Available online at www.sciencedirect.com

ScienceDirect

Vision Research 48 (2008) 235–243

Vision Research

www.elsevier.com/locate/visres

A machine learning predictor of facial attractiveness revealing human-like psychophysical biases

Amit Kagian [a], Gideon Dror [b], Tommer Leyvand [a], Isaac Meilijson [c], Daniel Cohen-Or [a], Eytan Ruppin [a,d,*]

[a] School of Computer Sciences, Tel-Aviv University, Tel-Aviv 69978, Israel
[b] School of Computer Sciences, The Academic College of Tel-Aviv-Yaffo, Tel-Aviv 64044, Israel
[c] School of Mathematical Sciences, Tel-Aviv University, Tel-Aviv 69978, Israel
[d] School of Medicine, Tel-Aviv University, Tel-Aviv 69978, Israel

Received 16 June 2007; received in revised form 28 September 2007

Abstract

Recent psychological studies have strongly suggested that humans share common visual preferences for facial attractiveness. Here, we present a learning model that automatically extracts measurements of facial features from raw images and obtains human-level performance in predicting facial attractiveness ratings. The machine's ratings are highly correlated with mean human ratings, markedly improving on recent machine learning studies of this task. Simulated psychophysical experiments with virtually manipulated images reveal preferences in the machine's judgments that are remarkably similar to those of humans. Thus, a model trained explicitly to capture a specific operational performance criteria, implicitly captures basic human psychophysical characteristics.

Keywords: Face perception; Facial attractiveness; Machine learning; Aesthetics; Computational neuroscience

1. Introduction

Philosophers, artists and scientists have been trying to capture the nature of beauty since the early days of philosophy. Although in modern days a common layman's notion is that judgments of beauty are a matter of subjective opinion alone, recent findings suggest that people share a common taste for facial attractiveness and that their preferences may be an innate part of our primary constitution. Several experiments have shown that 2–8 months old infants prefer looking at faces rated by adults as more attractive (Langlois et al., 1987). In addition, attractiveness ratings show very high agreement between groups of raters belonging to the same culture and even across cultures (Cunningham, Roberts, Wu, Barbee, & Druen, 1995). Such findings give rise to the quest for common factors which determine human facial attractiveness. Accordingly, various hypotheses, from cognitive, evolutional and social perspectives, have been put forward to describe and interpret the common preferences for facial beauty.

Inspired by Sir Francis Galton's photographic method of composing faces (Galton, 1878), Langlois and Rogmann have created averaged faces by morphing multiple images together. Human judges found these averaged faces to be attractive and rated them with attractiveness ratings higher than the mean rating of the component faces composing them, proposing that averageness is the answer for facial attractiveness (Langlois & Roggman, 1990; Rubenstein, Langlois, & Roggman, 2002). Investigating symmetry and averageness of faces, Grammer and Thornhill concluded that symmetry was more important than averageness in facial attractiveness (Grammer & Thornhill, 1994). Other

* Corresponding author. Address: School of Computer Sciences and School of Medicine, Tel-Aviv University, Tel-Aviv 69978, Israel. Fax: +972 3 640 9357.
E-mail addresses: amit.kagian@gmail.com (A. Kagian), ruppin@post.tau.ac.il (E. Ruppin).

0042-6989/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.visres.2007.11.007

studies have agreed that average faces are attractive but claim that faces with certain extreme features, such as extreme sexually dimorphic traits, may be more attractive than average faces (Little, Penton-Voak, Burt, & Perrett, 2002). Yet other researchers have suggested various conditions which may contribute to facial attractiveness such as neonate features, pleasant expressions and familiarity (Zebrowitz & Rhodes, 2002). Finally, Cunningham et al. have suggested a multiple fitness model in which there is no single constructing line that determines attractiveness (e.g. perception of fitness as implying an ideal romantic partner). Instead, different categories of features signal different desirable qualities of the perceived target (Cunningham, Barbee, & Philhower, 2002). Even so, the multiple fitness model agrees that some facial qualities are universally physically attractive to people.

Apart from eliciting the facial characteristics which account for attractiveness, modern researchers have aimed to describe the mechanisms underlying these preferences. Many contributors refer to the evolutionary origins of attractiveness preferences (Andersson, 1994; Møller & Swaddle, 1997; Thornhill & Gangsted, 1999). According to this view, facial traits signal mate quality and imply chances for reproductive success and parasite resistance. Some evolutionary theorists suggest that preferred features might not signal mate quality but that the "good taste" by itself is an evolutionary adaptation (individuals with a preference for attractiveness will have attractive offspring that will be favored as mates) (Thornhill & Gangsted, 1999). Another mechanism explains attractiveness' preferences through a cognitive theory—a preference for attractive faces might be induced as a by-product of general perception or recognition mechanisms (Rubenstein et al., 2002; Zebrowitz & Rhodes, 2002): attractive faces might be pleasant to look at since they are closer to the cognitive representation of the face category in the mind. Halberstadt and Rhodes have further demonstrated that not just average *faces* are attractive but also birds, fish, and automobiles become more attractive after being averaged with computer manipulation (Halberstadt & Rhodes, 2003). Such findings led researchers to propose that as perceivers can process an object more fluently, aesthetic response becomes more positive (Reber, Schwarz, & Winkielman, 2004). A third view suggests that facial attractiveness originates in a social mechanism, where preferences may be dependent on the learning history of the individual and even on his social goals (Zebrowitz & Rhodes, 2002).

