Do Expression and Identity Need Separate Representations?

Garrison W. Cottrell (gary@cs.ucsd.edu)
Kristin M. Branson (kbranson@cs.ucsd.edu)
Computer Science and Engineering; 9500 Gilman Drive
La Jolla, CA 92093-0114 USA
Andrew J. Calder (andy.calder@mrc-cbu.cam.ac.uk)
MRC Cognition and Brain Sciences Unit
15, Chaucer Road
Cambridge, CB2 2EF
UK

Abstract

Recent work has shown that expression recognition shows holistic processing effects much like face recognition (Calder et al., ress). We extend our previous model of facial expression recognition (Dailey et al., 2000) to account for these results. We show that our model, with small modifications to the training procedure, can account for the systematic biases between upper and lower facial expression recognition, and the holistic/configural processing effect. Finally, we show that results that seem to support the idea that separate representations are necessary for emotion and identity processing can be accounted for by a single representation model. This latter effect is demonstrated in subjects by constructing chimeric faces by taking the top half of one face and the bottom half of another.

Background

In recent years a consensus has emerged that face processing is "holistic" in nature (Tanaka and Farah, 1993; Farah et al., 1998). Here "holistic" means "configural": that is, there is an effect of the whole in recognition of the parts of a face. One way to show this is to construct composite faces by combining the upper and lower halves of different faces. If the subject's task is to recognize the identity of the upper half of the face, there is interference if the lower half of the face is from a different person. However, if the lower half of the face is misaligned with the upper half, there is no difference in subjects' responses when the lower half is the same person or a different person (Young et al., 1987).

Recently, it has been shown that this effect extends to facial expression recognition (Calder et al., ress). Calder et al. first replicate a well-known effect that certain expressions are more easily recognized from their top or bottom halves. A summary of the data is shown in Table 1. Based on this data, Calder et al. constructed composite faces from the same subject using a top-biased emotion in the top half and a bottom-biased emotion in the bottom half. When subjects were asked to identify the emotion in one half, their reaction times were slower when the two halves were aligned versus when they were misaligned.

In further experiments, Calder et al. present data that they interpret as showing that there must be separate representations for facial identity and facial expression processing. The experiment in this case was to show several kinds of composite images to the subjects, and to give them two tasks: identity judgements (after training on the identities of the subjects), and expression judgements. There were three kinds of composite images: 1) same identity, different emotion; 2) different identity, same emotion; and 3) different identity, different emotion. Consider the case of identity judgements. The subjects are asked to judge the identity of the subject in the bottom half of the image. If the top of the image is the same subject, but a different emotion, the reaction times are faster than if the top half of the image is a different subject. This is expected based on the above results on configural processing. However, when the top was a different subject, their reaction times did not differ between the case where the expression of the different subject was the same or different. The interpretation is that identity processing is not affected by affect processing. Identical results in the emotion identification task support the idea that expression processing is not affected by identity processing. The conclusion that two representations must therefore be in play makes intuitive sense, but should be tested in a model. In the following, we show that a single representation suffices to obtain these results.

The Model

We performed three experiments that paralleled as closely as possibly three of the experiments reported in (Calder et al., ress). In each experiment, we used the same images from the Pictures of Facial Affect (POFA) dataset (Ekman and Friesen, 1976), although normalized as described below. Also, we constructed our own versions of Calder's hand-constructed split images. All experiments used a similar model and data. The details of these experiments are described in this section.

Classification Model

In all experiments, our classification model employs image filtering, principal components analysis

Expression	Human Top	Human Bottom	Net Top	Net Bottom
Happy	0.20 (.09)	0.01 (.01)	0.40	0.00
Sad	0.19 (.05)	0.34 (.08)	0.28	0.40
Afraid	0.33 (.08)	0.56 (.09)	0.28	0.70
Angry	0.28 (.06)	0.49 (.09)	0.29	0.65
Surprised	0.06 (.21)	0.33 (.07)	0.00	0.21
Disgusted	0.62 (.10)	0.04 (.14)	0.20	0.00

Table 1: Fraction of test examples incorrectly identified for each expression. The Human Top and Human Bottom results correspond to the results reported for expression recognition by (Calder et al., ress). The Net Top and Net Bottom results correspond to the results achieved by our classification model. The number in parentheses is the standard error for the humans.

