UNIVERSITY OF SOUTHAMPTON

Human Identification using Soft Biometrics

by

Daniel A. Reid

A thesis submitted in partial fulfillment for the
degree of Doctor of Philosophy

in the
Faculty of Physical and Applied Sciences
Department of Electronics and Computer Science

April 2013

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF PHYSICAL AND APPLIED SCIENCE
DEPARTMENT OF ELECTRONICS AND COMPUTER SCIENCE

Doctor of Philosophy

by Daniel A. Reid

Humans naturally use descriptions to verbally convey the appearance of an individual. Eyewitness descriptions are an important resource for many criminal investigations. However, they cannot be used to automatically search databases featuring video or biometric data - reducing the utility of human descriptions in the search for the suspect. Soft biometrics are a new form of biometric identification which uses physical or behavioural traits that can be naturally described by humans. This thesis will explore how soft biometrics can be used alongside traditional biometrics, allowing video footage and biometric data to be searched using a description.

To permit soft biometric identification the human description must be accurate, yet conventional descriptions comprising of absolute labels and estimations are often unreliable. A novel method of obtaining human descriptions will be introduced which utilizes comparative categorical labels to describe the differences between subjects. A database of facial and bodily comparative labels is introduced and analysed.

Prior to use as a biometric feature, comparative descriptions must be anchored. Several techniques to convert multiple comparative labels into a single relative measurement are explored. Recognition experiments were conducted to assess the discriminative capabilities of relative measurements as a biometric.

Relative measurements can also be obtained from other forms of human representation. This is demonstrated using several machine learning techniques to determine relative measurements from gait biometric signatures. Retrieval results are presented showing the ability to automatically search video footage using comparative descriptions.

Contents

Abstract i

Abbreviations v

Declaration Of Authorship vi

Acknowledgements vii

1 Context and Contributions 1

2 On Human Descriptions and Soft Biometrics 4

2.1 Human Descriptions . . . 4

2.1.1 Content of Descriptions . . . 4

2.1.2 Accuracy of Descriptions . . . 6

2.1.3 Human Descriptions in Policing . . . 7

2.2 Soft Biometrics . . . 9

2.2.1 Forensic Anthropometry . . . 10

2.2.2 Incorporating Soft Biometrics in a Fusion Framework . . . 12

2.2.3 Human Identification Using Soft Biometrics . . . 14

2.2.3.1 Imputation . . . 19

2.2.4 Applications . . . 22

2.2.4.1 Continuous Authentication . . . 22

2.2.4.2 Surveillance and Re-Identification . . . 22

2.3 Conclusions . . . 24

3 Comparative Human Descriptions **25**

3.1 Relative Information . . . 26

3.2 Defining and Evaluating Human Comparisons . . . 27

3.3 Body Comparisons . . . 29

3.3.1 Traits . . . 29

3.3.2 Data Acquisition . . . 30

3.3.3 Data Analysis . . . 32

3.4 Facial Comparisons . . . 34

3.4.1 Traits . . . 35

3.4.2 Data Acquisition . . . 36

3.4.3 Data Analysis . . . 37

3.5 Conclusions . . . 41

4 Identification using Comparisons 43
4.1 Relative Measurements . . . 44
4.1.1 Pairwise Comparisons . . . 44
4.1.2 Thurstone's Model . . . 45
4.1.3 Elo Rating System . . . 48
4.1.4 Relative Measurement Accuracy . . . 50
4.2 Identification using Bodily Comparisons . . . 54
4.2.1 Technique . . . 54
4.2.2 Accuracy . . . 55
4.3 Identification using Facial Comparisons and Descriptions . . . 57
4.3.1 Technique . . . 57
4.3.2 Accuracy . . . 58
4.4 Conclusions . . . 60

5 Retrieval from Video Footage 61
5.1 Gait . . . 62
5.1.1 Gait Signatures . . . 63
5.1.1.1 Average gait signature . . . 64
5.1.1.2 Differential gait signature . . . 65
5.1.1.3 Unwrapped gait signature . . . 65
5.1.1.4 Measured gait signature . . . 66
5.2 Converting Gait Signatures to Soft Feature Vectors . . . 67
5.2.1 k Nearest Neighbours . . . 68
5.2.2 Support Vector Machine . . . 68
5.2.3 SVDImpute . . . 69
5.2.4 Accuracy . . . 71
5.2.4.1 Relative Measurements . . . 72
5.2.4.2 Absolute Traits . . . 75
5.3 Retrieval . . . 76
5.3.1 Genetic Algorithm Trait Weighting . . . 77
5.3.2 Retrieval Results . . . 78
5.4 Conclusions . . . 80

6 Memory and Human Comparisons 81
6.1 Eyewitness Memory . . . 81
6.2 Short Time Delays and Interference . . . 82
6.3 Long Time Delays . . . 84
6.4 Discussion and Conclusions . . . 86

7 Conclusions and Future Work 88
7.1 Conclusions . . . 88
7.2 Future Work . . . 88
7.2.1 Facial Retrieval . . . 88
7.2.2 Additional Comparisons . . . 89
7.2.3 Exploring Memory . . . 89
7.2.4 Bodily Retrieval Improvement . . . 90
7.2.5 Relative Measurement Refinements . . . 91

7.2.6 Imputation for Human Comparisons 91
7.2.6.1 Imputation of Absolute Descriptions 92

A Soton Gait Database 97

References 99

Abbreviations

ANOVA	Analysis of Variance
CCR	Correct Classification Rate
CCTV	Closed Circuit Television
CDF	Cumulative Distribution Function
Correlation	Pearson's Product-Moment Correlation Coefficient
EER	Equal Error Rate
FRS	Face Rating Schedule
k**NN**	k Nearest Neighbours
LSA	Latent Semantic Analysis
MAE	Mean Absolute Error
PNC	Police National Computer
RBF	Radial Basis Function
SGDB	Soton Gait Database
SVD	Singular Value Decomposition
SVM	Support Vector Machine

Declaration Of Authorship

I, Daniel Reid declare that this thesis titled, "Human Identification using Soft Biometrics" and the work presented in it are my own and has been generated by me as the result of my own original research. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University;
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
- Where I have consulted the published work of others, this is always clearly attributed;
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
- I have acknowledged all main sources of help;
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;

Signed: ____

Date: ____

,

Acknowledgements

I would like to thank my supervisor Professor Mark Nixon for giving me the opportunity to perform this research and for his help throughout the last three years. His support and guidance made this research possible, along with his knowledge and experience of dealing with new research areas.

My thanks also go to Dr. Sarah Stevenage whose advice has been critical to the psychological aspects of this project.

Researching into a new form of human description requires a lot of annotations. I thank my friends, family and the many psychology students who have given their time to help me collect human comparisons, without these descriptions this research would not have been possible.

Finally, I would like to thank my wife and family for their encouragement and support before and during my academic life.

Chapter 1

Context and Contributions

Biometrics provide an automated method to identify people based on their physical or behavioural characteristics. Previously, this consisted of biometrics which required cooperation from the individual. Biometrics such as fingerprints and DNA have been extensively used by the police. In recent years, the increased threat of terrorist activities and the ever growing surveillance infrastructure has driven the development of biometrics which operate at a distance. These have the ability to recognize people from surveillance footage without their cooperation. This is crucial in quickly identifying known criminals or suspects. Face, ear and gait biometrics are the most popular long distance biometrics.

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. Due to the differences between how humans and computers identify people, descriptions cannot be utilized to automatically identify an individual. This is known as a semantic gap. This project aims to use soft biometrics to bridge this gap. Jain et al. [1] defined soft biometric traits as 'characteristics that provide some information about the individual, but lack the distinctiveness and permanence to sufficiently differentiate any two individuals'.

In this thesis we will redefine soft biometrics. *Soft biometric traits are characteristics which people can naturally describe.* We will show how descriptions of soft biometric traits can be used to accurately identify people. Furthermore, we will show how the semantic gap can be bridged, allowing a video database to be searched automatically using human descriptions. Underpinning these advancements is the use of an innovative form of human description - comparative labels. 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 explore a new method to provide reliable and robust descriptions.

The background information referred to throughout this thesis is presented in chapter 2.

The history of forensic anthropometry and the modern day uses of human descriptions in policing are discussed. The content and accuracy of human descriptions is explored and the findings used to justify the main contribution of this project - comparative descriptions. We survey the current soft biometric techniques used in fusion architectures and standalone identification approaches as well as the current applications of soft biometrics [c2].

The notation of using comparative descriptions to accurately describe individuals is introduced in chapter 3 [c3]. We discuss why comparative descriptions are needed and what problems we hope to solve. Justification of the benefits of relative information is found in other studies within the field of image description and retrieval. Based on previous studies in soft biometrics, eyewitness description analysis and psychology, a set of physical features, both bodily and facial, are defined for use in this project. The comparative description database used throughout this research is introduced in this chapter. The experiments and websites used to collect descriptions are detailed and their designs are justified. An analysis of the data obtained from the experiments is shown and we present an in depth look into the correlations between the physical traits.

Chapter 4 explores the discriminative capabilities of comparative descriptions by identifying individuals [c4,c6]. Before exploiting comparative descriptions they must first be anchored to provide a single measurement - known as a relative measurement. We explore three different methods of converting comparative descriptions to relative measurements. We investigate the impact of subjectiveness on comparisons by examining the correlation between relative and real world measurements. The identification results demonstrate the biometric properties of relative measurements.

One of the main goals of this research is to bridge the semantic gap and use human descriptions to search a biometric database. In chapter 5 we explore how videos can be automatically searched using comparative descriptions [c5]. This is achieved by converting gait signatures, obtained from video footage, to relative measurements which can than be queried. In this chapter we explain gait biometrics and the various gait signatures experimented with. We introduce the machine learning techniques used to calculate relative measurements from gait signatures. Finally, the retrieval performance achieved from querying a video database with a description of an individual is presented.

When utilizing eyewitness descriptions special consideration must be given to the effects of memory. The affects of time delay and interference on comparative descriptions is analysed in chapter 6. Two experiments were conducted to explore these issues, one focusing on short time delays and the other on long time delays. We compare the performance of both absolute and comparative descriptions with limited exposures to the individual being described.

Future research directions are discussed within chapter 7. Exploration of the capabilities of comparative descriptions in application scenarios and under realistic memory conditions are encouraged, a potential experiment examining both aspects is suggested. Imputation techniques may offer an approach to increase the robustness of soft biometrics in real world scenarios where verbal descriptions and visual signatures may be missing data [c1]. Given the success of the facial comparisons within the identification experiments, we recommend the development of facial retrieval. The police store mugshots of suspects within the police national computer, implementing facial retrieval would allow these mugshots to be searched using facial descriptions - potentially identifying suspects. These three major research directions will bring comparative soft biometrics closer to being useful in practical applications.

The papers resulting from this research are listed below in chronological order:

[c1] D. A. Reid and M. S. Nixon, "Imputing human descriptions in semantic biometrics," in Proceedings of the 2nd ACM workshop on Multimedia in forensics, security and intelligence, 2010.

[c2] D. A. Reid, S. Samangooei, C. Chen, M. S. Nixon, and A. Ross, "Soft biometrics for surveillance: An overview," in Handbook of Statistics vol 31. Elsevier, In Press.

[c3] D. A. Reid, M. S. Nixon, and S. V. Stevenage, "Identifying humans using comparative descriptions," in International Conference on Imaging for Crime Detection and Prevention (ICDP), 2011.

[c4] D. A. Reid and M. S. Nixon, "Using comparative human descriptions for soft biometrics," in International Joint Conference on Biometrics (IJCB), 2011.

[c5] D. A. Reid, M. S. Nixon, and S. V. Stevenage, "Soft biometrics; human identification using comparative descriptions," IEEE Transactions on Pattern Analysis and Machine Intelligence, Submitted.

[c6] D. A. Reid and M. S. Nixon, "Human Identification using Facial Comparative Descriptions," in International Conference on Biometrics (ICB), 2013.

Chapter 2

On Human Descriptions and Soft Biometrics

2.1 Human Descriptions

To allow identification from human descriptions, the physical properties described must be accurate, salient and reliable. Human descriptions generally consist of two forms of description: categorical labels and continuous estimations. Labels are predominantly used to describe inherently categorical traits like ethnicity and gender, but they can also be used to describe continuous traits. For example height descriptions can include short, medium and tall. Estimations of continuous traits are more commonly described using measurements detailing the feature's length, width or weight. Much research has been conducted into obtaining accurate human descriptions due to their importance in many criminal investigations. This section introduces the psychological research conducted within the field of human description and explores how human descriptions are currently collected and used by the police.

2.1.1 Content of Descriptions

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.

One of the first studies into the most frequently mentioned descriptors was by Kuehn [2]. This paper studied the descriptions provided by 100 victims immediately after a violent crime (cases of rape, bodily injury or robbery). Nine physical traits were recorded by the police department, although it was not clear whether the descriptions were a result of free speech or questioning. Eight of the nine traits were mentioned by 70% of the victims, these in rank order were : gender, age, height, build, race, weight, complexion, and hair colour. On average a victim described seven physical characteristics and over 85% of the victims mentioned six or more of the traits. Eye colour was the least mentioned trait only being recalled 23% of the time. It was concluded that victims have a general impression of their assailant but cannot recall discrete features like hair and eye colour.

Van Koppen and Lochun [3] performed a large study into the content of 1313 human descriptions. The descriptions were obtained from written statements given by eyewitnesses following a robbery. The features described were categorized into 43 traits, describing bodily and facial features, as well as clothing and accents. On average the median amount of trait descriptions present in a description was eight, of which permanent features (such as gender, height and skin colour) were mentioned more often with a median of five per description. Of the 43 trait categories only nine were described by more than 30% of the witnesses, these include in rank order: gender, height, appearance (which includes race), skin colour, age, build, hair colour, type of hair and accent. It was discovered that only 5% of descriptions contained any inner facial features (for example eye colour, nose, mouth, eye shape and teeth), concurring with the conclusions made by Kuehn [2].

Sporer [4] analysed the content of 139 descriptions obtained from 100 witnesses. It was found that 22% of descriptions detailed physical (race, age, height) and movement features. Another 31% of the descriptors described clothing, 29.6% explained facial features, 5% mentioned personality inference, and 12% 'other' features (including jewellery, dialect, disguise and smell). Of the facial features described 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 [2] and the lack of vocabulary to describe inner facial features [5, 6]. Research has also suggested that facial perception is a holistic process [7] - identification is performed based on the whole face rather than individual features. This could possibly explain why eyewitnesses find it difficult to describe individual facial features.

Based on these studies bodily (height, weight, build) and global (race, gender, skin colour) traits appear to be the most frequently mentioned features. This implies they are the most memorable and easily identified features in criminal situations. An interesting experiment conducted by MacLeod et al. [8] provided an in-depth analysis of the reliability and saliency of bodily traits described using bipolar scales. The most reliable descriptors were discovered in a two step process. The experiment started by requiring participants to exhaustively describe people within still images and videos. From this process 687 descriptors were generated from still images and 1,238 from video. Of the video descriptors 84% described the general physique of the person whilst the remainder described the movement of the subject. From these descriptors 23 of the most common and distinct were selected to produce either 5 point bipolar scales or dichotomous items. Two groups of participants then labelled videos of people based on these 23 descriptors to discover their reliability. The responses of each group were averaged per subject per trait. The product moment correlation was calculated based on the two sets of means and used as a reliability metric for each descriptor. To improve distinction between descriptors, redundancy was discovered and removed. This was achieved by performing a factor analysis on the data, followed by a principal component analysis. This resulted in a reduced set of 13 of the most reliable descriptors. The 5 most reliable descriptors were found to be weight, height, leg thickness, chest size and leg length.

2.1.2 Accuracy of Descriptions

To identify individuals, descriptions must be accurate and reliable. This subsection will focus on analysing the accuracy of continuous estimations and categorical labels for describing humans.

Estimates of height, weight and age are commonly mentioned in human descriptions although are often found to be inaccurate. Yuille and Cutshall [9] showed that estimates of height, weight and age were incorrect 50% of the time based on 95 cases (considered accurate if within 2 inches, 5 pounds and 2 years respectively of the actual measurement). Van Koppen and Lochun [3] found that 52% of 1617 height estimations (within roughly 7cm of the actual height) and 61% of 1258 age estimations (within roughly 7 years) were correct. This inaccuracy was accredited to the witnesses' lack of training and experience at providing accurate estimations [9].

The effect of anchoring is the second largest source of errors when estimating height, weight and age [5]. Anchoring is a cognitive bias that occurs during decision making where judgements are affected by one piece of information. In the case of estimating physical traits, both the estimator's own trait measurements and their knowledge of population averages bias the decision making process. The first study into this bias was performed by Hinckley and Rethlingshafer [10]. 500 twenty-one year old college students were asked to guess the average height of men in America and estimate 28 heights using a nine point scale (representing nine equally spaced height ranges between 4'8" and 6'11"). It was shown that smaller judges estimated the average height of men in America to be significantly less than the estimates provided by tall judges. Short judges also constantly over-estimated the 28 heights presented. This finding confirms that anchoring directly affects the estimation of height. A further study by Flin and Shepherd [11] asked 588 individuals to estimate the height and weight of 14 targets. It was found that the participants used their own height and weight as a reference to judge the target. Descriptions also 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 judgements.

Continuous estimations have huge descriptive potential, although are often estimated inaccurately. The combination of anchoring and the skill required to accurately estimate measurements, makes this form of description unreliable and not suitable for soft biometrics.

Absolute labels are predominantly used to describe inherently categorical traits, however they are often subjective. Gender and ethnicity are the two most common categorical traits and have been shown to be easy to distinguish and annotate correctly (100% and 86% accuracy respectively) [3, 12]. Categorical descriptions of the colour and style of hair and clothing are frequently included in eyewitness descriptions. Van Koppen and Lochun [3] found that erroneous descriptions were common when describing dialect and type of hair. These inaccuracies were linked to the subjective nature of the characteristics. Yuille and Cutshall [9] also noted significant errors on descriptions of colour and style of both hair and clothing.

Absolute labels have also been used to describe inherently continuous traits [13], the results from this study are discussed fully in section 2.2.3. Absolute labels are ideal for traits which feature little subjectivity (like gender) but are often inaccurate due to the lack of a standardized meaning.

2.1.3 Human Descriptions in Policing

In 1974 the UK police department began recording information about stolen vehicles in the Police National Computer (PNC) database. Additional databases were added to the PNC, including the ‘names’ database which records details about people who have been previously convicted, cautioned, wanted, missing or recently arrested. The PNC can be accessed by all UK police forces and many other organizations like the Secret Intelligence Service (MI6) and HM Revenue and Customs. Each person entry aims to at least provide age, name, gender, ethnicity and height [14]. Table 2.1 shows additional fields stored within the names database. The characteristics field allows a free description of the suspect’s dress, jewellery, habits and skills.

The QUEST (Queries Using Extended Search Techniques) system [15] was developed to allow the names database to be searched with the aim of generating a list of possible suspects for a crime or event. One of the possible search methods is using a description of the suspect, the search options can be seen in table 2.1. It is evident that the system favours global soft biometric traits and has little information about inner facial or bodily features.

Typically in serious crimes, facial descriptions and composites are used for identification in addition to bodily and global trait descriptions. Facial composites are graphical rep-

Table 2.1: Subset of information recorded within the PNC names database

Trait	Search options
Name	Given, surname or nickname/alias
Age	Range e.g. 30-40 years old
Gender	Male, female or unknown
Ethnic Appearance	Six categorical labels including black, Asian and middle eastern
Height	Range in either metric or imperial
Eye Colour	Labels including mixed
Handedness	Left, right or ambidextrous
Build	Stocky, medium or thin
Shoe Size	British or European size
Nationality	3 digit country code
Hair Type	Labels including receding and shoulder length
Facial Hair Type	Beard, moustache, sideburns or clean shaven
Hair Colour	13 labels, including whether the hair is dyed
Hair Features	Additional features of the subject's hair e.g. pony tail
Marks/Scars/Tattoos	Code recording type of mark and location
Characteristics	Cannot be searched

resentations 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 [16]. These composites were created based on descriptions of the suspect's individual facial features. Research into their effectiveness highlighted two problems. Faces are generally remembered as holistic representations, using descriptions of individual features is not an ideal form of description. Secondly, it has been shown 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, solving both problems experienced with previous composite approaches. EvoFIT [17] is a popular software package which has been successfully exploited by UK police forces [18]. Initially the eyewitness is presented with a grid of 18 random computer generated faces. The eyewitness is required to click on the face which most resembles the suspect. Evolutionary algorithms create a new selection of faces using mutation and recombination based on the face chosen by the user. Additional manual tweaks can be performed by the user if required. A gradual convergence towards the suspect's appearance is achieved over many iterations of user feedback.

It can be seen that human descriptions of soft biometric traits still play a large role in law enforcement. The QUEST system allows the names database to be searched using a human description, but is limited by the amount of features available and the inaccuracies associated with labels and estimations. Modern facial composite systems, like EvoFIT, allow accurate composites to be created based on an eyewitness' memory. However, these composites cannot be used to automatically search the vast amount of mugshots available in the PNC.

2.2 Soft Biometrics

Traditional biometric techniques identify people using distinct physical or behavioural features. These features are clearly discriminative although they can rarely be described using linguistic labels. This restricts identification to situations in which the subject's biometric signature can be obtained and only permits identification of those subjects whose biometric signature has previously been recorded. Soft biometrics are a new form of biometric identification which concerns traits that people naturally use to describe each other. Although each trait can have reduced discriminative capability, they can be combined for identification [13, 19] and fusion with traditional 'hard' biometrics [20, 21].

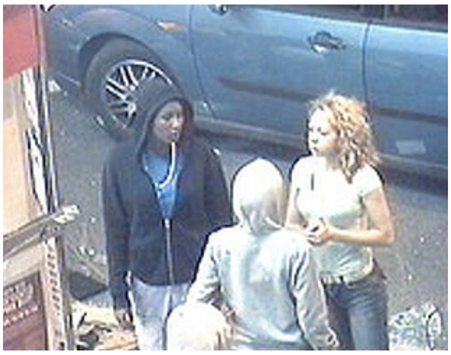


FIGURE 2.1: Surveillance frame displaying common surveillance problems[1]

Though face and gait are the only practical biometrics at a distance, in surveillance scenarios they can suffer from low frame rate and/or resolution. Figure 2.1 shows an example of a typical CCTV video frame. It can be observed that although the picture is at low resolution a detailed human description of the subjects can still be given. In comparison, automatic facial recognition would struggle with the low resolution and non-frontal viewpoint. Soft biometric traits can be obtained from the data derived from low quality sensors, including surveillance cameras. Soft biometrics also require no cooperation from the subject and are non-invasive - making them ideal in surveillance applications.

One of the main advantages of soft biometric traits is their relationship with conventional human descriptions [22]; humans naturally use soft biometric traits to identify and describe each other. Humans are unable to provide detailed descriptions of traditional biometric features resulting in a semantic gap between how machines and people recognize humans. Soft biometrics bridge this gap, allowing conversions between human descriptions and biometrics. Very often, in eyewitness reports, a physical description

[1] Metropolitan Police Flickr Account

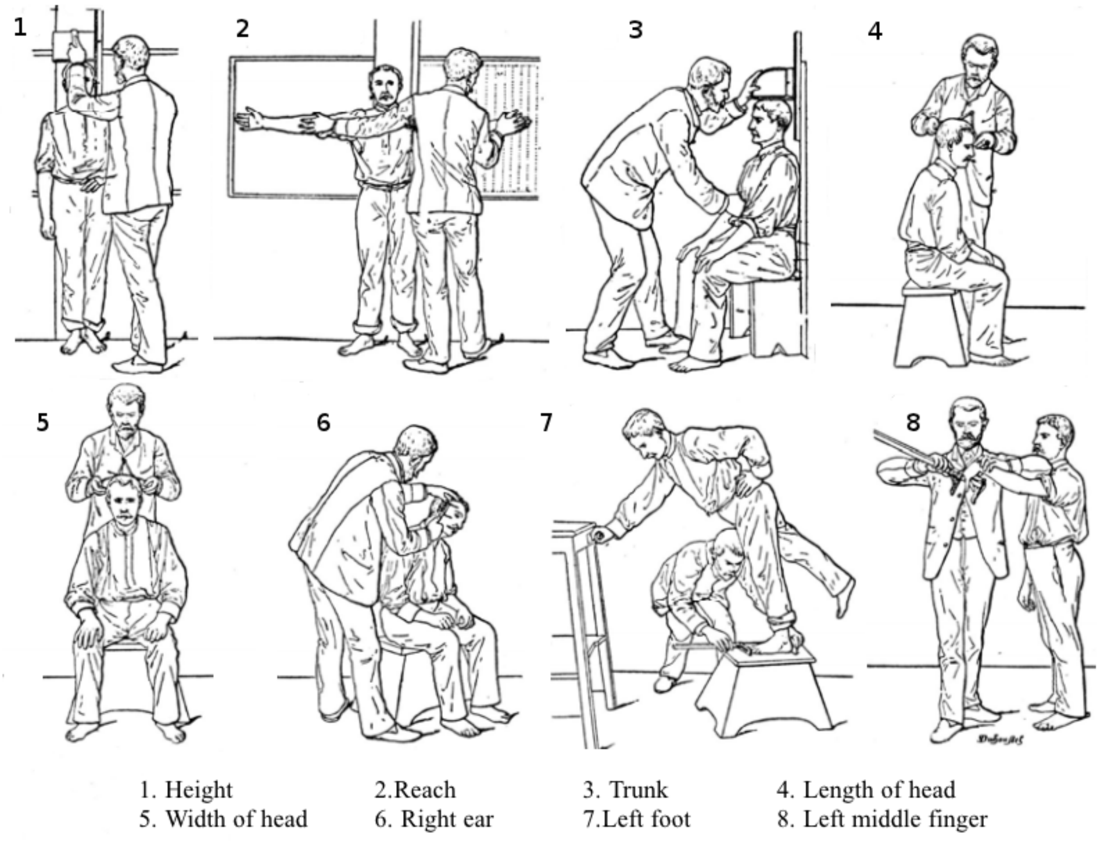


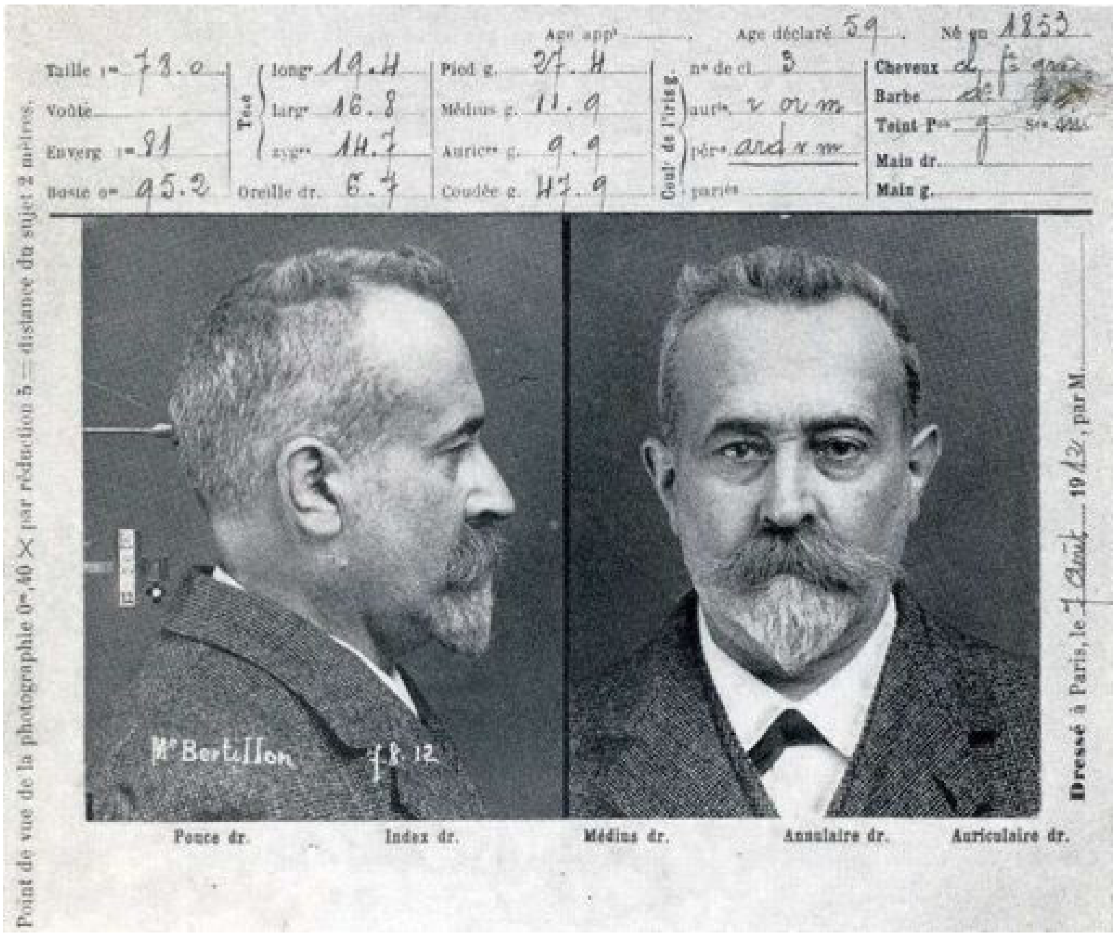
Figure 2.2: Techniques for obtaining accurate bodily measurements [24]

of a suspect may be available. By converting this description to a soft biometric feature vector, biometric databases and possibly surveillance footage could be searched automatically.

2.2.1 Forensic Anthropometry

The field of anthropometry refers to the measurement of the human body. The first use of anthropometrics as a form of identification was introduced in 1883 by Alphonse Bertillon [23] to identify repeat criminal offenders. Prior to 1832 it was legal to identify repeat offenders by clipping their ears or branding them - this procedure was abolished leaving the police system with no systematic re-identification method. The criminal records at the time contained a photograph and a vague description of the person. Problems arose when attempting to identify offenders. The photographs could only be indexed by the individual's name and this could be easily falsified, often resulting in a time extensive search of hundreds of photographs. Likewise, the physical descriptions recorded were subjective and did not enforce a limited vocabulary, making identification difficult especially when a person's appearance could be changed so easily.

The Bertillonage system was introduced to allow identification of repeat offenders using



Age app^t ...... Age déclaré 59. Né en 1853

Taille 1m 73.0
Voûte
Enverg 1m 81
Buste 0m 95.2

Tête long^r 19.4
larg^r 16.8
zyg^re 14.7
Oreille dr. 6.7

Pied g. 27.4
Médius g. 11.9
Auric^re g. 9.9
Coudée g. 47.9

Coul. de l'iris g. n° de cl. 3
auré^e r ov m
pér^ie ard r m
partic^tés

Cheveux [illegible]
Barbe [illegible]
Teint P^on g S^on [illegible]
Main dr.
Main g.

Mr Bertillon 1.8.12

Pouce dr. Index dr. Médius dr. Annulaire dr. Auriculaire dr.

Point de vue de la photographie 0m,40 × par réduction 5 — distance du sujet 2 mètres.

Dressé à Paris, le 1 août 1912, par M.

