Identifying Humans using Comparative Descriptions

Daniel A. Reid*, Mark S. Nixon* and Sarah V. Stevenage+

* School of Electronics and Computer Science | + School of Psychology
University of Southampton, Southampton SO17 1BJ, UK

Abstract

Soft biometrics is a new form of biometric identification which utilizes human descriptions of a subject's physical appearance. Although these descriptions intuitively have less discriminatory capability than traditional biometric approaches, they are able to retrieve and recognize subjects based solely on a human description. To permit soft biometric identification the human description must be accurate, yet conventional human descriptions comprising of absolute labels and estimations are often unreliable. In this paper we introduce a novel method of human description which utilizes comparative descriptors derived from visual comparisons between subjects. This innovative approach to obtaining human descriptions has been shown to counter many problems associated with absolute categorical labels. Comparative categorical labels are objective and can be used to infer descriptive continuous relative measurements. The resulting biometric signatures have been demonstrated to differ significantly from absolute descriptions allowing improved retrieval of subjects and could even be used to increase the accuracy of witness description in crime analysis.

1 Introduction

Soft biometrics exploit physical or behavioral features which can be described by humans. Although each attribute can have reduced discriminative capability, they can be combined for identification [1, 2] and fusion with traditional 'hard' biometrics [3,4]. One of the main advantages of soft biometrics is their relationship with human description; humans naturally use soft biometric traits to identify and describe one an other. This permits identification and retrieval based solely on a human description of the subject, possibly obtained from an eyewitness. This contrasts with traditional biometric techniques which restrict 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.

Biometric techniques which allow identification from a distance, like face and gait recognition, can suffer in surveillance applications from low frame rates and/or resolution. Figure 1 shows suspects of the murder of a Hamas commander in Dubai in 2010. The frame is at a low resolution and the subjects' physical features cannot wholly be seen. However, a detailed human description of the suspects 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.

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

Human descriptions must be accurate and reliable to allow identification but are often considered unreliable [5]. Conventional human descriptions consist of categorical labels (i.e. 'tall') or continuous estimations of human characteristics (i.e. 6'2"). Categorical labels are subjective and naturally lack detail whilst estimates of trait attributes can be inaccurate. This paper will introduce a new form of description which exploits visual comparisons between subjects. Objective comparative labels are used to compare a single suspect to multiple subjects. In application settings an eyewitness would be asked to compare the observed target to multiple subjects obtained from a database. A set of descriptive continuous relative measurements describing the target can be inferred from multiple comparisons. We show that relative measurements can be used to identify subjects with a 95% accuracy.

The paper is structured as follows. Section 2 will explore the current methods of human description and the advantages of human comparisons. The process of building a human comparison dataset is detailed in section 3. Analysis of the human comparison dataset is presented in section 4. Section 5 will explore the discriminatory capability of relative measurements which are inferred from multiple human comparisons.

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

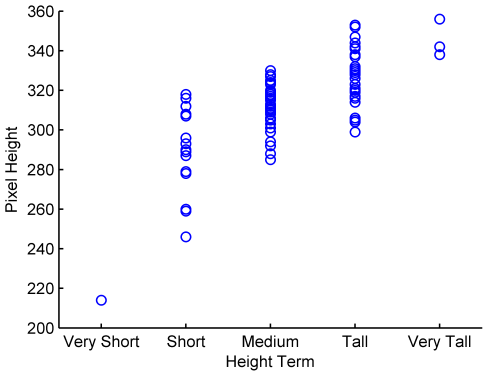

Figure 2. The relationship between pixel height and absolute labels

2 Human Descriptions

To allow identification from human descriptions, the physical properties described must be accurate, salient and reliable. However, absolute categorical labels are inherently subjective; a label's meaning is based on the annotator's own attributes and their perception of population averages and variation. This can vary, making subjective labels unreliable.

A study exploring the content of human descriptions identified the color and style of hair and clothing as the most frequent use of categorical labels [6]. Clothing and hair although mentioned frequently can be inaccurate. [7] notes errors between 20% and 44% when describing the color and style of hair and clothing. These inaccuracies were linked to the subjective nature of the characteristics.

