CS280A Project:
Automatic Prediction of Human Attractiveness

Ryan White, Ashley Eden, Michael Maire

February 17, 2004

Abstract

We apply computer vision methods to the task of automatically predicting human attractiveness from frontal face images. A dataset of thousands of images and corresponding scores was obtained from a popular website that asks viewers to rate the attractiveness of the people appearing in the images. Using a combination of radial basis functions and specialized feature detectors, we achieved moderate success in estimating female attractiveness. However, male attractiveness proved more difficult to predict. We believe significant improvements are possible through future development of new feature types.

1 Introduction

In this project, we create an automatic predictor of facial attractiveness. Our methods involve linear regression on features extracted from the data set. We approximated factors relating to attractiveness using readily available vision techniques. These factors are based on research in evolutionary and social psychology.

1.1 Psychology of Human Attractiveness

Human facial attractiveness has long been a topic of study in psychology. Although no absolute metrics have been cited, there has been empirical analysis as to the importance of certain attributes.

Evolutionary psychology points to three major characteristics that contribute to attractiveness cross-culturally for both genders: symmetry, averageness, and nonaverage sexual dimorphic features. All three reflect a healthy development, and thus a high probability of desirable genetic information. Bilateral symmetry, given by the similarity between the left and right sides of the face, may be a direct effect of an individual's ability to defend against parasites and environmental change during development, and is thus a good indicator of "good genes". (In particular, heterozygosity [3].) There is some debate as to whether symmetry by itself is an indicator of attractiveness, or whether its existence highly correlates with the existence of other attractive features. In a test where subjects were asked to rate images of only half a face, thus removing all symmetry cues, there was still a relationship between the ratings and the faces which, when viewed in whole, were more symmetric.

Averageness is another characteristic which is difficult to dissociate from other features such as symmetry. Directly, averageness may again indicate heterozygosity and parasite resistance. Studies show that computer-generated composite faces are usually rated higher than any of the individual faces. The composite, however, blurs features, resulting in smooth skin and the disappearance of features which may lead to asymmetry. In addition, moving some features of the average face away from the average may result in a more attractive face, especially if those deviations imply desirable hormone markers [10].

The features relating to hormone markers, i.e. nonaverage sexual dimorphic features, are different for each gender. High levels of testosterone in males and estrogen in females are toxic, and thus the more exaggerated the features affected by the hormones, the more healthy the immune system of the individual. (High levels of estrogen also signify fertility in women.) Examples of masculine features

in men are: large eyebrow ridges, large jaw and lower face, and cheekbones. Feminine features are: prominent cheekbones, smooth skin texture, high eyebrows, thick lips, and a childlike jaw.

Interestingly, while extremely feminine faces are considered attractive, women vary in their preference of masculine male faces. Many studies seem to show that women prefer more masculine faces when they are most likely to get pregnant, but otherwise would prefer a less masculine mate. This implies that they desire offspring with desirable genes, but since testosterone is often linked with antisociality and a decreased likelihood of actually helping to take care of the child, they would prefer a more feminine partner [6].

Other attractive features in women may include: narrower facial shape, less fat, narrower nose, thinner eyelids, and a slightly larger distance between eyes. Additional attractive male features are: narrower facial shape, less fat, broader top of face, no wrinkles between the nose and the corner of the lips, and fuller lips. It hasn't been fully studied, however, how these features relate to genetic desirability, and how much they're culturally influenced [2]. There are also some studies relating the placement of facial features on the face to attractiveness which show that low or medium height of features is more desirable than high placement. Plus there are features with more sociological implications, such as gaze [4].

1.2 The Dataset

Our dataset consists of images and scores extracted from the website hotornot.com [5]. Using a webcrawler to download images from the website, we collected over 30000 images split roughly evenly between the two genders. Figure 1 shows a snapshot of the hotornot website. We extracted the following items: the image of the subject being rated, the average score for the image, and the number of votes. The website allows users who post images to self categorize based on two criteria. Because most users (and thus most of the images) are young, we limited our work to the male and female 18 to 25 year old categories. While we were able to download 30,000 images over the course of several days, the website claims to have had 7 billion votes and 9.1 million photos submitted.

We pruned this dataset based on a number of metrics:

- Quality of face detector response. The faces were found using the Mikolajczyk-Schmid face finder [7]. By selecting only images where the face detector responded highly, we can be reasonably sure that the face had been found accurately. Scores of 10 or greater were used.
- Size of face in image. Faces at least 100 pixels (and less than 400 pixels) on a side were selected so that the future steps would have reasonably high resolution images to process.
- High rectification score. Rectification code based on [1] and [8] found important facial features to rectify all of the faces to a common view using an affine transform. Only reliable rectifications (with scores above 2.5) were used. Again, this is mostly to guarantee that subsequent processing will have reliable images with which to work.
- More than 50 votes. According to the hotornot website, "Scores do not tend to change much after as few as 30 to 50 votes."

