# Real-time Emotion Recognition from Facial Expressions

Minh-An Quinn, Guilherme Reis, Grant Sivesind, {minhan, greis, gsivesin}@stanford.edu
Stanford University, CS229, Autumn 2017

## Introduction

We built several models capable of recognizing emotions from facial expressions. Using the FER-2013 dataset of non-posed grayscale images, we achieve 47.8% accuracy using an SVM and 66.5% using a CNN; on the CK+ dataset, we achieve 99.5% accuracy.

We then built a real-time system to detect faces from a video feed and continuously classify them using our model, demonstrating the ability to transfer skills learned on the static datasets.

## Data

We trained our model on two datasets:

FER-2013 Dataset [1]
- 28,000 labeled emotions in training set, 3,500 labeled emotions in development set, and 3,500 labeled emotions in test set
- Images are posed and un-posed headshots: 48x48 pixels, grayscale
- 7 emotions: angry, disgust, afraid, happy, sad, surprised, neutral
- In Kaggle competition, top accuracy for FER-2013 was 71%

CK+ (extended Cohn-Kanade) Dataset [3]
- 5,876 labeled images of posed individuals
- Images are posed headshots: 640x490 pixels, mostly grayscale
- 8 emotions: angry, disgust, afraid, happy, sad, surprised, contempt, neutral

## SVM Model

For our baseline, we made both a one vs one (OVO) (rbf kernel) and a one vs all (OVA) (linear kernel) SVM. We experimented with raw features, scaled features, HOG features, and also tried reducing the feature space using PCA. [2]

## CNN Model

We trained our CNN on the FER2013 dataset, and experimented with a variety of techniques and architectures. Data augmentation helped considerably: we randomly rotate, shift, flip, crop, and sheer our training images.

The CNN's architecture is reminiscent of LeNet, but with more parameters:
Conv(32, 5x5) → Conv(64, 5x5) → MaxPool(2x2) → Conv2D(128, 3x3) → Dropout(0.1) → Maxpool(2x2) → FullyConnected(2048) → Dropout(0.5) → FullyConnected(1024) → Dropout(0.5) → Softmax(num_emotions)

All Convolutional and FC layers use ReLU activation. Dropout is used to prevent overfitting; together with the randomized data augmentation, train accuracy is actually kept below dev accuracy. We found the best optimizer to be Adadelta, using categorical cross-entropy loss.

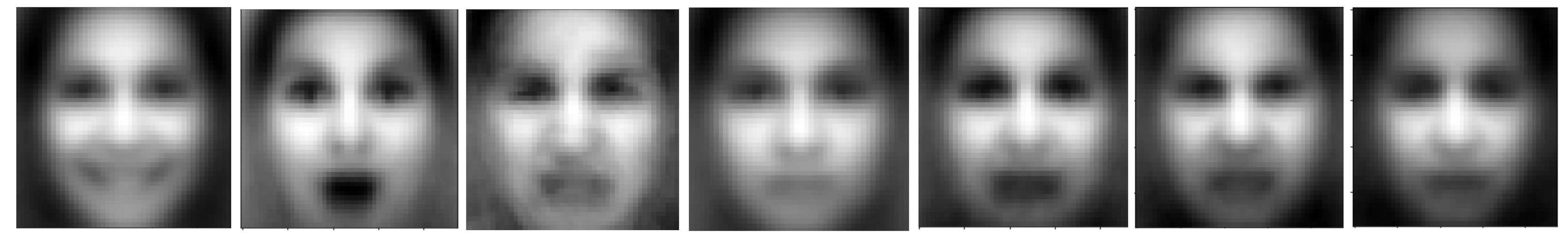

**Figure 1:** Average faces of each emotion from FER-2013
From left to right: happy, surprise, disgust, neutral, fear, anger, sad

