manuscript number 3033

Facial Attractiveness: Beauty and the Machine

Yael Eisenthal
School of Computer Science
Tel-Aviv University
Tel-Aviv 69978, Israel
yre2101@columbia.edu

Gideon Dror
Senior lecturer
Department of Computer Science
Academic College of Tel-Aviv-Yaffo
Tel-Aviv 64044, Israel
Tel: 972-3-5211857
Fax: 972-3-5211871
gideon@mta.ac.il

Eytan Ruppin
Professor of Computer Science and Medicine
School of Computer Science
Tel-Aviv University
Tel-Aviv 69978, Israel
Tel: 972-3-6405528
Fax: 972-3-6409357
ruppin@post.tau.ac.il

Corresponding author:
Gideon Dror
gideon@mta.ac.il

None of this material has been published or is under consideration for publication elsewhere.

Abstract

This work presents a novel study of the notion of “facial attractiveness” 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 attractiveness of a face. The trained predictor achieves a significant correlation of 0.65 with the average human ratings. The results clearly show that facial beauty is a universal concept, which can be learned by a machine. Analysis of the accuracy of the beauty prediction machine as a function of the size of the training data indicates that a machine producing human-like attractiveness rating could be obtained given a moderately larger dataset.

1 Introduction

In this work, we explore the notion of facial attractiveness through the application of machine learning techniques. We construct a machine which learns from facial images and their respective attractiveness ratings to produce human-like evaluation of facial attractiveness. Our work is based on the underlying theory that there are objective regularities in facial attractiveness to be analyzed and learned. In the introduction, we first briefly describe the psychophysics of facial attractiveness and its evolutionary origins. We then provide a review of previous work done in the computational analysis of beauty, attesting the novelty of our work.

1.1 The Psychophysics of Beauty

1.1.1 Beauty and the Beholder

The subject of visual processing of human faces has received attention from philosophers and scientists, such as Aristotle and Darwin, for centuries. Within this framework, the study of human facial attractiveness has had a significant part - “Beauty is a universal part of human experience, it provokes pleasure, rivets attention, and impels actions that help ensure survival of our genes” (Etcoff, 1999). Various experiments have empirically shown the influence of physical

attractiveness on our lives, both as individuals and as part of a society; its impact is obvious by the amounts of money spent on plastic surgery and cosmetics each year. Yet, the face of beauty, something we can recognize in an instant, is still difficult to formulate. This outstanding question regarding the constituents of beauty has led to a large body of ongoing research by scientists in the biological, cognitive and computational sciences.

Over centuries, the common notion in this research has 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. Several rating studies by Perrett et al. and other researchers have demonstrated high cross-cultural agreement in attractiveness rating of faces of different ethnicities (Cunningham, Roberts, Wu, Barbee & Druen, 1995; Jones, 1996; Perrett, May & Yoshikawa, 1994; Perrett, Lee, Penton-Voak, Rowland, Yoshikawa, Burt, Henzi, Castles & Akamatsu, 1998). This high congruence over ethnicity, social class, age and sex has led to the belief that perception of facial attractiveness is data-driven, i.e. that the properties of a particular set of facial features are the same irrespective of the perceiver. If different people can agree on which faces are attractive and which are not, when judging faces of varying ethnic background, then this suggests that people everywhere are using similar criteria in their judgements.

This belief is further strengthened by the consistent relations, demonstrated in experimental studies, between attractiveness and various facial features, with both

male and female raters. Cunningham et al. showed a strong correlation between beauty and specific features, which were categorized as neonate (features such as small nose and high forehead), mature (e.g. prominent cheekbones) and expressive (e.g. arched eyebrows). They concluded that beauty is not an inexplicable quality which lies only in the eye of the beholder (Cunningham, 1986; Cunningham, Barbee & Pike, 1990).

A second line of evidence in favor of a biological rather than an arbitrary cultural basis of physical attractiveness judgements comes from studies of infant preferences for face types. Langlois et al. showed pairs of female faces (that had been previously rated for attractiveness by adults) to infants only a few months old (Langlois, Roggman, Casey, Ritter, Rieser-Danner & Jenkins, 1987). The infants preferred to look at the more attractive face of the pair, indicating that even at two months of age, adult-like preferences are demonstrated. Slater et al. demonstrated the same preference in newborns (Slater, Von der Schulenberg, Brown, Badenoch, Butterworth, Parsons et al., 1998). The babies looked longer at the attractive faces, regardless of the gender, race, or age of the face.

The "owner vs observer" hypothesis was further studied in various experiments. Zaidel explored the question of whether beauty is in the perceptual space of the observer or a stable characteristic of the face (Chen, German & Zaidel, 1997). Results showed that facial attractiveness is more dependent on physiognomy of the face than on a perceptual process in the observer - both for male and female observers.

1.1.2 Evolutionary Origins

Since Darwin, biologists have studied natural beauty's meaning in terms of the evolved signal content of striking phenotypic features. Evolutionary scientists claim that the perception of facial features may be governed by circuits shaped by natural selection in the human brain. Aesthetic judgements of faces are not capricious but, instead, reflect evolutionary functional assessments and valuations of potential mates (Thornhill & Gangestad, 1993). These “Darwinian” approaches are based on the premise that attractive faces are a biological "ornament" that signals valuable information; attractive faces advertise a "health certificate", indicating a person's "value" as a mate (Thornhill & Gangestad, 1999). Advantageous biological characteristics are probably revealed in certain facial traits, which are unconsciously interpreted as attractive in the observer’s brain. Facial attributes like good skin quality, bone structure and symmetry, for example, are associated with good health and, therefore, contribute to attractiveness. Thus, human beauty standards reflect our evolutionary distant and recent past and emphasize the role of health assessment in mate choice, or, as phrased by anthropologist Donald Symons, “Beauty may be in the adaptations of the beholder".

Research has concentrated on a number of characteristics of faces, which may honestly advertise health and viability. Langlois and others have demonstrated a preference for average faces: composite faces, a result of digital blending and averaging of faces, were shown to be more attractive than most of the faces used to create them (Grammer & Thornhill, 1994; Langlois & Roggman, 1990;

Langlois, Roggman & Musselman, 1994; O´Toole, Price, Vetter, Bartlett & Blanz, 1999). Evolutionary biology holds that in any given population, extreme characteristics tend to fall away in favor of average ones, therefore, the ability to form an average-mate template would have conveyed a singular survival advantage (Symons, 1979; Thornhill & Gangestad, 1993).