From these pruning techniques, we reduced our dataset to approximately 4000 images, divided almost evenly by gender. Most of the experiments presented in this paper are drawn from the rectified face images, as opposed to the original images. The rectified images were all downsampled to 86 pixels by 86 pixels with three color channels.

1.2.1 Understanding the Dataset

Before attacking the problem of creating an automatic method to rate attractiveness from images, we try to gain an understanding of the nature of the dataset itself. Figure 2 shows 16 randomly chosen rectified faces from the male 18-25 category.

To complicate matters slightly, the dataset contains a huge gender bias. The scores associated with male subjects tend to be much higher, with a lower variance: male subjects received an average score of

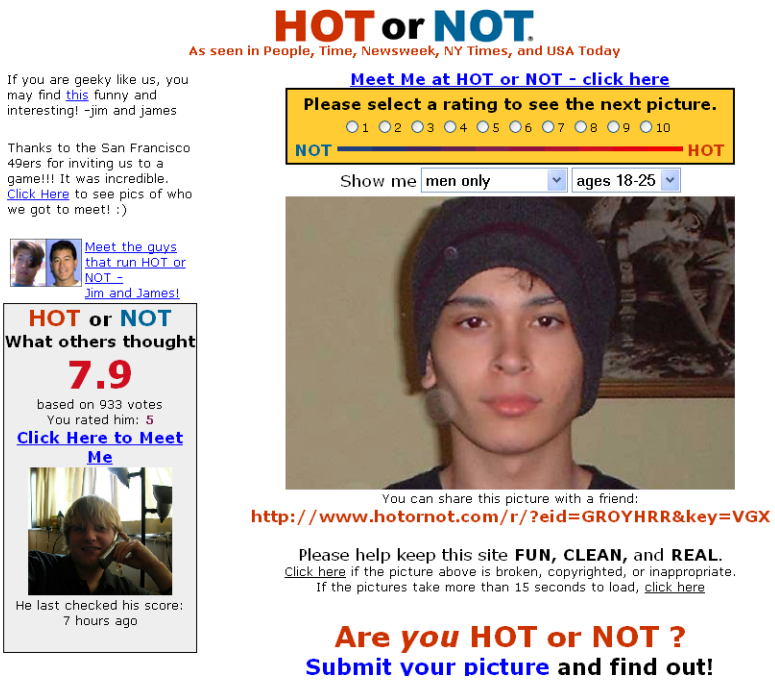

Figure 1: An example page from the website www.hotornot.com. Note that this includes several key pieces of information for building a dataset: an image of the person, their average score and the number of votes that they received.

8.4 with a variance of 0.9 while female subjects received an average score of 7.4 with a variance of 2.2. A histogram of the score by gender is presented in Figure 3. In addition to the bias in the actual scores, the number of votes received by subjects in the two categories differs significantly. Female subjects received 1250 votes on average while male subjects received 420 votes on average.

To illustrate that rectification worked well, we combined all of the images in the rectified dataset (separately for each gender). The pixel by pixel averages are shown in Figure 4. To characterize the dataset even further, we sorted the images by score, and did a pixel by pixel average for the highest and lowest scores. Results are shown in Figure 5.

1.2.2 Validating Assumptions

There is some concern that face finding, rectifying and downsampling could remove the important information contained in the dataset. In order to gauge this, we generated random pages of five images with two possible scorings: the real scores or scores randomly chosen from the dataset. Our local expert was able to correctly label 27 of 28 such pages (taken from the female 18-25 group).

2 Statistical Techniques

Since the goal of this paper is to build a program that can predict human attractiveness scores, it is important to define a reasonable error metric. Our metric is the Mean Squared Error Ratio (MSER):

$$\begin{aligned} MSER &= \frac{\frac{1}{N}\sum_{i=1}^{N}(P_i - S_i)^2}{\frac{1}{N}\sum_{i=1}^{N}(\overline{S} - S_i)^2} \\ &= \frac{\sum_{i=1}^{N}(P_i - S_i)^2}{\sum_{i=1}^{N}(\overline{S} - S_i)^2} \end{aligned}$$

Figure 2: A sample of 16 typical faces for males in the age range 18 to 25. These images have been rectified as described in the text.

