

Learning Facial Attractiveness

Yael Eisenthal¹ Gideon Dror² Eytan Ruppin¹

¹School of Computer Science, Tel-Aviv University, Tel-Aviv 69978, Israel

eisentha@post.tau.ac.il, ruppin@post.tau.ac.il

²Department of Computer Sciences, Academic College of Tel-Aviv-Yaffo, Tel-Aviv
64044, Israel

gideon@mta.ac.il

DRAFT (03/06/2004)

Abstract

In this work we study of the notion of “attractiveness” of faces in a machine-learning context. To this end, we collected human beauty ratings for datasets of facial images and used various techniques for learning the average attractiveness of a face. The results clearly show that beauty is a universal concept, which can be learned by a machine. Due to the limited size of the dataset, most of the information about the target is extracted from features that are simply correlated with facial beauty.

1. Introduction

The subject of human facial attractiveness has received attention from scientists for centuries, yet, the face of beauty, something we can recognize in an instant, is still difficult to formulate. This outstanding question has led to a large body of ongoing research by scientists in the biological, cognitive and exact sciences.

The common notion in research has always been that beauty is “in the eye of the beholder”, that individual attraction is not predictable beyond our knowledge of a person's particular culture, historical era or personal history. However, more recent work suggests that the constituents of beauty are neither arbitrary nor culture bound. Numerous studies have demonstrated high congruence over ethnicity, social class, age and sex [1][2], suggesting that properties of facial features are the same irrespective of the perceiver, and that people everywhere are using similar criteria in their judgements. This is further strengthened by the consistent relations demonstrated in experimental studies between attractiveness and various facial features [3], as well as by studies demonstrating that even infants and newborns show a preference for more attractive faces [4].

Research has found certain features and characteristics to be positively related to facial attractiveness (e.g. symmetry, averageness, full lips, large eyes), yet the relative importance of these traits and their interactions with other facial attractiveness determinants are still unknown.

Different studies have examined the relationship between subjective judgements of faces and their objective regularity. Morphing software has been used to create average and symmetrized faces [1], as well as attractive and unattractive prototypes, in order to analyze their characteristics. Others have produced attractive faces from a collection of golden ratios, from a fractal geometry based on powers of two or by evolution using an interactive genetic algorithm [5]. Previous work has mainly involved averaging and morphing of digital images and geometric modeling to construct attractive faces. In this work

we explore the notion of facial attractiveness using machine learning techniques: using only the images themselves, we try to learn and analyze the mapping from two dimensional facial images to their attractiveness scores, as determined by human raters.

2. Data

In order to reduce the effects of age, skin color, facial expression and other irrelevant factors, subject choice was confined to young Caucasian females. Images were constrained to frontal views with neutral expressions, with no accessories or obscuring items (e.g. jewelry). Furthermore, to get a good representation of the notion of beauty, the dataset was also required to encompass both extremes of facial beauty: very attractive as well as very unattractive faces.

We obtained two datasets, which met the above criteria, both of relatively small size of 92 images: one contained images of young American women, and the second - of Israeli girls aged approximately 18. The distributions of the two datasets were found to be too different for combining the sets, and, therefore, all our experiments were conducted on each dataset separately. Images were converted to grayscale to lower data dimension and to simplify the computational task.

Attractiveness ratings were collected for both datasets; 28 observers rated the images in the first dataset and 18 rated those in the second. Each facial image was rated on a discrete scale between 1 (very unattractive) and 7 (very attractive) by each rater. Ratings were tested for adequacy and consistency. The final attractiveness rating of a facial image used in subsequent analysis was the average of its collected ratings.

3. Work and Results

3.1 Face Representation

In our analysis we adopted two different representations. The first is the raw grayscale pixel values, in which all relevant factors, such as texture, shading, pigmentation and shape, are represented, though not in form simple to extract. The second representation, motivated by golden ratio arguments, is based on the manual measurement of 37 facial feature distances that reflect the geometry of the face. These include, for example, the distance between eyes and width and length of mouth and of each eye. The facial feature points, according to which distances are defined, are shown in Fig. 1. All raw distance measurements, which are in units of pixels, are normalized by the distance between pupils, which serves as a robust and accurate length scale. To these purely geometrical features we added several non-geometrical measured features: The average face hue, the average hair color, and an estimate of skin smoothness. The latter was estimated by applying standard edge detection filters on the cheeks and forehead.

