Predicting Beauty

CS221: Autumn 2015

Malina Jiang
malinaj@stanford.edu

Diana Le
dianale@stanford.edu

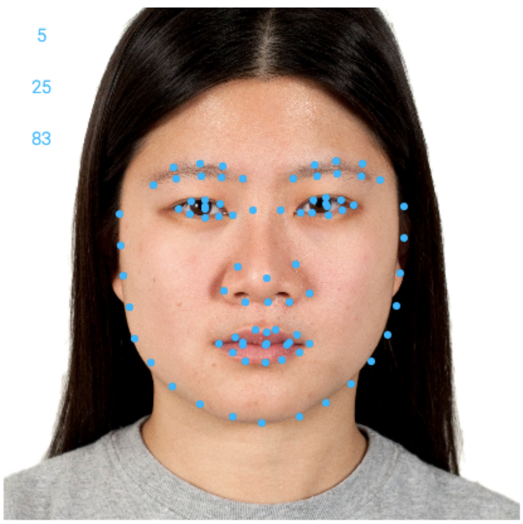
Evan Lin
elin13@stanford.edu

I. Introduction and Literature Review

Though it has often been suggested that "Beauty is in the eye of the beholder," recent studies on the nature of beauty have shown that there is a high amount of cross cultural agreement in ratings of attractiveness, suggesting that there are some objective criteria for determining facial beauty [1]. However, there is some dispute as to these exact criteria. Rubenstein and Langlois have shown that mathematically averaged faces tend to be considered attractive, while studies by Little et al have shown that extreme sexually dysmorphic features are considered ideal [2][3]. The goal of this project is to create a predictor of facial attractiveness based on facial distances and ratios such as the length of the mouth to the width of the face, placement of the eyes in comparison to face length, etc. Using tools such as linear regression, stochastic gradient descent, and k-means clustering, we hope to find the features that will most likely predict the attractiveness of a face.

II. Data

A dataset of 597 faces, all varying in age, race, and gender, was used as training and testing data for the linear regression algorithm. All of the faces had a rating on the Likert scale, a numbering scale from 1 to 7, and each had about 10-15 ratings from independent judges. To prevent overfitting of the training data and to also facilitate the development of the optimal feature vector, the data was separated into three sets (following the example shown in lecture): training, validation (or development), and testing. The training data would be used to train the weight vector while the development set would be used to minimize the error on the feature vector and to tweak the features. 199 of the faces were used for training, another 199 were used for development, and the last 199 were used for testing. Cross-validation could not be used because repartitioning the data set into the different sub-samples would have required several calls to the Face++ API, which was often unreliable.



To obtain the features, a face recognition API from Face++ was used to generate 83 landmarks on each face, as shown by the figure. These landmarks were then used to generate facial ratios, which would serve as the features for each face's feature vector. Examples of ratios included distance between the eyes versus the width of the face and length of the nose bridge versus the length of the face. A list of 30 possible features was generated; their descriptions are noted in the appendix.

III. Methodology

Using stochastic gradient descent, we trained our predictor with 199 faces, changing the weight vector with each face. With 500 iterations over the entire training set and a step size of 0.00001, the weight vector was then used to make predictions on the 199 faces of the development set. Though it is generally recommended that the step size be decreasing with the number of iterations, using a decreasing step size did not allow the weight vector to converge; the earlier iterations where the step size is 1 or slightly less than 1 likely caused the algorithm to "jump" around the function too much and would not allow for a convergence onto a local minimum. The conservative step size made the algorithm run much more slowly, but it allowed for a convergence, and with 199 training vectors, the added time was not considered a hindrance.

We then applied the weight vector to our development data and calculated the absolute error for each face. Any predictions that were within a 10% error margin, or about 0.7, of the actual ratings were considered to have been accurate, and the percentage of accurate predictions was then calculated, serving as a metric for how well the predictor performed. Using the error percentages from the development set, we toggled the features to obtain the optimal feature vector for the data set. The final feature vector was then used against the test data set to calculate the final error.

Additional analysis on standards of beauty was also done through k-means clustering. By clustering based on similar features and finding the average ratings of each cluster, we were able to find out the general ratios and features that were considered attractive.

IV. Results

4.1 Linear Regression

Our first attempt at creating the optimal feature vector involved 30 different facial distances and ratios, inspired by Cunningham et al. [4]. After testing features on the development set, a final list of 24 features (listed in the appendix) was selected to predict attractiveness, and our accuracy on the development set was 71.4%. Our final accuracy on the test set was 67.3%, which is just shy of our goal of 70% accuracy.

The final features were surprisingly made up of a large number of absolute distances, as opposed to ratios. However, the features that did contribute the most to a face's attractiveness tended to be ratios, such as the distance between the eyes compared to the width of the face at the eyes. This suggests that ratios are significant in determining attractiveness, but there are only a handful that matter in terms of contributing to the overall rating. It may also be the case that our dataset can reliably use absolute distances since the framing of each photo is highly consistent, so comparing the absolute face length of one photo to another may be somewhat accurate in determining the attractiveness of both subjects.

If you are curious about testing out an image for yourself, you may run the rate_me.py file with the link to an image of a face. For best results, you will need a JPG file placed on ImageShack. Also note that the Face++ API server is unreliable and may not be able to identify the landmarks for the photo.