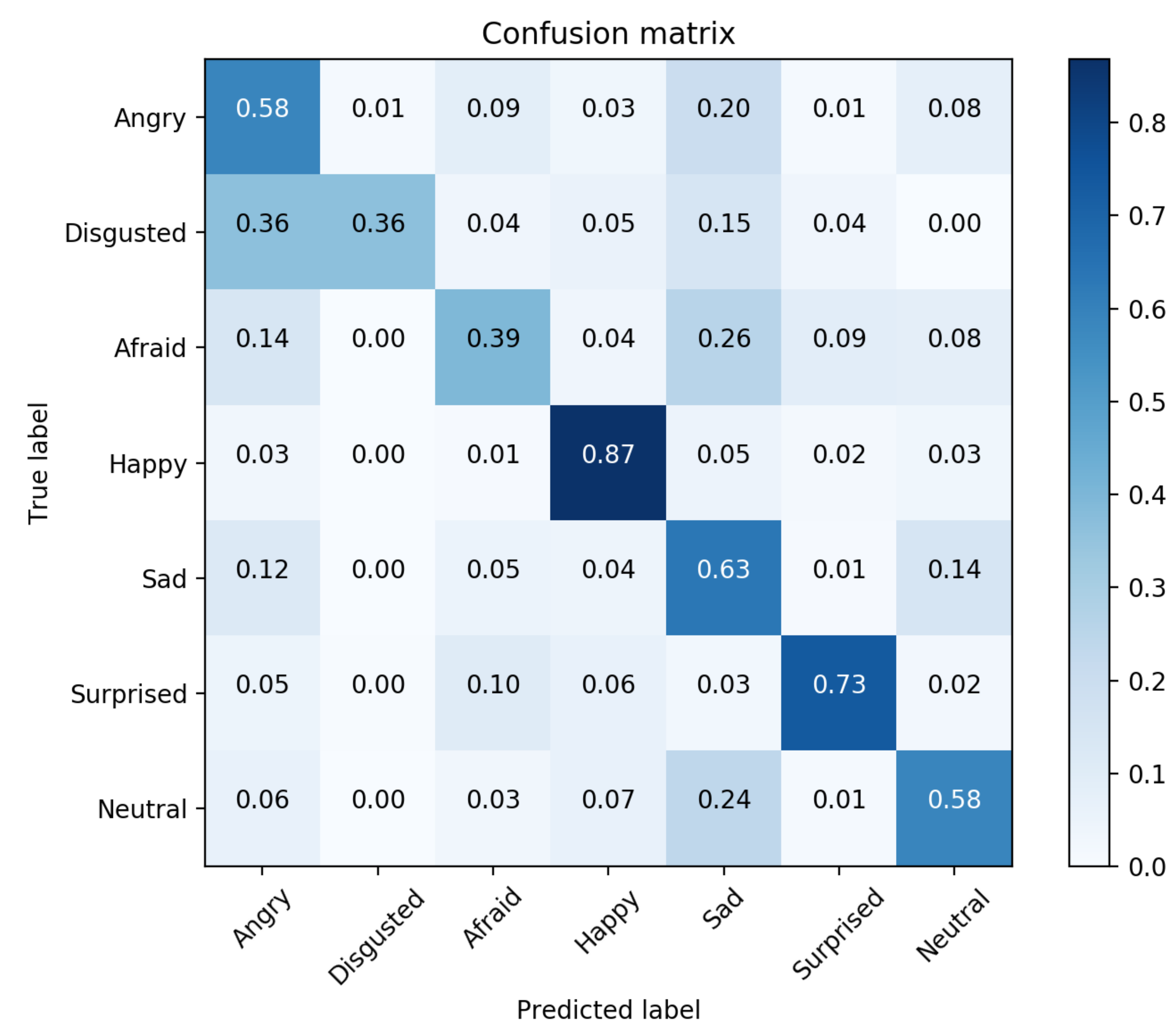


**Figure 2:** Confusion matrix for CNN on FER-2013

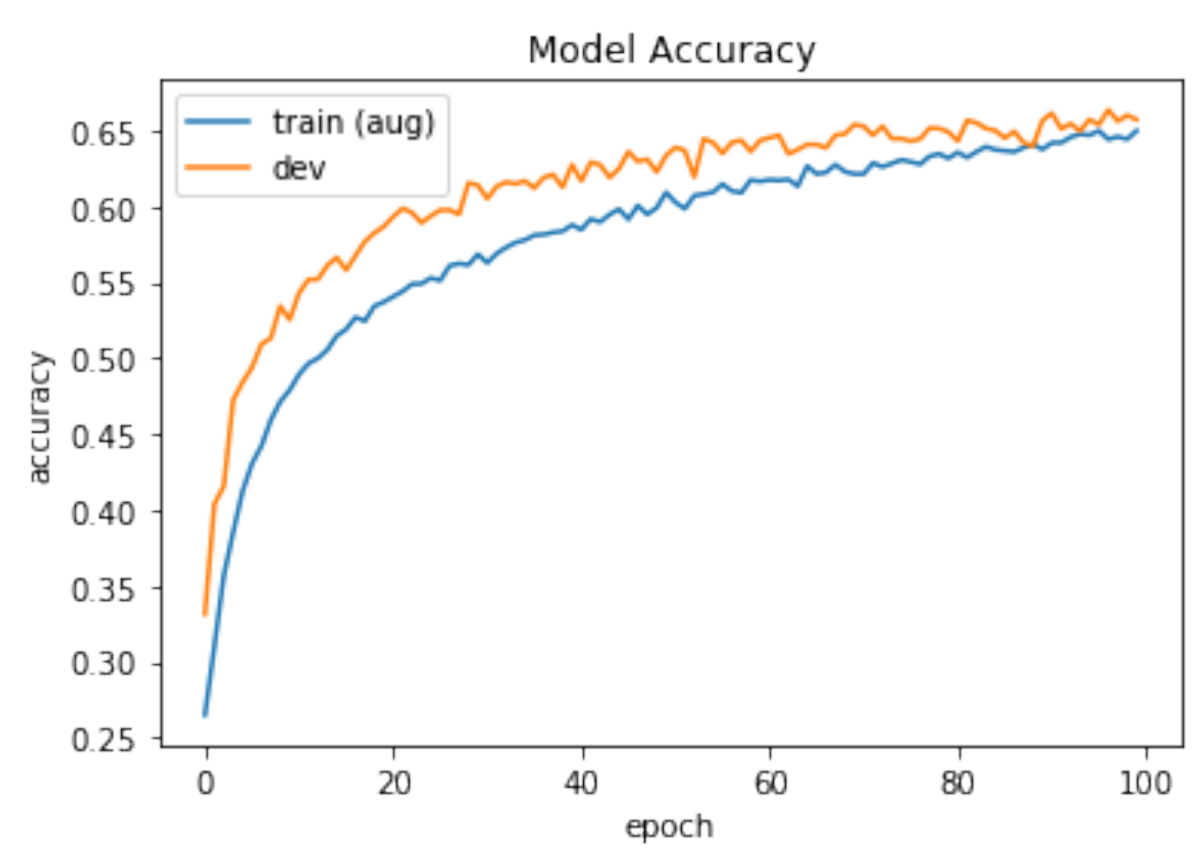


**Figure 3:** Model accuracy during training for CNN on FER-2013

## Results (FER-2013)

| Optimization | Featurization/ Hyperparameters | Training Accuracy | Test Accuracy |
|---|---|---|---|
| SVM (OVO) | Scaled pixels | 43.4% | 38.6% |
| SVM (OVO) | Scaled pixels, PCA – 25 comps | 41.9% | 38.6% |
| Linear SVM (OVA) | HOG (4,4) pixels/cell | 67.2% | 47.8% |
| CNN | See CNN Model Section | 65.0% | **66.5%** |

## Real-time Classification

We use a webcam's video feed and OpenCV's implementation of Haar Cascades to detect a square face region. We extract, grayscale, and resize the face region to be 48x48. We then use the CNN model to predict a probability distribution over emotions, and display that visually. Using GPU acceleration, this works with no lag in real-time.

## Discussion

On FER-2013, we achieve 66.5% accuracy using the CNN, well above guessing the most common class (24%) but a little below the top Kaggle score (71%). Human scores on FER-2013 are 65% +/- 5%. On the CK+ dataset, we achieve an accuracy of 99.55% using an SVM - which although near-perfect, is in line with what is reported in literature. All accuracies were computed on blind test sets.

We decided to focus on the FER-2013 datasets because its more diverse, un-posed images more closely reflect the distribution of images in real-time video capture. Real-time classification better exposed our model's strengths. Neutral, Happy, and Surprised are consistently well-detected. It also revealed a bias in our dataset: certain people have their emotions detected much more accurately. Artifacts such as glasses, beards, and poor illumination significantly affect performance.

