UC San Diego

UC San Diego Previously Published Works

Title

Are Face and Object Recognition Independent? A Neurocomputational Modeling Exploration

Permalink

https://escholarship.org/uc/item/9hg7194z

Journal

Journal of Cognitive Neuroscience, 28(4)

ISSN

0898-929X

Authors

Wang, Panqu
Gauthier, Isabel
Cottrell, Garrison

Publication Date

2016-04-01

DOI

10.1162/jocn_a_00919

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Are Face and Object Recognition Independent? A Neurocomputational Modeling Exploration

Panqu Wang[1], Isabel Gauthier[2], and Garrison Cottrell[1]

Abstract

■ Are face and object recognition abilities independent? Although it is commonly believed that they are, Gauthier et al. [Gauthier, I., McGugin, R. W., Richler, J. J., Herzmann, G., Speegle, M., & VanGulick, A. E. Experience moderates overlap between object and face recognition, suggesting a common ability. *Journal of Vision, 14,* 7, 2014] recently showed that these abilities become more correlated as experience with nonface categories increases. They argued that there is a single underlying visual ability, v, that is expressed in performance with both face and nonface categories as experience grows. Using the Cambridge Face Memory Test and the Vanderbilt Expertise Test, they showed that the shared variance between Cambridge Face Memory Test and Vanderbilt Expertise Test performance increases monotonically as experience increases. Here, we address why a shared resource across different visual domains does not lead to competition and to an inverse correlation in abilities? We explain this conundrum using our neurocomputational model of face and object processing ["The Model", TM, Cottrell, G. W., & Hsiao, J. H. Neurocomputational models of face processing. In A. J. Calder, G. Rhodes, M. Johnson, & J. Haxby (Eds.), *The Oxford handbook of face perception*. Oxford, UK: Oxford University Press, 2011]. We model the domain general ability v as the available computational resources (number of hidden units) in the mapping from input to label and experience as the frequency of individual exemplars in an object category appearing during network training. Our results show that, as in the behavioral data, the correlation between subordinate level face and object recognition accuracy increases as experience grows. We suggest that different domains do not compete for resources because the relevant features are shared between faces and objects. The essential power of experience is to generate a "spreading transform" for faces (separating them in representational space) that generalizes to objects that must be individuated. Interestingly, when the task of the network is basic level categorization, no increase in the correlation between domains is observed. Hence, our model predicts that it is the type of experience that matters and that the source of the correlation is in the fusiform face area, rather than in cortical areas that subserve basic level categorization. This result is consistent with our previous modeling elucidating why the FFA is recruited for novel domains of expertise [Tong, M. H., Joyce, C. A., & Cottrell, G. W. Why is the fusiform face area recruited for novel categories of expertise? A neurocomputational investigation. *Brain Research, 1202,* 14–24, 2008]. ■

INTRODUCTION

Understanding how visual object recognition is achieved in the human visual cortex has been an important goal in various disciplines, such as neuroscience, neurophysiology, psychology, and computer science. Among all object classes, because of their social importance, faces have been studied most extensively, especially since the fusiform face area (FFA) was discovered (Kanwisher, McDermott, & Chun, 1997; Sergent, Ohta, & MacDonald, 1992). Some research suggests that the FFA is a domain-specific "module" processing only faces (Grill-Spector, Knouf, & Kanwisher, 2004; Kanwisher et al., 1997; McCarthy, Puce, Gore, & Allison, 1997); however, the FFA responds to nonface object categories of expertise, including birds, cars (McGugin, Van Gulick, Tamber-Rosenau, Ross, & Gauthier, 2014; Xu, 2005; Gauthier, Skudlarski, Gore, & Anderson, 2000), chessboards (Bilalić, Langner, Ulrich, & Grodd, 2011), and even artificial objects when participants are sufficiently trained in the laboratory (Gauthier, Tarr, Anderson, Skudlarski, & Gore, 1999). High-resolution fMRI in the FFA and neurophysiology in macaque's brain reveal the existence of highly selective face areas within the FFA or its likely homologue in monkeys, but no reliable selectivity for nonface objects (Grill-Spector, Sayres, & Ress, 2006; Tsao, Freiwald, Tootell, & Livingstone, 2006). However, when behavioral expertise is taken into consideration, more recent work found a reliable correlation between behavioral car expertise and the response to cars in the FFA, which remains reliable even in the most face-selective voxels in this region (McGugin, Newton, Gore, & Gauthier, 2014; McGugin, Gatenby, Gore, & Gauthier, 2012). They suggest that experience individuating members of a category may be sufficient to create this activation.

A more novel approach to study the relationship between face and object recognition is that of individual differences in behavioral performance. With the development of the Cambridge Face Memory Test (CFMT; Duchaine

[1]University of California, San Diego, [2]Vanderbilt University

© 2016 Massachusetts Institute of Technology

& Nakayama, 2006), reliable individual differences in face recognition abilities have been characterized in the normal population. Using a classical twin study design, Wilmer et al. (2010) provided evidence that face recognition ability is highly heritable. These authors also reported that face recognition ability (CFMT scores) shared very little variance (6.7%) with a test of visual memory for abstract art. In other work, performance on the Cambridge Car Memory Test was found to share only 13.6% of the variance with the CFMT, although the two tests are very similar in format (Dennett et al., 2011). These results suggested that the ability to recognize faces has very little to do with the ability to recognize nonface objects.

Gauthier et al. (2014) challenged this conclusion by gathering evidence for the following hypothesis: Face and object recognition share a domain-general visual ability, v, for discriminating visually similar objects, and this ability will only be expressed in performance when an individual has sufficient experience, E, for a given category. In brief, $Performance_{\text{cat}} \propto v \cdot E_{\text{cat}}$, where the subscript denotes a particular object category. The authors assumed that, for faces, E is generally saturated and makes little contribution to performance (as on the CFMT for instance). For objects, however, they expected E to vary much more across individuals, and as a result, performance should not be as good a measure of v. However, because they conceived of v as the ability that allows people to learn from experience with a category, they predicted that v would be expressed most directly in performance with objects in those people with the most experience. To test this hypothesis, the authors collected three measures from 256 participants: (1) performance on the CFMT, (2) performance with eight nonface categories on the Vanderbilt Expertise Test (VET; McGugin, Richler, Herzmann, Speegle, & Gauthier, 2012), and (3) a self-rating of experience with faces and the eight VET object categories (O-EXP, 1–9).

For the CFMT, participants studied six target faces and finished an 18-trial learning phase. They were then tested with 30 three-alternative forced-choice (3AFC) test displays to determine which faces were among the studied faces. They then studied the target faces again and were tested over 24 test trials, where the stimuli were presented in Gaussian noise. For the VET, participants studied six target exemplars and then performed 12 3AFC training trials with feedback. Finally, they studied the six exemplars again and performed 36 3AFC (without feedback). In these trials, new exemplars from the target categories were used to test whether their learning generalized to new objects within the category.

Participants were divided into six groups based on their level of reported experience with all VET object categories. According to their hypothesis, if the common visual ability v is expressed through experience, then their performance on the VET (O-PERF) should also be more correlated with their performance on the CFMT as experience (E) grows. As predicted, a regression analysis found that, as experience grows, the shared variance between the CFMT and O-PERF increased monotonically from essentially 0 to 0.59 along the six groups (see Figure 2A). The result indicated that the correlation is indeed moderated by experience: When participants had sufficient experience with nonface objects, if they were found to perform poorly (well) with faces, they were found to also perform poorly (well) with nonface objects. This result suggests that data showing no or little correlation between object and face performance result from not taking into account the participant's level of experience with the objects.

These results are consistent with a neurocomputational model of face processing ("The Model" [TM]; Cottrell & Hsiao, 2011; Dailey & Cottrell, 1999). TM has been used to explain how and why an area of visual expertise for faces (the FFA) could be recruited for other nonface object categories: The resources in the face network can be shared with other object processing, provided that this processing is at the subordinate (expertise) level task (Tong, Joyce, & Cottrell, 2008; Joyce & Cottrell, 2004).

The present implementation of TM is similar to the expert network described in Tong et al. (2008): (1) images are preprocessed by Gabor filters, modeling V1; (2) the Gabor representation is analyzed by PCA, which we consider to correspond to representations in the occipital face area; and (3) a neural network with one hidden layer is trained to recognize individual faces. The model is then trained on object categories at the subordinate level. That is, we assume that experience with a category leads to recognition at the subordinate level (e.g., white, brown, and portobello mushrooms).

Because this is an individual differences study, one network corresponds to one participant. We used individual behavioral data from Gauthier et al. (2014), including CFMT scores, VET scores, and VET category experience scores. Because Gauthier et al. (2014) found self-report of faces to be less reliable than that for objects, we simply assumed that all participants have a very large amount of experience with faces, so that their CFMT score represents their domain general ability v. We therefore identify v with the CFMT score and map that score to the number of hidden units. We map the self-rated experience score E to the number of appearances of individual items within a specific category during training. As described above, we first train the network on faces to simulate the ability expressed by the CFMT performance and then train on three nonface object categories (butterflies, cars, and leaves) to simulate the abilities tested by the VET. We show that the shared variance between the recognition accuracy on faces and the average recognition accuracy on nonface objects increases as experience with the nonface object categories increases, consistent with Gauthier et al.'s data.

In Gauthier et al., the correlation with VET scores was not obtained when they used data from a single category on the VET. Instead, they had to average over the experience with all VET categories, which we replicated here. However, when we increased the number of participants

(networks), we found correlations based on single categories. Consequently, we predict that the correlation between scores on the CFMT and on the VET will be observed depending only on experience with a single category, if enough participants are tested. This prediction of the model has yet to be tested.

Furthermore, we show that the effect of experience moderating the correlation between VET and CFMT scores is not observed in our model if it is only trained to make basic level categorizations; hence, we predict that this effect is carried by the FFA. This suggests that CFMT scores should have the increasing correlation with VET scores based on not only mere experience with a category but also the kind of experience with a category, where members of the category are processed at the subtype level.

Finally, we run an analysis on the net input of hidden units in two networks with different levels of experience and show that the power of experience is to expand the representational space to a larger region, where each individual object is more separated. The experience moderation effect is a direct reflection of this power. This phenomenon is also consistent with previous research using TM that demonstrates why the FFA is recruited for other domains of expertise (Tong et al., 2008).

