Personalized Facial Attractiveness Prediction

Jacob Whitehill
University of California, San Diego
La Jolla, CA 92093
jake@mplab.ucsd.edu

Javier R. Movellan
University of California, San Diego
La Jolla, CA 92093
movellan@mplab.ucsd.edu

Abstract

We present an approach to learning the personal preferences of individual users directly from example images. The target application is computer assisted search of partners in online dating services. The proposed approach is based on the use of ε-SVMs to learn a regression function that maps low level image features into attractiveness ratings. We present empirical results based on a dataset of images collected from a large online dating site. Our system achieved correlations of up to 44 % (Pearson rank correlation) on the attractiveness predictions for individual users. We show evidence that the approach learned not just on a universal sense of attraction shared by multiple users, but capitalized on the preferences of individual subjects. Our results are promising and could already be potentially used to facilitate the personalized search of partners in online dating.

1. Introduction

Every day, millions of users log on to Internet dating sites to make one of the most important decisions in their life: choosing a partner. To this end users are confronted with the difficult task of datamining the wealth of choices available in dating sites. These sites usually contain textual information, like age, hobbies, and personal interests to assist users making their choices. However critical sources of information like a person physical attractiveness are beyond current search tools. This is because many of the things that determine our attraction towards specific faces are subtle and difficult to verbalize.

Recent machine vision research [1, 2, 3, 4] has tackled the problem of predicting the attractiveness of faces averaged over a large universe of observers. While there are faces that many people find particularly attractive, in practice people differ greatly on their individual preferences. In fact some users from online dating sites may not be interested on partners that may appear attractive to too many people. In this paper, we explore the use computer vision and machine learning methods to learn the preferences of individual users directly from labeled face images. While computer vision and machine learning applications to face processing have a recent history of success, to our knowledge the problem of learning to predict individual user's attraction for other face images has never previously been approached in the literature. Thus, it is unclear whether this problem is solvable with current methods, and if so, which image representations, learning algorithms, and training set sizes are most appropriate. In this paper we present empirical results comparing the performance of a variety of low level representations such as PCA, Gabor filter banks, and Gaussian RBFs. We also investigate image representations based on higher-level features, like automated analysis of facial expressions.

2. Related Work

Relatively little machine learning research has been conducted on the specific task of attractiveness prediction, and all the existing literature has focused on prediction of *universal* attractiveness. Aarabi, et al [3] used k-nearest neighbors to classify face images as belonging to one of four ratings of beauty. The feature vectors consisted of 8 geometric ratios of distances between certain fiducial points (eyes, brows, and mouth) of the face. On a validation set of 40 images, their system achieved 91% correct classification. When fiducial points were inaccurately registered, however, the performance sank to 37%.

White, et al [2] used ridge regression, a Gaussian RBF kernel, and textural facial features to predict the mean attractiveness scores assigned to 4000 face images downloaded from hotornot.com. They experimented with several textural features and achieved their best performance, corresponding to a Pearson coefficient of 0.37, using kernel PCA on the face pixels.

Eisenthal, et al compared three alternative classification methods - SVMs, k nearest neighbors, and standard linear regression - to classify face images as either "attractive" or "unattractive." Linear regression and SVMs performed best and were reported to exhibit similar accuracy. They

found that geometric features based on pairwise distances between fiducial points were superior to textural features such as eigenface projections (Pearson coefficients of 0.6 and 0.45, respectively). They tested on two small databases, each containing 92 images, of young women from the USA and Israel posing neutral facial expressions.

Kagian, et al [1] used standard linear regression to predict mean attractiveness scores on the same female Israeli database. The human coders labeled each image with an attractiveness score in 1-7 range. By using 90 principal components of 6972 distance vectors between 84 fiducial point locations, some of which were labeled manually, they achieved a Pearson correlation of 0.82 with mean attractiveness scores labeled by humans.