The averageness hypothesis, however, has been widely debated. Average composite faces tend to have smooth skin and be symmetric; these factors, rather than averageness per se, may lead to the high attractiveness attributed to average faces (Alley & Cunningham, 1991). Both skin texture (Fink, Grammer & Thornhill, 2001) and facial bilateral symmetry (Grammer & Thornhill, 1994; Mealey, Bridgstock & Townsend, 1999; Perrett, Burt, Penton-Voak, Lee, Rowland & Edwards, 1999) have been shown to have a positive affect on facial attractiveness ratings. The averageness hypothesis has also received only mixed empirical support. Later studies found that, although averageness is certainly attractive, it can be improved upon. Composites of beautiful people were rated more appealing than those made from the larger, random population (Perrett et al., 1994). Also, exaggeration of the ways in which the prettiest female composite differed from the average female composite resulted in a more attractive face (O'Toole, Deffenbacher, Valentin, McKee, Huff & Abdi, 1997; Perrett et al., 1994, 1998); these turned out to be sexually dimorphic traits, such as small chin, full lips, high cheekbones, narrow nose and a generally small face. These sex-typical, estrogen dependent characteristics in females may indicate youth and

fertility, and are, thus, considered attractive (Perrett et al., 1998; Symons, 1979, 1995; Thornhill & Gangestad, 1999).

1.2 Computational Beauty Analysis

The previous section clearly indicates the existence of an objective basis underlying the notion of facial attractiveness. Yet the relative contribution to facial attractiveness of the aforementioned characteristics and their interactions with other facial beauty 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 (Langlois & Roggman, 1990; Perrett et al., 1994, 1999), as well as attractive and unattractive prototypes (http://www.beautycheck.de), in order to analyze their characteristics. Other approaches have addressed the question within the study of the relation between aesthetics and complexity, which is based on the notion that simplicity lies at the heart of all scientific theories ("Occam's Razor" principle). Schmidhuber created an attractive female face composed from a fractal geometry based on rotated squares and powers of two (Schmidhuber, 1998).

Exploring the question from a different approach, Johnston produced an attractive female face using a genetic algorithm, which evolves a "most beautiful" face according to interactive user selections (Johnston & Franklin, 1993). This

algorithm mimics, in an oversimplified manner, the way humans (consciously or unconsciously) select for features they find the most attractive. Measuring the features of the resulting face showed it to have "feminized" features. This study and others, which have shown attractiveness and femininity to be nearly equivalent for female faces (O'Toole et al., 1997), have been the basis for a commercial project, which uses these sex-dependent features to determine the sex of an image and predict its attractiveness (http://www.intelligent-earth.com).

1.3 This Work

Previous computational studies of human facial attractiveness have mainly involved averaging and morphing of digital images and geometric modeling to construct attractive faces. In general, computer techniques used include delineation, transformation, prototyping and other image processing techniques, most requiring fiducial points on the face. In this work, rather than attempt to morph or construct an attractive face, we explore the notion of facial attractiveness through the application of 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. The cross-cultural consistency in attractiveness ratings demonstrated in many previous studies has led to the common notion that there is an objective basis to be analyzed and learned.

The remainder of this paper is organized as follows: Section 2 presents the data used in our analyses – both images and ratings, where section 3 describes the representations we chose to work with. Section 4 describes our experiments with learning facial attractiveness, presenting prediction results and analyses. Finally, section 5 consists of a discussion of the work presented and general conclusions. Additional details are provided in Appendix A.

2 The Data

2.1 Image Datasets

To reduce the effects of age, gender, skin color, facial expression and other irrelevant factors, subject choice was confined to young Caucasian females in frontal view with neutral expression, without 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 each:

1. **Dataset #1** contains 92 young Caucasian (American) females in frontal view with neutral expressions, face and hair comprising the entirety of the

picture. The images all have identical lighting conditions and nearly identical orientation, in excellent resolution, with no obscuring or distracting features, such as jewelry and glasses. The pictures were originally taken by Japanese photographer Akira Gomi. Images were received with attractiveness ratings.

2. **Dataset #2** contains 92 Caucasian (Israeli) females, aged approximately 18, in frontal view, face and hair comprising the entirety of the picture. Most of the images have neutral expressions, but, in order to keep the dataset reasonably large, smiling images in which the mouth was relatively closed were also used. The images all have identical lighting conditions and nearly identical orientation. This dataset required some image preprocessing and is of slightly lower quality. The images contain some distracting features, such as jewelry.

The distributions of the raw images in the two datasets were found to be too different for combining the sets, and, therefore, all our experiments were conducted on each dataset separately. Dataset #1, which contains high-quality pictures of females in the preferred age range, with no distracting or obscuring items, was the main dataset used. Dataset #2, which is of slightly lower quality, containing images of younger women with some distracting features (jewelry, smiles), was used for exploring cross-cultural consistency in attractiveness judgement and in its main determinants. Both datasets were converted to grayscale to lower the dimension of the data and to simplify the computational task.

2.2 Image Ratings

2.2.1 Rating Collection

Dataset #1 was received with ratings, but, to check consistency of ratings across cultures we collected new ratings for both datasets. To facilitate both the rating procedure and the collection of the ratings, we created an interactive html-based application, which was used by all our raters. This provided a simple rating procedure, in which all participants received the same instructions and used the same rating process. The raters were asked to first scan through the entire dataset (in grayscale), to obtain a general notion of the relative attractiveness of the images, and only then to proceed to the actual rating stage. They were instructed to use the entire attractiveness scale, and to consider only facial attractiveness in their evaluation. In the rating stage, the images were shown in random order to eliminate order effects, each on a separate page. A rater could look at a picture for as long as he or she liked and then score it. The raters were free to return to pictures they had already seen and adjust their ratings.

Images in dataset #1 were rated by 28 observers - 15 male, 13 female, most in their twenties. For dataset #2, 18 ratings were collected from 10 male and 8 female raters of similar age.

Each facial image was rated on a discrete integer scale between 1 (very unattractive) and 7 (very attractive). The final attractiveness rating of a facial image was the mean of its ratings across all raters.

2.2.2 Rating Analysis

In order to verify the adequacy and consistency of the collected ratings, we examined the following properties:

- **Consistency of ratings:**

The raters were randomly divided into two groups. We calculated the mean ratings of each group and checked consistency between the two mean ratings. This procedure was repeated numerous times and consistently showed a correlation of 0.9-0.95 between the average ratings of the two groups for dataset #1 and a correlation of 0.88-0.92 for dataset #2. The mean ratings of the groups were also very similar, for both datasets, and a t-test confirmed that the rating means for the two groups were not statistically different.