(PCA), and a single-layer neural network to classify the expression and identity of an input pixel image of an actor posing an expression (Dailey et al., 2000).

Preprocessing of these images begins by aligning the images so that the eyes and mouth of all images are in the same location, then cropping the images to eliminate the background. After each image is aligned, it is convolved with a grid of two-dimensional Gabor jets. Each jet is composed of 40 Gabor filters of five different sizes and eight different orientations. Each jet is centered on a pixel of the aligned image. This image filtering was chosen because it is similar to filtering done in the striate cortex of cats and has previously been shown to improve expression recognition in neural networks (Dailey et al., 2000). Applying Gabor filters to a subsampled 240 x 292 pixel image results in a 40,600 component vector. These vectors are then z-scored (transformed to 0 mean, unit std. dev.) on an individual filter basis, resulting in the Gabor pattern.

As our experiments required a method of directing the classification model's attention to just one half (bottom or top) of the face stimulus, the other half of the face stimulus is attenuated. Each component in the half of the Gabor pattern to be attenuated is multiplied by 0.125. The factor 0.125 was chosen after comparing the results of attenuating by different fractions. An attenuation factor of 0.125 resulted in the error of the model's recognition of expression in half face training data (described below) most closely resembling the error of the human's recognition of expression in half face images. However, we found little variance in the model's error with attenuation factors between 0.5 and 0.125. We define a Gabor filter as being within the half the image to be attenuated if the pixel it is centered on is in that half, or if it is within two times the standard deviation of the Gaussian of the Gabor filter. This way, even filters in the attended half will be attenuated if their receptive field overlaps the other half of the image.

50 principal components of each Gabor pattern are extracted from the training data. These are also z-scored. Finally, a soft-maxed single-layer neural network is trained using the cross-entropy error criterion.

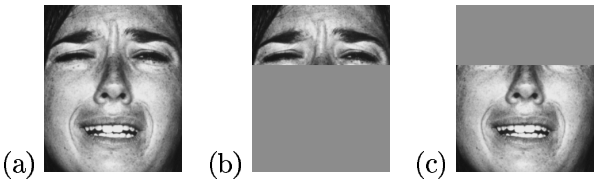

Figure 1: Whole and Half Face Training Data: (a) Aligned and cropped pixel image; (b) Pixels corresponding to the top half; (c) Pixels corresponding to the bottom half.

Training Data

All stimuli are derived from pixel images from the POFA database. This data set includes images of 14 actors posing 6 expressions: happiness, sadness, fear, anger, surprise, and disgust.

In all experiments, the principal components of the Gabor patterns are extracted from whole face and half face Gabor patterns, and the neural network is trained upon those. We assume that during the learning of the emotions, subjects also attend to just one half of the face at different times. We can think of this as either an attentional process or as a crude simulation of eye movements. In addition, splitting the images between the top half and bottom half, as defined by Calder, resulted in the top half of the image being smaller in area than the bottom half of the image. Extracting principal components while attenuating one half of the image allowed more equal representation of each half in the principal components. Whole face Gabor patterns are created by convolving the original pixel image with the Gabor filters, as described above, then z-scoring these patterns. Half face Gabor patterns are created in a similar manner: the original pixel image is convolved with the Gabor filters and z-scored, then one half of the pattern is attenuated to a fraction of 0.125. Thus, when creating a top half face Gabor

pattern, components corresponding to the bottom half of the face are multiplied by the fraction 0.125. An alternative method would have been to start with a pixel image of just one half of the face, and then convolve the image with Gabor filters. We did not do this because the Gabor filters would have given a strong response at the edge in the image between the zeroed half and the non-zeroed half, and we were concerned that this signal might confuse the network (this may have been an unnecessary worry). An example of an aligned and cropped pixel image and bottom and top half images are shown in Figure 1.

