Fully automated facial picture evaluation using high level attributes

Arnaud Lienhard, Alice Caplier, Patricia Ladret

Fully Automated Facial Picture Evaluation Using High Level Attributes

Arnaud Lienhard, Alice Caplier, Patricia Ladret
GIPSA-Lab, Grenoble, France

Abstract— **People automatically and quickly judge a facial picture from its appearance. Thus, developing tools that can reproduce human judgments may help consumers in their picture selection process. Previous work mostly studied the position of facial keypoints to make predictions about specific traits: trustworthiness, likability, competence, etc. In this work, high level attributes (e.g. gender, age, smile) are automatically extracted using 3 different tools and are used to build models adapted to each trait. Models are validated on a set of synthetic images and it is shown that using attributes increases significantly the correlation between human and algorithmic evaluations. Then, a new dataset of 140 images is presented and used to demonstrate the relevance of high level attributes for evaluating faces with respect to likability and competence. A model combining both facial keypoints and attributes is finally proposed and applied to picture selection: which picture depicts the most likable face for a given person?**

I. INTRODUCTION

Social psychology studies have shown that people evaluate faces automatically and quickly [1], and these first impressions predict social outcomes such as online dating [2] or electoral success [3], [4]. If evaluating a person's expression is subjective and seems difficult to automatize, it is an active research domain leading to many concrete applications: industries may want to analyze people reactions and emotions to an advertisement or a new product.

With the widespread use of digital cameras and smartphones, selecting the best facial picture of a particular person for a given application is a time consuming challenge. Thus, a system providing automatic feedback about images would be an interesting and useful tool. Embedded algorithms such as automatic face and smile detection are already helping consumers to make satisfying shots. However, there are many other cues that have to be considered. A smiling face surrounded by a colorful background would certainly be a good choice when sharing images with friends, while pictures required for professional purposes (visiting cards, resumes) often show a straight face in front of a uniform background. Since these applications mostly imply frontal and good quality images, the proposed model is limited to frontal faces and to images of satisfying quality, for which facial feature detection is possible.

The global evaluation of a facial picture is the combination of all the subjective judgments that a person makes when looking at a face. Studies have shown that face evaluation can be approximated by two dimensions [5], [6]: the first trait corresponds to trustworthiness, which is highly correlated to all positive judgments (likability, attractiveness). Dominance is the other dimension and describes how much a face is evaluated as threatening or mean.

In face images, the subject is evaluated with respect to his/her expression, face shape and other cues such as make-up or face adornments. However, at the best of our knowledge, many of these attributes are not considered in automatic face evaluation systems. A first attempt to create a data-driven model of several evaluation traits is discussed in [7]. In their work, 300 faces are generated by the Facegen Modeller software (http://www.facegen.com) with different shape parameters. A subjective experiment is conducted, where participants evaluate each face with respect to a particular trait: aggressiveness, attractiveness, threat, etc. Finally, parameters are fitted to the ground truth scores provided by participants to build a regression model for each social judgment. Several sets of synthetic faces, generated with respect to the models created in [7], are used for experiments and validation in this work.

Besides, behavioral studies have shown that shape is not enough to evaluate a face and reflectance (cues such as skin illumination and texture) plays an important role in face perception [8]. A more complete model including reflectance parameters is elaborated and validated in [9]. However, the faces considered in all their experiments are synthetic and without facial hair, make-up or accessories. Real 3-D scanned faces have been used in [10] to identify relevant shape and reflectance features. Even in recent attempts of automated face expression evaluation in videos [11], the use of facial keypoints is still predominant. The disadvantage of these models is that it only takes into account the position of facial keypoints and reflectance parameters.

High level attributes are defined as abstract and global concepts describing an image. They correspond to descriptors that cannot directly be obtained by extracting visual data due to the semantic gap between information contained in pixels and human analysis. A small set of attributes provides more significant information than the relative positions of many facial keypoints: converting pixel level information to high level attributes allows to fill the semantic gap between low level computer understanding and human comprehension. Smiles, presence of glasses, gender or age are example of these features in the particular case of facial images. In this article, evaluation is made with respect to a given context: does the face look friendly? Is it a straight face? Thus, the problem is slightly different from facial expression recognition and the proposed method considers high level attributes instead of low-level descriptors (e.g. Gabor wavelets) that are traditionally used in emotion recognition.

