Semantic Face Signatures: Recognizing and Retrieving Faces By Verbal Descriptions

Nawaf Yousef Almudhahka, Mark S. Nixon, Jonathon S. Hare

***Abstract*—The adverse visual conditions of surveillance environments and the need to identify humans at a distance have stimulated research in soft biometric attributes. These attributes can be used to describe a human's physical traits semantically and can be acquired without their cooperation. Soft biometrics can also be employed to retrieve identity from a database using verbal descriptions of suspects. In this paper, we explore unconstrained human face identification with semantic face attributes derived automatically from images. The process uses a deformable face model with keypoint localisation which is aligned with attributes derived from semantic descriptions. Our new framework exploits the semantic feature space to infer face signatures from images and bridges the semantic gap between humans and machines with respect to face attributes. We use an unconstrained dataset, LFW-MS4, consisting of all the subjects from View-1 of the LFW database that have four or more samples. Our new approach demonstrates that retrieval via estimated comparative facial soft biometrics yields a match in the top 10.23% of returned subjects. Furthermore, modelling of face image features in the semantic space can achieve an equal error rate of 12.71%. These results reveal the latent benefits of modelling visual facial features in a semantic space. Moreover, they highlight the potential of using images and verbal descriptions to generate comparative soft biometrics for subject identification and retrieval.**

***Index Terms*—Semantic attributes, soft biometrics, face recognition, identity retrieval, computer vision.**

I. Introduction

THE increased awareness of the role of surveillance systems for public safety and security has driven vast deployments of CCTV networks around the globe. This has motivated research in human identification using semantic descriptions based on eyewitnesses statements with a view to enable searching a database of subjects through verbal descriptions, as illustrated in Fig. 1. These semantic descriptions are based on soft biometrics, which refer to physical and behavioural attributes that can be used to identify people. Human identification has been traditionally based on hard biometrics such as iris, DNA, and fingerprint, which require subjects' cooperation. On the other hand, soft biometrics can be acquired at a distance while having more robustness to the challenging visual conditions of surveillance such as occlusion of features, viewpoint variance, low resolution, and changes in illumination [1]–[3]. Accordingly, soft biometrics can play a significant role in criminal investigations, where it is required to retrieve the identity of a suspect from a database of subjects (e.g. mugshots or surveillance footage) using a verbal description (i.e. eyewitness statement).

N. Y. Almudhahka, M. S. Nixon, and J. S. Hare are with the School of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, U.K. (email: {nya1g14, msn, jsh2}@ecs.soton.ac.uk).

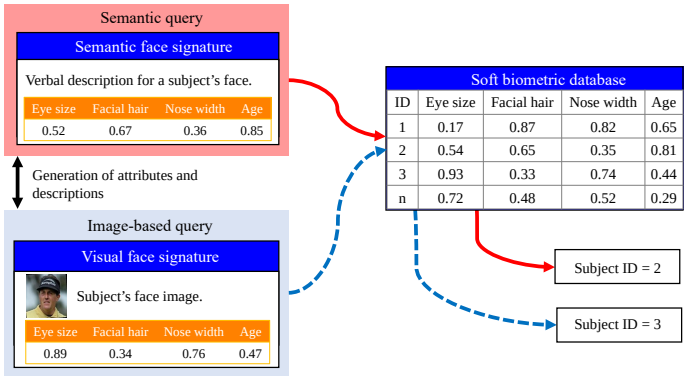


Fig. 1. Using comparative semantic attributes for subject identification and retrieval via face signatures.

Existing work on soft biometrics has studied the face [4]–[7], body [1], [8], [9], and clothing [10] as sources of attributes that enable human identification. Commonly, soft biometrics have been expressed in categorical format [1], [4], [5], [11], for example a person's age might be described as "*young*" or "*old*". However, recent research has shown that describing a person's attributes relative to another person (e.g. person-A is "*older*" than person-B) yields a better recognition accuracy [2]. Moreover, relative attributes have demonstrated richer semantic descriptions and more accurate visual interpretations as compared with categorical attributes [12].

The earliest exploration of semantic facial attributes emerged in [5], where face verification using the Labelled Faces in the Wild (LFW) database [13] was examined via 65 automatically estimated attributes that represent the presence or absence of facial features (i.e. categorical soft biometrics). A key study in facial soft biometrics is that of Reid and Nixon [6], which was the first study to investigate human face identification using comparative soft biometrics. In their study, 27 comparative facial attributes were defined, and annotations were collected for subjects from the Southampton Gait Database [14]. The experiments showed that comparative facial soft biometrics significantly outperform categorical facial soft biometrics with regards to recognition accuracy. Klare et al. [4] conducted a detailed examination of categorical facial attributes for suspect identification. They proposed a method for automatically extracting facial attributes, and identification experiments were conducted using the FERET database [15] with all possible probe-gallery combinations (i.e. human vs. machine) via 46 facial attributes. In [16], Tome et al. proposed shape and size features of facial traits as categorical soft biometric attributes that can be used in conjunction with

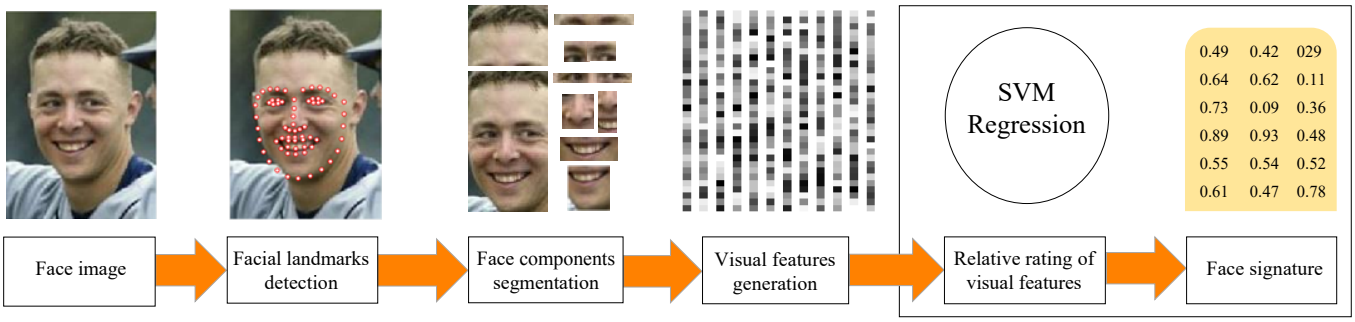


Fig. 2. Vision-based retrieval of face signatures from facial images.

traditional face recognition systems to improve recognition performance. Their experiments, which used the ATVS [16] and MORPH [17] databases, demonstrated the significant discriminative power of categorical soft biometrics and the performance gain that can result from using them in fusion with traditional face recognition systems. Samangouei et al. [18] proposed a method for face authentication on mobile devices that is based on categorical (binary) attributes. Using the MOBIO [19] and AA01 [18] databases, their approach has shown that facial attributes can improve face verification performance considerably. Recently, approaches that are based on deep learning have been proposed [20]–[22] to automatically estimate the 40 binary facial attributes (e.g. *pointy nose* and *big lips*) of the CelebA dataset [23], and their experiments have revealed the capabilities of deep techniques in accurately estimating binary facial attributes.

