# **Stress Testing Facial Recognition with Adversarial Examples: Progress Report Feedback Response**

Emily Strong

March 30, 2018

CSYE 7245 Spring 2018

## **Introduction**

My project is on testing the sensitivity of face recognition models to input designed to trick it. The feedback on my in-class progress report was that since I was having problems with the Labeled Faced in the Wild data set having too few images of my test subjects and more generally not having enough images of people who are not white men for selecting test subjects I should consider other data sets. I have decided to switch to the FaceScrub data set, which I describe below in *Data Set*. Additionally, I needed to clarify what my test cases are for stress testing image quality manipulations. To help with the clarification I have separated my test cases into two parts – Image Quality and Targeted Examples. My planned test cases are below in *Stress Tests*.

## **Methodology Changes**

### Data Set

I initially planned to use the Labeled Faces in the Wild data set, however after problems with validating the face recognition workflow shown in the results section I am changing to FaceScrub [1]. FaceScrub is another famous people data set with 265 men and 265 women, with ~200 images of each person. Most of the subjects are white, a problem I found with all face image data sets that are both large and public, however it does not have the gender balance problems of LFW.

The data set was created in 2014 and for copyright reasons it is a list of URLs rather than the actual images. I wrote a web scraping script to obtain the images and found that because the URLs are several years old many are no longer valid. After cleaning up the files to remove the placeholders some websites use when a particular image is no longer available, the final data set has over 66,000 images with individuals having counts ranging from 40 to 197. For test cases I selected people with more than 100 images in the scraped data set. Keira Knightley and Natalie Portman are not in the data set, so I scraped images for them in a similar manner. These images I am using as the training data. Images for test cases are from Wikimedia Commons.

### Stress Tests

The test cases I am planning are as follows:

*Part A: Image Quality*

I will use the imgaug library [2] to implement these test cases. All test cases will be run with an image of Colin Firth and an image of George Lopez using a model trained on the full FaceScrub data set.

1. Invert and invert per channel (RGB)
2. Add and add per channel
3. Multiply and multiply per channel
4. Gaussian blur
5. Sharpen
6. Additive Gaussian noise
7. Salt and pepper noise
8. Contrast normalization

The script will track at what parameter value the face is no longer correctly identified and at what value a face is no longer detected, or will terminate when the maximum/minimum for that parameter is reached.

*Part B: Targeted Examples*

All test cases will be run with a model trained only on the people being used in the test cases and on a model trained on the full FaceScrub set. For test cases 1-5, each will be run twice per model, switching which person is being modified and which is the target.

1. People with similar faces
   1. Subjects: Keira Knightley, Natalie Portman
2. People with dissimilar faces (same race, same gender)
   1. Subjects A: Keira Knightley, Kristin Chenoweth
   2. Subjects B: Colin Firth, Matthew Perry
3. People with different genders (same race)
   1. Subjects A: Kristin Chenoweth, Colin Firth
   2. Subjects B: Tatyana M Ali, Samuel L Jackson
4. People of different races (same gender)
   1. Subjects A: Colin Firth, George Lopez
   2. Subjects B: Samuel L Jackson, Ken Watanabe
   3. Subjects C: Kristin Chenoweth, Tatyana M Ali
5. People of different races and genders
   1. Subjects A: Kristin Chenoweth, George Lopez
   2. Subjects B: Tatyana M Ali, Ken Watanabe
6. Non-human face
   1. Subjects: a cat, Colin Firth
7. No face
   1. Subjects: a bicycle, Colin Firth

## **Results**

I have validated that the OpenFace pipeline correctly identifies each of the test subjects with images that have not been modified for both a limited training set using only the test subjects and the full FaceScrub set. For comparison, I have included the LFW results from my progress report.

*Table 1: Model validation*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Subject** | **Original Test Subjects** | **Full LFW** | **New Test Subjects** | **Full FaceScrub** |
| Keira Knightley | 51%  88%  96% | 5%  2%  9% | 51%  99%  99% | 34%  85%  96% |
| Natalie Portman | 64%  98%  87% | 3%  13%  6% | 66%  98%  98% | 44%  95%  94% |
| Meryl Streep | 86% | 1% |  |  |
| Tom Hanks | 71% | Predict Mahmoud Abbas with 0.01 confidence |  |  |
| Morgan Freeman | 79% | Predict Kofi Annan with 0.02 confidence |  |  |
| Kristin Chenoweth |  |  | 99% | 87% |
| Tatyana M Ali |  |  | 94% | 73% |
| Matthew Perry |  |  | 99% | 92% |
| Colin Firth |  |  | 98% | 84% |
| George Lopez |  |  | 98% | 87% |
| Samuel L Jackson |  |  | 97% | 80% |
| Ken Watanabe |  |  | 95% | 53% |
| cat | Exception: Unable to find a face | Exception: Unable to find a face | Exception: Unable to find a face | Exception: Unable to find a face |
| bicycle | Exception: Unable to find a face | Exception: Unable to find a face | Exception: Unable to find a face | Exception: Unable to find a face |

## **Future Steps**

My next steps are to write the scripts for parts A and B. For part B I will need to determine if I can change the output from the SVM model used in the OpenFace pipeline to list multiple matches to use as feedback for a particle swarm, otherwise I will have to brute force the changes.

## **References**

[1] <http://vintage.winklerbros.net/facescrub.html>

[2] <https://github.com/aleju/imgaug>