AI Meets Online Dating
User Demographics and Profile Pictures as a Predictor of Attractiveness in Online Dating Profiles
Project Checkpoint

Tucker Leavitt (tuckerl), Duncan Wood (duncanw), Huyen Nguyen (huyenn)

Abstract— **This is where we will summarize our project.**

I. INTRODUCTION

Although judgments of beauty are subjective, the prevalence of beauty contests, cosmetic surgery and homogeneous ratings of photos on dating websites suggest there might be a common denominator for attractiveness perception. This project is an attempt to quantify that denominator.

There have been several experiments with a similar goal. In 2006, Yael Eisenthal and his team at Tel-Aviv University, Israel trained their model on 2 datasets, one contains 92 images of Austrian females and one contains 92 images of Israeli females. Using K-nearest neighbors and SVM with both quadratic and linear kernels, their model achieved a correlation of 0.65 with the average human ratings. The same year, Amit Kagian and his team, also from Tel-Aviv University, trained their model on a dataset of 91 facial images of American females. By using 90 principal components of 6972 distance vectors between 84 fiducial point locations, they achieved a Pearson correlation of 0.82 with human ratings. In 2009, two students in CS229, Hefner and Lindsay used a training set of 147 photos and a test set of 50 photos, all obtained from HotOrNot and got the maximum correlation with linear regression estimation. Most recently, in July 2015, Avi Singh from Indian Institute of Technology, Kanpur used Guassian Process Regression, K-nearest neighbor, Random Forest, SVM and Linear Regression with leave-one-out cross-validation on a dataset of 500 Asian females. He achieved correlations of 0.52, 0.6, 0.64, 0.22, 0.64 respectively.

Its noticeable that all the experiments have been done on small and racially homogeneous datasets with relatively low correlation. Through this project, we want to achieve two goals:

1) We want to see how our model generalizes with racially diverse datasets. For example, we want to train our model of a data which consists solely of Asian females and test our hypothesis on datasets of Caucasian or African females.
2) We want to train our hypotheses on a racially diverse dataset to see if it will increase or decrease the correlation.

II. DESCRIPTION OF DATA

For the initial phase, we used the SCUT-FBP dataset which contains 500 frontal images of 500 different Asian females with neutral expressions, simple backgrounds, and minimal occlusion. These factors are conducive to facial beauty perception in both geometry and appearance. Each image is rated by 75 users, both males and females, on a discrete scale from 1 to 5. We computed the average rating score for each image and used these as the labels .

As the SCUT-FBP dataset contains only of images of Asian females and we are interested in attractiveness rating across cultures, we plan to obtain more data points in the future by scraping data from HotOrNot. We chose HotOrNot because each HotOrNot profile photo has a rating we can easily access. We will then manually select photos from the obtained photos to ensure that each photo meets the following requirements:

1) It is a frontal image
2) Minimal occlusion with the whole face and hair
3) Neutral expression
4) Enough lighting to make all features visible

III. METHODS

As an initial proof-of-concept test, we implemented a quick-and-dirty multilabel classifier using the SCUT-FBP dataset. To construct our feature set, we manually tagged 19 different points on 30 different faces using pythons ginput feature. We then used the pairwise distances between all pairs of points as the input features of our classifier.

We used a multiclass logistic regression classifier without regularization to predict the average rating of each picture (as an integer between 1 and 5), using Pythons sklearn library.

We used a leave-one-out testing scheme to test our model. We trained our model on all but one of the pictures in our dataset, then used our model to predict the average rating of the left-out picture. We repeated this procedure for each picture in the model.