With respect to relative facial attributes, Almudhahka et al. [24] have recently studied the automatic estimation of comparative labels using 400 subjects from LFW and 20 attributes. The approach is simple, effective, and demonstrated significant accuracy of estimating comparative labels from face images. However, a major issue is that its time and space complexities grow exponentially with dataset size because it involves estimating comparisons rather than rating individual samples. Also, the approach is limited to binary comparative labels, whereas some applications might involve multi-level comparative labels. Taken together, these issues motivate the proposal of a compact pointwise approach that is invariant to labels levels and can efficiently scale to larger datasets.

From this literature review, it can be seen that there has been no detailed investigation of face retrieval by verbal descriptions using comparative soft biometrics. Furthermore, and to the best of our knowledge, no previous work has explored the semantic gap between humans and machines with respect to relative facial soft biometrics. Although some studies have addressed the retrieval of relative body [25] and clothing [10] attributes for subject identification, retrieval of relative facial attributes has only been studied for non-biometrics objectives [12], [26]. Therefore, the purpose of this paper is to explore subject retrieval using verbal descriptions or sample face images in a database of face images. Both of these have an important practical impact for identification of subjects in surveillance databases. In addition, we explore the semantic gap between humans and machines in rating facial attributes. The main contributions of this paper are:

- Exploring the rating of facial attributes from images and analysing the attribute significance for identification.
- Discovering the semantic gap between humans and machines in interpreting relative facial attributes.
- Investigating the impact of semantically modelling face features on identification and verification performance.
- Proposing a framework for face retrieval and verification using relative attributes.

The remainder of the paper is dedicated to exploring unconstrained face retrieval using comparative soft biometrics, and the automatic retrieval of relative attributes from face images. Section II introduces the soft biometric attributes and describes the dataset that is used in this paper. Section III explains the generation of semantic face signatures for unconstrained identification through relative rating from comparative labels. Section IV details the prediction of attributes from face images to construct visual face signatures and explores the correspondence between humans and machines estimations of comparative attributes. Section V presents and discusses the results of retrieval and verification experiments. Finally, Section VI summarises the findings and presents the conclusions. The terms “soft biometrics” and “semantic attributes” are used synonymously throughout the paper.

II. Facial Soft Biometrics

Due to its richness of features and details as compared with body and clothing, the human face is considered the most informative source of attributes for identification at short distances [3], [27]. This informative richness can result in the generation of numerous attributes that semantically describe facial features. However, facial features vary in their importance for recognition by humans [28]. Also, people differ in their perception, understanding, and ability to semantically describe facial attributes [6], [29]. In criminal investigations, the accuracy and completeness of verbal descriptions play a pivotal role in suspect identification [30]–[32]. Therefore, the definition of attributes from the human face should consider the relative importance of facial features in addition to human subjectivity in understanding and describing those attributes.

A. Attribute Definition

We use our previously defined comparative facial soft biometrics, which were introduced in [29], for the purpose of

TABLE I
SOFT BIOMETRIC ATTRIBUTES AND COMPARATIVE LABELS.

Attribute	Labels
Chin height	[More Small, Same, More Large]
Eyebrow hair colour	[More Light, Same, More Dark]
Eyebrow length	[More Short, Same, More Long]
Eyebrow shape	[More Low, Same, More Raised]
Eyebrow thickness	[More Thin, Same, More Thick]
Eye-to-eyebrow distance	[More Small, Same, More Large]
Eye size	[More Small, Same, More Large]
Face length	[More Short, Same, More Long]
Face width	[More Narrow, Same, More Wide]
Facial hair	[Less Hair, Same, More Hair]
Forehead hair	[Less Hair, Same, More Hair]
Inter eyebrow distance	[More Small, Same, More Large]
Inter pupil distance	[More Small, Same, More Large]
Lips thickness	[More Thin, Same, More Thick]
Mouth width	[More Narrow, Same, More Wide]
Nose length	[More Short, Same, More Long]
Nose septum	[More Short, Same, More Long]
Nose-mouth distance	[More Short, Same, More Long]
Nose width	[More Narrow, Same, More Wide]
Spectacles	[Less Covered, Same, More Covered]
Age	[More Young, Same, More Old]
Figure	[More Thin, Same, More Thick]
Gender	[More Feminine, Same, More Masculine]
Skin colour	[More Light, Same, More Dark]

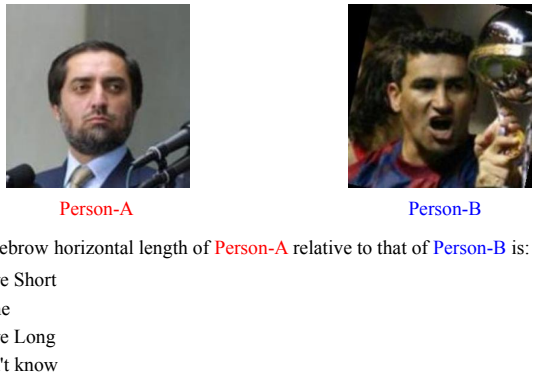


Fig. 3. Example crowdsourced comparison.

TABLE II
CROWDSOURCING JOB STATISTICS FOR THE LFW-V1 DATASET.

	Collected	Inferred	Total
Attribute comparisons	241560	132879504	133121064
Subject comparisons	10065	5536646	5546711
Average comparisons per subject	4.98	1371.1	N/A
Annotators (contributors)	9901	N/A	N/A

identifying humans in surveillance databases, as it has demonstrated its significance and effectiveness for unconstrained identification. The soft biometric set consists of two groups of attributes: (1) *facial*, which describe face components (e.g. eyebrows and nose) and; (2) *global*, which describe demographic information that can be inferred from faces (e.g. age and gender). The selection of the *facial* attributes has taken into consideration three aspects: (a) maintaining coverage of the major face components from forehead down to chin; (b) giving priority to size features, as shape features have demonstrated low discriminative power [7]; and (c) emphasising the role of eyebrows and eyes in face recognition by humans, as they demonstrated the highest impact on face recognition [28], [33]–[35]. This resulted in the definition of 24 attributes (20 *facial* and 4 *global*), where each attribute X is associated with three comparative labels that represent the difference in the attribute between two subjects as: "*Less X*", "*Same X*", and "*More X*". The attributes are listed in Table I.

B. The LFW-MS4 Dataset

LFW is a well-known database for studying unconstrained face recognition [36]. The LFW database consists of 13233 sample face images for more than 5000 subjects extracted from the web. The images of LFW exhibit challenging visual conditions such as variances in the pose, facial expressions, and illumination, in addition to low resolution, which make it suitable to study human identification using soft attributes. LFW has a defined protocol for the pair matching problem and consists of two major subsets: *View 1*, which is dedicated to training and model selection, and *View 2*, which is dedicated to performance reporting. As the purpose of this paper is to discover face retrieval based on verbal descriptions, we constructed the LFW-MS4 dataset*, which is a subset of *View-1* of the LFW database that consists of all the subjects represented by at least four sample face images in the database. The LFW-MS4 dataset has a total of 430 subjects and 1720 sample images. We have used LFW-MS4 to start the exploration of the automatic estimation of comparative facial soft biometrics because larger datasets require many pairwise annotations, as the number of comparative labels exponentially grows with the dataset size. Thus, comparative soft biometrics require careful investigation with a smaller dataset before launching a very large annotation project. The samples in LFW-MS4 were all aligned using deep funnelling [37] and normalised to an inter-pupil distance of 50 pixels. From the 1720 samples, four galleries were constructed by randomly assigning a sample of each subject to one of the galleries, which resulted in 430 samples per gallery. The comparative labels for the subjects were extracted from our previously crowdsourced labels in [29], which included 4038 subjects from the *View 1* subset of the LFW database (LFW-V1). Table II presents an overview on the crowdsourcing of comparative labels, and Fig. 3 shows an example crowdsourced labelling for an attribute.