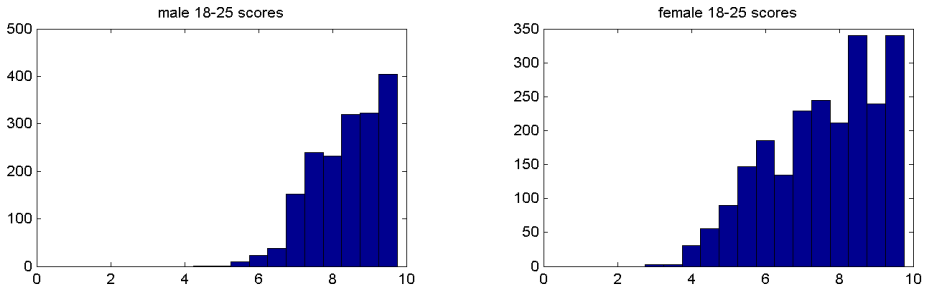

Figure 3: A histogram of the scores received in the rectified dataset. Note that the distribution differs by gender.

Figure 4: The average male (left) and female (right) faces. These are pixel by pixel averages on the rectified images.

Figure 5: The average attractive (right) and unattractive (left) male (bottom) and female (top) faces. The images on the right reflect the average taken over the 40 images with the lowest attractiveness score and the images on the right reflect the average taken over the 40 images with the highest attractiveness score.

where S_i is the score of image i, P_i is the predicted score for image i and $\overline{S}$ is the average score over all images. In this metric, a 1 indicates that the prediction algorithm does nothing and a 0 indicates perfect prediction.

2.1 Predicting the Score

The basic prediction algorithm for this paper will be a linear combination of the features. This can expressed as follows:

$$P_i = \alpha + \sum_{f=1}^{M} \beta_f x_{f,i}$$

where f indexes features in the array of features $x_{f,i}$. Note that the equation is linear in the features and uses the same coefficients β_f and α for all images. Given example images, the optimal coefficients can be found using least squares:

$$\begin{aligned} \alpha &= \frac{1}{N} \sum_{i=1}^{N} S_i \\ \beta &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \end{aligned}$$

Here the $x_{f,i}$ have been merged into the matrix $\mathbf{X}$, the β_f into β

Least squares, however, leads to overfitting. To prevent overfitting, we can do a penalized least squares that penalizes the square of the magnitude of regression coefficients (aka Ridge Regression). This is useful, because large coefficients tend to occur when the regression is magnifying noise. The solution for α remains the same, but for β:

$$\beta = (\mathbf{X}^\top \mathbf{X} + \lambda \mathcal{I})^{-1} \mathbf{X}^\top$$

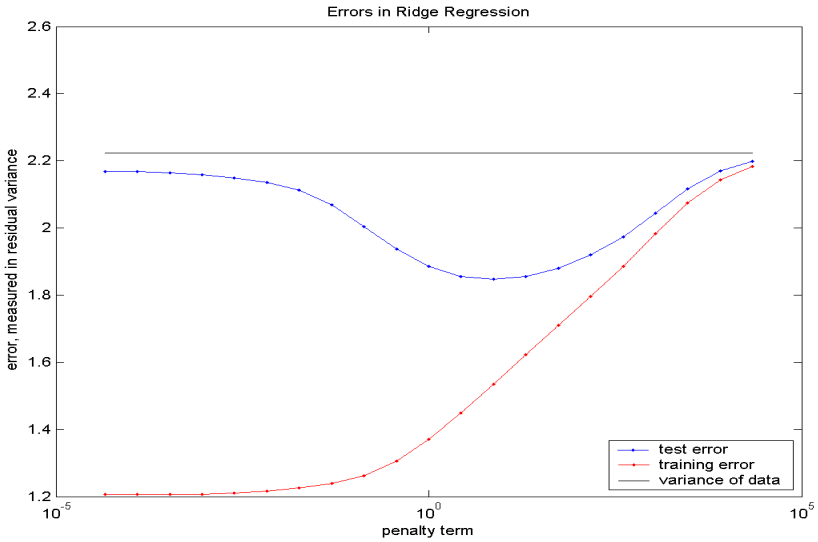

Figure 6: An example of ridge regression performance on some of the better features in the female dataset. Even at the optimal value of the penalty term, there is a large gap between the minimum test error and training error, indicating overfitting.

Figure 6 shows the results of ridge regression on a sample set of parameters taken from the female dataset. The overfitting is still non-zero, so we tried another technique to minimize the overtraining called the lasso. Instead, this technique penalizes the sum of the absolute values of the regression coefficients. The details won't be discussed here, but can be found in [11]. The technique reduces to an iterated quadratic program. Experimentally, this technique was found to be inferior to ridge regression (see Figure 7), but had the significant upshot that it produced a large number of 0 coefficients.

