A Deep Siamese Neural Network Learns the Human-Perceived Similarity Structure of Facial Expressions Without Explicit Categories

Sanjeev Jagannatha Rao (sjrao@ucsd.edu)
Department of Computer Science and Engineering, University of California San Diego
9500 Gilman Dr, La Jolla, CA 92093 USA

Yufei Wang (yuw176@eng.ucsd.edu)
Department of Electrical and Computer Engineering, University of California San Diego
9500 Gilman Dr, La Jolla, CA 92093 USA

Garrison W Cottrell (gary@ucsd.edu)
Department of Computer Science and Engineering, University of California San Diego
9500 Gilman Dr, La Jolla, CA 92093 USA

Abstract

In previous work, we showed that a simple neurocomputational model The Model, or TM) trained on the Ekman & Friesen Pictures of Facial Affect (POFA) dataset to categorize the images into the six basic expressions can account for wide array of data (albeit from a single study) on facial expression processing. The model demonstrated categorical perception of facial expressions, as well as the so-called facial expression circumplex, a circular configuration based on MDS results that places the categories in the order happy, surprise, fear, sadness, anger and disgust. Somewhat ironically, the circumplex in TM was generated from the similarity between the categorical outputs of the network, i.e., the six numbers representing the probability of the category given the face. Here, we extend this work by 1) using a new dataset, NimsStims, that is much larger than POFA, and is not as tightly controlled for the correct Facial Action Units; 2) using a completely different neural network architecture, a Siamese Neural Network (SNN) that maps two faces through twin networks into a 2D similarity space; and 3) training the network only implicitly, based on a teaching signal that pairs of faces are in either in the same or different categories. Our results show that in this setting, the network learns a representation that is very similar to the original circumplex. Fear and surprise overlap, which is consistent with the inherent confusability between these two facial expressions. Our results suggest that humans evolved in such a way that nearby emotions are represented by similar appearances.

Keywords: facial expressions; similarity structure; deep siamese neural network; multidimensional scaling (MDS); facial expression circumplex

Introduction

According to Darwin, facial expressions of emotion evolved and adapted to prepare the organism to deal with its environment and to also serve to communicate the internal state of the organism (Darwin, 1872; Hess & Thibault, 2009). If facial expressions of emotion are an outward manifestation of an internal state, then similar internal states should lead to similar expressions, in order to make the outward manifestations consistent and easy to understand. At the same time, expression of different emotions should also be sufficiently distinguishable in order to make it possible to properly respond to them.

How are facial expressions represented in the brain? There are two competing theories. One theory is based on experimental evidence of categorical perception of expressions of emotion, suggesting that the representation of facial expressions is divided into discrete categories. Once an expression has been categorized, the subtleties of the expression are lost.

An opposing theory suggests that perception of facial expressions is not as discrete as suggested by data supporting categorical perception. This notion of facial expression perception suggests that while some facial expressions have full membership in one of the six basic emotion classes (happy, disgust, angry, sad, fear, surprise), that nevertheless there is an underlying similarity structure to the expressions. Russell is the strongest advocate of this view, and has presented results that support this notion of perception of facial expressions (Russell & Bullock, 1986; Russell, 1980; Russell, Lewicka, & Niit, 1989). This and other related research suggests that there is a continuous underlying multidimensional perceptual space in which there are clear neighborhood relationships between expression categories, where each expression is closer to some expressions than others.

Dailey et al. (2002) developed a neural network model trained to classify facial expressions into six basic emotions (this model is referred to as "The Model" (TM) in (Cottrell & Hsiao, 2011)). The model was able to fit data usually taken to support each of the two competing theories of facial expression recognition (Young et al., 1997). It displayed categorical perception as well as graded reaction times near category boundaries, and responses indicating that the model was sensitive to mixed-in emotions even on the opposite side of the category boundary.

Dailey et al. performed MDS on the human forced-choice responses published by (Ekman & Friesen, 1976) and on their model's responses to the same stimuli. These are shown in Figures 7 and 8, respectively. They showed that the ordering of emotions is the same in both the cases, a result that is unlikely to have occurred by chance (the probability of this outcome is 1/60, or 0.017). This reflects clear neighborhood relationships between facial expressions.