Samangooei and Nixon [1] developed a soft biometric system which used 23 absolute categorical labels to describe a subject's physical appearance. The 23 traits were chosen to be universal, distinct, easily discernible at a distance and largely permanent. The selected traits featured both naturally categorical attributes, like hair color, and characteristics generally associated with value metrics, like height - both were described using categorical labels. The descriptions of value metric traits were found to be unreliable. Figure 2 shows the relationship between pixel height and the median absolute height label used to describe the subjects. Overlaps exist between the short, medium and tall labels, this is caused by the undefined and therefore subjective nature of the labels.

Traits which are explained using continuous measurements, like weight and height, are generally estimated by the annotator. Although continuous measurements are very descriptive, humans can be poor at estimating the trait's attributes. [7] showed that estimates of height, weight and age were incorrect 50% of the time based on 95 cases. Inaccurate estimates have been accredited to an own anchor effect, where the witness's own characteristics were used as a reference to judge the suspect [8]. It was also found that descriptions tended towards the witness's perception of the population average - estimating shorter people as taller and vice versa. This was thought to occur due to the witnesses shying away from extreme judgments. This finding was confirmed by [5].

This paper will study the advantages of comparative descriptors as opposed to absolute labels. Comparative descriptors (for example 'shorter') are less subjective than absolute labels, resulting in robust descriptions. A continuous relative measurement can be inferred from multiple human comparisons. This measurement is more descriptive than absolute labels whilst avoiding asking the user to estimate a continuous attribute.

3 Human Comparison Dataset

The method used to obtain comparisons from a user is an important consideration when developing a new form of human description. The practical limitations of human memory and the ability of humans to compare bodily attributes must be considered and explored. An experiment was designed to answer the following questions:

- Are human comparisons more robust against errors originating from subjectiveness?
- Do the resulting relative measurements reflect the subject's physical attributes?
- Do relative measurements provide more discriminatory information than absolute labels?
- Is the developed method of obtaining human comparisons practical?

When applied to application environments an eyewitness would be asked to compare the observed suspect to other subjects. Ideally we would seek to compare against the minimum number of subjects to achieve accurate relative measurements of the suspect. The most informative and practical method of presenting the subjects to the eyewitness would be videos obtained from a database. After a series of comparisons the relative measurements of the suspect's attributes could be inferred and used to identify the suspect. To validate this approach the experiment will be designed to mimic the application procedure.

The first experiment explored the benefits of comparative annotations in ideal settings and the second investigated the application potential of comparisons. Initially the user was asked to compare two subjects whilst both were visible to the user. This removes all problems with memory and validates the effectiveness of comparative descriptions. Five subjects were compared to a single subject (known as the target) - this simulates the idea of comparing a suspect against a selection of subjects. A single human comparison consists of 19 traits comparisons (shown in table 1), each using one of five categorical labels. It can be observed that three traits (gender, ethnicity and skin color) were annotated using absolute 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. These absolute annotations are not considered when analyzing the comparative annotations but are utilized when identifying 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 Color	Absolute	[White, Tanned, Oriental, Black]
Hair Color	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]

Table 1. Soft traits used to compare subjects

The next part of the experiment tested the application potential of comparative annotations. Memory is a huge problem in eyewitness descriptions [9] and its effects on comparative and absolute annotations must be explored. A continuous set of videos showing a target walking, was presented to the user. These videos were the only opportunity the user had to observe the target, simulating a limited exposure. The videos continued until the user was ready to begin. The user was then asked to compare five subjects with the target. When comparing the subjects the user was prevented from viewing the target again. This examines how memory affects the comparative descriptions over time. Finally the user was asked to describe the target using absolute categorical annotations, discovering the effects of memory on absolute human descriptions. The time between viewing the target and completing the six annotations (five comparative annotations and one absolute annotation) was on average twelve minutes.

Videos of subjects from the Soton gait database [10] were used within this experiment. The gait database includes videos of 100 people walking in a plane normal to the view of the camera. Previously absolute categorical labels had been collected for the same database [1] - allowing comparisons between the two forms of description. The videos were displayed within a website shown on a computer monitor. Identifying the scale and size of the subjects from video data can be difficult, especially when attempting to compare an observed suspect to subjects within videos. This could be overcome by projecting the video to form a full size representation of the subject. Projection was not required within this experiment as each subject was filmed identically relative to background cues. The benefits of projection will be explored in future research.