By using the lasso to pick important coefficients and then ridge regression to pick the actual values of the coefficients, the overfitting was reduced while maintaining a roughly constant error. The resulting ridge regression is shown in Figure 8.

3 Fishing for Human Attractiveness

3.1 Simple Cues

Our first test of the dataset was to check how various characteristics of imaging and voting affected the scores. Figure 9 shows some ridge regressions on these simple cues. When the number of features was incredibly small, quadratic terms were introduced. The numbers displayed in the table are the Mean Squared Error Ratio (MSER) on the test dataset. Numbers for the training set were typically a few percent better.

3.2 Kernelizing

One common method for finding features in a high dimensional dataset is to use radial basis functions to represent how the data varies with respect to other elements of the dataset. The basic expression for a radial basis function is:

$$K_{i,j} = e^{-\frac{\|\mathcal{I}_i - \mathcal{I}_j\|}{\sigma^2}}$$

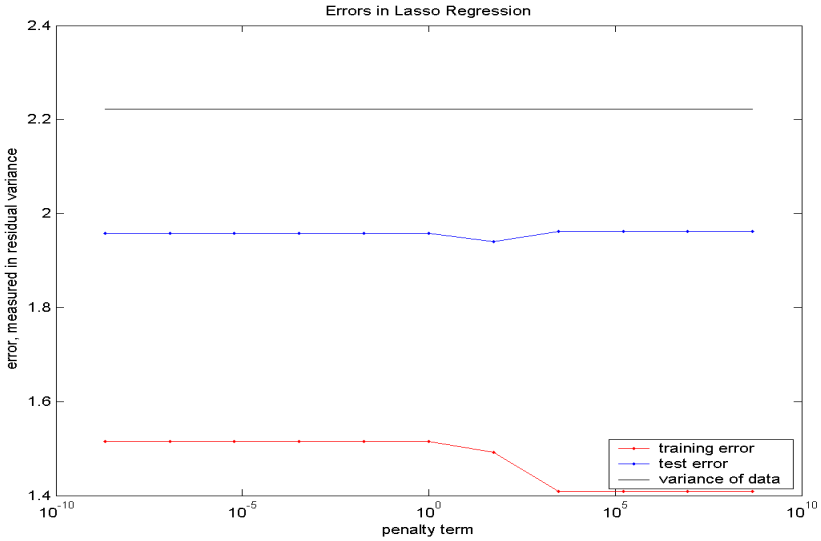

Figure 7: The lasso regression also shows problems with overtraining.

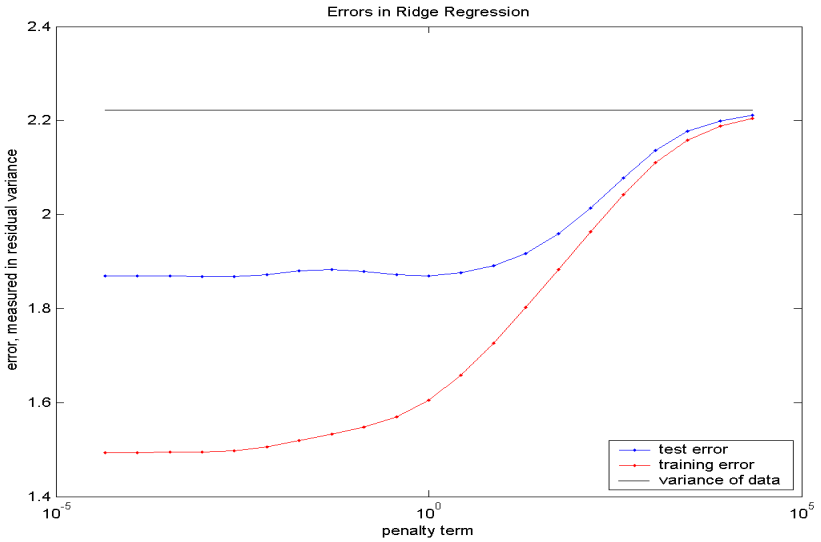

Figure 8: Ridge regression results when using only features that were important in the lasso regression. The number of features was reduced from 257 to 128 with little increase in test error. The regression is much less sensitive to the value of the penalty parameter.

Feature	*Male MSER*	*Female MSER*
number of votes	0.965	0.98
face size	0.972	0.97
rectification score	1.00	0.99
image size	0.988	0.98

Figure 9: Ridge regression results for several imaging and voting cues.