Our claim is that the addition of high level attributes significantly improves the models of facial picture evaluation, which are, at the best of our knowledge, only based on

faces described by their keypoints and other pixel level information (texture, illumination). Thus, this work focuses on demonstrating the need of using high level attributes to build efficient facial picture evaluation models. Many attributes (age, gender, presence of glasses, beard, smile, etc) can be extracted automatically by machine learning algorithms. They have already been successfully used in various research domains such as face recognition or verification [12] and portraiture aesthetics [13]. 3 tools analyzing faces are considered: Betaface (*http://betaface.com*), SkyBiometry (*http://skybiometry.com*) and SHORE [14]. Each tool takes a picture as input and provides a set of attributes describing the face.

Sec. II introduces the main steps of the proposed method, including the presentation of databases, feature sets and learning algorithms. Sec. III demonstrates the relevance of high level attributes in the context of face evaluation and Sec. IV applies the method on natural images. Further analysis is given in Sec. V.

II. PROPOSED METHOD

In order to prove that high level attributes are efficient for automatic face evaluation, it is necessary to collect ground truth data and tools to compute the attributes. Synthetic and natural data are described below, followed by the list of attributes obtained from 3 different face analyzers. Learning algorithms are finally applied to fit attributes to the ground truth scores and build facial picture evaluation models.

A. Datasets

1) Synthetic Faces used for Validation: Many datasets containing facial pictures have been built and annotated with respect to criteria such as trustworthiness or dominance. In this section, only synthetic faces are considered, without any extra-facial cues such as hairstyle, beard, glasses or jewelry.

Using human-rated synthetic facial pictures, 7 models of face evaluation are computed in [7] with respect to the following traits: attractive, competent, dominant, extroverted, likable, threatening, trustworthy. For each model, a dataset of 25 distinct faces is created. These faces are manipulated along the respective traits to generate seven variations corresponding to seven levels of the considered dimension, producing sets of $25 \times 7 = 175$ images. Subjective experiments revealed that these synthetic faces are greatly correlated with the models [9]. 3 variations of trustworthiness for a given face are presented in Fig. 1. In our experiments, 4 datasets corresponding to the following traits are considered: trustworthy, dominant, likable and competent.

2) Human Study on Natural Images: To evaluate our model on natural images, 140 frontal and centered pictures of 20 different persons [1] (10 men and 10 women) have been gathered, mostly from the LFW [15] dataset. A large variety of gaze and expression, facial hair styles or accessories are considered for each identity. Samples are shown in Fig. 2. Two distinct experiments involved participants that were

[1]Unfortunately, the French law does not allow us to make this database publicly available.

Fig. 1. Examples of synthetic faces manipulated for trustworthiness. From left to right: untrustworthy, neutral and very trustworthy.

Fig. 2. Examples of faces considered in the proposed dataset.

asked to evaluate each image in the same viewing conditions, with respect to either competence or likability. Images were presented in a random order, after a preliminary learning process where participants had to rate images that are not part of the dataset. A discrete scale from 1 (not at all competent/likable) to 6 (very competent/likable) has been considered, which appeared to be relevant since the Cronbach's Alpha [16] value is above 0.96 for both experiments. Finally, 27 participants aged from 20 to 55 rated each image for both criteria. Scores average and standard deviation are respectively 3.3 and 0.47 for competence evaluation, 3.37 and 0.79 for likability.

B. High Level Attributes for Face Evaluation

In this work, attributes extraction is performed by 3 applications provided "as is": the SHORE software [14] and two free cloud based applications: Betaface and SkyBiometry. Each tool T returns a total of N_T distinct features. Values may either be discrete (is it a male or a female?) or continuous (how much is the person smiling?). Some features have both a discrete component ("yes" or "no") and a continuous component ("how much?"): Does this person smile (yes or no)? How much (from 0 to 1)?

A total of 63 attributes is gathered: 37 from Betaface, 20 from SkyBiometry and 6 from Shore. A simplified list of these attributes is given in Tab. I. Note that Betaface and SkyBiometry also return a list of facial keypoint positions (respectively 94 and 73 points); detection examples are given in Fig. 3. Thus, it is possible to compare the performance between the use of keypoints and high level attributes (III-B). Attribute sets are first tested separately (III-C) in order to compare the performance of each tool. Then, the 3 sets of values are fused with the keypoints to create a global model (III-D). This global model is finally tested and validated on natural images (IV).

