Automatic Point-based Facial Trait Judgments Evaluation

Mario Rojas Q.[1], David Masip[1,2], Alexander Todorov[3], Jordi Vitrià[1,4]
[1]Computer Vision Center, Edifici O, Campus UAB, Spain
[2]Universitat Oberta de Catalunya, Rambla del Poblenou 156, 08018, Barcelona, Spain
[3]Department of Psychology, Princeton University, Princeton, New Jersey, USA 08540
[4]Department de Matematica Aplicada i Analisi, Universitat de Barcelona, Spain
mrojas@cvc.uab.es, dmasipr@uoc.edu, atodorov@princeton.edu, jordi.vitria@ub.edu

Abstract

Humans constantly evaluate the personalities of other people using their faces. Facial trait judgments have been studied in the psychological field, and have been determined to influence important social outcomes of our lives, such as elections outcomes and social relationships. Recent work on textual descriptions of faces has shown that trait judgments are highly correlated. Further, behavioral studies suggest that two orthogonal dimensions, valence and dominance, can describe the basis of the human judgments from faces. In this paper, we used a corpus of behavioral data of judgments on different trait dimensions to automatically learn a trait predictor from facial pixel images. We study whether trait evaluations performed by humans can be learned using machine learning classifiers, and used later in automatic evaluations of new facial images. The experiments performed using local point-based descriptors show promising results in the evaluation of the main traits.

1. Introduction

One of the most exciting research areas in the field of multi-modal intelligent interfaces is the design of better "human-centered" modules of interaction with machines. In this context, the processing of facial images has become an important research topic where automatic face classification techniques have been explored in a wide spectrum of applications ([2, 5, 6]). Automatic evaluation of faces is a human mechanism that evolved influenced by the process of assessment of potential threats which has thus shaped the saliency of facial cues. It is a mechanism that is to some extent capable of delivering rapid inferences about a variety of trait dimensions[1], which can be related to important social outcomes. Oosterhof and Todorov mention in [11] how trait judgments intervene in the assessment of face trustworthiness. Humans rapidly perform these trait judgments from faces and the results of this unconscious facial evaluations can determine the results of important social events, such as an electoral process [1, 9] or courts' sentences [3].

In [11], Oosterhof and Todorov identify the most important underlying trait dimensions used by humans to evaluate faces. They show that there is a large degree of correlation among these trait dimensions, and using a data-driven approach from behavioral data they determine that a 2D model suffices to explain the behavioral model of face evaluation. The two orthogonal dimensions were denominated *valence* and *dominance*, and constitute the basis that accounted for more than 80% of the variance of the behavioral data collected.

The objectives of the current research are twofold. Firstly study which traits can be computationally learned based on a limited amount of information. That is, in contrast to humans, that use a significant amount of visual data to assess a face, we intend to use the information contained in the structure of a scarce number of facial salient points to determine the discriminability of facial traits. Secondly perform a subject oriented evaluation with respect to a predefined set of studied traits, deriving an *automatic predictor* of the human inferences performed on pixel images.

The remainder of the document is organized as follows. Section 2 describes the relevance of trait judgments and the studies supporting the research. In section 3 the automatic facial salient point detection algorithm is outlined. The details of the methodology of the study are given in section 4 and the experiments performed as well as the results obtained are described in section 5. Section 6 concludes this paper.

[1]Strong evidence supporting this fact is still required in order to prove the accuracy of these inferences

978-1-4244-6985-7/10/$26.00 ©2010 IEEE

2. Facial Traits Judgments

In this paper, we base the automatic learning of the facial trait judgments evaluations on the behavioral data collected in [11]. This section briefly describes the facial trait judgments collected, the databases used and the acquisition protocol. In a first step, the facial trait dimensions were identified in an experiment involving 55 undergraduate students from Princeton University. Each student wrote an unconstrained description from a set of 66 standardized faces from the Karolinska [10] amateur actors face database. 1134 descriptions were collected, and two researchers independently classified the attributes from the descriptions into broad categories (discrepancies were solved by a third party). The reseachers' classification of the unconstrained descriptions, resulted in 14 selected categories.

In a second step, the 66 faces were rated on a continuous scale by a separate group of 327 participants based on their first impression, and faces were presented three times in separate blocks. The question "How [trait] is that person?" was presented altogether with the centred face, and a response (in the range 1 to 9) was to be given.