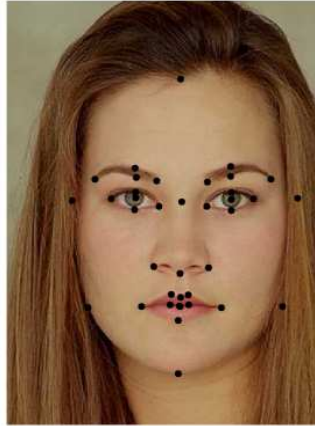


Figure 1: Marked points are the facial landmarks used to compute the measurements for the feature-based representation (e.g. distance between pupils, face width at eye level, face length).

3.2 Mutual Information Maps

We began our work with a preliminary study of our data and its relation to facial attractiveness using various data analysis methods. Within this framework, we applied a Mutual Information analysis to identify the features that are most "informative" for facial attractiveness determination. For this, we recast the problem of predicting facial attractiveness as a simpler one of discerning 'attractive' faces from 'unattractive' faces. The reduced problem can naturally be translated into the binary classification task of separating the 25% highest rated images, which comprise the class of "attractive" faces, from the 25% lowest rated images, which comprise the class of "unattractive" faces. Not surprisingly, humans also find this task to be much simpler and very high correlations between human judgements are observed. Feature values were also made discrete by binning.

A novel visualization of the facial regions important for attractiveness determination emerged in the calculation of the mutual information between the (binned) raw pixel values and the target classes. Since the mutual information is a function of the pixels, the whole array of mutual information values can be represented as a face-like image, as shown in Fig. 2. The lighter areas are the pixels that were found to be more "informative" for facial attractiveness determination.

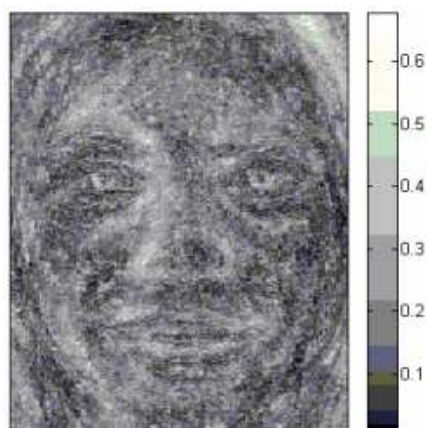


Figure 2: Pixel mutual information map (values can be seen in the colorbar on the right of the image)

The eyes (shape and hue), eyebrows, nose (length and width) and mouth are all clearly visible. Additional important features are cheeks, contour of lower face and hair, which was also very highly correlated with the ratings.

Results of a mutual information analysis on the constituents of the feature-based representation were consistent with the pixel mutual information maps.

3.3 Learning facial attractiveness

Using the facial images in both representations and their respective human ratings, we perform learning and prediction of facial attractiveness. As the dimension of the initial pixel image data is extremely high, of the order of 100,000, dimension reduction was required. Therefore, we reduced the dimension of the image data with Principal Component Analysis (PCA), a global dimension reduction technique, shown to relate reliably to human performance on face image processing tasks [6].

To produce sharper eigenfaces, all images were aligned before undergoing PCA. Faces were translated and rescaled to fix the location of the pupils at predefined positions. We further aligned the images according to a fixed vertical location of the center of the mouth. As this latter alignment changed face height to width ratio, the vertical scaling factor was added to the low-dimensional representation of each face.

PCA was also performed on the feature-based measurements, in order to decorrelate the variables in this representation, as well. This was important since strong correlations, stemming, for example, from left-right symmetry, were observed in the data.

Feature Selection

To enable good prediction of ratings we performed feature selection, for both representations. We selected features by ranking them according to their correlation with the human ratings. We experimented with other ranking criteria, such as chi-square and mutual information, but those produced somewhat inferior results.

Interestingly, in the pixel representation, features found most correlated with the attractiveness ratings were those pertaining to intermediate eigenvalues. This is in contrast to many face analysis applications, in which the largest eigenvalues are selected. Fig. 3 shows the eigenvectors from PCA on pixel images from the main dataset, where 3(a) shows those pertaining to the highest eigenvalues and 3(b) shows the highest correlated ones. While the former show mostly general features of hair and face contour, the latter also clearly show lip size, nose tip and eye size and shape as important features. This feature selection improved results considerably: correlation of predicted ratings with the human ratings with KNN, for example, rose from 0.25 to 0.45.



Figure 3a: Eigenfaces with highest eigenvalues

Figure 3b: Eigenfaces with highest correlations with attractiveness ratings

Prediction

The data vectors were projected onto the top m eigenvectors from the feature selection stage, where m is a parameter with which we performed optimization. These new projection vectors were the low-dimensional, information-preserving representation of the data output from the PCA stage to the learners in the prediction stage.