Kernel Feature	*Male MSER*	*Female MSER*
Average Faces	0.995	0.93
60 Random Faces	0.969	0.93
200 KPCA coeffs	0.953	0.86
Upper Cheek	0.997	0.94
Lower Cheek	0.996	0.98
Left Eye	0.989	0.97
Mouth	0.993	0.94
Nose	0.995	0.97

Figure 10: Some kernelizing results. The last five (Upper Cheek, Lower Cheek, Left Eye, Mouth and Nose) were created by selecting 50 images from the dataset at random and then kernelizing the rest of the dataset with respect to these images. All values computer on a test dataset.

A common technique is to kernelize a subset of the dataset, and then use principle components to pick the highest variance vectors. When kernelizing faces, we mask the face to exclude regions in the image that are likely to be distractors.

In addition to kernelizing the whole face, we kernelized separate regions of the face. In these experiments, a small region of each image was selected based on a reasonable guess of importance. Results in Figure 10 show that different regions of the images show varying ability to predict the attractiveness score. We thought that searching a local window around the region for the optimal match would improve results, since the faces were not rectified perfectly (the center of the eye varies by several pixels between images). Experimental work showed that this did not improve prediction.

In addition to kernelizing against images that were in the dataset, we also kernelized against the three average images presented in previous sections: the average of all faces, the average of the most attractive faces and the average of the least attractive faces.

3.3 Gender

In order to check whether the degree of femininity and masculinity correlated with attractiveness, we decided to gender classify the faces and give each a score based on the distance from the gender border. Using a radial basis function network to first kernelize the data, we implemented the [9] support vector machine classification method. Then we performed linear regression on the gender score for each image.

3.4 Skin Detection

In comparing the attractiveness scores of the images to our personal opinions, we noticed that sometimes, especially for the females, there was a direct correlation between the score and the amount of skin they were showing. We decided a skin detector would thus be beneficial to our predictor. Specifically, we wanted to detect the number of skin pixels in the original image (not just the rectified face).

We first tried a histogram based approach, which discretized the image into 32 bins, and, for each pixel in the image, indexed into 2 histograms with the bin placement of the R, G, and B for that pixel. The histograms stored how many pixels in different ranges of the colorspace were skin pixels and non-skin pixels respectively. If the ratio of the value given by the skin histogram to the non-skin histogram was greater than some constant threshold, it would be marked as skin.

In practice, this method appeared to be greatly affected by background color and lighting, so we decided to try a different, threshold based approach using both the rectified face and original images. First, we converted all images from RGB to LAB, since we thought that comparing pixel values over a uniform colorspace would be more beneficial.

In the rectified image, we were able to define a constant patch location that was universally the cheek (and thus contained only skin pixels.) We then found the mean of the pixels in the patch (separately for L, a, and b) and created a distance matrix containing the absolute value of the difference between each pixel value and the mean (again, separately for L, a, and b.) We did this so that we could tweak

Figure 11: 16 original images, pre-face finding and rectification.

the threshold differently depending on how different the skin pixels were in the patch, hoping that if the patch itself had a lot of variation, that meant that the image itself had a lot of variation, and we could be more tolerant. More specifically, we added the L, a, and b components of the distance matrix for each pixel, and set the threshold to be some constant times the maximum sum. We created a similar distance matrix for each pixel in the original image, and those pixels whose L, a, and b distances summed to less than the threshold were considered skin pixels.

This method also proved to be not as robust as we would like, but it seemed to be proportionally accurate across images. Figure 11 shows 16 original images, and Figure 12 shows the detected skin in those images using the threshold method.

3.5 Eye Color

In order to determine whether eye color was correlated with attractiveness scores, we designed a mechanism for extracting a model of the eye color from the rectified face images. Although the rectification procedure transforms each image to place the eyes in approximately the same location, it was necessary to further localize the position of the pupil. Rectification assisted this process only by restricting the general region in which the eyes could be located. Determining the exact pupil location within the eye region consisted of a two step process. First, we computed the probability that each pixel in the eye region was a skin pixel. Second, we searched for an assembly of pixels in a circular shape that were different from the skin color.

Skin color was modeled by extracting patches from the cheeks of the rectified face as shown in Figure 13(e-f). The algorithm converted these patches from RGB colorspace to the uniform Lab colorspace. Skin color was defined by the mean and covariance of the pixel values in Lab colorspace. Using this Gaussian model for the skin, the probability density of matching the skin color was computed for each pixel in the eye region. Taking the log of the result and convolving it with a mask for the eye gives, for each pixel, an estimate of the log probability that an eye region centered at that pixel is actually skin. The locations within the left and right eye regions that minimize this distribution are the maximum likelihood estimates of the pupil positions. Figure 13(b) illustrates a typical result of the pupil localization process. Given the pupil position estimates, eye color was modeled by a Gaussian distribution with mean and covariance determined by the pixels in an annulus around each pupil. Figure 13(d) shows the mean eye color (transformed back into RGB colorspace) extracted from the original face image in