In this work, we aim to reproduce these results from MDS, albeit under more restrictive training conditions. In particular, the model is only told which faces are in the same category

and which faces are in different categories and is not explicitly given the categories themselves. To the best of our knowledge, this is the first time the circumplex has been shown to arise from facial expression data under such restrictive training conditions.

We design a siamese neural network and train it to learn a 2D representation of facial expressions. The network is trained on pairs of images with a binary label that indicates if the two images belong to the same or different facial expression categories. In essence, this model is not explicitly trained on the number of facial expressions categories in the underlying data, or on the relative relationships among the facial expressions. The low dimensional representation produced by the siamese network replicates to a large extent the circumplex found by Dailey et al. (Dailey, Cottrell, Padgett, & Adolphs, 2002).

Our results suggest that facial expressions of emotion have evolved to make their appearance easily discriminable, and that compatible inner states produce similar expressions. The similarity structure in the low-dimensional space discovered by the network indicates that human expressions of similar emotions are closer to each other when compared to the dissimilar ones. The inherent confusability between our perception of facial expressions is explained by the overlapping clusters in our representations. For example, the siamese network overlaps the Fear and Surprise clusters, which are known to be prone to confusion. The fact that we obtain distinct clusters in our similarity structure demonstrates our ability to express dissimilar emotions in a differentiable way.

Siamese Neural Network Model

Dimensionality Reduction and Siamese Neural Networks

Two classical methods for dimensionality reduction are Principal Component Analysis (PCA) and Multidimensional Scaling (MDS). PCA finds a linear projection of the input data to a low dimensional space that maximizes the explained variance. MDS arranges the data in the low dimensional space in a manner that best preserves the pairwise distances between input points. However, facial expression images pose several challenges, similar to those posed in any computer vision application. Changes in lighting can make images of dissimilar emotions more similar, and similar ones different. In emotion recognition, the identity of the individual is a confound; identity is noise with respect to expression, and vice-versa. This suggests that a nonlinear embedding is required. MDS provides this, but it does not provide a mapping of new data into the same space, so it is difficult to check for generalization.

We require a dimensionality reduction technique that is robust to these changes in input, and that provides a way to generalize to new images in order to check that the embedding is consistent. In this work, we aim to learn the low dimensional structure of facial expressions data without relying on the total number of categories in our data and without associating explicit category labels to each input data point. Siamese neural networks fit these modeling requirements perfectly.

Siamese neural networks are comprised of two neural networks that take a pair of images as input and share a common contrastive loss function. Like siamese twins who share organs, the two networks of a siamese neural network are identical to each other in their architecture, and they share the same weights.

Figure 1 shows the layout of our siamese neural network. It receives a pair of images that are resized to 227 x 227 as input in its first layer. Each of these inputs is then processed through a dedicated 6 layer feed forward network as shown in the Figure. The first three layers are convolutional and the last three layers are fully connected. The activations of the last layer from each network are used to compute the loss.

The loss function is an energy-based one that is designed to move the representations of pairs inputs that are supposed to be "the same" closer together, and ones that are supposed to be different farther apart. We use the loss developed in (Hadsell, Chopra, & LeCun, 2006). Let X_1 and X_2 be two images presented to the system, one to each network. Y is a binary label assigned to the pair, with $Y = 0$ if the images supposed to be similar, and $Y = 1$ if they are supposed to be different. G_1 and G_2 are the activation vectors of the last layer of each network, just before the contrastive loss function in Figure 1. Let $D_w = ||G_1 - G_2||$ be the Euclidean distance between these vectors, where the subscript indicates the dependence on the weights W of the network. Then the loss function is:

$$L = (1-Y)\frac{1}{2}(D_w)^2 + Y(\frac{1}{2}max(0, m - D_W))^2 \qquad (1)$$

where $m > 0$ is a margin. This loss function is inspired by an analogy to springs, where minimizing the first term corresponds to a spring pulling G_1 and G_2 closer together, and the second term corresponds to a repulsing spring, pushing G_1 and G_2 farther apart. This loss function can be optimized by gradient descent. In order to map the faces into a two-dimensional space, G_1 (and hence G_2) are composed of two units.

Siamese networks have been shown to work well in face verification (Chopra, Hadsell, & LeCun, 2005), where the categories are not known in advance, since there are an unbounded number of faces. The networks in this case map faces of the same person to nearby places in the representational space (the last layer), and faces of different people farther apart. Siamese neural networks have also been shown to work well at dimensionality reduction (Hadsell et al., 2006). We take our loss function from the latter publication. Both these models use deep convolutional neural network architectures to extract features from the input images.