- **Clustering of raters:**

The theory underlying the project is that individuals rate facial attractiveness according to similar, universal standards. Therefore, our assumption was that all ratings are from the same distribution. Indeed, clustering of raters produced no apparent grouping. Specifically, a chi-square test that compared the distribution of ratings of male versus female raters showed that there are no statistically significant differences between these two groups. In addition, the correlation between the average female ratings and average male ratings was very high: 0.92 for dataset #1 and 0.88 for dataset #2. The means of the female and male ratings

were also very similar, and a t-test confirmed that the means of the two groups were not statistically different. The results show no effect of observer gender.

An analysis of the original ratings for dataset #1 (collected from Austrian raters) vs. the new ratings (collected from Israeli raters) shows a high similarity in the images rated as most and least attractive. A correlation of 0.82 was found between the two sets of ratings. These findings strongly reinforce previous reports of high cross-cultural agreement in attractiveness rating.

3 Face Representation

Numerous studies in various face image processing tasks (e.g. face recognition and detection) have experimented with various ways to "specify" the physical information in human faces. The different approaches tried have demonstrated the importance of a broad range of shape and image intensity facial cues (Bruce & Langton, 1994; Burton, Bruce & Dench, 1993; Valentine & Bruce, 1986).

The most frequently encountered distinction regarding the information in faces is a qualitative one between feature-based and configurational-based information, i.e. discrete, local, featural information vs. spatial interrelationship of facial features. Studies suggest that humans perceive faces holistically and not as individual facial features (Baenninger, 1994; Haig, 1984; Young, Hellawell & Hay, 1989), yet experiments with both representations have demonstrated the importance of features in discriminative tasks (Bruce & Young, 1986; Moghaddam & Pentland, 1994). This is a particularly reasonable assumption for beauty judgement tasks, given the correlation found between features and attractiveness ratings. Our work uses both kinds of representations.

In the configurational representation, a face is represented with the raw grayscale pixel values, in which all relevant factors, such as texture, shading, pigmentation and shape, are implicitly coded (though difficult to extract). A face is represented by a vector of pixel values created by concatenating the rows or

columns of its image. The pixel-based representation of a face will be referred to as its "pixel image".

The featural representation is motivated by arguments tying beauty to ideal proportions of facial features such as distance between eyes, width of lips, size of eyes, distance between the lower lip and the chin etc. This representation is based on the manual measurement of 37 facial feature distances and ratios that reflect the geometry of the face (e.g. distance between eyes, mouth length and width). The facial feature points, according to which these distances were defined, are displayed in figure 1. The full list of feature measurements is given in Appendix A, along with their calculation method. All raw distance measurements, which are in units of pixels, were normalized by the distance between pupils, which serves as a robust and accurate length scale. To these purely geometric features we added several non-geometric ones: average hair color, an indicator of facial symmetry and an estimate of skin smoothness. The feature-based measurement representation of a face will be referred to as its "feature vector". The pixel-based representation of a face will be referred to as its "pixel image".

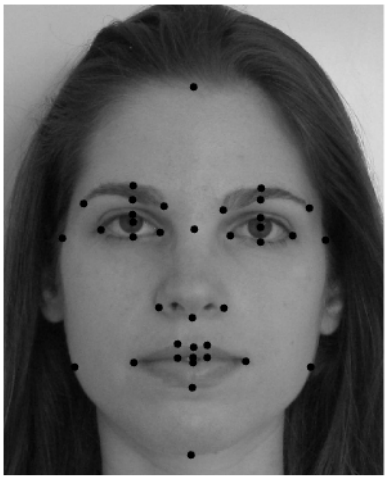

Figure 1: Feature landmarks used for feature-based representation

4 Learning Attractiveness

We turn to present our experiments with learning facial attractiveness, using the facial images and their respective human ratings. The learners were trained with the pixel representation and with the feature representation, separately.

4.1 Dimension Reduction

The pixel images are of an extremely high dimension, of the order of 100,000 (equal to image resolution). Given the high dimensionality and redundancy of the visual data, the pixel images underwent dimension reduction with Principal Component Analysis (PCA). PCA has been shown to relate reliably to human performance on various face image processing tasks, such as face recognition (O'Toole, Abdi, Deffenbacher & Valentin, 1993, Turk & Pentland, 1991) and race and sex classification (O'Toole, Deffenbacher, Abdi & Bartlett, 1991), and to be semantically relevant: The eigenvectors pertaining to large eigenvalues have been shown to code general information, such as orientation and categorical assessment, which has high variance and is common to all faces in the set (O'Toole et al., 1993; O´Toole, Vetter, Troje & Bülthoff, 1997; Valentin & Abdi, 1996). Those corresponding to the smaller eigenvalues code smaller, more individual variation (Hancock, Burton & Bruce, 1996; O'Toole et al., 1997).

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

4.1.1 Image Alignment

For PCA to extract meaningful information from the pixel images, the images need to be aligned, typically by rotating, scaling and translating, to bring the eyes to the same location in all the images. To produce sharper eigenfaces, we aligned the images according to a second point as well – the vertical location of the center of the mouth, a technique known to work well for facial expression recognition (Padgett & Cottrell, 1997). This non-rigid transformation, however, involved changing face height to width ratio. To take this change into account, the vertical scaling factor of each face was added to its low-dimensional representation.

As the input data in our case is face images and the eigenvectors are of the same dimension as the input, the eigenvectors are interpretable as faces and are often referred to as "eigenfaces" (Turk & Pentland, 1991). The improvement in sharpness of the eigenfaces from the main dataset as a result of the alignment can be seen in figure 2. Each eigenface deviates from uniform gray where there is variation in the face set. The left column consists of two eigenfaces extracted from PCA on unaligned images; face contour and features are blurry. The middle column shows eigenfaces from images aligned only according to eyes. The eyes are indeed more sharply defined, but other features are still blurred. The right

column shows eigenfaces from PCA on images aligned both by eyes and vertical location of mouth – all salient features are much more sharply defined.