Other studies have used computational methods to analyze facial attractiveness. In several cases faces were averaged using morphing tools (e.g. Perrett, May, & Yoshikawa, 1994; Rubenstein et al., 2002). Laser scans of faces were put into complete correspondence with the average face in order to examine the relationship between facial attractiveness, age and averageness (ÓToole, Price, Vetter, Bartlett, & Blanz, 1999). A genetic algorithm, guided by interactive user selections was programmed to evolve a "most beautiful" female face (Johnston & Franklin, 1993). Machine learning methods have been used recently to investigate whether a machine can predict attractiveness ratings by learning a mapping from facial images to their attractiveness scores (Eisenthal, Dror, & Ruppin, 2006). The latter predictor achieved a correlation of 0.6 with average human ratings, demonstrating that facial beauty can be learned by a machine, at least to some moderate extent. However, as human raters significantly outperform the predictor of Eisenthal et al., the challenge of constructing a facial attractiveness machine predictor with human-level accuracy has remained open.

A primary goal of this study is to surpass these results by developing a machine which obtains human-level performance in predicting facial attractiveness and, thus, passes what Kurzweil calls a *subject matter expert turing test* (SME TT) (Kurzweil, 2005). Having accomplished this, our second main goal is to conduct a series of simulated psychophysical experiments and study the resemblance between human and machine judgments. This latter task carries two potential rewards: first, to determine whether the machine can aid in understanding the psychophysics of human facial attractiveness, capitalizing on the ready accessibility of manipulating and studying its performance, and second, to study whether learning an explicit operational ratings prediction task also entails learning implicit human-like biases, at least for the case of facial attractiveness.

In the past decades machines have achieved human-level performance in rule-based systems such as playing games (Schaeffer & Herik, 2002) and in various expert systems (Slezak, 1991). Impressive progress has been displayed in simulating various tasks which involve face perception, such as face detection (Hjelmas & Low, 2001), face recognition (Becker, 1999; Zhao, Chellappa, Rosenfeld, & Phillips, 2000) and tasks of facial category learning such as emotion (Dailey, Cottrell, Padgett, & Adolphs, 2002) and gender (Graf, Wichmann, Bülthoff, & Schölkopf, 2006) recognition. The task of evaluating human attractiveness ratings adds the notion of *judgment of taste* to the previous achievements in machine perception of faces. Learning the concept of facial attractiveness could form an important demonstration of a computer's ability to learn to master a quantitative, basic, human judgment task.

To this end we have collected human scores of facial attractiveness for a given dataset of female facial images. We developed an algorithm for automatic extraction of a very large set of geometric facial features, which, combined with a set of global features, yields a principled representation of each facial image via a set of image-features in an appropriate dimension-reduced space. Using this data of facial representations and their associated rating scores, we have employed standard supervised learning algorithms to construct a facial attractiveness prediction machine. Given a new, unseen face, this machine predicts its human attractiveness score in an accurate manner. We then turned to performing a series of simulated psychophysical experiments, modeled after known experiments in the psycholog-

ical literature, to study the resemblance between human and machine preferences. These experiments are particularly interesting since the machine is trained on an explicit operational ratings prediction task with no defined instructions specifying the human-like biases in question.

2. Materials and methods

2.1. Rating the facial database

The chosen database was composed of 91 facial images of American females, taken by the Japanese photographer Akira Gomi. All 91 samples were frontal color photographs of young Caucasian females with a neutral expression. All samples were of similar age, skin color and gender. The subjects' portraits had no accessories or other distracting items such as jewelry. We focused on female faces since experimental results shows that there is a greater agreement on human ratings of female faces while male face preferences are more largely influenced by the menstrual cycle and self-perceived attractiveness of the raters (Little et al., 2002). All 91 facial images in the dataset were rated for attractiveness by 28 human raters (15 males, 13 females) on a 7-point Likert scale (1 = very unattractive, 7 = very attractive). Ratings were collected with a specifically designed html interface. Each rater was asked to view the entire set before rating in order to acquire a notion of attractiveness scale. There was no time limit for judging the attractiveness of each sample and raters could go back and adjust the ratings of previously rated samples. The images were presented to each rater in a random order and each image was presented on a separate page. The final attractiveness rating of each sample was its mean rating across all raters. To validate that the number of ratings collected adequately represented the "collective attractiveness rating" we randomly divided the raters into two disjoint groups of equal size. For each facial image, we calculated the mean rating on each group, and calculated the Pearson correlation between the mean ratings of the two groups. This process was repeated 1000 times. The mean correlation between two groups was 0.92 ($\sigma = 0.01$). It should be noted that the split-half correlations reported were high in all 1000 trials (as evident from the low standard deviation) and not only over the average. This correlation corresponds well to the known level of consistency among groups of raters reported in the literature (e.g. Cunningham et al., 1995). Hence, the mean ratings collected are stable indicators of attractiveness that can be used for the learning task. The facial set contained faces in all ranges of attractiveness. Final attractiveness ratings range from 1.42 to 5.75 and the mean rating was 3.33 ($\sigma = 0.94$).

2.2. Data preprocessing and representation

Preliminary experimentation with various ways of representing a facial image (e.g. Eisenthal et al., 2006) have systematically shown that features based on measured proportions, distances and angles of faces are most effective in capturing the notion of facial attractiveness. To extract facial features we developed an automatic engine that is capable of identifying eyes, nose, lips, eyebrows and head contour. In total, we measured 84 coordinates describing the locations of those facial features (Fig. 1). Several regions are automatically suggested for sampling mean hair color, mean skin color and skin texture. The feature extraction process was basically automatic but some coordinates needed to be manually adjusted in some of the images. The facial coordinates are used to create a *distances-vector* of all 3486 distances between all pairs of coordinates in the complete graph created by all coordinates. For each image, all distances are normalized by face length (as measured from the coordinate at the top of the forehead to the coordinate at the bottom of the chin). In a similar manner, a *slopes-vector* of all the 3486 slopes of the lines connecting the facial coordinates is computed. Central fluctuating asymmetry (CFA) is calculated from the coordinates as well. CFA corresponds to the sum of the absolute values of the differences of the midpoints of adjacent horizontal lines which connect matching bilateral facial coordinates