A data-driven model for the evaluation of facial trait inferences was built. Subsequently a PCA resulted in two prevalent orthogonal dimensions accounting for over 80% of the data variance[2] denominated *valence* and *dominance* respectively that we included in our study.

In a third step, a synthetic face database was generated using the FaceGen software (`http://facegen.com`). The software used a statistical model based on a large set of 3-D lasers scans of real faces, where the shape of each face is represented as a mesh of 3-D vertices. A Principal Component Analysis was performed on these coordinates, preserving the 50 components that account for most of the data variance. Faces were randomly generated using this model, where small changes on each PCA coefficient produce holistic changes on the vertex coordinates of the face image. Images were randomly generated bounding the software to generate Caucasian faces with neutral expression. A total of 300 synthetic faces were generated and rated following the protocol defined above. We used this data set as a starting point in our experiments, evaluating the following traits: Attractive, Competent, Trustworthy, Dominant, Mean, Frightening, Extroverted, Threatening and Likeable.

The main motivation of this study is to analyze the pixel images in terms of human trait inferences. We tackle the problem of predicting trait judgments from images using classic feature extraction and machine learning classifiers. The behavioral data collected in [11] is used to learn the set of image features that determine each trait judgment.

[2] According to [11], the third PC accounted for less than 6% of the data variance and had no clear interpretation

3. Facial Point Detection

An automatic salient facial point detector was derived following a similar preprocessing scheme to the one in [13]. The detector comprises 3 phases: (i) the face detection stage, where the Viola and Jones face detector [12] has been directly run in the face database; (ii) the detection of the facial zones, i.e. eyes and mouth zones; and (iii) the detection of the salient points inside each zone.

3.1. Eye Detection

Once the region of the face is located, we implemented an eye detector that is used to compute the rotation angle and the scaling factor of the face. Subsequently we normalize the image so that the interocular distance is constant among the images, and the line joining the eyes has an angle of zero degrees with the horizontal. Taking advantage of the bounding box retrieved by the face detector,we search the coordinates of the center of the eye in the upper half of the face. In the training stage, the coordinates of the center of each eye are used to extract a patch around their immediate neighborhood as positive examples and to generate randomly distributed patches around two different radii from the center. The extracted data patches are normalized and projected to the Non-parametric Discriminant Analysis subspace (see [4] for more details on NDA) before they are passed to the Gentleboost classifier [7]. To generate a more robust detector a hierarchical cascaded approach is followed, discarding first the patches that are less likely to be of the eye class i.e. those with more uniform intensity. We used 630 images from the Cohn Kanade database [8] to train the eye detection scheme, using patches of size 22×34 pixels to cover the entire region of the eye. We followed a cross validation protocol using the 630 images, and we found that the percentage of eye coordinates with normalized error[3] smaller than 20% represents over 96% of the samples.

3.2. Detection of the Facial Zones and Salient Points

The next step is the localization of the mouth. First we use the horizontal coordinates of the detected eyes and the bottom half of the facial bounding box to generate a search area covering the nostrils and the mouth. Within this region, a combination of a horizontal edge map and the integral image are used to obtain the coordinates of the center point of the mouth.

These three reference points (eyes and mouth coordinates) serve as a base to heuristically generate 17 regions that constitute the center of the searching areas for the salient points (as shown in figure 1). For each salient point a 300 dimensional NDA projected descriptor is generated

[3] Computed as the difference between the detected coordinates and the ground truth, normalized by the distance between the two eyes

from the intensity patches (which are of the same size of the eye patches), to train a GentleBoost region classifier. The performance of the system on the synthetic database (using 200 images) is superior to 87.5% requiring less than 13% of the points to be manually adjusted.

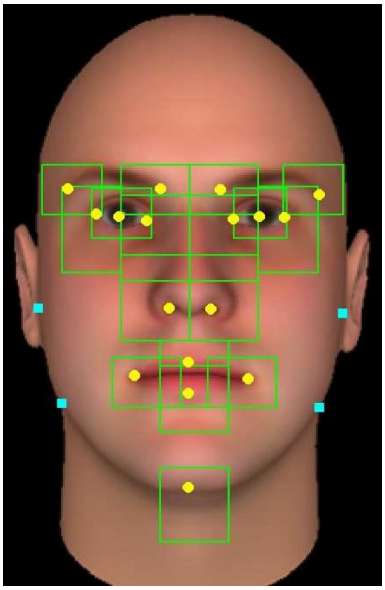

Figure 1. Synthetic image with points used to describe the facial structure and search areas labeled. The 17 round markers are automatically detected by the algorithm described in section 3

4. Automatic Facial Traits Evaluation

Judgment of faces as mentioned in [11] is based on subtle variations of the facial features in a holistic frame i.e. it is not constrained by assumptions about specific facial features. Based on this idea we used potential salient facial points and used their spatial relations to establish a facial structure descriptor built on three types of relations.

