Unconstrained Human Identification Using Comparative Facial Soft Biometrics

Nawaf Y. Almudhahka Mark S. Nixon Jonathon S. Hare
University of Southampton
Southampton, United Kingdom
{nya1g14,msn,jsh2}@ecs.soton.ac.uk

Abstract

Soft biometrics are attracting a lot of interest with the spread of surveillance systems, and the need to identify humans at distance and under adverse visual conditions. Comparative soft biometrics have shown a significantly better impact on identification performance compared to traditional categorical soft biometrics. However, existing work that has studied comparative soft biometrics was based on small datasets with samples taken under constrained visual conditions. In this paper, we investigate human identification using comparative facial soft biometrics on a larger and more realistic scale using 4038 subjects from the View 1 subset of the LFW database. Furthermore, we introduce a new set of comparative facial soft biometrics and investigate the effect of these on identification and verification performance. Our experiments show that by using only 24 features and 10 comparisons, a rank-10 identification rate of 96.98% and a verification accuracy of 93.66% can be achieved.

1. Introduction

Systems for the automatic identification of people have traditionally been based on hard biometrics, such as iris images, fingerprints and DNA, which require the individual cooperation to be acquired [1]. The need to identify people at distance, and under challenging visual conditions, has motivated research in soft biometrics, which are physical and behavioural attributes that can be used to identify humans [1]. Soft biometrics bridge the semantic gap between human descriptions and biometrics, enabling the identification of humans in databases based solely on a semantic descriptions (i.e. an eyewitness statement) with soft traits that address invariance and subjectivity issues in human face descriptions. Several recent studies investigating the use of soft biometrics for human identification have included: face [2, 3, 4, 5], body [1, 6], and clothing [7].

Most of the literature surrounding soft biometrics focuses on categorical labelling, where individual traits are assigned to specific classes (e.g. a subject's *jaw shape* might be labeled as *square* versus *round*). It has however been shown that using comparative soft labels for human identification and retrieval improves the performance as compared to traditional categorical labels [6]. Comparative labels are generated by comparing two subjects (e.g. subject A has a *more rounded jaw* than subject B). The use of comparative facial soft biometrics for identification was studied in [4, 5], however, the datasets that were used in these studies [8, 9] were relatively small and the images were taken under constrained conditions. Real surveillance scenarios scale to a much larger population size with high variability in pose, illumination, facial expressions, resolution, and demographics. Categorical soft biometrics for unconstrained face verification have been previously investigated by different researchers (e.g. [10, 11]). However, to the best of our knowledge, no study has examined the use of comparative soft biometrics for face identification or verification in unconstrained conditions. A study of this nature is essential to assess the reliability and scalability of comparative soft biometrics for large and unconstrained datasets. The purpose of this paper is to study human identification and verification via comparative facial soft biometrics in large and unconstrained datasets using the Labelled Faces in The Wild (LFW) database [12]. In addition, we investigate the significance of different soft biometric features. Normally, a standard LFW experiment will use View 1 for algorithm development and View 2 for performance reporting; each

Figure 1: CCTV for three men sought by police in connection with racist attack by Chelsea fans on the Paris Metro in 2015. http://www.bbc.co.uk/news/uk-31558168

of these views contains pairs of images (either matching or mismatching). This arrangement is not well suited to our experiments as we are not interested in learning low level features or building a model to verify unseen subjects. Rather, we wish to characterize how well comparative soft biometrics perform when used to identify a target on the basis of comparative labels against a set of other subjects - in essence we wish to model how well the comparative biometrics work with respect to performing identification on the basis of an eyewitness statement formed by comparing the eyewitnesses mental model of the target against a set of other subjects. It should be noted that our method takes a target image (and associated set of soft biometric labels) and as output produces a ranking against a database of subjects; the aim is to have the target subject (represented by a different image of the same person) appear as high as possible in the ranking. As such we perform our experimental validation only on a subset of View 1 from LFW. Throughout the paper we use the terms ”trait”, ”soft biometric”, and ”feature” synonymously. Our main contributions are summarized as follows:

- Investigation of the performance of human identification using comparative soft biometrics at a larger scale under more challenging and realistic conditions than previous works.
- Provision of a framework for studying human identification using comparative soft biometrics for large unconstrained datasets.
- Assessment of the significance of a new set of comparative facial soft biometrics with respect to identification and verification performance.
- Creation of a dataset of 241560 crowdsourced comparisons that describe 4038 subjects from View 1 of LFW, based on 24 traits, that will be made publicly available.