FIGURE 2.3: A Bertillonage identity card showing Alphonse Bertillon [26]

records indexed by ten physical measurements: height, stretch (left shoulder to middle finger of raised right arm), bust (torso from head to seat when seated), head length (crown to forehead) and width (temple to temple), width of cheeks and the length of the right ear, left foot, middle finger and cubit (elbow to tip of middle finger). Each distance was chosen to be simple to measure by selecting easily identified features for the start and end points. This enabled trained individuals to obtain accurate measurements. The process for obtaining each measurement was meticulously detailed within Bertillon's manual [23] and a sample of the various procedures can be seen in figure 2.2. Additional descriptions including skin, hair and eye colour, facial feature shapes, clothing, race, voice, language and any marks, tattoos or scars were also recorded to confirm the identity of the individual [25]. This was known as the 'spoken portrait' and was recorded using a standardized shorthand. The measurements, descriptions and a standardized photograph of the individual (now known as a 'mug shot') was recorded on a card, an example can be seen in figure 2.3. The cards were indexed in drawers each representing a specific range of the 10 metrics. This allowed hundreds of records to be quickly searched based on a set of measurements.

Although successful the system had many problems. Practitioners required rigorous training and often measurements varied between trained technicians due to slight differences in measurement technique. The tools used in the measurement process needed frequent recalibration and maintenance which required skill and was time extensive. It was also found that measurements changed as an individual aged. Furthermore, it was shown that Bertillon measurements could not discriminate between twins (epitomized by the famous West vs. West case [27]). Due to these issues the system was replaced with fingerprint analysis which could be reliably recorded and could be collected at crime scenes. Although the Bertillonage system was replaced, it represented the first systematic biometric system used for identification in forensic applications, leading the way for modern day biometrics and identification methods. Furthermore, the system utilized soft biometric traits to allow identification, therefore representing the first biometric system which utilized soft biometric features.

2.2.2 Incorporating Soft Biometrics in a Fusion Framework

Primary biometric traits such as face, fingerprints and iris can suffer from noisy sensor data, non-universality and lack of distinctiveness. Further, in certain applications, these traits may fail to achieve high recognition rates. Multi-modal biometric systems [28] can solve these problems by combining multiple biometric traits, resulting in a biometric signature that is robust and more distinctive. Multi-modal systems offer improved performance, but the time taken to verify users can drastically increase thereby causing inconvenience to the subjects and reducing the throughput of the system. Soft biometric traits have been investigated to solve this problem [1].

Jain et al. [29, 20, 1] experimented with the integration of soft biometrics in a biometric system. The primary biometric system compares the input biometric signature obtained from a user against each subject in the database. The secondary soft biometric system uses one or more soft traits to confirm the output of the primary biometric system. The authors used height, gender and ethnicity for this purpose. Gender and ethnicity were automatically obtained from facial images using the technique discussed in [30]. The height data was not available within the test data and, hence, a random height was assigned to each user. The soft biometric feature vector updates the probability that each subject within the database is the same individual as the user.

Experiments were performed on a 263-subject database using multi-modal and uni-modal primary biometric systems. The authors first considered the fusion of a fingerprint-based uni-modal biometric system with a single soft biometric trait (one of height, gender and ethnicity). It was observed that fusion resulted in improved accuracy compared to the fingerprint system. Height was seen to be more discriminative compared to gender and ethnicity, leading to a 2.5% increase in rank-1 retrieval accuracy - although this could be a result of the random generation of heights. Fusing all three soft biometric traits with fingerprints resulted in a 5% increase in rank-1 accuracy compared to using fingerprints alone. Finally, soft biometrics were used to improve a multi-modal system featuring face and fingerprints. An improvement in rank 1 accuracy of 8% (over individual modalities) was observed when combining gender, height and ethnicity information.

Jain et al.'s system showed the advantages of using soft biometric fusion in the context of uni-modal and multi-modal biometric systems. Increasing the number of soft and primary biometric traits increases the uniqueness of a user's signature, leading to better discrimination between subjects. [31] obtained similar success using body weight and fat measurements to improve fingerprint recognition, reducing the total error rate of 62 test subjects from 3.9% to 1.5%.

One concern, however, is the need for an automated technique to weight the soft biometric traits [29]. Marcialis et al. [32] observed that certain soft biometric traits are only useful for a limited set of users. Their work only used soft biometric fusion when the user exhibited an uncommon soft trait thereby bypassing difficulties involved in weighting individual traits. It was assumed that the uncommon soft biometric feature could help in identifying a user from a set of possible candidate identities retrieved using primary biometric traits. An experiment fusing facial biometric signatures with ethnicity and hair colour was developed to verify this assumption. When fusing face and hair colour (when uncommon), the equal error rate (EER) on a database of 100 subjects fell from 6% to 4.5%. This paper clearly detailed the importance of uncommon traits and their ability to identify people. However, the use of hair colour limits this technique to small databases and opens itself to spoof attacks.

The idea of utilizing uncommon traits was extended by [21, 33] to identify people using facial marks. These marks include features such as scars, moles, freckles, acne and wrinkles. The system proposed by the authors utilized facial marks, ethnicity and gender to improve uni-modal face recognition. One of the major advantages of facial marks is their utility (compared to automated facial matching) in courts of law since they are more descriptive and human understandable. In [21], marks were characterized as salient localized regions on the face. Blob detectors based on the Laplacian of Gaussian were used to detect such regions. A commercial facial recognition system's EER was reduced from 3.85% to 3.83% using facial marks. While this is a small reduction in EER, it demonstrates that the addition of soft biometrics can improve highly discriminative hard biometrics. Facial marks are especially beneficial when dealing with occluded or off-frontal face images. In their work, the authors artificially generated several examples of occluded face images, all of which were not recognized by the commercial facial recognition system. Upon using facial marks, the identities of subjects were correctly retrieved on average at rank 6. This demonstrates the benefit of utilizing uncommon traits and marks for human recognition in operational scenarios.

Thus, soft biometric fusion, when appropriately designed, can improve the accuracy of primary biometric systems with minimal inconvenience to the user. Soft traits can be used to either confirm results obtained from a classical biometric system or reduce the search space by filtering large databases. Soft biometric fusion is well suited for incorporation in security applications where speed and convenience are important. Further, in forensic applications, soft biometrics may help in confirming the identity of a subject.

2.2.3 Human Identification Using Soft Biometrics

As stated in the introduction, soft biometric traits were originally defined as features which lack the distinctiveness and permanence to accurately identify a person. This definition remains true when dealing with single traits, but has been shown to be partially overcome when dealing with multiple soft biometric traits. Dantcheva et al. [34] 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. However, by agglomerating many such features a reasonably unique signature can be constructed. Soft traits have some advantages compared to classical uni-modal and multi-modal systems.

One advantage of soft biometric systems is the bridging of the semantic gap between biometric traits and human descriptions. Soft biometric traits use human understandable descriptions (for example height, hair colour and gender) and as a result can be naturally searched and understood. This also negates the requirement of obtaining biometric data of subjects before identification, allowing previously unencountered subjects to be identified using human descriptions. This presents exciting possibilities such as searching surveillance footage and databases based solely on an eyewitness' description.

The two most popular traits for identification-at-a-distance are face [35] and gait [36]. These can suffer from the poor sensor quality of most CCTV cameras. Low resolution can seriously impair facial recognition, and low frame rates (sometimes even time-lapse cameras) obscure the motion of the human body required for gait recognition. In contrast, soft traits can often be obtained from very poor quality video or images. This has huge potential for immediate real-world use without upgrading the vast surveillance infrastructure.

Ailisto et al. [19] presented a soft biometric system aimed at addressing concerns of privacy, identity theft and the obtrusive nature of previous biometric solutions. Their system used unobtrusive and privacy preserving soft traits, including height, weight and body fat percentage. The system had applications in low-risk convenience scenarios where a relatively small number of people required identification, such as homes, small offices and health clubs. Height, weight and body fat were obtained from 62 subjects to mimic the target application environment. Single modalities were shown to be very weak, with weight being the most distinctive resulting in a 11.4% total error rate (total false accepts and rejects). A combination of weight and height resulted in a 2.4% total error rate and the rank-5 retrieval accuracy was 100%. Using just three easy-to-obtain soft features allowed a database of 62 subjects to be sufficiently differentiated for the target application.

Sridharan et al. [37] proposed a facial image retrieval system that is queried using verbal descriptions. Queries can include up to 14 defined features, composed of 5 Boolean descriptors (e.g. presence of beard) and 9 categorical labels (e.g. nose width, face length, hair colour). The soft biometric database is automatically generated from frontal facial images. This is achieved using feature localization followed by parametrization of the various facial features. The continuous traits are discretized to 3 labels using predefined thresholds, for example nose length can be described as short, medium and long. A Bayesian approach determines the probability that a facial image matches the provided description, allowing the facial images to be ordered based on their similarity to the query. 25 users were asked to describe a subject from a 125 subject database. The average number of feature descriptions required to achieve a rank 5 retrieval was 6.6 out of the possible 14. This result shows that facial features are very discriminative when described accurately.

Table 2.2: Semantic traits and corresponding terms

Trait	Terms
Arm Length	[Very Short, Short, Average, Long, Very Long]
Arm Thickness	[Very Thin, Thin, Average, Thick, Very Thick]
Chest	[Very Slim, Slim, Average, Large, Very Large]
Figure	[Very Small, Small, Average, Large, Very Large]
Height	[Very Short, Short, Average, Tall, Very Tall]
Hips	[Very Narrow, Narrow, Average, Broad, Very Broad]
Leg Length	[Very Short, Short, Average, Long, Very Long]
Leg Shape	[Very Straight, Straight, Average, Bow, Very Bowed]
Leg Thickness	[Very Thin, Thin, Average, Thick, Very Thick]
Muscle Build	[Very Lean, Lean, Average, Muscly, Very Muscly]
Proportions	[Average, Unusual]
Shoulder Shape	[Very Square, Square, Average, Rounded, Very Rounded]
Weight	[Very Thin, Thin, Average, Fat, Very Fat]
Age	[Infant, Pre-Adolescence, Adolescence, Young Adult, Adult, Middle Aged, Senior]
Ethnicity	[Other, European, Middle Eastern, Far Eastern, Black, Mixed]
Sex	[Female, Male]
Skin Colour	[White, Tanned, Oriental, Black]
Facial Hair Colour	[None, Black, Brown, Blond, Red, Grey]
Facial Hair Length	[None, Stubble, Moustache, Goatee, Full Beard]
Hair Colour	[Black, Brown, Blond, Grey, Red, Dyed]
Hair Length	[None, Shaven, Short, Medium, Long]
Neck Length	[Very Short, Short, Average, Long, Very Long]
Neck Thickness	[Very Thin,Thin,Average,Thick,Very Thick]

Samangooei and Nixon [13] developed a soft biometric system which identifies subjects

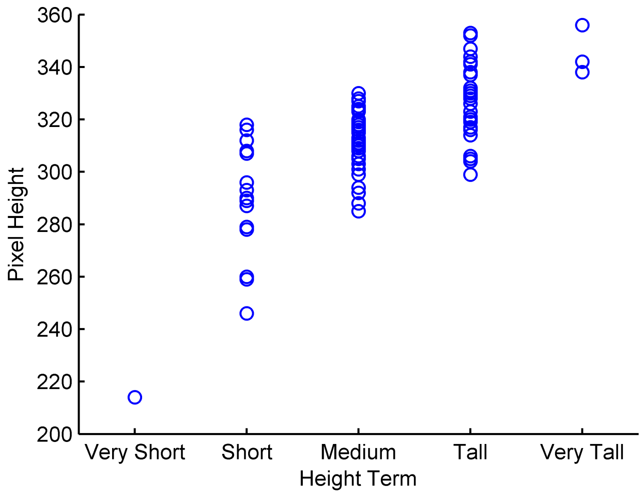


FIGURE 2.4: The relationship between pixel height and absolute labels provided by annotators

from video footage (Soton gait database [38]) based solely on a verbal human description. This description was composed of 23 absolute categorical labels (table 2.2) which were chosen to be universal, distinct, easily discernible at a distance and largely permanent. The selected soft biometric traits featured both intrinsically categorical attributes, like hair colour, and characteristics generally associated with value metrics, like height - both were described using absolute labels.

Initially 959 descriptions of the 115 subjects from the Soton gait database were obtained and used to build a database of soft biometric feature vectors which described the given descriptions. Initial analysis of the descriptions showed that the categorical labels used to describe the subjects were unreliable, especially when describing traits generally associated with value metrics. Figure 2.4 shows the relationship between the height of the subjects (obtained from the video footage and represented in pixels) and the median absolute height label used to describe the subjects (on average each subject was described by 8 individual annotators). Large overlaps between the short, medium and tall labels were observed resulting in a statistically significant ($p < 0.0001$) Pearson's correlation of 0.71. This incorrectness between actual and labelled height is due to the categorical nature and subjectiveness of the labels.

Latent semantic analysis (LSA) [39], which is extensively used in document analysis, was employed to learn the structure between the soft biometrics and gait signatures which were obtained from video footage. By learning the relationships between the visual gait signature and the soft biometric features, the technique can be used to automatically label people based on their physical characteristics - thus converting gait signatures to human descriptions automatically. The results from this approach were modestly successful showing an accuracy of 68% when determining semantic labels automatically from gait signatures [40].

Table 2.3: F-ratio of several soft biometric traits based on the Soton gait database [41]

Trait	F-ratio	Trait	F-ratio
Sex	383.70	Neck Thickness	14.73
Skin Colour	149.44	Arm Thickness	13.90
Ethnicity	96.10	Leg Length	13.68
Hair Length	79.05	Muscle Build	12.85
Age	57.02	Leg Thickness	11.61
Hair Colour	52.18	Hips	10.55
Facial Hair Length	25.72	Arm Length	5.74
Height	25.14	Facial Hair Colour	5.61
Weight	20.75	Leg Shape	3.25
Figure	20.69	Proportions	2.77
Chest	18.32	Shoulder Shape	2.54
Neck Length	15.57		

An interesting statistical analysis of these soft biometric traits was presented in [41]. Each trait used to describe a person should be meaningful and provide additional information which differentiates the person from others. This property can be tested by determining the trait's ability to significantly separate the subjects within the database. If the subjects are not separated by a trait, then it could be said that the trait lacks any discriminative power and is not beneficial to the description (for the given set of subjects). To assess the discriminative power of each trait individually, one-way ANOVA (analysis of variance) was used to generate a statistic called the F-ratio:

$$\text{F-ratio} = \frac{\text{total between-group variance}}{\text{total within-group variance}} \tag{2.1}$$

$$= \frac{\sum_i n_i(\bar{X}_i - \bar{X})^2/(K-1)}{\sum_{ij}(X_{ij} - \bar{X}_i)^2/(N-K)}. \tag{2.2}$$

Here, X_{ij} represents the j^{th} observation of the soft biometric of the i^{th} user and n_i denotes the number of observations of the i^{th} subject. $\bar{X}_i$ is the mean of the i^{th} user's observations and $\bar{X}$ is the mean across all subjects' observations. K represents the number of subjects while N represents the number of traits. Table 2.3 shows each trait's F-ratio, where a higher F-ratio indicates traits which are more successful at separating individuals.

It can be observed that "global" traits like gender, ethnicity and skin colour have more discriminative power than physical traits, like leg thickness. This is most likely due to the difficulty of labelling continuous physical traits compared to the categorical nature of the global traits. Traits like shoulder shape, proportions and leg shape have been shown to be non-discriminative thereby revealing their inability to distinguish between users. This important statistical analysis identifies the significance of each trait within a description and can be used to remove traits that do not contribute to additional information.

The database of soft biometric feature vectors was analysed to assess the discriminatory power of the descriptions. Recognition experiments were conducted by retrieving subjects from the database to assess the uniqueness of each subject's soft biometric feature vector and the variance between multiple descriptions of the same subject. Each subject's feature vector consisted of the most commonly used label to describe a subject's soft biometric trait (on average each subject was described by 8 individual annotators). Each possible label was represented by a boolean value within the feature vector. If the label was assigned to the subject the corresponding boolean value was set to true. A leave-one-out validation approach was used to evaluate recognition performance. The probe, which was used to query the database, was formed from a single verbal description of the subject given by a single annotator. The mode of the remaining descriptions of the subject were used as the gallery, the feature vector within the database being searched. The feature vectors within the database were ordered based on their similarity with the probe feature vector. The Hamming distance metric was used to assess the similarity between two feature vectors. The position of the probe subject's gallery feature vector within the ordered list represents the retrieval performance of the system.

Figure 2.5 shows the results. The rank 1 retrieval performance (i.e. the recognition accuracy) was found to be 48%. Retrieval performance increased to 90% at rank 15. Subject interference [34] is a known problem when using labels and occurs when two subjects are indistinguishable from each other due to the limited number of labels available. This obviously has a drastic effect when attempting to identify a subject and would explain the poor recognition results. This highlights the lack of distinctiveness between subjects due to the limited information conveyed using categorical labels. As such, absolute labels can be used to recognize people but are limited in accuracy leading to a limited recognition capability.

A statistical analysis of soft biometric systems utilizing categorical descriptions of physical traits was performed in [34, 42] to determine the reliability of such a system in larger operational settings. When using categorical labels, it is important to consider the likelihood of a subject being indistinguishable from other subjects in the database: this is referred to as inter-subject interference [34]. Obviously the interference has a huge impact on the soft biometric system's performance and the number of traits recorded directly affects the probability of interference between subjects. The system developed within the project identified nine semantic traits, mainly focusing on facial soft biometrics. These include: the presence of a beard, moustache and glasses, each containing two terms; the colour of the skin, eye and hair composed of three, six and eight terms, respectively; body mass index consisting of four terms defined by population norms. Further, the colour of clothing on the torso and legs were determined, each being labelled based on a set of eleven terms.

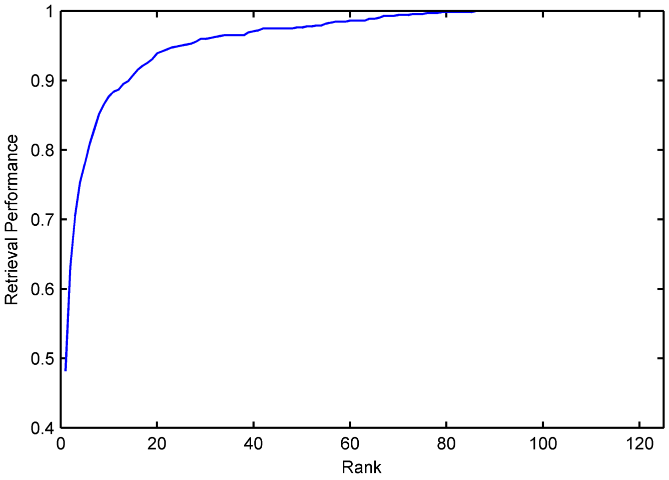


FIGURE 2.5: Retrieval accuracy of absolute descriptions from a soft biometric database

Figure 2.6 shows the likelihood of interference occurring with N subjects where N ranges from 0 to 1000 subjects. The figure shows the probability of interference, $p(N)$, within a database of subjects and the probability of a randomly chosen subject from the database interfering with another subject(s), $q(N)$. Figure 2.6 clearly shows that with only 49 people a 50% chance of interference exists. This likelihood of interference can be reduced by increasing the uniqueness of each subject's trait signature. Increasing the amount of possible combinations of terms is one possible method for achieving this - only if the new term combinations further discriminate between the subjects. This can be achieved by either increasing the amount of traits or the detection of more terms per trait. In comparison, Samangooei et al. [22]'s soft biometric system, featuring 23 traits, has 3.7x10^{15} possible combinations of semantic terms - potentially decreasing the likelihood of interference. This important work clearly identified the need for maximizing the amount of term combinations and its effects on interference and ultimately the performance of the soft biometric system. Further statistical studies are required to identify the optimal number of term combinations for target application environments, taking into account the expected distributions across different soft traits.

2.2.3.1 Imputation

Human physical traits and appearance inherently contain structure, features frequently co-occur or have fixed relationships with other features. This occurs either due to social aspects (long hair is common on females), genetics (black hair is common within people of Asian descent) or the morphology of the human body (taller people are more likely to have longer legs). Imputation techniques are a statistical approach used to predict missing variables. Using such techniques missing soft biometric features can be predicted utilizing the structure within human appearance. This structure offers a basis to improve the recognition of soft biometric traits and to make soft biometric systems more robust

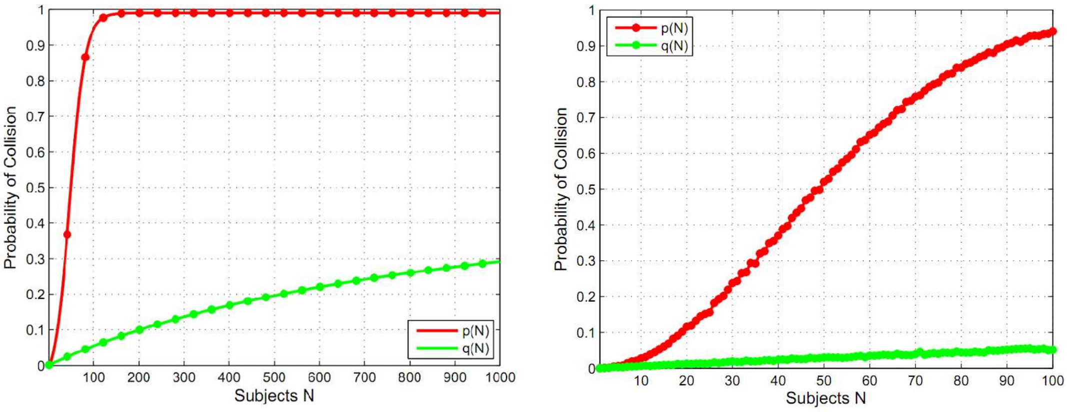


FIGURE 2.6: Interference probability in a N sized population ranging from 0 - 1000, and a magnified version showing 0 - 100 [34]

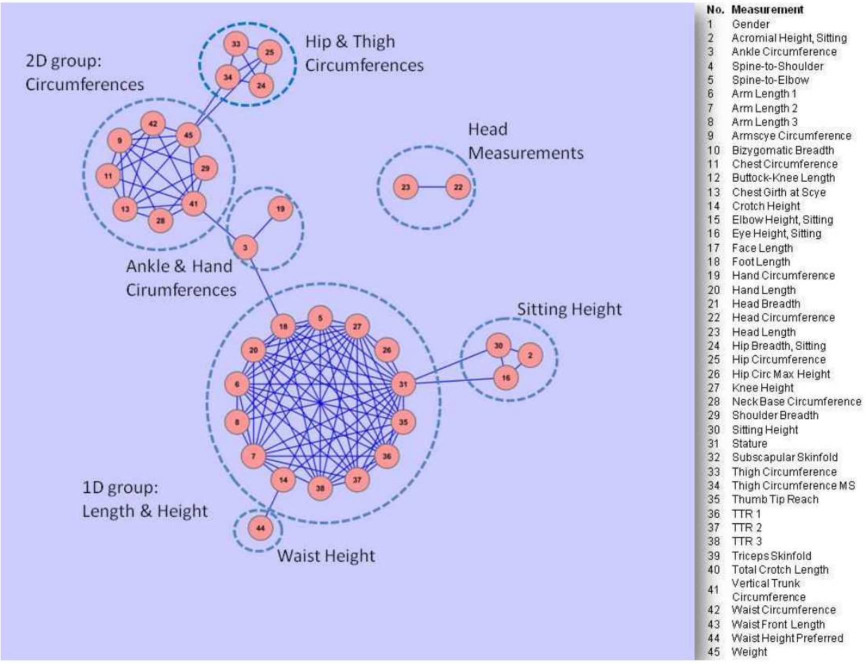
to missing traits or occluded visual features.

Adjeroh et al. [43] studied correlation and imputation in human appearance analysis. Data was gathered from the CAESAR anthropometric dataset which comprised of 45 continuous physical measurements for 2369 subjects. The relationships between the human measurements was first assessed using the Pearson correlation coefficient. To visualize the correlation, a correlation graph was created - shown in figure 2.7. This graph shows connections between traits if the correlation was stronger than a threshold value. This clearly confirms the structure within human appearance and highlights clusters of traits with strong correlation. It can be observed that the measurements generally fall into two groups, both of which have physical meaning: the 2D group which contains circumferences of body parts and the 1D group containing lengths and heights. These clusters suggest that only a few measurements would have to be known to predict the majority of the other traits.

The metrology predictability network was developed to predict missing traits based on the most suitable subset of observed traits. Correlations between the missing and present traits were initially used to define a suitable subset. Using the correlation graph any nodes linked to the missing node are used in the prediction process. Traits which have been shown to accurately predict the missing trait are also considered. The expected error is assessed using multiple linear regression on training data from the CAESAR dataset.

31 prediction models were constructed, each varying the order of the model, the number of variables and the variable combinations. Using a training set of measurements, the error predicting a trait using a single prediction model is assessed. These errors are used to create a predictability graph (similar to figure 2.7) denoting the ability of a trait to predict another trait accurately, where edges denote errors which are below a threshold.

The measurements obtained from the subject, called the seed measurements, are used


Hip & Thigh Circumferences
2D group: Circumferences
Head Measurements
Ankle & Hand Cirumferences
Sitting Height
1D group: Length & Height
Waist Height

No.	Measurement
1	Gender
2	Acromial Height, Sitting
3	Ankle Circumference
4	Spine-to-Shoulder
5	Spine-to-Elbow
6	Arm Length 1
7	Arm Length 2
8	Arm Length 3
9	Armscye Circumference
10	Bizygomatic Breadth
11	Chest Circumference
12	Buttock-Knee Length
13	Chest Girth at Scye
14	Crotch Height
15	Elbow Height, Sitting
16	Eye Height, Sitting
17	Face Length
18	Foot Length
19	Hand Circumference
20	Hand Length
21	Head Breadth
22	Head Circumference
23	Head Length
24	Hip Breadth, Sitting
25	Hip Circumference
26	Hip Circ Max Height
27	Knee Height
28	Neck Base Circumference
29	Shoulder Breadth
30	Sitting Height
31	Stature
32	Subscapular Skinfold
33	Thigh Circumference
34	Thigh Circumference MS
35	Thumb Tip Reach
36	TTR 1
37	TTR 2
38	TTR 3
39	Triceps Skinfold
40	Total Crotch Length
41	Vertical Trunk Circumference
42	Waist Circumference
43	Waist Front Length
44	Waist Height Preferred
45	Weight

Figure 2.7: The relations between measurements based on correlation (>0.81) [43]

to predict the missing traits. Some traits will be easy to predict due to their strong relationship with the seeds. These highly correlated traits are also used to predict traits with a weak relationship with the seeds. Principal component regression is used to predict the missing traits. An initial prediction is made using regression on the seed measurements, principal component analysis is used to reduce the measurements needed and then regression is applied to these features (figure 2.8). Experiments were conducted on the CAESAR dataset and 23 subjects from the CMU motion capture database. 4 seed measurements were used - arm length, knee height, shoulder breadth and standing height. Based on these seeds the remaining traits were predicted with an average mean absolute error of 0.041. Another experiment predicted all 41 measurements from just 3 seeds and used these measurements to predict the gender, resulting in a 88.9% correct gender classification rate.

It has been shown that human appearance contains an inherent structure and just a few seed measurements are required to accurately predict the remaining features. This redundancy is vital when dealing with occlusion in visual data.

Structure is inherent within human appearance and is echoed in human descriptions and biometric representations. Imputation is crucial in operational settings where visual data is often occluded and human descriptions often erroneous or incomplete. By utilizing structure within the soft traits, issues with view invariance and the subjective and unreliable nature of human descriptions can be addressed.

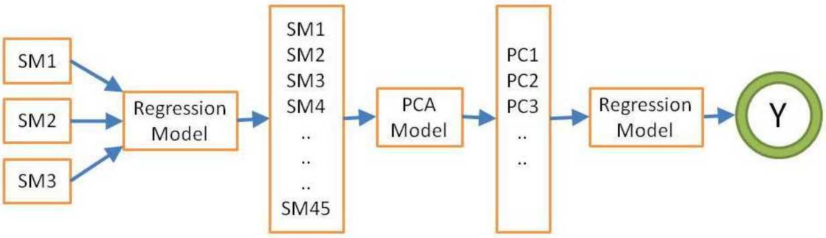


Figure 2.8: Two step prediction, using 3 seed measurements [43]

2.2.4 Applications

2.2.4.1 Continuous Authentication

Most existing computers only authenticate users at the beginning of a session, leaving the system open to imposters until the user logs out. Continuous user authentication provides a method to continually confirm the identity of the user. Conventional biometric modalities such as face and fingerprint are either inconvenient for continuous operation or difficult to capture when the user is not explicitly interacting with the sensor. Soft biometrics offers a potential solution to this problem [44] by using features like the colour of the user's clothes and facial skin.

When the user is initially authenticated using facial recognition and a password, soft biometric traits are obtained and recorded. Throughout the session the user is authenticated using these traits, without enforcing a strict posture or requiring constant verification. Facial recognition is also used periodically, when the biometric data is available, to guard against spoof attacks. Histograms of the various colours are gathered and the Bhattacharyya coefficient [45] is used to calculate the similarity of two histograms, by measuring the amount of overlap. In one experiment, a database of 20 subjects was constructed. Each subject was asked to perform 6 actions including turning their heads, leaning back in their chair, stretching arms and walking away from the computer. The average false rejection and acceptance over all the recorded actions were 4.16% and 0%, respectively. Soft biometrics has been shown to provide secure continuous user authentication whilst being robust to the user's posture and not requiring manual registration of the soft biometric traits for each session.

2.2.4.2 Surveillance and Re-Identification

CCTV cameras have been widely introduced and accepted [46, 47]. Their primary role within society is to assist in the fight against crime [46]. This involves deterring and detecting crime, reducing the fear of crime and to provide evidence when crime does occur. There has been considerable investment into the CCTV infrastructure (particularly in the UK) but currently these cameras (and the ensuing footage) are still generally monitored only by human operators. Due to the number of cameras within most cities, operators cannot monitor the data in intricate detail. This means looking for a single person can be time consuming and prone to mistakes. Soft biometrics can potentially solve these problems by providing a method for searching surveillance footage using human descriptions.

Soft biometrics offers several benefits over other forms of identification-from-a-distance. Face recognition often requires good resolution images and gait recognition requires good frame-rates. In comparison, certain soft biometric traits can be obtained from low resolution and low frame-rate videos, and from an arbitrary viewpoint of the subject. The human compliant nature of soft biometric traits can also be exploited to allow searches based solely on a human description - possibly obtained from an eye witness. This allows for the use of soft biometrics when primary biometric identifiers cannot be obtained or when only a description of the person is available.

Denman et al. [48] used soft biometric traits to identify people using previous observations or human descriptions when traditional biometrics are unavailable. The height and colour of the torso, legs, and head are used to model subjects. Identifying these three body components is done by first locating the person using background segmentation and then analysing the colour of moving pixels in each row. Large colour differences can often be found between the head, torso and legs due to clothing that can be easily identified by examining colour gradients. Average body proportions were used to identify the most likely colour gradients representing the three desired regions. After the regions are located, a colour histogram is recorded and the real world height estimated. Heights are matched using average height and standard deviations, and colour histograms are matched using the Bhattacharyya coefficient. The PETS 2006 surveillance database was used to test the system. This dataset features four cameras monitoring a train station: four recordings of 25 people were obtained. The system achieved an equal error rate of 6.1% when evaluated using the leave-one-out cross-validation scheme. These recordings included videos from two different viewpoints, demonstrating the view invariant nature of the selected soft traits. In comparison, primary biometric traits such as face, typically only work from one viewpoint. Similar studies show successful retrieval results using facial features [49] and clothing colour [50].