4.2 K-Means Clustering

All 597 faces were clustered into 10 clusters based on the similarity of their features. The average attractiveness and the averages of each of the features were then calculated. Using this data, we were able to glean some surprising conclusions.

Clusters were not clearly segmented upon race, and a diverse set of features could be found in every ethnicity. Though two of the clusters were predominantly made of black females and males, the other seven clusters were a mixture of Latino, Asian, and white males and females. These clusters can be categorized into types of faces, where each cluster may hold general characteristics such as a smaller eye height or width or a longer face length. These types of faces also have a rating associated with them, and features from the most attractive cluster inspired some of the features in the final feature vector in the linear regression portion of the project such as lip thickness and eyebrow thickness.

The most attractive cluster, with an average rating of 3.69, consisted mostly of black females, contrary to popular surveys that suggest that white females and males are considered the most attractive (likely due to the dominance of Eurocentric ideals of beauty) [5]. The faces within this cluster tended to have longer face lengths, larger eyes, and larger distances between the eyebrow and the eye. The average ratios found within this cluster also did not emulate the Golden Ratio, but for future research, the average ratios found from clustering could be incorporated into the feature vector.

V. Future Work

Improvements to the project include using a dataset that involves faces of several different sizes and using a dataset that has several more data points. The former guards against the possibility that the absolute distance features in the feature vector will give people with larger faces higher scores; it would force the weight vector to weight ratios more heavily, and the predictor would likely give more accurate ratings for faces that did not come directly from the dataset. The latter would make a more accurate feature vector with more faces to train on. In particular, many more faces at the extremes of the attractiveness scale should be included in the dataset as our predictor is still likely to give an average rating of 2-4, despite being given a highly attractive or highly unattractive face.

To account for differences in standards of beauty for males and females, two different weight vectors should likely be used to predict attractiveness for each sex. Sexually dysmorphic features are often considered the most attractive to the opposite sex, so certain features may contribute significantly to the attractiveness of one sex, but not the other (e.g., larger eyes are considered attractive on a female but not on a male). Though we were still able to achieve 67% accuracy with a single weight vector for both sexes, we would likely be able to achieve an even higher accuracy with two separate weight vectors.

Additional work can be done to better integrate the k-means clustering and linear regression portions of our project. Through the clustering, we were able to find rough estimates of ratios that were considered more attractive than others. Adding in boolean features into our vector to account for the existence of these ratios could lead to a more accurate predictor.

Lastly, the notion of averages as mentioned by the study by Rubenstein et al. could also be incorporated. Because a "mathematically averaged" face is considered the most attractive, future projects could use the training data to create a composite face, and additional boolean features in the feature vector would include whether or not a face has the same features as the composite.

VI. Appendix

1. Length of the face
2. Width of the face at the cheekbones
3. Distance from the cheek to the middle of the mouth
4. Eye height
5. Eye width
6. Width of the face at the eyes
7. Distance between the eyes at the outer corners
8. Distance between the eyes at the inner corners
9. Distance between the eyes at the inner corners vs. width of the face at the eyes
10. Length of the nose
11. Length of the nose vs. length of the face
12. Width of the nose at the tip
13. Width of the nose at the widest point
14. Length of the chin
15. Width of the chin
16. Distance between the eyes at the idle of the pupils
17. Height between the eyebrow and the pupil of each eye
18. Height between the eyebrow and the pupil of each eye vs. length of the face
19. Upper lip thickness
20. Lower lip thickness

21. Lower lip thickness vs. length of the face
22. Width of lips at widest point
23. Eyebrow thickness
24. Nose width

VII. References

[1] V. Coetzee, J. Greeff, I. Stephen and D. Perrett, 'Cross-Cultural Agreement in Facial Attractiveness Preferences: The Role of Ethnicity and Gender', PLoS ONE, vol. 9, no. 7, p. e99629, 2014.

[2] Rubenstein, A.J., Langlois, J.H & Roggman, L.A. (2002) What makes a face attractive and why: The role of averageness in defining facial beauty. In Rhodes, G. & Zebrowitz, L.A. (eds.), Advances in Visual Cognition, Vol. 1: Facial Attractiveness, pp. 1-33. Westport, CT: Ablex.

[3] Little, A.C., Penton-Voak, I.S., Burt, D.M. & Perrett, D.I. (2002) Evolution and individual differences in the perception of attractiveness: How cyclic hormonal changes and self-perceived attractiveness influence female preferences for male faces. In Rhodes, G. & Zebrowitz, L.A. (eds.), Advances in Visual Cognition, Vol. 1: Facial Attractiveness, pp. 59-90. Westport, CT: Ablex.

[4] M. Cunningham, A. Roberts, A. Barbee, P. Druen and et al, '"Their ideas of beauty are, on the whole, the same as ours": Consistency and variability in the cross-cultural perception of female physical attractiveness.', Journal of Personality and Social Psychology, vol. 68, no. 2, pp. 261-279, 1995.

[5] M. Lewis, 'Who is the fairest of them all? Race, attractiveness and skin color sexual dimorphism', Personality and Individual Differences, vol. 50, no. 2, pp. 159-162, 2011.