The 100 subjects from the Soton gait database were assigned as one of either 20 targets or 80 subjects. Half of the subjects were used for each part of the experiment. Previously, when obtaining absolute labels, multiple annotations of the same subject were gathered to counter the subjectiveness of the labels. Comparative annotations are believed to be less subjective and hence the number of duplicate descriptions is

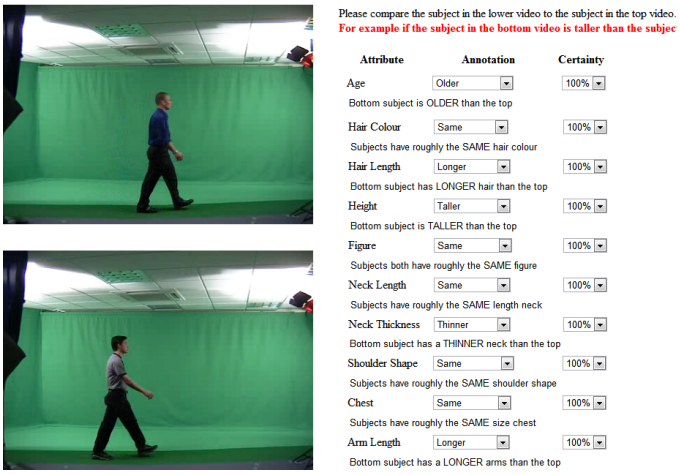

Figure 3. Developed website showing the first part of the experiment

of less importance. Subjects were assigned to users in a way which maximized the number of descriptions comparing different subjects and targets. Performing comparisons between a large group of subjects and a small group of targets also allowed us to infer annotations between subjects. If two subjects were both compared against the same target then the comparison between the two subjects could be inferred, reducing the amount of comparisons required.

A website was developed to record the annotations given by users and is shown in figure 3. The website was designed to allow videos of both the subject and target to be visible onscreen. This allows users to make direct comparisons without memory issues or uncertainties concerning the scale of the videos. Drop down boxes for each trait allowed users to select how the subject differed from the target. The chosen label was emphasized by constructing a sentence explaining the given annotation - ensuring the user was comparing the subject to the target and not the other way around. Eyewitness descriptions can be influenced by providing a default answer to a question. This is

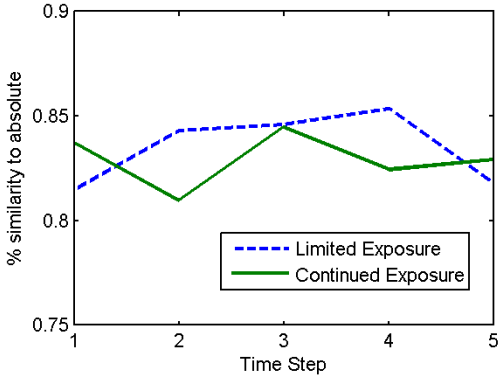

Figure 4. The similarity of comparative and categorical annotations. Time steps represent the five subjects compared to each target

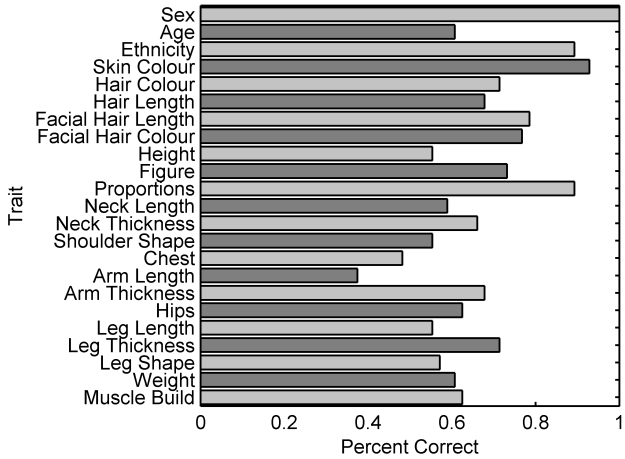

Figure 5. Accuracy of categorical labels after limited exposure

known as anchoring [11]. To avoid anchoring, all drop down boxes were initially set as empty. The design of the website ensured the comparative descriptions were as accurate and correct as possible.

