Learning Facial Emotion Representations across Cultures

Safoora Yousefi
Department of Computer Science
Emory University
safoora.yousefi@emory.edu

Robert Thorstad
Department of Computer Science
Emory University
robert.thorstad@emory.edu

Eugene Agichtein
Department of Computer Science
Emory University
eugene@mathcs.emory.edu

Abstract

In this paper we train a classifier for facial emotion classification on two different cultures and compare the representations it learns from each culture in terms of face parts that maximally activate intermediate filters.

1 Introduction and Related Work

The Basic Emotions Hypothesis in psychology states that there are 6 basic emotional expressions that are shared across cultures, a hallmark of which is universal distinctive facial expressions across cultures. [1, 2]. However, the universality of basic emotions is debated. There have been several reports of basic emotions perceived differently across cultures, for example fear perceived as threat [3]; anger confused with disgust [4]. Moreover, cultural differences have been discovered in features used to classify emotion category, ie emotion recognition. Easterners, for instance, focus more on eyes than the whole face [5].

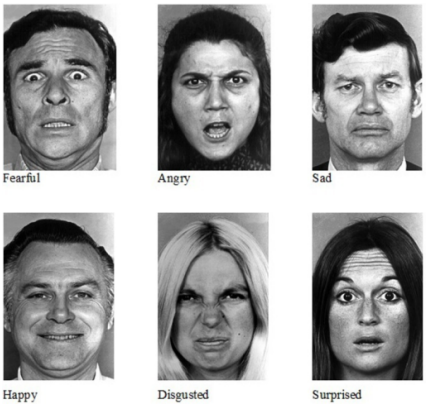

Figure 1: Examples of basic emotion photographs from the Picture of Facial Affect Dataset [6]. Across a wide range of cultures, participants presented with these pictures correctly label these pictures with the emotion displayed [7].

One problem may be over-reliance on posed images from Western posers: culture of poser affects classification accuracy, the in-group advantage [8], and posed images may under-represent cultural differences in emotion production, so-called emotional dialects [9].

We aim to test the basic emotions hypothesis using classification of naturally occurring images by a convolutional neural network. If basic emotions are universal, then we hypothesize that the accuracy of a classifier trained to recognize emotion category from images will not significantly differ as a function of culture of emotional images. Also, the features used by a classifier to categorize emotional expressions will not significantly differ as a function of culture. In this work, we aim to collect a cross-cultural dataset and design and implement experiments to test these hypotheses. The data collection and clean up process turned out to be a time consuming and labor intensive task and by the project due date, we were only able to collect a small pilot dataset to experiment with. Please regard this project as the development of techniques and sofware to later use with the final version of the dataset when it is ready.

Neural networks have been used in the past for the purpose of facial emotion recognition [10]. In [11] the authors employ a semi-supervised feature learning strategy to disentangle factors of variation in emotion recognition. While their strategy involves deep autoencoders, Khorrami et al outperform them using convolutional networks [12]. Authors in [13] use boosted deep belief networks for feature learning and emotion classification. Both [12] and [13] attempt to interpret the attention of their models, the latter feeds images to their network one patch at a time while the former uses the more artificially intelligent method of filter visualization proosed in [14] and [15].

The rest of the paper is organized as follows: In section 2 we explain the neural net architecture, training and interpretation methods. In section 3 we introduce our own collected dataset and a benchmark dataset that we experiment on. Sections 4 and 5 include results and discussion, respectively.

2 Methods

2.1 Classifier Architecture and Training

We implemented a deep convolutional network in Tensorflow, inspired by the work of [12]. The model consists of three convolutional layers each followed by Relu non-linearity and 2 x 2 max pooling. The convolutional layers are followed by a fully connected layer which in turn is followed by a Softmax layer to output class assignment probabilities. A visualization of the model is given in Figure 2.