C. Learning from the Data

The statistical models presented in this article have been computed by the OpenCV implementation of the Gradient

TABLE I

SIMPLIFIED LIST OF HIGH-LEVEL ATTRIBUTE CATEGORIES COMPUTED BY EACH TOOL. THE NUMBER OF VALUES (DISCRETE OR CONTINUOUS) DESCRIBING EACH CATEGORY IS REPORTED IN THE GREEN CELLS.

	Gender	Age	Smile	Mood	Beard	Mustache	Glasses	Eyes	Mouth	Eyebrows	Nose	Skin	Hair	Shape
Betaface	2	1	3		3	3	3	4	3	3	2	2	5	3
SkyBiometry	2		2	8			4	2	2					
SHORE	1	1		1				2	1					

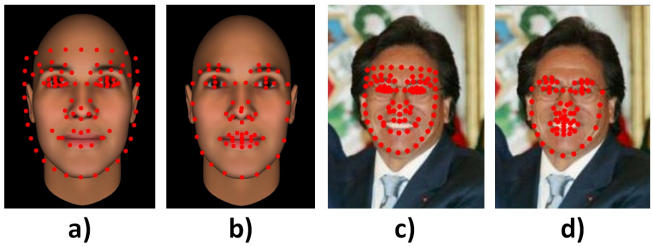

Fig. 3. Examples of facial keypoint positions for synthetic and natural images. a) and c) show Betaface points, b) and d) SkyBiometry points.

Boosted Trees (GBT) algorithm. This choice was motivated by the fact that this algorithm is quite robust to the noise introduced in the datasets, including human subjective ratings and errors due to missing or erroneous attribute values. GBT has shown more promising results in our study than Neural Networks (NN), Random Forests (RF) and SVM (see discussion in Sec. V).

Testing the models is performed by 10-fold cross validations for both classification and regression problems. To avoid sampling bias, each experiment is repeated 10 times with randomly chosen images and the final average performance is reported.

III. VALIDATION ON SYNTHETIC FACES

In a first set of experiments, facial keypoints are used to create face evaluation models able to classify faces in seven categories, corresponding to seven levels of the considered trait. Since the synthetic faces are generated by keypoints distortion, models using these features should provide high classification performance (III-B).

Then, attributes provided by each library are tested separately on synthetic faces and compared to the reference models using keypoints (III-C). Attributes are finally concatenated and used to create a new and coherent model, confirming that high level features are promising in the case of facial picture evaluation (III-D).

A. Performance Criteria for Classification

Evaluating performance of classification algorithms needs the use of appropriate criteria, describing both the number and the type of errors that the classifier makes. In the case of multi-class categorization, performance is evaluated by the Cross-Category Error (CCE)

$$CCE_i = \frac{1}{N_t} \sum_{n=1}^{N_t} I(\hat{c}_n - c_n = i)$$

TABLE II

EXPERIMENTAL RESULTS FOR 7-CLASS CATEGORIZATION USING EITHER BETAFACE OR SKYBIOMETRY FACIAL KEYPOINTS

	Betaface		SkyBiometry	
	GCR	MCE	GCR	MCE
Trustworthy	**0.40**	128	0.39	**116**
Dominant	0.39	117	**0.49**	**92**
Likable	0.28	158	**0.34**	**139**
Competent	0.31	152	**0.39**	**121**

and the Multi-Category Error (MCE)

$$MCE = \sum_{i=-(N_c-1)}^{N_c-1} |i| CCE(i)$$

where N_t is the number of test images, N_c the number of classes, $\hat{c}_n$ the ground truth class, c_n the predicted class. i is the difference between ground truth and predicted class and $I(.)$ is the indicator function. Finally, the Good Classification Rate (GCR) is defined by the ratio between the number of images that are correctly classified (CCE_0) and the total number of images (N_t). Ideally, GCR should be the highest (close to 1) and MCE the lowest possible (close to 0).

B. State of the Art : Key Points

The synthetic datasets described in Sec. II contain faces that are categorized in 7 levels of each considered traits: trustworthy, dominant, likable and competent. Our first attempt in creating a face model for each trait requires the facial keypoints provided by either Betaface or SkyBiometry. Applying the GBT algorithm to each dataset, the results presented in Tab. II are obtained.

