Human Identification using Facial Comparative Descriptions

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

Eyewitness descriptions are vital for many criminal investigations, although typically still require manual discovery of possible suspects. Soft biometrics introduce a possibility to automatically search databases based on biometric features obtained from verbal descriptions. In this paper we introduce the use of comparative human descriptions for facial identification. Twenty-seven comparative traits are used to accurately describe facial features. The Elo rating system is utilized to determine continuous biometric features from multiple comparative descriptions. Experiments on the Soton gait database demonstrate a 96.7% identification accuracy with just three comparisons.

1. Introduction

Throughout history the use of human descriptions obtained from eyewitnesses has instigated the identification and apprehension of suspects. Humans naturally use labels and estimations of physical attributes to describe people. Typically eyewitness descriptions are used to coarsely search criminal databases using broad descriptions like gender, race and height [10]. A manual search of the retrieval results is required to identify potential suspects.

Soft biometric features [12] are characteristics which people can naturally describe. This new form of biometrics has allowed individuals to be automatically identified from criminal databases based on bodily descriptions [11]. This greatly speeds up the process of finding possible suspects.

In this paper we will explore whether facial descriptions can be used to identify individuals. Although facial descriptions are not mentioned as often in eyewitness descriptions compared to bodily descriptions, they are vital in many serious crime investigations. For this reason the identification of possible suspects is an important problem.

We will show how descriptions of facial soft biometric traits can be used to accurately identify individuals from a criminal database containing soft biometric information. Underpinning this advancement is the use of an innovative form of human description - comparative labels. Previously absolute bodily descriptions were utilized to identify individuals [13], achieving an identification accuracy of 48% [11]. Absolute labels were shown to be a poor form of description, suffering from subjectivity and subject interference [11]. Comparative labels have been found to be less subjective than traditional forms of description and are preferred by the majority of annotators. Furthermore, informative continuous relative measurements can be inferred from multiple comparisons, providing the level of detail required for identification. Previous research studying bodily comparisons of soft biometric traits achieved identification accuracies of 92%, demonstrating the advantages of comparative descriptions [11]. We exploit the ease of making comparisons to explore a new method to provide reliable and robust facial descriptions.

2. Facial Descriptions in Policing

Ideal physical traits for use within a soft biometric system would be easily identifiable at a distance and memorable. Traits which are frequently mentioned within eyewitness descriptions are most likely to adhere to these two requirements.

Van Koppen and Lochun [17] performed a large study into the content of 1313 human descriptions. The descriptions were obtained from written statements given by eyewitnesses following a robbery. It was discovered that only 5% of descriptions contained any inner facial features (for example eye colour, nose, mouth, eye shape and teeth). Sporer [15] analysed the content of 139 descriptions obtained from 100 witnesses. It was found that 29.6% of the descriptions explained facial features, of which the majority of the descriptors described the hair and facial hair of the suspect rather than inner facial features. Inner facial features are not frequently mentioned in eyewitness descriptions. This has been accredited to eyewitnesses not being able to recall discrete features [7] and the lack of vocabulary to describe inner facial features [9, 18].

Typically in serious crimes, facial descriptions and composites are used for identification. Facial composites are

graphical representations of a face generated from descriptions provided by eyewitnesses. Composites were initially created by an artist or by combining images of facial features from an image database [8]. These composites were created based on descriptions of the suspect's individual facial features. Research into their effectiveness highlighted that describing a face is difficult due to a lack of vocabulary, so relying on techniques which require descriptions is not ideal. Modern composites use evolutionary techniques to 'evolve' faces to match the eyewitness' memory. These techniques do not require descriptions and present an entire face to the user. EvoFIT [5] is a popular software package which has been successfully exploited by UK police forces [4].

It can be seen that human descriptions of facial soft biometric traits still play a large role in law enforcement. The lack of vocabulary to describe facial features represents a barrier when collecting and exploiting facial descriptions. Traditional facial descriptions consist of categorical labels which are very subjective and hence often inaccurate. In this paper we utilize comparative labels. Comparative descriptions describe the differences between faces reducing the subjectivity associated with categorical labels.

