Using Comparative Human Descriptions for Soft Biometrics

Daniel A. Reid and Mark S. Nixon
School of Electronics and Computer Science
University of Southampton, Southampton SO17 1BJ, UK
{dar1g09|msn}@ecs.soton.ac.uk

Abstract

Soft biometrics is a new form of biometric identification which utilizes labeled physical or behavioral traits. Although these traits intuitively have less discriminatory capability than mensurate approaches, they offer several advantages over traditional biometric techniques. Soft biometric traits can be typically described as labels and measurements which can be understood by people, allowing retrieval and recognition based solely on human descriptions. Although being a key component of eyewitness evidence, conventional human descriptions can be considered to be unreliable. A novel method of obtaining human descriptions will be introduced which utilizes visual comparisons between subjects. The Elo rating system is used to infer relative measurements of subjects' traits based on the comparative human descriptions. This innovative approach to obtaining human descriptions has been shown to counter many problems associated with categorical (absolute) labels. The resulting soft biometric signatures have been demonstrated to be robust and allow accurate retrieval of subjects in video data and show that elapsed time can have little effect on comparative descriptions.

1. Introduction

Traditional biometric techniques identify people using distinct physical or behavioral features. These features are very discriminative although can rarely be described using labels which can be understood by people. This restricts identification to situations where the subject's biometric signature can be obtained and only permits identification of those subjects whose biometric signature has previously been recorded. Soft biometrics concerns labels which people use to describe each other. Although each trait/ label can have reduced discriminative capability, they can be combined for identification [8, 1] and fusion with traditional 'hard' biometrics[4, 7]. Dantcheva et al. [2] likens this to obtaining a single ridge of a fingerprint or a small section of the iris, these would not be unique enough to identify a subject but by gathering many small features we are able to build a unique signature.

One of the main advantages of soft biometrics are their relationship with human description; humans naturally use soft biometric traits to identify and describe each other. Beyond identification, soft biometrics also allow retrieval. This is achieved by bridging the semantic gap between biometric measurements and human descriptions.

Figure 1. Surveillance frame displaying common surveillance problems[1]

Though face and gait are the only possible biometrics at a distance, in surveillance scenarios they can suffer from low frame rate and/or resolution. Figure 1 shows an example of a typical CCTV video frame. This frame shows suspects of the murder of a Hamas commander in Dubai in 2010. It can be observed that although the picture is at low resolution and the subjects' physical features are occluded, a detailed human description of the subjects can still be determined especially when viewing the video from which this frame was derived. Soft biometric traits can be obtained from the data derived from low quality sensors, including surveillance cameras. They also require less computation compared to hard biometrics, no cooperation from the subject and are non-invasive - making them ideal in surveillance applications.

[1] Arabian Business http://www.arabianbusiness.com/interpol-issues-notice-for-hamas-murder-suspects-40450.html

To allow identification from human descriptions, physical properties must be accurately described. Conventional human descriptions represent an important element of eyewitness evidence, although can be considered inaccurate and unreliable [5, 6]. Previously categorical labels have been used to describe soft biometric traits [8]. Humans can be inaccurate when predicting measurements [10] and labels were seen as a more robust method of obtaining human descriptions. The major problem associated with absolute categorical labels is their subjective nature. A label's meaning is based on the person's own attributes and their perception of population averages and variation. This can vary, making subjective labels less reliable. Categorical labels naturally lack detail, resulting in biometric signatures which have lower discriminatory capability. This paper introduces a new method for obtaining human descriptions which are more robust with improved discriminatory capability when compared with the use of absolute labels.

Labeling, or estimating attributes, is required to convey visual information verbally. Visual information, can be very difficult to convert to dimensions or labels - resulting in inaccuracy or ambiguity. Comparing the appearance of two subjects is a natural method of comparing bodily attributes, bypassing the need to label the visual information. Intuitively it is very easy to say whether a person is taller than someone else, but labeling or estimating the height can be difficult. On the other hand, relative measurements of a suspect's traits can be accurately inferred by visual comparison. We use this notion to solve problems associated with absolute labels and measurements, to provide reliable and robust descriptions.

The rest of this paper will explore the effectiveness of human comparisons and how they can be applied to soft biometric representation and retrieval. Section 2 will introduce a dataset of human comparisons used throughout this study. A modified Elo rating system, used to calculate relative measurements, is presented within section 3. Section 4 will demonstrate the discriminative capabilities of the relative measurements derived from the comparative analysis, detailing results for retrieval from video data.