METHODS

Architecture of TM

In general, TM is constructed using four layers that represent the human visual system from low-level features to high-level object categorizations (Figure 1). Given an input (retina level), we first pass the stimuli through a layer of classical Gabor filter banks, which represent the receptive fields of V1 complex cells (Daugman, 1985). The Gabor filters are composed of five spatial scales and eight orientations. In the second layer, the Gabor filter responses are processed using PCA to reduce the dimensionality and perform efficient coding. The PCA layer models the information extraction process beyond primary visual cortex, up until lateral occipital region (LOC). We think of this layer as the structural description layer from the classic model of Bruce and Young (1986), that is, the level where the representation is suitable for face recognition and facial expression analysis. Because PCA can be implemented using a Hebbian learning rule (Sanger, 1989), we consider this step to be biologically plausible. The next layer is the hidden layer in the neural network. We consider the number of hidden units as the available resources for the task. At this layer, features are learned through backpropagation that are useful for the task. For example, if the task is to discriminate different faces, this layer will learn face-related representations adaptively through learning, and we can assume this layer corresponds to the FFA. If the task is to classify basic level object categories, the layer will learn basic-level-related representations, modeling those in the LOC. The fourth layer is the output layer, which represents the categories of the different objects. This simulates the category cells in pFC. At each layer of the preprocessing network, there is a normalization step before giving them to the next layer. Each image pixel value is z scored independently across the image set; the Gabor filters are normalized to be a percentage of the total responses of the eight orientations for each location, scale, and image; and each principal component value is z scored across the data set.

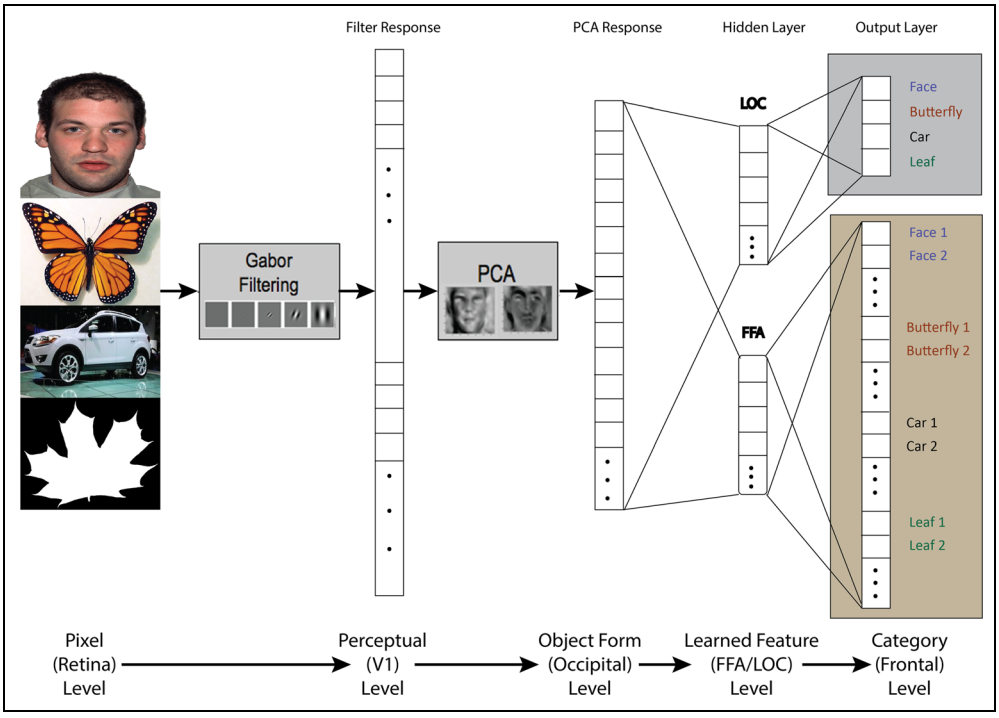

Figure 1. Model architecture.

Data Set and Preprocessing

We use four object categories in all of our experiments: butterflies, cars, faces, and leaves. The three nonface object categories are three of the eight VET categories. The reason we chose these three domains is that there are readily available data sets for these VET categories that include subordinate level labels. We collected the images from four separate data sets: (1) faces, the NimStim Face Stimulus Set (has 646 images across 45 individuals; Tottenham et al., 2009); (2) butterflies, the Leeds Butterfly Data Set (has 832 images across 10 species; Wang, Markert, & Everingham, 2009); (3) cars, the Multi-view Car Data Set (has approximately 2000 images for 20 models; Ozuysal, Lepetit, & Fua, 2009); and (4) leaves, the One-hundred Plant Species Leaves Data Set (has 1600 images for 100 categories; Mallah, Cope, & Orwell, 2013). For every object category, we randomly chose 16 images from each of 10 randomly selected subordinate level categories to form the training set (12 images per individual) and test set (four images per individual[1]). We first transform all images to grayscale and crop them to a uniform size of 64×64 pixels. We then process them through Gabor filter banks as defined in Lades et al. (1993), with eight different orientations ranging from 0 to $\frac{7\pi}{8}$ and five spatial scales. To make the filter response values in the same range, we normalize them across orientations for each scale on a per-image basis, so there is a low-frequency to high-frequency representation of the image. We normalize them across orientations for each scale on a per-image basis, so there is a low-frequency to high-frequency representation of the image. We normalize the response this way because we hypothesize that the downstream cells perform similar normalizations as the retina, which performs contrast normalization. In addition, this representation equalizes the power across spatial frequencies, so none dominate the representation. We sample the 40 Gabor filter responses in an 8×8 grid over the image, resulting in a 2560-dimensional vector to represent a single image. The PCA step removes the redundancy of this representation by decorrelating the filter responses and generates a lower dimensional vector for efficient further processing. We perform PCA separately on the five scales, keep the eight eigenvectors with the largest eigenvalues for each scale, and project all Gabor filter responses for each image onto the corresponding eigenvectors. The 40 projections are z scored by dividing by the square root of the corresponding eigenvalue before presentation to the neural network.

As in previous work (Tong et al., 2008), the label we give to the hidden layers (LOC or FFA) depends on the level of categorization. We hypothesize that LOC performs basic level categorization and FFA is involved in fine level discrimination. As we showed in previous work, this changes the representation at the hidden layer dramatically, in that hidden units in the LOC model clump categories into small regions of representational space, whereas the hidden units in the FFA model increase within-category distance, spreading members of a category out into different locations in representational space.

Mapping and Network Training

To model Gauthier et al.'s experiment, we represent each participant by one network. The data for each of the 256 participants are used to set the parameters for each network. In the psychological experiment performed by Gauthier et al. (2014), there are two key variables: the domain general visual ability, v, and the self-reported experience of the object categories, E. On the basis of Gauthier et al.'s theory, we write the following relation: $Performance_{\text{cat}} \propto v \cdot E_{\text{cat}}$. That is, v is only expressed in performance via experience with a category. We can nominally think of E as a number between 0 and 1, although this is transformed in our model to a more relevant range of experience for the networks. We assume that the maximum value of E is the value for faces (every participant has maximal experience with faces), which means we can measure v directly from each participant's data as their performance on the CFMT (a number from 0 to 1). E is given directly in the self-report data (a number from 1 to 9).

We assume that v is based on the available representational resources of the participant for processing faces and objects; hence, we map v to the number of hidden units in each network using a simple function. With more hidden units, the network in general will generate higher dimensional and more accurate features for a given object category, thus improving the classification performance. We choose the particular mapping through cross-validation so that we do not use too many hidden units for the size of our data set, which would result in poor performance from overfitting.

We use a linear function to map the reported experience (i.e., $3 + E$) to the frequency of individual exemplars in an object category in the training set. In Gauthier et al. (2014), the test–retest reliability for the self-reported experience measure, O-EXP, for nonface object categories is much higher than that of faces (.60). As noted above, we imagine that face experience is maximal for each participant, and for the other categories, we use a linear mapping from the self-reported O-EXP, as the simplest possible unbiased estimate of the relationship between reported experience and training examples. Because, in our database, we have 12 images each of 10 subordinate categories for each type (faces, cars, leaves, and butterflies), if a participant has Experience level 1 with leaves, they will see four exemplars of each leaf or 40 images of leaves. If they have Experience level 9, they will see all 12 exemplars. We repeat the smaller number of exemplars to match the number of training instances in a model network's "day." Hence, we are mapping O-EXP to the variety of experience with an object category.

For the faces, always use all 120 images of 10 people in the training set. The scaling above is calibrated to reach

480 updates of the weights per epoch, again, providing each network with an equal length "day." Hence, given a fixed training time (e.g., one epoch), different object categories have a different variety of training examples based on their level of experience. This mapping is reasonable given that more experience with a category should lead to more variety of experience with a category. Consider, for example, that a good chef will know many different varieties of mushroom, where a less experienced cook may know only two or three.

As a result, our variable mapping and general training process of the network are as follows: We map v to the number of hidden units and E to the amount of training examples that appear at each training iteration. For each network, we train on subordinate level face identification first to simulate the process of gaining expertise on faces. This is intended to reflect the fact that, before humans become familiar with the various species of butterflies, for example, they had expertise on faces. After training on individuating faces, we add the three nonface object classes (butterflies, cars, and leaves) into the network by adding extra sets of output nodes and new training examples. In Experiments 1 and 2 in the next section, as the task is to discriminate the 10 individuals in each category, all networks have 40 output nodes. In Experiment 3, as we only perform basic level categorization of the nonface categories, the network only has 13 output nodes, with 10 for individual faces and three for each nonface object category. We measure the recognition accuracy on the test set for each object when we finish training and use this score to model the VET performance.

RESULTS

We will describe three simulations in this section. The first experiment is intended to model directly the psychological experiment performed by Gauthier et al. (2014) that showed that the correlation between performance on the VET and the CFMT increases with experience with objects. In that experiment, the level of experience was averaged across categories because they did not find a correlation between performance on the VET for a single category based on experience with that category. The second experiment provides a prediction that, if more participants were used, the correlation would emerge at the single category level. In the first and second experiments, the networks were trained to be "experts" in the categories, that is, they were trained to individuate people, car models, and butterfly and leaf species. This suggests that the correlation emerges as a result of shared variance within the FFA. The third experiment predicts that we would not see the experience moderation effect based on basic level experience—expertise is necessary. Finally, we analyze networks trained to be experts to show why the experience moderation effect appears when using the same hidden units, counter to the intuition that there should be a competition for shared resources.