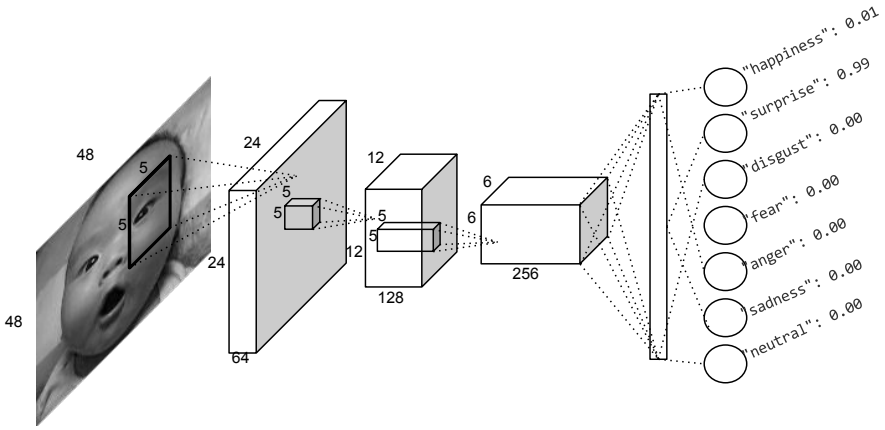

Figure 2: Emotion classification convolutional network. Code available at: `https://github.com/safooray/EmotionClassificationCNN`

We use gradient descent with exponential learning rate decay. We add the momentum term to the optimization to reduce the fluctuations due to the fact that we train the model one mini-batch at a time.

Figure 3: So far we have collected 400 facial expression images per each emotion {happy, suprised} per each culture {American, Korean}. After manual clean up for culture and emotion labels, the number of usable images is much smaller, as reported in Table 1

2.2 Feature Visualization

With a strong discriminative model at hand and inspired by the work of [15], we utilize a similar method to gain insight into discriminative features, that is, to answer questions such as "Where in the face does the model look at to decide about the emotion being expressed?". We aim to use this insight to study differences in emotion production between different cultures, by looking at what the model *looks at* in Korean faces compared to American faces in order to perform well in the classification task.

In the Softmax layer of the model, we have as many nodes as there are emotions in the training set. By calculating the gradient of the class probability with respect to the image and visualizing it, we can see what the model looks at in the image in order to classify it.

If we take one step further in the back-propagation and actually apply the above mentioned gradients to the image, we are able to achieve interesting emotion modification tasks that we explain in section 4.3.

3 Datasets

3.1 Natural Faces on the Web

Currently available datasets are highly reliant on Western posers. The NimStims dataset, as an instance of cross cultural datasets, relies on actors receiving their MA in Drama at NYU Tisch Program were shown examples of facial expressions and then asked to pose each facial expression. Muscles were adjusted until the desired expression was achieved, which might have led to under-representation of cultural differences in emotion production.

In an attempt to collect naturally occurring facial expression images labeled by culture and emotion expressed, we used Google image search using {culture, emotion} word pairs, such as "happy American". We call this dataset the *"Natural Faces on the Web (NFW)"*. Some samples of this dataset are shown in Figure3.

For non English speaking cultures we used queries in the corresponding language. We limited the search results to faces only and extracted the face rectangles from them using Microsoft Cognitive Services Face API. We have collected a pilot dataset for the purposes of periliminary experiments and manually cleaned up the labels for both culture and emotion. The number of samples in each class is shown in Table 1. We were able to collect data on 5 basic emotions but were not able to collect high quality images for the "disgusted" class yet.

Culture	Angry	Afraid	Happy	Sad	Surprised	Total
Korean	21	21	107	54	86	289
American	132	60	153	161	288	794

Table 1: Number of samples in each emotion class separated by culture.

	Angry	Afraid	Happy	Sad	Surprised	Total
Angry	**10.4**	0.6	1.2	1.4	2	15.6
Afraid	0.4	**1.2**	0.4	1.2	2.2	5.4
Happy	2	0	**12.4**	1.2	0.8	16.4
Sad	1.6	1.2	1.4	**10**	1.2	15.4
Surprised	1.6	1.2	0.6	1.6	**21**	26
Total	16	4.2	16	15.4	27.2	**78.8**

Table 2: Confusion matrix on testing set of **American** Natural Faces on the Web.The model performs best on Happy and Surprised classes due to more samples available in those classes.

We have an Amazon Mechanical Turk task ready to get correct labels for a larger dataset of several cultures. All the results in this report are based on the pilot dataset and the benchmark dataset introduced in the following subsection.

3.2 Toronto Face Dataset

Toronto Face Dataset (TFD) is an amalgamation of several smaller facial recognition and expression datasets with 4,178 expression labeled and 112,234 unlabeled images. Each image aligned and rescaled to 48 x 48 pixels [16].