From this literature several trends can be observed: First, all previous approaches to attractiveness prediction focused on learning attractiveness scores averaged across people. In contrast, here we focus on the potentially more difficult problem of learning an individual's notion of facial attractiveness. Second, all of the above described systems except [2] were tested on very small image databases of less than 100 individuals. For our application, we are interested in heterogeneous datasets consisting of at least 1000 people. Finally, there exists a performance/robustness trade-off when using geometric features: while image features based on relative positions of fiducial points achieved the highest accuracy in [4] and [1], some of the facial feature points needed to be manually adjusted. Moreover, [3] reported that accuracy can suffer considerably when fiducial points are inaccurately located. In our research, since we are interested in approaches that can work automatically with current technology, we thus concentrate more on *textural* features (e.g., eigenface projections, Gabor decompositions).

3. Our Approach

The approach investigated in this paper is to learn a personalized attractiveness function based on example face images labeled by individuals for their degree of attractiveness. The images were 36x36 pixels, and the attractiveness function was learned using ε-SVMs applied to the following image feature types (see Figure 1):

1. *Eigenface projections* [5] have a long-standing history in the field of face analysis. Typically, each face image to be analyzed is projected onto the N eigenface projections with highest associated eigenvalues, where N is optimized for the particular application.

2. *Gabor filters* have a proven track record in object detection [6] and expression recognition [7]. Their use in automated beauty detection was also suggested by [4].

3. *Edge orientation histograms* (EOH) have garnered considerable interest for object recognition [8] as well as object detection [9] in recent years. Particular forms of EOH features can capture local symmetry in face appearance; given the reported importance of symmetry in facial beauty [10], they may be particularly useful for attractiveness detection.

4. *Geometric* features: We experimented with geometry based features based on the pairwise distances and angles between fiducial points. We used fiducial points which can be reliably detected using current technology, i.e., corners of the eyes, and centers of mouth and nose.

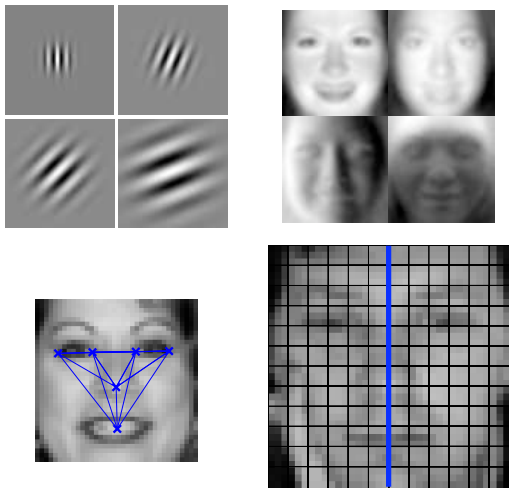

Figure 1. An illustration of the various alternative feature types we used for personalized attractiveness prediction. Top-left: four sample Gabor filters. Top-right: the four eigenface projections with highest associated variance. Bottom-left: Geometric features derived from relative distances between fiducial points. Bottom-right: grid of cells within which histogram of oriented gradient (EOH) features are extracted.

3.1. Regression

During preliminary experimentation with standard linear regression, ridge regression, multinomial logistic regression, and ε-SVM regression, we found that ε-SVMs worked best and thus used ε-SVMs in our experiments. In its linear form $y = \mathbf{w}^T\mathbf{x} + b$, the ε-SVM seeks to minimize the L_2 norm of the weight vector $\mathbf{w}$ subject to the constraints that each predicted output value y can be no more than ε distance from its true value in the training set. So long as the error between the predicted and actual output values is at most ε, the exact error is irrelevant [11, 12].

3.2. Implementation Details

All face patches were downscaled from 96x96 (for human coding) to 36x36 for automated analysis. Then, one or more sets of features were extracted, and either a regressor

was trained (training phase) or the attractiveness score of the face patch was predicted (prediction phase). When predicting attractiveness scores, the real numbers returned by the regressor were linearly scaled to integers in $\{-1, 0, 1, 2\}$ based on the minimum and maximum real values in the particular cross-validation fold.

For the extraction of EOH features, each face image was decomposed into an 12x12 grid of square cells, each 3 pixels wide, as illustrated in Figure 1. For the first EOH feature type, a histogram spanning 8 different orientation bins (22.5deg) was calculated within each cell, and the histograms over all cells were concatenated to form the feature vector. For the second EOH feature type, the same histograms were computed in all cells. Each cell was paired with its bilaterally symmetric neighbor (reflected across the middle vertical line in the figure). For each orientation bin $i \in \{1, \ldots, 8\}$ in the left cell in each pair, one feature value was computed as the difference in total gradient in bin i of the left cell and bin $8 - (i - 1)$ of the right cell.