Experiments

Experiment 1

The goal of the first experiment was to determine whether our network gave the same results as Calder et al.'s subjects in terms of which expressions are top-biased and which are bottom-biased. An expression is *top-biased* if it is more accurately identified by our model from the top half of the face than the bottom. An expression is *bottom-biased* if the opposite is true. Calder et al. used 10 of the POFA actors for this experiment. We used the same 10 actors.

The general procedure for this experiment was to classify half face examples using the classification model described above, then compare the classification error for top and bottom test half face examples for each expression. We used "hold one actor out" cross validation, so we trained on nine and tested on the tenth. Each of these were repeated ten times using different initial random weights. The test half-face examples classified by the network are the result of convolving a whole face image with the Gabor filters, z-scoring, attenuating one half using the multiplier 0.0 (thus there is no output from that half), and projecting onto the principal components. Training is stopped when error on the training set most closely corresponds to the human confusion matrix reported by Ekman.

Results: The average results of classifying each test example 10 times are shown in Table 1. The network responses do not vary much over networks. These results show that for our classification model, happy and disgusted are bottom-biased while sad, afraid, angry and surprised are top-biased. These results are very similar to the human results: happy and disgusted are bottom-biased and sad, afraid, and angry are top-biased. The only difference is that our classification model finds surprised to be top-biased while the human results find surprised to be unbiased, due to subject variance. In addition, large differences between the network's classification error fractions of the top- and bottom-half stimuli correspond to large differences between humans' classification error fractions of the top and bottom-half images.

Experiment 2

The goal of the second experiment was to determine if incorrect configural information disrupts the model's expression recognition. This involves comparing the model's accuracy on identifying the expression in one half of two different types of stimuli: composite and noncomposite. A composite example is the result of aligning the top half of one face with the bottom half of another (Figure 2(a)). A noncomposite example is the result of misaligning the top half of one face with the bottom half of another. When performance degrades on composite faces compared to non-composites, this is taken to be an indicator of configural processing (Young et al., 1987). That is, when the two halves are aligned, subjects are unable to ignore the information in the other half of the face, even though they are judging only one half.

In this experiment, both halves of the composite and noncomposite examples correspond to the same actor but different expressions. In Calder et al.'s experiment, reaction times were slower for composite images than for non-composite. Composite images are created by aligning the bottom half of one happy, surprised, or disgusted face image with the top half of one sad, afraid, or angry face image. Following Calder et al., images from only four of the actors are used.

Composite Gabor patterns are the result of convolving these composite images with Gabor filters and z-scoring. As the model must identify the expression in one half of the example, one half of the composite Gabor pattern is attenuated by a fraction of 0.125. Noncomposite Gabor patterns are created from the composite Gabor patterns. The half of the pattern that is attenuated is misaligned with the other half by replacing the components of the attenuated half corresponding to the right side of the face with the components corresponding to the left side and zeroing the components corresponding to the left side. These are projected onto the principal components of the same training set as the first experiment. The network is trained as in the previous experiment, and tested on identifying the expression in one half of both composite and noncomposite examples.

Results: The results are shown in Figure 3. These results indicate that our classification performs better on noncomposite test examples than composite test examples. As the top and bottom face halves are aligned in the composite examples, there is incorrect configural information in these examples that is not present in the noncomposite examples. Therefore, this result indicates that incorrect configural information disrupts our model's expression recognition. The trend in classification error for these two types of test examples is similar to the trend reported for human response time for these two types of examples, as shown in Figure 3.

Figure 2: Different stimuli types. (a) Composite/Non-composite images; composites with (b) same identity/different expression; (c) different identity/same expression; and (d) different identity/different expression.

