experts were still non-experts at Greebles after training on them. Further training on the Non-experts would have widened the gap between training times on the Greebles for Experts and Non-experts even more. On the other hand, for the Experts, there was a positive correlation between the training set accuracy and the number of training epochs, suggesting the differences in training epochs between the Experts would have narrowed if the training set accuracy on the newly added task had been controlled. Being an expert on some task prior to learning another expert task was clearly advantageous.

Experiment 2: Fixed Exposure Pretraining

Not surprisingly, the Non-experts maintained lower RMSE than all the Experts during the pretraining phase (Figure 4). Among the four groups of expert networks, the face experts had the most difficult pretraining task, followed by the can experts, book experts, and finally cup experts.

For the secondary task training, there was a crossover in effects at 1280 epochs: Fewer epochs meant the non-Experts had an advantage; more meant Experts had an advantage (Figure 5). If the networks were pretrained long enough, the improvement on the error for the pretrained portion of the task became negligible compared to the error due to the newly added task. In this case, we can safely argue that the epochs of training required in the second phase are fully determined by the learnability of the newly added task. If the pretraining stopped prematurely, however, the networks must improve their performance on the prior task as well as the newly added task to achieve the second phase RMSE threshold. All of the networks that were pretrained for at least 1280 epochs achieved a pretraining RMSE that was an order of magnitude lower than the second phase RMSE threshold. Therefore, the advantage that Experts with this amount of pretraining gained must be due solely to their performance in learning the new task.

Analysis: Network Plasticity We hypothesized that the advantage of learning a fine-level discrimination task would be due to greater plasticity in the hidden units. That is, we expected the activations of the hidden units to end up in the linear range of the squashing function in order to make the fine discriminations. This is also a good place to be for back propagation learning, as the higher the slope of the activation function, the faster learning occurs. We therefore analyzed how the features extracted at the hidden layer were tuned by measuring the plasticity (average slope) of the pretrained networks. Our findings surprised us.

We defined a network's plasticity as the value of the derivative of the activation function at the activation level averaged across all hidden layer units and all patterns in

a given set of input patterns:

$$P(S) = \frac{1}{N}\sum_{s\in S}\frac{1}{n}\sum_{i\in I} g'(x_{si})$$

where $g(x)$ is the activation function, S a set of patterns, N the number of patterns in S, I the set of hidden layer units, n the number of hidden layer units, and x_{si} the activation of unit i in response to pattern s. In the online backpropagation learning rule, $g'(x)$ scales the weight changes of the hidden layer units with respect to the errors computed for the output layer. The plasticity of neural networks is usually taken to be predictive of the ability to learn new tasks (Ellis & Lambon Ralph, 2000).

As the activation function, all the hidden layer units in our neural networks used the logistic sigmoid function:

$$g(x) = \frac{1}{1+\exp(-x)}$$

where x is the weighted sum of the inputs, or the inner product of the input vector $\vec{z}$ and the weight vector $\vec{w}$:

$$x \equiv \vec{z}\cdot\vec{w}.$$

The first derivative of $g(x)$ can be written as

$$g'(x) = \frac{\exp(-x)}{(1+\exp(-x))^2} = g(x)(1-g(x)).$$

For $x \in (\infty, -\infty)$, $g(x)$ ranges over $(0,1)$ and $g'(x)$ over $(0, 0.25]$. $g'(x)$ is a bell-shaped function with a global maximum at $g'(0) = 0.25$. By our definition of plasticity, the networks that produce intermediate responses at the hidden layer level would have higher plasticity than networks with bimodal responses. Networks with higher plasticity are generally more likely to learn new tasks faster since the hidden layer units would change their weights more rapidly in response to the errors propagated from the newly added output units.

Network plasticity can also be considered as a measurement of *mismatch* between the hidden layer weights and the input patterns. If the input patterns and the weights were orthogonal, x would be near 0, resulting in maximal plasticity. If, however, the weights tuned for some task matched the input patterns of a new stimulus class, $|x|$ would have a larger value resulting in lower plasticity. The issue is whether these features will be advantageous for learning the new stimulus class. When a network with a low plasticity (high match) measured on novel patterns learns faster on those patterns than a network with higher plasticity, this suggests that the highly matched features are efficacious for classifying the new stimuli.