Gabor features were extracted by filtering each face patch with 40 Gabor filters at 5 different spatial frequencies and 8 orientations, as in [13]. The Gabor decompositions for each face patch were concatenated and then downsampled by a factor of 36.

For eigenface analysis, we varied the number of highest-variance eigenface projections over $\{100, 150, 200, 250, 300, 500\}$ and found that 250 yielded the best results.

Finally, for the ε-SVM regression, we used the libsvm SVM implementation [14]. We let $\varepsilon = 0.001$ and used a Gaussian RBF kernel.

4. Experiment Design

4.1. Data Set

Face images for our experiment were taken from the *GENKI* database [15], which consists of nearly 70,000 face images collected from the Web. The persons in GENKI span a wide range of ages (though all persons are at least 18 years old), imaging conditions, and ethnicities. All images are labeled for 3-D pose (yaw, pitch, roll) and 6 fiducial point locations on the face (inner and outer eye corners, nose, and mouth). We randomly selected 1000 males and 1000 females from this database whose yaw, pitch, and roll parameters were all in $[-4, 4]$ degrees of frontal. In this sample, approximately 80% of the individuals were white, 10% were Asian, 5% were Latin, and 5% were African/black. Using the labeled fiducial point coordinates, the faces were converted to grayscale, rotated to a canonical rotation, and scaled to 96x96 pixels. Since then we have developed a facial feature finder that could have been used to automate the image registration process.

4.2. Procedure

4.2.1 Data Collection

We collected facial preference data from 8 human coders, 7 of whom are graduate students and 1 of whom is a system administrator at our institution. The ages of the coders varied from early 20's to mid 30's. Six of the coders were male, and two were female. Two of the male coders labeled faces of other males (homosexual preference). The other six coders labeled faces of the opposite sex. Seven of the coders were Caucasian, and one was a Latin male.

Preference data consisted of attractiveness ratings assigned to 1000 different faces of the user's preferred gender. The coders did not have access to any biographical information (e.g., hobbies, interests, occupation, age) of the people they were rating. Ratings consisted of integers in $\{-1, 0, 1, 2\}$ with the following interpretation:

-1	*Definitely not* interested in meeting the person
0	Not interested in meeting the person
1	Interested in meeting the person
2	*Definitely* interested in meeting the person

Each coder assigned attractiveness ratings to all 1000 images of his/her preferred gender by clicking the mouse on the corresponding face image. The default rating assigned to each face (before the coder adjusted it using the mouse) was 0.

During the coding process, each human coder was presented with 40 montages of 25 faces each (pre-cropped and normalized) in 5x5 tile format. Only one montage appeared on the screen at a given moment. The coder could cycle through the different attractiveness ratings of each face by clicking the mouse. The user could advance to the next montage by using the keyboard. No time limit was imposed, but coders typically required about 45 minutes to label 1000 faces of their preferred gender.

In order to measure the consistency of each subject relative to him/herself, we asked each coder three months later to re-label a subset of 100 randomly selected faces (from the original 1000). These second sets of labels were used only for self-consistency assessment, not for training or validation of the attractiveness regression function.

4.2.2 Training and Validation

For each user, the 1000 labeled images were divided into five disjoint cross-validation folds of 200 faces each. Each regressor was trained on four of the five folds and then tested on the remaining fold. For each person, performance was computed as the average (over all five folds) Pearson correlation of predicted attractiveness with the human-labeled ratings.

Correlations with Human Labels of Attractiveness		
Coder #	**Self-Correlation**	**ε-SVM Predictions**
Users Preferring Males		
1	0.64	0.36
2	0.57	0.19
3	0.67	0.32
4	0.49	0.28
Users Preferring Females		
5	0.38	0.08
6	0.43	0.32
7	0.64	0.38
8	0.78	0.44
Avg	*0.58*	*0.30*

Table 1. Correlation coefficients of predicted attractiveness scores for each human coder. Results shown are for PCA-based features and ε-SVMs.