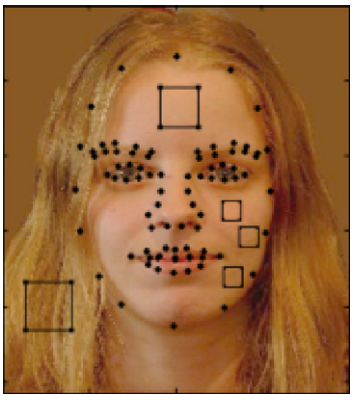

Fig. 1. Facial coordinates with hair and skin sample regions as represented by the facial feature extractor. Coordinates are used for calculating geometric features and asymmetry. Sample regions are used for extracting color values and smoothness. (The sample image, used for illustration only, is of T.G. and is presented with her full consent.)

(see Grammer & Thornhill, 1994). The application also provides, for each face, hue, saturation and value (HSV) values of hair color and skin color, and a measurement of skin smoothness. Smoothness of skin was calculated with an edge-detection algorithm in which many detected edges suggest a low level of skin smoothness. Combining the distances-vector and the slopes-vector yields a vector representation of 6972 *geometric features* for each image. Since strong correlations are expected among the features in such representation, principal component analysis (PCA) was applied to the 6972 geometric features, producing 90 principal components that span the sub-space defined by the 91 image vector representations. The geometric features are projected on those 90 principal components to produce 90 decorrelated *eigenfeatures* representing the geometric features of the images. Eight non-geometric measured features were not included in the PCA analysis, including CFA, smoothness, hair color coordinates (HSV) and skin color coordinates. These features are assumed to be directly connected to human perception of facial attractiveness and are hence kept at their original values. These 8 features were added to the 90 geometric eigenfeatures, resulting in a total of 98 *image-features* representing each facial image in the dataset.

2.3. Predictor construction and validation

We experimented with several induction algorithms including simple Linear Regression, Least Squares Support Vector Machine (LS-SVM) both linear as well as non-linear (Suykens, Van Gestel, De Brabanter, De Moor, & Vandewalle, 2002) and Gaussian Processes (GP) (Rasmussen & Williams, 2006). However, as the LS-SVM and GP showed no substantial advantage over Linear Regression, the latter was used and is presented in the sequel.

A key ingredient in our method is to use a proper image-features selection strategy. To this end we used subset feature selection, implemented by ranking the image-features by their Pearson correlation with the target. Other ranking functions produced no substantial gain. To measure the performance of our method we removed one sample from the whole dataset. This sample served as a test set. We found, for each left out test sample, the optimal number of image-features by performing leave-one-out-cross-validation (LOOCV) on the remaining samples and selecting the number of features that minimized the absolute difference between the algorithm's output and the targets of the training set. Ranking of features was conducted independently for each held out image and performance was measured by aggregating together the scores of all images. In other words, while setting aside a test sample, we used LOOCV on the remaining training samples in order to optimize the number of features to select, m_i, and afterwards used this number of features to predict a single, fixed attractiveness score for the left out test sample, that is, the score for a test exam-

ple was predicted using a single model based on the training set only. This process was repeated $n = 91$ times, once for each image sample, resulting with a vector of attractiveness predictions for all images. In order to avoid overfitting, the entire learning procedure (including feature selection) is repeated from scratch for each data partition, so that a different number of features are selected for each data partition. The number of selected features, m_i, ranges between 50 and 77. $m_i = 67$ features were most frequently selected. To examine the influence of resetting the number of selected features at each fold, we tested the predictor in a leave-one-out-cross-validation, on the entire dataset, while keeping the number of selected features constant in all iterations (and not reselecting it each time). The fixed number of selected features ranged between $m = 1$ (a single feature) and $m = 98$ (all features). The best Pearson correlation of 0.87 was achieved with $m = 64$. These 64 features include 7 of the 8 non-geometric features (all but Hair-hue). The remaining 57 geometric eigenfeatures explain 96% of the variance of the geometric features. For all $54 \leqslant m \leqslant 74$, Pearson and Spearman correlations were above 0.8. As is evident, a fixed number of selected features can yield better performance than the one our method of reselection produced. Still, we did not use a fixed value for the number of selected features, m, since we did not want to rely on test performance when choosing the constant value of m in order to avoid overfitting. It should be noted that we tried to use the same feature selection and training procedure with the original geometric features (without PCA) instead of using the eigenfeatures. This, however, has failed to produce good predictors due to strong correlations between the original geometric features (the maximal Pearson correlation obtained was 0.26).

Once the predictor was constructed and validated, we turn to simulate a number of psychophysical experiments that were previously conducted with human subjects.

2.3.1. Experiment 1

We created virtual face composites for the machine to rate by simulating a morphing technique similar to the one used by Rubenstein et al. (2002). Coordinate values of the original component faces were averaged to create a new set of coordinates for the virtual face composite. These coordinates were used to calculate the geometrical features and CFA of the averaged face. Smoothness and HSV values for the composite face were calculated by averaging the corresponding values of the component faces (HSV values are converted to RGB before averaging). To study the effect of the number of component faces, n_c, on the attractiveness score of face composites we produced 1000 virtual morph images for each value of n_c between 2 and 50, and used our attractiveness predictor to compute the attractiveness scores of the resulting composites.