Further work in soft biometrics has provided a technique to recognize subjects moving between multiple surveillance cameras in order to generate a rough framework for facial recognition [51]. The technique uses gender, ethnicity and session-based soft biometrics (skin colour, upper and lower body clothing colour and hair colour). Session-based soft biometrics are features which are reasonably constant for a short time period. These features, although not permanent, allow subjects to be identified when moving between different cameras. Once a person has been identified in the surveillance footage, the directional pose is determined. If the person is walking towards the camera, the face is analysed to deduce ethnicity and gender, which is combined with the colour-based traits that are extracted automatically. When a camera observes a new subject, their session-based features are compared to that of people previously observed by the camera network. If a match is found, the subject is given the same identity tag.

A custom low-resolution surveillance dataset was constructed featuring 100 subjects. An average correct classification rate of 60% and 83%, for gender and ethnicity, respectively, was observed using just a resolution of 66x61 (pixels) facial images obtained from the video dataset. Gender and ethnicity were also used to partition the database of observed faces to speed up queries. The gender and ethnicity of the facial query were obtained and only faces featuring the same soft traits within the database were tested. The soft biometric partitioning reduced the time required for face recognition queries by almost a factor of 6 on a 600 subject database. Session-based soft biometrics are ideal for tracking people between cameras due to the speed in trait acquisition and their view invariant nature. Additional traits would allow for tracking in more crowded areas and would reduce the reliance on colour, which is problematic if the cameras are not calibrated. Additional traits could also be used to partition the database further thereby reducing the time taken for primary biometric queries. Denman et al. [52] exploit soft biometrics to track customers through a multiple camera surveillance network with the aim to observe customer behaviour and dwell times in commercial applications.

2.3 Conclusions

In this chapter we have introduced the field of soft biometrics and explored the accuracy, content and police usage of human descriptions.

Human descriptions are generally made up of categorical labels or continuous estimations - both have advantages and limitations. Categorical labels are easy to use but are typically subjective (especially when describing naturally continuous features like height) and lack detail. Previous soft biometric systems have shown a low correlation between labels and actual measurements and low discriminative capabilities due to the limited range of labels available. Continuous annotations are very descriptive but have been shown to be incorrect 50% of the time (when describing age, height and weight). This has been accredited to the inexperience of the annotators and self anchoring.

These findings have spurred research into more reliable forms of description. The next chapter introduces comparative descriptions which aim to reduce subjectivity and infer a discriminative continuous measurement whilst not requiring continuous estimations.

Chapter 3

Comparative Human Descriptions

In this section we describe a new method for obtaining human descriptions which exploits the process of making visual comparisons between subjects. 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 2.1.2 we discussed the issues with conventional forms of human description. Comparative categorical labels present a solution to these problems:

Obtaining the necessary level of detail to allow identification is problematic with current forms of description:

- Continuous estimations are informative although frequently inaccurate [3, 9] due to the witnesses' lack of training and experience at providing accurate estimations [9].
- Absolute labels require little skill to annotate but due to their categorical nature have less discriminative capabilities (demonstrated in section 2.2.3) and are prone to subject interference [34].

Comparative descriptions exploit categorical labels which are easy to understand and annotate. Furthermore, informative continuous relative measurements can be inferred from *multiple* comparisons, providing the level of detail required for identification. Comparative descriptions can convey accurate and descriptive information whilst avoiding asking the user for continuous estimations.

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.

This chapter will justify the use of relative information and introduce the comparative databases used throughout this research. In section 3.1 we explore other studies which have benefited from relative measurements. An introduction to human comparisons is presented in section 3.2. The traits, method of annotation and evaluation of bodily and facial comparative databases are discussed in sections 3.3 and 3.4 respectively.

3.1 Relative Information

Relative information has recently been explored to improve human descriptions of objects within images. Several techniques have exploited similarities between objects as a form of description. Kumar et al. [53] have explored similarities between faces to identify and explain facial attributes. The developed 'simile classifiers' recognize similarities between a face (or regions of a face) and a set of specific reference subjects. This allows descriptions such as '*lips like Barack Obama*' or '*a nose like Owen Wilson*'. The advantage of this system is the ability to produce descriptions of features which are generally hard to describe. Wang et al. [54] exploits similarities between objects to allow recognition with few or no examples. Descriptions such as '*a zebra is similar to a horse in shape and a crosswalk in texture*', allows the approach to identify a zebra with no training examples. Exploiting descriptions of similarity between objects has been shown to improve recognition of objects within images with few training examples. Both of these techniques utilize relative information to improve descriptions, although they differ significantly from our approach. Similarity between reference subjects or other objects provides a method of description, whereas the comparison of subjects provides an ordering based on the specific trait being compared. Although different, these techniques show the benefits of relative information especially when describing features or attributes which are normally difficult to communicate.

Image descriptions have been further improved by determining order based on the strength of a specific attribute, allowing such comparisons as '*lions are larger than dogs*' [55]. Given a set of images and a partial set of comparisons detailing the relative strength of a certain attribute, the technique determines a complete ordering of the images. This was approached as an optimization problem where the comparisons were treated as constraints. A ranking support vector machine was used to determine a ranking function which fitted a weight vector to maximize the number of constraints satisfied - this was based on ranking algorithms used within search engines [56]. The ranking function could then be used to determine the ordering between all of the images. Zero-shot learning from relationships was introduced based on this ordering approach, allowing previously unseen objects to be identified based on comparisons with observed objects. The zero-shot learning results show that the relative descriptions convey stronger discriminatory power compared to binary descriptions.

3.2 Defining and Evaluating Human Comparisons

In this chapter we will introduce facial and bodily human comparisons as a new form of human description. Before examining the details of bodily and facial comparisons, this section will define human comparisons and discuss how comparisons are evaluated and utilized.

A *human 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 human comparison will describe the differences between the target and subject in terms of individual traits, such as height, weight and nose length. A *trait comparison* is a comparison of an individual soft trait. Each soft biometric trait comparison is represented by a single categorical label taken from a set of five ordered labels, for example ‘much shorter’, ‘shorter’, ‘same’, ‘taller’ and ‘much taller’. Each of the five labels are assigned a value, ranging from -2 to 2, based on their order; such that -2 represents a ‘much less’ comparison (e.g. ‘much shorter’) and +2 a ‘much more’ (e.g. ‘much taller’). A trait comparison, C_{st}, between a target, t, and a subject, s, can be described as follows:

$$C_{s,t} \in \{-2, -1, 0, 1, 2\} \tag{3.1}$$

In sections 3.3.3 and 3.4.3, absolute annotations are used to examine the differences in information provided by comparative and absolute labels. This evaluation is described here for simplicity.

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 were labelled as 'short' and 'tall' respectively and the comparative descriptor provided was 'taller', we would consider both annotations as concurring. The absolute annotations obviously lack detail; two people labelled as 'tall' are unlikely to be exactly the same height. 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. Such that:

$$concurrence(A_t, A_s, C_{s,t}) = \begin{cases} 1 & A_s < A_t \; and \; C_{s,t} < 0 \\ 1 & A_s > A_t \; and \; C_{s,t} > 0 \\ 1 & A_s = A_t \\ 0 & otherwise \end{cases} \tag{3.2}$$

Where A is a value representing an ordered absolute label. The concurrence between absolute and comparative labels is assessed for *each soft trait individually* and considers all the comparisons collected which describe the soft trait. The trait concurrence is expressed as the proportion of comparisons which concur with the absolute labels, such that:

$$trait\,concurrence = \frac{1}{n}\sum_{i=1}^{n} concurrence(F_{t_i}, F_{s_i}, C_{s_i,t_i}) \tag{3.3}$$

where n is the total number of trait comparisons obtained describing a particular trait. The ith trait comparison details the difference (in respect to the particular trait) between a target, t_i, and subject, s_i, such that the ith trait comparison is annotated C_{s_i,t_i}. The most frequently annotated absolute label describing the target and subject are shown as F_{t_i} and F_{s_i} respectively. The mode of the absolute labels was utilized to reduce the subjectivity associated with individual absolute labels.

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. Given two subjects, s_i and s_j, who are both compared against the same target, t, the inferred subject-subject comparison is obtained by finding the difference between the two subject-target comparisons:

$$C_{s_i,s_j} = C_{s_i,t} - C_{s_j,t} \tag{3.4}$$

$$C_{s_j,s_i} = C_{s_j,t} - C_{s_i,t} \tag{3.5}$$

Once the difference has been found it can be scaled between -2 and 2, representing the difference in attribute strength between the two subjects.

Inferring comparisons does introduce errors. If two subjects are both labelled as 'taller' than the target, the inferred comparison would be 'same'. The likelihood is that the subjects are not the same height and we are losing resolution with this assumption. Although inaccurate, this approach allowed us to fully exploit the comparisons we obtained from limited experiments. The subsequent chapters in this thesis will utilize the subject-subject comparisons.

3.3 Body Comparisons

Bodily and global traits, such as height, weight, race and gender, are the most frequently mentioned descriptions in eyewitness reports. This implies that they are memorable and salient. Unfortunately, the methods used to describe these traits are often inaccurate and unreliable. Problems include the subjectivity of absolute labels and the experience required to accurately estimate continuous measurements. Comparative descriptions may offer a solution to these two major problems.

As well as being mentioned frequently, bodily comparisons could also be utilized to search surveillance footage. Bodily and global traits are ideal for surveillance applications due to their saliency and size, allowing trait descriptions to be obtained even from low resolution and low framerate footage.

3.3.1 Traits

We have shown in section 2.1 that descriptions of some traits are more salient and reliable than others. Samangooei and Nixon [13] explored the use of absolute descriptions for soft biometrics. The traits chosen in this study were largely based on MacLeod's work [8] and hence reflect the optimal bodily traits for human description. For this reason they were used in this research also allowing comparisons between the two approaches. Several traits were excluded. The leg shape trait was removed as it was hard to detect the trait from side on video footage. The facial hair traits (colour and length) were only applicable to a few subjects within the database and hence were removed. Finally the proportions trait was excluded due to its low significance and discriminatory capability [41].

A single human comparison consists of 19 traits. 16 of which are trait comparisons (shown in table 3.1), each described using one of five comparative labels. It can be observed that three traits (gender, ethnicity and skin colour) 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 criterion. These absolute annotations are not considered when analysing the comparative annotations and are used only for recognition and retrieval.

Trait	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]

Table 3.1: Soft traits used to compare subjects

3.3.2 Data Acquisition

The method used to obtain descriptions from an observer is an important consideration when exploring a new form of human description. In the case of human comparisons 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:

- Do relative measurements provide more discriminatory information than absolute labels?
- Are the resulting relative measurements highly correlated with the subject's physical attributes?

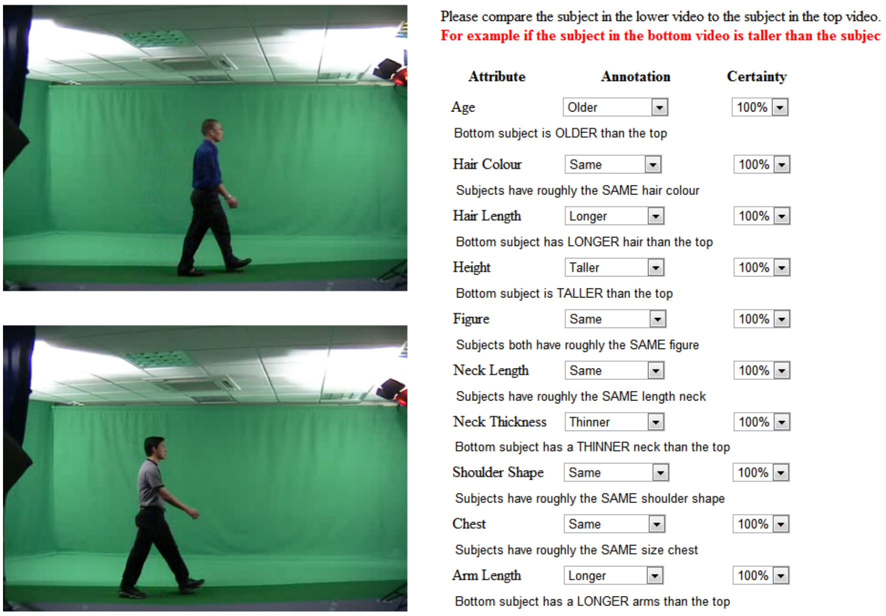
- Is the developed method of obtaining human comparisons practical?
- Are human comparisons more robust against errors originating from subjectiveness?

As mentioned in section 3.2, multiple comparisons must be available to infer a robust description, with each comparison allowing the description of the target to be refined. A practical method of obtaining comparisons, between a *target* subject (representing the suspect in application settings) and multiple *subjects*, is to present videos of the subjects to the annotator. This permits multiple comparisons with minimal equipment and personnel. To validate this approach the experiment will present videos of individuals from the SGDB to the annotator. The gait database includes videos of 100 people walking in a plane normal to the view of the camera, more information can be found in appendix A. Previously absolute categorical labels had been collected for the same database [13] - allowing comparisons between the two forms of description.

The experiment was split into two parts. The first part explored the benefits of comparative annotations in ideal settings and the second investigated the application potential of comparisons. Initially volunteers were asked to compare two subjects whilst both were visible. This removes all problems with memory and validates the effectiveness of comparative descriptions. Five subjects were compared to a single target - this simulates the idea of comparing a selection of subjects against a suspect.

The next part of the experiment tested the application potential of comparative annotations. Memory is a huge problem in eyewitness descriptions [57] 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 user was then asked to compare five subjects with the target. Finally, the user was asked to describe the target using absolute categorical annotations. The results of this stage of the experiment will be discussed fully in chapter 6. Until chapter 6, the comparative descriptions from the first and second parts of this experiment will be combined and treated as a single database due to the small error observed in the delayed comparisons.

The 100 subjects from the SGDB 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 of less importance. Subjects were assigned to users to gather the most comparisons describing different pairs of subjects and targets. Performing comparisons between a large group of subjects and a small group of targets also allowed inference of annotations between subjects as shown in section 3.2.



Please compare the subject in the lower video to the subject in the top video.
For example if the subject in the bottom video is taller than the subjec

Attribute	Annotation	Certainty
Age	Older	100%
Bottom subject is OLDER than the top		
Hair Colour	Same	100%
Subjects have roughly the SAME hair colour		
Hair Length	Longer	100%
Bottom subject has LONGER hair than the top		
Height	Taller	100%
Bottom subject is TALLER than the top		
Figure	Same	100%
Subjects both have roughly the SAME figure		
Neck Length	Same	100%
Subjects have roughly the SAME length neck		
Neck Thickness	Thinner	100%
Bottom subject has a THINNER neck than the top		
Shoulder Shape	Same	100%
Subjects have roughly the SAME shoulder shape		
Chest	Same	100%
Subjects have roughly the SAME size chest		
Arm Length	Longer	100%
Bottom subject has a LONGER arms than the top		

FIGURE 3.1: Bodily comparative label collection

Comparisons were gathered using the website shown in figure 3.1. The website was designed to allow videos of both the subject and target to be presented to the annotator simultaneously. This allows users to make direct comparisons without memory demands or uncertainties concerning the scale of the videos. It should be noted that the video footage used does contain static objects which could be used as a reference point by which to assess the height of the individuals. In application settings the scale of the subjects would need to be conveyed to allow the eyewitness to accurately compare them to the observed suspect. Hence, the experiment mimics the proposed application. Further research would need to performed into assessing how subjects were presented to the eyewitness to allow accurate comparisons.

Drop-down boxes for each trait allowed users to describe how the subject differed from the target. The chosen label was emphasized by constructing a sentence explaining the given annotation - ensuring the annotator was comparing the subject to the target instead of vice versa. Eyewitness descriptions can be influenced by providing a default answer to a question, this is known as anchoring [58]. To avoid anchoring, all drop down boxes were initially void - forcing a response from the annotator.

3.3.3 Data Analysis

There have been 558 comparisons between the 80 subjects and 20 targets. These comparisons were collected from 57 annotators. From these comparisons 6783 inferred comparisons were calculated, detailing the differences between the 80 subjects. More information about the collected comparisons can be found in table 3.2.

	Collected	**Inferred**
Total trait comparisons	10602	128877
Total human comparisons	558	6783
Average human comparisons per subject	6.9	84.7
Average human comparisons per target	27.9	N/A
Average human comparisons per subject-target pair	0.69	N/A
Average human comparisons per subject-subject pair	N/A	1.05

Table 3.2: The number of collected and inferred bodily comparisons

The comparative annotations were compared with the absolute categorical labels gathered by Samangooei and Nixon [13] (see section 3.2 for details of the evaluation method). This comparison between annotation techniques will not show which is better, only how much each technique differs from the other. It was found that comparative annotations differed from absolute descriptions on 17% of occasions. This does not necessarily mean that the comparative annotations are better - just that they are considerably different.

Figure 3.2 shows the average difference between absolute and comparative annotations for each trait (using equation 3.3). The F-ratios, derived by ANOVA analysis, presented within [13] clearly show that absolute labels describe some features better than others. 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 on average by 30% compared to absolute labels. Given the inaccuracy of absolute labels for this trait, the difference *could be* indicative of more accurate 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 are on average 5% in respect to hair length, which shows that the comparisons are largely the same as the successful data obtained from the absolute annotations.

Figure 3.3 shows the correlation between the collected bodily comparisons. The correlations between traits were calculated using Elo relative measurements deduced from the comparative labels (introduced in section 4.1), the correlation results are presented here for completeness. The white cells within the figure represent traits with high correlation and the black cells represent traits with no correlation. It can be observed that large amounts of correlation occur within the 16 traits. The strongest correlations are between traits which describe some form of width or thickness, for example figure, chest, arm thickness and weight. Obviously weight and thickness are almost synonymous and strong correlations would be expected. This suggests that many of these traits could be excluded without much loss of information. Surprisingly the traits describing heights and lengths do not follow this pattern. Leg length has the strongest correlation with height at 0.76, whilst arm and neck length have correlations with height of 0.4 and 0.15

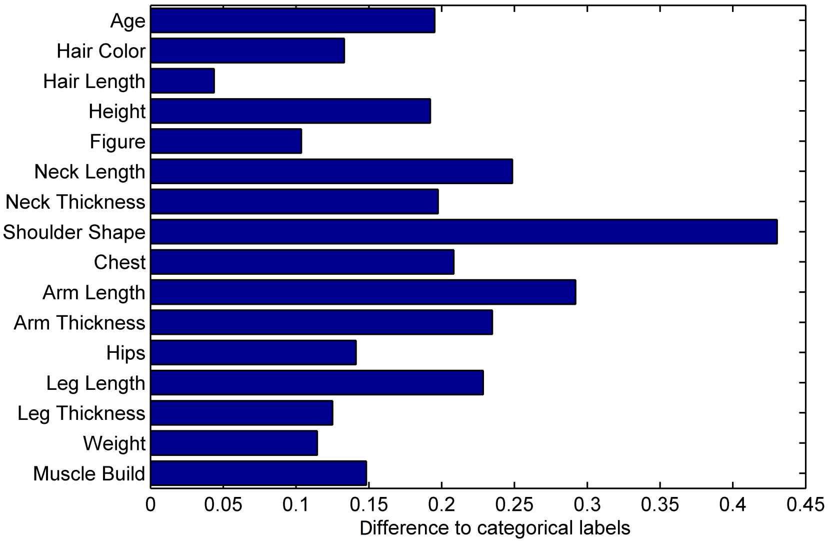


Figure 3.2: The average difference of each comparative trait and categorical annotation

respectively. Neck length is prone to a range of covariates such as long hair and collars, this may account for the weak correlation with height.

3.4 Facial Comparisons

Psychological research has determined that descriptions of faces, particularly inner facial features, are often inaccurate and are infrequently mentioned in descriptions of suspects. This is believed to be the result of a lack of vocabulary to describe facial features [5] and the inability to recall discrete features [2] (see section 2.1.1 for more details).

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 [17].

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.

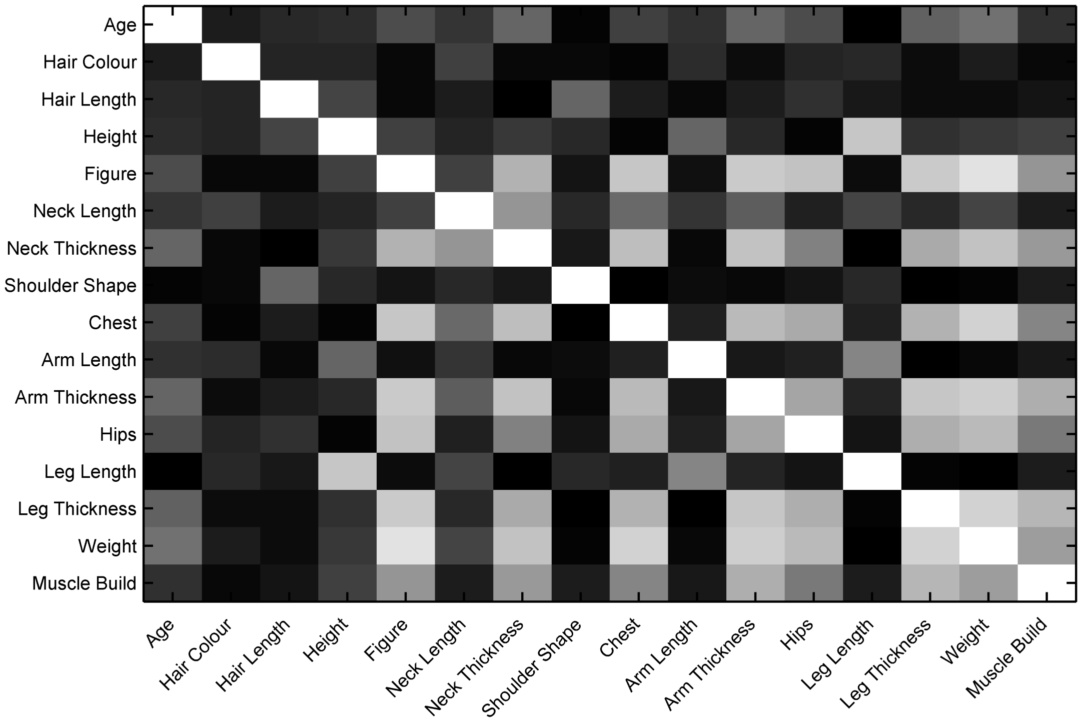


FIGURE 3.3: Correlation between bodily comparisons. White cells represent strong correlations. Black cells represent weak correlations.

3.4.1 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) [59] were used in this research. The FRS features a comprehensive selection of traits and has been used in other studies [60, 61]. The FRS contains 53 absolute traits, the majority described using 5 point bipolar scales. The modified FRS introduced in [60] 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 the bodily comparison experiment 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 3.3. 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 3.4).

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

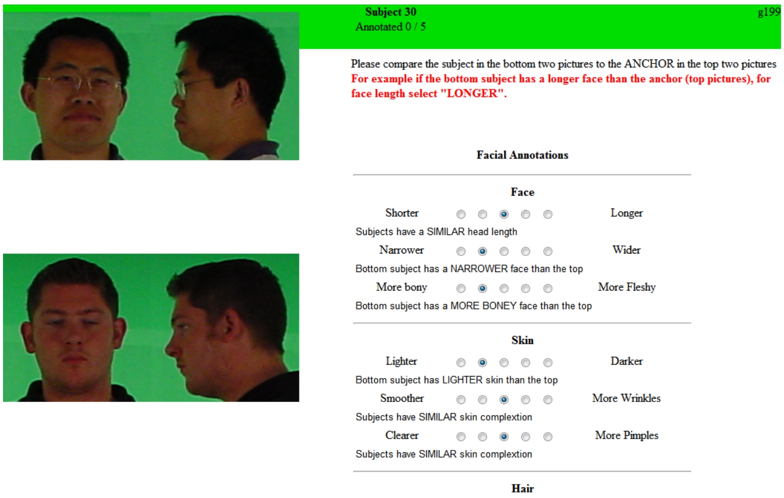
Table 3.3: Facial features used to compare subjects

3.4.2 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.

The SGDB used in the bodily comparison experiments is also comprised of facial images of the subjects, featuring both frontal and side images (more information can be found in appendix A). Using this database allows the accuracy of facial and bodily comparative descriptions to be compared using the same subjects. It also allows us to investigate any correlations between body and facial features.

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 3.3, except absolute labels were assigned



Subject 30
Annotated 0 / 5

g199

Please compare the subject in the bottom two pictures to the ANCHOR in the top two pictures
For example if the bottom subject has a longer face than the anchor (top pictures), for face length select "LONGER".

Facial Annotations

Face

Shorter ○ ○ ● ○ ○ Longer
Subjects have a SIMILAR head length

Narrower ○ ● ○ ○ ○ Wider
Bottom subject has a NARROWER face than the top

More bony ○ ● ○ ○ ○ More Fleshy
Bottom subject has a MORE BONEY face than the top

Skin

Lighter ○ ● ○ ○ ○ Darker
Bottom subject has LIGHTER skin than the top

Smoother ○ ○ ● ○ ○ More Wrinkles
Subjects have SIMILAR skin complextion

Clearer ○ ○ ● ○ ○ More Pimples
Subjects have SIMILAR skin complextion

Hair

Figure 3.4: Website used to obtain facial comparisons

to the extremes of the scales. The second section asked users to compare five subjects to a single target. The advantages of collecting comparisons in this way have already been discussed in section 3.2. 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 subjects.

Comparisons and absolute descriptions were collected using the website shown in figure 3.4. 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 [58] 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.

3.4.3 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 (see section 3.2) is shown in table 3.4. Further information about the absolute annotations can be seen in table 3.5.

	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 3.4: The number of collected and inferred facial comparisons

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

Table 3.5: The number of collected absolute facial annotations

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 3.5. 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 3.6 shows the correlation between the facial comparative features. The correlations between traits were calculated using Elo relative measurements deduced from the comparative labels (introduced in section 4.1), the correlation results are presented here for completeness. 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, especially when compared to the correlation present between bodily traits (figure 3.3). 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.

The correlation between facial and bodily comparisons is presented in figure 3.7. There is little correlation between the two sets of features showing that collecting both facial and bodily comparisons increases the amount of information available to identify the suspect. The lack of correlation also means that imputation methods would not work across the two sets of traits. The strongest correlations are present with the hair colour trait. Hair colour has been shown to be highly correlated with ethnicity and race [40], we can see in this figure that skin colour (Skin - Light/Dark) is highly correlated with hair colour as expected. Other traits with a strong correlation with hair colour include nose-narrow/wide, nose-flat/protruding, eyebrows-low/high and eyes-slanted/round suggesting that these traits may also be correlated with race and ethnicity.

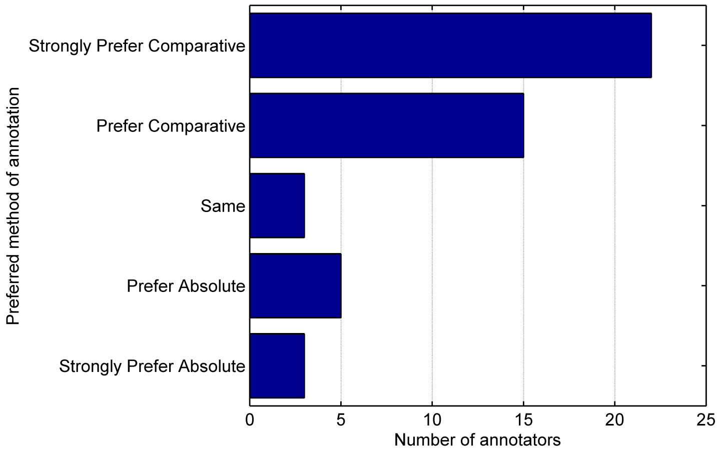


FIGURE 3.5: Annotators' preferred form of facial annotation

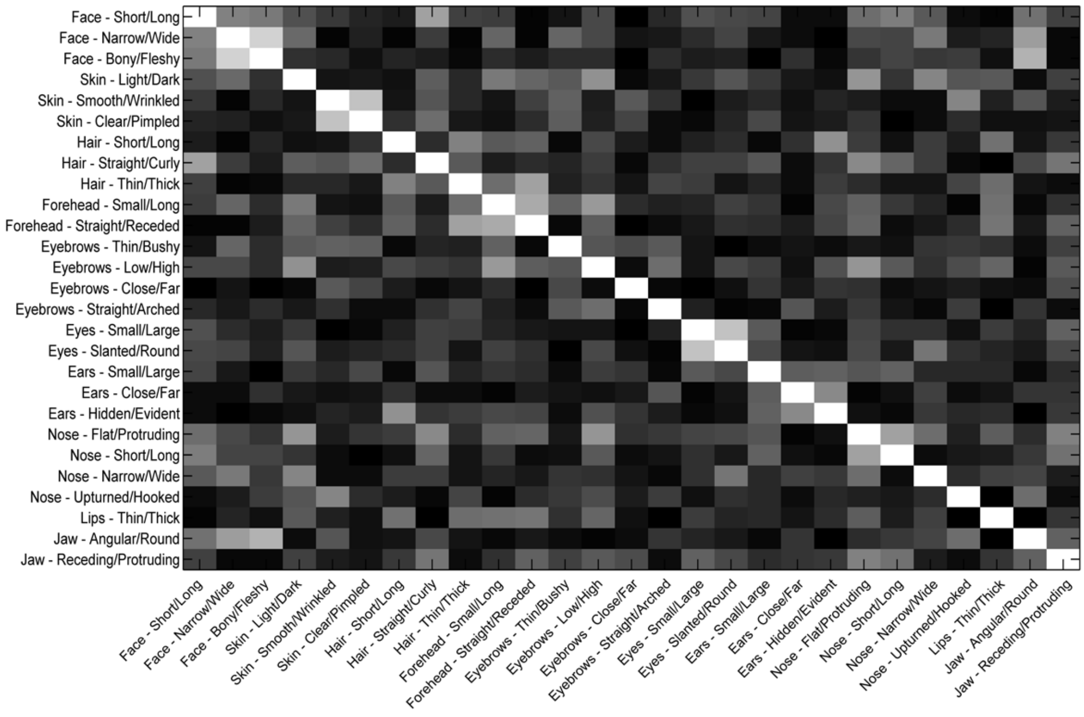


FIGURE 3.6: Correlation between facial comparisons. White cells represent strong correlations. Black cells represent weak correlations.