Dataset

We use facial images corresponding to the six basic emotions, Happy, Surprise, Fear, Sad, Anger and Disgust from

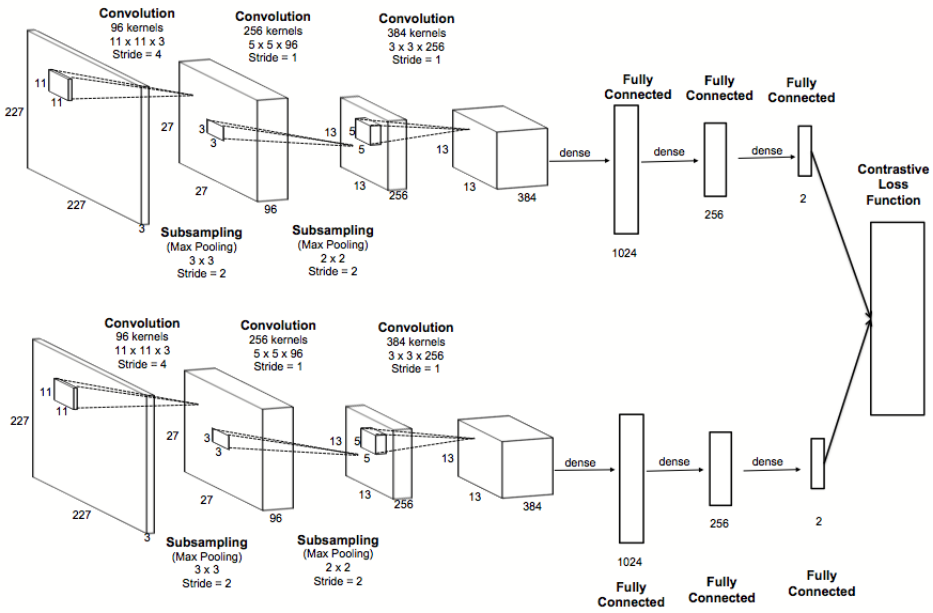


Figure 1: Siamese Neural Network Architecture

the NimStim dataset (Tottenham et al., 2009) for our analysis. We create all possible pairs from the images corresponding to these six basic emotions and use that as input to our siamese network. In all, we train on 126,756 pairs of images. The breakdown of category-wise pair counts is given in Table 1. Originally, we tried to balance the number of similar and dissimilar pairs, however, we ended up losing a significant amount of data and the model did not generalize well to unseen data. Hence we used all of the data, as shown in the table.

Approximately 10% of the subjects in the dataset are set aside for a validation set and 10% for a test set. The remaining 80 percent of the subjects contribute to the training set.

(a) Angry (b) Disgust (c) Fear

(d) Happy (e) Sad (f) Surprise

Figure 2: Sample Images from NimStim

Dealing with Limited Data

Deep convolutional neural networks (CNN) are trained on several hundred thousands of images. A large data set is required to learn the large number of parameters in the network. We are constrained by the relative small size of our dataset. A workaround for a small dataset is to initialize our model with a pre-trained model that will generalize to our problem.

The winning model of the ImageNet LSVRC-2012 contest (Krizhevsky, Sutskever, & Hinton, 2012) (dubbed "Alexnet") broke new ground in CNNs by using a 8 layer deep convolutional neural network. This model was trained to classify natural images into 1000 different categories. This model, along with its weights are publicly available. We use the first three convolutional layers as a starting point to build and train our siamese neural network.

Architecture

We experimented with several architectures, and we report the one that gave the minimum loss on the training set. Though we present only one architecture, all architectures that resulted in a significant reduction in loss during training yielded essentially the same representation. The network contains 6 layers, the first three are convolutional, and the next three are fully connected. Two such networks together form our siamese architecture. The output of the last fully connected layer serves as input to our loss function. We build our first three convolutional layers from the pre-trained weights on the ImageNet LSVRC-2012 dataset. We found this initialization to work really well for our purposes and has helped us cope with our limited dataset.

1. The first convolutional layer filters the $227 \times 227 \times 3$ input

image with 96 kernels of size $11 \times 11 \times 3$ with a stride of 4 pixels.