2.3.2. Experiment 2

In order to further examine the importance of symmetry on the machine's attractiveness judgments of averaged composites, we repeated the virtual composites experiment (Experiment 1) using perfectly symmetric faces as image components. Perfectly symmetric virtual versions of the original images were created by a similar technique to the one used by Rhodes, Sumich, and Byatt (1999), that is, each original face was virtually morphed with its mirror image in order to create a perfectly symmetric version of it. Averaging together perfectly symmetric component faces produces perfectly symmetric face composites. In the same manner as in Experiment 1, 1000 virtual composites were created for each number of components, n_c, between 2 and 50, and the machine rated them for attractiveness.

2.3.3. Experiment 3

Analogously to Zaidel, Chen, and German (1995) who have created chimeric facial composites by attaching one-half of the face to its mirror image, Right–right and left–left virtual chimeric composites were produced from the extracted coordinates of all original images and the machine was used to predict their attractiveness ratings. Learning was repeated for each chimeric composite with the original image used for each chimeric composition being excluded from the training set, to avoid a misleading positive bias as a consequence of the fact that the original image contains many features which are identical to those of the matching composite.

2.4. Facial features and attractiveness

The original measured facial features were ranked according to their correlation to human and machine ratings and the Spearman rank correlation between the two rankings was calculated. This analysis was repeated three times: (a) with 6980 predictor-features, (b) with 28 features investigated in previous studies and (c) with 13 features previously found to be significantly related to facial attractiveness. To determine the P-value of the rank correlation we repeated the features ranking 100 times with shuffled machine and human ratings. In none of the shuffled trials was the rank correlation as high as with the actual ratings. The 28 features we focused on are (features marked with an * were previously found to be significantly correlated with facial attractiveness): (1) forehead height, (2) eye height*, (3) eye width*, (4) separation of eyes*, (5) nose tip width, (6) nostril width*, (7) nose length (to eye top), (8) nose area*, (9) upper lip thickness, (10) lower lip thickness, (11) chin length*, (12) cheekbone width*, (13) jaw (cheek) width*, (14) mid-face length, (15) eyebrow height*, (16) mouth width* (features 1–16 are taken from Cunningham, 1986), (17) outer eye corner width*, (18) inner eye corner width, (19) cheek width, (20) cheekbone prominence*, (21) lower face proportions (features 17–12 are taken from Grammer & Thornhill, 1994), (22) forehead height (to eyes), (23) brow height, (24) brow curvature, (25) lower face length, (26) nose length, (27) mouth height*, (28) cheekbone height (features 22–28 are taken from Grammer, Fink, Juette, Ronzal, & Thornhill, 2002). (See *Supporting Information* for a detailed description of the calculation of these 28 feature measurements according to the raw coordinate representation.)

3. Results

3.1. Prediction accuracy of facial attractiveness

Machine attractiveness ratings of all sample images obtained a high Pearson correlation of 0.82 (P-value $< 10^{-23}$) with the mean ratings of human raters (the learning targets), corresponding to a normalized mean squared error of 0.39. This accuracy is a marked improvement over the recently published performance results of a Pearson correlation of 0.6 on a similar dataset (Eisenthal et al., 2006). The average correlation of an individual human rater to the mean ratings of all other human raters in our dataset is 0.67 and the average correlation between the mean ratings of groups of human raters is 0.92 (see Section 2). The Spearman rank correlation between machine and mean human ratings is 0.83. To further validate the correlation measures we removed the most attractive 12% and the least attractive 12% of the samples from the dataset and recalculated the correlations. Correlation values remained high with 0.80 (Pearson) and 0.81 (Spearman). To get a notion of the contribution of the 8 global, non-geometric features to attractiveness prediction, we have trained the predictor while removing them one at a time. This resulted in correlations of 0.68 when excluding asymmetry, 0.80 when excluding smoothness, 0.77 when excluding hair color (3 attributes) and 0.77 when excluding skin color (3 attributes). Excluding all non-geometric features and using geometric features alone yielded a correlation of 0.74.

3.2. Similarity of machine and human judgments

The ratings of each rater (28 human raters and the machine predictor) form a 91 dimensional *rating vector*

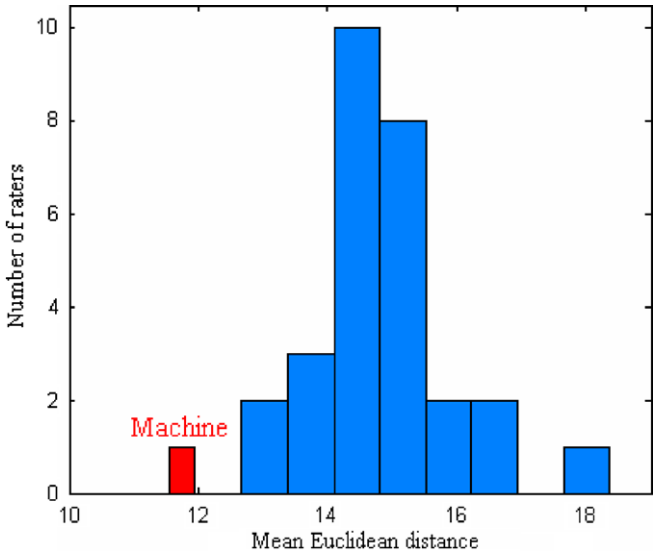

Fig. 2. Distribution of mean Euclidean distance from each human rater to all other raters in the *ratings space*. The machine's average distance from all other raters (left bar) is smaller than the average distance of each of the human raters to all others.

describing its attractiveness ratings of all 91 images. These vectors can be embedded in a 91 dimensional *ratings space*. The Euclidean distance between all raters (human and machine) in this space was computed. Compared with each of the human raters, the ratings of the machine were the closest, on average, to the ratings of all other human raters (Fig. 2).