3. Comparative Descriptions

Comparing the appearance of two subjects is a very natural process. Intuitively it is easy to say whether one person is taller than another, but labelling or estimating the height in absolute terms can be much more difficult. We exploit the ease of making comparisons to provide reliable and robust descriptions.

In section 2 we discussed the issues with conventional forms of facial description. Absolute labels require little skill to annotate but due to their categorical nature have little discriminative capability [11] and are prone to subject interference [2]. Comparative descriptions exploit categorical labels which are easy to understand and annotate. Collecting *multiple* comparisons allows informative continuous measurements to be inferred, providing the level of detail required for identification.

Human descriptions are inherently subjective; the process of selecting an estimate or label is based on the individual. However, absolute labels can be considered *highly* subjective due to the subjective internal benchmark by which the label is being assigned. Generally a label is based on the annotator's understanding of population averages and variation - this varies making the absolute labels unreliable. Comparative labels are less subjective as the benchmark is external and specified. If two annotators were asked to compare the same pair of subjects, both would annotate based on the same benchmark leading to descriptions which are more robust over different annotators.

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. The Elo rating system is utilized to convert facial comparisons to continuous relative measurements. The Elo system is based on estimating distributions from a limited set of pairwise comparisons producing a ranked scale. More information about this approach can be found in [11].

4. Facial Comparisons

Psychological research has determined that descriptions of inner facial features suffer in accuracy due to a lack of vocabulary [9]. Visual comparisons allow features to be described in a natural way using comparative labels. This offers a defined vocabulary whilst avoiding subjective absolute labels, like 'big'. Although this does not make the features more memorable it could facilitate accurate descriptions for cases where the eyewitness has observed and encoded the suspect's face. This could be exploited for searching databases of mugshots or the description could be used to seed the generation of composites in programs like EvoFIT [5].

Although facial features are not as common in eyewitness descriptions as bodily and global traits, they are vital in many serious crime investigations. Exploring the capabilities of visual comparisons could present solutions to the lack of objective vocabulary for describing facial features.

4.1. Defining Facial Comparisons

This section will define facial comparisons and discuss how comparisons are evaluated and utilized.

A *facial comparison* is a set of individual soft trait comparisons describing the differences between two subjects. In application settings, an eyewitness would compare the previously observed suspect to other subjects (possibly obtained from a video or image database). This allows information about the suspect to be inferred from the appearance of the subject and the comparison describing the differences between the two individuals.

Although descriptive, a single comparison between a suspect and another person will only explain the differences between the two. Thus, the inferred physical traits of the suspect will depend on the subject they were compared to. Multiple comparisons must be available to infer a more robust description, with each comparison allowing the description of the suspect to be refined. Therefore, ideally multiple comparisons should be obtained between the observed suspect and multiple subjects.

The experiments within this chapter replicate this application scenario by collecting multiple comparisons between

a *target* subject (representing the suspect in application settings) and multiple *subjects*.

A single facial comparison will describe the differences between the target and subject in terms of individual traits, such as nose length and face width. A *trait comparison* is a comparison of an individual soft trait.

4.2. Traits

Selecting optimal traits is vital in obtaining accurate descriptions and conveying as much information about a face as possible. A subset of traits from the Aberdeen University face rating schedule (FRS) [3] were used in this research. The FRS features a comprehensive selection of traits and has been used in other studies [16, 6]. The FRS contains 53 absolute traits, the majority described using 5 point bipolar scales. The modified FRS introduced in [16] was used as a base for the traits used in this study.

Several modifications were made to the FRS. Many traits, which recorded the presence of facial hair, glasses and jewellery, have been excluded as they describe temporary features and do not lend themselves to the comparative nature of the experiment. Traits describing colour were also excluded, hair colour had been explored in previous experiments [11] and the facial images used in this experiment are too low resolution to accurately identify eye colour.