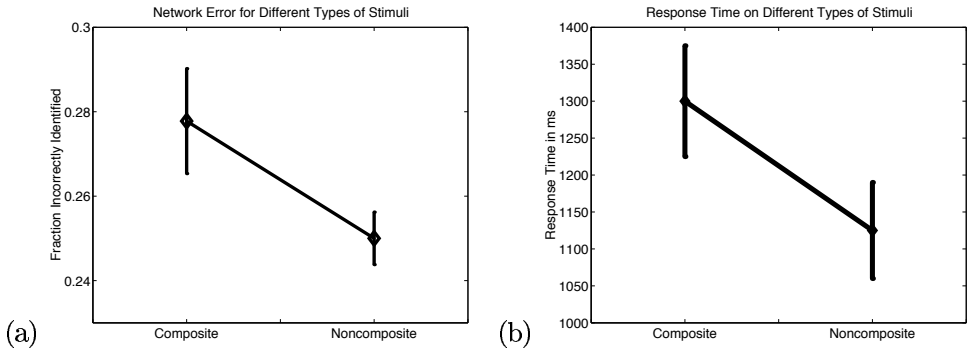

Figure 3: Experiment 2 results. (a) Average error proportion for composite, noncomposite, and half face test examples. Vertical bars indicate standard deviations. (b) Human response times for expression classification of composite and noncomposite test examples

Experiment 3

In Calder et al.'s Experiment 4, they investigated the interference between identity and expression processing. They noticed that when two happy expressions were combined from different individuals, the new image looked like a happy person who was a new individual, different from the two source individuals (see Figure 2(b)). They hypothesized from this that "the configural information used to encode facial identity may be different from that used to code facial expression." They suggested that if the two kinds of processing could be selectively disrupted, then that would be support for this two-representation model. Note that our model has one representation, corresponding to the principal components layer. If we can obtain the same results, then we will show that their conclusion, that there are two representations, is unwarranted.

This experiment involved three types of composite stimuli. Same identity with different expression (SID/DE) composite examples require that the identity of the actor in both halves be the same but the expression in each half be different. Different identity with same expression (DID/SE) composite examples require that the identity of the actor in each half be different but the expression in each half be the same. Different identity with different expression (DID/DE) composite examples require that the identity of the actor in each half be different but the expression in each half be the same. Figure 2(b-d) shows examples of all three types.

The model is trained on all the whole and half face examples for ten actors. Following Calder et al., the model must classify the expression and identity in the bottom half of the composite test examples constructed from three of the actors. In order to test for identity performance, Calder et al. trained their subjects on both identity and expression for the three actors they used to create the composite test stimuli. Hence these stimuli were included in the training set. Three additional localist outputs were included to learn the identity of these three actors. The seven remaining actors were trained to produce all 0's on these units. Following Calder et al., for the three test subjects, the composite examples constructed from them for testing purposes only use the happy, surprised, and disgusted expressions. We stopped training by using a holdout set composed of the remaining four actors from POFA that were not used in any of the Calder et al. experiments. The network was trained until error on the holdout set was minimal, and tested on all three types of composite examples.

Results: The results are shown in Figure 4. While both expression and identity were classified without error on all types of composites, there was a significant difference in a standard measure of reaction time from feed-forward networks: 1 - the maximum output. Figure 4(a) shows the reaction time of our network on the relevant stimuli.

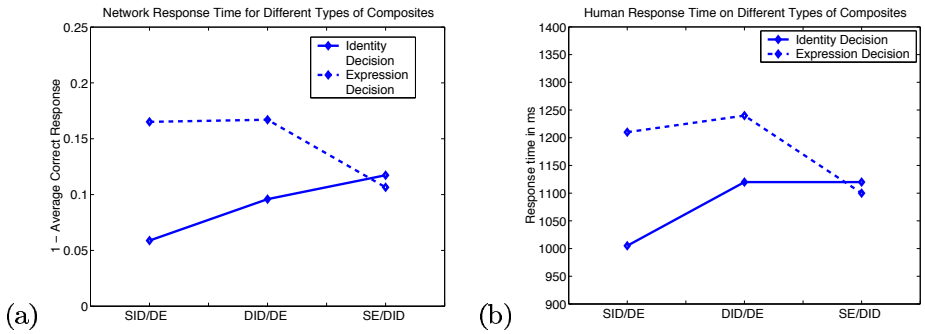

Figure 4: Experiment 3 results. Panel (a) shows the effect of composite type on our network reaction time for identity and expression recognition. Panel (b) is derived from (Calder et al., ress).