The comparisons were obtained from a predominantly female class of 50 psychology students. Currently there have been 558 comparisons between subjects and targets. An additional 519 subject-subject comparisons have been inferred from the subject-target comparisons.

4 Analysis of Comparisons

Initial analysis directly compared the comparative annotations with the absolute categorical labels gathered by Samangooei and Nixon [1]. This comparison between annotation techniques will not show which is better, only how much each technique differs from the other. To determine the similarity of the descriptions the comparative label is compared against the absolute labels used to annotate the subject and target. If the absolute labels differ and the comparative label reflects this difference the annotations are recorded as concurring. The absolute annotations obviously lack detail - two people labeled as 'tall' are unlikely to be exactly the same height. Small differences can be described using comparative annotations but not absolute labels. In the case of both the subject and target having the same absolute label, the similarity of the comparative annotation cannot be determined. In this case the comparative annotation was recorded as concurring - this ensures we do not overestimate the difference between absolute and comparative annotations.

Figure 4 shows the similarity of the comparative annotations in respect to the absolute. We can see that the comparative annotations differ from the absolute 20% of the time. This does not necessarily mean that the comparative annotations are better - just that they are considerably different. Figure 4 also shows the differences between the two stages of the experiment. It was expected that over time the annotations obtained from the second stage of the experiment would include more errors, since human memory is subject to both decay and interference. It can be observed that the similarity to the absolute annotations was constant over both parts of the experiment - implying that memory did not significantly affect the comparative annotations. This may be the result of a relatively short delay between observing the target and subjects. Future work would be well directed to examining the effects of longer delays.

Figure 5 shows the accuracy of the absolute labels gathered at the end of the second experiment. The annotations described the target, who had not been seen for ten minutes on average. These annotations were compared to annotations of the same subject collected by Samangooei and Nixon [1]. [1] collected descriptions of a single subject from multiple users (on average 9 users) which reduced the influence of the subjective errors, for this reason these annotations were treated as a 'ground truth'. The annotation was deemed to be correct if it matched the mode of the ground truth labels used to describe the subject. Large errors of 32% were present within the annotations when compared to the ground truth. This indicates that absolute categorical labels are actually prone to error after relatively short time periods unlike the comparative descriptions.

The F-ratios, derived by ANOVA analysis, presented within [1] clearly show that absolute labels describe some features better than others. Figure 6 shows the average similarity between absolute and comparative annotations for each trait. Large differences between absolute and comparative labels for traits demonstrated to be difficult to describe using absolute labels would be indicative of potential improvements when using comparative labels. It can be seen that comparative annotations of arm length (one of the hardest traits to explain categorically) differs by 30% compared to absolute labels. Given the inaccuracy of absolute labels in regards to this trait, the difference suggests that the comparative annotations contain new and more detailed information. Conversely, small differences for traits which were accurately described using absolute annotations, for example hair length, demonstrate that the trait is reliably described using both approaches. It can be observed that the difference between absolute and comparative annotations

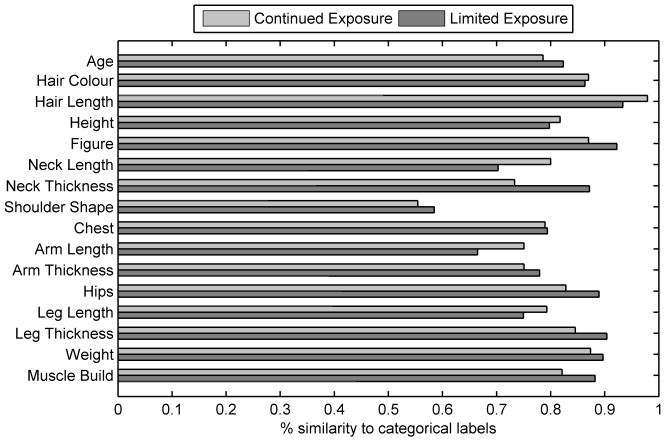

Figure 6. The average similarity of each comparative trait and categorical annotations