The final set of 27 comparative traits are presented in table 1. Each trait is described using a 5 point bipolar scale, the extremes of which are represented by two labels (an example of this can be seen in figure 1).

4.3. Data Acquisition

An experiment was designed to assess the advantages of comparative descriptions when describing facial features. In particular whether comparative labels improve the accuracy of inner facial feature descriptions, by reducing the subjectivity associated with absolute labels and providing a defined and understandable vocabulary.

Comparisons were made between frontal and side facial images of the 100 subjects in the Soton gait database (SGDB) [14]. The experiment was split into two parts. The first section asked users to provide absolute descriptions of five subjects from the SGDB. The absolute descriptions were composed of the same 27 traits which were presented in table 1, except absolute labels were assigned to the extremes of the scales. The second section asked users to compare five subjects to a single target, replicating the application scenario of comparing a observed suspect to multiple subjects within a database. Collecting both absolute and comparative descriptions allows the accuracies of both to be directly compared. The 100 subjects within the dataset were halved and assigned to one of the two parts of the experiment. The 50 subjects selected for the comparative facial experiment were designated as one of either 10 targets or 40

Figure 1. Website used to obtain facial comparisons

subjects. Performing comparisons between a large group of subjects and a small group of targets allows comparisons to be inferred between subjects. If two subjects were both compared against the same target, a comparison between the two subjects can be *inferred*, reducing the amount of comparisons required.

Comparisons and absolute descriptions were collected using the website shown in figure 1. The website was designed to display the frontal and side images of both subjects at the same time avoiding any issues with memory. The bipolar scales were implemented using radio buttons which required minimal user input and were found to be very easy to interpret. To avoid anchoring [1] the radio buttons were initially empty, forcing an input from the user. Annotations were emphasized by constructing a sentence explaining the given comparison - ensuring the annotator was comparing the subject to the target instead of vice versa. At the end of the experiment the annotators were encouraged to submit a small feedback form asking which form of annotation they preferred - absolute or comparative.

4.4. Data Analysis

Absolute and comparative descriptions were collected from 63 users. 302 absolute descriptions (describing 50 subjects) and 297 comparisons (comparing 40 subjects to 10 targets) were collected. More information about the collected comparisons and the resulting inferred facial comparisons is shown in table 2. Further information about the absolute annotations can be seen in table 3.

	Collected	Inferred
Total trait comparisons	8019	66501
Total human comparisons	297	2463
Average human comparisons per subject	7.3	61.5
Average human comparisons per target	29.1	N/A
Average human comparisons per subject-target pair	0.73	N/A
Average human comparisons per subject-subject pair	N/A	1.6

Table 2. The number of collected and inferred facial comparisons

Feature	Low Label	High Label	Feature	Low Label	High Label
Face	Shorter	Longer	Ears	Smaller	Larger
Face	Narrower	Wider	Ears	Closer to head	Further from head
Face	More Bony	More Fleshy	Ears	More Hidden	More Evident
Skin	Lighter	Darker	Chin and Jaw	More Angular	More Round
Skin	Smoother	More Wrinkles	Chin and Jaw	More Receding	More Protruding
Skin	Clearer	More Pimples	Lips	Thinner	Thicker
Eyebrows	Thinner	Bushier	Nose	Flatter	More Protruding
Eyebrows	Lower	Higher	Nose	Shorter	Longer
Eyebrows	Closer Together	Further apart	Nose	Narrower	Wider
Eyebrows	Straighter	More Arched	Nose	More Upturned	More Hooked
Forehead	Smaller	Larger	Eyes	Smaller	Larger
Forehead	Straighter Hairline	More Receded Hairline	Eyes	More Slanted	Rounder
Hair	Shorter	Longer			
Hair	Straighter	Curlier			
Hair	Thinner	Thicker			

Table 1. Facial features used to compare subjects

	Collected
Total trait annotations	8154
Total human annotations	302
Average human annotations per subject	6.2