The performance of a random classifier is approximately $GCR = 14\%$ and $MCE = 400$. Low MCE values reported in the table (between 100 and 150) and high GCR (between 30 and 50%) indicate that not only our classifier is able to classify correctly many faces (high GCR), but also makes only minor mistakes (low MCE). This is confirmed by the CCE measures, revealing that the classifier never makes errors of more than 2 levels. It is noticeable that SkyBiometry facial keypoints are slightly more efficient than Betaface's: there are significant differences between both detectors in the case of synthetic images (see Fig. 3). Finally, note that this process can be fully automated without the use of computed generated faces [9] or 3-D face scanners [10].

C. Contribution of High Level Features

The principal contribution of this article is the use of high level attributes (see Tab. I) for facial picture evaluation. In order to have an idea of the ability of high level attributes to classify face pictures, the datasets presented above are going

TABLE III

AVERAGE EXPERIMENTAL RESULTS FOR 7-CLASS CATEGORIZATION. ONLY HIGH LEVEL ATTRIBUTES ARE CONSIDERED.

	Betaface		SkyBiometry		SHORE	
	GCR	MCE	GCR	MCE	GCR	MCE
Trustworthy	0.31	145	**0.34**	**143**	0.23	198
Dominant	**0.44**	**105**	0.40	114	0.34	145
Likable	0.30	159	**0.36**	**137**	0.19	230
Competent	**0.36**	**134**	0.35	137	0.25	212

to be tested using one tool at a time. Classification results are presented for each feature set in Tab. III. Again, results are significantly higher than chance. In all the experiments, the values of CCE_i for $i < -1$ and $i > 1$ are very close to 0, showing that if there are some classification mistakes, they do not exceed one category.

Note that Betaface and SkyBiometry produce equivalent results, while SHORE's attributes are less discriminant. This is due to the lack of many relevant facial attributes in SHORE, especially the presence of a smile. Performance is slightly lower using attributes instead of facial keypoints on synthetic datasets since they were originally created by keypoints manipulation and are therefore naturally efficiently described with keypoints. The following paragraph shows that global performance can be enhanced by concatenating the three sets of attributes and the keypoints information.

D. Combining Attributes to Enhance the Model

In the following experiments, high level attributes obtained from different tools are concatenated and used for learning. This is called "early fusion" and is opposed to "late fusion", which consists in getting first 3 models based on each set of features and combining the results. In our case, early fusion produces better results since each attribute set contains exclusive relevant attributes.

It is possible to increase global performance by adding the locations of the facial features in the global feature set, taking both low and high level features into account. Fig. 4 shows the classification errors for the model on Likable and Competent datasets and compares performance of keypoints, attributes and the entire set (both keypoints and attributes). Fused high level attributes are as efficient as facial keypoints, resulting in approximately the same good classification rate ($GCR = 35\%$) and 1-category error (50%). Combining features increases performance for both datasets ($GCR = 40\%$) and reduces 2-category error (from 10 to 2%). This last result is promising since it shows that high level attributes provide additional information on synthetic datasets that are based on keypoints manipulation, and are thus likely to be of great help for natural images.

IV. APPLICATION ON NATURAL IMAGES

The dataset of natural images presented in Sec. II provides a final validation of the proposed model. Likability and competence have been chosen as face evaluation criteria since we have the intuition that they fit to several applications: likable faces have more success on social media, while competent looks are likely to be used in professional networks. Since each picture has a ground truth likability or competence score, it is possible to build regression models. This section validates the use of high level attributes in the case of natural faces, and shows that facial keypoints are significantly less efficient (IV-B). An example of picture selection method is then presented, showing that it is possible to automatically and efficiently select or remove images fitting a given application (IV-C).

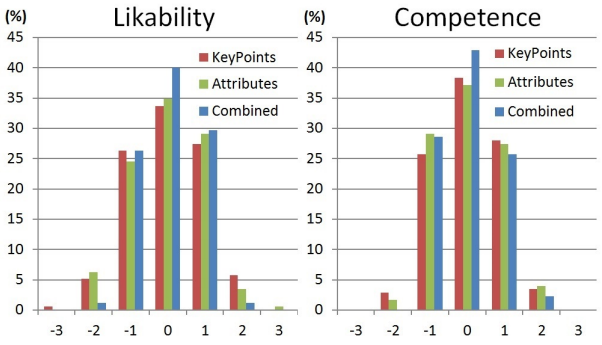

Fig. 4. Classification errors (CCE) for likability and competence, using different feature sets: facial keypoints (red), high level attributes (green) and global feature set (attributes and keypoints, blue).