III. SEMANTIC FACE SIGNATURES

The aim of the comparative soft biometric attributes used in this paper is to collectively create a descriptor that can uniquely identify individuals based on their facial traits. This descriptor is referred to as the semantic face signature, which is a feature vector that consists of the relative rate (i.e. strength) of each of the 24 comparative attributes shown in Table I. The creation of face signatures can be achieved either through the relative rating of attributes based on semantic labels, as

*The dataset is publicly available at http://github.com/almudhahka/lfw-ms4

explained in this section, or based on visual features generated from facial images as explained in Section IV.

A. Relative Rating of Attributes from Comparative Labels

The relative rate of an attribute of a particular subject, A, can be inferred from pairwise comparisons between A and other subjects in the dataset. We use the Elo rating system [38], which is a well-known algorithm for rating chess players, to predict the relative rates of the attributes from comparative labels, as its applicability and effectiveness for comparative soft biometrics have already been demonstrated [2], [7]. The rating process in the Elo system starts by initialising the rates of all players in a tournament to a default value. Then, for a game between players A and B with the initial rates R_A and R_B correspondingly, the expected scores, E_A and E_B, are calculated as follows:

$$E_A = \left[1 + 10^{(R_B - R_A)/400}\right]^{-1} \quad (1)$$

$$E_B = \left[1 + 10^{(R_A - R_B)/400}\right]^{-1} \quad (2)$$

Subsequently, based on the game's outcome (i.e. loss, win, or draw), the new rates, $\bar{R}_A$ and $\bar{R}_B$, for players A and B, respectively, are:

$$\bar{R}_A = R_A + K(S_A - E_A) \quad (3)$$

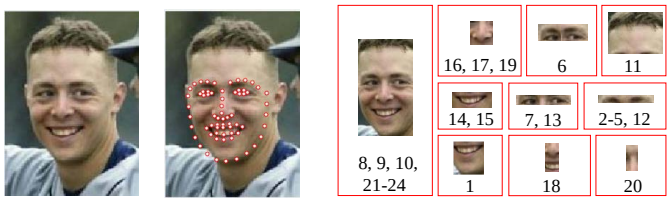
$$\bar{R}_B = R_B + K(S_B - E_B) \quad (4)$$

where S_A and S_B are scores that are set depending on the game outcome as: 0 for a loss, 1 for a win, and 0.5 for a draw, while K is the score adjustment factor that determines the sensitivity of rate update. The Elo rating system can be used to rate comparative attributes in a similar scheme to chess rating. Thus, by considering the subjects of the dataset as players in a tournament, and assuming that a comparison between two subjects, A and B, for a particular comparative attribute, X, is a game between two players, correspondingly, this comparison can result in one of three possible outcomes for each of the two players: "*Less X*", "*More X*", or "*Same X*", for example: "subject A has a *more thick* eyebrow than subject B".

The Elo rating system was used for each subject in the LFW-MS4 dataset and for each comparative attribute, to construct a semantic database that includes the semantic face signature of each subject as a vector of 24 relative rates, which correspond to the attributes in Table I. This semantic database was used for the retrieval experiments as we will see later in Section V.

IV. Visual Face Signatures

The existing work on comparative semantic facial attributes has addressed identification in semantic space (i.e. a semantic probe in a semantic database) [6], [29]. However, there has been no detailed investigation of the semantic gap between human and machine vision with respect to comparative facial attributes. Also, the efficacy of relatively rating attributes from visual features extracted from face images needs to be explored for identification and verification. In this section, we discover



(a) (b)

Fig. 4. (a) Facial landmarks estimated using the deformable model [39]. (b) Segmentation of face components with the attributes indices (as in Table I).

the gap between semantic and visual facial attributes. We present a framework to infer face signatures from images and examine the correspondence between semantic and visual attributes. The generation of visual face signatures, which is illustrated in Fig. 2, undergoes two major steps: (1) extracting visual features from face images; and (2) rating of visual features. The following subsections explain the process and discuss the findings in detail.

A. Extracting Visual Features from Faces

We follow a component-based approach that is similar to the approaches used in [40] and [4] to extract visual features from face images. The extraction of visual features from faces involves: (1) the estimation of facial landmarks locations; (2) the segmentation of face component images based on the estimated facial landmarks; and (3) the generation of visual features from the images of the segmented face components (as illustrated in Fig. 2). The estimation of facial landmarks locations can be achieved using a deformable model. The most well-known deformable models that have been used to estimate the locations of facial landmarks are the Active Shape Models (ASMs) [41], the Active Appearance Models (AAMs) [42] and the Constrained Local Models (CLMs) [43], [44]. All these models are based on flexible templates and exploit global shape constraints derived from training data to locate a set of facial landmarks on a new face image. The ASMs iteratively fit a shape model, which consists of the points making up the shape obtained from training images to a new image by increasing the match between the image and the model. The AAMs combine both shape and texture to match a model to an image, and have demonstrated more robustness as compared to the ASMs [45]. The CLMs follow a part-based approach in which a face image is sampled into a set of regions and the corresponding response maps, which represent the likelihood of having a particular landmark point at each pixel, are generated. Then, the parameters of the shape model are tuned to find the set of points that optimises the cost of the response maps. In this paper, a more recent and efficiently trained framework for face alignment in-the-wild [39] was used to locate facial landmarks and, accordingly, enable the segmentation of face components. This framework combines the shape model of the AAM approach with response maps that are generated by trained detectors (as in the CLMs) for face alignment. The approach has demonstrated its capability of handling face alignment in-the-wild. Also, it has shown a high efficiency. This framework was utilised to locate 66 points (shown in

Fig. 4(a)) on the LFW-MS4 dataset samples, which guided the segmentation of the face components that correspond to the soft biometric attributes as shown in Fig. 4(b).

The last step in the visual features extraction process is the generation of the GIST features [46] from the images of face components. The generation of the GIST features involves intensity normalisation for the image, and processing it through a series of Gabor filters in four scales and eight orientations per scale that yield 32 orientation maps. Each orientation map is divided into a 4×4 grid, and the average intensity is calculated for each block in the grid to form a vector of 512 features. The process was performed for each face component, which corresponds to a particular attribute (as illustrated in Fig. 4(b)).