Table 3. The number of collected absolute facial annotations

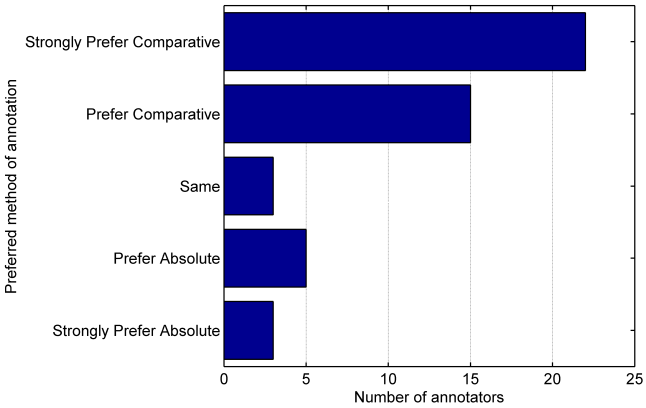

Figure 2. Annotators' preferred form of facial annotation

48 annotators chose to submit the feedback form at the end of the experiment stating which form of annotation they preferred. The results can be seen in figure 2. It is clear to see that the majority of the annotators (77%) preferred comparisons over absolute annotations. Only 16.6% of the annotators preferred absolute annotations. The inclination towards comparative annotations may be due to the simplicity of objective comparative labels.

Figure 3 shows the correlation between the facial comparative features. The correlations between traits were calculated using Elo relative measurements deduced from the comparative labels. The white cells within the figure represent traits with high correlation and the black cells represent traits with no correlation. It can be seen that there is very little correlation between the features. The lack of correlation highlights the independence of each facial trait, this is ideal for identification as each trait comparison conveys new and potentially discriminatory information. It should be noted that the low correlation does not mean that there is not a relationship between the features only that it is not prevalent within the dataset currently being used.

Comparing absolute and comparative labels allows us to observe the differences between the two forms of description. To determine the difference between 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 - for example if the target and subject noses were labelled as 'short' and 'long' respectively and the comparative descriptor provided was 'longer', we would consider both annotations as concurring. The absolute annotations obviously lack detail; two people labelled as having 'long noses' are unlikely to have exactly the same length nose. Thus, 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 difference between absolute and comparative facial descriptions. On average the descriptions differ by 26.3%. The traits which are most similar to absolute descriptors are prominent facial features, including traits like skin-light/dark, face-bony/fleshy and hair-short/long. These traits are easily recognized due to their prominence and therefore individuals have an understanding of the traits' averages and variation, this could explain why the absolute descriptions of these traits are comparatively similar to the comparative annotations. Traits such as face-short/long, ears-small/large and eyebrows-straight/arched may suffer from a lack of noticeable variation leading to large differences between the two forms of description. Small variations are difficult to describe using absolute labels and may not even be noticed due to the trait looking 'normal' or 'average'. Comparisons allow variation