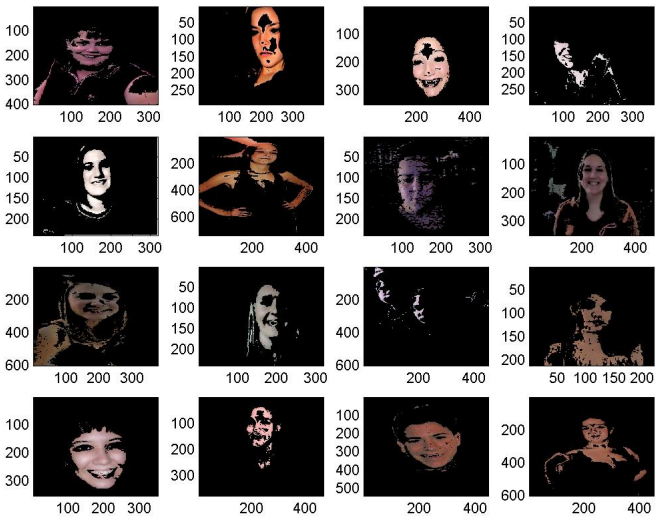

Figure 12: Detected skin in the 16 original images in Figure 11.

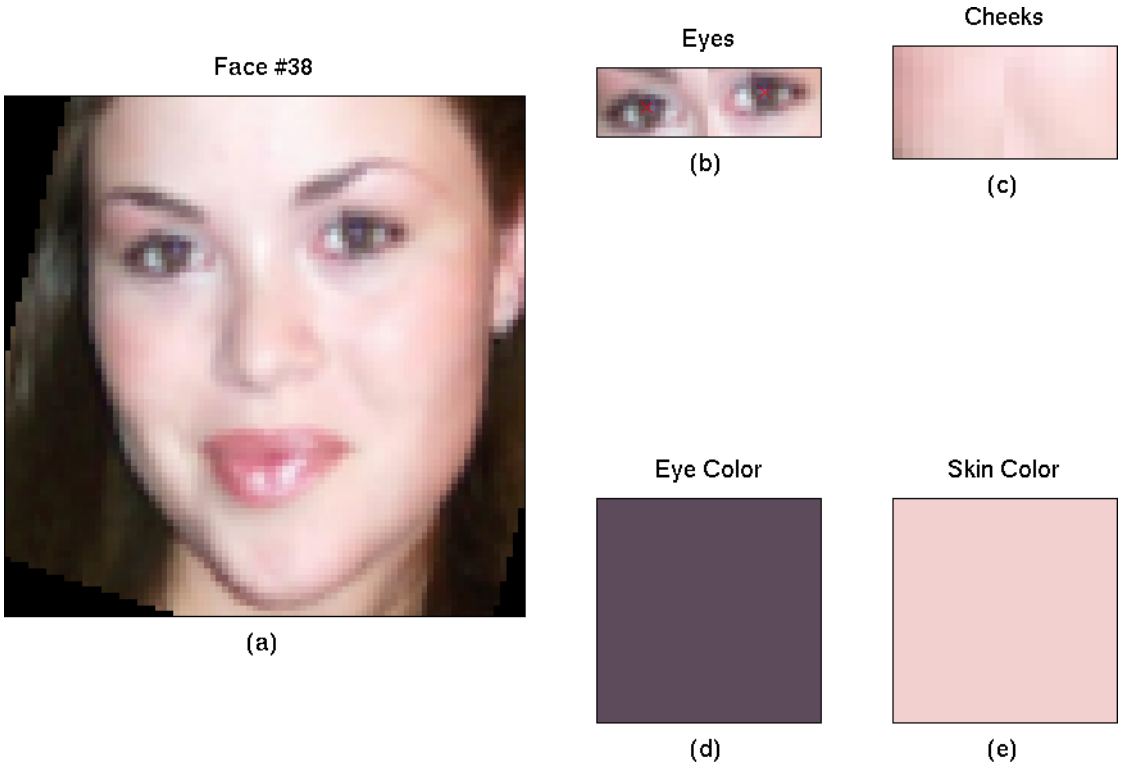

Figure 13: (a) Rectified face image. (b) Localization of pupils (marked with red x's) within the eye region using our eye model. (c) Skin sample from cheeks used for filtering out skin pixels in the eye region. (d) Eye color derived by applying a part mask around the pupil locations. (e) Average skin color derived for the skin model.