Experiment 1: Modeling Gauthier et al. (2014)

Gauthier et al. hypothesized a single underlying visual ability, v, that is only expressed through experience. This visual ability can be measured by performance on a face recognition test like the CFMT, as we all have a great deal of experience with faces. If v is a shared ability, it should become expressed in performance as a function of experience with nonface objects.

To model their experiment and results, we make a one-to-one mapping of v and E to our neural networks, with each network representing one human participant. Because $Performance_{\text{cat}} \propto v \cdot E_{\text{cat}}$ (according to Gauthier et al.'s hypothesis) and every human participant is assumed to have high and relatively similar experience with faces, their v is explicitly expressed by their face recognition score on the CFMT. We therefore initialized the network based on the participant's CFMT score by mapping that number to the number of hidden units according to the following formula:

$$N_{\text{hidden}}(s_{\text{net}}) = \lfloor 34 \times \text{CFMT}(s_{\text{h}}) - 14 \rfloor$$

where s_{h} represents a particular human participant, s_{net} is the corresponding network modeling that participant, $CFMT(s_{\text{h}})$ is the percent correct of s_{h} on the CFMT, and $N_{\text{hidden}}(s_{net})$ is the number of hidden units for that participant's network. The CFMT scores in Gauthier et al.'s data range from 0.4722 to 1, so N_{hidden} ranges from 2 to 20. As, in general, N_{hidden} must be matched to the size of the data set for good generalization, our range of hidden units is chosen by cross-validation to ensure that there are sufficient resources at the maximum number to provide good generalization without overfitting.[2]

Similarly, the formula for mapping self-rated experience (O-EXP) to the number of training samples for each subordinate object category is as follows:

$$N_{\text{sample}}(\text{category}, s_{\text{net}}) = 3 + \text{O-EXP}(\text{category}, s_{\text{h}})$$

As O-EXP ranges from 1 to 9, then the number of training samples ranges from 4 to 12 (12 is the maximum number of individual training samples in the data set for each individual). Hence, we use a fraction of the data set to learn each object when the participant has lower experience, whereas we use the full data set to train the networks with the highest experience. For faces, we assume O-EXP is 9. Note that, as described above, we must ensure that the networks are trained with the same total amount of images per epoch so that every network has the same number of updates. That is, there is the same number of "hours in the day" for each network. We set this number to 480, as this is the size of the most diverse training set (120 images of 10 individuals for four categories). We use N_{sample} to compute a proportion of the data set. That is, assuming leaves and cars are the only two object categories for the moment, N_{sample} if for leaves and cars is 6 and 12, respectively (with $N_{\text{sample}} = 12$ by definition for faces), the

proportion of the training set that are leaf images is 6/(6 + 12 + 12) or 20%.

We use stochastic gradient descent (online backpropagation) to train the network. A learning method with equivalent results to backpropagation, contrastive Hebbian learning, can be implemented in a biologically plausible way (O'Reilly, 1996; Plaut & Shallice, 1993). Although less biologically plausible, backpropagation training is much more efficient than contrastive Hebbian learning. The input vectors are z scored, and the weights are drawn uniformly from the range of −0.5 to 0.5. In all experiments, we set the learning rate to 0.015 and momentum to 0.01. As mentioned in the Methods section, we train the network on individuating faces first. We stop the face network training in either one of two conditions: If it hits the stopping threshold (mean squared error of 0.005, determined using cross-validation to provide the best generalization) or if the number of training epochs reaches 100, when we assume the network has gained sufficient expertise on face recognition as the training time is enough. We then start the second training phase by introducing the three nonface object categories into the training set and add 30 output nodes, corresponding to subordinate level categorization of the 10 individuals in the three categories. The network is trained until the error is below 0.005 or training epoch reaches 90. At the end of the training process, we measure the recognition accuracy on the test set for all four object categories and calculate the correlation between the score on faces and averaged nonface objects. We show the result in Figure 2B.

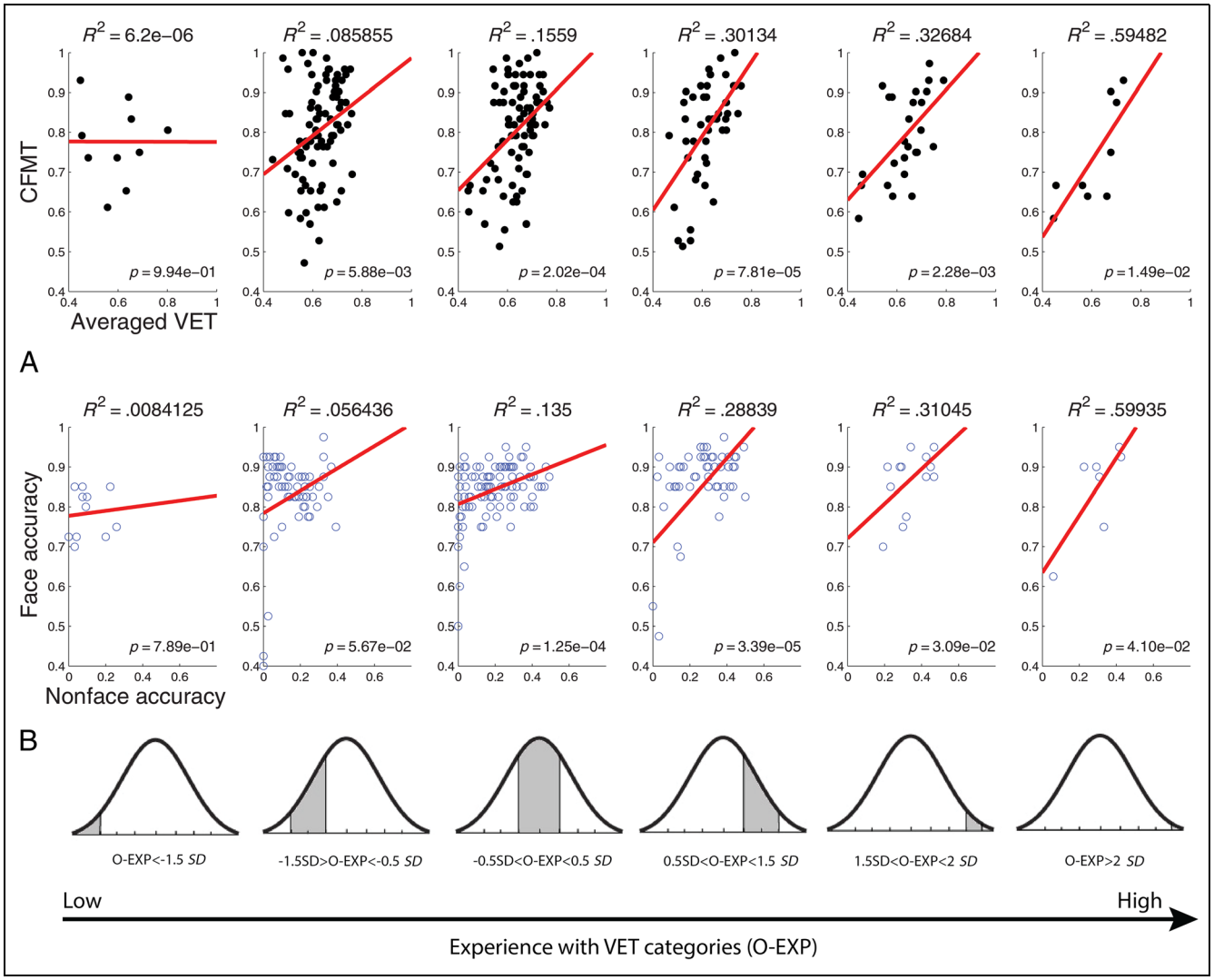

Figure 2. Results of Experiment 1. The first row (A) shows the experimental data from Gauthier et al. (2014). The second row (B) shows our modeling result. Each dot in B represents a single participant network whose parameters (v and E) are calculated based on the corresponding human participant. Each line in the graph represents the regression for each group between their CFMT scores against their VET or nonface object recognition scores. The bottom row shows how the participants are divided into six groups based on their self-rated experience score in VET object categories (O-EXP). For example, the second column (top row) shows the data from participants (dots) whose O-EXP score is between 1.5 and 0.5 *SD*s below the mean.

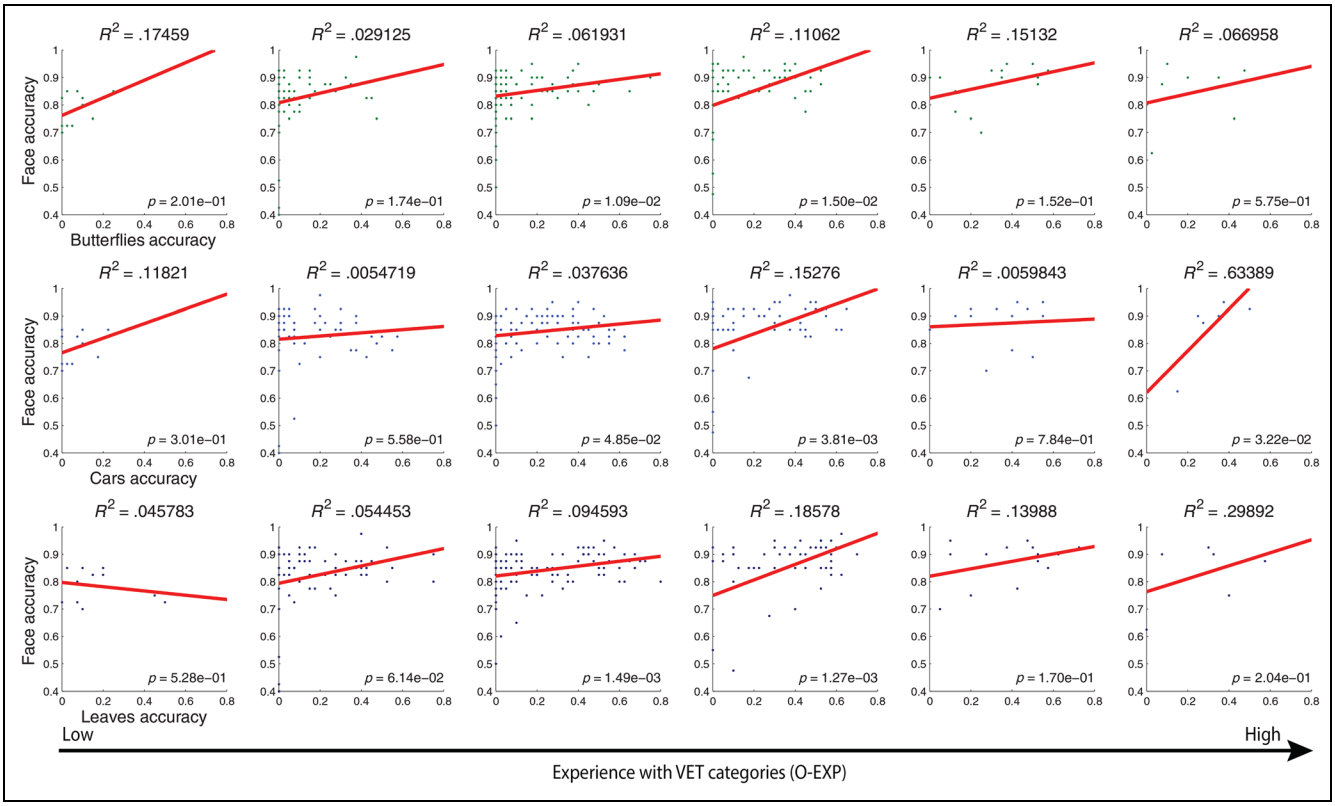

Figure 3. Result showing the correlation between the networks' face recognition performance and single nonface object recognition performance (butterflies, cars, and leaves) in Experiment 1, as a function of experience. Interestingly, although there appears to be an overall trend of increasing correlation (especially for the leaves), it is generally smaller and not monotonic when compared with the result using averaged performance (Figure 2B).