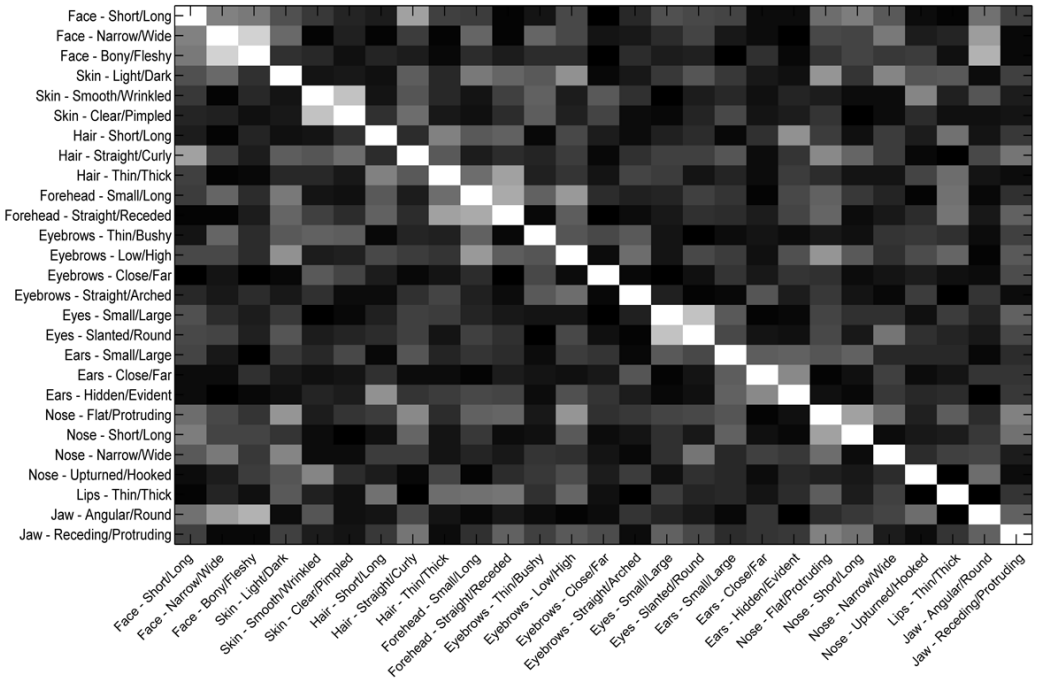

Figure 3. Correlation between facial comparisons. White cells represent strong correlations. Black cells represent weak correlations.

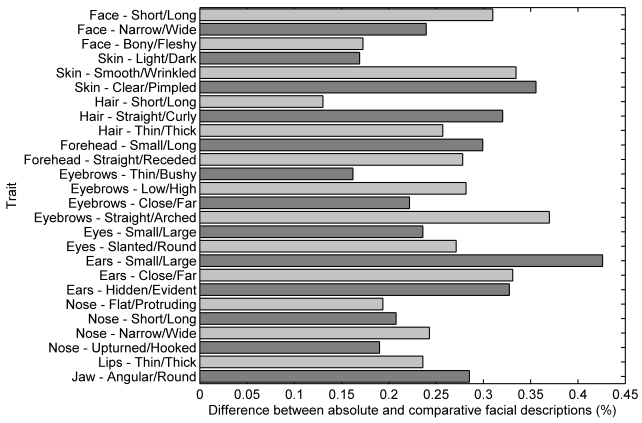

Figure 4. Differences between absolute and comparative facial descriptions

to be identified and accurately described leading to vast differences between absolute and comparative descriptions.

5. Identification using Facial Comparisons and Descriptions

5.1. Technique

Facial recognition was conducted using both the comparative and absolute descriptions collected, allowing the performance of each to be compared.

Comparative facial recognition aims to retrieve a suspect from an 40 subject database. The biometric signatures within the database consist of all the 27 traits (table 1), where comparative traits are represented as Elo relative measurements. The process starts by selecting a *suspect* from the database. n randomly sampled comparisons between the suspect and other subjects were removed from the database and used to infer the suspect's biometric signature used to query the database (known as the probe). This replicates the eyewitness comparing the suspect to n subjects from the database. n was varied to investigate how many comparisons are required to retrieve a suspect accurately. The suspect's remaining comparisons were used to produce the biometric signature stored within the database (known as the gallery). The remaining 39 subjects' feature vectors within the database were determined from all the available comparisons (excluding any comparisons used to construct the suspect's probe feature vector). The similarity between the probe and gallery feature vectors was assessed using the sum of the Euclidean distance. The subjects were ordered based on their similarity to the probe. The position of the suspect's gallery biometric signature within the ordered list shows the retrieval performance of the system. If the suspect's gallery signature is first in the ordered list the suspect has been successfully identified. This process was repeated 100 times for each subject and for each n.