The remainder of the paper is structured as follows: Section 2 presents our facial soft biometrics and comparative labels. Section 3 describes the collection of the comparative labels on the dataset, in addition to providing statistical analysis. Section 4 describes the experimental design and presents the results with discussions. Finally, Section 5 summarizes our findings and their implications.

2. Facial Soft Biometrics

As the objective behind using soft biometrics for identification is to enable searching a database based on an eyewitness statement, a soft biometric feature is required to be: understandable, memorable, and describable. The human face is rich in features that can be used to identify people at distance [1], although they differ in the extent to which they can be semantically described. On the basis of previous work [4], and considering these requirements, we have created a set of 24 soft biometric traits (20 facial and 4 global) that covers the major facial components (eyes, eyebrows, nose, and mouth) with an emphasize on eyebrow, due to its role in human face recognition [13], and nose, as it is high invariance to expressions [14]. All traits used to compose our soft biometrics set are comparative, including *facial hair* and *spectacles*, which are binary in nature but were expressed in a comparative format in our set. Also, global soft biometrics (*age*, *figure*, *gender*, and *skin colour*) were included in the set based on their high discriminative power [1]. Table 1 lists the soft biometrics used in this study along with their comparative labels. The comparative labels associated with each trait represent the strength of the trait in a person relative to a counterpart person as: *”More X”* or *”Less Y”*; *”Same”*; or *”More Y”*. Each comparative label is assigned a numerical value based on three-point bipolar scale that ranges from -1 to 1. The label value is used in computing the relative strength of the subjects’ traits based on the Elo rating system [15] as explained in Section 3.

No.	Trait	Labels
1	Chin height	[More Small, Same, More Large]
2	Eyebrow hair colour	[More Light, Same, More Dark]
3	Eyebrow length	[More Short, Same, More Long]
4	Eyebrow shape	[More Low, Same, More Raised]
5	Eyebrow thickness	[More Thin, Same, More Thick]
6	Eye-to-eyebrow distance	[More Small, Same, More Large]
7	Eye size	[More Small, Same, More Large]
8	Face height	[More Short, Same, More Long]
9	Face width	[More Narrow, Same, More Wide]
10	Facial hair	[Less Hair, Same, More Hair]
11	Forehead hair	[Less Hair, Same, More Hair]
12	Inter eyebrow distance	[More Small, Same, More Large]
13	Inter pupil distance	[More Small, Same, More Large]
14	Lips thickness	[More Thin, Same, More Thick]
15	Mouth width	[More Narrow, Same, More Wide]
16	Nose length	[More Short, Same, More Long]
17	Nose septum	[More Short, Same, More Long]
18	Nose-mouth distance	[More Short, Same, More Long]
19	Nose width	[More Narrow, Same, More Wide]
20	Spectacles	[Less Covered, Same, More Covered]
21	Age	[More Young, Same, More Old]
22	Figure	[More Thin, Same, More Thick]
23	Gender	[More Feminine, Same, More Masculine]
24	Skin colour	[More Light, Same, More Dark]

Table 1: Soft biometrics used in comparative labelling.

Figure 2: Example question from the crowdsourced job launched to collect comparative labels.

	Collected	Inferred	Total
Traits comparisons	241560	132879504	133121064
Subjects' comparisons	10065	5536646	5546711
Average number of comparisons per subject	4.98	1371.1	N/A
Number of annotators	9901	N/A	N/A

Table 2: The number of collected and inferred comparative labels.

3. Dataset and Label Acquisition

3.1. Labeled Faces in The Wild (LFW)

Labeled Faces in the Wild (LFW) [12] is a well-known database for studying unconstrained face recognition that consists of more than 13000 facial images derived from the web. LFW images reflect reasonably realistic surveillance conditions as they are greatly affected by variations in lighting, expressions, and to some extent pose, in addition to occlusion and low resolution. In this paper, we have used the 4038 subjects of the training set from the View 1 subset of the LFW database with all the images aligned using deep funneling [16]. A single sample was extracted for each person, and random selection was applied whenever multiple samples for a person exist. All the images were normalized in size to have an inter-pupil distance of 50 pixels, by following a similar approach to [17]. This was to ensure consistent comparisons between the subjects of the dataset.