2. The second convolutional layer takes as input the (response-normalized and pooled) output of the first convolutional layer and filters it with 256 kernels of size $5 \times 5 \times 96$.

3. The third convolutional layer has 384 kernels of size $3 \times 3 \times 256$ connected to the (normalized, pooled) outputs of the second convolutional layer.

4. The fourth, fifth, and sixth layers are fully connected with 1024, 256, and 2 units, respectively.

The loss function takes its inputs from the 2 units of the sixth layer from each of the two individual networks. We have chosen to have two units in the last fully connected layer in order to extract a two dimensional representation of our data.

The siamese neural network was developed using the Caffe deep learning framework (Jia et al., 2014). We use ReLU (recitified linear units) activation functions throughout our architecture. We use a base learning rate of 0.0001, a step learning rate policy with a step size of 5000. The margin m in Equation 1 is set to 1. We use dropout after the fourth and the fifth fully connected layers. Training the model on a single GPU took around 5 hours. The layers were initialized with Xavier initialization and stochastic batch gradient descent was used during training.

Training

We started with pre-trained weights on the first three convolutional layers and trained only the subsequent layers. We stopped training when there was no additional improvement in the performance on the validation set. We then fine tuned the first three convolutional layers as well. Fine tuning was done for 5000 epochs at which point the loss did not reduce any further. The representations learned through the process of training are shown in figure 3. We plot the activations in the last layer of the network for each image in the training set to generate these plots.

Evaluation

Figure 3 shows how the different categories are organized by the network during the course of training. Each point in the plots represents one image in the NimStim dataset corresponding to one of the six basic emotions of happy, sad, angry, fearful, surprised and disgusted. At the start of training, the network is unable to differentiate the facial expressions as shown in Figure 3(a). The six basic emotions begin to form clusters around the 4000 epoch mark (Figure 3(d)), and become distinct after 5000 epochs as shown in Figure 4. Until this point the convolutional layers were fixed at their initial values, and at 5000 epochs the loss reached its minimum. At this point, we started fine-tuning the pre-trained layers and there is a further drop in loss. As expected, the representation becomes more distinct after fine tuning as seen in Figure 5.

The model generalizes to unseen data within the NimStim dataset. Its performance on test set, shown in Figure 6, is similar to that on the training set.

Table 1: Image Pairs by Emotion Catogory

Emotion	Angry	Disg.	Fear	Happy	Sad	Surpr.
Angry	3741	7134	6960	11049	7308	3828
Disgust		3321	6560	10414	6888	3608
Fear			3160	10160	6720	3520
Happy				8001	10668	5588
Sad					3486	3696
Surprise						946

The reader should compare the organization of the facial expressions in Figure 5 with that seen in Figures 7 (the MDS of human responses) and 8 (the MDS of The Model's responses). The human MDS is derived from the human subjects' averaged six-alternative forced choice responses for each face in the POFA dataset, as published by (Ekman & Friesen, 1976) and on The Model's responses to the same stimuli as reported by Dailey et al. (2002).

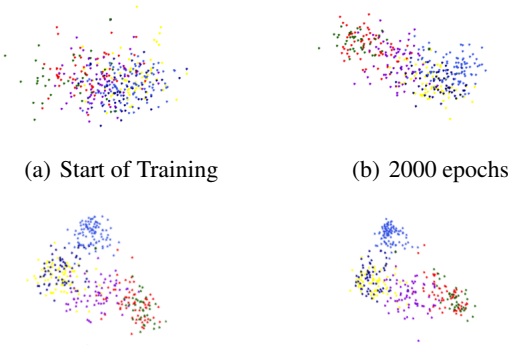
Of particular interest here is the ordering of the representations of facial expressions in two dimensions. The ordering of the emotions in the results reported by Dailey et al. and in the representations produced by the siamese network model are very similar. However, we do not get a perfect circumplex. Surprise and fear images completely overlap in our representation, which is consistent with the inherent confusability between them. Disgust and anger are just barely separated, and these too are expressions that are confused by human subjects.

The resulting order is unlikely to have happened by chance. The probability of a random ordering of six emotions matching the representation in Figure 7 is only 1/60: starting with any emotion, there are 5 to choose from next, 4 after that, etc. This gives 120 possibilities, but whether they are clockwise or counterclockwise does not matter, so there are 60 possible events.