Identification using absolute facial descriptions utilized

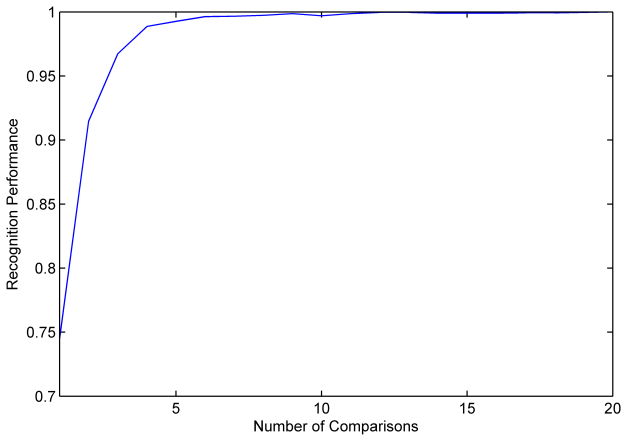

Figure 5. Facial recognition accuracy using relative measurements obtained from different numbers of comparisons

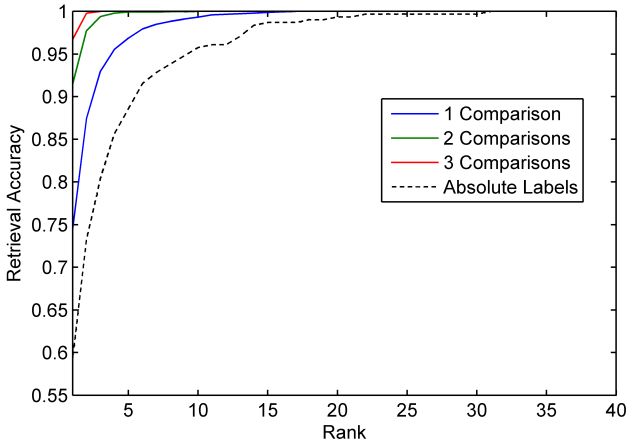

Figure 6. Face retrieval accuracy of absolute labels and relative measurements inferred from 1-3 comparisons

the same 27 traits, each being described using absolute ordinal labels (represented using a value ranging from -2 to 2). A leave-one-out validation approach was used to evaluate the recognition performance. Every description given was individually used to probe the database. The probe feature vector was formed from a single verbal description of a subject given by a single annotator. The remaining descriptions of the subject were used to produce the feature vector present within the database being searched. On average each subject was described by 6 users, the most frequently used label to describe a trait was used to produce the biometric signature describing the subject. The database consisted of 50 subjects, none of which were included within the comparative facial database. The Euclidean distance metric was used to evaluate the similarity between the probe and gallery feature vectors - this was possible due to the ordinal nature of the labels. The subjects were ordered based on their similarity to the probe. The position of the suspect's gallery biometric signature within the ordered list shows the retrieval performance of the system.

The identification results shown in this research are obtained from exhaustively calculating the similarity between the probe and each gallery signature. For larger databases this process could be accelerated by filtering the subjects based on soft biometric features which are reliably and accurately described.

5.2. Accuracy

The facial recognition accuracy over varying numbers of probe comparisons is shown in figure 5. It can be seen that facial comparative descriptions vastly outperform bodily descriptions [11], achieving a 74.5% identification accuracy with a single comparison (compared to a 47% accuracy with bodily comparisons). A 99.3% recognition accuracy is obtained with just five comparisons, reaching a maximum of a 100% accuracy at 20 comparisons.

Facial descriptions have three benefits which aid in identification when compared to bodily descriptions. It was shown in section 4.4 that facial features have little correlation, resulting in more independent information available for identification. This increases the feature space by many dimensions, typically making each subject more distinctive and easier to identify. Body comparisons can be effected by many types of covariates. In the SGDB baggy clothes often hide features from the annotator. Faces have far fewer covariates. Glasses are a very common covariate within the SGDB (around 47 people wear glasses) but these rarely interfere with the observation of features, whilst only 6 people have facial hair within the database. This results in the features being very evident and easy to describe - improving the descriptions. Finally faces have much more features to describe. We collect 27 facial trait descriptions which results in typically more distinctive descriptions allowing greater accuracy when identifying subjects.