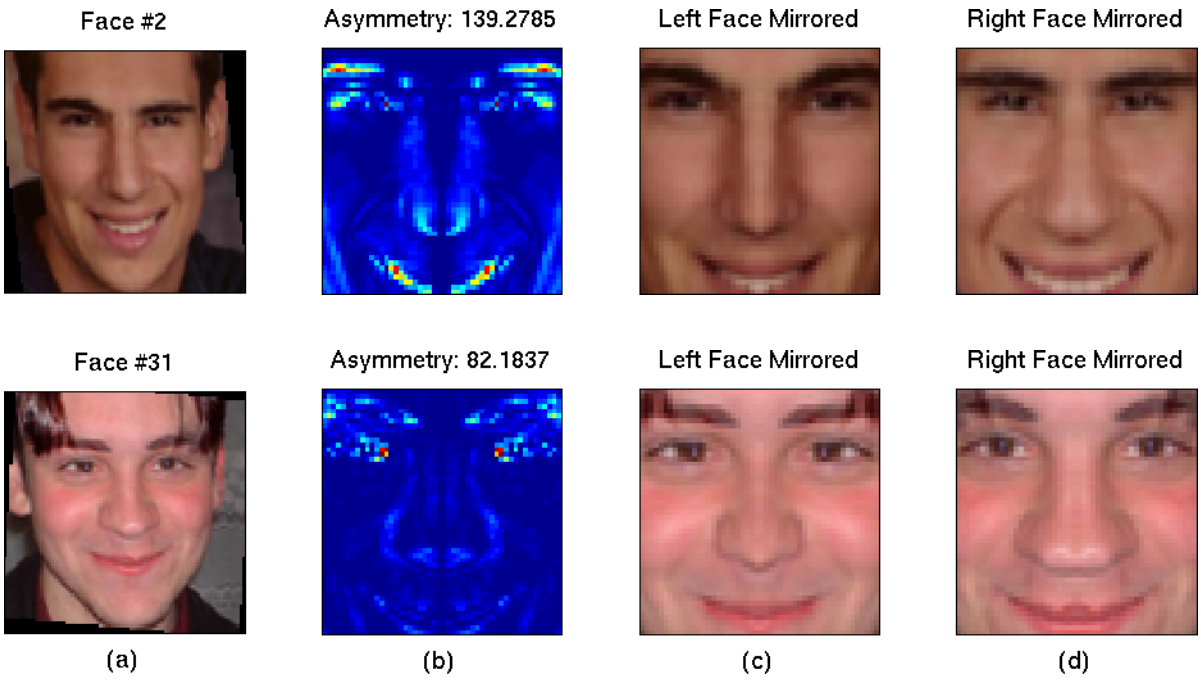

Figure 14: (a) Rectified face image. (b) Squared difference of color at each location from that at its mirror location. Color differences are computed in the uniform Lab colorspace. The total asymmetry score is the sum of the squared differences over the entire face. (c) Reconstruction of face obtained by mirroring the left side onto the right. (d) Reconstruction of face obtained by mirroring the right side onto the left.

Figure 13(a).

Unfortunately, including the mean and covariance for each face's eye color as parameters in the regression procedure used to predict scores offered no significant benefit. This could indicate that eye color is not correlated with attractiveness or that the eye regions in much of the dataset were not of high enough resolution to influence the rankings.

3.6 Gaze

We wanted to test the gaze of the person, so we took the rectified images, converted them to grayscale so that color information wouldn't cloud gaze information, and found a constant patch location which universally cut out the eyes. Then, instead of using the rectified image, we performed kernel PCA on the patch.

Because the rectification wasn't perfect, and the center of the eye wasn't in the same place for all images, this method encoded more than just the gaze. We decided that was fine, however, since we were trying to find all relevant features anyway.

3.7 Facial Symmetry

As previously discussed, research on human psychology suggests symmetry may be an indicator of attractiveness. We compute a simple symmetry statistic for each rectified face by measuring the mean color difference between corresponding pixels on the left and right sides of the face as shown in Figure 14. Symmetry was computed using a central region of the rectified image that consisted entirely of face pixels. Supplying the regression procedure with the mean and covariance of the symmetry for each face did little to improve prediction of attractiveness. It is possible that a feature-based approach that uses ratios of distances between various facial features on corresponding sides of the face would be more successful.