A. Performance Criteria for regression

Regression performance is measured by Pearson and Spearman correlations. Pearson correlation R is defined as

$$R = \frac{\sum_{n=1}^{N_t}(\hat{s}_n - \bar{\hat{s}}) \cdot (s_n - \bar{s})}{\sqrt{\sum_{n=1}^{N_t}(\hat{s}_n - \bar{\hat{s}})^2} \cdot \sqrt{\sum_{n=1}^{N_t}(s_n - \bar{s})^2}}$$

where $\bar{\hat{s}} = \frac{1}{N_t}\sum_{n=1}^{N_t}\hat{s}_n$ and $\bar{s} = \frac{1}{N_t}\sum_{n=1}^{N_t}s_n$. Spearman rank correlation ρ is computed by

$$\rho = 1 - \frac{6\sum_{n=1}^{N_t} d_n^2}{N_t(N_t^2 - 1)}$$

where $d_n = \hat{s_n} - s_n$ is the rank difference between variables. Pearson's measure quantifies how much the data are linearly correlated and Spearman's rank correlation tells if the data are monotonically correlated. Values close to 1 result from correlated data and 0 means that the data are not correlated.

B. Impact of High Level Attributes on Natural Images

In the case of natural pictures, it is much less efficient to rely exclusively on facial keypoints: many other facial cues play a role in face evaluation: hairs, beard, glasses, etc. This is confirmed by the experimental results presented in Tab. IV. For both likability and competence, using attributes instead of keypoints enhances the performance: correlation increases respectively from 0.6 to 0.8 and 0.4 to 0.5.

Another observation is that the performance of the competence model is lower than the likability model. Several reasons can explain this. First, likability scores present a

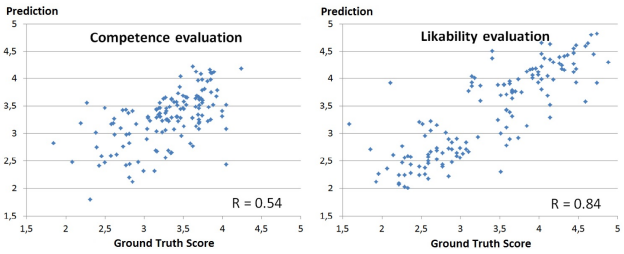

Fig. 5. Predictions for both competence and likability models.

TABLE IV

AVERAGE PERFORMANCE FOR NATURAL IMAGE REGRESSION, USING EITHER KEYPOINTS OR COMBINED HIGH LEVEL ATTRIBUTES

	B. Keypoints		S. Keypoints		HL Attributes	
	R	ρ	R	ρ	R	ρ
Likable	0.63	0.64	0.50	0.51	**0.83**	**0.83**
Competent	0.36	0.38	0.38	0.40	**0.5**	**0.53**

higher standard deviation (0.8 versus 0.5 for competence), showing that even humans better identify likable faces than competent ones. Plus, in the case of likability, some of the high level attributes (smile, mood) are extremely discriminant, whereas many cues that play a significant role in competence evaluation are not considered in the feature set (eg. clothes, background). This is discussed in Sec. V.

By combining attributes and facial keypoints, it is possible to enhance the competence evaluation model and obtain a correlation of 0.54 (0.5 with attributes only). Likability correlation between ground truth and predicted scores reaches a value of 0.84 (0.83 with attributes only). Fig. 5 presents the regression cloud points and shows the correlation between ground truth and predicted scores. In the case of likability, the prediction seems to be very accurate, even if there are still several outliers. These prediction errors are not only due to some missing attributes in the feature space (background or reflectance cues are not considered), but also to some erroneous measurements. Examples of erroneous measurements are given in the following paragraph, which presents an example of a possible application of this method and its limits.

C. Application to Picture Selection

Automated picture selection of a given person for a particular application is a practical example that may benefit from the proposed method and its results. Some people have hundreds of pictures from which they want to select a small set that is relevant for a given application: facebook profile picture, professional purposes like visiting cards, etc.

While our model is accurate when dealing with likability (Pearson and Spearman correlations above 0.8), many biases make the competence evaluation more difficult (correlations between 0.5 and 0.6). This is discussed in Sec. V, and only the likability trait is considered in this section.

Scoring pictures with high level attributes enables the selection of a small number of images: what are the 10 most likable faces in the entire set? Using the dataset containing 7 different pictures for each of the 20 people, the following

Fig. 6. A set of 7 images from the same person (LFW dataset).