The retrieval accuracy of the facial absolute labels is shown in figure 6, along with the retrieval accuracy of facial comparisons inferred from 1-3 comparisons. It can be seen that comparisons outperform the absolute facial labels even with just one comparison. The identification performance (i.e. the rank 1 retrieval accuracy) of absolute labels was found to be 59.3% compared to 74.5% achieved with relative measurements inferred from one comparison. The identification performance increases with additional comparisons, achieving a 96.7% identification accuracy with only 3 comparisons. These results, obtained under ideal conditions, show the potential of facial comparisons.

6. Conclusions

Facial descriptions, although infrequently mentioned in eyewitness statements, play a large role in many serious crime investigations. The lack of vocabulary to describe facial features is a major cause of inaccurate and unreliable facial descriptions. In this paper we have introduced the concept of comparative facial descriptions which produce accurate and robust annotations by exploiting a defined and objective vocabulary. The Elo rating system is utilized to produce continuous discriminative biometric features from verbal comparisons, allowing biometric identification of individuals. With a single comparison an identification accuracy of 74.5% is achieved. Obtaining more comparisons improves the identification accuracy, achieving 99.3% with five comparisons. These results show that comparative facial descriptions can be used to automatically find possible suspects. Future research will explore the application potential of identification using facial comparisons, considering larger databases, inaccuracy resulting from memory decay and the accuracy of comparing observed suspects to videos of subjects.

References

[1] 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*, pages 120–138, 2002.

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

[3] H. D. Ellis. Face recall: A psychological perspective. *Human Learning: Journal of Practical Research & Applications; Human Learning: Journal of Practical Research & Applications*, 1986.

[4] C. D. Frowd, P. J. B. Hancock, V. Bruce, A. H. McIntyre, M. Pitchford, R. Atkins, A. Webster, J. Pollard, B. Hunt, E. Price, S. Morgan, A. Stoika, R. Dughila, S. Maftei, and G. Sendrea. Giving Crime the 'evo': Catching Criminals Using EvoFIT Facial Composites. In *International Conference on Emerging Security Technologies (EST)*, pages 36–43, 2010.

[5] C. D. Frowd, P. J. B. Hancock, and D. Carson. EvoFIT: A holistic, evolutionary facial imaging technique for creating composites. *ACM Trans. Appl. Percept.*, 1(1):19–39, July 2004.

[6] J. Kabzińska and A. Niedźwieńska. The effect of providing descriptions of perpetrators on their identification by eyewitnesses and investigative bodies. *Problems of Forensic Sciences*, (84):326–335, 2011.

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

[8] K. R. Laughery and R. H. Fowler. Sketch artist and Identi-kit procedures for recalling faces. *Journal of Applied Psychology*, 65(3):307, 1980.

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

[10] National Policing Improvement Agency. *PNC User Manual, Volume 2*, Nov. 2009.

[11] D. A. Reid and M. S. Nixon. Using Comparative Human Descriptions for Soft Biometrics. In *International Joint Conference on Biometrics (IJCB)*, pages 1–6, 2011.

[12] D. A. Reid, S. Samangooei, C. Chen, M. S. Nixon, and A. Ross. Soft Biometrics for Surveillance: An Overview. In *Handbook of statistics*, volume 31. Elsevier, In Press.

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

[14] 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.

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

[16] S. L. Sporer. Person descriptions as retrieval cues: Do they really help? *Psychology, Crime & Law*, 13(6):591–609, Dec. 2007.

[17] P. J. Van Koppen and S. K. Lochun. Portraying perpetrators; the validity of offender descriptions by witnesses. *Law and Human Behavior*, 21(6):661–685, 1997.

[18] M. S. Wogalter. Effects of Post-exposure Description and Imaging on Subsequent Face Recognition Performance. In *Human Factors and Ergonomics Society Annual Meeting Proceedings*, volume 35. Human Factors and Ergonomics Society, 1991.