3.3 Data Augmentation

We randomly flip and rotate images to a random amount throughout the stochastic gradient descent procedure. We only used flip and rotate transformations because we did not want to add unnecessary variations such as scale and crop that are not naturally present in the data set. One reason for this is to avoid complicating the feature visualization and model interpretation task.

4 Results

4.1 Training and validation

Our fist goal was to make sure the model is working and is a good discriminative model to build the rest of our research upon. To this end, we reproduced the state of the art results reported on the TFD in [12], as shown in Figure 4. We trained the model for 10,000 epochs (going over all data) but only show the first 1000 training steps (going over one mini-batch) in Figure 4 since no change was visible on the training and validation errors after this point. Testing error was measured after 10,000 training epochs.

4.1.1 Training on NFW

We trained the model on American and Korean faces separately, and tested on the corresponding culture only. We repeated each experiment with 5 different random allocations of samples to train, validation and test sets. The confusion matrices of the testing samples averaged over 5 trials are given in tables 3 and 2 for both NFW datasets. Table 4 shows overall and class specific misclassification rate in both cultures in NFW.

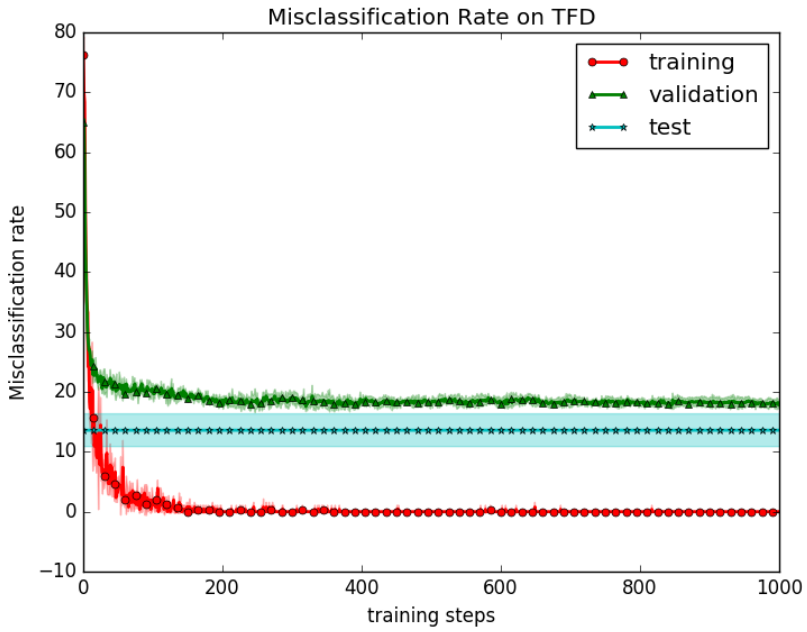

Figure 4: Performance of our convolutional network on the TFD dataset averaged over 4 cross validation folds. Shaded areas show standard variation and markers show the mean of the error over 4 folds.

	Angry	Afraid	Happy	Sad	Surprised	TOTAL
Angry	0.4	0	0.6	0	0.8	1.8
Afraid	0.2	0.6	1	0.2	1.2	3.2
Happy	0	0.2	**10**	1.4	1	12.6
Sad	0.4	0.6	1	2.2	1	5.2
Surprised	0	0.4	1.6	1.6	**5.6**	9.2
TOTAL	1	1.8	14.2	5.4	9.6	**32**

Table 3: Confusion matrix on testing set of **Korean** Natural Faces on the Web.The model performs best on Happy and Surprised classes due to more samples available in those classes.

4.1.2 Training on NFD after pre-training on TFD

We are postponing this step to after we acquire the full 7 class data with decent size, because in order to accomplish this task using the current NFW data we would have to discard some classes in TFD and train a model that will not be useful in the future with the final NFW data.

4.2 Feature Visualization Results

Applying the visualization technique proposed in section 2.2, we obtained saliency maps of images showing most discriminative features present in the image. We applied this method on the American and Korean models using images from the corresponding culture. We backpropagated to each image

% Misclassified	Angry	Afraid	Happy	Sad	Surprised	Total
Korean	60.0	66.6	29.5	59.2	41.6	**41.2**
American	35.0	71.4	22.5	35.1	22.8	**30.2**

Table 4: Percentage of misclassifies samples in each class and the overall misclassification rate of the model separated by culture.

only from its corresponding class node in the Softmax layer. Saliency map results of a random set of images are shown in Figures 5 and 6.