From Figure 2B, we can clearly see that, as experience (O-EXP) grows, the shared variance between face recognition performance and the averaged nonface object recognition accuracy increases monotonically from 1.2×10^{-4} to 0.60585. This result matches those of Gauthier et al. qualitatively and demonstrates that our network training strategy and variable mapping of v and E are reasonable. The mapping of v to various numbers of hidden units in the network spans the accuracy of face recognition (y axis of Figure 2B), suggesting the hypothesis that the variance across individual participants in the domain-general object recognition ability is the amount of representational resources in cortex (hidden units in the neural network). The mapping of E to the number of training examples on nonfaces spans the accuracy of nonface object recognition (x axis of Figure 2B), clearly illustrating that higher experience will generally facilitate object recognition performance by moving them from all being relatively low to a range of scores, expressing the underlying computational resources.

Experiment 2: Correlation with a Single Category

In Gauthier et al. (2014), the increasing trend of correlation was not observed for any individual category. Rather, it only appeared for the averaged VET score (O-PERF) against the CFMT score. This is theoretically problematic because, according to their hypothesis, v is a domain-general visual ability and face recognition should not be independent of any nonface object category when people have sufficient experience in that category. In the original study, this situation was attributed to the fact that self-reports were likely very imperfect measures of experience with a category. However, in the present simulations, experience had a very direct mapping to each network's training, and yet, we also did not see the phenomenon clearly in our simulations when using individual categories (see Figure 3). One possible explanation is that more participants are required to show the effect as there are few "experts" in the general population. In this experiment, we use a much larger number of participant networks and ability levels. We expect to see the same experience moderation effect as in the averaged category result if our assumption is true.

In this experiment, we use 1000 different networks rather than the 256 in the previous experiment. To produce a larger range of network performance, we extended the range of hidden unit numbers and experience levels. We manually created the initialization of the values of v and E for the participant networks. We map v to the range $N_{\text{hidden}} \in \{1,3,5,7,9,12,15,18,21,24,28,32,36\}$. We determined in advance that there is still no overfitting with up to 36 hidden units. For E, we set the range of experiential variety to $N_{\text{sample}} \in \{2,4,6,8,10,12\}$. As before,

higher numbers of samples indicate more varied experience with that category. The number of participant networks at each level of E and v is determined by a Gaussian distribution, and the number of training examples falls in the given interval from 2 to 12. This approach tends to assign more members to the middle value in the set, simulating the fact that most people should have intermediate level of E and v. The training procedure, data set we use, and network parameter settings are the same as in Experiment 1. We show our result in Figure 4.

As can be seen from Figure 4, as experience grows, the shared variance (R^2) between face and all three individual nonface objects increases monotonically, from a value near zero ($p > .1$) up to a value greater than 0.7 ($p < 5 \times 10^{-5}$). Not surprisingly, when we calculate R^2 between face and averaged nonface performance, the increasing correlation trend still exists, from .048 ($p = .1873$) to .829 ($p < 10^{-6}$). We ran the experiment 10 times, and the increasing correlation trend is very robust. The number of participants is one factor in observing the experience moderation effect at the single-category level. A possible explanation for this finding is that using the averaged category experience leads to an aggregation effect (Rushton, Brainerd, & Pressley, 1983). At the single-category level, the smaller amount of data at any level of experience will be more variable because of factors such as different initial random weights, different local minima, noise, and so forth. With several categories, these uncorrelated sources of noise are reduced. With more participants at any given level of experience, we can also eliminate this nuisance variance, as long as it is not correlated across different participants with similar experience, in the same way as it was not correlated across different categories for the same participants. Our

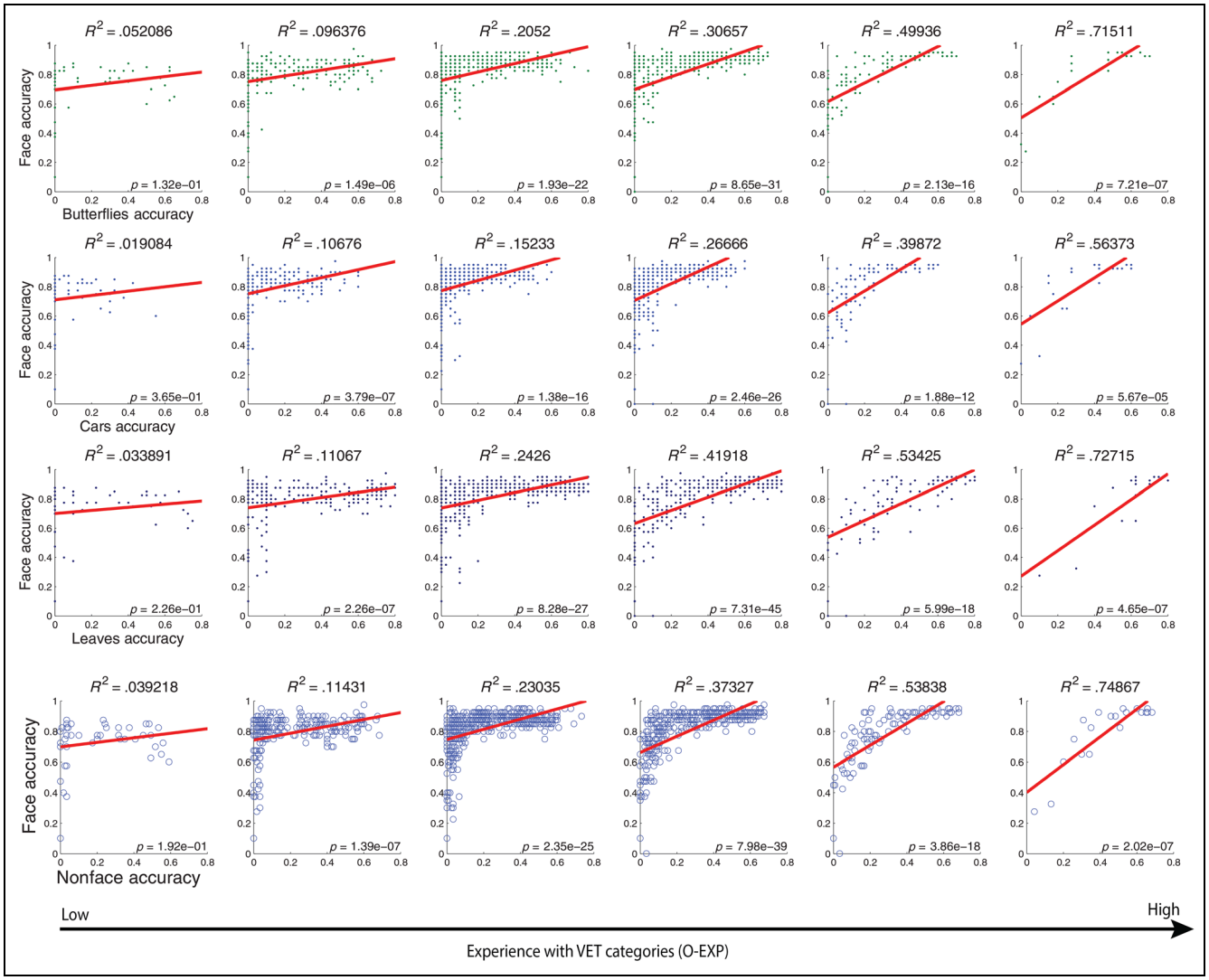

Figure 4. Results of Experiment 2. The top three rows show the trend of shared variance between face recognition accuracy (y axis) and single nonface object recognition performance (x axis; butterflies, cars, and leaves for each row), as a function of experience. The last row shows the correlation on averaged nonface object recognition performance. Each dot represents a participant network, and the red regression curve is also plotted for each group. As we can see, the correlation is monotonically increasing when experience grows, regardless of whether individual or averaged performance is used.

finding predicts that, if more participants were recruited, the experience moderation effect would be found at the single category level in actual behavioral data.

Experiment 3: Basic Level Classification

In Experiments 1 and 2, the networks representing each participant are trained and tested on subordinate level classification tasks, which means their job is to discriminate individuals (Is it Tom/John/Mary…or Benz/Ford/Toyota…?) within a category (faces or cars). That is, the networks are trained to become experts in these specific tasks. On the basis of our previous modeling (Tong et al., 2008) and fMRI (Wong, Palmeri, Rogers, Gore, & Gauthier, 2009) work, we would expect the FFA to be a main site for such subordinate level learning. This however begs the question, does the overlap in abilities we and Gauthier et al. (2014) have measured depend on expertise at the subordinate level? In other words, would we see the same result of experience moderating the relationship between face and object recognition if the networks were instead trained on basic level categorization?