Although, by construction, the machine's rating vector lies near the mean of human ratings, it may still be very different from any individual human rating vector. This may happen, e.g. when the distribution of human ratings forms several clusters or is non-convex. To assure this is not the case, we counted the number of human ratings vectors within small multidimensional spheres around each human rater as well as the rating of the machine. The machine had more human neighbors than the mean number of neighbors that a human rater had, even when the radiuses of the spheres were very small, testifying that it does not fall between clusters. Finally, to visualize the machine ratings among human ratings we applied PCA to machine and human ratings in the rating space and projected all ratings onto the resulting first 2 and 3 principal components. Indeed, the machine is well placed in a mid-zone of human raters (Fig. 3).

3.3. Human-like biases in the machine's performance

3.3.1. Experiment 1: The averageness hypothesis: A preference for averaged face composites

Rubenstein et al. (2002) discuss a morphing technique to create mathematically averaged faces from multiple face images. They report that averaged faces made of 16 and 32 original component images were rated by humans higher in attractiveness than the mean attractiveness ratings of their component faces and higher than composites consisting of fewer faces. In their experiment, 32-component composites were found to be the most attractive. In accordance with these experimental results, the predictor manifests a human-like bias for higher scores for averaged composites over their components' mean score. Fig. 4a shows the percent of components which were rated as less attractive than their corresponding composite, for each number of components n_c. As evident, the attractiveness rating of a composite surpasses a larger percent of its components' ratings as n_c increases. Fig. 4a also shows the mean scores of 1000 composites and the mean scores of their components, for each n_c (scores are normalized to the range $[0,1]$). Their actual attractiveness scores are reported in Table 1. As expected, the mean scores of the component images are independent of n_c, while composites' scores increase with n_c (see *Supporting Information* for a more detailed analysis of the difference between composites and components scores).

Recent studies have provided evidence that skin texture influences judgments of facial attractiveness (Fink, Grammer, & Thornhill, 2001). Since blurring and smoothing of faces occur when faces are averaged together (Rubenstein et al., 2002), the smooth complexion of composites may underlie the attractiveness of averaged composites. In our experiment, a preference for averageness is found even though our method of virtual-morphing does not produce the smoothening effect and the mean smoothness value of composites corresponds to the mean smoothness value in the original dataset, for all n_c (see Fig. 4b). Researchers have also suggested that averaged faces are attractive since they are exceptionally symmetric (Alley & Cunningham, 1991). Fig. 3a and b shows that the mean level of asymmetry (CFA, see Section 2) is indeed highly correlated with the mean scores of the composites (Pearson correlation of -0.91, P-value $< 10^{-19}$). However, examining the correlation between the rest of the image-features and the composites' scores reveals that this high correlation is not at all unique to asymmetry. In fact, as the images are being morphed, the changes in 45 of the 98 image-features are strongly correlated with the changes in attractiveness scores ($|\text{Pearson correlation}| > 0.9$). The high correlation between these numerous features and attractiveness scores of averaged faces indicates that symmetry level is not an exceptional factor in the machine's preference for averaged faces. Instead, it suggests that averaging causes many features to change in a direction which causes an increase in attractiveness.

It has been argued that although averaged faces are found to be attractive, very attractive faces are not average (Alley & Cunningham, 1991). A virtual composite made of the 12 most attractive faces in the set (as rated by humans) was rated by the machine with a high score of 5.6 while 1000 composites made of 50 faces from random levels of attractiveness got a *maximum* score of only 5.3. (Their mean score was only 3.94 as reported in Table 1.) This type of preference resembles the findings of Perrett et al. (1994)

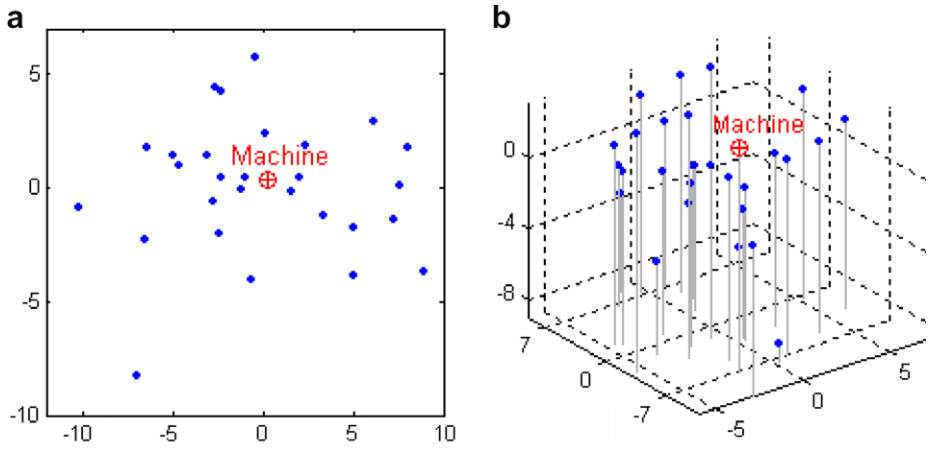

Fig. 3. Location of machine ratings among the 28 human ratings. Ratings were projected into 2 dimensions (a) and 3 dimensions (b) by performing PCA on all ratings and projecting them on the first principal components. The projected data explain 29.8% of the variance in (a) and 36.6% in (b).