Figure 5: Rows from top to bottom depict salient features in **Korean** faces belonging to classes angry, afraid, happy, sad, and surprised. Features were obtained by guided back-propagation from the Softmax node corresponding to the class.

4.3 Side achievement: Emotion Modification

Using gradients from Softmax nodes corresponding to each emotion, we can intensify that emotion in any given image. This method can be used either to intensify the emotion that is present in the image, or to create an emotion that is absent in the image by reinforcing features in the image that activate the desired class. Results for both emotion intensification and emotion conversion are given in Figure 7.

5 Discussion and Future Work

In order to cast doubt on the Basic Emotion Hypothesis we have to be able to show difference in saliency maps obtained from different cultures. This seems impractical with the current size of our cross-cultural dataset but is definitely one of the most important things we aim to do with the NFW when fully collected.

Furthermore, instead of qualitatively examining the saliency maps, we plan to measure the proximity of high gradients to facial landmarks and therefore be able to offer a qualitative measure of how different emotion production is from one culture to another.

References

[1] Paul Ekman. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200, 1992.

Figure 6: Rows from top to bottom depict salient features in **American** faces belonging to classes angry, afraid, happy, sad, and surprised. Features were obtained by guided back-propagation from the Softmax node corresponding to the class.

[2] Carroll E Izard. Basic emotions, relations among emotions, and emotion-cognition relations. 1992.

[3] Carlos Crivelli, James A Russell, Sergio Jarillo, and José-Miguel Fernández-Dols. The fear gasping face as a threat display in a melanesian society. *Proceedings of the National Academy of Sciences*, page 201611622, 2016.

[4] James A Russell. Is there universal recognition of emotion from facial expressions? a review of the cross-cultural studies. *Psychological bulletin*, 115(1):102, 1994.

[5] Masaki Yuki, William W Maddux, and Takahiko Masuda. Are the windows to the soul the same in the east and west? cultural differences in using the eyes and mouth as cues to recognize emotions in japan and the united states. *Journal of Experimental Social Psychology*, 43(2):303–311, 2007.

[6] Paul Ekman and Wallace V Friesen. Measuring facial movement. *Environmental psychology and nonverbal behavior*, 1(1):56–75, 1976.

[7] Paul Ekman, Wallace V Friesen, Maureen O'Sullivan, Anthony Chan, Irene Diacoyanni-Tarlatzis, Karl Heider, Rainer Krause, William Ayhan LeCompte, Tom Pitcairn, Pio E Ricci-Bitti, et al. Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of personality and social psychology*, 53(4):712, 1987.

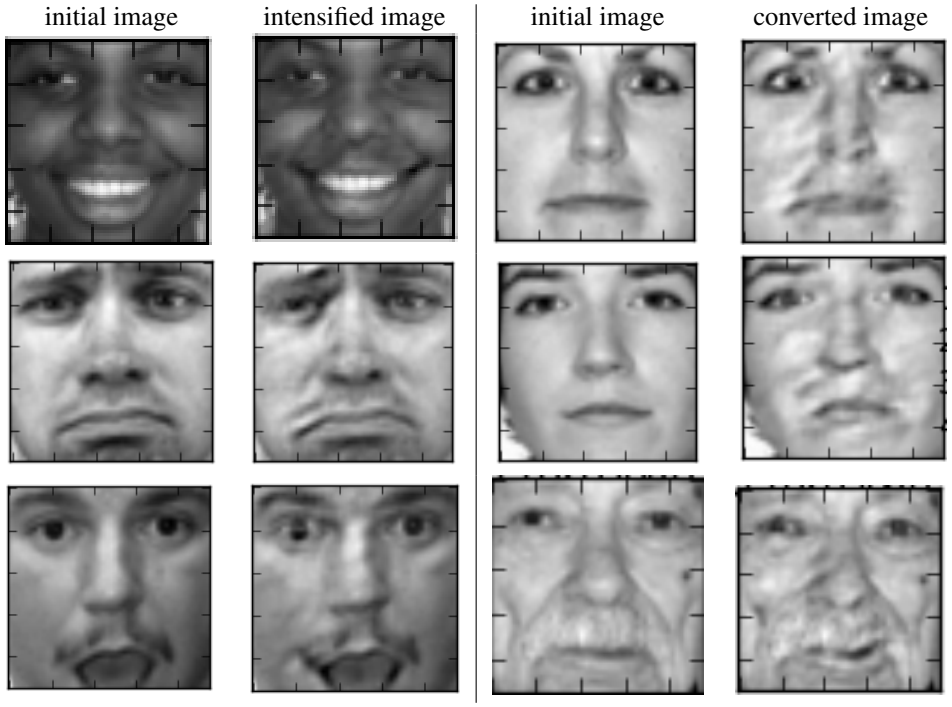

Figure 7: Images on the left were processed for emotion intensification, while in the images on the right we have converted the emotion from happiness to fear, happiness to sadness, and neutrality to happiness as you look from top to bottom.