B. Relative Rating of Visual Features

To predict the rates of the attributes from visual features, we use Support Vector Regressor (SVR) [47], which is a machine learning tool that is effective in high dimensional spaces and has high versatility through the use of various kernel methods for adapting to different feature spaces [48]. Given a dataset of n samples with the feature space, $x_n \in \mathbb{R}^m$, and the associated labels, y_n, correspondingly, the objective of the SVR model is to find the coefficients α_m and $\alpha_m{}^*$ that minimise the following loss function:

$$L(\alpha) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)K(x_i, x_j) + \epsilon\sum_{i=1}^{n}(\alpha_i + \alpha_i^*) - \sum_{i=1}^{n} y_i(\alpha_i - \alpha_i^*) \tag{5}$$

subject to

$$\begin{aligned} \sum_{k=1}^{n}(\alpha_k - \alpha_k^*) &= 0 \\ 0 \le \alpha_k \le C \quad \forall \quad 1 \le k \le n \\ 0 \le \alpha_k^* \le C \quad \forall \quad 1 \le k \le n \end{aligned} \tag{6}$$

where C determines the penalty of misclassification out of the margin ϵ, while $K(x_i, x_j)$ is the radial basis kernel function that is defined for two samples, x_i and x_j, as:

$$K(x_i, x_j) = \exp\left(-||x_i - x_j||^2 / \sigma^2\right) \tag{7}$$

Accordingly the relative rate of a new sample, x, can be predicted based on the following function:

$$f(x) = \sum_{k=1}^{n}(\alpha_k - \alpha_k^*)K(x_k, x) + b \tag{8}$$

By training the SVR model using the GIST features (as x_n) along with the corresponding normalised relative rates, which are inferred from the semantic space using the Elo rating system, as the training labels, y_n, the relative rate of an attribute for a new (unseen) subject can be predicted from the visual features extracted from the subject's face image as outlined in Fig. 2 and using Equation 8.

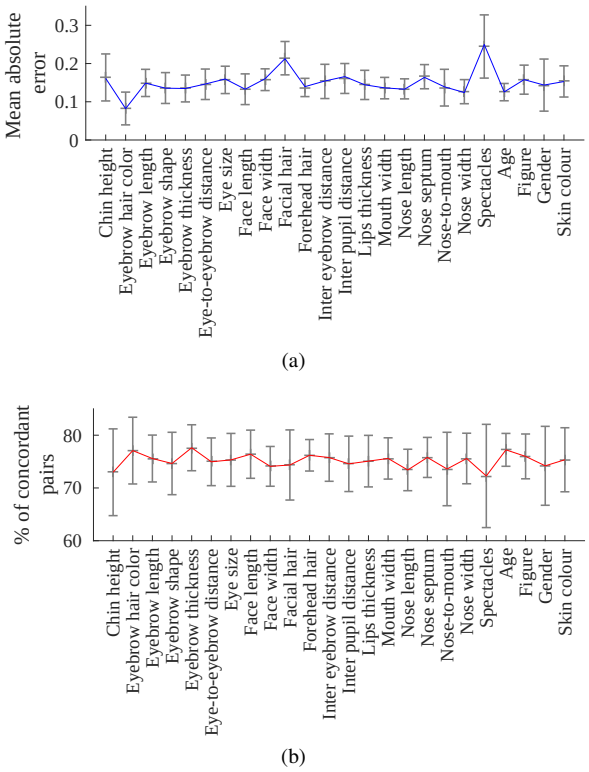


Fig. 5. Correspondence between visual and semantic space for the attributes reported using: (a) mean absolute error; and (b) level of concordance.

C. Vision-based Retrieval of Attributes

1) Experimental Design: The experimental design used to investigate vision-based retrieval of attributes from facial images follows a 10-fold cross validation with each of the 24 attributes defined in Table I. The 430 subjects of LFW-MS4 were randomly divided into ten folds, where each fold consists of 43 subjects. It is important to emphasise that the subjects in the test fold and the training folds are mutually exclusive (i.e. new test subjects). The training subjects with their samples (4 samples per subject) were used to train the SVR model based on the GIST visual features of the facial components associated with the attributes being retrieved as well as the normalised relative rates deduced using the Elo rating from the crowdsourced comparisons (as explained in Section III). Finally, the relative rates of the attribute being retrieved were inferred for each sample of each test subject using the trained SVR and the GIST visual features, while the performance was reported as the grand average outcome across the ten test folds and the four samples of each test subject.

2) Correspondence Analysis: The semantic correspondence of visual attributes, which were predicted using the SVR model, was determined using the following two measures:

- Mean Absolute Error (MAE). The mean absolute error for an attribute is computed for each test fold as the absolute difference between the normalised ground truth relative rate, r, and the predicted rate from visual features, $\hat{r}$, for each subject in the test set as follows:

(a) Age (b) Figure

(c) Gender (d) Skin colour

Fig. 6. Illustration for the outcomes of ranking the *global* attribute in both the semantic and the visual spaces, which are denoted as S and V respectively. For each attribute, the images in the top row represent the *top* rated subjects, while the images in the bottom row represent the *least* rated subjects with respect to the corresponding attribute. The semantic labels of the *top* and *least* for each attribute are listed in Table I.

$$\text{MAE} = \sum_{i=1}^{l} |r_i - \hat{r}_i| \,/\, l \tag{9}$$

where l is the number of subjects in the test fold. The MAE was evaluated for each test subject, and the average of the ten folds was used to report the result.

- Level of concordance. The purpose of this measure is to assess the accuracy of the approach in ranking subjects according to attribute strength. Given a ground truth rank, r, and a predicted rank, $\hat{r}$, for a pair of subjects, i and j, the pair is considered concordant if the following condition is satisfied:

$$r_i > r_j \quad \text{and} \quad \hat{r}_i > \hat{r}_j \tag{10}$$

or

$$r_i < r_j \quad \text{and} \quad \hat{r}_i < \hat{r}_j \tag{11}$$

while all the other possible relations between the ground truth and the predicted rank for a subject are considered discordant. For each subject, the level of concordance can be defined as the ratio between the concordant pairs and all the pairs that involve the test subject with the subjects of the training folds. This measure was evaluated for each subject in the dataset and reported as the average of the ten folds.

Fig. 5 shows the correspondence between semantic and visual attributes measured using MAE and level of concordance. Fig. 5(a) shows that the binary-like *facial* attributes (i.e. *facial hair* and *spectacles*) have the lowest accuracy in terms of MAE, and this might be due to the modelling of binary attributes in a relative continuous space. Another interesting finding is the high accuracy of *eyebrow hair colour* estimates, which can be attributed to the fact that *eyebrow hair colour* is an intensity-based attribute, and thus, it can be learned by machines with high correspondence. The most notable finding from the level of concordance analysis (shown in Fig. 5(b)) is the high concordance of *eyebrow thickness* and *eyebrow hair colour*, which support the emphasis on eyebrows in [29]. Also, the analysis demonstrates the significant uncertainty in the level of concordance among the dataset, which implies the challenging visual conditions of the dataset. Fig. 6 shows some examples that demonstrate the semantic gap between humans and machines in rating the *global* soft biometric attributes. By observing the outcomes of *age* rating in Fig. 6, we can see that the oldest subject in the visual space looks much older than the oldest subject in the semantic space. Also, the most masculine/least feminine subject in the visual space via relative *gender* has more *facial hair*, which is associated with masculinity, as compared with the equivalent subject in the semantic space. These findings might indicate that machines are effective in learning the visual features that are significant for rating *age* and *gender*. Additionally, Fig. 6 shows that the machine predictions of *skin colour* from a face can be very sensitive to noise. This can be seen from the image of the top-rated subjects in terms of *skin colour* in the visual space. It might be that the dark hat covering the subjects forehead has significantly affected the *skin colour* rating.