The main predictors we worked with were KNN and SVM. We also used linear regression, which serves as a baseline for the other methods.

Due to the relatively small sample sizes, we evaluated the performance of the predictors using cross-validation; predictions were made in leave- n -out, with $n = 1$ for linear regression and KNN and $n = 5$ for SVM. Predicted ratings were evaluated according to their correlation with the human ratings.

The output of the KNN predictor for a test image was the weighted average of the targets of the image's k nearest neighbors, where the weight of a neighbor was the inverse of its distance from the test image. The predictor was run with k values ranging from 1 to 45.

For the SVM method, we tried several kernels: linear, polynomial of degree 2 and 3 and gaussian with different values of γ , where $\log_2 \gamma \in \{-6, -4, -2, 0\}$. We performed a grid search over the values of slack parameter, c , and the width of regression tube, w , such that $\log_{10} C \in \{-3, -2, -1, 0, 1\}$ and $w \in \{0.1, 0.4, 0.7, 1.0\}$. In all runs we use a soft-margin SVM implemented in *SVMlight* [7].

Fig. 4 depicts the results of the predictors on the images of one of the datasets. The top figure shows the correlations reached with the pixel-based representation, where KNN was run with $k = 31$, and the bottom figure shows the results for the feature-based representation, where KNN correlations are for $k =$

42. The results for the pixel images show a peak near $m=25$ features, where the maximum correlation, achieved with KNN, is approximately 0.45. The figure for the feature-based representation shows a maximum value of nearly 0.6 at $m=15$ features, where the highest correlation is achieved both with SVM and linear regression. Results obtained on the second dataset were very similar. Highest SVM results in both representations were reached with a linear kernel.

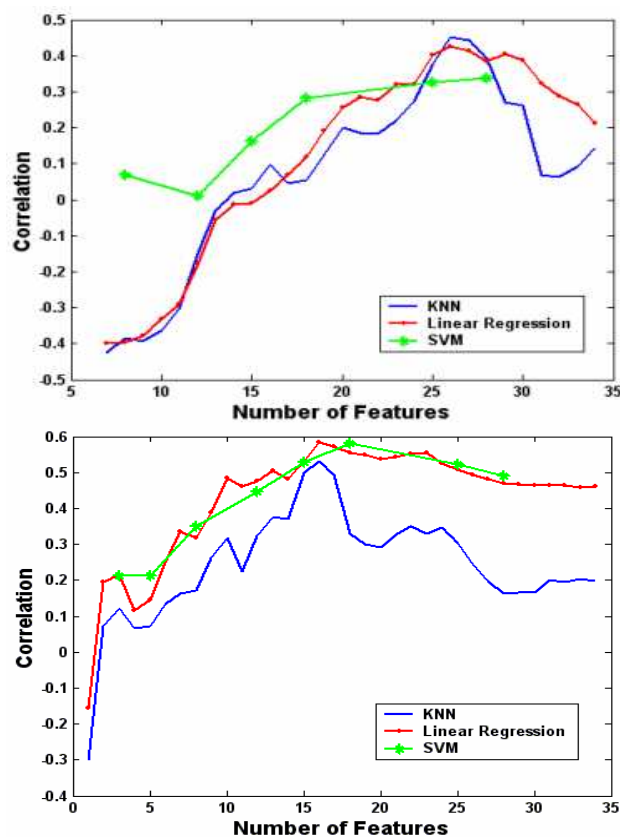


Figure 4: Results of all three predictors for the pixel images (top figure) and for the feature-based representation (bottom figure)

It is interesting to note that the simple linear regression predictor is as good and in certain cases, such as in the feature representation, even better than the KNN predictor. SVM performance was, for the most part, as good as and slightly better than that of the other methods.

In general, performance with the feature-based measurements was better, enabling a correlation of nearly 0.6 vs. a correlation of 0.45 with the pixel images. This implies that a feature-based representation, more "natural" than the pixel values, might be more informative for facial attractiveness determination.

4. Discussion

In this work we present a predictor of facial attractiveness, trained with female facial images and their respective average human ratings. Images were represented both as raw pixel data and as measurements of key facial features. Prediction was carried out using KNN, SVM and linear regression, and ratings predicted achieved a correlation of approximately 0.6 with the human ratings.

Given the high dimensionality and redundancy of visual data, the task of learning facial attractiveness is undoubtedly a difficult one. Nonetheless, our predictor achieved significant correlations with the targets. Yet, given the results of the prediction process and additional data analysis, we believe our success was limited by a number of hindering factors.