To compute the probability of the current results, we can consider that we have a failure to separate two emotions, so the results are consistent with an ordering of Happy, Surprised, Fearful, Sad, Angry and Disgusted, or Happy, Fearful, Surprised, Sad, Angry and Disgusted. Since each of these ordering have a probability of 1/60, both together have a probability of 1/30, or 0.033.

Dailey et al. (2002) found that happy faces were the easiest to classify fear faces the most difficult to classify, consistent with the literature (Katsikitis, 1997; Ekman & Friesen, 1976; Matsumoto, 1992). The results of the siamese network model are consistent with these patterns. Happy images have been pushed into a tight cluster in the two dimensional representation, such that they are essentially linearly separable from the others, even in this very low dimensional space, and fear is completely overlapping with surprise, making it impossible to separate from the others.

The siamese network model has not been trained to classify the emotions into the six categories used by humans; rather it



(a) Start of Training (b) 2000 epochs

(c) 3000 epochs (d) 4000 epochs

Figure 3: Representations during training. Refer Figure 4 for legend.

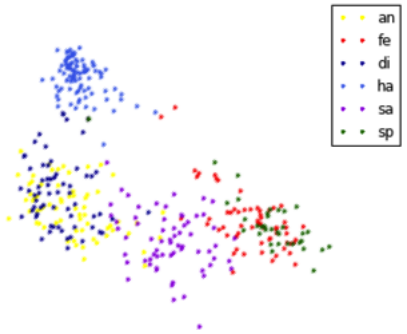


Figure 4: Representation after training for 5,000 epochs without fine tuning. Legends used in Figures 3-6: an: Angry, fe: Fear, di: Disgust, ha: Happy, sa: Sad, sp: Surprise.

has simply been trained on what humans consider "same" or "different" categories. These results, therefore, suggest that the similarity structure learned by the network is inherent in the similarity structure of the faces and the fact that some are different from others. We further hypothesize that, not only have the expressions evolved to be discriminable, but similar emotions have similar expressions.

Conclusions

We have presented a siamese neural network model that derives low dimensional representations of facial expressions under restrictive training conditions.

The network is only given same/different information about the images, it is not given any similarity information, so the structure of the clusters reflects the similarity between the expressions themselves. In Dailey et al., 2002, the circumplex was derived from the softmax output of the network, which also reflects confusability, but here, the network is deriving the similarity structure solely from the images and the information that some are the same, and some are different. It is never told which *categories* are similar to each other, so that is induced by the network from the similarity in the images. If Darwin is correct, and emotional expressions signal

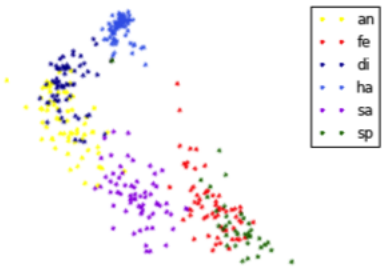


Figure 5: Representation after fine tuning. Refer Figure 4 for legend.

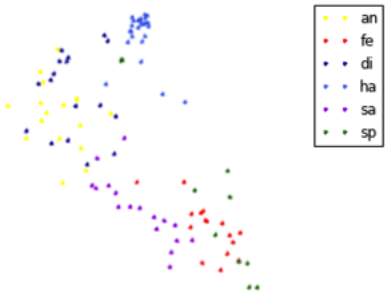


Figure 6: Test set representation. Refer Figure 4 for legend.

internal states, then the model predicts that anger and disgust have similar internal states, and fear and surprise also signal similar internal states.

Our results suggest that, through evolution, our facial expression of emotions and their perception have developed to communicate an inner state that is easily differentiable, and that associated emotional states are communicated similarly. Disgust and anger are often combined in everyday life, and in more exciting, if unfortunate, circumstances, fear and surprise are highly compatible and tend to co-occur. Our network has no access to these notions, no access to culture, yet it places these pairs of emotions either next to each other (as in disgust and anger), or overlapping (as in fear and surprise). The fact that we obtain relatively distinct clusters in our similarity structure suggests that our emotional expressions are inherently differentiable.