When classifying expression (dashed line), the model responds more quickly when the expression is the same in both halves of the composite example. Crucially, the model is not slowed more in the DID/DE case than the SID/DE case. The same result is true for identity classification. When classifying identity, the model is faster when the identity is the same in both halves of the composite example. Crucially, the model is not slowed more in the DID/DE case than the DID/SE case. In fact, it is slightly faster. The pattern of these results are equivalent to the pattern of the response times in the human experiments (Figure 4(b)).

Analysis

In this section we examine the kind of representation that the network is using for this task. In particular, we examine the principal components of the Gabor filters. Calder et al. showed that the principal components of gray scale images taken from the POFA database show a separation in terms of identity and expression. That is, there are components that best separate expressions, and different components that best separate the identities of the models in the POFA.

The principal components that best discriminate a data set, assuming the data samples are normally distributed, are those that minimize the variance of the samples within each class, while maximizing the total variance of all the samples. Therefore, we can rank the approximate discriminating power of each principal component by the Wilk's Λ value: the within-class scatter divided by the total scatter of the samples. The smaller the Wilk's Λ, the greater the discriminating power.

The two Wilk's Λ values for each principal component were computed for the two face recognition tasks compared in this paper: expression recognition and identity recognition. Only two components out of the highest 10 ranked principal components for each task overlapped. The 5th and the 19th principal components were ranked in the top 10 for both expression and identity recognition, but other than that the top ten components for the two tasks were disjoint.

This suggests that the principal components used to encode expression are separate from those used to encode identity. Figure 5(a) and (b) show the projections of the identity classes and expression classes on the most discriminating expression principal components. It is apparent that these components separate expression better than identity. Figure 5(c) shows the identity Wilk's Λ value plotted against the expression Wilk's Λ value, for corresponding principal components.

Discussion

Our results suggest that our simple neural network model can explain a variety of effects in psychological research in expression recognition. We found that our model showed nearly the same pattern of results in discriminability of expressions from half faces. In order to obtain this result, we had to change our model in two ways. First, we modeled the attentional process as an inhibition of irrelevant information, an approach that is well supported in the literature. Second, we had to add training on half faces to our model. We suggest that this modification is independently motivated by the fact that people foveate on different parts of the face when performing such tasks. Future research should concentrate on actually implementing an eye movement mechanism that is modulated by the task. Our previous model of scan paths (Yamada and Cottrell, 1995) only used bottom up information. Top down, task-related information can be incorporated by using the mutual information between the features and the categories – essentially, feature selection.

Second, we showed that our model's error patterns when shown composite versus non-composite faces follow the reaction time pattern in the human subjects data. That is, where our model makes more errors, the humans have longer reaction times. In an unreported experiment, we found the correct pattern