4.1. Image Descriptors

Twenty one predefined points $\mathbf{P} = \mathbf{p}_1, \ldots, \mathbf{p}_{21} \in R^2$ from each face are detected following the process defined in the previous section, and the mean coordinate values $\mathbf{M} = \mathbf{m}_1, \ldots, \mathbf{m}_{21} \in R^2$ of the database are computed (figure 1). Using this information a 1134-dimensional structural feature vector of the face is computed as follows:

1. The first 42 values of the descriptor consist of the difference of each point $\mathbf{p}_i$ to its corresponding mean $\mathbf{m}_i$ $(i = 1, \ldots, 21)$. In order to extract more information on the difference the computation is done in polar coordinates, and the values for angle and radius are stored.

2. The second set encodes the spatial relations between each salient point $\mathbf{p}_i$ of the face and all the points of the mean face image $\mathbf{m}_i$ $(i, j = 1, \ldots, 21)$ in terms of radius and angle, hence generating a $21\text{x}(21+21)$,882 dimensional sub vector.

3. The third set encodes the intra face structural relationships, and consists of 210 values with the euclidean distances of each point $\mathbf{p}_i$ to all the other points in the same image $\mathbf{p}_j$ $(i, j = 1, \ldots, 21)$.

4.2. Evaluation Framework

In order to establish appropriately the potentiality of the salient points as cues for facial trait judgment, an adequate framework is to be set up. We performed two different experiments, the first one dealing with controlled data synthetically acquired using the FaceGen Software (the images are frontal expressionless portraits of the faces against a black background), and the second one using real pixel images from the Karolinska face database. The selection of these data sets has been motivated by the availability of annotated ground truth data on different trait judgments. We have, for each face image, analyzed a label vector where a floating-point score is given for each facial trait judged. In addition, we also used the first and second principal component of the traits information as suggested in [11].

It should be pointed out that given the nature of the domain of the scores used to describe the different traits, a protocol to fit the problem into a binary class had to be derived. In this case, from the sorted scores for each trait, only one third of the highest and one third of the lowest values were kept and labeled as being or not from each class. This binarization of the continuous data represented in the numerical scores ensures that the bulk of values that could be mislabeled did not introduced noise in the system.

Moreover, the experiments are performed using the 1134-dimensional structural feature vectors defined above for each face. In previous studies, we evaluated the salient point detection system with different descriptors, obtaining from NDA the best results. In order to increase the reliability of the experiments, we trained a bank of classifiers to determine the performance of the afore mentioned descriptors. More specifically, the GentleBoost classifier [7] (trained with 100 rounds), the SVM using a Radial Basis Function[4] and finally a Knn (K parameters 3 and 5) are used in the experiments.

In addition, and given the high dimensionality of the structural descriptors, we replicated the experiments after performing a Principal Component Analysis (PCA) feature extraction. From the analysis of a subset of the data, it was established that the first 50 eigen vectors kept over 99% of the data variance information, thus the same figure was used as a basis for posterior analysis.

[4] Parameters σ and C were set by cross validating the training set.

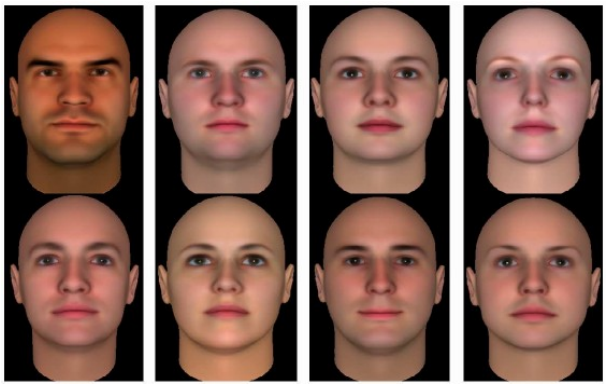

Figure 2. Faces portraying from left to right the highest and lowest rated samples of the traits with best and worst accuracy. Top row trait: Dominant, Bottom row trait: Competent

5. Experiments and Results

The initial evaluation was done on the database of 300 synthetic faces. The system was tested using a 10 fold cross validation scheme, where the database was divided in 10 sets so each time, 9 were used to train and the remaining one to test.