5. Results

5.1. Correlation with Humans

Using eigenface features and ε-SVM regression, we achieved the correlations with human ratings shown in Table 1. The average correlation (over all 8 subjects) of predicted attractiveness scores with human ratings was 0.30. This result is based on a fully automatable system and is comparable with the average inter-subject rank correlation between attractiveness labels (Spearman correlation of 30%). The average self-consistency correlation was 0.58, which underlines the difficultly of the personalized attractiveness prediction problem. The accuracy of the automated prediction was highly correlated with the self-consistency of the subjects($r = 0.77$), i.e., it was easier to predict the preferences of those subjects that had make more consistent choices . For coders 7 and 8, the ε-SVM achieved a correlation substantially higher than inter-human correlation - our system achieved 38% and 44% correlation with ground-truth rankings, respectively. While this result can certainly be improved upon, we believe it is a valuable first step towards viable personalized attractiveness prediction.

A plausible hypothesis is that our system had just learned features that are universally liked, rather than learning the preferences of individual users. To test this hypothesis we calculated for each person the correlation between his/her preferences and the preferences predicted by the system for other people who preferred the same gender. The results indicated that the system was tapping on personal preferences. The median correlation for predictors trained on other users was 0.235, whereas the median correlation for predictors trained on the same users was 0.32, a 36% improvement, which was statistically significant (Wilcoxon Sign Test =3, $p < 0.02$, one tail).

Feature Type Analysis	
Feature Type Set	**Correlation**
{ Gabor }	0.30
{ PCA }	0.30
{ EOH }	0.26
{ Geom }	0.06
{ PCA,Gabor }	0.30
{ PCA, EOH }	0.29
{ PCA, Geom }	0.28

Table 2. Analysis of the effect of feature type on the mean Pearson correlation (over all human coders). Eigenface (PCA) features and Gabor features tied for the best performance. Adding a second feature type did not increase performance.

5.2. Feature Type Analysis

Results of the feature type analysis are shown in Table 2. Eigenface projections and Gabor decompositions achieved equal accuracy at a correlation of 0.30. Since the principal component projections can be computed more quickly than the Gabor features, we chose eigenfaces as our primary feature. Then, to test whether performance increased by adding a second feature type, we ran an additional series of tests with feature sets: { PCA, Gabor }, { PCA, EOH }, and { PCA, Geom}. As shown in the table, performance did not increase by appending the additional features.

5.3. Finding the Attractive Faces

In an online dating setting, the user is usually more interested in finding "attractive" people rather than "unattractive" people. Typically, the user can specify certain search parameters regarding geographical location, age group, and preferred gender, and the dating website then returns a set of faces matching these basic criteria for the user's review. To our knowledge, no currently available online dating service uses any form of automated attractiveness detection; hence, beyond the initial search criteria, the probability that the querying user will find the returned set of faces attractive is equivalent to a random draw. Our system can be used to improve on this "random draw" significantly by either sorting the faces in decreasing order of attractiveness, or by setting a threshold and classifying faces as either "attractive" or "non-attractive." We took the latter approach and measured the accuracy of our system using the Proportion of Faces Correctly Predicted to be Attractive (PFCPA), which is equivalent to the true positive rate when the learned regression function is given a particular threshold for binary classification (attractive versus non-attractive).

Results using the ε-SVMs on eigenface (PCA) features are shown in Table 3. As shown in the table, the PFCPA using our system was higher than the baseline (random sample, as in contemporary online dating sites) for all human

Proportion of Faces Correctly Predicted to be Attractive		
Coder #	**Predicted by ε-SVM**	**Baseline**
Users Preferring Males		
1	0.17	0.09
2	0.13	0.08
3	0.17	0.10
4	0.22	0.18
Users Preferring Females		
5	0.04	0.04
6	0.34	0.24
7	0.44	0.20
8	0.52	0.30
Avg	*0.25*	*0.15*

Table 3. Fraction of people correctly predicted to be attractive by our ε-SVM system using PCA features compared to the baseline probability of attractiveness obtained by a random sample (the status-quo of online dating sites). For several users, our system nearly doubled the proportion of attractive faces.

subjects in our experiment. For several users, the percentage of "attractive" people returned using our system was nearly double that of a random sample, The average increase in PFCPA over random selection was 0.10, which is statistically significant ($t(7) = 3.3, p < 0.05$) . These results suggest that personalized attractiveness prediction, though a nascent research topic, can already be effectively used to improve users' online dating experiences.