2. Human Comparison Dataset

Multiple comparisons are required to infer accurate relative measurements of a suspect's physical traits. To obtain multiple comparisons in application environments, the observed suspect can be compared to videos of multiple subjects obtained from a database. After a series of comparisons the relative measurements of the suspect's attributes can be inferred.

The experiment used to gather data for the human comparison dataset was designed to mimic an application environment, requiring a user to compare a single suspect with five different subjects. This will give an insight into the proposed method of obtaining comparisons and whether it is suitable for real world applications.

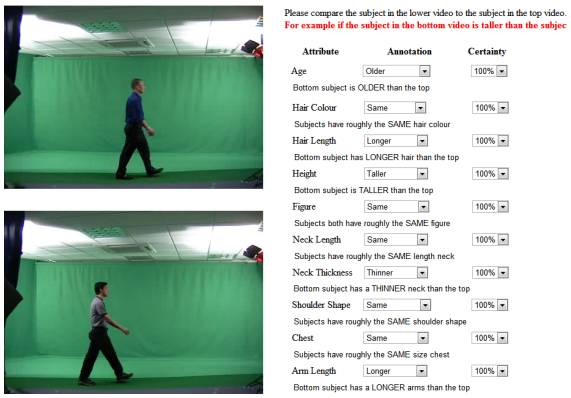

Figure 2. The website developed to collect comparisons

Comparisons were made between fronto-parallel videos of 100 people from the Soton gait database [9]. These people were assigned at random as one of either 20 'suspects' or 80 'subjects' and comparisons were derived by 57 human 'annotators'. At first, the annotator viewed both the suspect and the subject simultaneously. Later, the annotator viewed a limited exposure to a video of the the suspect before comparing the videos of five subjects, simulating application environments. Figure 2 shows the website used to gather the comparisons.

Annotations of 19 soft traits (table 1) were obtained for each human comparison. It can be observed that three traits were annotated using categorical labels. These three traits are unsuited to comparative annotations, either due to the inherently categorical nature of the trait or the lack of a suitable comparison criteria.

The resulting data included 558 suspect-subject comparisons. To maximize coverage of the comparison data, subject-subject comparisons were inferred when two subjects were compared to the same suspect.

The comparisons appeared to remain accurate after a limited exposure to the suspect, showing promise for eyewitness applications where memory is a real concern. More information and analysis of the comparisons, experimental procedure and its potential for use in crime applications will be presented in a future publication.

3. Relative Measurements

Comparative annotations must be anchored to convey meaningful subject invariant information. The resulting value is a relative measurement, providing a measurement of the specific trait in relation to the rest of the population. This can be used as a biometric feature allowing retrieval and recognition based on a subject's relative trait measurements.

Table 1. Soft traits used to compare subjects

Trait	Description Type	Labels
Arm Length	Comparative	[Much Shorter, Shorter, Same, Longer, Much Longer]
Arm Thickness	Comparative	[Much Thinner, Thinner, Same, Thicker, Much Thicker]
Chest	Comparative	[Much Smaller, Smaller, Same, Bigger, Much Bigger]
Figure	Comparative	[Much Smaller, Smaller, Same, Larger, Much Larger]
Height	Comparative	[Much Shorter, Shorter, Same, Taller, Much Taller]
Hips	Comparative	[Much Narrower, Narrower, Same, Broader, Much Broader]
Leg Length	Comparative	[Much Shorter, Shorter, Same, Longer, Much Longer]
Leg Thickness	Comparative	[Much Thinner, Thinner, Same, Thicker, Much Thicker]
Muscle Build	Comparative	[Much Leaner, Leaner, Same, More Muscular, Much More Muscular]
Shoulder Shape	Comparative	[More Square, Same, More Rounded]
Weight	Comparative	[Much Thinner, Thinner, Same, Fatter, Much Fatter]
Age	Comparative	[Much Younger, Younger, Same, Older, Much Older]
Ethnicity	Absolute	[European, Middle Eastern, Far Eastern, Black, Mixed, Other]
Gender	Absolute	[Female, Male]
Skin Colour	Absolute	[White, Tanned, Oriental, Black]
Hair Colour	Comparative	[Much Lighter, Lighter, Same, Darker, Much Darker]
Hair Length	Comparative	[Much Shorter, Shorter, Same, Longer, Much Longer]
Neck Length	Comparative	[Much Shorter, Shorter, Same, Longer, Much Longer]
Neck Thickness	Comparative	[Much Thinner, Thinner, Same, Thicker, Much Thicker]