Figure 2: Left column: eigenfaces from unaligned images, middle: eigenfaces from images aligned only by eyes, right: eigenfaces from images aligned both by eyes and mouth

4.1.2 Eigenfaces

PCA was performed on the input vectors from both representations, separately. Examples of eigenvectors extracted from the pixel images from the main dataset can be seen in figure 3. The eigenfaces in the top row are those pertaining to the highest eigenvalues, the middle row shows eigenfaces corresponding to intermediate eigenvalues and the bottom row presents those pertaining to the smallest eigenvalues. As expected, the eigenfaces in the top row

seem to code more global information, such as hair and face shape, while the eigenvectors in the bottom row code much more fine, detailed feature information.

Each eigenface is obviously not interpretable as a simple single feature (as is often the case with a smaller dataset), yet it is clearly seen in the top-row eigenfaces that the directions of highest variance are hair and face contour. This is not surprising as the most prominent differences between the images are in hair color and shape, which also causes large differences in face shape (due to partial occlusion by hair). Smaller variance can also be seen in other features, mainly eyebrows and eyes.

Figure 3: Eigenfaces from largest to smallest eigenvalues (top to bottom)

4.2 Feature Selection

The eigenfaces are the features representing the face set; they can be combined in a certain weighting to represent a specific face. A low-dimensional representation using only the first eigenvectors minimizes the squared-error between the face representation and the original image, and is sufficient for accurate face recognition (Turk & Pentland, 1991). However, omitting the dimensions pertaining to the smaller eigenvalues decreases the perceptual quality of the face (O'Toole et al., 1993, 1997). Consequently, we anticipated that using the first *m* eigenfaces would not produce accurate results in our attractiveness evaluation task. Indeed, these experiments resulted in poor facial attractiveness predictions. We therefore selected the eigenfaces most important to our task by sorting them according to their relevance to attractiveness ratings. This relevance was estimated by calculating the correlation of the eigenvector projections with the human ratings across the various images. Interestingly, in the pixel representation, the features found most correlated with the attractiveness ratings were those pertaining to intermediate and smaller eigenvalues. Figure 4 shows the eigenfaces, where the top row displays those pertaining to the highest eigenvalues and the bottom row presents the eigenfaces with projections most correlated with human ratings. While the former show mostly general features of hair and face contour, the latter also clearly show the lips, the nose tip and eye size and shape to be important features.

The same method was used for feature selection in the feature-based representation. The feature measurements were sorted according to their correlation with the attractiveness ratings.

It should be noted that, despite its success, using correlation as a relevance measure is problematic, as it assumes the relation between the feature and the ratings to be monotonic. Yet, experiments with other ranking criteria that do not make this assumption, such as chi-square and mutual information, produced somewhat inferior results.

Figure 4: Eigenfaces pertaining to highest eigenvalues (top row) and highest correlations with ratings (bottom row)

4.3 Attractiveness Prediction

The original data vectors were projected onto the top *m* eigenvectors from the feature selection stage (where *m* is a parameter on which we performed optimization) to produce a low-dimensional representation of the data as input to the learners in the prediction stage.

4.3.1 Classification into Two Attractiveness Classes

Although the ultimate goal of this work was to produce and analyze a facial beauty predictor using regression methods, we begin with a simpler task, on which there is even higher consistency between raters. To this end, we recast the problem of predicting facial attractiveness into a classification problem - discerning "attractive" faces (the class comprised of highest 25% rated images) from "unattractive" faces (the class of lowest 25% rated images). The main classifiers used were standard K-Nearest Neighbors (KNN) (Mitchell, 1997) and Support Vector Machines (SVM) (Vapnik, 1995).

The best results obtained are shown in Table 1, which displays the percentage of correctly classified images. Classification using the KNN classifier was good; correct classifications of 75%-85% of the images were achieved. Classification rates with SVM were slightly poorer, though, for the most part, in the same percentage range. Both classifiers performed better with the feature vectors than with the pixel images; this is particularly true for SVM. Best SVM results were achieved using a linear kernel. In general, classification (particularly with KNN) was good for both datasets and ratings, with success percentages slightly lower for the main dataset.

KNN does not use specific features, but rather averages over all dimensions, and, therefore, does not give any insight into which features are important for attractiveness rating. In order to learn what the important features are, we used a

C4.5 Decision Tree (Quinlan, 1986, 1993) for classification using feature vectors without preprocessing by PCA.

In most cases, the results did not surpass those of the KNN classifier, but the Decision Tree did give some insight into which features are "important" for classification. The features found most informative were those pertaining to size of the lower part of the face (jaw length, chin length), smoothness of skin, lip fullness and eye size. These findings are all consistent with previous psychophysics studies.

		Dataset #1	**Dataset #2**
pixel images	**KNN**	75%	77%
	SVM	68%	73%
feature vectors	**KNN**	77%	86%
	SVM	76%	84%

Table 1: Percentage of correctly classified images

4.3.2 The Learners for the Regression Task

Following the success of the classification task, we proceeded to the regression task of rating prediction. The predictors used for predicting facial beauty itself were, again, KNN and SVM. For this task, however, both predictors were used in their regression version, mapping each facial image to a real number that represents its beauty. We also used linear regression, which served as a

baseline for the other methods. Targets used were the average human ratings of each image.

The output of the KNN predictor for a test image was computed as the weighted average of the targets of the image's k nearest neighbors, where the weight of a neighbor is the inverse of its Euclidean distance from the test image. That is, let $v_1,\ldots,v_k$ be the set of k nearest neighbors of test image v with targets $y_1,\ldots y_k$, and let $d_1,\ldots,d_k$ be their respective Euclidean distances from v. The predicted beauty y for the test image v is then

$$y = \frac{\sum_i w_i y_i}{\sum_i w_i} \quad ,i = 1,2\ldots k$$

where $w_i = (d_i + \sigma)^{-1}$ is the weight of neighbor v_i and where σ is a smoothing parameter. On all subsequent uses of KNN we set σ=1. KNN was run with k values ranging from 1 to 45.

As a predictor for our task, KNN suffers from a couple of drawbacks. First, it performs averaging, and, therefore, its predicted ratings had very low variance, and all extremely high or low ratings were evened out and often not reached. In addition, it uses a Euclidean distance metric, which need not reflect the true metric for evaluation of face similarity. Therefore, we also studied an SVM regressor as an attractiveness predictor, a learner which does not use a simple distance metric and does not perform averaging in its prediction.

The SVM method, in its regression version, was used with several kernels: linear, polynomials of degree 2 and 3 and Gaussian with different values of γ, where $\log_2\gamma \in \{-6, -4, -2, 0\}$. γ is related to the width parameter σ by $\gamma = 1/(2\sigma^2)$.