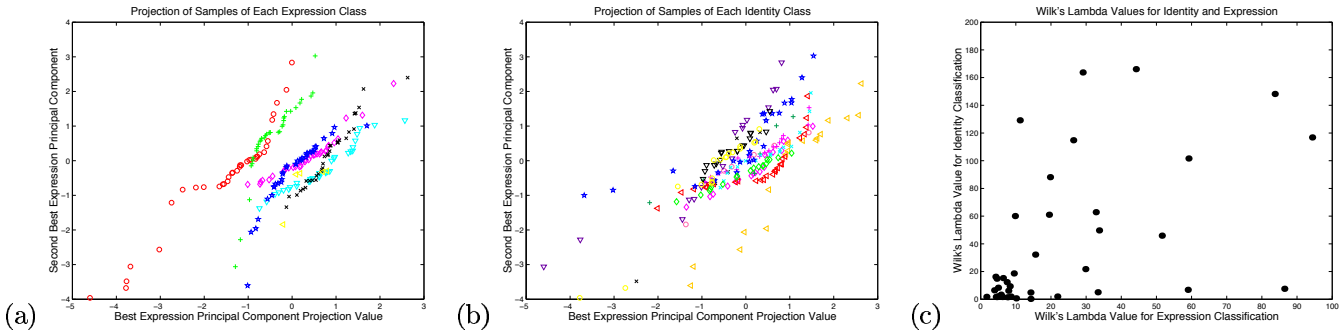

Figure 5: PC Analysis: (a) Plots of each class of expression on the two most discriminating expression principal components; (b) Plots of each class of identity on the two most discriminating expression principal components; (c) Identity and expression Wilk's Λ values for corresponding principal components.

of reaction times when the network is only shown a half face versus a composite face. This suggests that either the human subjects do not look at the other half of the face at all when they are misaligned, or that a greater degree of attenuation of the misaligned half should lead to results more in keeping with the RT data.

Third, our model showed that it is not necessary to posit two independent representations for identity and expression processing. Since the representation at the principal components layer is a set of orthogonal vectors, and the categorizer is a single layer perceptron, this suggests that each output unit is simply cutting off a different corner of the feature hypercube, and the learned hyperplane is simply orthogonal to the non-informative directions of variation. This is exactly what one would expect given this type of model, and so there is no mystery in our results.

This result might also have been expected given previous results examining the principal components of gray scale images of facial expressions directly (Calder et al., 2001), where it was found that expression and identity tended to load on different principal components. The main difference here is that our principal components are performed on a more biologically plausible representation of faces than gray scale images. However, we find that when we do a similar analysis to that carried out by Calder et al. (2001), we also find that the representation loads expression and identity on orthogonal components. Does this mean that we have two representations? If one thinks of each principal component as a linear "neuron," and the projection of an input on that component as its activation, then observing these activities will appear to show units that respond to identity, and other units that respond to expression. One could argue that these are then separate representations.

However, we argue that when viewed ontologically these representations were developed to be orthogonal *with respect to one another*. That is, due to the constrain that the principal components must be orthogonal, the representation of emotion is made in the context of and in competition with the representation of identity, as these are the major directions in which the data vary. On the other hand, one can say that this is a *functional* separation of the representations, and indeed, it is.

Acknowledgments We thank Gary's Unbelievable Research Unit (GURU) for discussion and comments. This research was supported by NIMH grant MH57075 to GWC.

References

Calder, A. J., Burton, A. M., Miller, P., Young, A. W., and Akamatsu, S. (2001). A principal component analysis of facial expressions. *Vision Research*, 41:1179–1208.

Calder, A. J., Young, A., Keane, J., and Dean, M. (In Press). Configural information in facial expression perception. *JEP:HPP*.

Dailey, M., Cottrell, G., and Adolphs, R. (2000). A six-unit network is all you need to discover happiness. In *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*, pages 101–106, Mahwah, New Jersey. Lawrence Erlbaum Associates.

Ekman, P. and Friesen, W. (1976). *Pictures of Facial Affect*. Consulting Psychologists, Palo Alto, CA.

Farah, M. J., Wilson, K. D., Drain, M., and Tanaka, J. N. (1998). What is "special" about face perception? *Psychological Review*, 105(3):482–498.

Tanaka, J. and Farah, M. (1993). Parts and wholes in face recognition. *Quarterly Journal of Experimental Psychology. A: Human Experimental Psychology*, 46(2):225–245.

Yamada, K. and Cottrell, G. (1995). A model of scan paths applied to face recognition. In *Proceedings of the 17th Annual Conference of the Cognitive Science Society*, pages 55–60, Mahwah, New Jersey. Lawrence Erlbaum Associates.

Young, A., Hellawell, D., and Hay, D. (1987). Configural information in face perception. *Perception*, 16:747–759.