Hence, in Experiment 3, we test this hypothesis by performing the same experiment on our networks but training the network that has been pretrained on faces to classify the objects at the basic level. In a previous modeling study (Tong et al., 2008), they analyzed the effect of both subordinate level and basic level classification tasks using the same neurocomputational modeling approach we use here and found that there is a large difference in the hidden layer representational space developed over training in basic versus expert level categorization. Here, we investigate the result of an expert network (a face identification network) being additionally trained to be a basic level categorizer, and we compute the correlation between face identification performance and basic level categorization performance within the same network. If we still observe the experience moderation effect, it would indicate that the experience moderation effect is not specific to the subordinate level; if not, it suggests that the experience moderation effect requires that participants' experience be at the level of subordinate level categorization or at least rules out that it works for any training task.

To model the basic level classification task, the only change we make from Experiment 2 is altering the number of output nodes and collapsing across individuals. We keep training the output nodes for faces to make sure the model remains effective at individuating faces. As we have 10 individuals for each of the four object categories, all networks in Experiment 2 have 40 output nodes; here, the networks only have 13 output nodes (10 faces + 3 nodes representing each nonface object category: butterflies, cars, and leaves). The variable mapping and training procedure are otherwise the same as Experiment 2. The result is shown in Figure 5.

As can be seen from Figure 5, as experience grows, we do not observe increasing correlation between face and nonface recognition performance, no matter whether experience is measured based on a single category or across categories. Instead, we observe a relatively constant correlation between performance in the two domains, regardless of how much experience the network has on objects. For the correlation results on single categories, we either find no correlation (leaves) or nonmonotonically increasing low correlation (butterflies and cars). When performance is averaged across categories, however, because of the effect of aggregation, the overall correlation increases to around .35; nevertheless, the correlation does not monotonically increase as experience grows.

This phenomenon is easily explained. In Experiments 1 and 2, the variation across domain-general visual ability (v) allows the networks to express the full range of face recognition ability, with the face recognition performance spread out between 0 and 1 (y axis in Figures 2, 3, and 4). However, because of the constraint of experience for the nonface objects, the network cannot express the full range of object recognition ability until the experience level is high. This can be seen from the results in Experiments 1 and 2 (x axis in Figures 2, 3, and 4), where the dots are "squeezing" around zero for low-experienced objects and gradually spread out when experience increases. In general, the cause for low recognition performance is either that the participant network has low v (few hidden units) or because the subordinate level task is very hard and the resources are not sufficient.

In basic level categorization, however, the task is easier (the networks only have to recognize all leaves as leaves, all butterflies as butterflies, etc.), and to do so, the networks do not need a large number of hidden units, nor do we need very much training. Hence, all of the networks (and by inference people) have enough resources to attain a relatively high score on basic level object recognition. This is shown clearly in Figure 5: Face recognition performance is spread out as usual (y axis), and object recognition performance (x axis) has much lower variance in general. This explains why the correlation in the low-experience bins is approximately the same as in the high-experienced bins, and the increased in correlation with face recognition performance from the lowest level of experience (.32) to highest level of experience (.41) is not as large as in subordinate level classification (Figure 4, from .05 to .83). Experience does not mediate performance in an easy task such as basic level recognition, as the performance is dominated by the relative easiness of the task.

Hence, we infer that the type of experience matters in deciding how abilities in different domains overlap: Knowing the kind of leaf, car, or butterfly leads to an increasing correlation of performance with face recognition, whereas only knowing that a leaf is a leaf and so forth does not. The level of task, even if both tasks involve

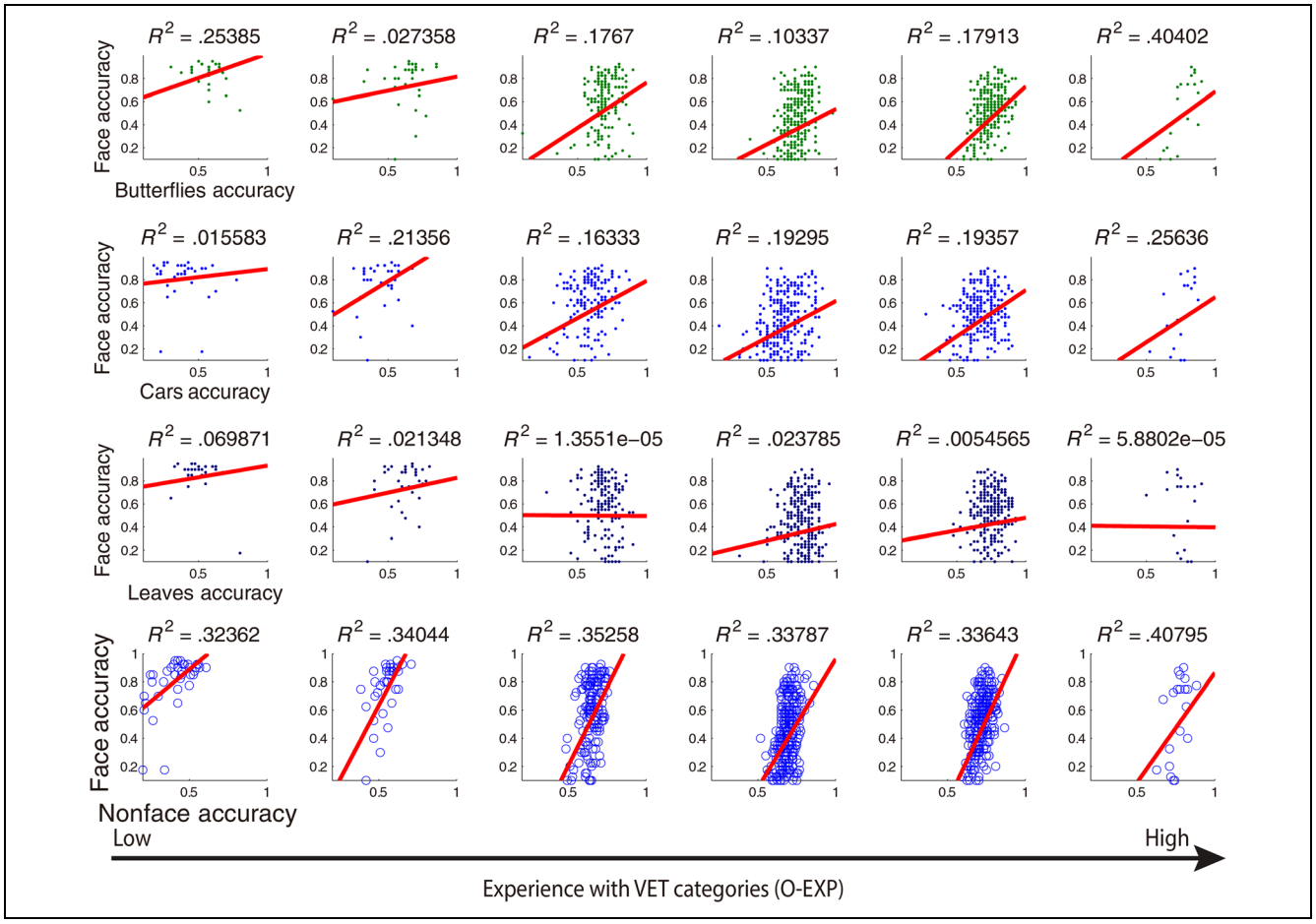

Figure 5. Result of Experiment 3 (basic level classification). The format is the same as Figure 4, with the top three rows showing the correlation between performance on face and single nonface objects and the last row showing the correlation on averaged nonface objects. There is no monotonically increasing correlation in either single or averaged category performance.

categorizing images, has significantly different impacts on the outcome of the experiment. The need to differentiate between individual objects within a visually homogeneous category, rather than placing them into categories that differ in the overall part structure, produces the moderation effect shown in Experiments 1 and 2.

Analysis: The Power of Experience

Given the finding that more experience leads to higher correlation between subordinate level classification tasks in Experiments 1 and 2, we may wonder why this happens. For example, it seems intuitive that, if the same hidden units are being used in both tasks, then there should be competition for these representational resources, and higher performance on faces should mean that more hidden units are dedicated to faces, which would result in lower performance on objects. This turns out not to be the case. Tong et al. (2008) showed that the hidden unit representation learned in a face identification task separates faces in hidden unit space, making it easy for a classifier to separate them. However, this same "spreading transform" generalized to novel categories. For example, they showed that, when novel objects ("Greebles") were presented to the trained network for the first time, without any training, they were already spread out in hidden unit space. In this experiment, using a similar analysis of the net inputs of the hidden units, we show how this effect develops as a result of experience.

More specifically, we analyze the hidden units on two participant networks with different levels of experience. Recall that we map experience (E) to the number of training examples per individual. For this analysis, we set the number of training examples per object for the two networks to be 3 and 12, respectively, representing low and high levels of experience. Both networks have 50 hidden units, so they have sufficient ability (v) to give the best performance. We train both networks on individuating faces first and continue training on recognizing mixed object categories. We measure the performance at the end of training. During training, we record the net input of the hidden units for all training examples at six different time points (see Figure 6), which enables us to observe the evolution of the internal representation. For data collected from each participant network, we

perform PCA on them and visualize the projections on the first and second principal components on a 2-D subspace. The result is shown in Figure 6. Note that, for Columns 1 and 2, the different colors represent different faces to show how the faces are separated in the space. Although some faces look like they are close in the space, they are separated by other dimensions. For Columns 3–6, the different colors represent different categories.

Several conclusions can be drawn from the results. First of all, for the networks trained on face recognition only (the first two columns), no difference in experience exists, so the representations that the networks develop are similar: Training on differentiating faces gradually separates each individual face in the subspace (second column), compared with the initial cluster at the center (first column). Second, when we take a close look at the third column, the nonface objects are already dispersed to the extent of the representational space formed by faces, even without training. This suggests that the projection into the hidden unit space learned for faces, which spreads out the faces, generalizes to novel objects, spreading them out as well. This is the same finding as in Tong et al. (2008), where it was shown that Greebles were spread out by the face network before it was even trained on Greebles. In that article, we also showed that there was nothing special about faces per se; rather, it is the task that is learned (individuation of similar looking items) that leads to this spreading transform. This result held for our model of the FFA, which suggests that the effects found in the Gauthier et al.'s (2014) article are also a reflection of expertise with the nonface categories. Finally, when training on multiple object categories, we find that more training generally produces a larger spreading effect for both networks (the change from the third column to the last column), but more experience spreads the objects to an even greater extent (compare the last columns in the two rows). In data not shown, both of these networks achieve 87.5% accuracy on face recognition, but the network with less experience with objects only achieves an average accuracy of 16.67% on nonface objects. This is well above chance but much lower than the more experienced network, which achieves an accuracy of 83.33% on objects. As a result, we can speculate that greater experience actually leads to a greater spread in the hidden units of the network, and this spreading transform positively correlates with performance on the object recognition task. Performance on objects and faces is similar in a network with more experience and very different in a network with less experience, as we saw in Figures 3 and 4. This is the power of experience.