3) Attribute Stability: Attribute stability can be defined as the consistency of an attribute rate among different samples (i.e. semantic or visual face signatures), which is vital in evaluating the attribute effectiveness and robustness as a semantic descriptor [29]. In this analysis, we measure the stability of the attributes in both the semantic and visual spaces. Given that LFW-MS4 has four samples, which contribute to four visual galleries, we have respectively created four semantic galleries by randomly dividing all the comparisons into four mutually exclusive groups, and generating four different semantic face signatures for each of the 430 subjects. Then, the stability of the semantic or visual attributes was measured as the Pearson correlation between all the possible pairs of the four galleries (i.e. six pairs) for each attribute.

Fig. 7 shows the average stability represented in terms of Pearson correlation coefficient, *r*. Closer inspection of the results in Fig. 7 reveals several important aspects of the modelling of the semantic attribute in visual space. First, the binary-like *facial* attributes (i.e. *facial hair* and *spectacles*) have low stability in the visual space, while they have high stability in the semantic space. A possible explanation for this might be that machines are inaccurate when interpreting binary features in a relative continuous format. Second, whereas *age* and *eyebrow hair colour* are among the lowest attributes in terms of semantic stability, they have significantly high stability in the visual space. This can be attributed to human subjectivity in estimating these two attributes as opposed to the consistency of machines predictions, which are based on the visual features. The significant consistency of machines

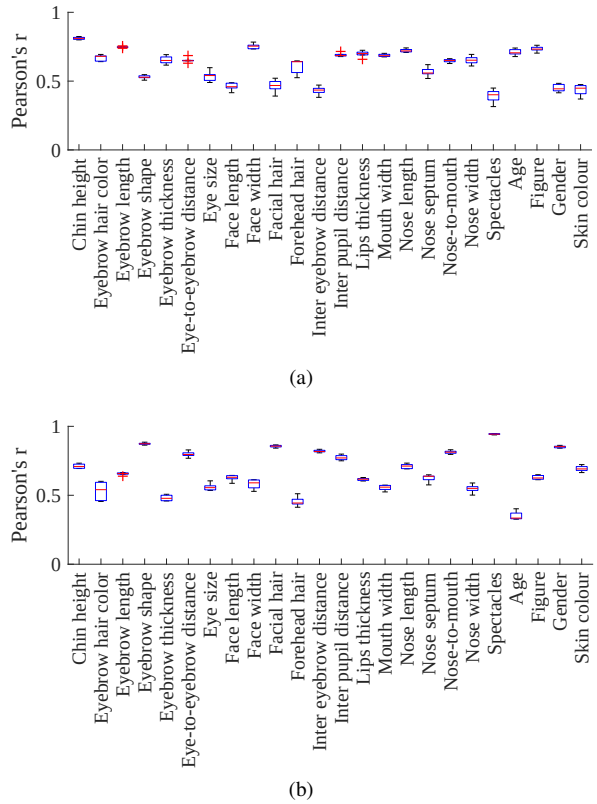


Fig. 7. Average attribute stability with the LFW-MS4 dataset: in (a) visual space; and (b) semantic space.

in estimating *age* as compared with humans agrees with the finding of Han et al. in [49] that the automatic estimates of *age* from face images are less biased than human estimates. Another interesting finding that highlights the gap between humans and machines is the high stability of *skin colour* in the semantic space as compared with its low stability in the visual space. This might be due to the precision of machines in extracting colour information relative to higher level interpretation of colour by humans [10]. Finally, the most important finding to emerge from the correspondence and stability analysis is that machines and humans differ considerably in their interpretation and evaluation of the human face attributes. Also, the modelling of semantic attributes in visual space results in a different impact on attribute consistency.

V. Experiments

The experiments in this paper simulate different scenarios that all aim to retrieve the identity of an unknown subject (probe) from a database. We explore human face identification via comparative soft biometrics using two types of databases:

- Semantic database, DB_s, in which each subject in the dataset is presented by its semantic face signature, which is inferred from the crowdsourced comparative labels.
- Visual database, DB_v, which contains the identity of each subject in the form of a visual face signature, which is predicted from the subject's face image as illustrated in Fig. 2, and thus, a visual database is equivalent to a database of face images.

Similarly, we use two types of probes to search a database:

- Semantic probe, PR_s, which represents a verbal description for an unknown subject's face to be identified, and it is constructed from crowdsourced comparisons using the Elo rating system. A semantic probe can also be considered as a semantic face signature that is deduced from a relatively small number of subject comparisons.
- Visual probe, PR_v, is a visual face signature that is used to search a visual database, DB_v, imitating a scenario in which a sample facial image is used to identify an unknown subject from a database of facial images.

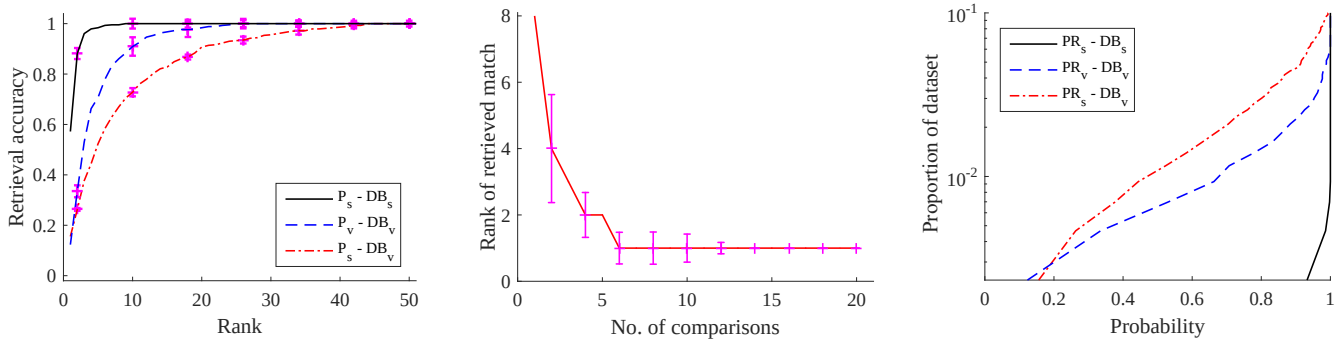
Two types of experiments are presented in this section: (a) subject retrieval from a semantic database DB_s using a semantic probe PR_s; and (b) subject retrieval from a visual database, DB_v, which is performed with semantic and visual probes (PR_s and PR_v respectively). As mentioned earlier, the experiments were conducted using the LFW-MS4 dataset.