3.1. Elo rating system

To produce relative measurements the comparisons between subjects must be analyzed to identify an ordering within the population in respect to an individual trait. This was achieved using an Elo rating system [3]. In essence the Elo rating system provides a method of inferring a relative measurement from comparisons. Elo ratings were designed to quantify the skill of chess players. The performance of a chess player cannot be measured absolutely, instead the player's (relative) skill level is inferred from matches against other players. This rating system solves a problem very similar to comparative annotations. In soft biometrics the absolute measurements of the traits cannot be directly observed due to the inaccuracy of human descriptions. Instead we can compare the traits to infer relative measurements, similar to how chess games compare two players' skill.

In the Elo rating system each player starts with a default skill rating, this is adjusted based on the result of any games played. The amount of adjustment is based upon the skill level of the opponent and the result of the match. Each game includes two players, each having a rating representing their inferred skill ratings. Based on these ratings the expected result of the match is determined, where E_a is the expected score for player A and E_b is the expected score for player B.

$$E_a = \frac{Q_A}{Q_A + Q_B} \tag{1}$$

$$E_b = \frac{Q_B}{Q_A + Q_B} \tag{2}$$

$$Q_A = 10^{R_A/U} \tag{3}$$

$$Q_B = 10^{R_B/U} \tag{4}$$

Where R_A and R_B are the current skill ratings of player A and B respectively and U is a constant determining how the current ratings affect the expected result. It can be observed that if a player has U rating advantage, the chance of winning is magnified ten times. These equations predict the expected outcome of a match based on the players' current inferred skill rating. Once the game has been completed the ratings of the players are updated using the following equation:

$$R_A^{'} = R_A + K(S_A - E_A) \tag{5}$$

Where S_A is the result of the match, generally set to 1 for a win, 0 for a loss and 0.5 for a draw. K is a constant which defines the maximum rating adjustment resulting from the match. If the expected result does not reflect the actual result, it is assumed the skill ratings of the players are incorrect. The skill ratings are adjusted based upon the extent of the error between the expected and the actual result.

In chess the unknown measurement is the skill of the chess player - in the case of comparative annotations the unknown is the relative measurement of the attribute being compared. Comparisons between subjects provide a measure of difference between the subjects' attributes, similar to how chess games compare the skill level of the players. This information is used to adjust the inferred relative measurements of the two subjects. A scoring system, similar to the win-draw-loss system used in chess, is required to compare the expected result to the actual result. Soft biometric traits are compared using five ordered labels, these are assigned a number ranging from 1 to 5 based on their order. The 'score' resulting from a comparison is obtained by normalizing the given label's value to within 0 and 1. If the actual result reflects the expected result the relative measurements are not adjusted. If the actual result disagrees with the expected result, the subjects' relative measurements are adjusted. The size of this adjustment is dependent on the error between the expected and actual results.

The main advantage of this system is that it does not re-

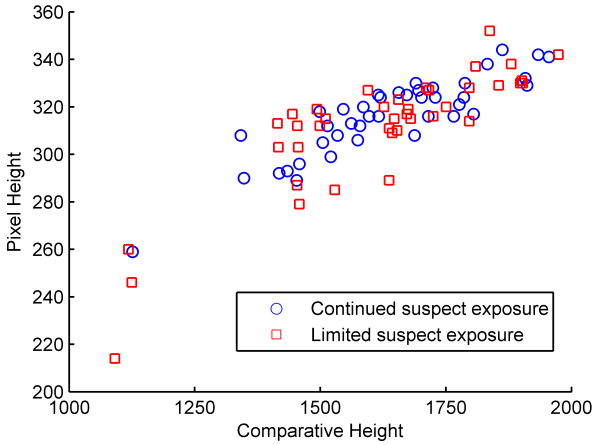

Figure 3. The relationship between pixel height and relative height

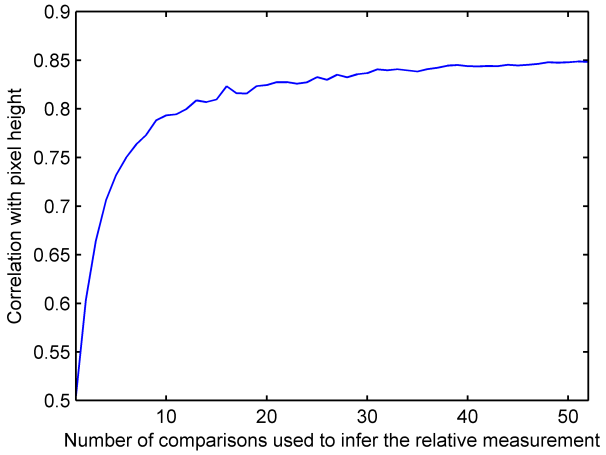

Figure 4. Correlation between pixel height and relative height with varying amounts of comparisons per subject.