The above analysis is based on the PCA projection of the net input on a 2-D space. Because there are 50 hidden units in the network, we want to explore whether the phenomenon could generalize along all dimensions. As we cannot visualize a 50-dimensional space, we take five

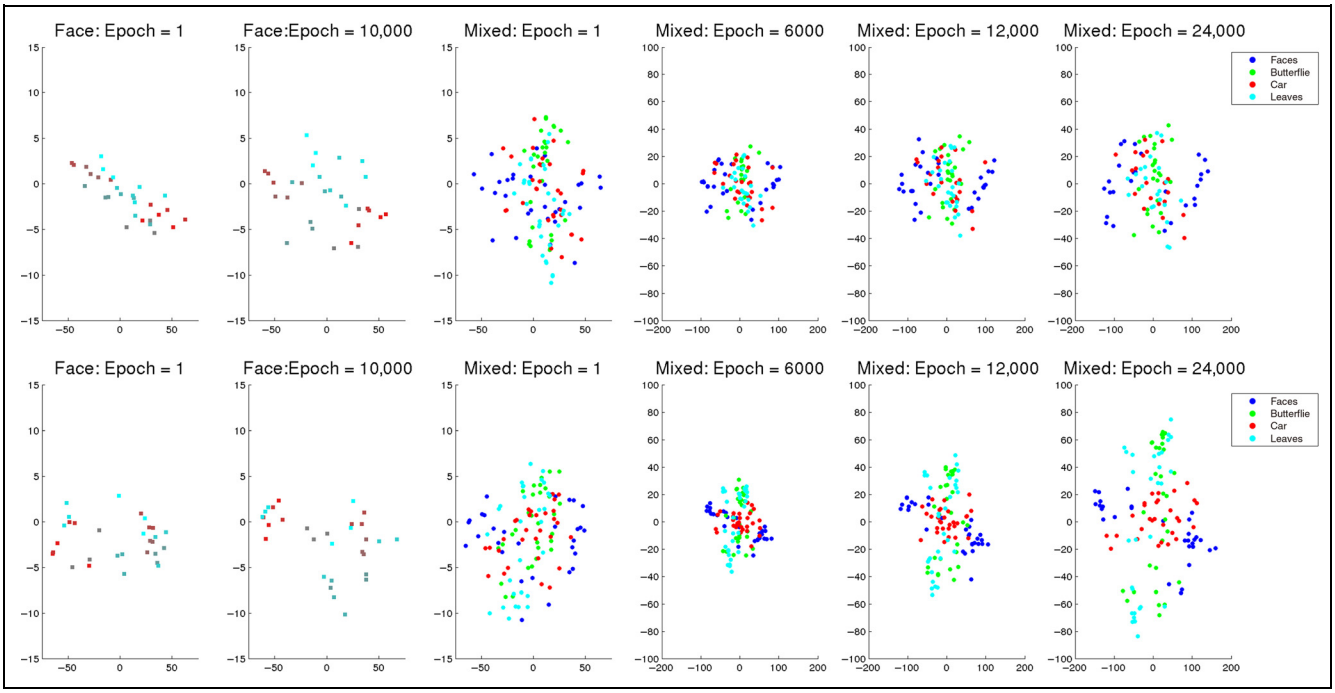

Figure 6. Visualization of the development of net input of hidden units over network training. First row: participant network with low experience (three training examples per individual). Second row: participant network with high experience (12 training examples per individual). Each column represents the data collected from corresponding training epoch (shown in the title). In the two left columns, the colored dots represent different individual faces. In the four right columns, the colored dots represent different object categories, shown in the legend. Note: The y axis changes from $[-15,+15]$ to $[-100,+100]$ in the fourth column for clarity.

measurements for each dimension to help us understand its behavior:

1. Max: the maximum value of the projection on a principal component dimension, for each single category (locally) and across all of the categories (globally).
2. Min: the minimum value along a dimension, for each single category and across all of the categories.
3. Var: the variance along each dimension, for each single category and across all of the categories.
4. Inter: the average between-class distance, measured using Euclidean distance between the center of each object within the same category to the center of the current category and averaged across all categories.
5. Intra: the average within-class distance, measured using Euclidean distance between each data point belonging to a single individual to the average of that individual's locations and averaged across all categories.

Among the five measurements, Max, Min, and Var are measured both globally (across all categories) and locally (for each object category), whereas Inter and Intra are only measured globally. Max, Min, and Var indicate how far the individual representations are spread out along each dimension, whereas Inter and Intra measure the behavior for each group. The results are shown in Figure 7.

From the local measurement results in Figure 7, we can clearly see that

1. For Max and Var, the value of high-experience network is always greater than low-experience network.
2. For Min, the value of high-experience network is always smaller than low-experience network.

These findings hold for all four object categories. These results demonstrate that, for individual representations, high levels of experience separate them along all dimensions in the space.

For the global measurement (combined all categories), we can see that

1. For Max, Min, and Var, their behaviors are the same as local measurements above.
2. For Inter and Intra, the value of high-experience network is mostly greater than low-experience network.

Imagine that each object forms a cluster in the space. The Inter and Intra results indicate that, as experience grows, each individual resident within that cluster will become further apart from its neighbors and the whole cluster itself will also move away from other clusters, like the "redshift" phenomenon in physical cosmology (Hubble, 1929). As this "redshift" of object representation happens in all dimensions of the hidden unit universe, it suggests that the essential power of experience is to generate a spreading transform for objects in the representational space and, accordingly, to facilitate a subordinate level classification task. The experience moderation effect, as can be seen in our experiment, is a direct outcome/reflection of this internal power, in a large population of participants.

In addition, we measured the entropy of the net input of all the hidden units. Entropy is a measurement about how much information is preserved in the hidden units, and it is scale free. If the data are highly scattered, the variance will be high, and more information will be carried. To calculate the entropy, we obtained the net input value of all hidden units across examples in the training set. We then calculated the entropy for each of the hidden units by getting the probability distribution (the normalized histogram) of the values, thereby computing p_i for each bin, and then summing $p_i \log p_i$ over the bins (the results were robust across various bin sizes). We then averaged the entropy over all of the hidden units. To examine how the entropy develops over time, we plot its value as a function of training iterations, as shown in Figure 8. As we can see, although both networks show a general increasing trend of the entropy, the network with more varied experience always has higher entropy. This result is expected based on the PCA visualization in Figure 6, as the representations for both face and nonface objects become more separated as training proceeds. Again, this result demonstrates that the power of experience is to learn a more separated representation for objects to facilitate the subordinate level classification task.

Furthermore, when looking into the local and global measurement of variance in Figure 7, we can see that, for the more experienced network, a larger number of dimensions accounts for more variance than for the less experienced network. This suggests that the more experienced network contains more complex information that must be decomposed into several different dimensions, which provides another way of measuring how the network is spreading out the representation of the categories.

DISCUSSION

Neurocomputational models can provide insight into behavioral results by suggesting hypotheses for how the results came about. We can then analyze the models in ways that are difficult or impossible in human participants or animals. In this article, we explored how a neurocomputational model can explain the experience moderation effect observed by Gauthier et al. (2014). We trained networks to perform the same tasks humans have to perform, that is, to recognize objects and faces. We used one network per participant, setting their parameters based on the individual participant data. We mapped domain-general visual ability, v, to the number of hidden units and experience, E, to the number of training examples per individual. We showed that the model fits the human data quite well: As the networks gain more experience with the object categories, the correlation