A. Retrieval from a Semantic Database using a Verbal Description

This experiment simulates a realistic scenario in which a semantic database DB_s is searched to identify an unknown subject using a verbal description for the subject's face as illustrated in Fig. 1. The identification performance evaluation follows a 10-fold cross validation, where the 430 subjects of the LFW-MS4 dataset were randomly divided into ten equal subsets, and each subset was used for testing while the remaining nine folds were used for training. For each subject in the test set, a probe semantic face signature PR_s was generated from the relative rates of the 24 attributes (listed in Table I). The relative rates were computed using the Elo rating system and based on comparisons between the probe and c other randomly selected subjects from the training folds. The remaining comparisons, after excluding those used for generating PR_s, were used to generate a semantic face signature for each subject, which made up the database DB_s to be searched. Then, the distance, d_p, between the probe and each subject in DB_s was calculated using the Pearson correlation coefficient as follows:

$$d_p = 1 - \frac{\sum_{i=1}^{t}(PR_s(i) - \overline{PR}_s)(S_c(i) - \overline{S}_c)}{\sqrt{\sum_{i=1}^{t}(PR_s(i) - \overline{PR}_s)^2}\sqrt{\sum_{i=1}^{t}(S_c(i) - \overline{S}_c)^2}} \tag{12}$$

where $S_c \in DB_s$ is the semantic face signature of the counterpart subject in the database DB_s that is being compared against the probe, and t is the number of attributes composing the semantic face signature. The rank of the closest match (i.e. shortest distance) to the probe was used to report the identification performance via a Cumulative Match Characteristic (CMC) curve. This cross validation ran over the ten folds and was repeated 100 times. The mode rank among the 100 trials was selected to report the performance, as it relates to the majority voting by eyewitnesses on a subject description. Fig. 8(a) shows the CMC curve resulting from this experiment (the P_s in DB_s scenario) and it can be seen that using ten subject comparisons to generate the probe semantic



(a) (b) (c)

Fig. 8. (a) retrieval performance of attributes in LFW-MS4 reported with different combinations of probe, *PR*, and database, *DB*, where *S* and *V* refer to semantic and visual entities respectively (b) the effect of the number of comparisons used for generating a semantic probe, PR_s on retrieval performance in a semantic database DB_s; and (c) compression achieved in search range with the probability of finding a correct match within the range.

face signature, which is the ideal size of identity parade [2], a rank-5 identification rate of 98.37% can be achieved. Moreover, a correct match will always be returned in the top 9 retrieved subjects (the top 2.09% subjects of the dataset). Fig. 8(b) shows the effect of the number of comparisons used for generating the probe PR_s on the retrieval performance in terms of the rank at which a correct match is retrieved, it can be noted that the accuracy and certainty of the outcomes of the retrieval increase as more comparisons are used for generating the probe.

The implications of using comparative semantic attributes on face identification can also be seen from another interesting perspective, which is compression of the search range in a database. Thus, narrowing down search range becomes vital for the efficiency of identification in a large database. In addition, when verbal descriptions are not sufficiently accurate, search compression can lead to filtering out a long list of suspects, making subject retrieval more efficient. Fig. 8(c) demonstrates the compression in search range that can be achieved in the LFW-MS4 dataset using the comparative facial soft biometrics. It shows that using ten comparisons only; the search range can be narrowed down to four subjects with probability $p = 1$ that a correct match with the subject will be found. In [4], a set of 46 attributes (19 binary and 27 categorical) resulted in an accuracy of 22.5% at rank-1 using 1196 subjects from the FERET database. Our approach can reach a rank-1 accuracy of 57.21% using ten subject comparisons and 24 comparative attributes only. These performance gains are due to the advantage of using comparative attributes [2].

B. Retrieval from a Visual Database

The objective of this experiment is to explore subject retrieval from a database of facial images DB_v in which each subject is represented by its visual face signature. The retrieval is examined here using semantic face signatures PR_s as well as visual face signatures PR_v as probes. These two scenarios simulate searching a database of facial images (e.g. mugshots) to retrieve the identity of an unknown subject based on a verbal description of the subject's face or a sample face image. For this purpose, we use the LFW-MS4 database, which has 430 subjects with four sample face images per subject. The experimental design that was used to deduce the visual face signatures, which construct the visual database DB_v, for both scenarios (i.e. semantic-based and image-based retrieval) follows a 10-folds cross validation as previously explained in Section IV-C, where the test subjects are all new (unseen). In addition, as each subject in the LFW-MS4 dataset has four sample facial images, four visual databases were constructed correspondingly in these two retrieval experiments $\{DB_{v1}, ..., DB_{v4}\}$.

1) Retrieval by Semantic Descriptions: In this experiment, a semantic probe, PR_s, was generated for each test subject using ten randomly selected counterpart subjects from the training set. As explained previously, the number of comparisons was chosen as ten because it is the average size of an ideal identity parade [2]. Then, retrieval was performed by measuring the L_1 distance between PR_s and the visual signature of each subject in the visual database DB_v. The returned subjects were ranked according to their similarity with the probe, and the rank of the correct match was reported for performance evaluation. The experiment was repeated 100 times with each of the four visual databases $\{DB_{v1}, .., DB_{v4}\}$. In each of the 100 trials, different subject comparisons were randomly selected and used to generate the probe PR_s. For each subject, the mode rank among the 100 trials and the 4 visual databases was used for the performance reporting via the CMC curve as shown in Fig. 8(a), from which it can be observed that although the retrieval accuracy is very low at rank-1, it dramatically increases to 72.79% at rank-10 and reaches 90.7% at rank-20. Also, from Fig. 8(c), we can see that the approach can result in narrowing the search range for the probe subject in the database to 10.23%.

2) Retrieval by Face Image: Retrieval by face image is evaluated by using a visual probe PR_v to identify an unknown subject from the four visual databases $\{DB_{v1}, .., DB_{v4}\}$ following the same experimental design detailed in Section IV-C. Thus, the visual face signature of each subject in the test set is zero-shot learned. The experiment was conducted using all the possible combinations of the four visual databases, and thus, it involves six different trials, (DB_{vi}, DB_{vj}), where

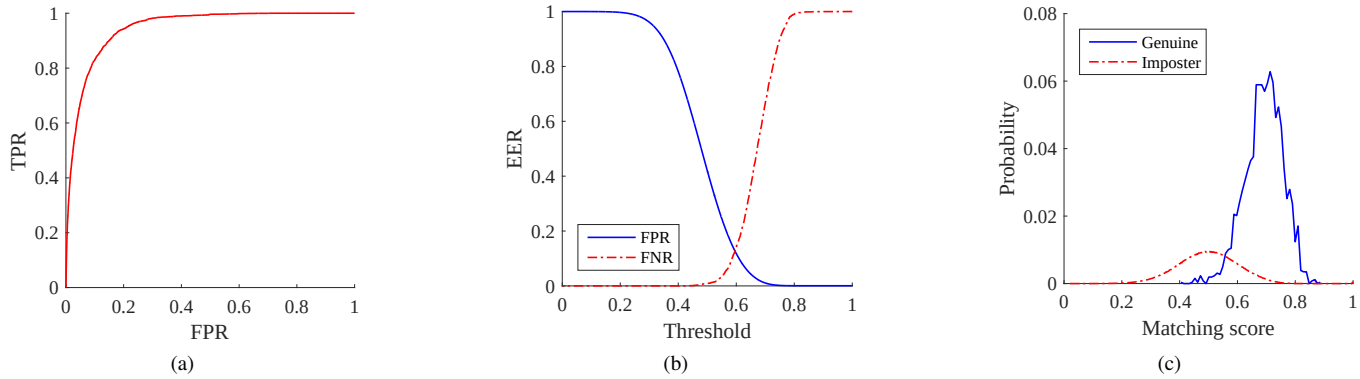


Fig. 9. Verification performance of attributes in LFW-MS4 reported using the four visual galleries. (a) Receiver Operating Characteristic (ROC) curve in terms of True Positive Rate (TPR) and False Positive Rate (FPR); (b) error curves in terms of False Positive Rate (FPR) and False Negative Rate (FNR); and (c) genuine and imposter distributions.