Table 1 shows the mean results of the three classification rules for all 9 traits and the first and second Principal Components of the labels (attributed to the mean *valence* and *dominance*) using the full structural descriptor. To attest the reliability of these results the confidence interval computed as

$$I = 1.96 \cdot \sigma / \sqrt{10} \tag{1}$$

is shown in parenthesis.

Results indicate that consistently among two of the three classification rules, there is a strong probability of detection - above 85%, for one of the traits and a reasonable good detection (over 75% consistently for two classification rules), for 5 traits, including the first principal component. There are also two traits of which the performance is close to chance, suggesting a more semantic nature of their definition than perceptual one, i.e. traits that may not depend on the structural face description but be contextually dependent.

Examples of the faces assigned to the traits with highest ("dominant") and lowest ("competent") performance among the three classification rules are depicted in figure 2. The dominance prediction is significantly correlated with our intuitive notion of the trait.

Subsequently as mentioned before, a principal component analysis on the 1134-dimensional descriptors is done. The second experiment consisted of a PCA projection of the data into the first 50 PCs, followed by the same classifiers testing protocol defined above. Table 2 shows the performance of the three classification rules. It can be seen that the performance decreases slightly in general with respect to the full descriptor, although some of the traits (such as *trustworthy, frightening, threatening and likeable*) remain basically unchanged among the three classification rules, suggesting that the data present on the weakest eigenvectors may represent a source of noise to the facial structure. As could be expected, the two first principal components (traits 10 and 11), remain virtually unchanged within their confidence intervals.

Results show that the system performance is higher for traits portraying behavioral condition (e.g. threatening) in comparison with those portraying adaptive significance (e.g. likeable). This indicates that the system is better suited to infer a person's ability to implement a given intention, hence reflecting the structural nature of the problem. Regarding the contextual applicability of the model, our results are consistent with those reported in [11], i.e. best performance was achieved when evaluating traits where intrinsic facial features exemplified clearly both ends of a given dimension (e.g. dominant, figure 2).

5.1. Dependency on the Sample Size

Given the small dataset size of the real faces database, we inquired into the complexity of the problem. Driven by this question, a validation of the proposed point-based facial structure descriptor was done by studying the performance of the two classification rules that provided the best results in the previous experiment, as a function of the sample size. We trained a GentleBoost and SVM classifiers with incremental sample size, on the first two Principal Components mentioned in section 4.1. Thus an experiment using the synthetic faces database and increasing the number of samples from 180 to 300 was carried out, where each image is ultimately represented by its 1134 structural descriptor computed from the 21 facial salient points.

Results show in figure 3 that as data size increases, the performance increases accordingly for both classification rules, increasing the accuracy between 20% and 30% for a 67% growth of the sample size, which leads us to believe that an improvement on any result is likely to arise, should the dataset of rated face images grow.

5.2. Experiments with Real Data

In order to validate the previous findings, and see the prediction capabilities of the proposed scheme on real data, the experiments were repeated on the Karolinska faces database [10], which also has rating information from the psychological study presented in [11]. Given that the two databases only share 5 annotated traits in the ground truth data (*Attractive, Trustworthy, Dominant, Mean and Threatening*), the experiment was performed using these labels and the first two principal components. The results obtained from the study are shown in table 3. As mentioned in the previous section, the low availability of annotated data seriously

Table 1. Mean accuracy and confidence interval (in percentage), of the three classification rules for the 9 traits and the first two Principal Components of the ground truth.

Trait	Attractive	Competent	Trustworthy	Dominant	Mean	Frightening	Extroverted	Threatening	Likeable	1PC	2PC
GB	70.76 (6.7)	66.08 (10.8)	67.25 (5.1)	85.38 (5.4)	77.78 (5.6)	70.76 (4.6)	81.29 (3.7)	83.04 (8.0)	66.08 (4.3)	81.87 (6.3)	69.59 (8.0)
SVM	72.51 (4.2)	57.89 (5.2)	76.61 (5.2)	91.23 (3.6)	79.53 (6.0)	79.53 (5.8)	81.29 (5.7)	80.12 (9.2)	66.67 (6.1)	83.04 (5.8)	72.51 (5.8)
3NN	67.84 (6.6)	64.33 (9.2)	63.16 (7.5)	75.44 (7.5)	69.01 (3.4)	63.74 (8.5)	71.35 (5.2)	72.51 (10.2)	66.67 (8.6)	75.44 (5.2)	55.56 (6.3)
5NN	69.59 (9.0)	61.40 (8.0)	62.57 (6.0)	74.27 (5.3)	66.08 (5.9)	59.65 (7.7)	68.42 (4.6)	74.27 (7.2)	60.82 (9.0)	73.10 (6.2)	60.23 (5.7)