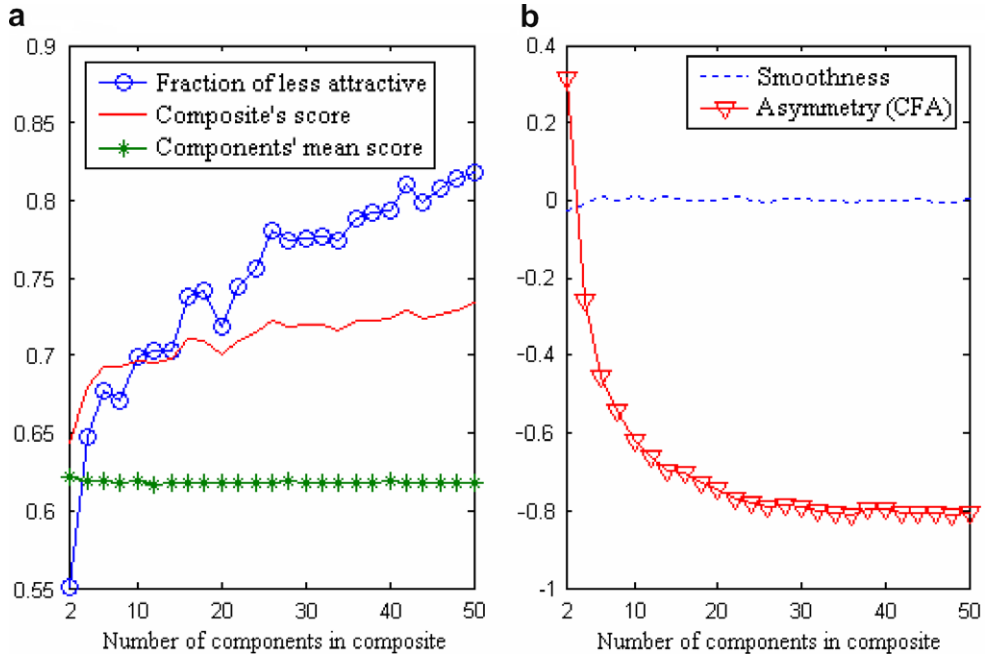

Fig. 4. Experiment 1: The averageness hypothesis: (a) percent of components that were rated as less attractive than their corresponding composite, accompanied with mean scores of composites and the mean scores of their components (scores are normalized to the range [0, 1], actual attractiveness scores are reported in Table 1). (b) Mean values of smoothness and asymmetry of 1000 composites for each number of components, n_c.

Table 1
Mean results over 1000 composites made of varying numbers of component images

Number of components in composite	Mean composite score	Mean components score	Components rated lower than composite (%)
2	3.46	3.34	55
4	3.66	3.33	64
12	3.74	3.32	70
25	3.82	3.32	75
50	3.94	3.33	81

in which a highly attractive composite, morphed from only attractive faces, was preferred by humans over a composite made of 60 images of all levels of attractiveness.

3.3.2. Experiment 2: Perfectly symmetric averaged faces

Rhodes et al. (1999) inquired whether changes in attractiveness produced by manipulating the averageness of individual faces should disappear when all the images are made perfectly symmetric. They created perfectly symmetric composites by morphing original images with their matching mirror images. In their experiment human subjects showed a preference for averaged face composites even when the effect of symmetry is controlled for. Similarly, in our experiment, the effect of symmetry was neutralized by using only perfectly symmetric component faces which yielded perfectly symmetric composites (see Fig. 5b). It can be seen that the results presented in Fig. 5a are similar to those of Experiment 1 (Fig. 4a). That is, even though the

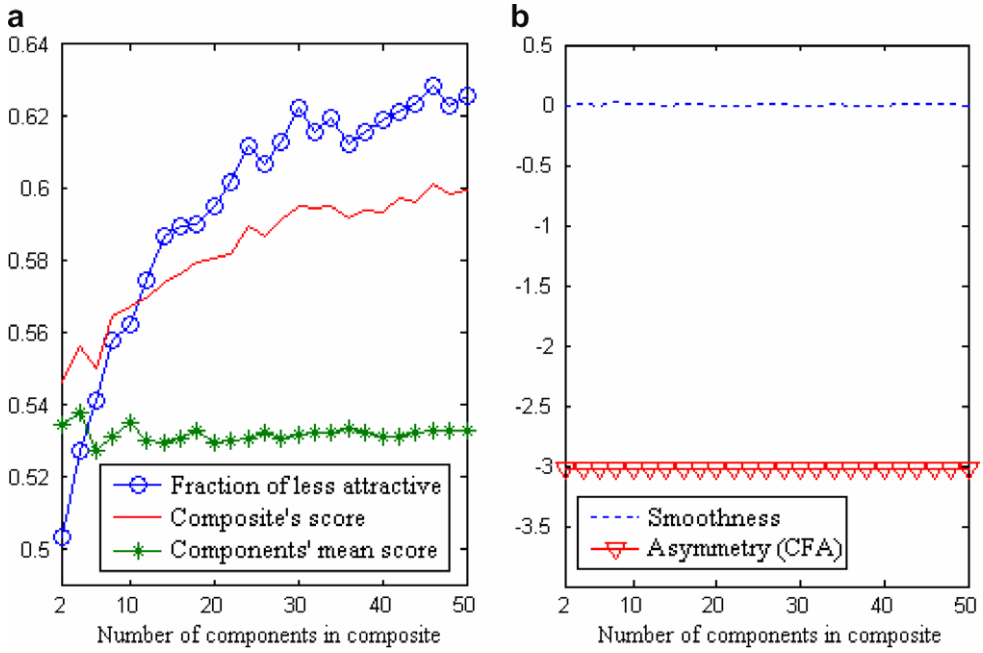

Fig. 5. Mean results over 1000 perfectly symmetric composites made of varying numbers of perfectly symmetric image components: (a) percent of components which were rated as less attractive than their corresponding composite, accompanied with mean scores of composites and the mean scores of their components (scores are normalized to the range [0, 1]). (b) Mean values of smoothness and asymmetry of 1000 composites for each number of components, n_c.

effect of symmetry is controlled for, attractiveness scores of averaged face composites increases with the number of components, n_c. Mean values of smoothness and asymmetry of the composites are presented in Fig. 5b. These results show that the machine's preference for averaged composites is not dependent on symmetry alone, in accordance with the experimental results of Rhodes et al. (1999) and with our conclusions from Experiment 1 (see *Supporting Information* for a more detailed analysis of the difference between perfectly symmetric composites and components scores).