3.2. Data Acquisition Through Crowdsourcing

The 4038 subjects in our dataset would result in over 8 million pairwise comparisons. Due to the infeasibility of collecting such a massive volume of comparisons, a map was drawn to relate the 4038 subjects in a way that involves each subject in at least 4 comparisons with other counterparts, whilst maximizing the potential of relation inference among the subjects. As a result, 10065 subject-to-subject comparisons were identified. Crowdsourcing was used to collect the comparative soft biometrics listed in Table 1 for each pair. Crowdsourcing was launched via the CrowdFlower platform (see Figure 2 for an example) and resulted in 241560 comparative labels for the different traits. This enables the inference of a much larger number of comparative labels. Additional statistics from the crowdsourcing are shown in Table 2.

3.3. Relative Rating of Traits

In Section 2, we mentioned that each comparative label is mapped to a numerical value that range between -1 and 1 according to the strength of the trait in subject A relative to subject B. The strength of each trait for any subject (relative rate) is calculated based on the Elo rating system [15], which is a popular system that is used for rating players in chess competitions based on the players expected and actual scores. For a game between two players A and B with the rates R_A and R_B respectively, the expected score, E, for each player is calculated as:

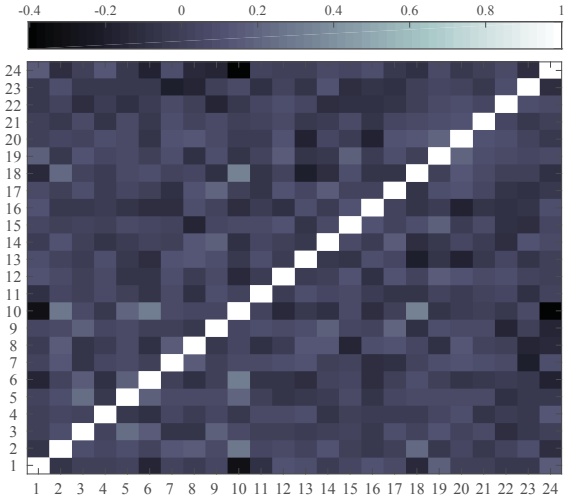


Figure 3: Soft biometrics correlation matrix.

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

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

Then, 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 is the actual score of the player (based on the game result: 0, 0.5, or 1.0 for loss, draw, or win respectively) and K is the score adjustment factor that determines the sensitivity of rate update. With regard to our study, a comparison between two subjects is considered as a game between two players, while the comparative label value, which ranges between -1 and 1, is normalized and set as the score, S, of the subject. The relative rate of each trait is used to create a biometric signature that describes a subject. This can then be used for identification as described in Section 4.

3.4. Trait Significance

Exploring the significance of the various soft biometrics is essential to revealing the extent to which each trait is contributing in discriminating the subjects in the dataset and to discover any statistical dependencies among the traits. In this subsection, three types of analysis are presented: (1) correlation analysis; (2) trait stability analysis, which reflects the level of agreement on a certain trait by different groups of annotators; and (3) trait discriminating power analysis, which shows the relative impact of a trait on human identification.

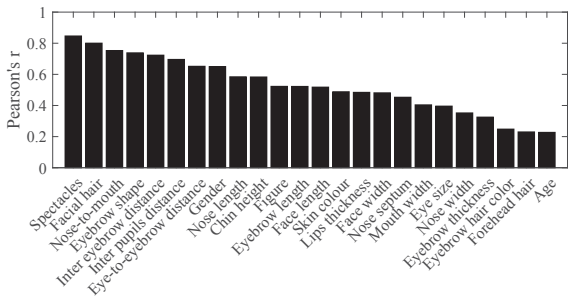


Figure 4: Trait stability based on Pearson's r.

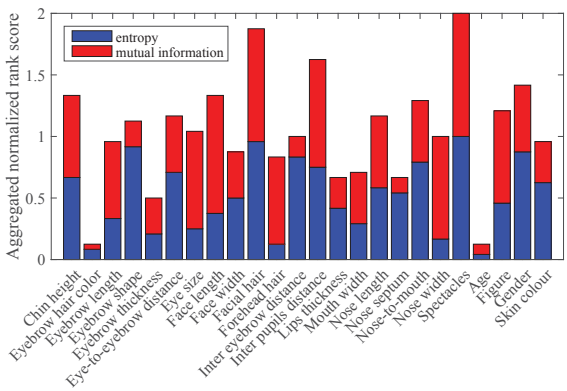


Figure 5: Discriminative power of the soft biometrics.