quire exhaustive comparisons between all the subjects to calculate an accurate relative measurement. Instead it adjusts the suspect's relative measurement based on any available comparisons, taking into account the relative measurements of the subjects which a suspect has been compared against.

3.2. Accuracy of relative measurements

The relative measurement details how the subject's trait compares to other subjects within the population. This naturally reveals population averages and trait distributions without enforcing strict labels. If the comparisons were accurate and the method of anchoring the comparisons was successful, the final relative measurements should represent the real world measurements of the traits. Determining the pixel height of a subject from the video data allowed us to explore the correlation between an actual trait's measurement and the inferred relative measurement. Figure 3 shows the relationship between the relative and actual height measurements. The correlation between pixel height and relative height was 0.87 - showing that the relative measurements strongly represent the physical traits.

The relative measurements shown in figure 3 were inferred from all the comparisons in the human comparison database. In application settings we would seek to compare against the minimum amount of subjects to achieve an accurate relative measurement. Figure 4 shows the correlation between relative height and pixel height for varying amounts of comparisons per subject. It can be seen that the correlation increases throughout the range presented (1-52 comparisons), clearly demonstrating that additional comparisons improve the accuracy of the resulting relative measurement. The correlation was within 10% of its terminal value after 9 comparisons, which is suitable in an application scenario.

Figure 3 includes results from the second part of the experiment, which simulated a limited exposure to the suspect, these results exhibit a weaker correlation with the pixel height. This implies that there were some errors within the comparisons. Although the correlation was weaker, the resulting relative measurements still represent the actual pixel height of the subjects. This shows great promise for the accuracy of comparisons after a limited exposure to the suspect. Further studies into this topic are critical to assess the suitability of comparative human descriptions for eyewitness applications.

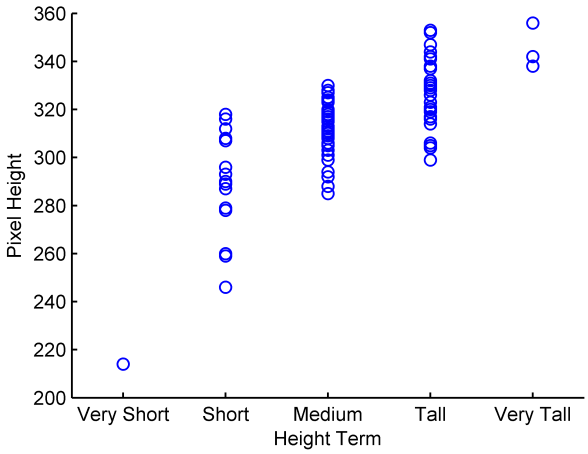

Figure 5. The relationship between pixel height and absolute labels

Figure 5 shows the relationship between pixel height and the absolute height labels used previously. Huge confusion exists between the short, medium and tall labels, this is caused by the undefined and therefore subjective nature of the semantic labels. The correlation (0.71) is much weaker mainly due to this ambiguity but also the categorical nature of the labels. Figure 3 highlights the continuous nature of the relative measurements, providing much more infor-

mation about the subjects' traits. Comparative categorical annotations can capture more accurate and descriptive information whilst avoiding asking the user for continuous estimations.

4. Retrieval

Biometric retrieval aims to identify an unknown subject by comparing their biometric signature to a database of biometric signatures. Currently the police collect labeled descriptions of suspects. These descriptions are stored and can be searched to retrieve subjects. Unfortunately the labels used to describe the subjects lack distinctiveness and are subjective. Relative measurements could be used to provide robust human descriptions allowing accurate retrieval. The following section will explore the distinctiveness and robustness of relative measurements compared to absolute categorical labels.

4.1. Retrieval using labeled descriptions

Previously, labels were used to describe traits [8]. Due to their categorical nature, the differences between subjects were often small. Subject interference [2] is a known problem when using labels and occurs when two subjects are indistinguishable from each other. When analyzing larger databases the probability of interference increases, especially if the traits' distributions are small (seen in fig. 5).