$(i, j) \in \{1, 2, 3, 4\}$. In each trial, the visual face signatures of the test subjects in one visual database DB_{vi} were used to retrieve the subjects' identities from the other visual database DB_{vj}. The similarity between the probe in DB_{vi} and each subject in DB_{vj} was measured based on the L_1 distance. The subjects of DB_{vj} were sorted according to their distance with the probe in DB_{vi} and the rank of the correct match was used to report the retrieval accuracy. This testing procedure was applied to the six possible combinations of the four galleries and the harmonic mean of the ranks resulting for each subject was reported as the experiment outcome as shown in Fig. 8(a). By using visual face signatures, a rank-10 retrieval accuracy of 90.93% can be attained. Moreover, a correct match will always be found in the top 5.81% returned subjects. By examining the verification performance, the visual face signatures can achieve an Equal Error Rate (EER) of 12.71% and an Area Under the Curve (AUC) of 94.53% (as can be seen in Fig. 9). The approach proposed in [16] used shape and size features of facial regions as soft biometrics and achieved rank-10 accuracies of 100% and 75% with the relatively constrained ATVS and MORPH databases respectively, while the best EER achieved in [16] was 3.06% and 12.27% for ATVS and MORPH databases correspondingly. In [18], categorical facial attributes were proposed for active authentication on mobile devices and achieved an EER of 14% with the MOBIO dataset, which is composed of video footage acquired by mobile cameras for 152 subjects. The retrieval and verification accuracies achieved by our approach with the unconstrained LFW-MS4 dataset demonstrate the impact of utilising comparative soft biometrics to create a visual space for facial attributes.

The experiments presented in this section demonstrate the impact of the semantic gap between humans and machines on the retrieval performance. This gap might be attributed to several psychological factors to which humans are exposed. First, humans are influenced by the *other-race* effect. Thus, it was shown that the recognition accuracy of humans improves when recognising faces of individual from the observers' races [50]. On the other hand, the automated approaches are passive with respect to race, as they deal with low-level visual features. Second, it was found that face familiarity significantly affects human face recognition accuracy [51], and this might also applies to the accuracy of human annotators in labelling attributes of faces that are familiar to them (e.g. public figures). Finally, the human visual system processes facial attributes in a holistic form [34], while the automatic approach in this paper predicts the attributes from visual features that were extracted from images of individual face components (as illustrated in Fig. 2).

VI. Conclusion

This paper examines identity retrieval by verbal descriptions using semantic and visual face signatures which are generated from comparative soft biometrics and images, respectively. It proposes a framework for exploiting the semantic space to infer facial attributes from images and investigates the correspondence between semantic and visual attributes. The experiments demonstrate that using verbal descriptions for identity retrieval from a visual database, constructed from visual face signatures, can yield a correct match in the top 10.23% returned subjects. Furthermore, the experiments demonstrate that modelling visual features of the human face in the semantic space shows promising retrieval and verification performance. The findings of this paper enhance our understanding of the semantic-visual correspondence of relative facial attributes and promote the exploration of more comparative facial soft biometrics that can contribute to further improvements in retrieval and verification performance.

References

[1] Sina Samangooei, Baofeng Guo, and Mark S. Nixon. The use of semantic human description as a soft biometric. In *Biometrics Theory, Applications and Systems (BTAS), 2008 IEEE 2nd International Conference on*, pages 1–7. IEEE, 2008.

[2] Daniel A. Reid, Mark S. Nixon, and Sarah V. Stevenage. Soft biometrics; human identification using comparative descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(6):1216–1228, 2014.

[3] Mark S. Nixon, Paulo L. Correia, Kamal Nasrollahi, Thomas B. Moeslund, Abdenour Hadid, and Massimo Tistarelli. On soft biometrics. *Pattern Recognition Letters*, 68:218–230, 2015.

[4] Brendan F. Klare, Scott Klum, Joshua C. Klontz, Emma Taborsky, Tayfun Akgul, and Anil K. Jain. Suspect identification based on descriptive facial attributes. In *Biometrics (IJCB), 2014 IEEE International Joint Conference on*, pages 1–8. IEEE, 2014.

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

[6] Daniel A. Reid and Mark S. Nixon. Human identification using facial comparative descriptions. In *Biometrics (ICB), 2013 International Conference on*, pages 1–7. IEEE, 2013.

[7] Nawaf Y. Almudhahka, Mark S. Nixon, and Jonathon S. Hare. Human face identification via comparative soft biometrics. In *Identity, Security and Behavior Analysis (ISBA), 2016 IEEE International Conference on*, pages 1–6. IEEE, 2016.

[8] Daniel Martinho-Corbishley, Mark S. Nixon, and John N. Carter. Soft biometric recognition from comparative crowdsourced annotations. *IET Biometrics*, pages 1–16, 2016.

[9] Daniel A. Reid and Mark S. Nixon. Using comparative human descriptions for soft biometrics. In *Biometrics (IJCB), 2011 International Joint Conference on*, pages 1–6. IEEE, 2011.

[10] Emad Sami Jaha and Mark S. Nixon. From clothing to identity: Manual and automatic soft biometrics. *IEEE Transactions on Information Forensics and Security*, 11(10):2377–2390, 2016.

[11] Anil K. Jain, Sarat C. Dass, and Karthik Nandakumar. Soft biometric traits for personal recognition systems. In *Biometric Authentication*, pages 731–738. Springer, 2004.

[12] Devi Parikh and Kristen Grauman. Relative attributes. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 503–510. IEEE, 2011.

[13] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, Technical Report 07-49, University of Massachusetts, Amherst, 2007.

[14] Jamie D. Shutler, Michael G. Grant, Mark S. Nixon, and John N. Carter. On a large sequence-based human gait database. In *Applications and Science in Soft Computing*, pages 339–346. Springer, 2004.

[15] P Jonathon Phillips, Hyeonjoon Moon, Syed A. Rizvi, and Patrick J. Rauss. The feret evaluation methodology for face-recognition algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(10):1090–1104, 2000.

[16] Pedro Tome, Ruben Vera-Rodriguez, Julian Fierrez, and Javier Ortega-Garcia. Facial soft biometric features for forensic face recognition. *Forensic Science International*, 257:271–284, 2015.

[17] Karl Ricanek and Tamirat Tesafaye. Morph: A longitudinal image database of normal adult age-progression. In *Automatic Face and Gesture Recognition (FG), 2006 IEEE 7th International Conference on*, pages 341–345. IEEE, 2006.

[18] Pouya Samangouei, Vishal M. Patel, and Rama Chellappa. Facial attributes for active authentication on mobile devices. *Image and Vision Computing*, 58:181–192, 2017.

[19] Christopher McCool, Sebastien Marcel, Abdenour Hadid, Matti Pietikäinen, Pavel Matejka, Jan Cernockỳ, Norman Poh, Josef Kittler, Anthony Larcher, Christophe Levy, et al. Bi-modal person recognition on a mobile phone: using mobile phone data. In *Multimedia and Expo Workshops (ICMEW), 2012 IEEE International Conference on*, pages 635–640. IEEE, 2012.

[20] Max Ehrlich, Timothy J. Shields, Timur Almaev, and Mohamed R. Amer. Facial attributes classification using multi-task representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 47–55, 2016.

[21] H. Han, A. K. Jain, S. Shan, and X. Chen. Heterogeneous face attribute estimation: A deep multi-task learning approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP(99):1–1, 2017.

[22] Emily M. Hand and Rama Chellappa. Attributes for improved attributes: A multi-task network utilizing implicit and explicit relationships for facial attribute classification. In *AAAI*, pages 4068–4074, 2017.