Correlation Analysis investigates any associations between traits, which aids in assessing their statistical independence. Figure 3 shows all the correlations between every possible pair of traits (see Table 1); computed using Pearson's correlation coefficient, r. The correlation matrix in Figure 3 shows that there is very minor correlation between the traits, and this indicates the independence of each individual trait and the unique informative value that it can add to the identification.

The **Trait Stability** reflects consistency among different samples (i.e. multiple comparisons for the same subject) that are produced by different annotators, and it is an indicator for the robustness of a feature. In order to measure trait stability, two galleries, which are composed of the Elo rates for traits of all subjects in the dataset, were constructed based on two different sets of randomly selected subjects comparisons, and Pearson's correlation coefficient, r, was calculated for each individual trait based on the two samples. Thus, the higher is the correlation coefficient, the more stable is the trait. The resulted Pearson's r values are shown in Figure 4, while the resulted $p-$values for all the traits correlations were all less than the significance level $(p \leq 0.05)$, which implies that the correlations are statistically significance. As can be seen in Figure 4, *spectacles* and *facial hair*, which are binary-like soft biometrics, have the highest stability among the other traits. Followed by, *nose-to-mouth distance* and a group of eye and eyebrow regions traits. On the other hand, *eyebrow hair colour*, *forehead hair*, and *age* have the lowest stability.

The **Trait Discriminative Power** measures the distinctiveness of traits, and hence has implications with respect to identification performance. To assess the discriminative power of the traits, two methods were used: entropy, which is the amount of information contained in the trait; and mutual information [18], which is a measure of information carried by each trait about the subjects' labels. Both methods were applied with the relative rate data of each trait. Figure 5 shows the normalized entropy and mutual information when used with each traits. We can see that both methods rated the binary-like features (e.g. *spectacles* and *facial hair*) with a high discriminative power beside *chin height* and *inter pupil distance*. On the other hand, *age* and *eyebrow hair colour* are found with the least discriminative power among the traits based on both methods. Interestingly, some traits like *eyebrow length*, *eye size*, *forehead hair*, *inter eyebrow distance*, and *nose width* have significant difference in their discriminative power rates with the two methods. In light of the analysis presented in this section, we can conclude that the binary-like traits such as *spectacles* and *facial hair* are distinctive in their stability and discriminative power. On contrast, *eyebrow hair colour*, and *age* are apparently low in their stability and discriminative power, the low significance of *age* can be referred to the infeasibility of humans in age estimation for large scale datasets [17]. The impact of the features with the lowest discriminative power on identification is further investigated in Section 4.

4. Experiments

The purpose of the identification experiments in the context of this study is to assess the reliability of the comparative facial soft biometrics for human identification and verification under realistic conditions, in addition to measuring the impact of different features on the identification performance. This section describes the experimental design used for this study, presents the results of identification and verification experiments, then discusses the results as well as their consequences.

4.1. Identification via Comparative Soft Biometrics

The identification performance evaluation in this study is based on 6-fold cross validation technique in which the dataset (4038 subjects) is partitioned into 6 equal subsets each of which is used as a test set while the remaining 5 folds are used as for training (the gallery). For each test subject, s, in a given test fold, C comparisons between s and other randomly selected subjects are removed from the total comparisons in the dataset and used to generate a probe biometric signature for each test subject using the Elo rating system. The comparisons remaining after excluding those

selected for the test fold subjects are used to construct the gallery in which each subject is described by a biometric signature produced using the Elo rating system. Then, Pearson correlation coefficient is used to measure the distance, d_P, between the biometric signatures of the probe and that of each of the subjects in the gallery as follows:

$$d_P = 1 - \frac{\sum_{i=1}^{T}(X(i)-\overline{X})(Y(i)-\overline{Y})}{\sqrt{\sum_{i=1}^{T}(X(i)-\overline{X})^2}\sqrt{\sum_{i=1}^{T}(Y(i)-\overline{Y})^2}} \quad (5)$$

where X is a vector that represents the biometric signature of the probe, Y is a vector that represents the biometric signature of the subject in the gallery that is being compared against the probe, and T is the number of soft biometric traits composing the biometric signatures.

The subject from the gallery that has the minimum Pearson's distance with the probe (i.e. the nearest neighbor) is considered as the rank-1 match. This experimental procedure was applied in a round robin manner over the 6 folds to cover the complete dataset, and the arithmetic mean of the identification performance rate at each rank over the 6 fold is considered as the experiment outcome. The 6-fold identification experiment was repeated until the harmonic mean of the identification rates over all trials converged. Figure 6 shows the identification performance (based on harmonic mean) achieved using 10, 15 and 20 subject comparisons in identification, and Figure 7 presents query retrieval examples using 20 comparisons.