3.3.3. Experiment 3: Asymmetry of facial attractiveness perception

A recent study examining the asymmetry of *attractiveness perception* has offered an intriguing relationship between facial attractiveness and hemispheric specialization (Zaidel et al., 1995). In this research, right–right and left–left chimeric composites (where 'left' refers to the subject's side of the face) were created by attaching each half of the face to its mirror image. Human subjects were asked to look at left–left and right–right composites of the same image and judge which one is more attractive. For women's faces, right–right composites, composed of the right half of the subject's face, got twice as many 'more attractive' responses than left–left composites. Interestingly, similar results to those were found in Experiment 3 in which we simulated this phenomenon by comparing the machine's rating of facial attractiveness for left–left and right–right composites. The machine gave 63 out of 91 right–right composites a higher rating than their matching left–left composite, while only 28 left–left composites were judged as more attractive. A paired *t*-test shows these results to be statistically significant with P-value $< 10^{-7}$ (scores of chimeric composites are approximately normally distributed). When rating composites created from a certain image the machine was trained without the original image in its training set. Since the machine representation of the images is completely symmetric, any asymmetric bias revealed is likely to be an implicit manifestation of a psychophysical bias of the human raters. It is interesting to see that the machine manifests the same kind of asymmetry bias reported by Zaidel et al. (1995), though it has never been explicitly trained for that.

3.4. Facial features and attractiveness

After establishing that our machine exhibits human-like biases, we turn to compare its processing with those reported in the pertaining human psychophysics literature. A number of studies have singled out facial features that are especially relevant to facial attractiveness, by identifying significant correlations between facial features measurements and human attractiveness ratings (Cunningham et al., 2002; Grammer & Thornhill, 1994; Little et al., 2002). In analogy, facial features that are significantly correlated with the machine's ratings may be considered as important in determining the machine's perception of attractiveness. In order to examine whether the important features according to the machine are similar to those of humans, we calculated the correlation between each of the 6980 features used by our predictor (6972 geometric

features and 8 non-geometric measurements) and the machine and human ratings (separately). The features were ranked according to their absolute correlation to attractiveness ratings which resulted with two feature rankings: human and machine. The Spearman rank correlation between human and machine ranking was 0.57 and significant (P-value < 0.01, see Section 2). To further compare between feature rankings of humans and machine, we repeated the above computation focusing on 28 facial features which were previously studied in the literature of human facial attractiveness (see Section 2). Those 28 features were now ranked according to machine and human ratings and the Spearman rank correlation between the two rankings was 0.68 (P-value < 0.01). Out of those 28 features, only 13 were found to be significantly related to facial attractiveness in the original studies (Cunningham et al., 2002; Grammer & Thornhill, 1994; Little et al., 2002). Ranking these 13 facial features according to machine and human ratings, yields a Spearman rank correlation of 0.75 (P-value < 0.01) between the rankings. These results provide further evidence of the human-like nature of the machine's perception of attractiveness, as they show that features that were previously related to facial attractiveness are ranked similarly according to the machine and according to human raters.

4. Discussion

In this work, we constructed a high quality training set for learning facial attractiveness of human faces. Using a combination of extensive automatic facial feature extraction, dimension reduction and feature selection, and supervised learning methodologies; we created the first accurate facial attractiveness predictor. Our results add the task of facial attractiveness prediction to the collection of abstract tasks that have been successfully accomplished with current machine learning techniques. While previous machines that successfully passed a subject matter expert turing test (SME TT) have dealt with rule-based cognitive systems, such as playing games, or perceptual tasks of category learning, such as emotion recognition, our machine predicts continuous facial attractiveness ratings and passes a perceptual SME TT that concerns simulating judgment of taste. Whether to compare the machine's performance in the task to the performance of an individual human rater or to a group of raters is an interesting issue: the machine is an 'individual rater' which learns 'group average ratings' and thus is essentially a hybrid between the two. For that reason we report on both benchmarks, that is, the human *individual-to-group* mean correlation of 0.67 and the human *group-to-group* mean correlation of 0.92, and indeed, we find the machine's performance (correlation of 0.82) between the two. One of the main improvements over previous similar works, such as the work of Eisenthal et al. (2006), is the much richer representation of 84 facial coordinates and 6972 distance and angle features (induced from the full graph on the facial coordinates). This suggests that improving the facial representation might be valuable for future research. One promising suggestion is to employ a non-metric facial representation which may expedite the learning of human facial attractiveness.

Examining the machine and human raters' representations in the ratings space identifies the ratings of the machine near the center of the distribution of human ratings, and closest, on average, to other human raters. The ranking of facial features according to their correlations with machine ratings is correlated significantly with the ranking of those features according to human ratings. The similarity between human and machine preferences has prompted us to further study the machine's operation. To this end, we have found that the machine favors averaged faces made of several component faces. While this preference is known to be common to humans as well, researchers have previously offered different reasons for favoring averageness. Our analysis has revealed that symmetry is strongly related to the attractiveness of averaged faces, but is definitely not the only factor in the equation, since about half of the image-features relate to the ratings of averaged composites in a similar manner as symmetry and since a preference for averaged faces was found even when the effect of symmetry was neutralized. This suggests that a general movement of features toward attractiveness, rather than a simple increase in symmetry, is responsible for the attractiveness of averaged faces. This movement suggests a convergence towards a prototypical facial representation that matches the cognitive explanations of the averageness hypothesis (Rubenstein et al., 2002). Obviously, this is true only for the machine, but given the human-like biases displayed by our predictor, this may extend also to human perception of facial attractiveness. Overall, it is quite surprising and pleasing to find that a model trained explicitly to capture a specific operational performance criteria such as attractiveness rating (weak AI), implicitly and concomitantly captures basic human psychophysical biases and demonstrates a wide range of human-level characteristics of facial attractiveness judgment (strong AI), as revealed by studying its ''psychophysics''.