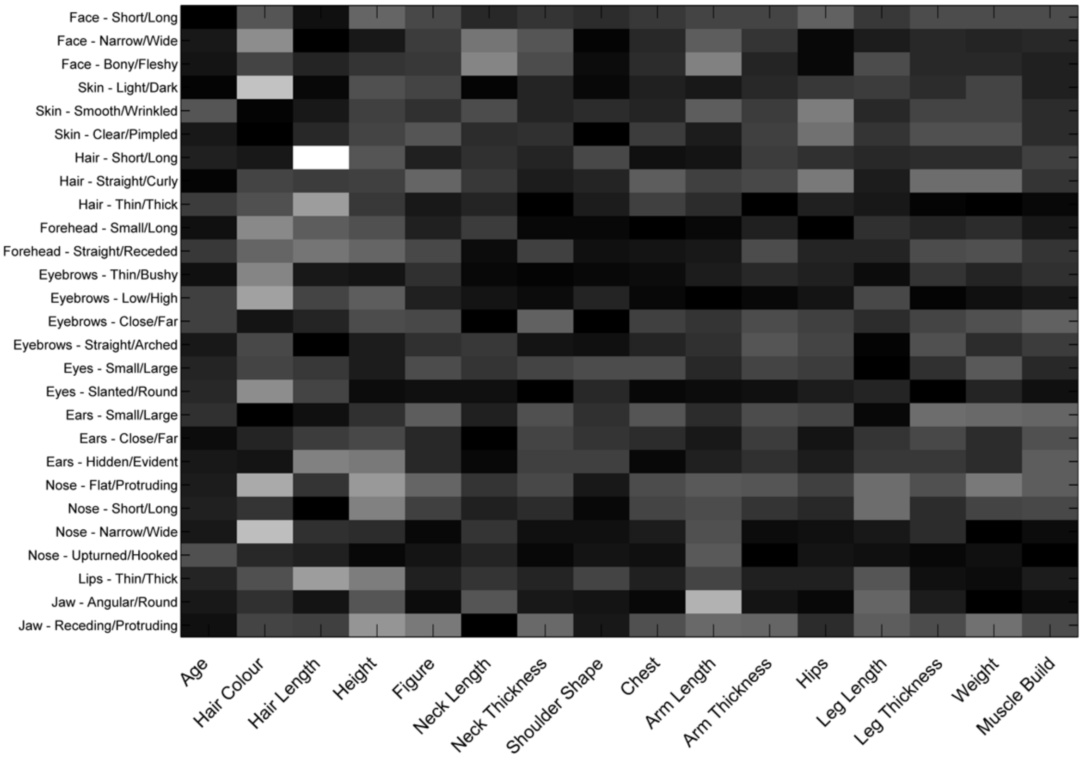


FIGURE 3.7: Correlation between facial and bodily comparisons. White cells represent strong correlations. Black cells represent weak correlations.

One unexpected observation from these results is the relationship between face-bony/fleshy and weight. The bony/fleshy trait was believed to be synonymous with weight but evidently this was not the case. This could indicate that people did not understand the meaning of the trait or that the relationship between weight and bony/fleshy face does not exist (in the SGDB). Further research into the relationship between head size and weight looked at relationships within the 1988 U.S. Army Anthropometry Survey (ANSUR) database [62]. The ANSUR database contains 34 anthropometric measurements of 3984 army personnel (male and female). The correlation between hip breadth and head breadth (0.219), head circumference (0.295) and head length (0.204) all suggest the lack of a strong relationship between head size and weight. Further examination of the ANSUR data shows little correlation between height and head breadth (0.122), head circumference (0.3452) and head length (0.3515) showing agreement with the low correlation of face-short/long and height seen in the database.

Figure 3.8 shows the difference between absolute and comparative facial descriptions. On average the descriptions differ by 26.3% which is slightly more than the difference between absolute and comparative bodily descriptions shown in figure 3.2. 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

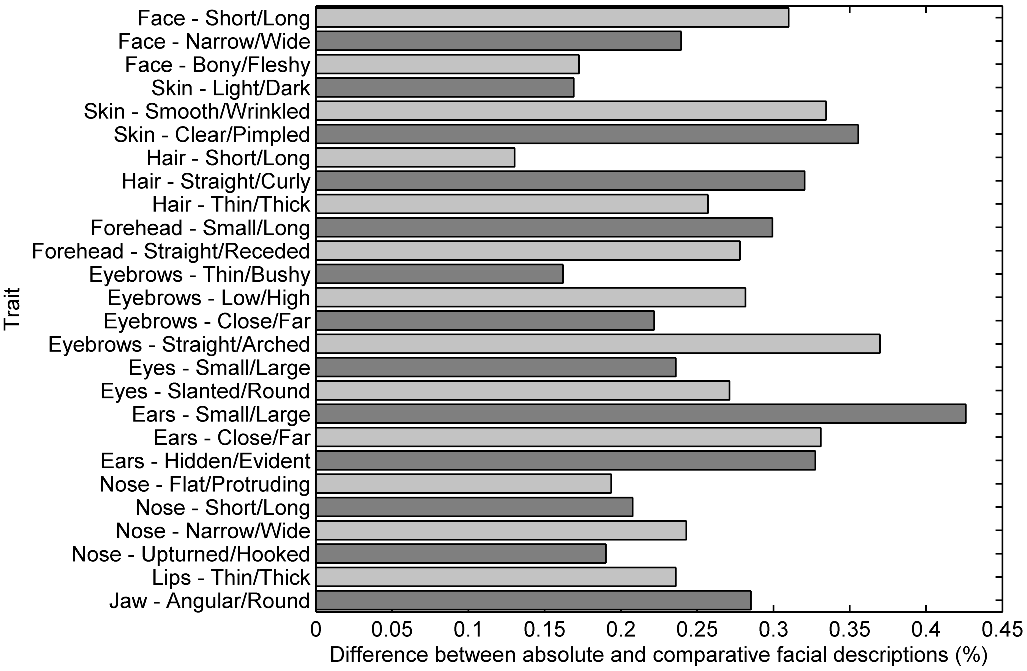


Figure 3.8: Differences between absolute and comparative facial descriptions

of the traits' averages and variation, this could explain why the absolute descriptions of these traits are comparatively similar to the relative 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 to be identified and accurately described leading to vast differences between absolute and comparative descriptions.

3.5 Conclusions

In this chapter comparative human descriptions were proposed as an alternative to continuous estimations and absolute labels for human description. Research into relative information in other fields has shown great advantages over absolute information - justifying the exploration of comparisons. A database of facial and bodily comparisons was collected using web based annotation forms which allow comparisons of subjects from the SGDB.

Analysis of the collected bodily comparisons show differences of 17% between absolute and comparative information. The largest differences are present in traits which have been shown to be difficult to describe absolutely - suggesting that the comparisons are providing new and more accurate information. Differences of 26.3% were observed between comparative and absolute facial features, further emphasizing the differing information contained within the two forms of annotation.

Correlation analysis between trait comparisons has shown strong structure between bodily traits and weak correlation between facial features. The additional traits and lack of redundant information within facial descriptions show that they should be more descriptive than bodily descriptions when available in criminal investigations. The weak correlations between bodily and facial traits clearly show that collecting both facial and bodily descriptions can drastically increase the amount of information available to identify the individual. In addition to the statistical results observed in this chapter, we have also shown that the majority of the participants of the facial experiment preferred comparative annotations over absolute.

The next chapter investigates how comparisons can be used as a biometric. Recognition experiments will confirm the discriminative nature of comparisons and identify advantages over other forms of human description.

Chapter 4

Identification using Comparisons

Comparisons have been introduced as a more robust method for gathering descriptions, but we must consider how they can be applied to identification applications. In addition to being a practical application for soft biometrics, identification experiments also explore the discriminative potential, accuracy and reliability (especially between different annotators) of comparative descriptions.

There are two separate biometric experiments we will consider. In this chapter we will identify subjects from a database of soft biometric signatures. In chapter 5 we attempt to retrieve subjects from a database of videos.

Soft biometric identification would be ideally suited to criminal investigations where an eyewitness description is available as well as a database of possible suspects each with soft biometric information, in this case obtained from previous human comparisons. The eyewitness would compare the suspect they observed to multiple subjects from the criminal database. Based on the given comparisons, a soft biometric feature vector representing the suspect would be inferred and used to query the database. The subjects within the database would be ordered based on their similarity to the feature vector. Figure 4.1 shows a diagram detailing the identification process. Querying criminal databases using physical descriptions is already common practice within police investigations, although currently it is performed using absolute labels and estimates of continuous traits rather than comparative descriptions [15].

Biometric recognition aims to identify an unknown subject by comparing their biometric signature to a database of biometric signatures. This type of identification is only possible when a database of biometric data is already available. A biometric database could be constructed using previous human comparisons or obtained from other forms of human representation. Sections 4.2 and 4.3 will focus on identifying a suspect from a soft biometric database formed from previous bodily and facial comparisons respectively. Later, chapter 5 introduces the automatic retrieval of a subject from video footage.

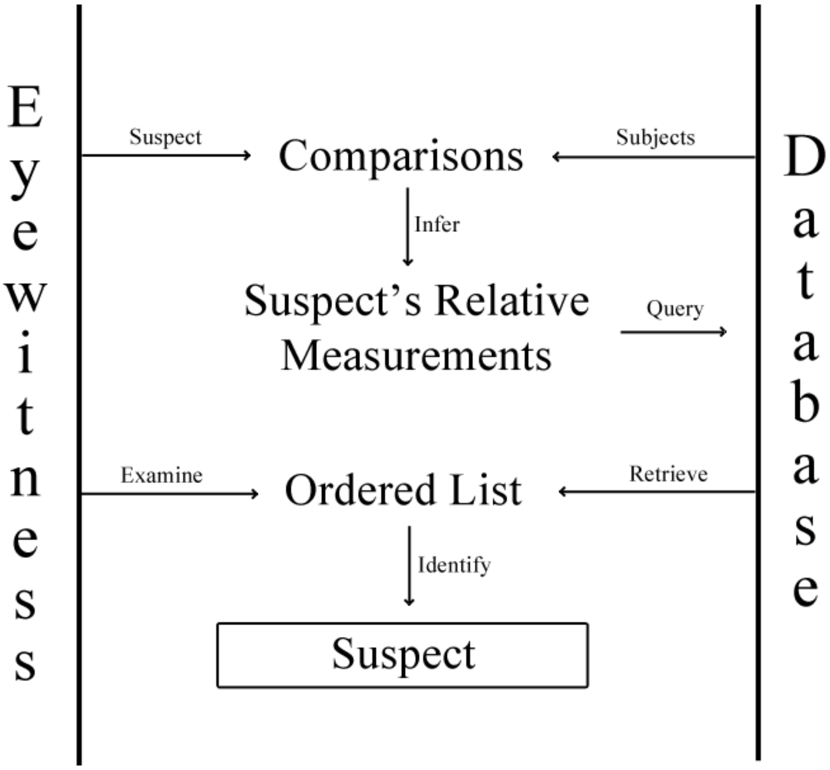


Figure 4.1: Verbal identification from soft biometric database

The first stage in both video retrieval and identification is to convert the comparative descriptions to relative measurements which can be used as a biometric signature, this is described in section 4.1.

4.1 Relative Measurements

Comparisons are inherently relative; each subject is described using another subject as a benchmark. Comparative annotations must be anchored to convey meaningful subject invariant information. The resulting value is defined as a *relative measurement*, providing a measurement of a 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.

4.1.1 Pairwise Comparisons

Comparisons between two entities, in respect to some property or attribute, are known as pairwise or paired comparisons. Each comparison describes the difference in 'strength' of the comparison criteria between two entities, for example the label 'taller' indicates that an entity has a stronger presence of the trait height than another. Typically, a pairwise comparison can result in one of three possible outcomes based on the strength of the comparison criteria exhibited by the compared entities, i and j:

- $i > j$: Entity i has a stronger presence of the comparison criteria
- $i = j$: Entities are equal in respect to the comparison criteria
- $i < j$: Entity j has a stronger presence of the comparison criteria

Multiple pairwise comparisons can be represented using a count matrix, $\mathbf{M}$, which records the number of times each entity was deemed to be 'better' than every other entity, such that $\mathbf{M}_{ij}$ represents the number of times $i > j$.

4.1.2 Thurstone's Model

In 1927, Thurstone introduced the law of comparative judgement [63], allowing the underlying strength of an entity's attribute (also known as the entity's *quality*) to be determined from pairwise comparisons. The model allowed the calculation of *quality scores* for a single pair of entities and was later extended to determine the quality of more than two entities. The law of comparative judgement revolutionized the field of psychometrics allowing information collected through pairwise comparisons to be quantified.

Thurstone's model employed Gaussian distributions to model pairwise comparisons. It was assumed that an individual's judgement of an entity's quality could be considered as a Gaussian random variable, modelling the subjective nature of assessing 'quality'. Therefore, the entity's quality score could be modelled by the mean quality of the resulting Gaussian distribution.

Given two entities, i and j, and their corresponding qualities:

$$i \sim \mathcal{N}\left(\mu_i, \sigma_i^2\right), \qquad j \sim \mathcal{N}\left(\mu_j, \sigma_j^2\right) \tag{4.1}$$

Thurstone states that an individual will compare the entities by drawing two realizations from the entities' quality distributions, shown in figure 4.2. The probability that the individual will choose i over j, $P(i > j)$, is dependent on whether their realization of i is greater than their realization of j, such that:

$$P(i > j) = P(i - j > 0) \tag{4.2}$$

Given that $i - j$ is the difference between two Gaussian random variables, $i - j$ is also

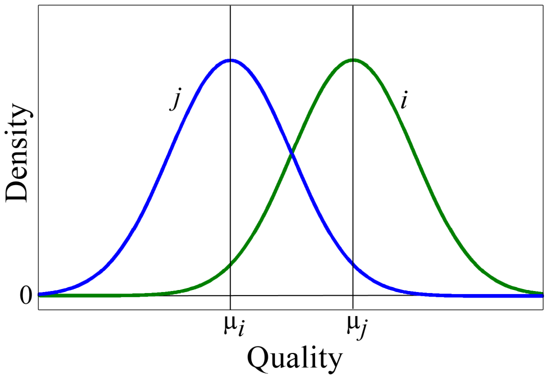


FIGURE 4.2: The PDFs of i and j

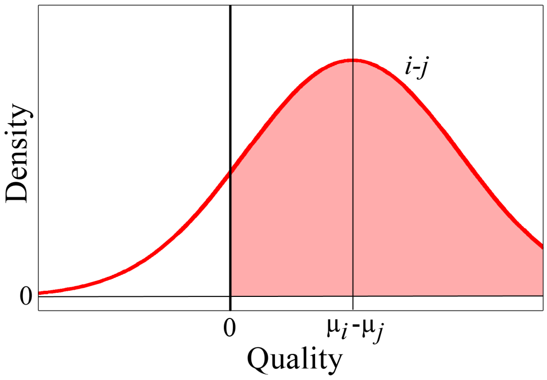


FIGURE 4.3: The PDF of $i - j$

a Gaussian random variable:

$$i - j \sim \mathcal{N}\left(\mu_i - \mu_j, \sigma_{i-j}^2\right) \tag{4.3}$$

$$\sigma_{i-j}^2 = \sigma_i^2 + \sigma_j^2 - 2\rho_{ij}\sigma_i\sigma_j \tag{4.4}$$

where ρ_{ij} is the correlation between i and j. The corresponding PDF can be seen in figure 4.3. The shaded area in figure 4.3 represents $P(i - j > 0)$ and can be calculated using:

$$P(i - j > 0) = 1 - \Phi\left(\frac{0 - (\mu_i - \mu_j)}{\sigma_{i-j}}\right) = \Phi\left(\frac{\mu_i - \mu_j}{\sigma_{i-j}}\right) \tag{4.5}$$

where $\Phi(x)$ is the standard normal cumulative distribution function (CDF). Once $P(i > j)$ is determined this can be inverted to find $\mu_i - \mu_j$, assuming σ_{i-j} is known:

$$\mu_i - \mu_j = \sigma_{i-j}\Phi^{-1}(P(i > j)) \tag{4.6}$$

where Φ^{-1} is the inverse of the standard normal CDF. Equation 4.6 is known as Thurstone's law of comparative judgement. Obviously in practical applications $\mu_i - \mu_j$ is not known and cannot be used to calculate $P(i > j)$, instead $P(i > j)$ must be approximated. Thurstone proposed that the proportion of people who favored entity i over entity j would be an accurate approximation of $P(i > j)$, such that:

$$P(i > j) = \frac{\mathbf{M}_{ij}}{\mathbf{M}_{ij} + \mathbf{M}_{ji}} \tag{4.7}$$

This approximation can be used in equation 4.6 to determine the difference between i and j, assuming the variance of i and j and the correlation between the two entities can be calculated or is known:

$$\mu_i - \mu_j = \sigma_{i-j}\Phi^{-1}\left(\frac{\mathbf{M}_{ij}}{\mathbf{M}_{ij} + \mathbf{M}_{ji}}\right) \tag{4.8}$$

Thurstone proposed five versions of the law of comparative judgement [63] which differ in approximations, assumptions and the level of simplicity. The most popular version is the *case V* model which assumes that the variance of i and j are equal and there is no correlation between the two entities:

$$\sigma_i^2 = \sigma_j^2 \tag{4.9}$$

$$\rho_{ij} = 0 \tag{4.10}$$

Resulting in Thurstone's case V model, where $\sigma = \sigma_i - \sigma j$ (Thurstone suggested setting $\sigma_i^2 = \sigma_j^2 = 0.5$ such that $\sigma = 1$):

$$\mu_i - \mu_j = \sigma\Phi^{-1}\left(P(i > j)\right) \tag{4.11}$$

As the variance and correlation cannot be accurately predicted, the case V model will be used throughout this thesis and will be referred to as the Thurstone model from now on. The value of sigma was set empirically, 1 was found to be a suitable value.

The law of comparative judgement provides a model to determine the quality of two entities based on pairwise comparisons. When comparing between more than two entities it is unlikely that a set of qualities will satisfy all of the available comparisons. For this reason an approximation must be made. The rest of this section will introduce a maximum likelihood solution [64][65] to this estimation problem. Initially we will consider a maximum likelihood solution for two entities, then this will be generalized to more than two entities.

Given two entities, i and j, and their corresponding comparison counts, $\mathbf{M}_{ij}$ and $\mathbf{M}_{ji}$, we would like to estimate their quality scores, μ_i and μ_j (and hence $P(i > j)$ as shown in equation 4.11). Considering the comparisons as a series of independent two option choices, the probability of $\mathbf{M}_{ij}$ and $\mathbf{M}_{ji}$ occurring given $P(i > j)$, can be calculated using the binomial distribution probability mass function:

$$P(\mathbf{M}_{ij}, \mathbf{M}_{ji} | P(i > j)) = \left(\begin{array}{c}\mathbf{M}_{ij} + \mathbf{M}_{ji}\\ \mathbf{M}_{ij}\end{array}\right) P(i > j)^{\mathbf{M}_{ij}}(1 - P(i > j))^{\mathbf{M}_{ji}} \tag{4.12}$$

where

$$\left(\begin{array}{c} n \\ k \end{array} \right) = \frac{n!}{k!(n-k)!} \tag{4.13}$$

The corresponding likelihood is as follows:

$$\mathcal{L}(P(i > j)|\mathbf{M}_{ij}, \mathbf{M}_{ji}) = P(\mathbf{M}_{ij}, \mathbf{M}_{ji}|P(i > j)) = \left(\begin{array}{c} \mathbf{M}_{ij} + \mathbf{M}_{ji} \\ \mathbf{M}_{ij} \end{array} \right) P(i > j)^{\mathbf{M}_{ij}} P(j > i)^{\mathbf{M}_{ji}} \tag{4.14}$$

Maximizing the likelihood leads to:

$$\mu_i - \mu_j = \sigma \Phi^{-1} \left(\underset{P(i>j)}{\arg\max}\ \mathcal{L}(P(i > j)|\mathbf{M}_{ij}, \mathbf{M}_{ji}) \right) \tag{4.15}$$

Given a count matrix, $\mathbf{M}$, and a vector of quality scores, $\mu = \{\mu_i | i = 1, ..., n\}$, this can be easily extended to n entities:

$$\mathcal{L}(\mu|\mathbf{M}) = P(\mathbf{M}|\mu) = \sum_{i,j}^{n} \left(\begin{array}{c} \mathbf{M}_{ij} + \mathbf{M}_{ji} \\ \mathbf{M}_{ij} \end{array} \right) \Phi \left(\frac{\mu_i - \mu_j}{\sigma} \right)^{\mathbf{M}_{ij}} \left(1 - \Phi \left(\frac{\mu_i - \mu_j}{\sigma} \right) \right)^{\mathbf{M}_{ji}} \tag{4.16}$$

resulting in the following optimization:

$$\underset{\mu}{\arg\max}\ \mathcal{L}(\mu|\mathbf{M}) \tag{4.17}$$

To ensure a unique solution a constraint such as $\sum_i \mu_i = 0$ can be enforced.

4.1.3 Elo Rating System

In essence, the Elo rating system provides a method of inferring a relative measurement from comparisons and is based on Thurstone's case V model [63]. Elo ratings were originally 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 traits to infer relative measurements, similar to how chess games compare two players' skill.

In the Elo rating system a 'match' is defined as a comparison between two players, A and B. This comparison could be a chess game or, in the case of soft biometrics, a visual comparison. The outcome of the match is a sample of how the two players differ from each other. The outcome is used to adjust the players' ratings to reflect the sample obtained from the match.

$$R_A^{'} = R_A + K(S_A - E_A) \tag{4.18}$$

$$R_B^{'} = R_B + K(S_B - E_B) \tag{4.19}$$

The system adjusts the players' ratings, R, based on the result of match S. The updated rating is derived from the difference between the result of a match, S (1 for a win, 0.5 for a draw and 0 for a loss), and the expected outcome, E, given the players' current ratings. This difference is controlled by K, which defines the maximum rating adjustment resulting from the match.

The expected outcome, E, is an adaption of equation 4.7 based on the Bradley-Terry-Luce model [66, 67] (which models $i{-}j$ as a logistic random variable), where Q represents a player's current rating. The constant U is chosen to reflect how a player's current rating can affect the expected result. This value was chosen empirically.

$$Q_A = 10^{R_A/U} \tag{4.20}$$

$$Q_B = 10^{R_B/U} \tag{4.21}$$

$$E_A = \frac{Q_A}{Q_A + Q_B} \tag{4.22}$$

$$E_B = \frac{Q_B}{Q_A + Q_B} \tag{4.23}$$

In chess the unknown measurement is the skill of the chess player - in the case of comparative annotations the unknown variable is the relative measurement of the attribute being compared. Comparisons between subjects provide a measure of difference between the subjects' attributes, just as chess games compare the skill level of the players. This information is used to adjust the inferred relative measurements of the two subjects.

To utilize the Elo rating system for human comparisons a new 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 -2 to 2 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 in the direction indicated by the comparison. The size of this adjustment is dependent on the error between the actual and expected results.

In chess the maximum rating adjustment variable, K, can be kept small and over many games the skill rating of a chess player can be slowly refined. In contrast, our application would benefit from obtaining accurate ratings from the least number of comparisons. This variable can be used to ensure that relative measurements obtained from large numbers of comparisons are comparable to those inferred from just a few comparisons.

To allow any form of retrieval or identification the gallery and probe biometric features (i.e. the relative measurements) must be comparable and similar. If K was a constant then the total rating adjustment possible for N comparisons would be $N * K$, this would mean that relative measurements inferred from a small number of comparisons would not be in the same range as those inferred from a large number of comparisons. To solve this K is adjusted based on the number of comparisons available. A maximum rating constant, m, is used to define $K = m/N$ allowing m to be fully explored by any number of comparisons.

The Elo rating system is used to calculate a single continuous variable, representing the relative strength of an attribute, from visual comparisons. In practice to generate a biometric feature vector describing a suspect, we must first obtain multiple human comparisons - comparing the suspect to multiple subjects (each with predefined Elo ratings). The rating system begins by setting the suspect's Elo ratings (one rating for each comparative trait) to a default value. Each comparison obtained is processed in turn, each time adjusting the suspect's Elo ratings. Once all the comparisons have been considered, a feature vector containing the Elo ratings is constructed.

The main advantage of this system is that it does not require exhaustive comparisons between all the subjects to calculate an accurate relative measurement. Instead it adjusts the target's relative measurements based on any available comparisons, taking into account the relative measurements of the compared subjects. In this way the ratings for a set of players can be inferred from a limited set of matches between them.

4.1.4 Relative Measurement Accuracy

Relative measurements detail how the subject's traits compare to other subjects within the population. We would expect that the relative measurements, if accurate, would be strongly correlated with the actual physical measurements of the traits. Determining the pixel height of a subject's gait signature from the SGDB video data allowed the correlation between an actual trait's measurement and the inferred relative measurement to be explored. The pixel height was calculated by averaging the silhouette height of a subject whilst in the midstance and midswing positions of the gait cycle (more information concerning the gait cycle can be found in section 5.1).

The accuracy of the Elo rating system, maximum likelihood Thurstone's model and comparative label averaging (details in following paragraph) were evaluated. In application settings we would seek to compare against the minimum amount of subjects to achieve an accurate relative measurement. This experiment assessed relative measurements generated using varying amounts of comparisons. For each technique n (ranging from 1 to 50) random comparisons were retrieved from the database and used to generate a subject's relative measurements. The correlation between a subject's relative height and pixel height was recorded. Due to the random element of comparison selection this process was repeated 250 times to determine an average correlation.

Comparative label averaging is the simplest approach to this problem and simply takes the average comparative label used to compare a subject to others. This is a naive approach as it does not consider the attributes of the subject being compared against. The average of the n comparisons was assigned as the subject's relative measurement.

The Thurstone method utilizes a count matrix to optimize a set of quality scores, these quality scores are the subjects' relative heights. Due to the count matrix approach each comparison is automatically utilized by both the subjects being compared, this results in each subject's relative height being generated from $2n$ comparisons *on average*, instead of n (each subject will have a minimum of n comparisons). For this reason the number of comparisons used to generate Thurstone relative heights should be treated as an average and results are shown only for even amounts of comparisons. To construct the count matrix n comparisons were retrieved for a subject, i. A single comparison, C_{ij}, compares subject i to subject j and is denoted with a value ranging from -2 to 2 - if positive subject i is taller and if negative subject j is taller. The count matrix, $\mathbf{M}$, is adjusted as follows:

$$\begin{aligned} \mathbf{M}_{ij} &= \mathbf{M}_{ij} + 1, \text{ if } C_{ij} > 0 \\ \mathbf{M}_{ji} &= \mathbf{M}_{ji} + 1, \text{ if } C_{ij} < 0 \end{aligned} \tag{4.24}$$

The maximum likelihood model was used to determine quality scores for each subject. The CVX matlab package was exploited to perform the required optimization.

The Elo rating system adjusts ratings based on both the result of the comparison and the current Elo ratings of the two subjects being compared. For this experiment each random comparison is only used to update a single subject's rating, this differs from a normal Elo implementation where both subjects have their ratings updated. This approach ensures that only n comparisons are used to infer the relative height of a subject. The Elo rating system begins by assigning each user a default rating of 1500, this value is arbitrary and the use of 1500 was chosen to reflect the standard value used in most Elo rating applications. Due to the default rating it is critical to update all of the subjects' ratings for each of the n comparisons in turn - this avoids basing rating adjustments on default Elo ratings (except for the first comparison of n).

The correlation between pixel height and relative measurements can be seen in figure 4.4. The best performing technique through out the range of comparisons is the Elo rating system, it achieves high correlations with low numbers of comparisons and matches the performance of the Thurstone method at higher numbers of comparisons.

The average labels actually outperform the Elo ratings with less than four comparisons, this is mainly because the Elo rating system needs multiple comparisons to produce

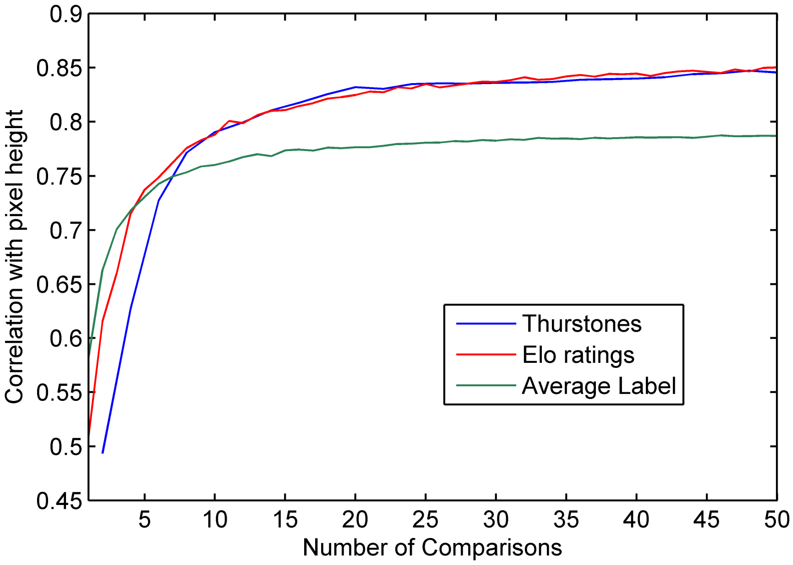


FIGURE 4.4: Correlation between pixel height and relative measurements generated by three techniques with varying amounts of comparisons

ratings which are not at the extremes of the Elo scale. Additionally, the Elo rating system can only utilize default ratings for the first comparison, which negates one of the main advantages of the Elo system. As expected the average labels perform worse than the other two techniques overall, this is because the technique does not consider the attributes of the subject being compared against. Although average labels did have a lower correlation (with more than 7 comparisons), the difference between average labels and Elo ratings was only 0.06 at 50 comparisons (drawn randomly from the comparison database), which is surprisingly successful for such a naive approach.

The Thurstone method does perform well at high numbers of comparisons, unfortunately, it does not perform as well with low numbers of comparisons. The Thurstone method does not utilize the extreme labels (the ‘much more’ and ‘much less’ labels) and only records which subject had a stronger presence of the comparison attribute - this may account for some of the inaccuracies at low numbers of comparisons. Additionally, there is little information to guide the optimization process when dealing with just a few comparisons.

The Elo rating system will be utilized throughout the rest of this thesis due to its performance and the low processing overhead. From now on the term relative measurement is synonymous with Elo ratings.

In figure 4.4 it can be seen that the correlation increases throughout the range presented (1-50 comparisons), clearly demonstrating that additional comparisons improve the accuracy of the resulting Elo relative measurements. The correlation was within 10% of its terminal value after 9 comparisons drawn randomly from the comparison database.

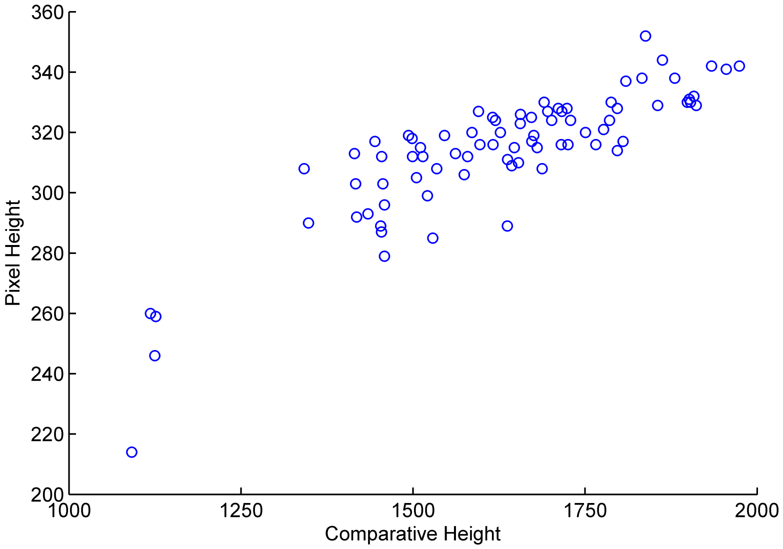


FIGURE 4.5: The relationship between pixel height and relative height

The Elo ratings shown in figure 4.5 were inferred from all the comparisons in the human comparison database. The correlation between pixel height and Elo relative height was statistically significant ($p < 0.0001$) at 0.87 - showing that the relative measurements inferred from human comparisons strongly represent the physical traits. This implies that the Elo rating system has inferred, from visual comparisons, an accurate ordering of the subjects based on height. The correlation between pixel height and the absolute height labels used previously (figure 2.4) was found to be 0.71. This is significantly ($p = 0.0018$ calculated using Fisher transformation) weaker than relative measurements mainly due to the highly subjective and categorical nature of the absolute labels.