Table 2. Mean accuracy and confidence interval (in percentage), of the three classification rules for the 11 traits on the first 50 Principal Components of the 1134 dimensional structural descriptor.

Trait	Attractive	Competent	Trustworthy	Dominant	Mean	Frightening	Extroverted	Threatening	Likeable	1PC	2PC
GB	59.65 (6.5)	65.50 (5.2)	70.18 (6.3)	82.46 (6.3)	72.51 (6.1)	68.42 (6.5)	73.68 (5.9)	78.36 (9.1)	60.23 (4.0)	73.10 (5.3)	63.74 (5.8)
SVM	64.91 (7.8)	64.33 (3.9)	77.19 (5.4)	89.47 (4.6)	80.12 (8.9)	78.95 (4.6)	81.87 (3.7)	79.53 (9.1)	66.67 (8.5)	81.29 (4.3)	69.01 (7.2)
3NN	67.84 (7.9)	63.74 (9.7)	62.57 (7.2)	74.85 (7.3)	67.25 (3.2)	62.57 (8.9)	70.18 (4.9)	71.93 (9.5)	66.08 (8.5)	76.02 (4.9)	55.56 (6.3)
5NN	69.01 (9.7)	61.40 (8.0)	62.57 (6.0)	74.27 (5.3)	66.08 (5.7)	60.23 (7.8)	68.42 (5.4)	73.68 (7.3)	61.40 (8.5)	71.93 (6.3)	60.82 (5.7)

Figure 4. Images of politicians sorted according to the predicted Trustworthiness. To the left, the highest score.

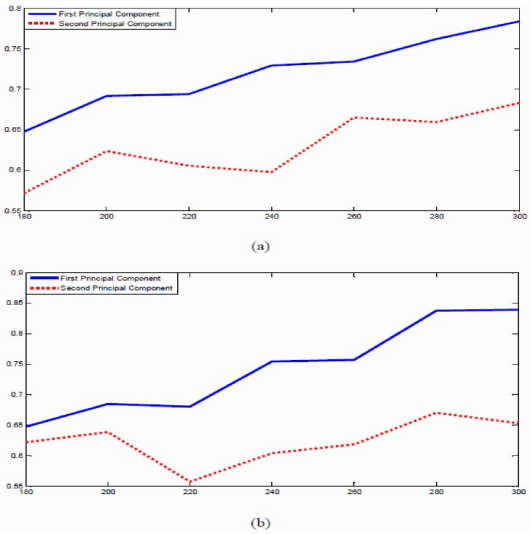

Figure 3. Performance with incremental dataset size for the first two Principal Components. (a) GentleBoost (b) SVM

Figure 5. Images of politicians sorted according to the predicted Valence. To the left, the highest score.

harms the automatic trait prediction. Nevertheless, 3 trait judgments seem to be predictable above chance.

As mentioned before, there is an important dependency on sample size which statistically invalidates any results obtained from experiments on real faces databases, so further analysis shall be restricted to an intuitive relation to those results obtained from the synthetic database.

We performed a final experiment of famous faces obtained from Internet on selected trait judgments. We evaluated the *Trustworthiness* and *Valence* judgments on politicians and *Dominance* and *Threatening* on a group celebrities. We took benefit of the GentleBoost classifier trained in the previous experiment and used it to automatically rate the new acquired faces. Figure 4 shows the images sorted according to the predicted trustworthiness, in figure 5 valence is reported and figure 6 depicts results for celebrities judgments. Prior knowledge and semantics involved in each image, seem to influence and modify eventual human face judgment, an effect that is not present in an automatic face-assessment system.

6. Conclusions

An automatic trait predictor based on a scare-point facial structure descriptor, has been proposed and preliminarily evaluated using a corpus of behavioral data ratings, in order to study the possibility that machine learning classifiers are able to simulate human rapid trait evaluation.