## Future Work

To expand upon our work, further work can be done to:
- Refine the CNN structure: replace redundant parameters with others in more useful places in the architecture; in adapting the learning rate decay schedule; in adapting the location and probability of dropout; and in experimenting with stride sizes.
- Diversify static datasets to more closely resemble real-time data distribution

## References

[1] "Challenges in Representation Learning: A report on three machine learning contests." I Goodfellow, D Erhan, PL Carrier, A Courville, M Mirza, B Hamner, W Cukierski, Y Tang, DH Lee, Y Zhou, C Ramaiah, F Feng, R Li, X Wang, D Athanasakis, J Shawe-Taylor, M Milakov, J Park, R Ionescu, M Popescu, C Grozea, J Bergstra, J Xie, L Romaszko, B Xu, Z Chuang, and Y. Bengio. arXiv 2013.

[2] Dumas, Melanie. "Emotional Expression Recognition Using Support Vector Machines." Machine Perception Lab, Univeristy of California, 2001.

[3] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar and I. Matthews, "The Extended Cohn-Kanade Dataset (CK+): A complete facial expression dataset for action unit and emotion-specified expression," in 3rd IEEE Workshop on CVPR for Human Communicative Behavior Analysis, 2010

[4] Goodfellow, I. J., Erhan, D., Carrier, P. L., Courville, A., Mirza, M., Hamner, B.,... Bengio, Y. (2015). Challenges in representation learning: A report on three machine learning contests. Neural Networks, 64, 59-63. doi:10.1016/j.neunet.2014.09.005