[23] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3730–3738, 2015.

[24] Nawaf Y. Almudhahka, Mark S. Nixon, and Jonathon S. Hare. Automatic semantic face recognition. In *Automatic Face and Gesture Recognition (FG), 2017 IEEE 12th International Conference on*, pages 180–185. IEEE, 2017.

[25] Daniel Martinho-Corbishley, Mark S. Nixon, and John N. Carter. Retrieving relative soft biometrics for semantic identification. In *Pattern Recognition (ICPR), 2016 23rd International Conference on*, pages 3067–3072. IEEE, 2016.

[26] Aron Yu and Kristen Grauman. Just noticeable differences in visual attributes. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2416–2424, 2015.

[27] Daniel Reid, Sina Samangooei, Cunjian Chen, Mark Nixon, and Arun Ross. Soft biometrics for surveillance: an overview. *Machine Learning: Theory and Applications*, pages 327–352, 2013.

[28] Nigel D. Haig. Exploring recognition with interchanged facial features. *Perception*, 15(3):235–247, 1986.

[29] Nawaf Y. Almudhahka, Mark S. Nixon, and Jonathon S. Hare. Unconstrained human identification using comparative facial soft biometrics. In *Biometrics Theory, Applications and Systems (BTAS), 2016 IEEE 8th International Conference on*, pages 1–6. IEEE, 2016.

[30] Lowell L. Kuehn. Looking down a gun barrel: Person perception and violent crime. *Perceptual and Motor Skills*, 39(3):1159–1164, 1974.

[31] Christian A. Meissner, Siegfried L. Sporer, and Jonathan W. Schooler. Person descriptions as eyewitness evidence. *Handbook of Eyewitness Psychology: Memory for People*, pages 1–34, 2013.

[32] Peter J. Van Koppen and Shara K. Lochun. Portraying perpetrators: The validity of offender descriptions by witnesses. *Law and Human Behavior*, 21(6):661, 1997.

[33] Javid Sadr, Izzat Jarudi, and Pawan Sinha. The role of eyebrows in face recognition. *Perception*, 32(3):285–293, 2003.

[34] Pawan Sinha, Benjamin Balas, Yuri Ostrovsky, and Richard Russell. Face recognition by humans: Nineteen results all computer vision researchers should know about. *Proceedings of the IEEE*, 94(11):1948–1962, 2006.

[35] Graham Davies, Hadyn Ellis, and John Shepherd. Cue saliency in faces as assessed by the 'Photofit' technique. *Perception*, 6(3):263–269, 1977.

[36] Gary B. Huang and Erik Learned-Miller. Labeled faces in the wild: Updates and new reporting procedures. *Dept. Comput. Sci., Univ. Massachusetts Amherst, Amherst, MA, USA, Tech. Rep*, pages 14–003, 2014.

[37] Gary Huang, Marwan Mattar, Honglak Lee, and Erik G. Learned-Miller. Learning to align from scratch. In *Advances in Neural Information Processing Systems*, pages 764–772, 2012.

[38] Arpad E. Elo. *The rating of chessplayers, past and present*. Arco Pub., 1978.

[39] Akshay Asthana, Stefanos Zafeiriou, Georgios Tzimiropoulos, Shiyang Cheng, and Maja Pantic. From pixels to response maps: Discriminative image filtering for face alignment in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(6):1312–1320, 2015.

[40] Hu Han, Brendan F. Klare, Kathryn Bonnen, and Anil K. Jain. Matching composite sketches to face photos: A component-based approach. *IEEE Transactions on Information Forensics and Security*, 8(1):191–204, 2013.

[41] Timothy F. Cootes, Christopher J. Taylor, David H. Cooper, and Jim Graham. Active shape models-their training and application. *Computer Vision and Image Understanding*, 61(1):38–59, 1995.

[42] Timothy F. Cootes, Gareth J. Edwards, and Christopher J. Taylor. Active appearance models. In *European Conference on Computer Vision*, pages 484–498. Springer, 1998.

[43] David Cristinacce and Tim Cootes. Automatic feature localisation with constrained local models. *Pattern Recognition*, 41(10):3054–3067, 2008.

[44] Jason M. Saragih, Simon Lucey, and Jeffrey F Cohn. Deformable model fitting by regularized landmark mean-shift. *International Journal of Computer Vision*, 91(2):200–215, 2011.

[45] Timothy F. Cootes, Gareth J. Edwards, and Christopher J. Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):681–685, 2001.

[46] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001.

[47] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[48] Alex J. Smola and Bernhard Schölkopf. A tutorial on support vector regression. *Statistics and Computing*, 14(3):199–222, 2004.

[49] Hu Han, Charles Otto, Xiaoming Liu, and Anil K. Jain. Demographic estimation from face images: Human vs. machine performance. *IEEE transactions on pattern analysis and machine intelligence*, 37(6):1148–1161, 2015.

[50] Alice J. O'Toole. Psychological and neural perspectives on human face recognition. In *Handbook of Face Recognition*, pages 349–369. Springer, 2005.

[51] A. Mike Burton, Stephen Wilson, Michelle Cowan, and Vicki Bruce. Face recognition in poor-quality video: Evidence from security surveillance. *Psychological Science*, 10(3):243–248, 1999.

Nawaf Yousef Almudhahka received the Masters Degree in Computer Engineering from North Carolina State University. He is currently working towards a Ph.D. at the University of Southampton researching into human identification using comparative facial soft biometrics. He is currently an Assistant Teacher in the College of Technological Studies, PAAET, Kuwait. His research interests include soft biometrics, machine learning, and computer vision.

Mark S. Nixon is currently a Professor of Computer Vision at the University of Southampton, U.K. His research interests are in image processing and computer vision. His team develops new techniques for static and moving shape extraction, which have found application in automatic face and automatic gait recognition and in medical image analysis. His team were early workers in face recognition, later came to pioneer gait recognition, and then joined the pioneers of ear biometrics. More recently his team have been pioneering soft biometrics for human identification. He has supervised many research projects with funding from sources including DARPA, US Army, EPSRC, NERC, EU, and General Dynamics. His vision textbook, with A. Aguado, Feature Extraction and Image Processing, Third Edition (Academic Press, 2012) and has become a standard text in computer vision. With T. Tan and R. Chellappa, their book Human ID Based on Gait is part of the Springer Series on Biometrics and was published in 2005. Mark has chaired/program chaired many conferences (BMVC, ICPR, IEEE BTAS, IAPR ICB, IAPR/IEEE IJCB, IEEE ISBA and most recently IAPR/IEEE IJCB 2017) and given many invited talks. Mark is the President of the IEEE Biometrics Council and a member of IAPR TC4 Biometrics. Mark is a Fellow of IET and the IAPR, and a Distinguished Fellow of the BMVA.

Jonathon S. Hare is a lecturer in Computer Science at the University of Southampton. He holds a BEng degree in Aerospace Engineering and Ph.D. in Computer Science. His research interests lie in the area of multimedia data mining, analysis and retrieval, with a particular focus on large-scale multimodal approaches. This research area is at the convergence of machine learning and computer vision, but also encompasses other modalities of data. The long-term goal of his research is to innovate techniques that can allow machines to understand the information conveyed by multimedia data and use that information for fulfil the information needs of humans.