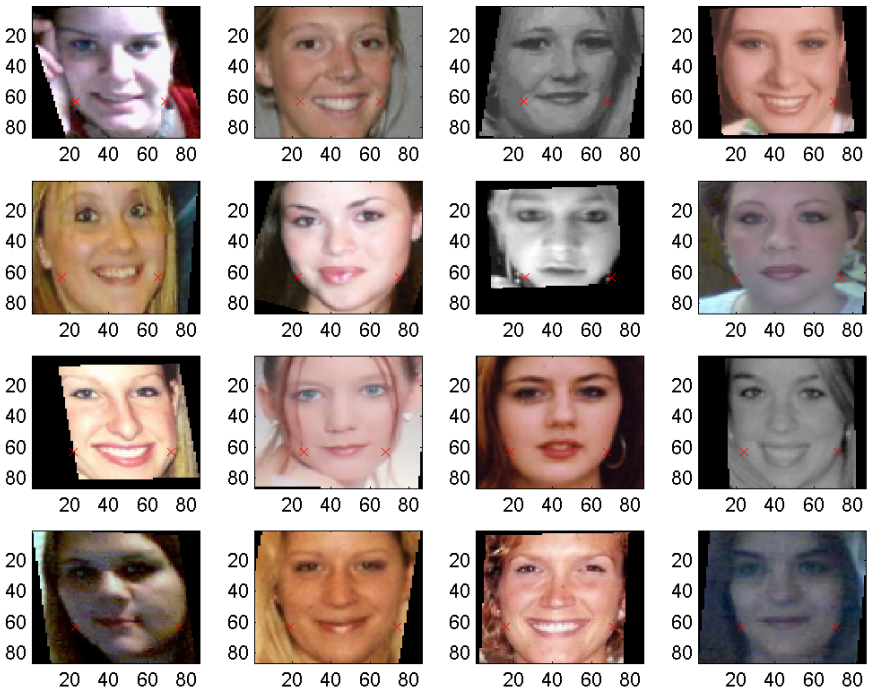

Figure 15: Detection of the sides of the face at the same vertical level as the mouth. The face boundaries across from the mouth are marked by red x's. We believe measuring distances between facial features (such as the length across the face) may allow better prediction of attractiveness scores in the future.

4 Future Work

As illustrated by Figure 10, kernelizing on the entire face or various face regions results in severely limited prediction accuracy. Furthermore, while our custom feature detectors were able to reliably extract eye and face color models, as well as symmetry scores, these features also offered only minimal predictive power. Part of the difficulty encountered when using the color features may be a result of large variance in lighting conditions and color quality across images in the dataset. A larger problem may be that all of our current feature types only weakly encode the characteristics by which people rate attractiveness.

Therefore, the most promising direction for future work is the development of new feature descriptors that better capture these characteristics. Since psychological research points to facial shape as a key indicator of attractiveness, explicit computation of facial dimensions may produce features more closely linked to attractiveness. In particular, the relative length and width of the face, as well as the distances between and sizes of major facial components (such as the eyes, nose, lips, mouth, forehead, cheeks, and chin) could be correlated with face scores. Initial work on determining distances between face components is shown in Figure 15 in which, we detect the width of the face across the mouth.

References

[1] Alexander Berg and Jitendra Malik. Geometric blur for template matching. In *Computer Vision and Pattern Recognition*, 2001.

[2] Christoph Braun, Martin Gruendl, Claus Marberger, and Christoph Scherber. Beauty check. http://www.uni-regensburg.de/Fakultaeten/phil_Fak_II/Psychologie/Psy_II/beautycheck/english/index.htm.

[3] Bernhard Fink and Ian Penton-Voak. Evolutionary psychology of facial attractiveness. *Current Directions in Psychological Science*, 11(5):154–158, 2002.

[4] Sybil Geldart, Daphne Maurer, and Heather Henderson. Effects of the height of the internal features of faces on adults' aesthetic ratings and 5-month-olds' looking times. *Perception*, 28:839–850, 1999.

[5] James Hong and Jim Young. http://www.hotornot.com.

[6] Anthony Little and David Perrett. Do women prefer 'manly' faces? http://www.firstscience.com/site/articles/perception.sap, 200.

[7] K. Mikolajczyk. *Detection of local features invariant to affine transformations*. PhD thesis, INPG Grenoble, 2002.

[8] Tamara Miller, Alexander Berg, Jaety Edwards, Michael Maire, Ryan White, Yee-Whye Teh, Erik Learned-Miller, and D.A. Forsyth. Faces and names in the news. Submitted, Computer Vision and Pattern Recognition, 2004.

[9] B. Moghaddam and M-H. Yang. Gender classification with support vector machines. In *Proc. of Int'l Conf. on Automatic Face and Gesture Recognition (FG'00)*, pages 306–311, Grenoble, France, March 2000.

[10] Ian Penton-Voak and David Perrett. Consistency and individual differences in facial attractiveness judgements: And evolutionary perspective. *Social Research*, 2000.

[11] R. Tibshirani. Regression shrinkage and selection via the lasso, 1994.