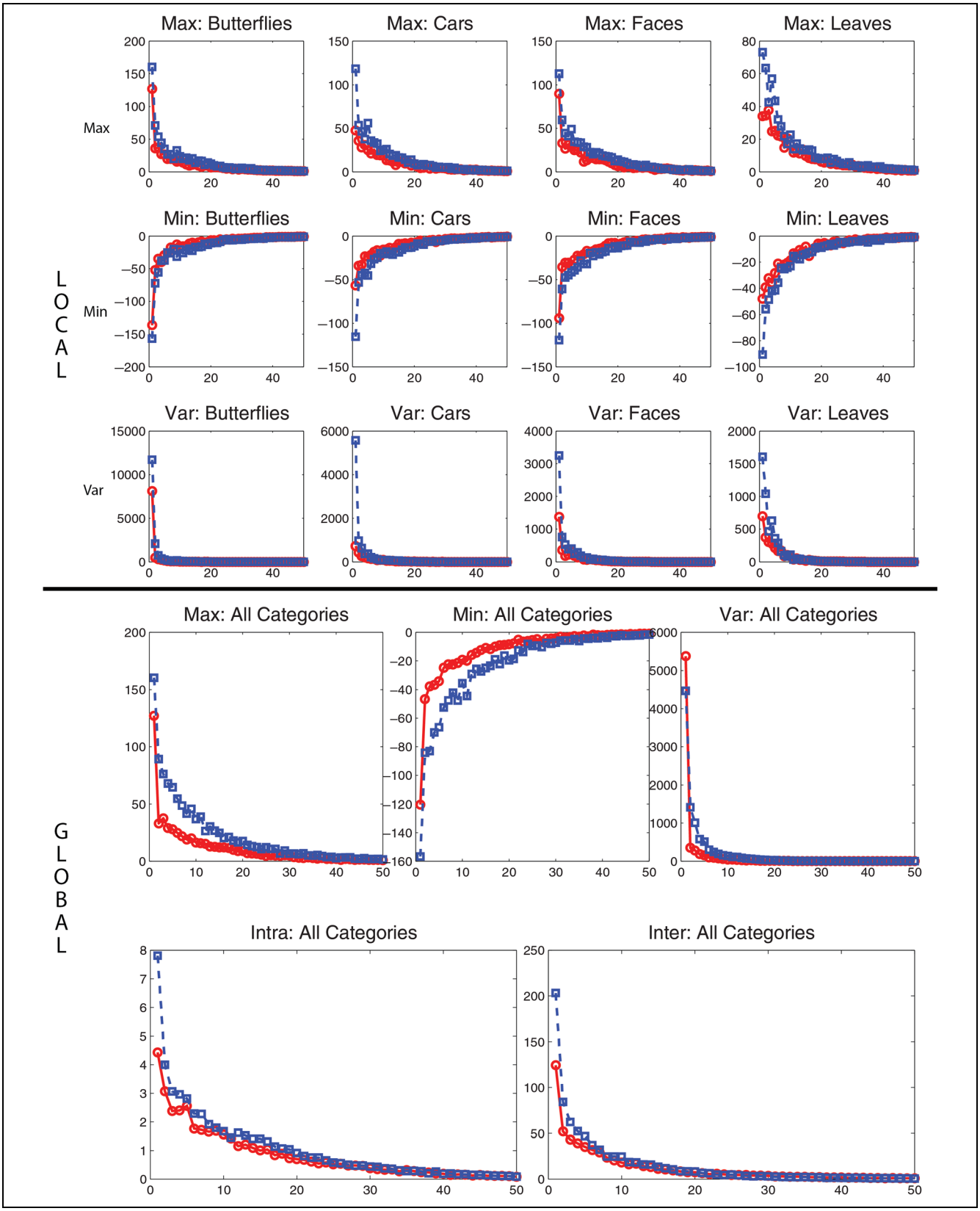

Figure 7. Visualization of the measurements taken on all 50 PCA dimensions of the hidden unit net inputs. In all graphs, the blue lines represent the network with a high level of experience, and the red lines represent the network with a low level of experience. We take five measurements: Max, Min, Variance, Intergroup distance, and Intragroup distance, as described in the text. The top three rows show the result of Max, Min, and Var on the four object categories (left to right: faces, butterflies, cars, and leaves). The last two rows show the result of all measurements on all categories.

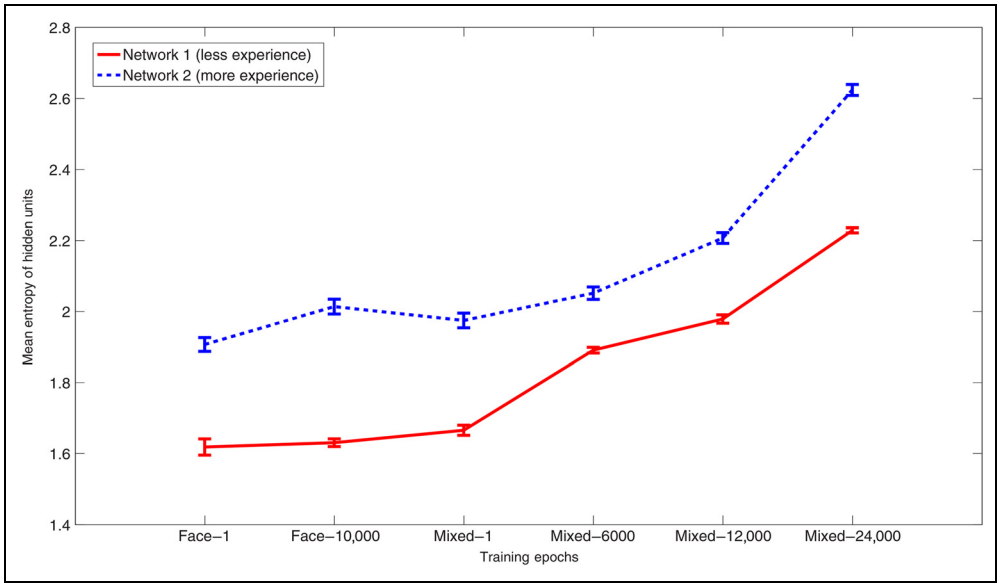

Figure 8. Entropy of hidden units as a function of training epochs. Blue dashed line: network with more experience. Red line: network with less experience. The network with more experience generally has greater entropy across training, suggesting that the representation is more separated. Error bars denote ±1 *SE*.

between performance on objects and performance on faces increases.

In Experiment 1, as in Gauthier et al. (2014), we had to average across category experience to obtain the correlation with face processing performance. That is, we could not significantly predict face recognition ability based solely on performance within a single category. In Experiment 2, we "recruited" more neural network participants and predicted that the effect should hold at the single-category level, provided that there are a sufficient number of participants that span all levels of visual ability and experience.

Finally, we also attempted to replicate the effect with networks that did not differentiate faces but simply placed objects and faces into basic level categories. Here, we did not find an experience moderation effect, suggesting that the type of experience and level of the task (basic or subordinate level discrimination) is an important factor to be considered in understanding these effects.

The conclusion that task matters in terms of the kind of perceptual expertise that is acquired and for the neural substrates recruited is supported by prior work. For instance, novel objects become processed in a holistic manner, like faces, if they are from a category for which participants practiced individuation but not categorization (Wong, Palmeri, & Gauthier, 2009). Likewise, brief individuation training improves discrimination of new faces from another race, whereas a different training task with the same faces that is as demanding but does not require individuation does not improve discrimination (McGugin, Tanaka, Lebrecht, Tarr, & Gauthier, 2011). Qualitatively different patterns of neural representations are observed after training with novel objects in tasks that produce different kinds of behavioral results (Wong, Folstein, & Gauthier, 2012; Wong, Palmeri, Rogers, et al., 2009).

This experiment predicts that the source of the experience moderation effect is not in regions of the brain that are sensitive only to category level, as opposed to regions that are associated with better performance in individuation for objects and faces, such as the FFA (McGugin et al., 2014; McGugin, Gatenby, et al., 2012; Furl, Garrido, Dolan, Driver, & Duchaine, 2011; Gauthier et al., 2000). One advantage of computational models is that we can analyze them in ways we cannot analyze human participants to provide hypotheses as to the underlying mechanisms of an effect. For example, an obvious question is, why isn't there a "zero-sum game" between the neurons allocated for each task? That is, how can the same features be used for both faces, leaves, cars, and butterflies?

Behavioral and neural studies show that face recognition and the recognition of other objects of expertise can compete. The N170 face-selective ERP is reduced for faces when they are shown in the context of objects of expertise (Rossion, Kung, & Tarr, 2004; Gauthier, Curran, Curby, & Collins, 2003). Behaviorally, nonface objects of expertise compete with faces but not with objects with which participants are not expert (McGugin et al., 2011; McKeeff, McGugin, Tong, & Gauthier, 2010). fMRI responses to cars in FFA predict behavioral expertise with cars when the cars are presented alone on the screen and, to some degree, still when shown among other objects, but not when the other objects are faces (McGugin et al., in press). What all these studies have in common is that interference occurs when faces and objects from another category of expertise have to be processed simultaneously or at least in very close temporal contiguity. Again, this suggests that they are sharing representations.

Our analysis shows why this would be the case. The nonlinear transformation by the network at the backpropagation-trained hidden layer displays a spreading transform that separates similar-looking objects. This transform generalizes to new categories. At the same time, as shown in the last four columns in Figure 6, the representation of faces is

interdigitated with the representations of other categories. Hence, the reason why we see interference in the human participant studies is because of this shared representation. In previous work (Tong et al., 2008), we hypothesized that the FFA contains features useful for fine level discrimination of faces and showed how these features generalize to the discrimination of novel categories. Here, we find the same result, shown in the third column of Figure 6, where we find that objects are already separated by the face features, that is, the transform that separates individual faces also separates individual objects even at the beginning of training on those objects. Given that our model is a model of the FFA, we hypothesize that the location of the experience moderation effect is in the FFA, but more generally, it could be in any area where face representations are more separated.

We conclude that the real power of experience at individuating objects within a homogeneous category is to separate the objects in all dimensions of the representational space spanned by the FFA and that the experience moderation effect is a direct reflection of this spreading transform. These results support the argument that face and nonface object discrimination are inherently correlated through the sharing of the same mechanism: The better one is at face individuation, and the better one will be at individuating objects, given sufficient experience with objects.

One may speculate that one may also find an experience moderation effect at the basic level of categorization. That is, if a participant shows high performance in simply discriminating object categories and has a great deal of experience in discriminating multiple categories, performance in multiple domains should be correlated. There is some evidence that a great deal of experience with basic level categorization, as in letter recognition, results in a different kind of expertise from that obtained for subordinate level experience—different both in behavior and neural substrate (Wong, Palmeri, Rogers, et al., 2009; Wong & Gauthier, 2007). One might hypothesize that multiple *v*s, that is, a basic level *v* and a fine level *v*, corresponding to different brain regions associated with these tasks. That is, there must be a constraint that the level of tasks be equalized before one can hope to find such a correlation. In our model, we use fine level *v*. Evidence for this hypothesis arises in recent work showing that a neural network that is good at differentiating the thousand categories of the Imagenet competition (Russakovsky et al., 2014) develops features that are useful in differentiating other categories (Wang & Cottrell, 2015; Zeiler & Fergus, 2014).

More recently, training backpropagation-based deep neural networks has been shown to achieve state-of-the-art performance on many computer vision tasks, such as image classification (Szegedy et al., 2014; Krizhevsky, Sutskever, & Hinton, 2012), scene recognition (Zhou, Lapedriza, Xiao, Torralba, & Oliva, 2014), and object detection (Girshick, Donahue, Darrell, & Malik, 2014). Researchers also have used deep neural networks to probe representations in neural data, especially in IT (e.g., Güçlü & van Gerven, 2015; Cadieu et al., 2014; Yamins et al., 2014). Remarkably, these studies have shown that the features learned in the neural networks can explain the representation in human and monkey IT. As these networks are also trained by backpropagation, they support our contention that our neurocomputational model is a reasonable model of FFA and LOC. As a result, it is a promising research direction to use deep neural networks to explain more cognitive/behavioral data and to model how the brain works.

In summary, we suggest that the correlation between visual processing of faces and objects is mediated by a common representational substrate in the visual system, most likely in the FFA, and that the reason for this mediation is that the FFA embodies a transform that amplifies the differences between homogeneous objects. This transformation is generic; it applies to a wide range of visual categories. The generic nature of this transform explains why there is a synergy between face processing and expert object processing.

Acknowledgments

This work was supported in part by NSF grant SMA 1041755 to the Temporal Dynamics of Learning Center (TDLC), NSF Science of Learning Center funding (G. W. C. and I. G.), NSF grant IIS-1219252 to G. W. C., a TDLC trainee grant to P. W., and a gift from Hewlett Packard to G. W. C.

Reprint requests should be sent to Panqu Wang, Department of Electrical and Computer Engineering, University of California, San Diego, 9500 Gilman Dr. 0407, La Jolla, CA 92093, or via e-mail: pawang@ucsd.edu, wangpanqumanu@gmail.com.