To function successfully as a biometric feature we require a small intra-class variance between relative measurements describing the same subject's features. Relative measurements are inferred from comparisons, a different set of comparisons will generate different relative measurements. This difference must be small to allow identification. An experiment was conducted to fully explore the stability and robustness of relative measurements. For each subject within the SGDB, n random comparisons were obtained. Relative measurements describing the subject's features were inferred from the n comparisons. This process was repeated 500 times for each subject and for each n. Ideally the relative measurements describing the same feature on a subject would be very similar and exhibit a low variance. The standard deviation of the 500 relative measurements describing the same feature was recorded.

The average standard deviation of the relative measurements over all the subjects and traits is presented in figure 4.6 with varying numbers of comparisons. This graph shows how the variance of relative measurements produced from n random comparisons decreases with more comparisons. As expected the relative measurements inferred from

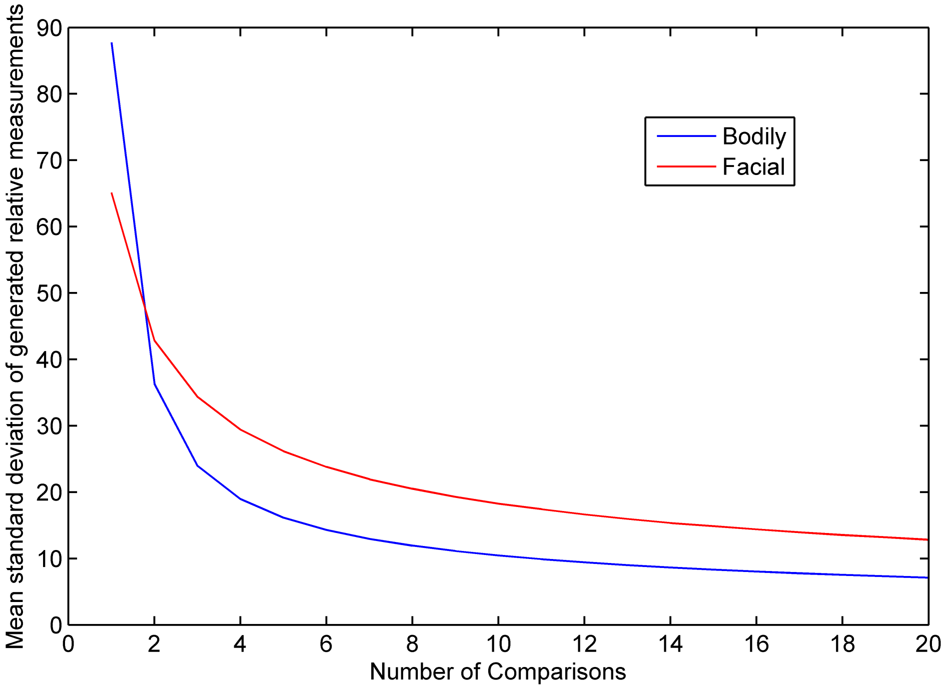


Figure 4.6: The average standard deviation of both bodily and facial relative measurements (for all of the comparative traits), describing the same subject, inferred from varying amounts of comparisons

a single comparison are very unreliable and would lead to inaccurate retrieval. This is because a single comparison only details the difference between the pair of subjects and does not provide much information about the subject in terms of the population. As more comparisons are considered more information is deduced about the relative strength of the subject's attributes within the population. The average standard deviation falls sharply with more comparisons and presents more robust descriptions of the subjects' traits - showing the more comparisons obtained, the more accurate and robust the inferred relative measurement. The same pattern can be seen for facial comparisons.

4.2 Identification using Bodily Comparisons

4.2.1 Technique

The identification experiment aims to retrieve a suspect from an 80 subject database (introduced in section 3.3.2). The biometric signatures within the database consist of all the 19 traits (table 3.1), where comparative traits are represented as relative measurements and absolute traits are represented by a value corresponding to the relevant categorical label. 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 79 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 three absolute traits within the bodily feature vector (gender, ethnicity and skin colour) were processed slightly differently due to fact that multiple absolute annotations can not be obtained from an individual. The absolute annotations gathered by Samangooei and Nixon [13] were used in this study. On average, each subject in the SGDB was described by 8 individual annotators. A single absolute annotation was obtained from the available annotations and used within the probe feature vector. The mode of the remaining absolute annotations was used to produce the suspect's gallery feature vector. Each absolute trait was represented within the feature vector using a single value representing the label assigned.

The similarity between the probe and gallery feature vectors was assessed using the sum of the Euclidean distance (for the relative measurements) and the Hamming distance (between absolute traits). 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.

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.

4.2.2 Accuracy

The recognition accuracy (i.e. rank 1 retrieval accuracy) over varying numbers of probe comparisons (n) is shown in figure 4.7. The recognition accuracy using just one comparison to construct the probe is 47%. Obviously one comparison only tells us how subjects differ and the resulting relative measurements are very inaccurate. Interestingly this result matches the recognition accuracy when using categorical labels, as seen in figure 2.5. As more comparisons are received, the accuracy of the probe's relative measurements increase, leading to improved recognition results. Strong similarities can be seen between figures 4.6 and 4.7, clearly the variance of the relative measurement directly impacts the retrieval performance. It can be seen with 9 comparisons a 91%

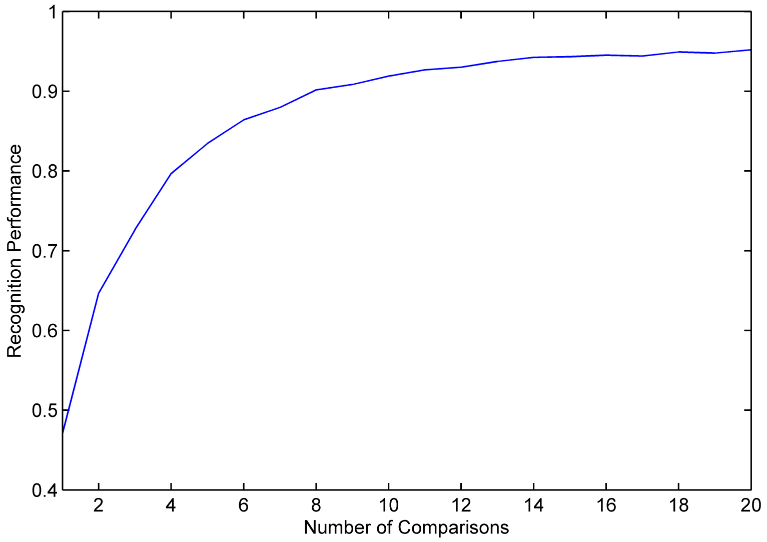


Figure 4.7: Bodily recognition accuracy using relative measurements obtained from different numbers of comparisons

correct recognition rate is achieved. Interestingly the Police and Criminal Evidence Act [68] states that an ideal identity parade should consist of 8 to 12 people, implying that a requirement of 9 comparisons would be suitable for application environments. The recognition performance continues increasing over the range shown, achieving a 95% correct recognition rate with 20 comparisons.

Figure 4.8 shows the retrieval performance of both relative measurements and absolute labels. Relative measurements inferred from just one comparison outperform absolute labels, achieving a 90% retrieval accuracy at rank 10 (i.e. 90% chance of the suspect being in the first 10 subjects returned from the database) compared to rank 15. As more comparisons are obtained relative measurements vastly outperform absolute labels, achieving a 99% retrieval accuracy at rank 5 with 10 comparisons.

Figure 4.9 shows the two most similar subjects within the SGDB in terms of their Elo ratings, and as such, they are often misidentified. It can be observed that the two subjects have almost identical bodily dimensions which are reflected within the Elo ratings. The major difference between the pair is skin colour but due to the coarse resolution of the trait's labels this difference was not reflected within the descriptions (both being labelled as ‘white’). In comparison, figure 4.10 shows a subject who was retrieved successfully even with only one comparison. The male subject has long hair, which is uncommon in 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 recognition capabilities of categorical labels with only one comparison. Recognition performance can be greatly improved by obtaining more comparisons.

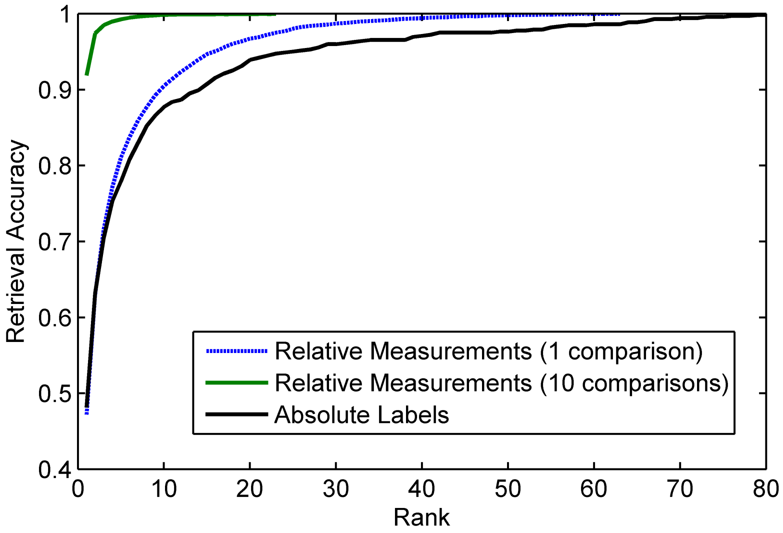


FIGURE 4.8: Retrieval accuracy of absolute labels and relative measurements inferred from 1 comparison and 10 comparisons

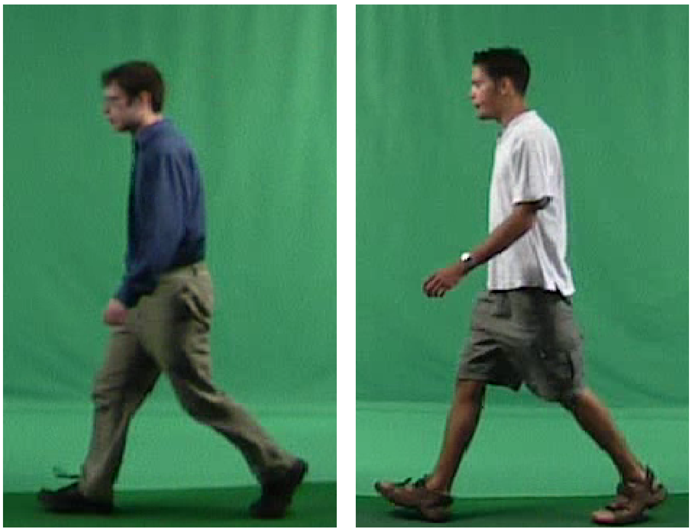


FIGURE 4.9: The most similar pair of subjects within the SGDB

4.3 Identification using Facial Comparisons and Descriptions

4.3.1 Technique

Facial recognition was conducted using both the comparative and absolute descriptions collected in section 3.4, allowing the performance of each to be compared.

Comparative facial recognition was performed in much the same way as the body recognition experiment (see section 4.2.1 for details). The only differences between the two experiments is that the facial biometric signatures were composed of 27 relative mea-

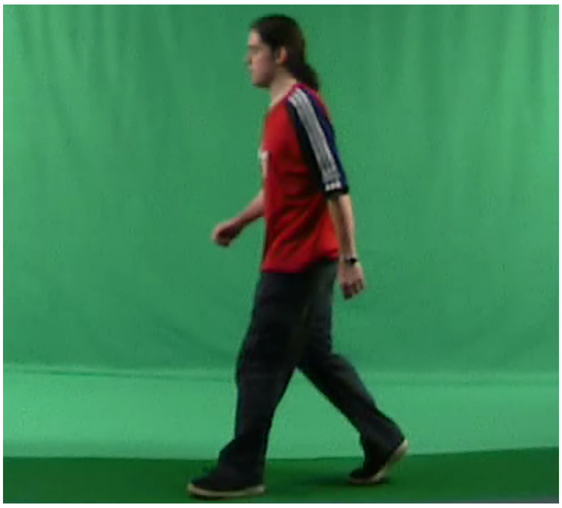


Figure 4.10: Subject achieved accurate retrieval due to uncommon traits

surements describing the facial features presented in table 3.3. The database used in this experiment is also smaller at only 40 subjects, the reasons behind this are discussed in section 3.4.2.

Identification using absolute facial descriptions utilized 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 experiment. 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.

4.3.2 Accuracy

The face recognition accuracy over varying numbers of probe comparisons is shown in figure 4.11. It can be seen that facial comparative descriptions vastly outperform bodily descriptions, achieving a 74.5% identification accuracy with a single comparison. A 99.3% recognition accuracy is obtained with just five comparisons, reaching a maximum of a 100% accuracy at 20 comparisons. It should be noted that the facial comparison database only contains 40 subjects compared to the 80 subject database used in the

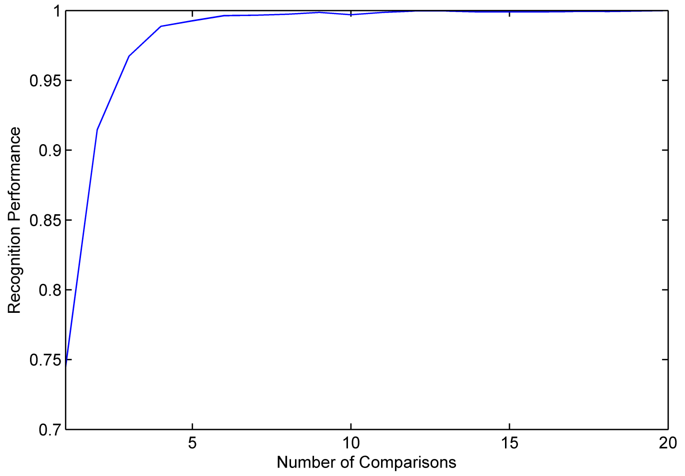


Figure 4.11: Facial recognition accuracy using relative measurements obtained from different numbers of comparisons

body recognition experiments.

Facial descriptions have three benefits which aid in identification when compared to bodily descriptions. It was shown in section 3.4.3 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 compared to only 19 bodily traits (a lot of which were highly correlated), which results in typically more distinctive descriptions allowing greater accuracy when identifying subjects.

The retrieval accuracy of the facial absolute labels (see section 3.4.2 for more details) is shown in figure 4.12, along with the retrieval accuracy of facial comparisons inferred from 1-3 comparisons. The accuracy of the facial absolute descriptions outperform the bodily absolute labels shown in figure 2.5, reinforcing the benefits of facial description over bodily. 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.

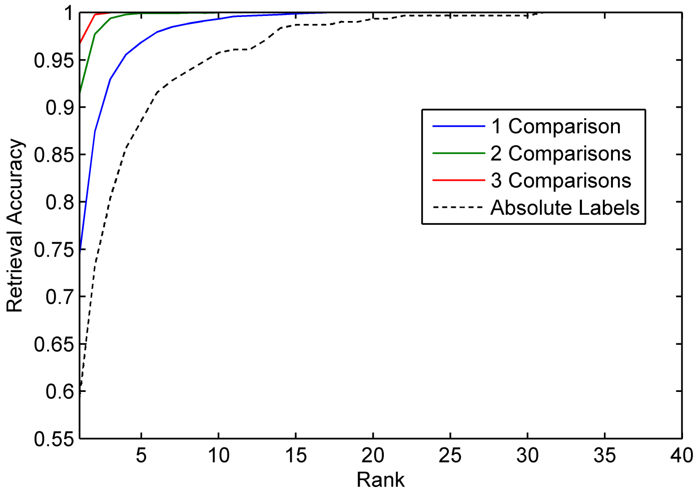


Figure 4.12: Face retrieval accuracy of absolute labels and relative measurements inferred from 1-3 comparisons

4.4 Conclusions

In this chapter we have studied the discriminatory capabilities of relative measurements. Three different techniques have been introduced to anchor comparative descriptions resulting in a single value which may be used in biometric signatures, know as a relative measurement. The Elo rating system has been shown to produce the most accurate relative measurements in the least amount of comparisons with a pixel height correlation of 0.783 after only 9 comparisons, reaching a maximum correlation of 0.87 with all the available comparisons. This correlation is 22% stronger than that observed with absolute labels, demonstrating the benefits of human comparisons.

We went on to show the recognition performance of Elo rating based relative measurements for both facial and bodily biometric signatures. Bodily comparisons achieved a 91% recognition accuracy with 9 comparisons demonstrating the discriminatory power of relative measurements. Facial relative measurements achieved a 99.8% recognition performance with 9 comparisons, outperforming bodily relative measurements. This was due to the lack of correlation between facial features, providing more independent information for which to identify an individual. The recognition results demonstrate the accuracy of relative measurements and the lack of variation across comparisons obtained from many different annotators.

The next chapter will focus on automatic person retrieval from video footage. Locating an individual within surveillance footage based on a description is a major aim of this project, bridging the semantic gap between biometric signatures and semantic descriptions.

Chapter 5

Retrieval from Video Footage

Biometric retrieval is the process of searching a database of subjects for an individual. In contrast to identification, retrieval aims to discover subjects who are most similar to the search query rather than confirm the identity of an individual. Traditional biometrics identify people by matching biometric signatures. This restricts identification and retrieval 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 are similar, in that it identifies people by matching signatures. The major difference is that a biometric signature based on relative measurements can be obtained from multiple sources. We have shown how relative measurements can be inferred from human descriptions (section 4.1). Many situations may require the described subject to be recognised based on images, surveillance footage, bodily measurements and different biometric signatures. This section will introduce how we can deduce relative measurements from visual and biometric representations, focusing on gait signatures. One exciting application of this technique is to retrieve subjects who match a human description from surveillance footage, this could allow the area surrounding a crime scene to be searched for an individual matching an eyewitness report.

Gait and face biometrics are among the only biometrics which can be obtained from a large distance, and as such they are ideally suited for surveillance applications. To accurately determine relative measurements the human representation must contain information about the soft traits which compose the soft biometric signature, in this case both facial and gait biometrics would suit the soft biometrics features we have explored. Facial recognition requires high resolution videos to capture the details required for identification and hence is not as robust as gait biometrics in surveillance applications. For this reason gait signatures were studied within this research. It is important to note that any human representation which encompassed the traits being described could potentially be used.

Figure 5.1 shows an example retrieval process. A soft biometric database containing rela-

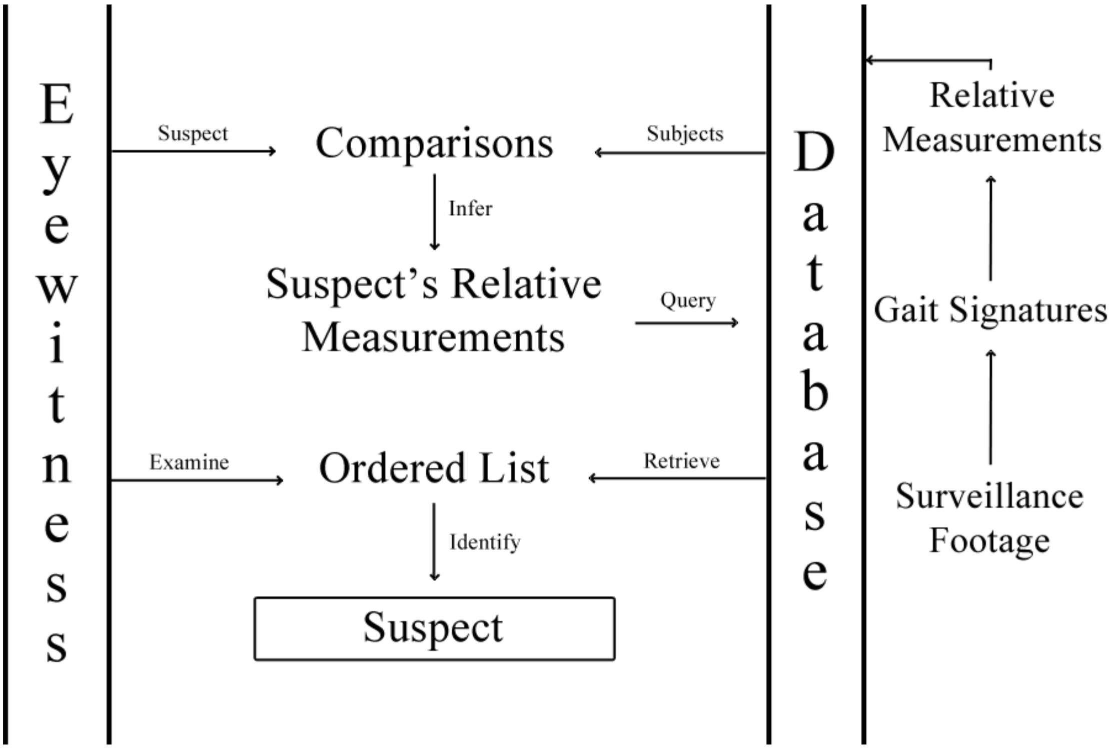


Figure 5.1: Verbal identification from video footage

tive measurements is obtained automatically from surveillance footage. This is achieved by converting the video to gait biometric signatures and then utilizing machine learning techniques to convert the gait representations to relative measurements (i.e. converting the measured biometric information to comparative semantic representations). In application scenarios, the witness would compare the observed suspect to multiple subjects within the database, with each comparison the relative measurements describing the suspect would be refined. When a sufficient number of comparisons have been made, the database (i.e. the surveillance footage) would be queried for individuals who are most similar to the suspect's relative measurements, returning an ordered list of possible matches.

This chapter will explore how this process is achieved. Section 5.1 briefly explains gait biometrics and introduces the various gait biometric signatures explored for video retrieval. The techniques utilized to convert gait signatures to relative measurements are explained in section 5.2. Finally sections 5.2.4 and 5.3 present the accuracy of the generated relative measurements and the retrieval performance of the system respectively.

5.1 Gait

Gait is the way in which an animal's limbs and body move to allow locomotion. Human gait differs between individuals [69] and it has been demonstrated that humans can

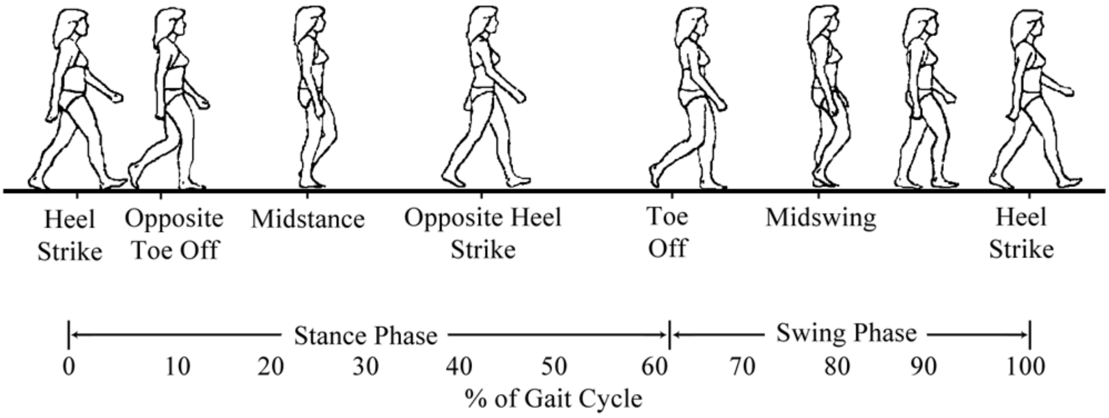


FIGURE 5.2: Gait cycle edited from [73]

recognize individuals from their gait [70, 71] suggesting that individual differences create a ‘unique’ pattern of movement. Gait biometrics has been studied in recent years and has shown that humans can be automatically recognized by the way they walk [36].

Gait biometrics has several advantages over other approaches. Gait can be identified over long distances and from low resolution imagery, making it ideal for surveillance applications. It is non-invasive and does not require cooperation from the individual. Finally, it is difficult to conceal (without hindering movement) unlike biometrics such as fingerprints and face recognition. However, gait is affected by covariates such as clothing (skirts, footwear and trench coats), walking surface and fatigue [72].

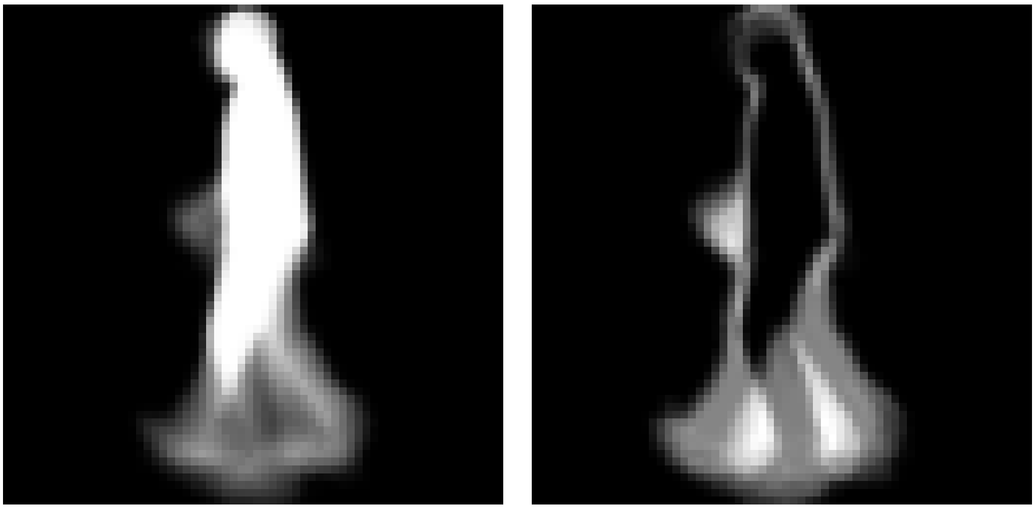
Human gait has a repeated pattern of movement, this is known as the gait cycle. A gait cycle begins with a heel strike (heel first touching the floor) with either foot and ends with the second heel strike of the same foot, i.e. a single gait cycle comprises of two steps. The cycle is shown in figure 5.2. The various positions within the gait cycle will be referred to later in this section.

This section will describe the four gait signatures which will be used to automatically determine the relative measurements.

5.1.1 Gait Signatures

Gait signatures are representations of an individual’s body shape and/or motion whilst walking. Gait signatures are comprised of two main types: model free and model based [36]. Model based signatures exploit the known dynamics of the human body, often focusing on how the limbs move. In contrast, model free signatures utilize the appearance of the body throughout the gait cycle.

This section will introduce the gait signatures used to automatically determine relative



(a)

(b)

Figure 5.3: Example gait signatures for the subject shown in figure 4.10. a) Average gait signature b) Differential gait signature

measurements. The bodily traits which compose the human comparisons are largely concerned with the appearance of the human body rather than its movement. For this reason the gait signatures studied in this project are model free representations.

All of the gait signatures introduced in this section are constructed based on binary silhouettes. These silhouettes are produced by removing the background from a video frame and converting the foreground elements (in this case a person) to a binary representation. The 'inside' scenario of the SGDB was recorded in front of a chroma keyed background to allow accurate background subtraction. Median background subtraction was applied to each frame followed by a conversion to a binary representation. To remove noise, connected component analysis is used to identify the largest set of connected foreground pixels - resulting in a binary silhouette of the individual.

5.1.1.1 Average gait signature

Average silhouette gait signatures describe the summation of a subject's binary silhouettes across one gait cycle [74]. This signature describes both the movement and appearance of the individual's body.

The average gait signature was constructed by first uniformly scaling a subject's silhouette to achieve a height of 64 pixels. Scaling both the height and width in proportion ensured the aspect ratio was preserved, this is critical when assessing the relationship between the subject's height and width. The scaling procedure removes absolute height information which effectively makes the signature distance invariant. The silhouette is translated so its centre of mass is centred on a 64x64 pixel image. The scaled and translated silhouettes over a single gait cycle (identified using silhouette width [75]) are summed resulting in a single 64x64 pixel signature, this can be expressed with the following:

$$\mathbf{A} = \sum_{i=0}^{n} \mathbf{S}(i) \tag{5.1}$$

where $\mathbf{A}$ is the 64x64 pixel average signature and $\mathbf{S}(i)$ represents the ith scaled and centred silhouette in the gait cycle comprising of n silhouettes. Finally each pixel's intensity is normalized within the range of 0 to 1. The pixels' intensities are a measure of how often the subject's body is in a certain location during the cycle - representing the subject's movement and their body shape. An example of an average gait signature can be seen in figure 5.3(a).

5.1.1.2 Differential gait signature

Veres et al. [76] identified the most critical features within the average gait signature for recognition. It was discovered that the majority of the important features were concentrated within the contours of the head and body. The legs, which contain the majority of the movement information, were found to play a small role in recognition performance. The study then considered the features of differential gait signatures, which are similar to average gait signatures although a differencing operation is used to combine silhouettes - focusing more on the movement of the silhouette over the gait cycle. The analysis showed that proportionally more of the important features were located within the leg features of the differential gait signature. Differential signatures also achieved the highest recognition rates in this study.

Differential gait signatures were constructed in the same way as average signatures although equation 5.1 was replaced with equation 5.2, where $\mathbf{D}$ is the 64x64 pixel differential signature. An example differential signature can be seen in figure 5.3(b).

$$\mathbf{D} = \sum_{i=0}^{n-1} |\mathbf{S}(i+1) - \mathbf{S}(i)| \tag{5.2}$$

5.1.1.3 Unwrapped gait signature

The unwrapped silhouette signature proposed by Wang et al. [77] utilizes pixel measurements of the silhouette. The advantage of this signature over the previous approaches is that many of the physical measurements described within the soft traits are explicitly measured rather than being implicit within the pixel data. The process begins by unwrapping the silhouette by stepping around the silhouette contour, recording the distance between the silhouette's centre of mass and the position of the n boundary pixels

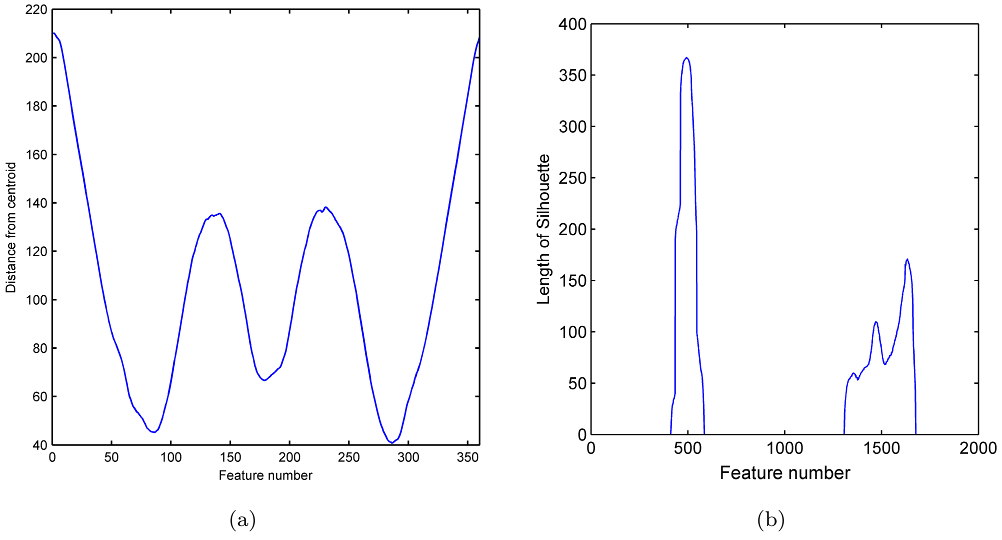


Figure 5.4: Example gait signatures for the subject shown in figure 4.10. a) Unwrapped silhouette signature at heel strike b) Measured gait signature

- resulting in a vector of n distances. This vector is treated as a one dimensional signal which can be used as a silhouette signature. The signal is normalized to a default length using sampling.