6. High-level Features: Facial Expression

In the research presented above, low-level image features were used to predict the attractiveness of faces for a particular user's taste. It is conceivable, however, that higher-level features such as facial expression cues may be more suitable for automatic attractiveness prediction. In this section, we investigate the usefulness of detecting the facial "action units" of the Facial Action Coding System (FACS, [16]) for predicting facial attractiveness. FACS decomposes human facial expressions into 46 component movements, which roughly correspond to individual facial muscles. These elementary movements are called action units (AUs) and can be regarded as the "phonemes" of facial expression. A particular facial expression can be described by the set of AUs it contains and their associated intensities, rated on a 5 point scale. Recent years have seen considerable progress in the development of automatic AU detectors; in this study, we use one particular such system, CERT, which achieves state-of-the-art accuracy both for a variety of AUs [7] and for the detection of social smiles [15].

Using CERT we automatically detected the intensities of AUs 1, 2, 4, 5, 9, 10, 12, 14, 15, 17, 20, and of the "smile"

Correlations of AU Intensities with Attractiveness Labels	
AU	**Correlation**
Users Preferring Males	
4	-0.09
9	-0.14
Users Preferring Females	
4	-0.17
5	$+0.11$
9	-0.18
17	-0.13

Table 4. The statistically significant correlations between Action Unit (AU) intensities and attractiveness labels for each gender. The correlations between each user's set of attractiveness labels and the AU intensities were averaged across all 4 coders for each gender.

value for all the face images in our experiment. Then, for each of the 4 human coders that labeled the attractiveness of a particular gender, we correlated the detected AU intensities with the human labels of attractiveness. To first gain a sense of any impact of AU intensity on universal perception of attractiveness, we averaged the correlations across all four coders for each gender and considered the average correlation to be statistically significant for a particular gender if $p < 0.05$ across all four coders.

Table 4 shows the AUs with which the 4 sets of attractiveness labels for each gender were significantly correlated. For both genders, there was a significant negative correlation with both AU 4, "brow lowerer," and AU 9, "nose wrinkler" (see Figure 2 for an illustration). Moreover, for female faces, there was also a negative correlation with AU 17. These AUs are all associated with negative emotion [16] and it is thus not unexpected that a negative correlation exists.

Action unit 5, with which female attractiveness labels were positively correlated, tends to expose the eyes; previous research has found a positive correlation between pupil size and sexual arousal of the observer [17], and it is possible that these two effects (eye widening and pupil dilation) are related.

To our surprise, the attractiveness labels were not significantly correlated with detected smile intensity, despite the fact that a wide range of intensity values, and substantial number of both smiles and non-smiles occurred in the dataset. However, recent empirical research has shown that the effect of smile on attractiveness perception is subtle and context dependent. For example, smiling faces are on average judged as being more attractive than neutral expression faces on images of faces with direct gaze. By contrast, for judgments of faces with averted gaze, attractiveness preferences are stronger for faces with neutral expressions than smiling faces [18]. In addition, smiles correlate with other

AU 4 (Brow lowerer)
Negative correlation

AU 5 (Upper lid raiser)
Positive correlation

AU 9 (Nose wrinkler)
Negative correlation

AU 17 (Chin raiser)
Negative correlation

Figure 2. Samples of each of the AUs with which the faces in our study were significantly correlated. Permission to reproduce images pending.

factors in interesting ways. For example, in our dataset of images from online dating sites, we found that older males tended to smile more than younger males. Moreover older males tended to be ranked as less attractive than younger males. Indeed, many of the most attractive individuals in the dataset seemed to adopt the "cool" facial expression typically seen in high-fashion models. This and other subtle effects may help explain the lack of empirical correlation between smile value and attractiveness.

6.1. Using Facial Expression for Personalized Prediction

In order to test the usefulness of AU features for personalized attractiveness prediction, we created a new feature type, "AU", and then ran the same experiment as in Section 4.2.1. Preliminary results indicate that AU features only improved marginally the accuracy of the prediction, from 0.30 for PCA alone to 0.31 for PCA + AU. We were surprised by the fact that the performance improvement was so small and are currently investigating whether this results holds when using other learning architectures.