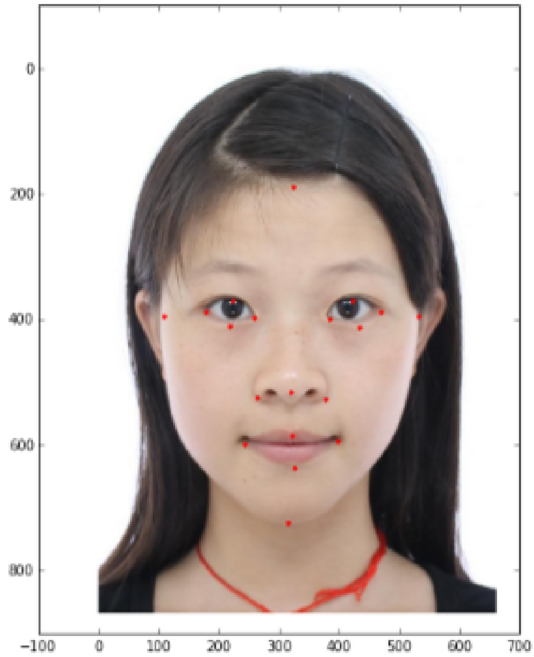
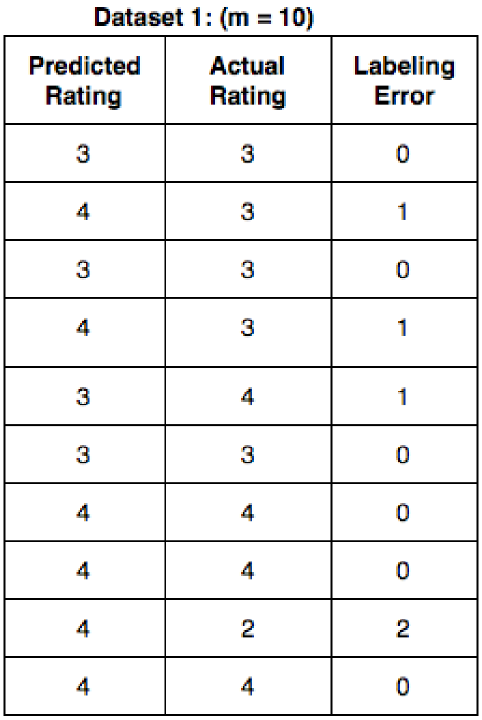


Figure 1: Example picture from the SCUT-FBP dataset, with added labels indicated.

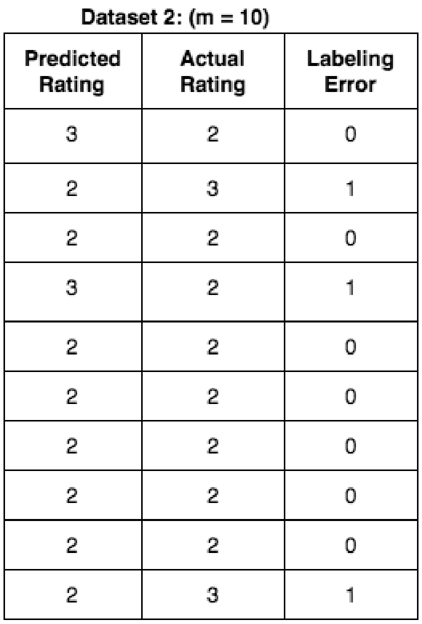
IV. Results

We split our dataset of 30 labeled faces into three sets of 10.



Dataset 1: (m = 10)

Predicted Rating	Actual Rating	Labeling Error
3	3	0
4	3	1
3	3	0
4	3	1
3	4	1
3	3	0
4	4	0
4	4	0
4	2	2
4	4	0



Dataset 2: (m = 10)

Predicted Rating	Actual Rating	Labeling Error
3	2	0
2	3	1
2	2	0
3	2	1
2	2	0
2	2	0
2	2	0
2	2	0
2	2	0
2	3	1

Our model correctly classified six of the ten data points in dataset 1 and seven of the ten data points in dataset 2. This puts our models success rate at well above chance (i.e. a 20% success rate). We were surprised at how well this preliminary model worked, particularly given the naivete of our feature selection and the small size or our training set.

V. Next Steps

The most pressing improvement we are looking to make at this point is data acquisition. As noted in the data section, we plan on acquiring data from the website HotOrNot, which simply contains pictures along with an aggregate rating given from users. More data should increase the accuracy with which we can predict the attractiveness of any face in general. However, we are also looking to not only generalize our training data to more models, but also to be able to cluster faces based on differentiating features such as sex, ethnicity, and age. It is conceivable that certain feature distributions will be more attractive in certain demographics; we would like to test this hypothesis, and see whether or not there is a universally appealing facial structure, or whether different groups have different optimally attractive facial structures.

The problem with acquiring more data comes in the analysis of the features. For the first few samples, we only analyzed 30 faces. We marked all of these by hand, and it was not a problem. If we are looking to expand our dataset to hundreds or thousands of faces, it will be beneficial to be able to analyze these faces automatically. More research is needed to see whether this task is feasible given our time constraints.

While automation will help save human time, we are considering further optimizations of our algorithms. Currently, we store a list of 2D points representing the facial structure for each photo. The features we run through the learning algorithm are every unique distance between points in this set. With 19 points for each face, that amounts to

171 features. This might not be a huge bottleneck, but it is reasonable to assume many of these features are nearly irrelevant, and somewhat redundant. It may make sense to do more analysis on the predictions made by our code and determine whether there is a smaller set of the most important features. For instance, many similar projects found online use on the order of 20.