Labeled descriptions were obtained from the Soton gait database [9]. 125 subjects were labeled by multiple users (average of 10 separate user annotations per subject) describing 23 traits [8]. A leave-one-out validation approach was used to evaluate performance. Each user description was used to retrieve the corresponding subject from the 125 subject dataset. Figure 6 shows the results. Rank 1 performance was found to be 48%. This result highlights how the subjective nature of the labels and the lack of information affects the retrieval performance.

4.2. Retrieval using relative measurements

The relative measurements introduced in this paper are continuous. This practically removes the problem of subject interference and increases the probability of differences between subjects. For this reason the biometric signatures should be more distinct, allowing accurate retrieval.

The retrieval experiment aims to retrieve a subject from an 80 subject database which was introduced in section 2. Retrieval will be performed using varying amounts of test comparisons, n. This investigates how many comparisons are required to accurately retrieve a subject. n comparisons will be randomly sampled for each subject. The n comparisons will be used to generate the relative measurement biometric signature which will be used to query the database, known as the probe. The subject's remaining comparisons will be used to construct the gallery. Random sampling will

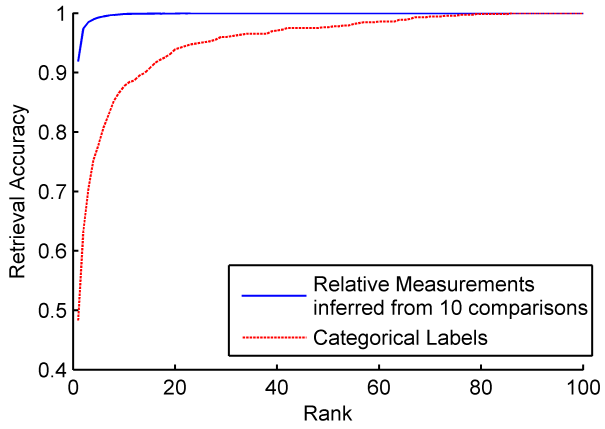

Figure 6. Retrieval Accuracy

be repeated until the retrieval accuracy remains constant for 10 random samples.

The biometric signatures within the database will consist of all of the 19 traits' relative measurements (see table 1). The Euclidean distance between two relative measurement signatures will be used to indicate their similarity. The retrieval results shown in this paper are obtained from exhaustively calculating the similarity between the probe and each gallery signature. The rank 1 retrieval accuracy over varying number of probe comparisons is shown in figure 7. The rank 1 performance using just one comparison to construct the probe is 47%. Obviously one comparison only tells us how the subject differs from another subject, the resulting relative measurements are very inaccurate. Interestingly this result matches the rank 1 retrieval accuracy when using categorical labels. As more comparisons are received the accuracy of the relative measurements increase, leading to improved retrieval results. With 10 comparisons a 92% rank 1 retrieval rate is achieved. This demonstrates that accurate relative measurements are very distinct. The retrieval accuracy continues increasing over the range shown, achieving a 95% retrieval accuracy with 20 comparisons.

Figure 8 shows an unsuccessful retrieval query where the two subjects were confused with each other. It can be observed that the subjects look very similar - both having a very similar build, hair length and skin color. The relative measurements of the subjects' traits reflect these similarities resulting in the confusion between the two. In comparison, figure 9 shows a subject who was retrieved successfully even with only one comparison. The male subject has long hair, which is not common within the Soton gait dataset, and is also particularly tall. This uncommon set of traits results in a distinct set of relative measurements making retrieval very successful.

It has been shown that the new relative measurements equal the retrieval capabilities of categorical labels with

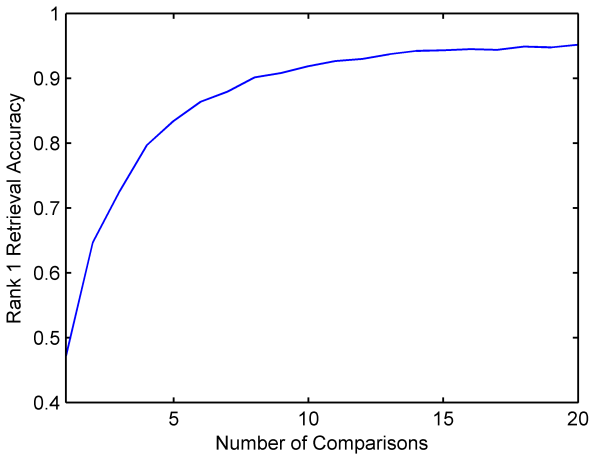

Figure 7. Rank 1 retrieval accuracy using relative measurements obtained from different amounts of comparisons