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 used a soft-margin SVM implemented in SVM*light* (Joachims, 1999).

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

4.3.3 Results of Facial Attractiveness Prediction

Predicted ratings were evaluated according to their correlation with the human ratings, using the Pearson correlation. The best results of the attractiveness predictors on the main dataset are shown in figure 5. The left figure shows the best correlations achieved with the pixel-based representation, and the right figure shows the best results for the feature-based representation. Prediction results for the pixel images show a peak near m=25 features, where the maximum correlation achieved with KNN is approximately 0.45. 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. Highest SVM results in both representations were reached with a linear kernel. Results obtained on the second dataset were very similar.

The normalized MSE of the best predicted ratings is 0.6-0.65 (vs. a normalized MSE of 1 of the "trivial predictor", which constantly predicts the mean rating). KNN performance was poor, significantly worse than that of the other regressors in the feature-based representation. These results imply that the

Euclidean distance metric is most probably not a good estimate for similarity of faces for this task. It is interesting to note that the simple linear regressor performed as good as or better than the KNN predictor. However, this effect may be attributed to our feature selection method, ranking features by the absolute value of their correlations with the target, which is optimal for linear regression.

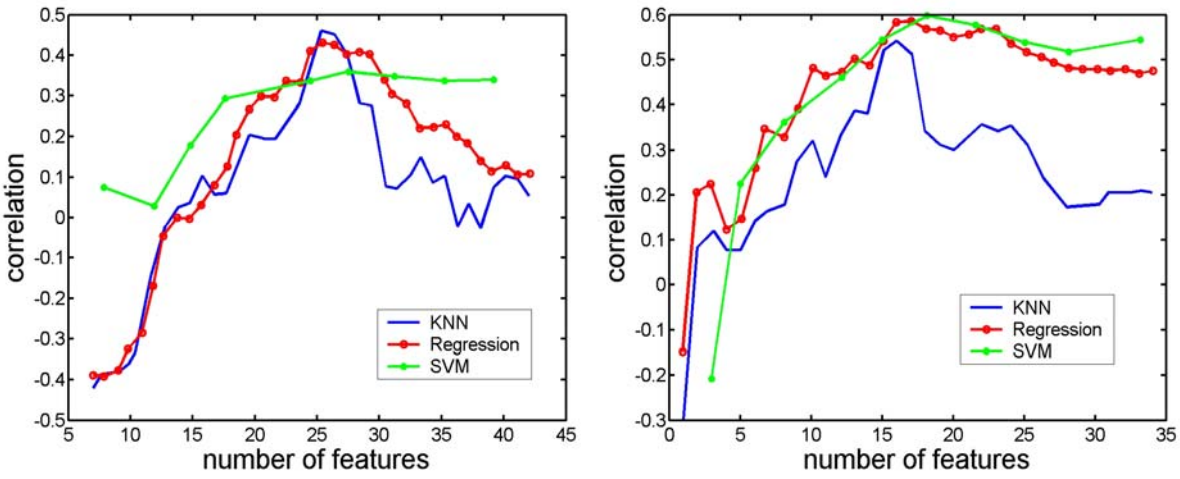

Figure 5: Prediction results obtained with pixel images (left) and with feature-based representation (right). Performance is measured by the correlation between the predicted ratings and the human ratings.

4.3.4 Significance of Results

All predictors performed better with the feature-based representation than with the pixel images (in accordance with results of classification task). Using the feature vectors enabled a maximum correlation of nearly 0.6 vs. a correlation of 0.45 with the pixel images. To check the significance of this score, we produced an empirical distribution of feature-based prediction scores with random ratings.

The entire preprocessing, feature selection and prediction process was run 100 times, each time with a different set of randomly generated ratings, sampled from a normal distribution with mean and variance identical to those of the human ratings. For each run, the score taken was the highest correlation of predicted ratings with the original (random) ratings. The average correlation achieved with random ratings was 0.28 and the maximum correlation was 0.47. Figure 6(a) depicts the histogram of these correlations. Using QQplot, we verified that the empirical distribution of observed correlations is approximately normal; this is shown in figure 6(b). Using the normal approximation, the correlation of 0.6 obtained by our feature-based predictor is significant to a level of α=0.001. The numbers and figures presented are for the KNN predictor. Correlations achieved with linear regression have different mean and standard deviation, but a similar z-value. The distribution of these correlations was also verified to be approximately normal and the correlation achieved by our linear regressor was significant to the same level of α=0.001. This test was not run for the SVM predictor due to computational limitations.

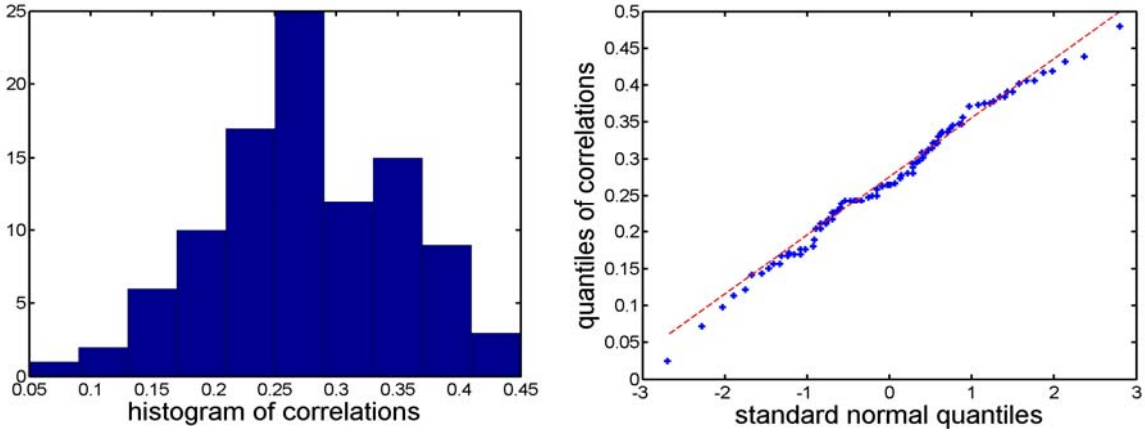

Figure 6: Correlations achieved with random ratings: (a) Histogram of correlations (left), (b) QQplot of correlations vs. standard normal distributed data (right). Correlations were obtained with KNN predictor.