Acknowledgments

We thank Dr. Bernhard Fink and the Ludwig-Boltzmann Institute for Urban Ethology at the Institute for Anthropology, University of Vienna, Austria, and Prof. Alice J. O'Toole from the University of Texas at Dallas, for kindly letting us use their face databases.

This work was supported by the internal research fund of The Academic College of Tel-Aviv-Yaffo.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.visres.2007.11.007.

References

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

Andersson, M. (1994). *Sexual selection*. Princeton, NJ: Princeton University Press.

Becker, S. (1999). Implicit learning in 3D object recognition: The importance of temporal context. *Neural Computation, 11*(2), 347–374.

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

Cunningham, M. R., Barbee, A. P., & Philhower, C. L. (2002). Dimensions of facial physical attractiveness: The intersection of biology and culture. In G. Rhodes & L. A. Zebrowitz (Eds.). *Advances in visual cognition, vol. 1: facial attractiveness*. Westport, CT: Ablex.

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 physical attractiveness. *Journal of Personality and Social Psychology, 68*, 261–279.

Dailey, M. N., Cottrell, G. W., Padgett, C., & Adolphs, R. (2002). EMPATH: A neural network that categorizes facial expressions. *Journal of Cognitive Neuroscience, 14*(8), 1158–1173.

Eisenthal, Y., Dror, G., & Ruppin, E. (2006). Facial attractiveness: Beauty and the machine. *Neural Computation, 18*, 119–142.

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

Galton, F. (1878). Composite portraits. *Journal of the Anthropological Institute of Great Britain and Ireland, 8*, 132–142.

Graf, A. B. A., Wichmann, F. A., Bülthoff, H. H., & Schölkopf, B. (2006). Classification of faces in man and machine. *Neural Computation, 18*, 143–165.

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

Grammer, K., Fink, B., Juette, A., Ronzal, G., & Thornhill, R. (2002). Female faces and bodies: N-dimensional feature space and attractiveness. In G. Rhodes & L. A. Zebrowitz (Eds.), *Advances in visual cognition, vol. 1: facial attractiveness*. Westport, CT: Ablex.

Halberstadt, J. B., & Rhodes, G. (2003). It's not just average faces that are attractive: Computer-manipulated averageness makes birds, fish, and automobiles attractive. *Psychonomic Bulletin and Review, 10*, 149–156.

Hjelmas, E., & Low, B. K. (2001). Face detection: A survey. *Computer Vision and Image Understanding, 83*, 236–274.

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

Kurzweil, R. (2005). *The singularity is near: When humans transcend biology*. Viking Penguin.

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

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.

Little, A. C., Penton-Voak, I. S., Burt, D. M., & Perrett, D. I. (2002). Evolution and individual differences in the perception of attractiveness: How cyclic hormonal changes and self-perceived attractiveness influence female preferences for male faces. In G. Rhodes & L. A. Zebrowitz (Eds.), *Advances in visual cognition, vol. 1: facial attractiveness*. Westport, CT: Ablex.

Møller, A. P., & Swaddle, J. P. (1997). *Asymmetry, developmental stability, and evolution*. Oxford: Oxford University Press.

ÓToole, A. J., Price, T., Vetter, T., Bartlett, J. C., & Blanz, V. (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.

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

Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian processes for machine learning*. The MIT Press. ISBN 0-262-18253-X.

Reber, R., Schwarz, N., & Winkielman, P. (2004). Processing fluency and aesthetic pleasure: Is beauty in the perceiver's processing experience? *Personality and Social Psychology Review, 8*, 364–382.

Rhodes, G., Sumich, A., & Byatt, G. (1999). Are average facial configurations attractive only because of their symmetry? *Psychological Science, 10*, 52–58.

Rubenstein, A. J., Langlois, J. H., & Roggman, L. A. (2002). What makes a face attractive and why: The role of averageness in defining facial beauty. In G. Rhodes & L. A. Zebrowitz (Eds.), *Advances in visual cognition, vol. 1: facial attractiveness*. Westport, CT: Ablex.

Schaeffer, J., & Herik, H. J. (2002). Games, computers, and artificial intelligence. *Artificial Intelligence, 134*, 1–7.

Slezak, P. (1991). Artificial experts: Essay review. *Social Studies of Science, 22*(1), 175–201.

Suykens, J. A. K., Van Gestel, T., De Brabanter, J., De Moor, B., & Vandewalle, J. (2002). *Least squares support vector machines*. World Scientific, Singapore. ISBN 981-238-151-1.

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

Zaidel, D. W., Chen, A. C., & German, C. (1995). She is not a beauty even when she smiles: Possible evolutionary basis for a relationship between facial attractiveness and hemispheric specialization. *Neuropsychologia, 33*(5), 649–655.

Zebrowitz, L. A., & Rhodes, G. (2002). Nature let a hundred flowers bloom: The multiple ways and wherefores of attractiveness. In G. Rhodes & L. A. Zebrowitz (Eds.), *Advances in visual cognition, vol. 1: facial attractiveness*. Westport, CT: Ablex.

Zhao, W. Y., Chellappa, R., Rosenfeld, A., & Phillips, P. J. (2000). Face recognition: A literature survey. UMD CfAR *Technical Report* CAR-TR-948.