In this experiment an unwrapped gait signature was composed of 5 unwrapped silhouette signatures recorded during a gait cycle. The five silhouettes were defined as the three heel strikes and two stances (midstance and midswing) featured within the gait cycle. These are easy to identify providing a standard signature. Each unwrapped silhouette signature was constrained to 360 features. The 5 silhouette signatures were combined resulting in a 1800 feature gait signature, an example can be seen in figure 5.4(a).

5.1.1.4 Measured gait signature

The measured gait signature was inspired by Johnson and Bobick [78] and focuses on explicit pixel measurements of gait silhouettes. Many of the bodily traits explored in this thesis describe either height or width, for this reason the measured gait signature records the height and width of gait silhouettes. Silhouettes at the three heel strikes throughout a gait cycle are identified, using silhouette width [75], and used to create the gait signature. These three frames feature the least self occlusion and the most information about arm and leg length. Each frame is processed individually. First the silhouette is centred within a 1000x1000 pixel image, such that pixels within the silhouette have an intensity of more than zero and the background pixels have an intensity of zero. The distance between the first silhouette pixel (i.e. a pixel with intensity greater than zero) and the last is recorded for each row and column of the image (zero if none present). The resulting 2000 pixel measurements are averaged over the three heel strikes and used as

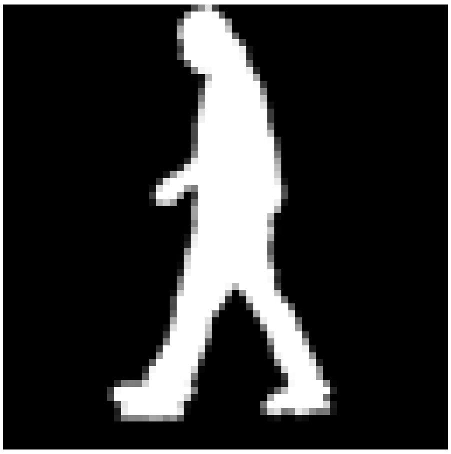


Figure 5.5: Cropped heel strike silhouette of the subject shown in figure 4.10

a gait signature.

Figure 5.5 shows a heel strike silhouette. It can be seen that recording pixel distances of the silhouette for each row and column will measure many of the bodily traits included within the relative measurements, including an indication of leg length (measurement of stride) and width (measurement of hips and upper thighs), arm length, weight, height, neck thickness, chest and hips. The signature is not scale invariant and relies on silhouettes being of similar size. The SGDB records all subjects from a set distance, allowing direct comparisons of heights and widths. For unconstrained environments an individual's height may be calculated from objects in the scene [79] or from the use of a calibrated camera [80], the resulting silhouette could then be scaled based on the individual's height. This signature is not optimal for unconstrained environments and is mainly studied as an investigation into ideal signature characteristics. An example signature is shown in figure 5.4(b).

5.2 Converting Gait Signatures to Soft Feature Vectors

To retrieve subjects from video footage the gait representations of the individuals must be converted to relative measurements. This allows the video database to be searched based on relative measurements deduced from comparative descriptions. To allow conversions we must learn the relationship between the visual gait signature and the soft biometric features. The majority of the soft features are relative measurements and hence require regression techniques. Three absolute features are also present within the soft feature vector requiring classification approaches. In this section we will introduce three machine learning techniques which will be deployed and the accuracy of the inferred soft biometric features.

5.2.1 k Nearest Neighbours

k nearest neighbours is a very simple technique suitable for both classification and regression. The approach works by comparing the distances between a test feature vector and the training data. The k closest training vectors are used to assign a class or calculate a relative measurement for the test vector. Classification was performed by assigning the most frequent class amongst the k nearest neighbours to the test vector, if there was not a single class with a majority the test vector was assigned to the class of its nearest neighbour. Relative measurements were calculated using a weighted average of the k nearest neighbours' relative measurements, weighted based on the inverse distance between neighbour and test vector.

The accuracy of the technique relies on the selection of a suitable value for k. If k is too large, the result will reflect either the most frequent class or the average relative measurement; if it is too small, the technique will be overly sensitive to noise. k was selected based on the misclassification or mean squared error of a 10 fold cross validation performed on the training data. The Euclidean distance metric was used to determine the k nearest neighbours, as it is perhaps the most popular, although other distance metrics could be used.

5.2.2 Support Vector Machine

Support vector machines (SVMs) [81] are a supervised learning technique suitable for regression and classification. SVMs construct hyperplanes which separate the training data with the maximum margin, this improves the model's ability to generalize to unseen data.

For two class linear classification problems a SVM constructs a hyperplane, $(w.x)+b = 0$, where x is a set of points, w is the normal vector to the plane and $\frac{b}{\|w\|}$ is the offset of the hyperplane from the origin. The hyperplane constructed aims to separate the two classes present within the training data, $(x_1, y_1), \ldots, (x_l, y_l), x \in \mathbb{R}^d, y \in \{-1, +1\}$. To separate the two classes the hyperplane must satisfy the following constraint:

$$y_i(x_i.w + b) \geq 1, \; i = 1, \ldots, l \tag{5.3}$$

To produce a maximal margin between the two classes the hyperplane must be an equal distance from both classes, the closest vectors to the hyperplane from both classes are known as support vectors and they satisfy $y_i(x_i.w + b) = 1$. The margin is measured using $2/\|w\|$ and the resulting value must be maximized (subject to equation 5.3) to achieve an optimal separation between the two classes (practically it is easier to minimize the following convex objective, $\frac{1}{2}\|w\|^2$). This constrained optimization problem is solved using Lagrange multipliers and involves finding the dot product between pairs of vectors within the training data.

Non-linear problems can be solved in the same way by utilizing a kernel [82]. The kernel trick [83] maps the training data to a higher dimensional feature space with the use of a mathematical mapping function like polynomials and radial basis functions. In the higher dimensional feature space the problem may move from a non-linear classification problem to a linear. This is possible as the constrained optimization problem is solved using dot products between vectors, the kernel simply changes the space in which the dot product is calculated.

The soft margin extension [84] was introduced to cope with otherwise infeasible constraints of the optimization problem, allowing linearly inseparable data to be classified (with errors). A non-negative slack variable is added to the minimization condition which acts as a penalty function for classification errors. A soft margin SVM searches for a hyperplane which splits the classes with the least error.

Regression can be performed using the same principles [85]. The aim is to approximate the training input, $(x_1, y_1), \dots, (x_l, y_l), x \in \mathbb{R}^d, y \in \mathbb{R}$, with a linear function of the form, $f(x) = (w.x) + b$. Minimizing $\frac{1}{2} \|w\|^2$ results in a simpler model that is most likely to generalize to unseen data and not overfit the training data. The ϵ-insensitive loss function defines an acceptable error rate which constrains this optimization. The resulting model will be the simplest model to describe the training data whilst keeping errors below ϵ. The soft margin extension can also be applied to regression problems, this allows (and penalizes) errors above ϵ to deal with otherwise unsolvable problems. This constrained optimization is solved using Lagrange multipliers and can exploit mapping functions allowing the construction of non-linear regression models.

Support vector machines were used to both classify the three absolute traits and regress the remaining 16 comparative soft traits. Regression used the ϵ-insensitive SVM with soft margins. Linear and radial basis function (Gaussian) kernels were experimented with. A grid search was used to set the various parameters required by the SVM (ϵ and cost) and the kernel mapping functions (sigma for the RBF kernel). A range of parameters values was evaluated using the training data and the best performing values were selected for use with the test data. The grid search's performance was measured using the mean squared error for regression and misclassification rate for classification tasks, performance was determined using 10 fold cross validation on the training data.

5.2.3 SVDImpute

Previously categorical human descriptions were automatically obtained from average silhouette gait signatures [13] using latent semantic analysis (LSA) [39]. LSA is a very popular technique for identifying structure between different types of data. Unfortunately this technique specializes in classification tasks and due to the continuous nature of the relative measurements this is not suitable. LSA uses singular value decomposition (SVD). This statistical technique can be used to approximate a co-occurrence matrix, identifying underlying structure. LSA utilizes this structure to create a vector space model used to classify data. This structure can also be exploited to perform regression, ideal for relative measurements.

SVDImpute [86] is a regression technique used frequently in predicting missing data within DNA microarrays [87]. The technique is based on SVD which allows the underlying structure within a co-occurrence matrix to be identified. This is ideal for identifying the structure between two types of representation, in our case the structure between gait signatures and relative measurements.

SVD was used to approximate a co-occurrence matrix which contains the occurrences of features (both soft and gait) for each training subject. Each soft trait was considered separately, emphasizing the relationship between a single trait and the gait signature. Each subject's feature vector contained the gait signature and a single trait's relative measurement. Each training subject's feature vector was combined to create the co-occurrence matrix $\mathbf{O}$.

The co-occurrence matrix will describe the relationship between the gait features and the relative measurement. This structure is obscured by a majority of irrelevant occurrences between features. By removing the irrelevant relationships (noise) the underlying semantic structure can be observed. Noise is removed by determining a rank reduced approximation of the occurrence matrix. SVD is utilized to factorize the matrix, allowing a rank reduced version to be determined. First factorizing the matrix $\mathbf{O}$ into three matrices such that:

$$\mathbf{O} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \tag{5.4}$$

Where $\mathbf{U}$ and $\mathbf{V}$ are orthogonal matrices and $\mathbf{\Sigma}$ is a diagonal matrix. $\mathbf{\Sigma}$ contains the singular values from $\mathbf{O}$ and the matrices $\mathbf{U}$ and $\mathbf{V}$ contain the left and right singular vectors of $\mathbf{O}$. By reducing the rank of these matrices the dimensionality of the problem is reduced, resulting in an approximation of $\mathbf{O}$. This approximation will ideally retain the most integral information within $\mathbf{O}$ and remove the noise. The reduced rank k determines how many dimensions the data is condensed to and ultimately how much information is lost. The diagonal matrix $\mathbf{\Sigma}$ consists of r diagonal values, these are ordered by size (and the corresponding row and column permutations applied to $\mathbf{V}$ and $\mathbf{U}$). By removing the smallest singular values the majority of the information is retained, resulting in an approximation of $\mathbf{O}$, $\mathbf{O}_k$ such that $\mathbf{O}_k = \mathbf{U}_k\mathbf{\Sigma}_k\mathbf{V}_k^T$.

SVDimpute was introduced to predict missing variables within data by utilizing the structure learnt using SVD. We can use this technique to predict a trait's relative measurement from a gait signature. First a feature vector, x, is constructed. The feature vector contains the information known about the subject (gait signature) and has empty features for the unknown information (relative measurement). The present data within this feature vector (in this case the gait signature) is regressed against the corresponding singular vectors within $\mathbf{V}_k$. $\mathbf{V}_k$ is shortened to reflect the missing variables within $\mathbf{x}$ becoming $\mathbf{V}_k^*$, the regression is performed as follows:

$$\hat{\boldsymbol{\beta}} = (\mathbf{V}_k^{*T}\mathbf{V}_k^*)^{-1}\mathbf{V}_k^{*T}\mathbf{x} \tag{5.5}$$

Once the scalars $\boldsymbol{\beta}$ have been resolved the missing elements of $\mathbf{x}$ (the relative measurement) can be found as $\mathbf{V}_k^{(*)}\hat{\boldsymbol{\beta}}$ where $\mathbf{V}_k^{(*)}$ are the elements of $\mathbf{V}_k$ representing the missing elements of $\mathbf{x}$. Using this technique any missing data within a feature vector can be rebuilt utilizing the structure implicitly learnt by the SVD technique.

SVDImpute was used to predict relative measurements from gait signatures. The reduced rank variable k determines the amount of data retained within the co-occurrence matrix and hence dictates the accuracy of the technique. k was selected based on the mean squared error of a 10 fold cross validation performed on the training data.

5.2.4 Accuracy

To assess the suitability of the proposed machine learning techniques the accuracy of predicted relative measurements and absolute traits must be considered. Soft biometric feature vectors were automatically determined from the four gait signatures presented in section 5.1.1. Ten fold cross validation split the 80 subjects from the SGDB into testing and training sets. The training set was used to train the relevant machine learning technique and define the various parameters required. Soft biometric feature vectors composed of the 19 traits shown in table 3.1 were generated from the gait signatures of the subjects within the test set. Each trait was regressed or classified individually and combined to create the subject's feature vector. The correct classification rate of generated absolute traits and the accuracy of generated relative measurements demonstrate the suitability of the 3 machine learning techniques and the 4 gait signatures.

Table 5.1: Proportion of error present in relative measurements obtained automatically from different gait signatures

		Gait Signature			
		Average	Unwrapped	Differential	Measured
Technique	kNN	0.139	0.150	0.144	0.149
	Linear SVM	0.139	0.135	0.135	0.122
	RBF SVM	0.141	0.135	0.136	0.126
	SVDImpute	0.120	0.132	0.123	0.125

5.2.4.1 Relative Measurements

The proportion of error between the actual and predicted relative measurements is shown in table 5.1, where the proportion is calculated as follows:

$$MAE = \frac{1}{n * m} \sum_{i=1}^{n} \sum_{j=1}^{m} |a_{i,j} - e_{i,j}| \tag{5.6}$$

$$Proportion = \frac{MAE}{MaxRelMes - MinRelMes} \tag{5.7}$$

where n is the number of subjects, m is the amount of comparative traits. $a_{i,j}$ and $e_{i,j}$ are the actual and estimated relative measurement respectively describing subject i's physical trait j. $MinRelMes$ and $MaxRelMes$ represent the minimum and maximum possible values for the relative measurement and hence represent the range of possible relative measurements. This representation was utilized to present an easily interpretable value which represents the percentage of error between actual and estimated relative measurements given the fixed bounds of the relative measurements.

Most physical traits have a Gaussian distribution, this is also reflected within Elo ratings, such that the majority of subjects are close to the average. The errors present within table 5.1 demonstrate errors of less than 15% given the normalized Elo range of 0-1, but considering the Gaussian distribution of the Elo ratings this may be misleading. Naively generating Elo ratings of 0.5 for every comparative trait results in a mean absolute error of 0.18, putting the values into perspective.

It is quite clear from the results that the kNN approach was the worst performing regression technique achieving an average error of 14.5%. The kNN approach relies on training examples which are similar to the test feature vector. In this experiment there were only 72 training examples which may have limited the approach. Better results may be obtained with a larger database. The two SVM techniques achieved similar error rates with 13.2% and 13.4% for linear and RBF respectively. It has been shown that moving to a non-linear feature space using the RBF kernel was detrimental to the regression performance. The data being used was already very highly dimensional (ranging from 1800 to 4096 features) reducing the need to move to a higher dimensionality space to discover a linear model. SVDImpute was shown to be the best performing technique

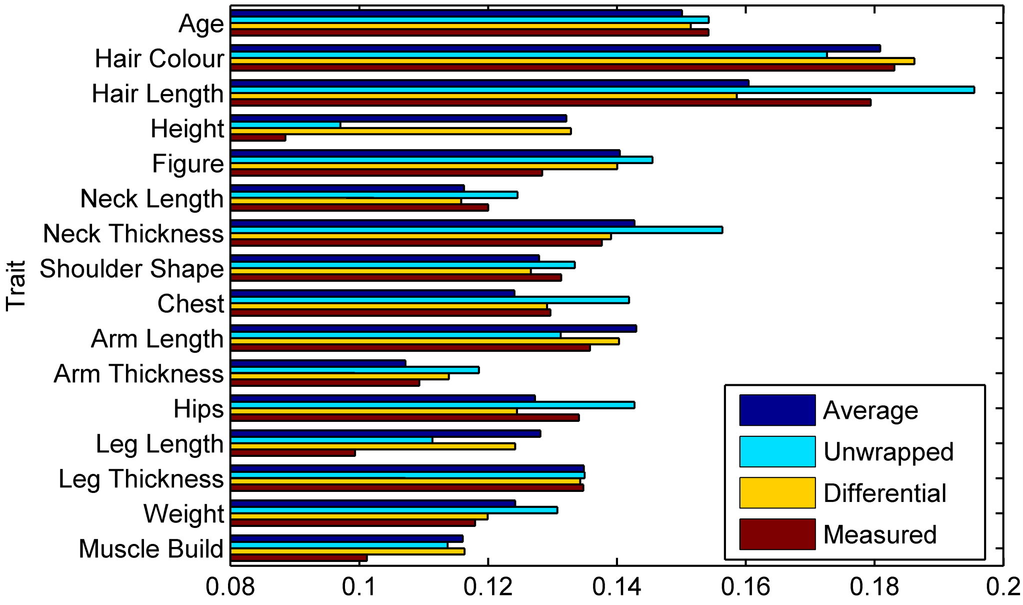


Figure 5.6: Proportion of error between actual and estimated relative measurements

with an average error rate of 12.5% over the four gait signatures. The variable k which controls the amount of data kept during the rank reduction varied during the experiments between 3 and 11. This reduction of the input information to just a few fundamental components allows the most important relationships between relative measurements and visual features to be identified and exploited.

The four gait signatures used in this experiment can be split into two categories: silhouette representations and silhouette measurements. The average and differential signatures represent two forms of silhouette representations consisting of pixels, both signatures achieved a 13.5% average error over the four machine learning approaches. The unwrapped and measured gait signatures consisted of measurements from around the silhouette and achieved a 13.8% and 13.1% error respectively. Unlike the machine learning approaches the accuracies of the gait signatures have remained comparatively constant.

Figure 5.6 shows the proportion of error between actual and estimated relative measurements over the four machine learning techniques for each of the four gait signatures.

Although the gait signatures appear to be almost equal there are variations in the individual trait performances which highlight the differences between the signatures. The first trait of interest is hair length, although the error was comparatively high for all the signatures, the silhouette measurement signatures performed significantly worse. Both measurement signatures take rough measurements from around the body which is not ideal for identifying traits, like hair length, which are conveyed within a few pixels (in the case of hair length generally the trait is identified using a few pixels at the back of the neck). In contrast, the silhouette representations will convey these small features. However, pixel representations suffer when inferring traits which are deduced from the structure of multiple pixels like height and leg length. Obviously height is implicitly contained within the pixel representations but the relationship is not as obvious. Measured and unwrapped signatures explicitly measure lengths and height leading to a strong relationship between the measurement and the relative measurement, ultimately leading to very accurate regression.

The measured gait signature outperformed the unwrapped signature on traits which describe thickness or weight, namely neck thickness, arm thickness, figure, chest, hips and weight. Figure 3.3 has shown that these traits are highly correlated and hence a single correct measurement of one of these traits would serve to annotate the rest successfully. The unwrapped signature measures the distance between the centre of the silhouette and 360 points around the silhouette. It is clear that the representation would record information about the width from the centre of the silhouette to the chest or stomach. Unfortunately, maybe due to differences in centroid location in respect to the silhouette or the variation in position of the 360 points on different subjects, width was not as accurately deduced from the signature. In contrast, the measured signature measures the width of the silhouette at every row, explicitly measuring many of the width and weight features mentioned previously. As such, it was not surprising that the unwrapped signatures led to the performance with the greatest error.

We have observed that silhouette representations and silhouette measurements signatures excel at different traits. The fusion of both types of gait signature was believed to allow more accurate relative measurements to be produced. SVDImpute has been shown to be the most successful and reliable method of predicting relative measurements and will be used in this fusion experiment. The best performing silhouette representation (average) and silhouette measurement (measured) signatures were fused. Fusion was achieved by simply adding the average gait signature vector to the end of the measured feature vector. This was then used to generate relative measurements using the experimental protocol introduced at the start of this section.

The fusion of the average and measured gait signatures resulted in a mean absolute error of 0.107 using the SVDImpute technique. The individual trait errors of the average and measured signatures and the fusion of the two can be seen in figure 5.7. It can be seen that the average and measured fusion performed more successfully than its individual components in almost all of the traits. The accuracy of traits like hair length and chest benefited from the advantages of the average gait signature. Likewise, traits such as height and leg length benefited from the inclusion of the measured gait signature. The fusion of the two signatures also improved the accuracy of traits like age, neck thickness and leg thickness which were comparatively inaccurate when predicted with either average or measured gait signatures. This suggests that the combination of both pixel intensities and measurements aid the regression of some traits.

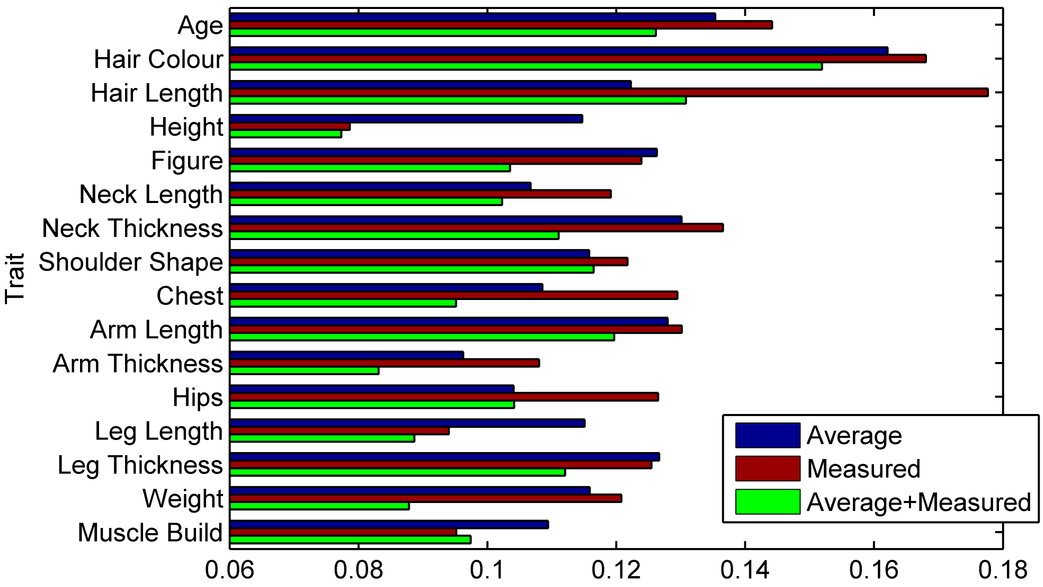


FIGURE 5.7: Proportion of error between actual and estimated relative measurements using SVDImpute

5.2.4.2 Absolute Traits

Each soft biometric feature vector is composed of 16 relative measurements and 3 absolute traits. This section will explore the accuracy of predicting absolute labels from the four gait signatures presented in section 5.1.1. Generating absolute labels from gait signatures was approached using the same experimental process explained previously for relative measurements. The trait's labels were each assigned a numerical class. The two machine learning techniques (SVDImpute was not compatible) classified the test vectors as one of the trait's classes. 10 fold cross validation split the population in a way which attempted to have at least one of each class in the training set.

The average correct classification rate of the three absolute traits can be seen in table 5.2. It can be observed that the correct classification rate is reasonably low (average of 74%) and quite consistent over the different gait signatures and machine learning approaches. The reason for this is that two of the three absolute traits, namely ethnicity and skin colour, cannot be accurately predicted from the four gait signatures due to the traits' reliance on colour (and smaller traits not visible in gait signatures). The best correct classification rate for ethnicity was 76.3%, whilst the most successful skin colour classification rate was 73.7%. On average ethnicity and skin colour was classified correctly 71% and 67% respectively. This could be greatly improved by including some representation of colour within the gait signatures.

To provide a better understanding of the gait signatures and machine learning perfor-

Table 5.2: Average correct classification rate of absolute labels (gender, ethnicity and skin colour) generated from different gait signatures

		Gait Signature			
		Average	Unwrapped	Differential	Measured
Tech.	*k*NN	73.2%	75.9%	72.4%	71.9%
	Linear SVM	73.2%	71.1%	81.1%	76.3%
	RBF SVM	73.2%	73.2%	74.6%	72.8%

mance table 5.3 shows just the gender correct classification rate. The average gender classification performance was 83%. It can be seen that the RBF SVM obtains the same CCR for all gait signatures suggesting that it did not identify a pattern between the visual features and gender. Like relative measurements, the addition of the RBF kernel to the SVM reduced the accuracy of the generated absolute labels when compared to the linear SVM. The linear SVM was the most successful technique (excluding the unwrapped result) obtaining the two best classification rates of 93%. The measured gait signature was the most successful gait signature over the three techniques. As mentioned previously the measured gait signature explicitly represents height which is strongly correlated with gender [88], allowing accurate classification. Additionally, the measured gait signature measures the width of every row of the silhouette, this will record information regarding the individual's chest which is obviously highly correlated with gender.

Table 5.3: Correct classification rate of gender generated from different gait signatures

		Gait Signature			
		Average	Unwrapped	Differential	Measured
Tech.	*k*NN	80%	90%	83%	88%
	Linear SVM	86%	68%	93%	93%
	RBF SVM	80%	80%	80%	80%

5.3 Retrieval

To determine the application potential of such a system, we must also consider the retrieval accuracy. The retrieval process is identical to that introduced in section 4.2 although all the subjects' soft biometric feature vectors within the gallery are generated automatically from gait signatures. The relative measurements were calculated using the SVDImpute technique based on the fusion of the average and measured gait signatures. The absolute labels were determined from measured gait signatures using a linear SVM. The gait signatures and machine learning techniques used to generate the soft biometric feature vectors were selected based on the results presented in section 5.2.4.

The database is composed of feature vectors which were automatically determined from gait signatures. This replicates a database of videos being automatically converted to soft biometric feature vectors for querying. A probe feature vector was constructed for each of the 80 subjects in turn, using all the available comparisons and absolute descriptions. The similarity between the probe and every subjects' biometric signature within the database was assessed using the sum of the Euclidean distance (for the relative measurements) and the Hamming distance (between absolute traits). 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. This experiment replicates the use case scenario of searching surveillance footage based on a description of a suspect.

5.3.1 Genetic Algorithm Trait Weighting

The errors shown in figure 5.7 demonstrate that some features are calculated from gait signatures with more accuracy. These features should be favoured when retrieving subjects, as they are more likely to be correct. Additionally some features are more discriminative or accurately described than others, this has been seen in sections 2.1.2 and 2.2.3. The more discriminative or accurate a trait, the more it should be favoured when retrieving subjects from a video database.

For these reasons, the similarity measures were weighted when used for retrieval. A genetic algorithm was used to discover the optimal weights. The genetic algorithm begins by generating a population of weight vectors. Each weight vector contains a weight for each of the 19 traits. Every member of the population is evaluated by calculating the sum of the retrieval accuracies over all of the 80 ranks - this is known as the individual's fitness. A new population is then created with the aim of producing fitter weight vectors. This drives the genetic algorithm to produce weight vectors which achieve high retrieval accuracies at low ranks.

Three methods were used to create the new population. The genetic algorithm was elitist so the top performing weight vector was automatically moved to the next population - this ensures we do not lose the best solutions found so far. Randomly changing some weight vectors allows new weight combinations to be explored, this is known as mutation and works by selecting a weight vector and changing some of its weights based on a random value. A uniform crossover technique was also used to combine weight vectors. Crossover and mutation was performed using a rank based roulette wheel selection method which favours the best performing weight vectors. 80% of the time crossover was performed over mutation. Crossover was performed more than mutation as it allows the genetic algorithm to optimize the population. Mutation allows occasional random exploration of the fitness landscape to introduce genetic diversity.

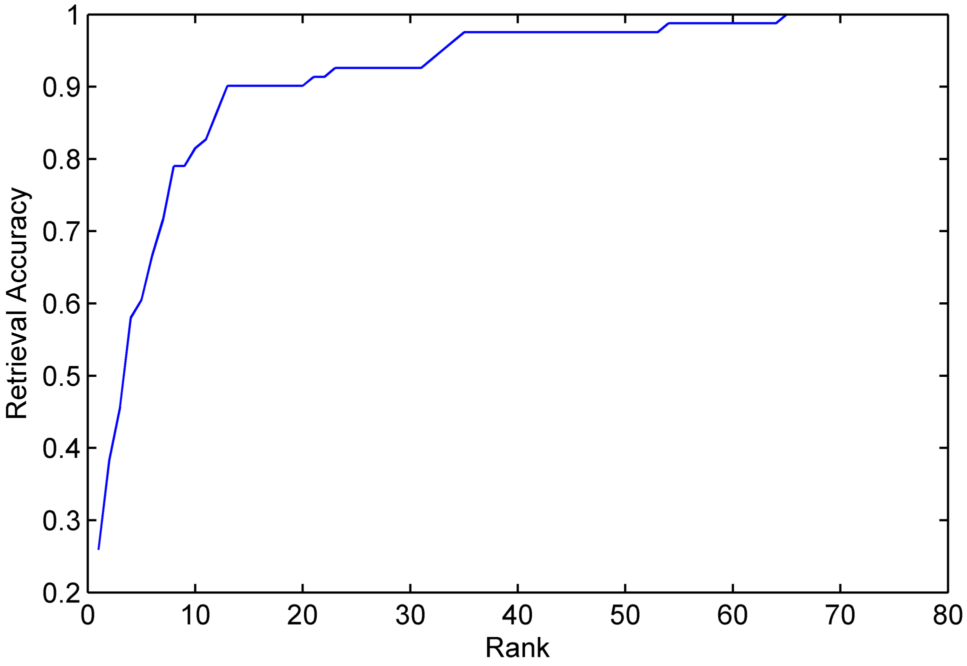


FIGURE 5.8: Retrieval performance of a soft biometric feature vector composed of relative measurements generated from a fusion (average and measured) gait signature using SVDImpute and categorical labels inferred from measure gait signatures using a linear SVM

5.3.2 Retrieval Results

The retrieval results can be seen in figure 5.8. Although the retrieval accuracy is only 26% at rank 1, it quickly increases achieving a 72% retrieval accuracy at rank 7 and 90% at rank 13. As such, there is a 90% chance that the correct subject is retrieved within the top 13 matches.