Notes

1. Note that, for faces, "individual" refers to a particular person; for butterflies and leaves, a particular species; and for cars, a particular make and model.
2. In general, the number of hidden units depends on the size of training set. In recent winner of ImageNet Large Scale Visual Recognition Challenge, the networks are trained with over 1.2 million images, and the final hidden layer has 4096 units (Krizhevsky et al., 2012). However, if the same network is trained on a smaller data set, the recognition accuracy is low because of overfitting (Zeiler & Fergus, 2014).

REFERENCES

Bilalić, M., Langner, R., Ulrich, R., & Grodd, W. (2011). Many faces of expertise: Fusiform face area in chess experts and novices. *Journal of Neuroscience, 31,* 10206–10214.

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

Cadieu, C. F., Hong, H., Yamins, D. L., Pinto, N., Ardila, D., Solomon, E. A., et al. (2014). Deep neural networks rival the representation of primate IT cortex for core visual object recognition. *PLoS Computational Biology, 10,* e1003963.

Cottrell, G. W., & Hsiao, J. H. (2011). Neurocomputational models of face processing. In A. J. Calder, G. Rhodes, M.

Johnson, & J. Haxby (Eds.), *The Oxford handbook of face perception*. Oxford: Oxford University Press.
Dailey, M. N., & Cottrell, G. W. (1999). Organization of face and object recognition in modular neural network models. *Neural Networks, 12,* 1053–1073.
Daugman, J. G. (1985). Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two dimensional visual cortex filters. *Journal of the Optical Society of America, 2,* 1160–1169.
Dennett, H. W., McKone, E., Tavashmi, R., Hall, A., Pidcock, M., Edwards, M., et al. (2011). The Cambridge Car Memory Test: A task matched in format to the Cambridge Face Memory Test, with norms, reliability, sex differences, dissociations from face memory, and expertise effects. *Behavior Research Methods, 44,* 587–605.
Duchaine, B., & Nakayama, K. (2006). The Cambridge Face Memory Test: Results for neurologically intact individuals and an investigation of its validity using inverted face stimuli and prosopagnosic subjects. *Neuropsychologia, 44,* 576–585.
Furl, N., Garrido, L., Dolan, R. J., Driver, J., & Duchaine, B. (2011). Fusiform gyrus face selectivity relates to individual differences in facial recognition ability. *Journal of Cognitive Neuroscience, 23,* 1723–1740.
Gauthier, I., Curran, T., Curby, K. M., & Collins, D. (2003). Perceptual interference supports a non-modular account of face processing. *Nature Neuroscience, 6,* 428–432.
Gauthier, I., McGugin, R. W., Richler, J. J., Herzmann, G., Speegle, M., & VanGulick, A. E. (2014). Experience moderates overlap between object and face recognition, suggesting a common ability. *Journal of Vision, 14,* 7.
Gauthier, I., Skudlarski, P., Gore, J. C., & Anderson, A. W. (2000). Expertise for cars and birds recruits brain areas involved in face recognition. *Nature Neuroscience, 3,* 191–197.
Gauthier, I., Tarr, M. J., Anderson, A. W., Skudlarski, P., & Gore, J. C. (1999). Activation of the middle fusiform face area increases with expertise in recognizing novel objects. *Nature Neuroscience, 2,* 568–573.
Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* (pp. 580–587).
Grill-Spector, K., Knouf, N., & Kanwisher, N. (2004). The fusiform face area subserves face perception, not generic within-category identification. *Nature Neuroscience, 7,* 555–562.
Grill-Spector, K., Sayres, R., & Ress, D. (2006). High-resolution imaging reveals highly selective nonface clusters in the fusiform face area. *Nature Neuroscience, 9,* 1177–1185.
Güçlü, U., & van Gerven, M. A. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience, 35,* 10005–10014.
Hubble, E. (1929). A relation between distance and radial velocity among extra-galactic nebulae. *Proceedings of the National Academy of Sciences, U.S.A., 15,* 168–173.
Joyce, C., & Cottrell, G. W. (2004). Solving the visual expertise mystery. In H. Bowman & C. Labiouse (Eds.), *In connectionist models of cognition and perception II: Proceedings of the Eighth Neural Computation and Psychology Workshop*. World Scientific.
Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience, 17,* 4302–4311.
Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems* (pp. 1097–1105).
Lades, M., Vorbrggen, J., Buhmann, J., Lange, J., von der Malsburg, C., Wrtz, R. P., et al. (1993). Distortion invariant object recognition in the dynamic link architecture. *IEEE Transactions on Computers, 42,* 300–311.
Mallah, C., Cope, J., & Orwell, J. (2013). Plant leaf classification using probabilistic integration of shape, texture and margin features. *Signal Processing, Pattern Recognition and Applications*. doi: 10.2316/P.2013.798-098.
McCarthy, G., Puce, A., Gore, J. C., & Allison, T. (1997). Face-specific processing in the human fusiform gyrus. *Journal of Cognitive Neuroscience, 9,* 605–610.
McGugin, R. W., Gatenby, J. C., Gore, J. C., & Gauthier, I. (2012). High-resolution imaging of expertise reveals reliable object selectivity in the FFA related to perceptual performance. *Proceedings of the National Academy of Sciences, U.S.A., 109,* 17063–17068.
McGugin, R. W., Newton, A. T., Gore, J. C., & Gauthier, I. (2014). Robust expertise effects in right FFA. *Neuropsychologia, 63,* 135–144.
McGugin, R. W., Richler, J. J., Herzmann, G., Speegle, M., & Gauthier, I. (2012). The Vanderbilt Expertise Test reveals domain-general and domain-specific sex effects in object recognition. *Vision Research, 69,* 10–22.
McGugin, R. W., Tanaka, J. W., Lebrecht, S., Tarr, M. J., & Gauthier, I. (2011). Race-specific perceptual discrimination improvement following short individuation training with faces. *Cognitive Science, 35,* 330–347.
McGugin, R. W., Van Gulick, A. E., Tamber-Rosenau, B. J., Ross, D. A., & Gauthier, I. (2014). Expertise effects in face-selective areas are robust to clutter and diverted attention, but not to competition. *Cerebral Cortex,* bhu060.
McKeeff, T. J., McGugin, R. W., Tong, F., & Gauthier, I. (2010). Expertise increases the functional overlap between face and object perception. *Cognition, 117,* 355–360.
O'Reilly, R. C. (1996). Biologically plausible error-driven learning using local activation differences: The generalized recirculation algorithm. *Neural Computation, 8,* 895–938.
Ozuysal, M., Lepetit, V., & Fua, P. (2009). Pose estimation for category specific multiview object localization. In *IEEE Conference on Computer Vision and Pattern Recognition, 2009* (pp. 778–785). IEEE.
Plaut, D. C., & Shallice, T. (1993). Deep dyslexia: A case study of connectionist neuropsychology. *Cognitive Neuropsychology, 10,* 377–500.
Rossion, B., Kung, C. C., & Tarr, M. J. (2004). Visual expertise with nonface objects leads to competition with the early perceptual processing of faces in the human occipitotemporal cortex. *Proceedings of the National Academy of Sciences, U.S.A., 101,* 14521–14526.
Rushton, J. P., Brainerd, C. J., & Pressley, M. (1983). Behavioral development and construct validity: The principle of aggregation. *Psychological Bulletin, 94,* 18–38.
Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., et al. (2014). ImageNet large scale visual recognition challenge. *International Journal of Computer Vision,* 1–42.
Sanger, T. D. (1989). Optimal unsupervised learning in a single-layer linear feedforward neural network. *Neural Network, 2,* 459–473.
Sergent, J., Ohta, S., & MacDonald, B. (1992). Functional neuroanatomy of face and object processing. A positron emission tomography study. *Brain, 115,* 15–36.
Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., et al. (2014). Going deeper with convolutions. *arXiv:* 1409.4842, 1–9.
Tong, M. H., Joyce, C. A., & Cottrell, G. W. (2008). Why is the fusiform face area recruited for novel categories of expertise? A neurocomputational investigation. *Brain Research, 1202,* 14–24.

Tottenham, N., Tanaka, J. W., Leon, A. C., McCarry, T., Nurse, M., Hare, T. A., et al. (2009). The nimstim set of facial expressions: Judgments from untrained research participants. *Psychiatry Research, 168,* 242–249.

Tsao, D. Y., Freiwald, W. A., Tootell, R. B., & Livingstone, M. S. (2006). A cortical region consisting entirely of face-selective cells. *Science, 311,* 670–674.

Wang, J., Markert, K., & Everingham, M. (2009). Learning models for object recognition from natural language descriptions. In *Proceedings of the British Machine Vision Conference*.

Wang, Y., & Cottrell, G. W. (2015). Bikers are like tobacco shops, formal dressers are like suits: Recognizing urban tribes with caffe. In *Proceedings of IEEE Winter Conference on Applications of Computer Vision (WACV)* (pp. 876–883). IEEE.

Wilmer, J. B., Germineb, L., Chabrisc, C. F., Chatterjeeb, G., Williamsd, M., Lokene, E., et al. (2010). Human face recognition ability is specific and highly heritable. *Proceedings of the National Academy of Sciences, U.S.A., 107,* 5238–5241.

Wong, A. C. N., & Gauthier, I. (2007). An analysis of letter expertise in a levels-of-categorization framework. *Visual Cognition, 15,* 854–879.

Wong, A. C. N., Palmeri, T. J., & Gauthier, I. (2009). Conditions for facelike expertise with objects becoming a Ziggerin expert—But which type? *Psychological Science, 20,* 1108–1117.

Wong, A. C. N., Palmeri, T. J., Rogers, B. P., Gore, J. C., & Gauthier, I. (2009). Beyond shape: How you learn about objects affects how they are represented in visual cortex. *PLoS One, 4,* e8405.

Wong, Y. K., Folstein, J. R., & Gauthier, I. (2012). The nature of experience determines object representations in the visual system. *Journal of Experimental Psychology: General, 141,* 682.

Xu, Y. (2005). Revisiting the role of the fusiform face area in visual expertise. *Cerebral Cortex, 15,* 1234–1242.

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences, U.S.A., 111,* 8619–8624.

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In *Computer vision–ECCV 2014* (pp. 818–833). Springer International Publishing.

Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., & Oliva, A. (2014). Learning deep features for scene recognition using places database. In *Advances in Neural Information Processing Systems* (pp. 487–495).