4.3.5 Hybrid Predictor

Ratings predicted with the two representations were very different; the best ratings achieved using each representation had a correlation of only 0.3-0.35. The relatively low correlation between the feature-based and pixel-based predictions suggests that results might be improved by using the information learned from both representations. Therefore, an optimal weighted average of the best feature-based and pixel-based ratings was calculated. We produced a hybrid machine that generates the target rating $y_{hybrid} = \alpha y_{feature} + (1-\alpha) y_{pixel}$, where $y_{feature}$ is the rating of the feature-based predictor, y_{pixel} is the prediction of the pixel-based machine and $0 \leq \alpha \leq 1$. Figure 7 shows the correlation between the hybrid machine ratings and the human ratings as a function of the weights tried (weights shown are those of the feature-based ratings, α). The hybrid predictor was constructed using the best feature-based and pixel-based ratings obtained with linear regression. As evident from the graph, the best weighted ratings achieve a

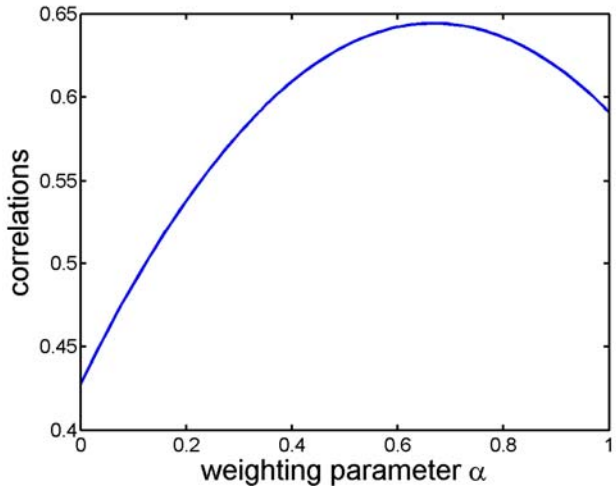

Figure 7: Correlation of the weighted machine ratings with human ratings vs. the weighting parameter, α.

correlation of 0.65 with the human ratings. The hybrid predictor with the optimal value of $\alpha = 0.65$ improves prediction results by nearly 10% over those achieved with a single representation. Its normalized MSE is 0.57, lower than that of the individual rating sets. These weighted ratings have the highest correlation and lowest normalized MSE with the human scores. Therefore, in subsequent analysis we use these weighted ratings as the best machine-predicted ratings, unless stated otherwise.

4.3.6 Evaluation of Predicted Rating Ranking

An additional analysis was performed to evaluate the relative image ranking induced by the best machine predictions. Figure 8 shows the probability of error in the predicted relative ordering of two images as a function of the absolute distance *d* in their original, human ratings. The distances were binned into 16 bins. The probability of error, for each distance *d*, was computed over all pairs of images with an absolute difference of *d* in their human ratings. As evident from the graph, the probability decreases almost linearly as the absolute difference in the original

ratings grows (the small peak observed at a distance of 3.6 is insignificant and stems from one erroneous relative ranking).

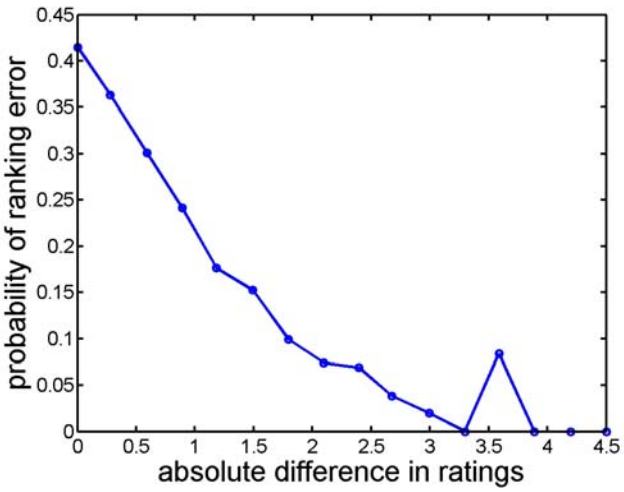

Figure 8: Probability of error in the predicted relative order of two images as a function of the absolute difference in their original, human ratings.

4.3.7 The Learning Curve of Facial Attractiveness

For further evaluation of the prediction machine, an additional experiment was run, in which we examined the learning curve of our predictor. We produced this curve by iteratively running the predictor for a growing dataset size, in the following manner: The number of input images, *n*, was incremented from 5 to the entire 92 images. For every *n*, the predictor was run ten times, each time with *n* different, randomly selected images (for n=92 all images were used in a single run). Testing was performed on the subsets of *n* images only, using leave-one-out, and results were evaluated according to the correlation of the predicted ratings

with the human ratings of these images. Figure 9 shows the results for the KNN predictor trained using the feature representation with k=16 and m=7 features. The correlations shown in the plot are the average over the ten runs. The figure clearly shows that the performance of the predictor improves with the increase in the number of images. The slope of the graph is positive for every $n \geq 50$. Similar behavior was observed with other parameters and learners. This tendency is less distinct in the corresponding graph for the pixel images.

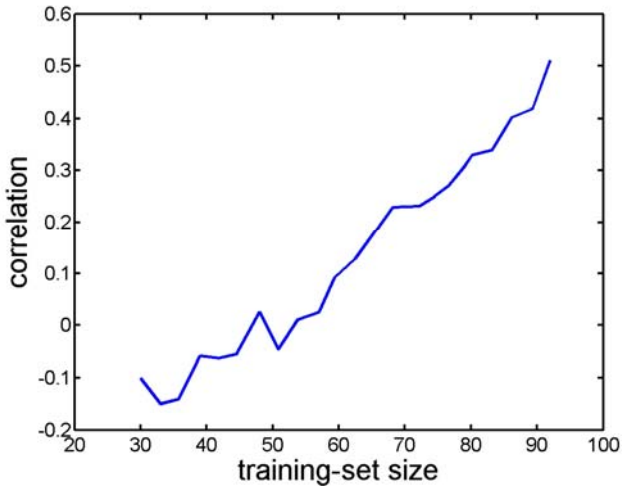

Figure 9: Accuracy of prediction as a function of the training-set size

5 Discussion

This paper presents 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 the best predicted ratings achieved a correlation of 0.65 with the human ratings. We consistently found that the measured facial features were more informative for attractiveness prediction, on all tasks tried.