TABLE V

GROUND TRUTH AND PREDICTION LIKABILITY FOR IMAGES IN FIG. 6.

Image	**a)**	**b)**	**c)**	**d)**	**e)**	**f)**	**g)**
Gr. Truth	1.59	1.96	2.22	3.38	3.56	3.59	4.04
Prediction	3.15	2.64	2.56	3.33	3.72	3.42	3.94

experiment is made. First, a model is learned using 19 persons (133 images). Then, the images of the last person are scored with respect to likability. Face likability ranking for a given person is displayed in Fig. 6 and Tab. V.

It can be observed that a face judged as likable by humans is evaluated as likable by the algorithm (pictures **d** to **f**). Smiling faces are considered as likable by both humans and the algorithm. Even faces where there is no obvious smile (**d**, **f**) are correctly evaluated by the algorithm. This can happen thanks to the use of other criteria, such as eyebrows shape and position. Picture **a** predicted score is far from ground truth since the algorithm misreads the facial expression and considers the presence of a smile. Faces **b** and **c**, considered as not likable by humans with ground truth scores of respectively 1.96 and 2.22, are correctly evaluated by the algorithm (respectively 2.64 and 2.56). Using the model and a particular threshold corresponding to a given score or pictures to select, it is possible to automatically remove unwanted pictures (in this case, the 3 images in the left), or to select a few good-looking images. Only few images are not correctly scored (picture **a**) due to the errors made through feature computation, which is one of the limits to the proposed model.

V. DISCUSSION

If GBT provides the best regression performance in the case of natural images for the proposed dataset (0.83 for likability), models based on RF, SVM and NN can also be considered and obtain, in the same experimental conditions, correlations of 0.81, 0.74 and 0.72. The same observation can be made for competence evaluation: GBT and RF output the optimal performance (0.5), ANN and SVM are slightly below (0.46). Optimizing the parameters may enhance the performance since only OpenCV default parameters have been considered in our experiments.

Using high level attributes to learn about face social perception is not only efficient, but may also be used to see what kind of features are helpful to create good models. The knowledge of relevant attributes and facial keypoints positions may be of great help in other applications like photo editing. By rating pictures and telling which attribute is missing or makes the picture perfectible, it is possible to perform automated corrections and produce modified facial pictures that fit nicely to a given application. To have an idea

TABLE VI

RELIEF VALUES OF THE MOST DISCRIMINATIVE ATTRIBUTE CATEGORIES FOR BOTH LIKABILITY AND COMPETENCE EVALUATION.

	Smile	Mood	Eyes	Beard	Gender	Age
Likability	**0.14**	**0.19**	0.08	0.08	0.09	0.08
Competence	0.11	0.11	**0.15**	0.08	0.09	0.06

about relevant features in our studies, dimension reduction algorithms have been applied and the most relevant features are presented below.

The *RReliefF* algorithm is implemented as described in [17] and is able to deal with both discrete and continuous data, assigning a weight to each attribute. Higher weights mean discriminant attributes. Tab. VI presents the most discriminant attributes and their weights for likability and competence evaluation for natural images. Synthetic images are not considered because several facial cues are missing: glasses, mustache, eyes closeness, etc.

As it has been observed in Sec. III, smile and mood measures are by far the most discriminant in the case of likability evaluation, followed by age estimation and gender. It is even possible to make accurate predictions using only these measures, resulting in $R = 0.76$, instead of 0.84 for the entire set. This is an interesting result since computing only 2 measures is easier and quicker, leading to many concrete applications involving real-time detection.

Smiles and emotions also play a role in competence evaluation, as well as criteria such as glasses, beard, followed by age and gender evaluation. Values related to eyes are the most relevant attributes in the case of competence evaluation. This is not surprising: images with closed eyes are not used in any real life application, especially when a straight face is required. Note that high correlation between face evaluation and a particular attribute does not mean that people having this attribute are likely to be more competent or likable. It can only be concluded that the algorithm is able to reproduce human preconceptions about face evaluation. There may be biases in the dataset as well: if several men are presented wearing suits, the model will predict men as more competent, and this cannot be generalized. This kind of bias would be reduced by the introduction of new features taking clothes, jewelry or background into account.

VI. CONCLUSION AND FUTURE WORK

In this article, high level attributes are used to evaluate facial pictures. It has been shown that these features are more efficient than keypoints. Even on images created artificially using face grids based on keypoints, facial attributes perform as well as models based on keypoint positions. Plus, combining both facial keypoints and high level attributes increases the classification performance on synthetic faces.