Acknowledgments

This work was supported in part by NSF grant SMA 1041755 to the Temporal Dynamics of Learning Center, an NSF Science of Learning Center. We are also grateful to members of Gary's Unbelievable Research Unit (GURU) for their help.

References

Chopra, S., Hadsell, R., & LeCun, Y. (2005). Learning a similarity metric discriminatively with application to face ver-

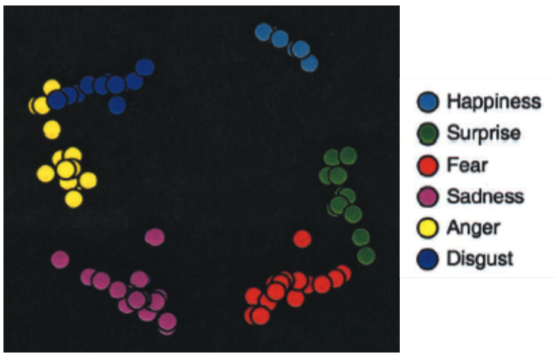


Figure 7: Human MDS representation of facial expressions.

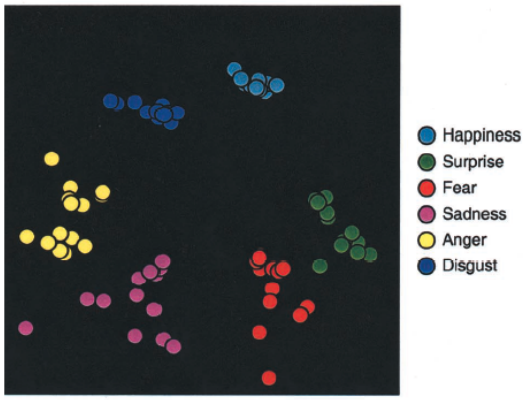


Figure 8: The Model's (Dailey et al., 2002) MDS Representation

ification. *Computer Vision and Pattern Recognition*, 539-546.

Cottrell, G., & Hsiao, J. (2011). Neurocomputational models of face processing. In *The Oxford Handbook of Face Perception.* Oxford, UK: Oxford University Press.

Dailey, M. N., Cottrell, G. W., Padgett, C., & Adolphs, R. (2002). Empath: A neural network that categorizes facial expressions. *Journal of Cognitive Neuroscience*, *14*(8), 1158-1173.

Darwin, C. (1872). *The expression of emotions in man and animals*. New York.

Ekman, P., & Friesen, W. (1976). Pictures of facial affect. *Consulting Psychologist Press*.

Hadsell, R., Chopra, S., & LeCun, Y. (2006). Dimensionality reduction by learning an invariant mapping. *Computer Vision and Pattern Recognition*, *2*, 1735-1742.

Hess, U., & Thibault, P. (2009). Darwin and emotion expression. *American Psychologist*.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., ... Darrell, T. (2014). Caffe: convolutional architecture for fast feature embedding. *Proceedings of the ACM International Conference on Multimedia*, 675-678.

Katsikitis, M. (1997). The classification of facial expressions of emotion: A multidimensional scaling approach. *Perception*, *26*, 613-626.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 1097-1105.

Matsumoto, D. (1992). American–japanese cultural differences in the recognition of universal facial expressions. *Journal of Cross-Cultural Psychology*, *23*, 72-84.

Russell, J. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*, 1161-1178.

Russell, J., & Bullock, M. (1986). Fuzzy concepts and the perception of emotion in facial expressions. *Social Cognition*, *4*, 309-341.

Russell, J., Lewicka, M., & Niit, T. (1989). A cross-cultural study of circumplex model of affect. *Journal of Personality and Social Psychology*, *57*, 848-856.

Tottenham, N., Tanaka, J., Leon, A., McCarry, T., Nurse, M., Hare, T., ... Nelson, C. (2009, August). *The nimstim face stimulus set.* Development of the MacBrain Face Stimulus Set was overseen by Nim Tottenham and supported by the John D. and Catherine T. MacArthur Foundation Research Network on Early Experience and Brain Development. Please contact Nim Tottenham at tott0006@tc.umn.edu for more information concerning the stimulus set.

Young, A. W., Rowland, D., Calder, A. J., Etcoff, N., Seth, A., & avid I. Perrett, D. (1997, June). Facial expression megamix: tests of dimensional and category accounts of emotion recognition. *Cognition*, *63*, 271–313.