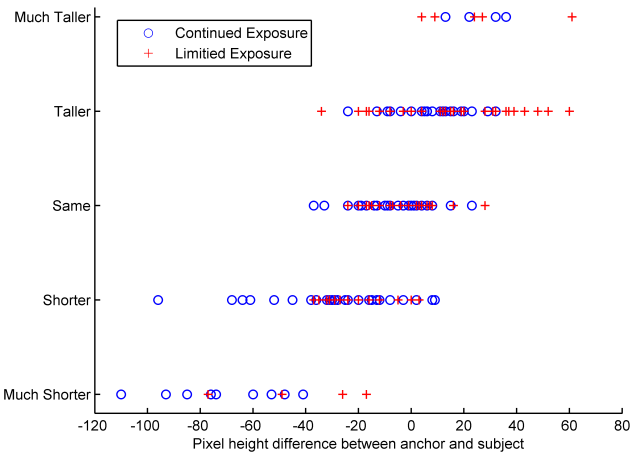

Figure 7. Differences between subject's and target's height - Actual pixel height difference against the height comparison

are 5% in respect to hair length, which shows that the comparisons are largely the same as the successful data obtained from the absolute annotations.

Evaluating human descriptions using real world measurements is very difficult. Trait descriptions are often based on multiple physical characteristics which are difficult to measure consistently. Height can be accurately and consistently measured from video data, a subject's pixel height can be automatically obtained from their gait signature. Figure 7 shows the difference in pixel height compared to the comparative annotations. The correlation (Pearson's R) between the difference in pixel height and the comparative labels is 0.75 - showing a statistically significant relationship between the labels and the difference in pixel height of the subjects. It is important to note that these comparative labels represent a single user's annotation, in comparison figure 2 shows the median absolute label from an average of nine user annotations (resulting in a correlation of 0.71). This implies that the comparative annotations strongly represent the subjects' physical attributes and are more robust against errors originating from subjectiveness. It was found that comparisons made after a limited exposure to the target were less correlated than the continued exposure comparisons (correlations of 0.68 and 0.77 respectively) - this implies that the limited exposure comparisons did not represent the differences in height as accurately.

5 Human Identification

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.

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 a standard algorithm which provides a method of inferring a relative measurement from comparisons [12]. The relative measurement details how the subject's trait compares to other subjects within the population. A subject's relative measurements can be treated as a biometric signature used for identification.

Biometric retrieval identifies an unknown subject by comparing their biometric signature to a database of biometric signatures. This can be used to evaluate the distinctiveness of a relative measurement feature vector. Retrieval will be performed on an 80 subject database using varying amounts of test comparisons, n. This investigates how many comparisons are required to accurately retrieve a subject. Ideally we would seek to minimize the number of comparisons required to identify a subject. n comparisons will be randomly sampled and used to generate the relative measurement 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. The biometric signatures within the database will consist of 19 relative measurements describing the traits shown in table 1. The Euclidean distance between two relative measurement signatures will be used to indicate their similarity. Random sampling will be repeated until the retrieval accuracy remains constant for 10 random samples.

The rank 1 retrieval accuracy over varying number of probe comparisons is shown in figure 8. 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 [1]. As more comparisons are exploited 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 com-

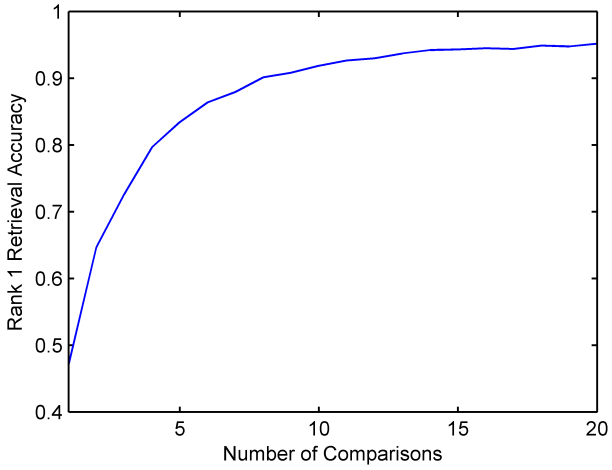

Figure 8. Rank 1 retrieval accuracy using relative measurements obtained from different amounts of comparisons

parisons. This promising result shows that relative measurements provide more discriminatory information than absolute descriptions.