In addition to learning facial attractiveness, we also examined some characteristics found correlated with facial attractiveness in previous experiments. In particular, we ran our predictor on the "average" face, i.e. the mathematical average of the faces in the dataset. This face received only an average rating, showing no support for the averageness hypothesis in our setting. This strengthens previous experiments that argued against the averageness hypothesis (as described in section 1.1.2). The high attractiveness of composite faces may be attributable to their smooth skin and symmetry and not to the averageness itself, explaining the fact that the mathematical average of the faces was not found to be very attractive.

Given the high dimensionality and redundancy of visual data, the task of learning facial attractiveness is undoubtedly a difficult one. We tried additional preprocessing, feature selection and learning methods, e.g. Wrapper (Kohavi & John, 1996), Isomap (Tenenbaum, de Silva & Langford, 2000) and Kernel PCA

(Scholkopf, Smola & Muller, 1999), but these all produced poorer results. The non-linear feature extraction methods probably failed due to an insufficient number of training examples, as they require a dense sampling of the underlying manifold. Nevertheless, our predictor achieved significant correlations with the human ratings. However, we believe our success was limited by a number of hindering factors.

The most meaningful obstacle in our project is likely to be the relatively small size of the datasets available to us. This limitation can be appreciated by examining figure 9, which presents a plot of prediction performance vs. the size of the dataset. The figure clearly shows improvement in the predictor's performance as the number of images increases. The slope of the graph is still positive with the 92 images used and does not asymptotically level off, implying that there is considerable room for improvement by using a larger, but still realistically conceivable, dataset.

Another likely limiting factor is insufficient data representation. While the feature-based representation produced better results than the pixel images, it is, nonetheless, incomplete; it includes only Euclidean distance-based measurements and lacks fine shape and texture information. The relatively lower results with the pixel images show that this representation is, likewise, not informative enough.

In conclusion, this work, novel in its application of computational learning methods for 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 results clearly show that facial beauty is a universal concept which can be learned by a machine. There are sufficient grounds to believe that future work with a moderately larger dataset may lead to an "attractiveness machine" producing human-like evaluations of facial attractiveness.

Acknowledgements

We thank Dr. Bernhard Fink and the Ludwig-Boltzmann Institute for Urban Ethology at the Institute for Anthropology, University of Vienna, Austria, for one of the facial datasets used in this research.

Appendix A

Feature-based representation

Following is a list of the measurements comprising the feature-based representation:

1. face length
2. face width – at eye level
3. face width – at mouth level
4. distance between pupils
5. ratio between 2 and 3
6. ratio between 1 and 2
7. ratio between 1 and 3
8. ratio between 4 and 2
9. right eyebrow thickness (above pupil)
10. left eyebrow thickness (above pupil)
11. right eyebrow arch – height difference between highest point and inner edge
12. left eyebrow arch – height difference between highest point and inner edge
13. right eye height
14. left eye height
15. right eye width
16. left eye width
17. right eye size = height * width

18. left eye size = height *width
19. distance between inner edges of eyes
20. nose width at nostrils
21. nose length
22. nose size = width * length
23. cheekbone width = (2 - 3)
24. ratio between 23 and 2
25. thickness of middle of top lip
26. thickness of right side of top lip
27. thickness of left side of top lip
28. average thickness of top lip
29. thickness of lower lip
30. thickness of both lips
31. length of lips
32. chin length – from bottom of face to bottom of lower lip
33. right jaw length – from bottom of face to right bottom face edge
34. left jaw length – from bottom of face to left bottom face edge
35. forehead height – from nose top to top of face
36. ratio of (distance from nostrils to eyebrow top) to (distance from face bottom to nostrils)
37. ratio of (distance from nostrils to face top) to (distance from face bottom to nostrils)
38. symmetry indicator (description follows)

39. skin smoothness indicator (description follows)

40. hair color indicator (description follows)

Symmetry indicator

A vertical symmetry axis was set between the eyes of each image and two rectangular, identically-sized windows, surrounding only mouth and eyes, were extracted from opposite sides of the axis. The symmetry measure of the image was calculated as $\frac{1}{N}\sum_i (X_i - Y_i)^2$, where N is the total number of pixels in each window, X_i is the value of pixel i in the right window and Y_i is the value of the corresponding pixel in the left window. The value of the indicator grows with the asymmetry in a face. This indicator is indeed a measure of the symmetry in the facial features, as the images are all consistent in lighting and orientation.

Skin smoothness indicator

The "smoothness" of a face was evaluated by applying a Canny edge detection operator to a window from the cheek/forehead area; a window representative of the skin texture was selected for each image. The skin smoothness indicator was the average value of the output of this operation, and its value monotonously decreases with the smoothness of a face.

Hair color indicator

A window representing the average hair color was extracted from each image. The indicator was calculated as the average value of the window, thus increasing with lighter hair.

References

Alley, T. R. & Cunningham, M. R. (1991) Averaged faces are attractive, but very attractive faces are not average *Psychological Science, 2*, 123-125.

Baenninger, M. (1994) The development of face recognition: Featural or configurational processing? *Journal of Experimental Child Psychology, 57*, 377-396.

Bruce, V. & Young, A. W. (1986) Understanding face recognition *British Journal of Psychology, 77*, 305-327.

Bruce V. & Langton S. (1994) The use of pigmentation and shading information in recognizing the sex and identities of faces *Perception, 23*, 803-822.

Burton, A. M., Bruce, V. & Dench, N. (1993) What's the difference between men and women? Evidence from facial measurement *Perception, 22*(2), 153-76.

Chen, A. C., German, C. & Zaidel, D. W. (1997) Brain asymmetry and facial attractiveness: Facial beauty is not simply in the eye of the beholder *Neuropsychologia, 35*(4), 471-476.

Cunningham, M. R. (1986) Measuring the physical in physical attractivenes: Quasi experiments on the sociobiology of female facial beauty *Journal of Personality and Social Psychology, 50*(5), 925-935.

Cunningham, M. R., Barbee, A. P., & Pike, C. L. (1990) What do women want? Facial metric assessment of multiple motives in the perception of male physical attractiveness *Journal of Personality and Social Psychology, 59*, 61-72.

Cunningham, M. R., Roberts, A. R., Wu, C. H., Barbee, A. P., & Druen, P. B. (1995) Their ideas of beauty are, on the whole, the same as ours: Consistency and variability in the cross-cultural perception of female attractiveness *Journal of Personality and Social Psychology*, *68*, 261–279.

Etcoff, N. (1999). *Survival of the Prettiest: The Science of Beauty.* First Anchor Books.

Fink, B., Grammer, K. & Thornhill, R. (2001) Human (homo sapien) facial attractiveness in relation to skin texture and color *Journal of Comparative Psychology, 115*(1), 92-99.