Figure 5.9 shows the weight assigned to each trait. The larger the weight the more influence it had in retrieving the subject. It can be seen there are five highly weighted traits, namely height, hair length, chest, arm length and weight. All of these traits are comparative, demonstrating the discriminatory capabilities of relative measurements and the accuracy of generating relative measurements from gait signatures. Surprisingly, the third most influential trait, hair length, is one of the most inaccurate traits to predict from gait signatures achieving an error rate of 0.13 (figure 5.7). Although inaccurate, the information may be strongly weighted due to its strong correlation with gender and its lack of correlation with the other 16 relative measurements - providing additional discriminative information about the subjects. As expected, the height and weight traits are the most favoured traits due to the accuracy of the generated relative measurements and their discriminative capabilities.

Traits with a weight of less that 0.2 have very little impact on the similarity measurement. Ethnicity and skin colour were assigned weights of 0.06 and 0.03 respectively. The low

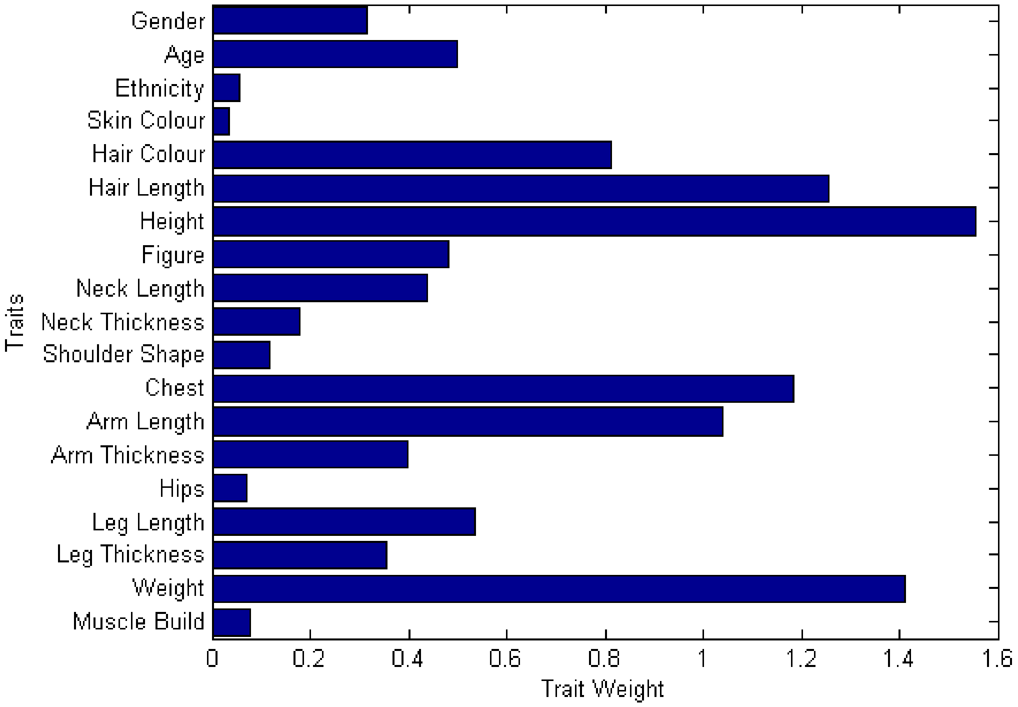


Figure 5.9: Calculated weights used to favour traits within the retrieval experiment shown in figure 5.8

weight assigned to these traits is due to the inaccuracy of predicting ethnicity and skin colour from gait signatures with no colour information. The remaining weak traits: neck thickness, shoulder shape, hips and muscle build, were accurately predicted from the gait signatures but evidently did not improve retrieval accuracy. Except for shoulder shape, these traits are highly correlated with the more favoured traits and hence only provide duplicate information. In comparison, shoulder shape achieved a moderate relative measurement error rate of 0.116 and has little correlation with other bodily traits, suggesting that the relative measurements are not discriminative possibly due to the inaccuracies of the comparisons (caused by the difficulty of observing shoulder shape from a side-on view point) or the innate indiscriminate nature of the trait.

In chapter 2 we introduced various psychological studies which explored the saliency and accuracy of described physical features. The weights presented in figure 5.9 give an indication of the importance of each trait. Traits were favoured not only based on their accuracy but also their discriminatory capability. MacLeod et al. [8] identified five of the most reliable descriptors, namely weight, height, leg thickness, chest size and leg length. The weights assigned to height, weight and chest coincide with this experimental study, however, our results differ in the importance of leg descriptions. Leg length and width are highly correlated with height and weight respectively and as such provide little additional discriminative information over the more accurate height and weight traits.

The retrieval experiment demonstrates the possibility of *automatically* filtering video data based on a description. Improvements may be found with the use of model-based gait signatures, which would provide a stronger relationship between the gait signatures and soft biometric relative measurements. This would increase the accuracy of the automatically generated relative measurements leading to improved retrieval results (see section 7.2.4 for more information).

5.4 Conclusions

Automatically determining soft biometric feature vectors from other forms of human representations is critical for many applications of soft biometrics. The most exciting of which is automatically searching CCTV footage and mugshots for people matching a description obtained from an eyewitness.

In this section we have explored the suitability of four different gait signatures and three machine learning approaches. The results have been successful allowing relative measurements to be determined with an accuracy of 10.7% and absolute labels to be classified with a CCR of 81%.

Of the four gait signatures the measured gait signature was the most successful. The explicit measurement of bodily traits led to strong relationships between visual and soft features resulting in accurate regression. Silhouette representation signatures were found to be successful at determining small traits, such as hair length, but were comparatively worse at larger traits like height and leg length. Unwrapped signatures performed similarly to the measured gait signatures except for traits which involved weight and width. The combination of average and measured gait signatures combined the benefits of silhouette representations and silhouette measurements, improving regression results by 12%.

One of the main aims of this project was to allow video footage to be searched using a human description. Video retrieval was conducted, achieving a 90% retrieval accuracy at rank 13. We believe these results represent a good start to this difficult problem and supports the possibility of automatically searching CCTV footage using comparative descriptions.

Chapter 6

Memory and Human Comparisons

6.1 Eyewitness Memory

Eyewitness identification is often treated as key evidence in criminal cases. However, memory can have a detrimental effect on the description and identification of observed suspects. A study of 205 cases of wrongful conviction showed that 50% were predominantly due to mistaken identification [89]. This evidence was bolstered by a recent review of 239 DNA exoneration cases, where mistaken identification played a role in more than 75% of the cases [90]. Although soft biometrics is concerned with description rather than eyewitness identification, the issues associated with memory must be considered.

Memory decay can be caused by time delays and/or interference. The passage of time was originally thought to be the sole cause of memory decay, as time passed the ability to recall memories reduced. Ebbinghaus proposed the forgetting curve [91] which states that memories are forgotten at an exponential rate based on time passed since the memory was encoded, this is widely accepted within the psychology community [92]. Interference is now believed to also contribute towards memory decay. Retroactive interference occurs when newly learnt information hinders previously learnt information being recalled [93].

Research has shown that the method in which eyewitnesses are questioned can have an affect on the accuracy of the resulting recalled information [94, 95, 96]. This project presents a unique opportunity to explore whether different forms of description could affect the accuracy of human descriptions after memory decay. This chapter represents introductory work into this interesting and novel question. We aim to explore how both interference and time delays affect comparative and absolute human descriptions.

Section 6.2 analyses data from an experiment studying interference and time delays of 2-10 minutes. Longer time delays of an hour are investigated in section 6.3. Finally, section 6.4 will conclude the results and discuss their implications.

6.2 Short Time Delays and Interference

In section 3.3.2 we introduced an experiment conducted to obtain bodily comparisons. The experiment was split into two parts. The first part explored the benefits of comparative annotations in ideal settings (subjects being compared were both visible to the annotator), whilst the second part investigated the affects of time delay and interference on the quality of the comparisons. In this section the results from the second half of this experiment will be explored.

The second part of the experiment was conducted as follows. A continuous set of videos showing one of the ten targets walking, was presented to the user. The videos continued until the user was ready to begin. These videos were the only opportunity that the user had to observe the target, simulating a limited exposure. The user was then asked to compare five subjects (out of forty) with the target. When comparing the subjects the user was prevented from viewing the target again. Comparing five subjects sequentially allowed us to observe how the accuracy of the comparisons changed over time. Furthermore, by showing multiple subjects to the annotator we could simulate interference and study its effects. Finally the user was asked to describe the target using absolute descriptions (using the traits and terms introduced in [13]), discovering the effects of memory on absolute human descriptions. The time between viewing the target and completing the six annotations (five comparisons and a single absolute description) was on average twelve minutes.

Initial analysis compared the comparative annotations to absolute categorical labels gathered in an ideal setting [13]. Samangooei and Nixon [13] collected descriptions of each subject from multiple users (on average 9 users) which reduced the influence of subjective errors. Figure 6.1 shows the similarity between comparative and absolute annotations, calculated using equation 3.2. The five time steps represent the five subjects shown sequentially to the user. Each subject-target comparison took on average two minutes. Figure 6.1 shows that the similarity between comparative and absolute annotations was alike for both continuous and limited target exposures. 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 - this experiment has shown that short time delays and interference did not significantly affect the comparative annotations when evaluated against previously collected absolute descriptions.

Figure 6.2 shows the accuracy of the absolute labels gathered at the end of the second

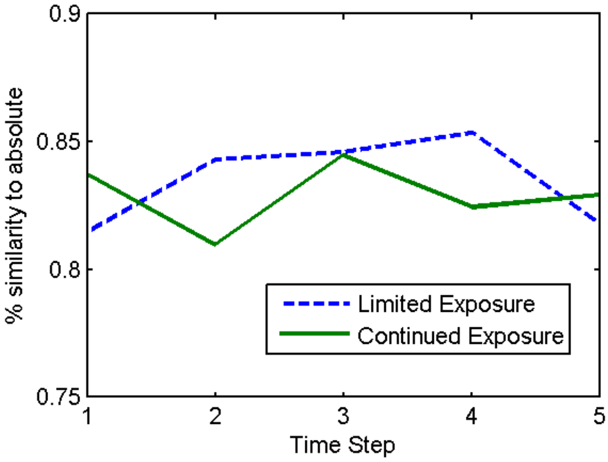


Figure 6.1: The similarity of comparative and categorical annotations. Time steps represent the five subjects compared to each target

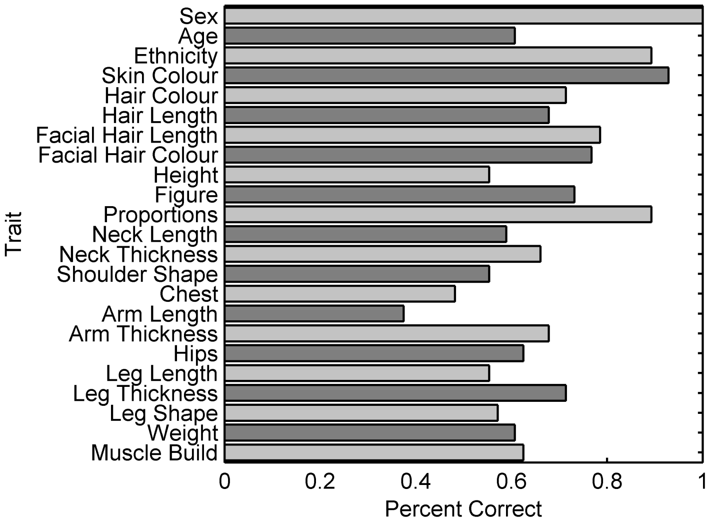


Figure 6.2: Accuracy of categorical labels after limited exposure

experiment. The annotations described the target, who had not been seen for ten minutes on average. These descriptions were compared to annotations of the same subject collected by Samangooei and Nixon [13]. An annotation was deemed to be correct if it matched the mode of the labels used to previously describe the subject. Errors of 32% were present within the delayed annotations when compared to the previously obtained labels. This indicates that absolute categorical labels are prone to error after relatively short time periods.

Analysis of the delayed comparisons must also be extended to the relative measurements produced from the annotations. Using the Elo rating system detailed in section 4.1.3, the delayed comparisons were converted to relative measurements. Figure 6.3 shows the relationship between the relative and actual height measurements. The comparisons obtained after a limited exposure to the target exhibit a slightly weaker correlation

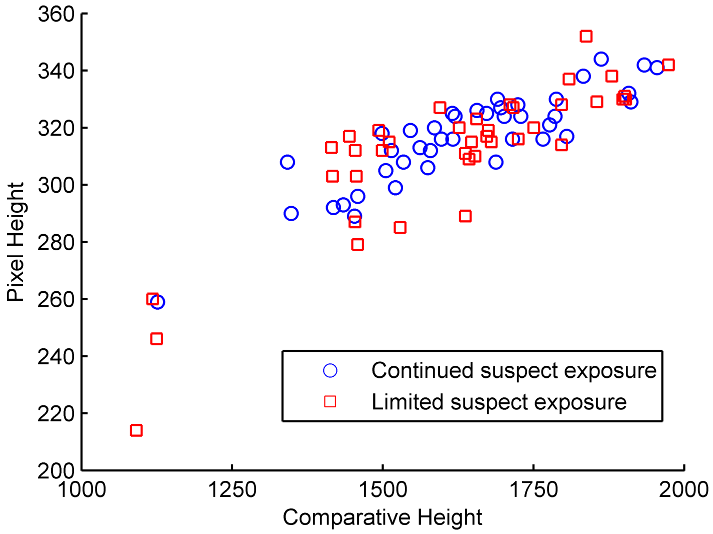


Figure 6.3: The relationship between pixel height and relative height

with the pixel height (0.85) when compared to the results obtained with a continued exposure (0.88). Although the correlation is weaker, the resulting relative measurements still represent the actual pixel height of the subjects.

The results within this section show great promise for the resilience of comparisons after short time delays and interference.

6.3 Long Time Delays

An experiment was conducted to explore memory effects over long time delays. A continuous video of a single target walking (shown in figure 6.4) was presented to a class of 55 psychology students (who did not take part in any previous experiments and were not aware that they would be included in an experiment providing descriptions of people). The target was chosen randomly from the Soton gait dataset. The video was projected onto a whiteboard for roughly 2 minutes. The students were only requested to look at the person walking and were not told that they would be required to describe the appearance of the target. After an hour delay (during which time the students were listening to a lecture introducing gait biometrics) each student was asked to compare the target to one subject from the Soton gait database. The comparison was made using the 16 comparative traits presented in table 3.1.

In total 55 comparisons were obtained between the target and 33 subjects from the Soton gait database. The 33 subjects used within this experiment were selected based on the number of previously obtained comparisons with the target (3.8 comparisons on average). The previously obtained comparisons had been given in ideal settings, with both the target and subject visible to the annotator (see section 3.3.2 for more details),

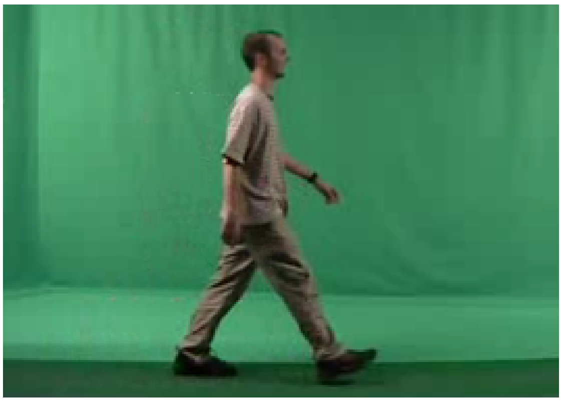


Figure 6.4: The target shown to the annotators

and hence are considered as a 'ground truth'.

The delayed comparisons were evaluated by first calculating the mode of the ground truth comparisons between the same target and subject pair. If the mode of the ground truth comparisons had the same trait label as the delayed comparison, that delayed trait annotation was considered correct.

Figure 6.5 shows the accuracy of the delayed comparisons. The average accuracy after an hour delay was found to be 54%. The absolute labels collected after a ten minute delay, shown in section 6.2, featured an accuracy of 60% (for the same 16 traits which were described with comparative labels in this experiment). This *implies* that comparative descriptions may be more resilient to memory loss than absolute labels due to the small difference in error compared to the large difference in time delay. However, the amount of interference must also be taken in to account. The absolute labels were gathered after seeing five additional subjects, whereas the comparative annotations were collected after an hour lecture on gait biometrics. Since both experiments are so different it is hard to assess the impact interference would have had on the results.

The accuracy of comparative descriptions after an hour delay has been show to be 54%. A delay of an hour before describing a suspect is quite realistic in application scenarios and this accuracy could be considered low. In this experiment we gave no indication to the annotators that they would need to later describe the target subject. This meant that many annotators did not really pay attention to the task or the target's appearance. This differs from the short time delay experiment where the annotators had been describing subjects for 10 minutes prior to the memory decay part of the experiment and were explicitly told that they would need to later describe the subject. For this reason the results obtained in this experiment are not an indication of application potential. However, the experiment does provide a performance metric which can evaluate descriptive techniques.

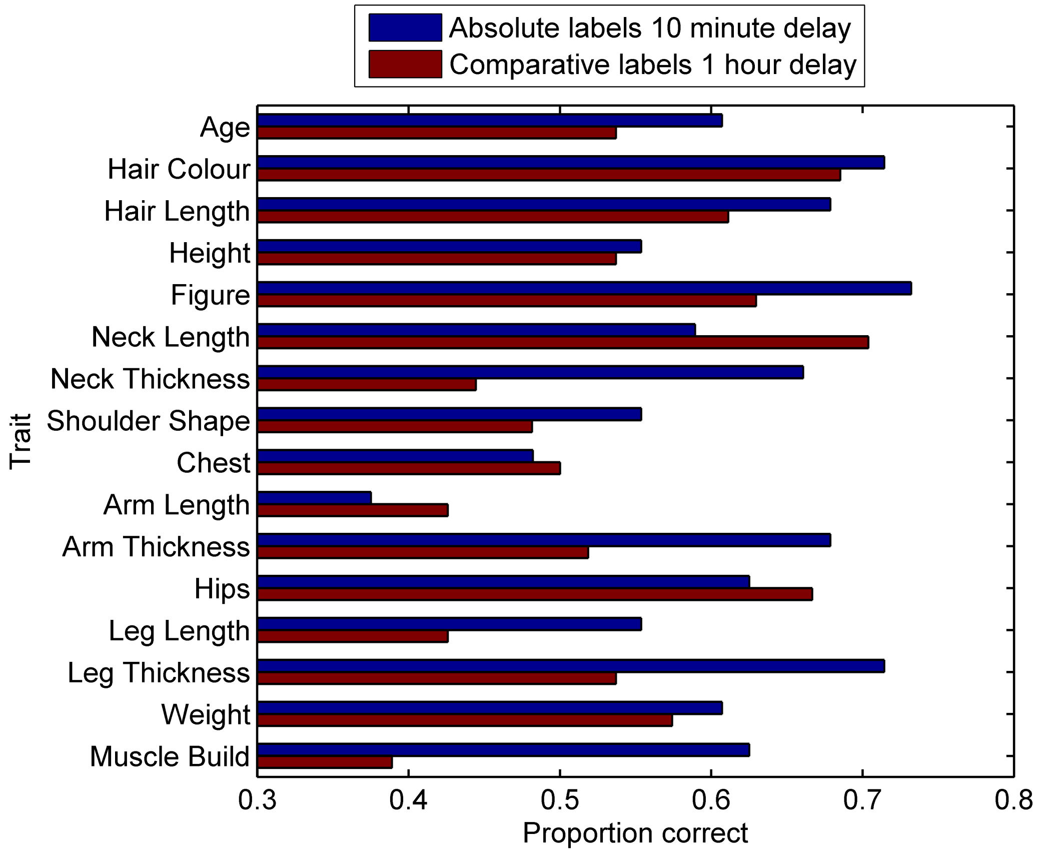


Figure 6.5: The accuracy of comparative descriptions after an hour delay compared to absolute annotations given after a 10 minute delay

6.4 Discussion and Conclusions

Memory is a key concern when obtaining descriptions from eyewitnesses. The performance of comparative descriptions after time delays and interference has been explored with two experiments. The first examined interference and time delays of 2-10 minutes, the second evaluated comparisons after time delays of an hour.

The first experiment explored interference by comparing a target to five subjects sequentially. The results showed that the accuracy of the comparisons (when compared to absolute annotations) did not decrease over the five subjects, suggesting that interference did not strongly affect the annotators. One possible explanation for these results may be the reduction in verbal overshadowing [97]. Verbal overshadowing occurs when an annotator describes an individual after exposure. The verbal description used to describe the individual overshadows the visual memory, becoming the primary source of any future descriptions or identifications - leading to reduced identification accuracy after a description has been provided. Visual comparisons could potentially avoid this problem by not absolutely describing the individual's features, instead only describing the differences between subjects. The verbal descriptions of differences between traits rather than the traits themselves may avoid overshadowing the visual memory - potentially reducing the effects of interference. Future work should study the effects of verbal overshadowing on comparative descriptions.

The results obtained in both experiments *suggest* that comparisons are more resilient to memory effects than absolute labels. These results, although promising, are far from conclusive. The lack of ground truth data made evaluation very difficult. Absolute annotations collected previously [13] provided an evaluation metric for delayed absolute descriptions. The comparative annotations collected in the short time delay experiment had to be evaluated by comparing them to absolute labels and the pixel height of the subjects being described. Furthermore, the lack of a standardized evaluation method meant that the absolute and comparative annotations were hard to directly contrast.

Future research in this area must define standard experiments which can be used to effectively compare and evaluate different descriptive methods along with 'ground truth' descriptions with which to evaluate the accuracy of delayed descriptions.

Chapter 7

Conclusions and Future Work

7.1 Conclusions

Soft biometrics is a new and exciting field of research, closing the semantic gap between human descriptions and biometrics. This thesis describes several advances to the state of the art. In chapter 3 we introduced the concept of comparative human descriptions which reduce the affects of subjectivity and self anchoring on human descriptions, resulting in increased accuracy. The correlation between measured height and described height was improved by 22% when using comparisons over absolute labels. Chapter 4 explored how comparative descriptions can be used as discriminative biometric signatures. Relative measurements were proposed and several techniques for their creation were evaluated. Recognition experiments confirmed the discriminatory capabilities of relative measurements achieving a 91% recognition accuracy with 9 bodily comparisons and a 99.8% recognition performance with 9 facial comparisons. Retrieval from video footage was discussed in chapter 5. We show how relative measurements can be automatically obtained from gait signatures, allowing video footage to be automatically searched for an individual matching a set of comparisons. Experiments exploring video retrieval accuracy demonstrated a 90% retrieval performance at rank 13, showing that video footage can be searched using descriptions. Finally, in chapter 6 we presented an introductory study into the affects of time delay and interference on different methods of description.

7.2 Future Work

7.2.1 Facial Retrieval

Facial descriptions have been shown to be discriminative and accurately described using comparative labels, achieving a 75% recognition rate with 1 comparison and increasing to 99.3% with just 5 comparisons. The recognition experiment conducted for facial descriptions relies on a database containing relative measurements. This requirement is obviously not suitable within application scenarios.

The UK police force attempt to record a photograph of every individual within the police national computer (PNC) [15]. This is generally taken within custody and features a frontal view of the individual's face [98]. The 'mugshots' are taken in controlled environments with strict requirements [98]. These images could be automatically converted to relative measurements allowing the PNC to be automatically searched for an individual which matches a set of facial comparisons. This would be approached in a similar way to the gait retrieval system shown in chapter 5. The quantity of faces to consider could be reduced by first filtering the database using the QUEST query system [15].

Although facial descriptions are not provided frequently by eyewitnesses, when they are available they have been shown to be highly discriminative. Automatically searching the PNC based on a set of facial comparisons could help to identify a suspect or at least provide a reduced set of individuals to consider.

7.2.2 Additional Comparisons

Inferred comparisons have been used throughout this project to deal with the limited data collected from volunteers. Although this has provided successful results, the inferred comparisons do contain errors which would not be seen in application environments. By collecting additional comparisons, inference would not be necessary - allowing the full benefit of comparisons to be observed.

Crowd sourcing services (for example Amazon's Mechanical Turk) could be utilized to hire individuals to compare subjects. This could potentially provide thousands of annotations for a minimal cost. This approach would also increase the diversity of the annotators, ensuring the annotation technique is suitable and accurate for any demographic.

7.2.3 Exploring Memory

Memory is a critical consideration when obtaining descriptions from eyewitnesses. In chapter 6 we introduced an initial exploration into time delays and interference. Unfortunately, firm conclusions could not be made due to the lack of data and the differences in experimental design between the multiple experiments. Future research must explore how memory affects comparative labels.

The experiment conducted by Geiselman et al. [94] could be adapted to assess the benefits of comparisons in real world scenarios and explore how memory affects comparisons.

Volunteers would be shown a video of a simulated violent crime. The video must be realistic and contain opportunities for the viewer to see the suspect. The volunteers should then be split into two groups. The first group should describe the suspect immediately after viewing, half of the volunteers within this group should use traditional descriptive methods (absolute labels and estimations) and the other half provide comparisons. The second group should be asked to return the following day to provide a description of the suspect, again half using traditional descriptions and half using comparisons. The results would explore three aspects. Firstly, whether comparisons outperform traditional descriptions in realistic scenarios. Secondly, throughout this thesis the only ground truth measurement available, for which to ascertain the accuracy of human descriptions, was height. This experiment would allow the actor portraying the suspect to be fully measured allowing the given descriptions to be evaluated using ground truth data. Finally, the effects of time delays could be assessed for both forms of description.

The effects of verbal overshadowing on visual comparisons was discussed in section 6.4. Potentially, visual comparisons may not overshadow visual memories due to their comparative nature. This would be hugely important when eyewitnesses are required to describe, then identify a suspect - a common practice when searching criminal databases and participating in identification parades. A reduction in verbal overshadowing could lead to less mistaken identifications. Experiments exploring the effects of verbal overshadowing on comparisons could determine any benefits.

7.2.4 Bodily Retrieval Improvement

Bodily retrieval experiments undertaken in this project (chapter 5) have indicated that retrieving a suspect from video footage is possible. There are obviously many ways in which the current approach could be improved.

Several model free gait signatures were considered within section 5.1.1. It can be seen that the best performances were achieved when measuring the individual's body (using the measured gait signature). Model based gait signatures may offer measurements with greater accuracy compared to those experimented with in this study. Structural models exploit the known movement of the body to accurately record properties of the individual's body [99]. This can include stride length, height and leg length [100, 101]. Limb measurements could allow relative measurements to be calculated with far greater accuracy leading to improved retrieval results. Additionally, model based signatures are generally invariant to different view points and scales which is crucial for unconstrained environments.

7.2.5 Relative Measurement Refinements

Relative measurements are key to utilizing comparisons as a biometric signature. They convert subject dependent comparisons to a single value which specifies the strength of an attribute in relation to the rest of the population. The Elo rating system offers a solution to this problem but is by no means the only approach. Before developing and experimenting with other rating systems, ground truth data is required. Relative measurements should be highly correlated with actual measurements to ensure they are accurate. In this study we have only had pixel height to validate the accuracy of different rating systems. Collecting a database of subjects each with detailed physical measurements would allow the accuracy and benefits of current and future rating systems to be ascertained.

7.2.6 Imputation for Human Comparisons

Human physical traits and appearance inherently contain structure, features frequently co-occur or have fixed relationships with other features. This occurs either due to social aspects (long hair common on females), genetics (black hair common within people of Asian descent) or the morphology of the human body (taller people more likely to have longer legs). This structure offers a basis to improve the robustness of the system in respect to missing soft feature descriptions or occluded visual features.

Extending automatic soft annotation to footage of unconstrained environments introduces problems resulting from occlusion. Visual features can be concealed by the scenery, the person's body (self occlusion) or covariates such as bags, hats and clothing. These occluded features can affect the automatic soft annotation of the biometric data, leading to inaccurate soft descriptions. By utilizing the structure present within soft biometric features we can compensate for missing visual features and correct erroneous soft descriptions. Likewise, human memory is often unreliable and can severely suffer under stressful situations. This can lead to incorrectly described features or missing feature descriptions. By exploiting the known structure it is possible to predict soft features which are uncertain or missing, refining the description.

Imputation techniques are a statistical approach used to predict missing variables. Section 7.2.6.1 demonstrates how a simple imputation technique, which exploits the known structure between features, can accurately predict missing absolute soft labels. This technique could also be applied to comparative labels allowing trait comparisons or relative measurements to be predicted. This could be beneficial for improving the accuracy of relative measurements calculated automatically from gait signatures or refining search queries when a complete description is not available.

7.2.6.1 Imputation of Absolute Descriptions

Although absolute human descriptions have been demonstrated to have less discriminatory power compared to relative descriptions, they are currently in use within the UK police national computer [15] and many other databases. A common problem faced within the UK police's criminal database is missing feature descriptions. This reduces the search possibilities available when querying for a specific individual. Using imputation and the known structure between human features, these missing descriptions could be predicted.

Two techniques have been explored to predict missing absolute soft descriptions. The experiment was designed to predict a single missing trait label using the subject's remaining trait labels. The database collected by Samangooei and Nixon [13] was used, this featured 100 subjects each having 23 soft traits (shown in table 2.2) described using a number of labels (also known as terms). Leave one out cross validation considered each of the 100 subjects in turn. Each trait was artificially removed from the test subject's feature vector. The missing trait was predicted and the correct classification rate of the rebuilt labels was analysed. A subject's feature vector is composed of a real number for each of the soft terms available to describe the 23 traits. Each value represents the percentage of people who chose that label to describe the corresponding trait.

To verify that structure is present within the soft features a correlation matrix was produced. This shows the correlation between soft traits based upon their occurrences within the SGDB. It is worth noting that some of the soft traits feature no ordering between the labels, for instance ethnicity and skin colour. When exploring the correlation of these unordered traits each possible ordering was enumerated and the maximum correlation was deemed to be the most representative of the relationship between the two traits.