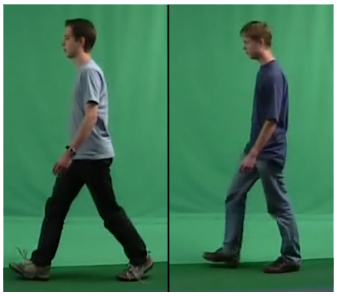

Figure 8. Incorrect retrieval with 10 comparisons. Left: Database probe. Right: Retrieved subject

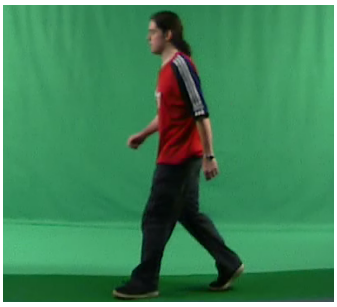

Figure 9. Subject achieved correct retrieval with only one comparison

only one comparison. Retrieval can be greatly improved by obtaining more comparisons. Subject interference, which limits the effectiveness of categorical labels, has been demonstrated not to affect relative measurements using an 80 subject database.

5. Discussion and Conclusions

Soft biometrics exploit labeled physical or behavioral traits to allow human identification. Humans naturally use these traits to identify each other, permitting a soft biometric signature to be determined based solely on a human description. Comparative annotations have been introduced as a new approach for gathering human descriptions. They offer several advantages over absolute labels. Most critically, comparisons do not use subjective labels, resulting in robust annotations which are constant between different people.

Comparisons between a suspect and videos of multiple subjects would be used to infer relative measurements using an Elo rating system. Relative measurements of 19 traits are combined to create a biometric signature describing the suspect. The accuracy of these relative measurements depends on the amount of comparisons received. Results comparing actual height to the inferred relative height showed a correlation of 0.87 was achieved. This strong correlation was achieved by avoiding subjective labels and inferring an informative continuous measurement.

Classic biometric retrieval was used to explore the distinctiveness and robustness of the relative measurements. Results showed that accurate retrieval was possible, allowing a 92% rank 1 retrieval performance with only 10 comparisons. This outperformed labeled descriptions, which achieved a rank 1 retrieval rate of 48%. Relative measurements have been shown to contain more discriminative information and do not suffer from subject interference, where two subjects are indistinguishable from each other.

References

[1] H. Ailisto, M. Lindholm, S. M. Makela, and E. Vildjiounaite. Unobtrusive user identification with light biometrics. In *Proc. NordiCHI*, pages 327–330, 2004. 1

[2] A. Dantcheva, J. Dugelay, and P. Elia. Soft biometrics systems: Reliability and asymptotic bounds. In *BTAS*, pages 1–6, Sept. 2010. 1, 5

[3] A. E. Elo. *The rating of chessplayers, past and present*. Batsford, 1978. 3

[4] A. K. Jain, K. Nandakumar, X. Lu, and U. Park. Integrating faces, fingerprints, and soft biometric traits for user recognition. In *BioAW*, volume LNCS 3087, pages 259–269, 2004. 1

[5] E. F. Loftus. *Eyewitness testimony*. Harvard U. Pr., 1996. 2

[6] C. A. Meissner, S. L. Sporer, and J. W. Schooler. Person descriptions as eyewitness evidence. *Handbook of eyewitness psychology*, 2:3–34, 2007. 2

[7] U. Park and A. K. Jain. Face Matching and Retrieval Using Soft Biometrics. *IEEE Trans on Information Forensics and Security*, 5(3):406–415, Sept. 2010. 1

[8] S. Samangooei and M. S. Nixon. Performing Content-based Retrieval of Humans using Gait Biometrics. *Multimedia Tools and Applications*, 49(1):195–212, 2010. 1, 2, 5

[9] J. Shutler, M. Grant, M. S. Nixon, and J. N. Carter. On a large sequence-based human gait database. In *Proc RASC*, pages 66–72. Springer Verlag, 2002. 2, 5

[10] J. C. Yuille and J. L. Cutshall. A case study of eyewitness memory of a crime. *Journal of Applied Psychology*, 71(2):291–301, 1986. 2