The most meaningful limiting factor was probably the relatively small size of the datasets. We confirmed this by iteratively running the predictor for a growing dataset size. Fig. 5 shows the results for KNN on the feature-based representation of one of the datasets with $k=16$ and $m=7$ features. Results shown are the average correlation of 10 runs with different subsets of images. The graph clearly shows improvement as the number of images increases. The slope of the graph is still positive with 92 images and does not asymptotically level off, implying that there is considerable room for improvement with a larger dataset. The size of the dataset also limited our ability to estimate the posterior distribution of the attractiveness ratings given feature value.

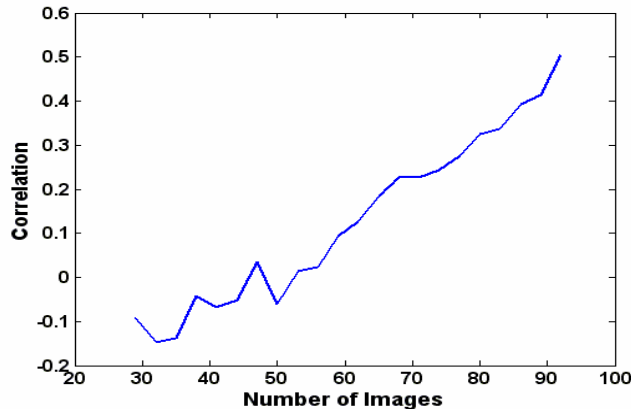


Figure 5: Correlation as a function of the number of images

Another limiting factor was most probably insufficient data representation. Our data analysis showed the relation of the features and pixels with the attractiveness ratings to be lower than initially expected, demonstrating the difficulty in learning from these data representations.

While producing better results than the pixel images, the feature-based representation is, nonetheless, insufficient; it includes only Euclidean distance-based measurements, and, therefore, lacks shape and texture information. Hues are totally omitted, and the shape of facial features is represented in a very coarse manner.

The relatively lower results with the pixel images show that this representation is, likewise, not informative enough for the discriminatory task of facial attractiveness evaluation. In addition to its high dimensionality and redundancy, the representation of a face in this vector space is “unnatural”, as, for example, an average of two faces will not necessarily produce a face-like image.

Future work should incorporate an encoding that is perceptually or cognitively more realistic, such as the output of Gabor filters or wavelets. Furthermore, as PCA operates independently of higher-level perceptions, combining feature extraction from pixel information with feature-based approaches, such as "eigenfeatures" [8], would probably result in improved performance.

In conclusion, our work, novel in its application of learning methods in the analysis of facial attractiveness, has produced promising results. Significant correlations with human ratings were achieved despite the difficulty of the task and several hindering factors. The entirety of our findings show promise of even better results in future research ameliorating these factors and overcoming the obstacles inherent in our work.

Acknowledgments

We thank the Ludwig-Boltzmann Institute for Urban Ethology at the Institute for Anthropology, University of Vienna, Austria for the dataset, which was of great assistance in our research.

References

1. D.I. Perrett, K.A. May and S. Yoshikawa, Facial Shape and Judgements of Female Attractiveness, *Nature*, 1994, Vol. 368, 239-242
2. D.I. Perrett, K.J. Lee, I. Penton-Voak, D. Rowland, S. Yoshikawa, D.M. Burt, S.P. Henzi, D.L. Castles and S. Akamatsu, Effects of Sexual Dimorphism on Facial Attractiveness, *Nature*, 1998, Vol. 394
3. M.R. Cunningham, Measuring the Physical in Physical Attractiveness: Quasi Experiments on the Sociobiology of Female Facial Beauty, *Journal of Personality and Social Psychology*, 1986, Vol. 50, No. 5, 925-935
4. J.H. Langlois, L.A. Roggman, R.J. Casey and J.M. Ritter, Infant preferences for attractive faces: Rudiments of a stereotype?, *Developmental Psychology*, 1987, Vol. 23, 363-369
5. V.S. Johnston and M. Franklin, Is Beauty in the Eye of the Beholder?, *Ethology and Sociobiology*, 1993, 14(3), 183-199
6. M. Turk and A. Pentland, Eigenfaces for Recognition, *Journal of Cognitive Neuroscience*, 1991, Vol. 3, No. 1
7. T. Joachims, Making large-Scale SVM Learning Practical. Advances in Kernel Methods - Support Vector Learning, B. Schölkopf and C. Burges and A. Smola (ed.), MIT-Press, 1999.
8. A. Pentland, B. Moghaddam, T. Starner, O. Oliyide and M. Turk, View-Based and Modular Eigenspaces for Face Recognition, Technical Report 245, M.I.T Media Lab, 1993