Figure 6 shows the plasticity of the pretrained networks in response to the training set used for the pretraining and to the set of new patterns which would be added into the training set for the second phase. For both the pretrained patterns and the unseen patterns, Non-experts

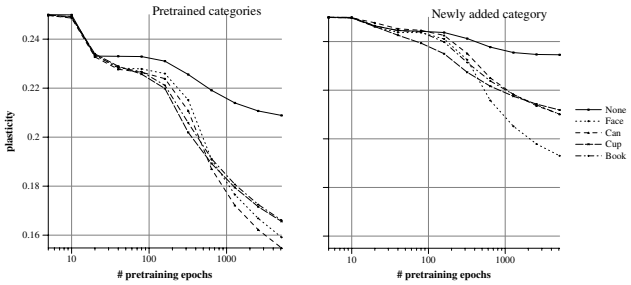

Figure 6: Plasticity of the pretrained networks

retained their plasticity better than all Experts. As we saw in the previous section, however, it was the Experts that eventually showed an advantage in learning the new task. All Experts learned the new task faster as they rapidly lost plasticity over pretraining time. Normally, we would expect the Experts to be generally poorer in learning new tasks due to their low plasticity. These results imply that the advantage the Experts gained in learning the Greebles task cannot be explained as part of a general capability of learning new tasks. Instead, it is more appropriate to interpret this to mean that the hidden unit features were well-matched to the new task.

The lower plasticity of the Experts for the pretrained data implies that the Experts had to finely tune their features to fit their respective expert category, while the Non-experts did not require fine tuning of the features to achieve the lower errors. The eventual advantage of the Experts can then be explained in terms of the features tuned for the expert tasks matching the Greebles data set as well. Given the strong trend regardless of the domain of expertise, we claim that the features useful for one subordinate classification task are general expert features that are good for discriminating individuals of other classes as well. Although other uncontrolled factors such as the length of the weight vectors can influence the plasticity of a network, they seem unlikely to explain our experimental results.

Experiment 2 Summary For longer pretraining epochs, the Non-experts took longer than any of the Experts to reach the final RMSE after the new task was added. While we would expect the Experts to have higher plasticity given their advantage in learning the new task, it was the Non-experts that retained higher plasticity. A comparison between Experts within each task also showed that the networks with longer pretraining and lower plasticity learned the new task faster. The results regarding network plasticity led us to interpret plasticity as a measurement of mismatch specific to a given set of patterns, rather than a predictor of the ability to learn arbitrary new tasks. Given these results, we claim that the underlying cause for the advantage gained by the Experts is the generality of the hidden layer features, fitting well with the subordinate classification task of other classes. This is remarkable in that *overtraining* on a prior task facilitates learning the new one.

Conclusion

Based on the recent studies that showed FFA's engagement in visual expertise of homogeneous stimuli is not limited to faces, we hypothesized that the features useful for discriminating individual faces are useful for the expert learning of other classes. Both of our experiments yielded results in favor of our hypothesis. Furthermore, while faces had a tendency to show the greatest advantage, the results were replicated with networks whose expertise was with other stimulus classes, including cups, cans and books.

The results of the two experiments showed that the possession of a fully developed expertise for faces or non-face objects is advantageous in learning the subordinate level classification of Greebles. Contrary to our expectation that expert networks would show greater plasticity, analyses of network plasticity for Experiment 2 showed that plasticity decreased for the expert networks over time, and it was lower than for non-Expert networks. Indeed, the *lower* the plasticity, the *less* time it took to learn the new task. By reinterpreting low plasticity to mean "high match," we take these results to mean that the features being learned not only match the Greeble stimuli well, but also are the *right* features for fine discrimination of Greebles. Since the choice of Greebles for the second experiment was arbitrary, this suggests that learning to discriminate one homogeneous visual class leads to faster learning in discriminating a new one. Therefore, we conclude that visual expertise is a general skill that translates well across a wide variety of object categories.

Future Work

Firm believers in the face specificity of the FFA might insist that it must be shown that individual neurons in the FFA can simultaneously code features for multiple classes of objects in order their theory to be rejected (Kanwisher, 2000). Even with the advances in brain imaging technology, monitoring each neuron in the FFA is infeasible. Simulations with computational models, however, allow us to monitor the behavior of every single unit in the network.

Naturally, then, one possible extension of the current research is to investigate in detail what the hidden layer units in the expert networks are encoding. Although our experimental results seem to suggest there are features that are useful for visual expertise of any object classes, it is unclear exactly what those features are. Visualization of these expert features would help understand how we develop visual expertise.

References

Buhmann, J., Lades, M., and von der Malsburg, C. (1990). Size and distortion invariant object recognition by hierarchical graph matching. In *Proceedings of the IJCNN International Joint Conference on Neural Networks*, volume 2, pages 411–416, New York. IEEE.

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.

De Renzi, E., Perani, D., Carlesimo, G., Silveri, 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.

Ellis, A. and Lambon Ralph, M. (2000). Age of acquisition effects in adult lexical processing reflect loss of plasticity in maturing systems: Insights from connectionist networks. *Journal of Experimental Psychology: Learning, Memory & Cognition*, 26(5).

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

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. (1997). Becoming a "greeble" expert: Exploring mechanisms for face recognition. *Vision Research*, 37(12):1673–1682.

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.

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.

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

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