Grammer, K. & Thornhill, R. (1994) Human facial attractiveness and sexual selection: The role of symmetry and averageness *Journal of Comparative Psychology, 108*(3), 233-242.

Haig, N. D. (1984) The effect of feature displacement on face recognition *Perception, 13*, 505 – 512.

Hancock, P. J. B., Burton, A. M. & Bruce, V. (1996) Face processing: Human perception and PCA *Memory and Cognition, 24*, 26-40.

Joachims, T. (1999). Making large-scale SVM learning practical. In B. Schölkopf, C. Burges & A. Smola (Eds.), *Advances in Kernel Methods - Support Vector Learning.* MIT Press.

Johnston, V. S. & Franklin, M. (1993) Is beauty in the eye of the beholder? *Ethology and Sociobiology, 14*, 183-199.

Jones, D. (1996). *Physical Attractiveness and the Theory of Sexual Selection: Results from Five Populations.* University of Michigan Museum.

Kohavi, R. & John, G. H. (1996) Wrappers for feature subset selection *Artificial Intelligence, 97*(1-2), 273-324.

Langlois, J. H., Roggman, L. A., Casey, R. J., Ritter, J. M., Rieser-Danner, L. A., & Jenkins, V. Y. (1987) Infant preferences for attractive faces: Rudiments of a stereotype? *Developmental Psychology, 23*, 363-369.

Langlois, J. H. & Roggman, L. A. (1990) Attractive faces are only average *Psychological Science, 1*, 115-121.

Langlois, J. H., Roggman, L. A., & Musselman, L. (1994) What is average and what is not average about attractive faces? *Psychological Science, 5*, 214-220.

Mealey, L., Bridgstock, R. & Townsend, G. C. (1999) Symmetry and perceived facial attractiveness: A monozygotic co-twin comparison *Journal of Personality and Social Psychology, 76*(1), 151-158.

Mitchell, T. (1997). *Machine Learning.* McGraw Hill.

Moghaddam, B. & Pentland, A. (1994) Face recognition using view-based and modular eigenspaces *Automatic Systems for the Identification and Inspection of Humans, SPIE, 2257*.

O'Toole, A. J., Deffenbacher, K. A., Abdi, H., & Bartlett, J. A. (1991) Simulating the "other-race effect" as a problem in perceptual learning *Connection Science Journal of Neural Computing, Artificial Intelligence, and Cognitive Research, 3*, 163-178.

O'Toole, A., Abdi, H., Deffenbacher, K. A. & Valentin, D. (1993) Low-dimensional representation of faces in higher dimensions of the face space *Journal of the Optical Society of America A, 10*(3), 405-411.

O'Toole, A. J., Deffenbacher, K. A., Valentin, D., McKee, K., Huff, D. & Abdi, H. (1997) The perception of face gender: The role of stimulus structure in recognition and classification *Memory and Cognition, 25*.

O´Toole, A. J., Vetter, T., Troje, N. F. & Bülthoff, H. H. (1997) Sex classification is better with three-dimensional head structure than with image intensity information *Perception* 26, pp. 75-84.

O´Toole, A. J., T. Price, T. Vetter, J. C. Bartlett & V. Blanz (1999) 3D shape and 2D surface textures of human faces: the role of "averages" in attractiveness and age *Image and Vision Computing, 18*, 9-19.

Padgett, C. & Cottrell, G. (1997) Representing Face Images for Emotion Classification *Advances in Neural Information Processing System*s*, 9*, Editors. M. Mozer, M. Jordan and T. Petsche.

Perrett, D. I., May, K. A. & Yoshikawa, S. (1994) Facial shape and judgements of female attractiveness *Nature, 368*, 239-242.

Perrett, D. I., Lee, K. J., Penton-Voak, I., Rowland, D. A. , Yoshikawa, S., Burt, D. M., Henzi, S. P., Castles, D. L. & Akamatsu, S. (1998) Effects of sexual dimorphism on facial attractiveness *Nature, 394*.

Perrett, D. I., Burt, D. M., Penton-Voak, I. S., Lee, K. J., Rowland, D. A. & Edwards, R. (1999) Symmetry and human facial attractiveness *Evolution and Human Behavior, 20*, 295-307.

Quinlan, J. R. (1986) Induction of decision trees *Machine Learning, 1*, 81-106.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning.* Morgan Kauffman.

Schmidhuber, J. (1998). Facial beauty and fractal geometry. http://www.idsia.ch/~juergen/locoface/newlocoface.html

Scholkopf, B., Smola, A. & Muller, K. R. (1999) Kernel principal component analysis. In B. Schölkopf, C. Burges & A. Smola (ed.), *Advances in Kernel Methods - Support Vector Learning.* MIT Press.

Slater, A., Von der Schulenberg, C., Brown, E., Badenoch, M., Butterworth, G., Parsons, S. *et al*. (1998) Newborn infants prefer attractive faces *Infant Behavior and Development, 21*, 345-354.

Symons, D. (1979). *The Evolution of Human Sexuality.* Oxford University Press.

Symons, D. (1995). Beauty is in the adaptations of the beholder: The evolutionary psychology of human female sexual attractiveness. In P. R. Abramson & S. D. Pinkerton (Eds.), *Sexual Nature, Sexual Culture* (pp. 80-118). Chicago: University of Chicago Press.

Tenenbaum, J., de Silva, V. & Langford, J. (2000) A global framework for nonlinear dimensionality reductio*n Science, 290*(550), 2319-2323.

Thornhill, R. & Gangestad, S. W. (1993) Human facial beauty: Averageness, symmetry and parasite resistance *Human Nature, 4*(3), 237-269.

Thornhill, R., & Gangestad, S. W. (1999) Facial attractiveness *Trends in Cognitive Sciences, 3*, 452-460.

Turk, M. & Pentland, A. (1991) Eigenfaces for recognition *Journal of Cognitive Neuroscience, 3*(1).

Valentine, T. & Bruce, V. (1986) The effects of race, inversion and encoding activity upon face recognition *Acta Psychologica, 61*, 259-273.

Valentin, D. & Abdi, H. (1996) Can a linear autoassociator recognize faces from new orientations? *Journal of the Optical Society of America, series A, 13*, 717-724.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. Springer-Verlag, New York.

Young, A. W., Hellawell, D. & Hay, D. C. (1989) Configurational information in face perception *Perception, 16*, 747-759.