By using the 24 traits with 10 comparisons, which is the average size of an ideal identity parade [6], a match will be found in the top ten results with probability of 97%. Furthermore, a match is always guaranteed in the top 1.83% results (i.e. rank-74 of the 4038 total subjects). Additionally, the results show the effect of increasing the number of comparisons in significantly improving the identification performance by a percentage increase of 47% at rank-1 when increasing comparisons from 10 to 15, and a percentage increase of 22% at rank-1 reaching an identification rate of 72% when increasing the comparisons from 15 to 20. To the best of our knowledge, the only published work that studied human face identification using human annotations for both probe and gallery with a large dataset is the work of Klare et al [3]. By comparing the identification performance obtained from our experiments with the results of [3], which achieved a rank-1 accuracy of 22.5% using 46 features (19 binary and 27 categorical) with 1196 subjects from the FERET database, there is an evident advantage by using comparative soft biometrics as compared to categorical soft biometrics. Thus, the comparative labels have resulted in a better identification performance (30.2% at rank-1) for a larger and more challenging database, using fewer attributes (24 only) and 10 subject comparisons only. We attribute the advantage to the use of comparative labels.

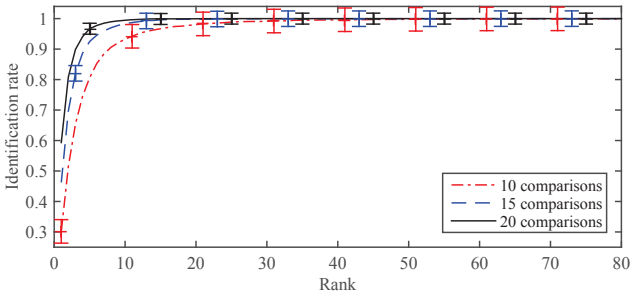


Figure 6: Identification performance for $C \in \{10, 15, 20\}$ comparisons per subject.

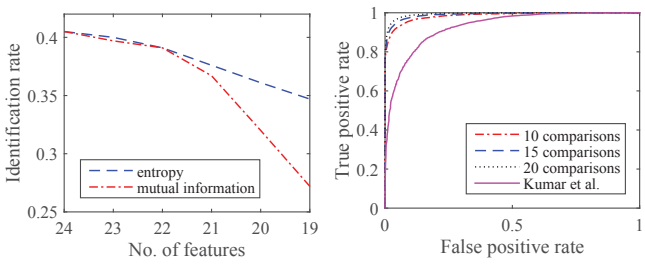
Figure 7: Example for query retrieval, left: a subject successfully retrieved at rank 1, right: a subject retrieved at rank 6.

4.2. Trait Impact on Identification

The discriminative power analysis presented in Section 3 has shown that some traits have a relatively low discriminative power (e.g. *age*, and *eyebrow hair colour*). To discover the impact of such traits on the identification performance, the 24 features were sorted in descending order based on their discriminative power. Then, five feature subsets were formed by selecting the top k features each time, allowing k to vary between 19 and 24. This procedure was applied with the two discriminative power ranks, which were generated based on entropy and mutual information. As shown in Figure 8(a), rank-1 identification performance decreases each time a feature is excluded from the identification process. Also, a steeper decrease in rank-1 identification performance is observed when excluding those features that are rated as the lowest in discriminative power based on mutual information. The most important finding to emerge from this experiment is that all the features of our soft biometric set are contributing in the identification performance, even those features with the lowest discriminative power. Also, the result shows that entropy is more efficient for feature discriminative power assessment than mutual information.