Figure 7.1 shows the correlation matrix where lighter cells represent more correlated features. The most prominent relationship is that between skin colour, hair colour and ethnicity, which can be seen in the top left corner. This relationship details the genetic likelihood that people from certain ethnic backgrounds are likely to have a certain skin colour and hair colour. Another interesting region within the figure is the lower right corner which details the relationship between physical bodily attributes. The strongest correlations are present between traits describing features concerned with weight or width. An individual's weight affects the width of their limbs creating a strong relationship between thickness and weight. It was expected that traits describing lengths, like height, arm length and leg length, would be equally strongly correlated, but comparatively the correlation is weaker than that of the 'weight' features. This may be due to the variation in length descriptions. Lengths could be described absolutely, in relation to the gender or height of the individual or based on the annotator's understanding of population averages. Differences in description would result in inaccurate and varying

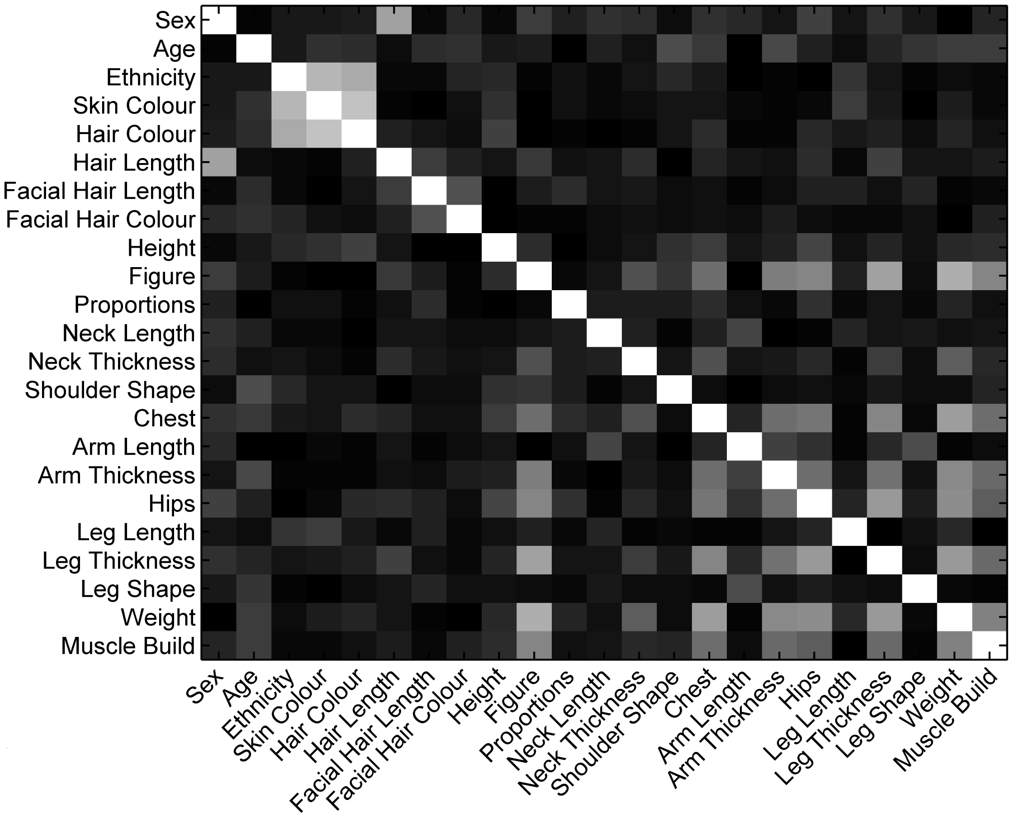


Figure 7.1: Correlation between absolute soft labels

trait labels - reducing the structure between the traits.

The first technique was developed to utilize the known correlation between different absolute soft traits. If the missing soft trait is highly correlated with another trait then it is beneficial to exploit this relationship to predict the missing term. The technique uses a similar method as the k nearest neighbour (kNN) classification technique and was inspired by work within [102]. To begin the k nearest training subjects are identified. Typically this comparison involves finding the distance between the two subjects' feature vectors. This has been modified to make use of the known correlation between traits. The similarity of each trait is weighted by the correlation between that trait and the test subject's missing trait. This favours neighbours with the same labels for traits with a strong relationship with the missing trait. The similarity between traits was determined using the Manhattan distance metric, although other distance metrics could be used. The weighted similarity between two subjects is determined as shown in equation 7.1 where X and Y are feature vectors representing the training and test subject respectively (for notation simplicity $X_{j,k}$ returns the percent of people who described trait j with label k). The trait i is missing from the test subject's feature vector and hence does not contribute towards the similarity metric. N is the total number of traits and T_m

Table 7.1: Observations of Hair Color and Skin Color

	Black	Blond	Brown	Grey	Red	Dyed
Black	1	0	0	0	0	0
Oriental	23	0	0	0	0	0
Tanned	6	0	1	0	0	0
White	1	17	54	2	1	2

Table 7.2: Percentage of observations of Hair Color and Skin Color

	Black	Blond	Brown	Grey	Red	Dyed
Black	1	0	0	0	0	0
Oriental	1	0	0	0	0	0
Tanned	0.86	0	0.14	0	0	0
White	0.01	0.22	0.7	0.03	0.01	0.03

is the total number of terms available to describe trait m. The matrix $\mathbf{C}$ contains the correlation between two traits (values range from [-1,1]). The missing trait is predicted by taking the mode of the corresponding trait within the k nearest neighbours.

$$Similarity(X,Y) = \sum_{j=1}^{N} |\mathbf{C}_{i,j}| \frac{1}{T_j} \sum_{k}^{T_j} 1 - |X_{j,k} - Y_{j,k}| \tag{7.1}$$

Correlation is adequate for determining linear relationships between traits, although it cannot determine relationships between terms and traits. Table 7.1 shows the observations of skin colour and hair colour obtained from the Soton gait database. It can be observed that some terms, for example white skin, show more variance when compared to other terms from the same trait, for example oriental skin. By determining the correlation over all terms within a trait, potentially strong ties between terms, for example oriental skin and black hair, are being lost. By observing a term's ability to predict a missing trait, better accuracy can be achieved. It can be seen that ideal terms to predict hair colour contain the least variance over their occurrences with hair colour. This important property can be used to estimate the ability of a term to predict a missing trait and can be used to weight the similarity when looking for the k nearest neighbours. If table 7.1 is converted to percentages showing the distribution of a term over the trait hair colour (see table 7.2) the variance can be easily identified. Calculating the entropy of all the elements within a row provides a measure of certainty. This shows how successful the term is at predicting the missing trait. The inversed entropy is used to weight neighbours' similarities, favouring low entropy. The similarity between two subjects is determined as shown in equation 7.3 where the matrices $\mathbf{O}$ and $\mathbf{P}$ contain the observations (table 7.1) and percentages of observations (table 7.2) respectively between terms, such that $\mathbf{O}_{k,l}$ details the observations of term k with term l. M_x is the maximum possible entropy for the x terms, this variable is used to invert the entropy.

$$H(k) = M_{T_i} + \sum_{l}^{T_i} \mathbf{P}_{k,l} log \mathbf{P}_{k,l} \tag{7.2}$$

$$Similarity(X, Y) = \sum_{j=1}^{N} \frac{1}{T_j} \sum_{k}^{T_j} H(k)(1 - |X_{j,k} - Y_{j,k}|) \tag{7.3}$$

Figure 7.2 shows results from an experiment testing both techniques' correct classification rate. It can be observed that the entropy approach achieved the most successful results, featuring a higher average correct classification rate of 79% compared to the 74% achieved using the correlation based approach.

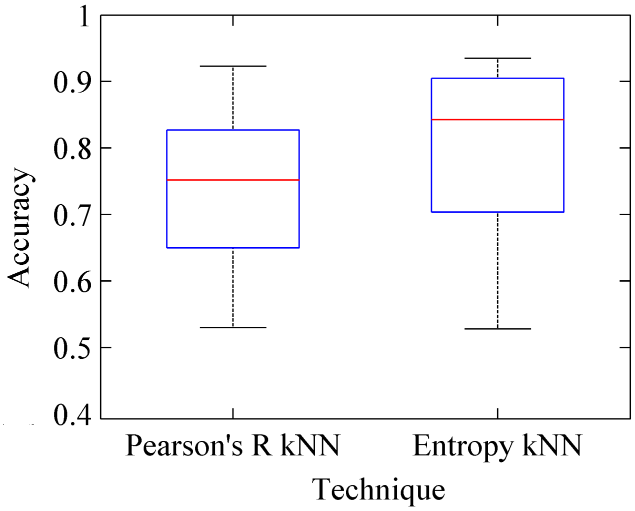


FIGURE 7.2: Results from rebuilding soft data based on remaining soft traits

Figure 7.3 shows the accuracy of rebuilding each soft trait using the entropy based approach. It can be observed that the least successful traits are those which include lengths and heights. As mentioned previously the correlation between 'length' traits is comparably weaker than the correlations between 'weight' traits. This lack of structure makes it difficult to predict the missing labels. The most successful traits are skin colour and ethnicity, this is likely due to their strong correlation with other traits allowing accurate predictions of missing data.

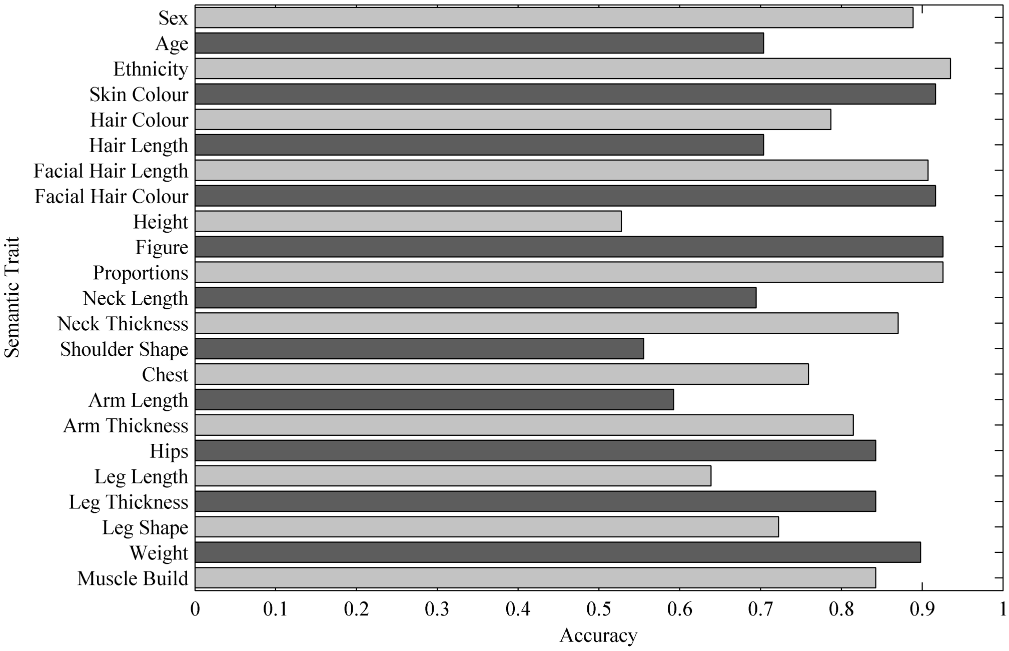


Figure 7.3: Rebuilding soft data using entropy based weighted kNN

Appendix A

Soton Gait Database

The Soton gait database (SGDB)[38] contains 118 subjects filmed in three scenarios with accompanying still images. This database is used in this project to collect human descriptions (both absolute and comparative) and within the video retrieval experiments. This section will introduce the videos and images from the SGDB which are used in this research.

The 'inside' scenario features videos of subjects walking from a side on viewpoint within a constrained environment. The filming setup for this scenario is shown in figure A.1. Each subject walks continuously around the circuit and is recorded from two viewpoints against a chroma-keyed background (an example frame from the normal camera is shown in figure A.2). Each subject was recorded walking over the central area of the circuit multiple times (between 6 and 20) either walking left to right or right to left. The gait signatures introduced in section 5.1.1 were obtained from the normal camera orientation which provides a side-on / fronto-parallel viewpoint. The bodily comparisons, introduced in section 3.3, were also obtained based on footage from the normal camera in the 'inside' scenario.

The still images within the database are high quality photos (4 megapixels) of each subject from a front and side on viewpoint, an example can be seen in figure A.3. The faces of each subject were manually extracted and centred within a 200x200 pixel image and used to obtain facial comparisons and descriptions within section 3.4.

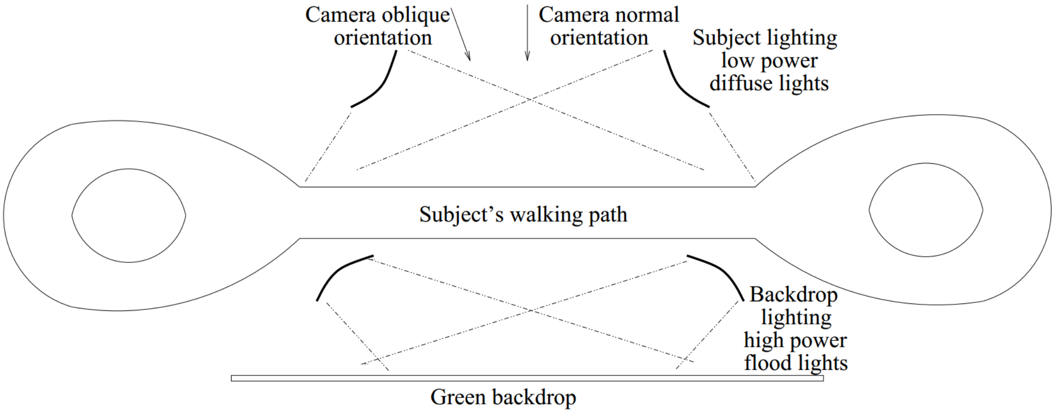

Camera oblique orientation
Camera normal orientation
Subject lighting low power diffuse lights
Subject's walking path
Backdrop lighting high power flood lights
Green backdrop

FIGURE A.1: The 'inside' scenario of the SGDB [38]

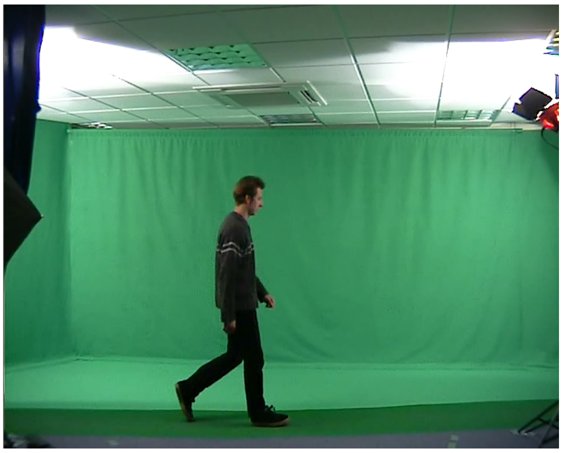


FIGURE A.2: A frame from the normal camera in the 'inside' scenario

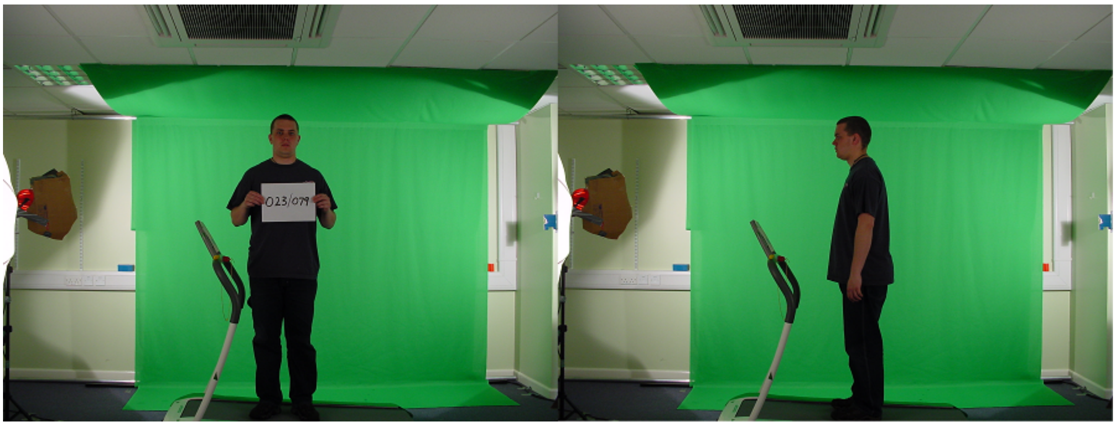


FIGURE A.3: The front and side still images within the SGDB

References

[1] A. K. Jain, S. C. Dass, and K. Nandakumar, "Soft biometric traits for personal recognition systems," in *International conference on Biometric Authentication*, 2004, pp. 731–738.

[2] L. L. Kuehn, "Looking down a gun barrel: Person perception and violent crime." *Perceptual and Motor Skills*, vol. 39, no. 3, pp. 1159–1164, 1974.

[3] P. J. Van Koppen and S. K. Lochun, "Portraying perpetrators; the validity of offender descriptions by witnesses," *Law and Human Behavior*, vol. 21, no. 6, pp. 661–685, 1997.

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

[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] M. S. Wogalter, "Effects of Post-exposure Description and Imaging on Subsequent Face Recognition Performance," in *Human Factors and Ergonomics Society Annual Meeting Proceedings*, vol. 35, no. 9. Human Factors and Ergonomics Society, 1991.

[7] J. W. Tanaka and M. J. Farah, "Parts and wholes in face recognition," *The Quarterly Journal of Experimental Psychology Section A*, vol. 46, no. 2, pp. 225–245, May 1993.

[8] M. D. MacLeod, J. N. Frowley, and J. W. Shepherd, "Whole body information: Its relevance to eyewitnesses," in *Adult eyewitness testimony: Current trends and developments*. Cambridge University Press, 1994, ch. 6.

[9] 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.

[10] E. D. Hinckley and D. Rethlingshafer, "Value judgments of heights of men by college students," *The Journal of Psychology*, vol. 31, no. 2, pp. 257–262, 1951.

[11] R. H. Flin and J. W. 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.

[12] I. A. Fahsing, K. Ask, and P. A. Granhag, “The man behind the mask: accuracy and predictors of eyewitness offender descriptions.” *Journal of Applied Psychology*, vol. 89, no. 4, p. 722, 2004.

[13] 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.

[14] National Policing Improvement Agency, *Guidance on the Management of Police Information*, 2nd ed., 2010.

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

[16] K. R. Laughery and R. H. Fowler, “Sketch artist and Identi-kit procedures for recalling faces.” *Journal of Applied Psychology*, vol. 65, no. 3, p. 307, 1980.

[17] C. D. Frowd, P. J. B. Hancock, and D. Carson, “EvoFIT: A holistic, evolutionary facial imaging technique for creating composites,” *ACM Trans. Appl. Percept.*, vol. 1, no. 1, pp. 19–39, Jul. 2004.

[18] 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)*, 2010, pp. 36–43.

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

[20] 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.

[21] 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.

[22] S. Samangooei, B. Guo, and M. S. Nixon, “The use of semantic human description as a soft biometric,” in *IEEE International Conference on Biometrics: Theory, Applications and Systems*, Sep. 2008, pp. 1–7.

[23] A. Bertillon, *Signaletic instructions including the theory and practice of anthropometrical identification.* The Werner Company, 1896.

[24] A. Bertillon, *Identification anthropométrique: instructions signalétiques*, 1893.

[25] S. A. Cole, *Suspect identities: A history of fingerprinting and criminal identification.* Harvard Univ Pr, 2002.

[26] H. T. F. Rhodes, *Alphonse Bertillon, father of scientific detection.* Abelard-Schuman, 1956.

[27] R. D. Olsen, "A fingerprint fable: The Will and William West case," *Identification News*, vol. 37, no. 11, 1987.

[28] A. A. Ross, K. Nandakumar, and A. K. Jain, *Handbook of multibiometrics.* Springer, 2006.

[29] A. K. Jain, S. C. Dass, and K. Nandakumar, "Can soft biometric traits assist user recognition?" in *Proceedings of SPIE*, vol. 5404, 2004, pp. 561–572.

[30] X. Lu and A. K. Jain, "Ethnicity identification from face images," in *Proceedings of SPIE*, vol. 5404, 2004, pp. 114–123.

[31] H. Ailisto, E. Vildjiounaite, M. Lindholm, S. Makela, and J. Peltola, "Soft biometrics - combining body weight and fat measurements with fingerprint biometrics," *Pattern Recognition Letters*, vol. 27, no. 5, pp. 325–334, 2006.

[32] G. L. Marcialis, F. Roli, and D. Muntoni, "Group-specific face verification using soft biometrics," *Journal of Visual Languages & Computing*, vol. 20, no. 2, pp. 101–109, Apr. 2009.

[33] A. K. Jain and U. Park, "Facial marks: Soft biometric for face recognition," in *IEEE International Conference on Image Processing*, Nov. 2009, pp. 37–40.

[34] A. Dantcheva, J. Dugelay, and P. Elia, "Soft biometrics systems: Reliability and asymptotic bounds," in *BTAS*, 2010, pp. 1–6.

[35] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *ACM Computing Surveys (CSUR)*, vol. 35, no. 4, pp. 399–458, 2003.

[36] M. S. Nixon and J. N. Carter, "Automatic Recognition by Gait," *Proceedings of the IEEE*, vol. 94, no. 11, pp. 2013–2024, 2006.

[37] K. Sridharan, S. Nayak, S. Chikkerur, and V. Govindaraju, "A probabilistic approach to semantic face retrieval system," in *Audio-and video-based biometric person authentication.* Springer, 2005.

[38] 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.

[39] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American society for information science*, vol. 41, no. 6, pp. 391–407, 1990.

[40] D. A. Reid and M. S. Nixon, "Imputing Human Descriptions in Semantic Biometrics," in *Proceedings of the 2nd ACM workshop on Multimedia in forensics, security and intelligence*, 2010, pp. 25–30.

[41] S. Samangooei, "Semantic biometrics," Ph.D. dissertation, University of Southampton, 2010.

[42] A. Dantcheva, C. Velardo, A. DAngelo, and J. Dugelay, "Bag of soft biometrics for person identification," *Multimedia Tools and Applications*, vol. 51, no. 2, pp. 739–777, Jan. 2011.

[43] D. Adjeroh, D. Cao, M. Piccirilli, and A. Ross, "Predictability and Correlation in Human Metrology," in *Proc. of IEEE International Workshop on Information Forensics and Security*, 2010, pp. 1–6.

[44] K. Niinuma, U. Park, and A. K. Jain, "Soft Biometric Traits for Continuous User Authentication," *IEEE Transactions on Information Forensics and Security*, vol. 5, no. 4, pp. 771–780, Dec. 2010.

[45] A. Bhattacharyya, "On a measure of divergence between two statistical populations defined by their probability distributions," *Bulletin of the Calcutta Mathematical Society*, vol. 35, pp. 99–109, 1943.

[46] C. W. W. Webster, "CCTV policy in the UK: reconsidering the evidence base," *Surveillance & Society*, vol. 6, no. 1, p. 10, 2009.

[47] F. Helten and B. Fischer, "What do people think about CCTV? Findings from a Berlin survey," *Urban Eye*, vol. 13, pp. 1–52, 2004.

[48] S. Denman, C. Fookes, A. Bialkowski, and S. Sridharan, "Soft-Biometrics: Unconstrained Authentication in a Surveillance Environment," *Digital Image Computing: Techniques and Applications*, pp. 196–203, 2009.

[49] D. A. Vaquero, R. S. Feris, D. Tran, L. Brown, A. Hampapur, and M. Turk, "Attribute-based people search in surveillance environments," in *Applications of Computer Vision (WACV), 2009 Workshop on.* IEEE, 2009.

[50] J. Thornton, J. Baran-Gale, D. Butler, M. Chan, and H. Zwahlen, "Person attribute search for large-area video surveillance," in *Technologies for Homeland Security (HST), 2011 IEEE International Conference on.* IEEE, 2011.

[51] M. Demirkus, K. Garg, and S. Guler, "Automated person categorization for video surveillance using soft biometrics," in *Biometric Technology for Human Identification VII*, 2010.

[52] S. Denman, A. Bialkowski, C. Fookes, and S. Sridharan, “Identifying customer behaviour and dwell time using soft biometrics,” *Video Analytics for Business Intelligence*, 2012.

[53] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar, “Attribute and Simile Classifiers for Face Verification,” in *ICCV*, 2009, pp. 365–372.

[54] G. Wang, D. Forsyth, and D. Hoiem, “Comparative object similarity for improved recognition with few or no examples,” in *CVPR*. IEEE, 2010, pp. 3525–3532.

[55] D. Parik and K. Grauman, “Relative Attributes,” in *ICCV*, 2011.

[56] T. Joachims, “Optimizing search engines using clickthrough data,” in *SIGKDD*. ACM, 2002, pp. 133–142.

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

[58] 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.

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

[60] S. L. Sporer, “Person descriptions as retrieval cues: Do they really help?” *Psychology, Crime & Law*, vol. 13, no. 6, pp. 591–609, Dec. 2007.

[61] J. Kabzińska and A. Niedźwieńska, “The effect of providing descriptions of perpetrators on their identification by eyewitnesses and investigative bodies.” *Problems of Forensic Sciences*, no. 84, pp. 326–335, 2011.

[62] C. C. Gordon, T. Churchill, C. E. Clauser, B. Bradtmiller, and J. T. McConville, “Anthropometric survey of US Army personnel: Summary statistics, interim report for 1988,” DTIC Document, Tech. Rep., 1989.

[63] L. L. Thurstone, “A law of comparative judgment.” *Psychological Review; Psychological Review*, vol. 34, no. 4, p. 273, 1927.

[64] J. Arbuckle and J. H. Nugent, “A general procedure for parameter estimation for the law of comparative judgement,” *British Journal of Mathematical and Statistical Psychology*, vol. 26, no. 2, pp. 240–260, 1973.

[65] K. Tsukida and M. R. Gupta, “How to Analyze Paired Comparison Data,” DTIC Document, Tech. Rep., 2011.

[66] R. A. Bradley and M. E. Terry, “Rank analysis of incomplete block designs: I. The method of paired comparisons,” *Biometrika*, vol. 39, no. 3/4, pp. 324–345, 1952.

[67] R. D. Luce, *Individual choice behavior.* John Wiley, 1959.

[68] M. Zander, *The Police and Criminal Evidence Act 1984.* Sweet & Maxwell, 2005.

[69] M. P. Murray, "Gait as a total pattern of movement: Including a bibliography on gait," *American Journal of Physical Medicine & Rehabilitation*, vol. 46, no. 1, p. 290, 1967.

[70] N. F. Troje, C. Westhoff, and M. Lavrov, "Person identification from biological motion: Effects of structural and kinematic cues," *Attention, Perception, & Psychophysics*, vol. 67, no. 4, 2005.

[71] S. V. Stevenage, M. S. Nixon, and K. Vince, "Visual analysis of gait as a cue to identity," *Applied Cognitive Psychology*, vol. 13, no. 6, 1999.

[72] I. Bouchrika and M. S. Nixon, "Exploratory factor analysis of gait recognition," in *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on.* IEEE, 2008.

[73] J. Rose and J. Gamble, *Human Walking*, 2nd ed. Williams & Wilkins, 1994.

[74] Z. Liu and S. Sarkar, "Simplest representation yet for gait recognition: Averaged silhouette," in *ICPR*, vol. 4. IEEE, 2004, pp. 211–214.

[75] R. T. Collins, R. Gross, and J. Shi, "Silhouette-based human identification from body shape and gait," in *Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on.* IEEE, 2002.

[76] G. Veres, L. Gordon, J. N. Carter, and M. S. Nixon, "What image information is important in silhouette-based gait recognition?" in *IEEE Computer Vision and Pattern Recognition conference*, 2004, pp. 776–782.

[77] L. Wang, T. Tan, H. Ning, and W. Hu, "Silhouette analysis-based gait recognition for human identification," *TPAMI*, vol. 25, no. 12, pp. 1505–1518, 2003.

[78] A. Y. Johnson and A. F. Bobick, "A multi-view method for gait recognition using static body parameters," *Lecture Notes in Computer Science*, pp. 301–311, 2001.

[79] A. Criminisia, A. Zisserman, L. Van Gool, S. Bramble, and D. Compton, "New approach to obtain height measurements from video," in *Proceedings of SPIE- The International Society for Optical Engineering*, vol. 3576, 1999.

[80] J. Lee, E. D. Lee, H. O. Tark, J. W. Hwang, and D. Y. Yoon, "Efficient height measurement method of surveillance camera image," *Forensic science international*, vol. 177, no. 1, 2008.

[81] V. N. Vapnik, *The nature of statistical learning theory.* Springer, 1995.

[82] B. E. Boser, I. M. Guyon, and V. N. Vapnik, “A training algorithm for optimal margin classifiers,” in *Proceedings of the fifth annual workshop on Computational learning theory.* ACM, 1992.

[83] A. Aizerman, E. M. Braverman, and L. I. Rozoner, “Theoretical foundations of the potential function method in pattern recognition learning,” *Automation and Remote Control*, vol. 25, pp. 821–837, 1964.

[84] C. Cortes and V. Vapnik, “Support-vector networks,” *Machine learning*, vol. 20, no. 3, 1995.

[85] A. Smola and V. Vapnik, “Support vector regression machines,” *Advances in neural information processing systems*, vol. 9, 1997.

[86] T. Hastie, R. Tibshirani, G. Sherlock, M. Eisen, P. Brown, and D. Botstein, “Imputing missing data for gene expression arrays,” Division of Biostatistics, Stanford University, Tech. Rep., 1999.

[87] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman, “Missing value estimation methods for DNA microarrays,” *Bioinformatics*, vol. 17, no. 6, pp. 520–525, 2001.

[88] Joint Health Surveys Unit of the National Centre for Social Research and Department of Epidemiology and Public Health at the UCL Medical School, “Health Survey for England - 2010,” 2011.

[89] A. Rattner, “Convicted but innocent: Wrongful conviction and the criminal justice system.” *Law and Human Behavior; Law and Human Behavior*, vol. 12, no. 3, p. 283, 1988.

[90] K. Ask and P. A. Granhag, “Perception of line-up suggestiveness: effects of identification outcome knowledge,” *Journal of Investigative Psychology and Offender Profiling*, vol. 7, no. 3, pp. 214–230, 2010.

[91] H. Ebbinghaus, *Memory: A contribution to experimental psychology.* Teachers college, Columbia university, 1913.

[92] S. M. Kassin, V. A. Tubb, H. M. Hosch, and A. Memon, “On the ’general acceptance’ of eyewitness testimony research: A new survey of the experts,” *American Psychologist*, vol. 56, no. 5, p. 405, 2001.

[93] K. A. Deffenbacher, B. H. Bornstein, and S. D. Penrod, “Mugshot exposure effects: Retroactive interference, mugshot commitment, source confusion, and unconscious transference,” *Law and Human Behavior*, vol. 30, no. 3, pp. 287–307, 2006.

[94] R. E. Geiselman, R. P. Fisher, D. P. MacKinnon, and H. L. Holland, “Enhancement of eyewitness memory with the cognitive interview,” *The American journal of psychology*, pp. 385–401, 1986.

[95] E. F. Loftus, “Leading questions and the eyewitness report,” *Cognitive Psychology*, vol. 7, no. 4, pp. 560–572, 1975.

[96] M. MacLeod, “Retrieval-induced forgetting in eyewitness memory: forgetting as a consequence of remembering,” *Applied Cognitive Psychology*, vol. 16, no. 2, pp. 135–149, 2002.

[97] C. S. Dodson, M. K. Johnson, and J. W. Schooler, “The verbal overshadowing effect: Why descriptions impair face recognition,” *Memory & Cognition*, vol. 25, no. 2, 1997.

[98] National Policing Improvement Agency, *Police Standard for Still Digital Image Capture and Data Interchange of Facial/Mugshot and Scar, Mark & Tattoo Images*, 2007.

[99] C. Yam and M. Nixon, “Model-based Gait Recognition,” *Enclycopedia of Biometrics*, 2009.

[100] A. F. Bobick and A. Y. Johnson, “Gait recognition using static, activity-specific parameters,” in *CVPR*, 2001, pp. 423–430.

[101] J. H. Yoo, M. S. Nixon, and C. J. Harris, “Extracting Gait Signatures based on Anatomical Knowledge,” in *Proceedings of BMVA Symposium on Advancing Biometric Technologies*, 2002.

[102] I. Wasito and B. Mirkin, “Nearest neighbour approach in the least-squares data imputation algorithms,” *Information Sciences*, vol. 169, no. 1-2, pp. 1–25, 2005.