6 Conclusions

Soft biometrics exploit human descriptions to allow human identification. The main advantage over traditional biometrics is the ability to recognize and retrieve subjects based solely on a human description. Conventional human descriptions are typically comprised of absolute labels and estimations of continuous attributes. These forms of description can be unreliable due to the errors common with estimations and the subjectiveness of absolute labels. Human comparisons offer a potentially more robust method of obtaining human descriptions by exploiting objective comparative labels which can be used to infer informative relative measurements.

Comparisons between subjects from the Soton gait database were collected. Each annotator was asked to compare a single target to multiple subjects. The first experiment explored the benefits of comparative annotations in ideal settings and the second investigated the robustness of comparisons over a time delay. It was found that comparative descriptions differed on average by 20% when compared against absolute categorical descriptions. This difference remained constant across both the first and second experiments, showing that elapsed time had little effect on comparative descriptions. In contrast absolute descriptions showed an error rate of 32% after a limited exposure to the subject.

Comparative annotations of traits which were poorly described using absolute labels were found to differ by up to 40%, suggesting the comparative annotations contained new and more detailed information. Relative measurements inferred from multiple comparisons were demonstrated to allow accurate recognition of subjects. After ten comparisons the rank 1 retrieval accuracy was found to be 92% from a database of 80 subjects and after 20 comparisons a 95% retrieval accuracy was achieved. In comparison absolute descriptions achieved a rank 1 retrieval rate of only 48%. Comparative descriptions have been shown to contain more discriminative information and present an innovative approach to obtaining robust human descriptions for soft biometrics and possibly eyewitness descriptions.

References

[1] S. Samangooei and M. S. Nixon, "Performing Content-based Retrieval of Humans using Gait Biometrics," *Multimedia Tools and Applications*, vol. 49, no. 1, pp. 195–212, 2010.

[2] H. Ailisto, M. Lindholm, S. M. Makela, and E. Vildjiounaite, "Unobtrusive user identification with light biometrics," in *Proc. NordiCHI*, 2004, pp. 327–330.

[3] A. K. Jain, K. Nandakumar, X. Lu, and U. Park, "Integrating faces, fingerprints, and soft biometric traits for user recognition," in *BioAW*, vol. LNCS 3087, 2004, pp. 259–269.

[4] U. Park and A. K. Jain, "Face Matching and Retrieval Using Soft Biometrics," *IEEE Trans on Information Forensics and Security*, vol. 5, no. 3, pp. 406–415, Sep. 2010.

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

[6] S. L. Sporer, "An archival analysis of person descriptions," in *Biennial Meeting of the American Psychology-Law Society in San Diego, California*, 1992.

[7] J. C. Yuille and J. L. Cutshall, "A case study of eyewitness memory of a crime," *Journal of Applied Psychology*, vol. 71, no. 2, pp. 291–301, 1986.

[8] Flin and Shepherd, "Tall Stories: Eyewitnesses' Ability to Estimate Height and Weight Characteristics," *Human Learning: Journal of Practical Research & Applications*, vol. 5, no. 1, pp. 29–38, 1986.

[9] S. A. Christianson, "Emotional stress and eyewitness memory: A critical review." *Psychological Bulletin*, vol. 112, no. 2, pp. 284–309, 1992.

[10] J. Shutler, M. Grant, M. S. Nixon, and J. N. Carter, "On a large sequence-based human gait database," in *Proc RASC*. Springer Verlag, 2002, pp. 66–72.

[11] G. B. Chapman and E. J. Johnson, "Incorporating the irrelevant: Anchors in judgments of belief and value." *Heuristics and Biases: The Psychology of Intuitive Judgment*, pp. 120–138, 2002.

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