4.3. Verification via Comparative Soft Biometrics

The objective of the verification experiment is to evaluate the effectiveness of our comparative soft biometrics for human verification. Since each subject in our dataset has been compared against many other subjects (see Table 2), multiple samples (biometric signatures) can be derived for each subject by using different set of comparisons. To assess the verification performance of our comparative soft



(a) (b)

Figure 8: (a) Rank-1 identification rate when excluding the features with low discriminative power. (b) Verification performance of comparative facial soft biometrics with $C \in \{10, 15, 20\}$ comparisons per subject.

biometrics, two samples were derived for each of the 4038 subjects in this study by using: 10, 15, and 20 comparisons with Pearson's correlation as a distance measure. The Receiver Operator Characteristic (ROC) is used to summarize verification performance in terms of true positive and false positive rates as shown in Figure 8(b). The resulted ROC accuracies are: 93.7%, 95.2%, and 96.2%; using 10, 15, and 20 comparisons respectively, while Kumar et al. have achieved an accuracy of 85.25% using trained classifiers for 73 attributes in [10]. Here, it is important to re-state that the performance evaluation in [10] was performed with the View 2 subset of LFW and the model selection was performed using the View 1 subset. Whereas we have performed our experiments with 4038 subjects from the the View 1 subset, as we are addressing a verification scenario for previously seen subjects based on semantic features, and our approach is aimed to assess the reliability of comparative facial soft biometrics for unconstrained human identification. Overall, as verification is performed with LFW using 24 traits, these results emphasize the effectiveness of our traits for verification. These results have important implications for verification scenarios based on comparative soft biometrics.

5. Conclusions

Our main goal has been to study human identification via comparative facial soft biometrics in large, unconstrained datasets. We have introduced a new set of facial soft biometrics with comparative labels of a reduced scale and analyzed the statistical significance and the discriminative power of each trait. The study confirmed the reliability of comparative facial soft biometrics and highlighted their effectiveness for identification as compared to the categorical soft biometrics considering the challenging visual conditions of the used dataset and the small resolution of the feature vectors used in identification and verification (24 features and comparative labels of 3 levels only). In addition, analyzing the impact of our soft biometrics has shown that every trait is significantly contributing with respect to identification performance irrespective of discriminative power. Also, the comparative facial soft biometrics have shown a potency in human verification. These findings have important implications for the understanding of how to further explore human identification using comparative soft biometrics. Finally, this study can serve as a base for future studies in unconstrained human identification using other comparative soft biometrics such as those based on body and clothing.

References

[1] Daniel Reid, Sina Samangooei, Cunjian Chen, Mark Nixon, and Arun Ross. Soft biometrics for surveillance: an overview. *Machine learning: theory and applications. Elsevier*, pages 327–352, 2013.

[2] Pedro Tome, Julian Fierrez, Ruben Vera-Rodriguez, and Mark S Nixon. Soft biometrics and their application in person recognition at a distance. *Information Forensics and Security, IEEE Transactions on*, 9(3):464–475, 2014.

[3] Brendan F Klare, Scott Klum, Joshua C Klontz, Emma Taborsky, Tayfun Akgul, and Anil K Jain. Suspect identification based on descriptive facial attributes. In *IEEE IJCB*, 2014.

[4] Nawaf Almudhahka, Mark Nixon, and Jonathon Hare. Human face identification via comparative soft biometrics. In *IEEE ISBA*, 2016.

[5] Daniel A Reid and Mark S Nixon. Human identification using facial comparative descriptions. In *IEEE ICB*, 2013.

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

[7] Emad Sami Jaha and Mark S Nixon. Soft biometrics for subject identification using clothing attributes. In *IEEE IJCB*, 2014.

[8] Richard D Seely, Sina Samangooei, Minhung Lee, John N Carter, and Mark S Nixon. The university of southampton multi-biometric tunnel and introducing a novel 3d gait dataset. In *IEEE BTAS*, 2008.

[9] 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.

[10] Neeraj Kumar, Alexander C Berg, Peter N Belhumeur, and Shree K Nayar. Describable visual attributes for face verification and image search. *IEEE TPAMI*, 33(10):1962–1977, 2011.

[11] 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.

[12] 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.

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

[14] Niv Zehngut, Felix Juefei-Xu, Rishabh Bardia, Dipan K Pal, Chandrasekhar Bhagavatula, and Marios Savvides. Investigating the feasibility of image-based nose biometrics. In *IEEE ICIP*, 2015.

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

[16] Gary Huang, Marwan Mattar, Honglak Lee, and Erik G Learned-Miller. Learning to align from scratch. In *NIPS*, 2012.

[17] Hu Han and Anil K Jain. Age, gender and race estimation from unconstrained face images. *Dept. Comput. Sci. Eng., Michigan State Univ., East Lansing, MI, USA, MSU Tech. Rep.(MSU-CSE-14-5)*, 2014.

[18] Baofeng Guo and Mark S Nixon. Gait feature subset selection by mutual information. *IEEE SMC*, 39(1):36–46, 2009.