[8] Hillary Anger Elfenbein and Nalini Ambady. On the universality and cultural specificity of emotion recognition: a meta-analysis. *Psychological bulletin*, 128(2):203, 2002.

[9] Hillary Anger Elfenbein, Martin Beaupré, Manon Lévesque, and Ursula Hess. Toward a dialect theory: cultural differences in the expression and recognition of posed facial expressions. *Emotion*, 7(1):131, 2007.

[10] Matthew N Dailey, Garrison W Cottrell, Curtis Padgett, and Ralph Adolphs. Empath: A neural network that categorizes facial expressions. *Journal of cognitive neuroscience*, 14(8):1158–1173, 2002.

[11] Salah Rifai, Yoshua Bengio, Aaron Courville, Pascal Vincent, and Mehdi Mirza. Disentangling factors of variation for facial expression recognition. In *European Conference on Computer Vision*, pages 808–822. Springer, 2012.

[12] Pooya Khorrami, Thomas Paine, and Thomas Huang. Do deep neural networks learn facial action units when doing expression recognition? In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 19–27, 2015.

[13] Ping Liu, Shizhong Han, Zibo Meng, and Yan Tong. Facial expression recognition via a boosted deep belief network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1805–1812, 2014.

[14] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*, pages 818–833. Springer, 2014.

[15] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.

[16] Joshua Susskind, Adam Anderson, and Geoffrey E Hinton. The toronto face dataset. *U. Toronto, Tech. Rep. UTML TR*, 1:2010, 2010.

[17] Paul Ekman and Daniel Cordaro. What is meant by calling emotions basic. *Emotion Review*, 3(4):364–370, 2011.

[18] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. Robust image sentiment analysis using progressively trained and domain transferred deep networks. *arXiv preprint arXiv:1509.06041*, 2015.

[19] Jianbo Yuan, Sean Mcdonough, Quanzeng You, and Jiebo Luo. Sentribute: image sentiment analysis from a mid-level perspective. In *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining*, page 10. ACM, 2013.

[20] Sanjeev Jagannatha Rao, Yufei Wang, and Garrison W Cottrell. A deep siamese neural network learns the human-perceived similarity structure of facial expressions without explicit categories.

[21] Yilin Wang, Suhang Wang, Jiliang Tang, Huan Liu, and Baoxin Li. Unsupervised sentiment analysis for social media images. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina*, pages 2378–2379, 2015.

[22] Joshua M Susskind, Adam K Anderson, Geoffrey E Hinton, and Javier R Movellan. *Generating facial expressions with deep belief nets*. INTECH Open Access Publisher, 2008.

[23] Andrew Ortony and Terence J Turner. What's basic about basic emotions? *Psychological review*, 97(3):315, 1990.

[24] Joshua M Susskind, Daniel H Lee, Andrée Cusi, Roman Feiman, Wojtek Grabski, and Adam K Anderson. Expressing fear enhances sensory acquisition. *Nature neuroscience*, 11(7):843–850, 2008.

[25] Lauri Nummenmaa, Enrico Glerean, Riitta Hari, and Jari K Hietanen. Bodily maps of emotions. *Proceedings of the National Academy of Sciences*, 111(2):646–651, 2014.

[26] Casey C Bennett and Selma Šabanović. The effects of culture and context on perceptions of robotic facial expressions. *Interaction Studies*, 16(2):272–302, 2015.

[27] Rachael E Jack, Caroline Blais, Christoph Scheepers, Philippe G Schyns, and Roberto Caldara. Cultural confusions show that facial expressions are not universal. *Current Biology*, 19(18):1543–1548, 2009.

[28] Rachael E Jack, Oliver GB Garrod, Hui Yu, Roberto Caldara, and Philippe G Schyns. Facial expressions of emotion are not culturally universal. *Proceedings of the National Academy of Sciences*, 109(19):7241–7244, 2012.