In the case of natural images, accurate models can be elaborated to judge faces on different traits. Likability evaluation is precise enough to be used in real life applications. Our algorithm can automatically assess a face's likability, except for a few images where it is difficult to measure efficiently facial attributes.

Competence is harder to evaluate accurately since our perception relies on more subtle cues which are not encoded in the feature set: clothes and background have to be considered in future work. Note that performance can be increased by the use of bigger datasets since it is difficult to build a generic model with only 140 images.

Finally, combining the proposed method with models of image quality and aesthetics would enable to assess both face evaluation and image visual appeal. The development of a standalone and fully automated software able to both analyze and provide feedback about facial pictures would make this research helpful for various applications, from face social perception to automatic photo selection or editing.

REFERENCES

[1] Janine Willis and Alexander Todorov, "Making Up Your Mind After a 100-Ms Exposure to a Face," *Psychological science*, vol. 17, no. 7, pp. 592–598, 2006.

[2] Jeffrey T. Hancock and Catalina L. Toma, "Putting Your Best Face Forward: The Accuracy of Online Dating Photographs," *Journal of Communication*, vol. 59, no. 2, pp. 367–386, June 2009.

[3] Alexander Todorov, Anesu N Mandisodza, Amir Goren, and Crystal C Hall, "Inferences of competence from faces predict election outcomes.," *Science (New York, N.Y.)*, vol. 308, no. 5728, pp. 1623–6, June 2005.

[4] Niclas Berggren, Henrik Jordahl, and Panu Poutvaara, "The looks of a winner: Beauty and electoral success," *Journal of Public Economics*, vol. 94, no. 1-2, pp. 8–15, Feb. 2010.

[5] Nikolaas N Oosterhof and Alexander Todorov, "The functional basis of face evaluation.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 32, pp. 11087–92, Aug. 2008.

[6] Alexander Todorov, Sean G Baron, and Nikolaas N Oosterhof, "Evaluating face trustworthiness: a model based approach.," *Social cognitive and affective neuroscience*, vol. 3, no. 2, pp. 119–27, June 2008.

[7] Alexander Todorov and NN Oosterhof, "Modeling Social Perception of Faces," *Signal Processing Magazine, IEEE*, , no. March, pp. 117–122, 2011.

[8] Volker Blanz and Thomas Vetter, "A morphable model for the synthesis of 3D faces," *Proceedings of the 26th annual conference on Computer graphics and interactive techniques - SIGGRAPH '99*, pp. 187–194, 1999.

[9] Alexander Todorov, Ron Dotsch, Jenny M Porter, Nikolaas N Oosterhof, and Virginia B Falvello, "Validation of data-driven computational models of social perception of faces.," *Emotion*, vol. 13, no. 4, pp. 724–38, Aug. 2013.

[10] Mirella Walker and Thomas Vetter, "Portraits made to measure: Manipulating social judgments about individuals with a statistical face model," *Journal of Vision*, vol. 9, pp. 1–13, 2009.

[11] David Masip, Michael S North, Alexander Todorov, and Daniel N Osherson, "Automated prediction of preferences using facial expressions.," *PloS one*, vol. 9, no. 2, pp. e87434, Jan. 2014.

[12] Neeraj Kumar, Alexander C Berg, Peter N Belhumeur, and Shree K Nayar, "Attribute and simile classifiers for face verification," *EEE 12th International Conference on Computer Vision*, pp. 365–372, Sept. 2009.

[13] Congcong Li, AC Loui, and T Chen, "Towards aesthetics: a photo quality assessment and photo selection system," in *Proceedings of the international conference on Multimedia*, 2010, pp. 10–13.

[14] Andreas Ernst, Tobias Ruf, and Christian Kueblbeck, "A modular framework to detect and analyze faces for audience measurement systems," in *2nd Workshop on Pervasive Advertising at Informatik*, 2009, pp. 75–87.

[15] GB Huang, M Mattar, Tamara Berg, and E Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Tech. Rep., 2007.

[16] Lee J. Cronbach, "Coefficient alpha and the internal structure of tests," *Psychometrika*, vol. 16, no. 3, 1951.

[17] M Robnik-Šikonja and I Kononenko, "Theoretical and empirical analysis of ReliefF and RReliefF," *Machine learning*, vol. 53, pp. 23–69, 2003.