7. Conclusion

Physical attraction plays a critical role in the choice of partners, yet current online dating services have no way to use this source of information to assist their users. This is partly due to the fact that many of the things that determine our attraction towards others are subtle and difficult to verbalize.

Here we explored an approach to learn a user's preferences based on example labeled images. Overall, our results, while preliminary, show that automated personalized prediction of facial attractiveness is feasible with current technology. Our system achieved Pearson correlations up to 44% with human ratings of attractiveness. While not perfect, the system could already be used to help users search for potential partners.

There are many obvious ways in which our results could be potentially improved in the near future. For example, collaborative filtering techniques could be used so that rankings of attractiveness can implicitly be learned from other users for which preference models already exist. Another technique likely to result in significant improvements is the use of active learning to ask the user to label only those images that may be particularly informative to learn his/her preferences. Overall, the results presented in this document are encouraging and suggest a new potential application of modern computer vision technology to assist millions of people make one of the most important decisions in their lives.

References

[1] A. Kagian, G. Dror, T. Leyvand, D. Cohen-Or, and E. Ruppin. A humanlike predictor of facial attractiveness. In *Advances in Neural Information Processing Systems*, 2007. 1, 2

[2] R. White, A. Eden, and M. Maire. Automatic prediction of human attractiveness. *UC Berkeley CS280A Project*, 2004. 1, 2

[3] P.Aarabi, D. Hughes, K. Mohajer, and M. Emami. The automatic measurement of facial beauty. In *Systems, Man and Cybernetics*, 2001. 1, 2

[4] Y. Eisenthal, G. Dror, and E. Ruppin. Learning facial attractiveness. *Unpublished manuscript*, 2004. 1, 2

[5] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 1991. 2

[6] Martin Lades, Jan C. Vorbrüggen, Joachim Buhmann, J. Lange, Christoph von der Malsburg, Rolf P. Würtz, and Wolfgang Konen. Distortion invariant object recognition in the dynamic link architecture. *IEEE Transactions on Computers*, 42:300–311, 1993. 2

[7] G. Littlewort, M. Bartlett, I. Fasel, J. Susskind, and J. Movellan. Dynamics of facial expression extracted automatically from video. *Image and Vision Computing*, 24(6):615–625, 2006. 2, 5

[8] David G. Lowe. Object recognition from local scale-invariant features. In *International Conference on Computer Vision*, 1999. 2

[9] Kobi Levi and Yair Weiss. Learning object detection from a small number of examples: The importance of good features. In *Proceedings of the 2004 IEEE Conference on Computer Vision and Pattern Recognition*, 2004. 2

[10] K. Grammer and R. Thornhill. Human (homo sapiens) facial attractiveness and sexual selection: The role of symmetry and averageness. *Journal of Comparative Psychology*, 1994. 2

[11] A.J. Smola and B. Schoelkopf. A tutorial on support vector regression. *NeuroCOLT2 Technical Report NC2-TR-1998-030*, 1998. 2

[12] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer, New York, 1995. 2

[13] Gianluca Donato, Marian Stewart Bartlett, Joseph C. Hager, Paul Ekman, and Terrence J. Sejnowski. Classifying facial actions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(10):974–989, 1999. 3

[14] Chih-Chung Chang and Chih-Jen Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`. 3

[15] Jacob Whitehill, Gwen Littlewort, Ian Fasel, Marian Bartlett, and Javier Movellan. Developing a practical smile detector. *Submitted to PAMI*, 2007. 3, 5

[16] P. Ekman and W. Friesen. *The Facial Action Coding System: A Technique For The Measurement of Facial Movement*. Consulting Psychologists Press, Inc., San Francisco, CA, 1978. 5

[17] Selina Tombs and Irwin Silverman. Pupillometry: A sexual selection approach. *Evolution and Human Behavior*, 25(4), 2004. 5

[18] C.A. Conway, B.C. Jones, L.M. DeBruine, and A.C. Little. Evidence for adaptive design in human gaze preference. *Proceedings of the Royal Society of London B*, 275(1630):63–69, 2007. 6