For this purpose, we trained an automatic facial salient point detector that uses a cascaded eye detector as core, and computes spatial coordinates of feature points using a 300

Table 3. Mean accuracy and confidence interval (in percentage) for the Karolinska dataset, using full descriptor

Trait	attractive	mean	trustworthy	dominant	threatening	1PC	2PC
GB	64.29 (24.2)	68.75 (25.8)	58.93 (22.4)	72.32 (28.1)	83.04 (16.5)	75.00 (18.5)	44.64 (22.4)
SVM	61.61 (24.8)	82.14 (17.1)	65.18 (15.7)	77.68 (25.3)	87.50 (16.0)	75.89 (17.9)	49.11 (18.6)

Figure 6. Images of celebrities sorted according to the following predicted traits (to the left, the highest score): top row Dominance, bottom: Threatening.

dimensional NDA projected descriptor. To attain trait evaluation, we computed a high dimensional structural descriptor from the facial salient points, and we trained a set of state-of-the-art machine learning classifiers using a synthetic annotated data base as a ground truth. The experiments show promising results on the prediction of 7 facial traits, specially the *dominance* (with figures as high as 91.23%) , *threatening* and *extroverted* traits. The study also reflects that other facial traits may depend more on the context than on the intrinsic structure of the face.

Finally, we propose the use of the structural descriptor in real images. Experiments performed on the only one annotated database available show also successful prediction capabilities, although the small sample size of the database does not allow for statistically significant results. The final validation of the proposed descriptor in uncontrolled pixel images, may successfully be done once the issue of available rated data is solved.

7. Acknowledgements

This work was partially supported by MEC grants TIN2009-14404-C02-01 and CONSOLIDER-INGENIO 2010 (CSD2007-00018).

References

[1] C. C. Ballew and A. Todorov. Predicting political elections from rapid and unreflective face judgments. *Proceedings of the National Academy of Sciences of the USA*, 104:17948–17953, 2007. 1

[2] M. Bartlett, G. Littlewort, I. Fasel, and J. Movellan. Real Time Face Detection and Facial Expression Recognition: Development and Applications to Human Computer Interaction. In *CVPR Workshop on Computer Vision and Pattern Recognition for Human-Computer Interaction*, 2003. 1

[3] I. Blair, C. Judd, and K. Chapleau. The influence of Afrocentric facial features in criminal sentencing. *Psychological Science*, 15(10):674–679, 2004. 1

[4] M. Bressan and J. Vitria. Nonparametric discriminant analysis and nearest neighbor classification. *Pattern Recognition Letters*, 24(15):2743–2749, 2003. 2

[5] R. El Kaliouby and P. Robinson. Real-Time Inference of Complex Mental States from Facial Expressions and Head Gestures. In *Proc. Intl Conf. Computer Vision & Pattern Recognition*, volume 3, page 154. Springer, 2004. 1

[6] A. Goneid and R. El Kaliouby. Enhanced Facial Feature Tracking of Spontaneous and Continuous Expressions. *Systems, Social and Internationalization Design Aspects of Human-computer Interaction*, 2001. 1

[7] J. Hastie and R. Tibshirani. Additive logistic regression: A statistical view of boosting. *The Annals of Statistics*, 28(2):337–374, 2000. 2, 3

[8] T. Kanade, J. Cohn, and Y. Tian. Comprehensive database for facial expression analysis. In *Fourth IEEE International Conference on Automatic Face and Gesture Recognition, 2000. Proceedings*, pages 46–53, 2000. 2

[9] A. Little, R. Burriss, B. Jones, and S. Roberts. Facial appearance affects voting decisions. *Evolution and Human Behavior*, 28(1):18–27, 2007. 1

[10] D. Lundqvist, A. Flykt, and A. Ohman. Karolinska Directed Emotional Faces (Department of Neurosciences, Karolinska Hospital, Stockholm, Sweden). *Karolinska Directed Emotional Faces*. 2, 4

[11] N. Oosterhof and A. Todorov. The functional basis of face evaluation. *Proceedings of the National Academy of Sciences*, 105(32):11087, 2008. 1, 2, 3, 4

[12] P. Viola and M. Jones. Robust real-time object detection. *International Journal of Computer Vision*, 57(2):137–154, 2002. 2

[13] D. Vukadinovic and M. Pantic. Fully Automatic Facial Feature Point Detection Using Gabor Feature Based Boosted Classifiers. In *Systems, Man and Cybernetics, 2005 IEEE International Conference on*, volume 2, 2005. 2