UNIVERSITY OF SOUTHAMPTON

Semantic Biometrics

by

Sina Samangooei

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

in the
Faculty of Engineering, Science and Mathematics
School of Electronics and Computer Science

May 7, 2010

"...a figure strongly built and with broad shoulder, though shorter than the average of men and now stooped with age, leaning on a thick rough-cut staff as he trudged along..."

- J. R. R. Tolkien, *"The History of the Hobbit"* describing Gandalf the Grey

UNIVERSITY OF SOUTHAMPTON

Abstract

Faculty of Engineering, Science and Mathematics
School of Electronics and Computer Science

Doctor of Philosophy

by Sina Samangooei

Gait and face biometrics have a unique advantage in that they can be used when images are acquired at a distance and signals are at too low a resolution to be perceived by other biometrics. Given such situations, some traits can be difficult to extract automatically but can still be perceived semantically using human vision. It is contended that such semantic annotations are usable as soft biometric signatures, useful for identification tasks. Feature subset selection techniques are employed to compare the distinguishing ability of individual semantically described physical traits. Their identification ability is also explored, both in isolation and in the improvement of the recognition rates of some associated gait biometric signatures using fusion techniques.

This is the first approach to explore semantic descriptions of physiological human traits as used alone or to complement primary biometric techniques to facilitate recognition and analysis of surveillance video. Potential traits to be described are explored and justified against their psychological and practical merits. A novel dataset of semantic annotations is gathered describing subjects in two existing biometric datasets. Two applications of these semantic features and their associated biometric signatures are explored using the data gathered. We also draw on our experiments as a whole to highlight those traits thought to be most useful in assisting biometric recognition overall.

Effective analysis of surveillance data by humans relies on semantic retrieval of the data which has been enriched by semantic annotations. A manual annotation process is time-consuming and prone to error due to various factors. We explore the semantic content-based retrieval of surveillance captured subjects. Working under the premise that similarity of the chosen biometric signature implies similarity of certain semantic traits, a set of semantic retrieval experiments are performed using well established Latent Semantic Analysis techniques.

Abbreviations

PCA Principal Components Analysis

SVD Singular Value Decomposition

CCR Correct Classification Rate

EER Equal Error Rate

LoO Leave-one-Out

CBIR Content Based Information Retrieval

CCTV Closed Circuit Television

GAnn Gait Annotation system

PyGAnn Python Gait Annotation system

UK United Kingdom

IC Identity Code

4CIF Common Intermediate Format

COTS Commercial Off-The-Shelf

ICA Independent Component Analysis

LDA Linear Discriminant Analysis

EBGM Elastic Bunch Graph Matching

DLDA Direct Linear Discriminant Analysis

SVM Support Vector Machines

KNN K Nearest Neighbours

ANOVA Analysis Of Variance

CL-LSI Cross Language Latent Semantic Indexing

LSA Latent Semantic Analysis

LSI Latent Semantic Indexing

mAP mean Average Precision

i-mAP improved mean Average Precision

PLSA Probabilistic Latent Semantic Analysis

EM Expectation Maximisation

HIDDB Southampton Large (A) HumanID Database

TunnelDB Southampton Multibiometric Tunnel Database

MVC Model, View, Controller

WSGI Web Server Gateway Interface

ORM Object-Relational Mapping

Pearson's r Pearson's product-moment correlation coefficient

FP False Positive

FN False Negative

ROC Receiver Operator Characteristic

SDD semi-discrete decomposition

RMS Records Management System

COMPENDEX Computerised Engineering Index

Contents

Abstract ii

Abbreviations iii

Declaration of Authorship viii

Acknowledgements ix

1 Context and Contributions 1

2 Semantic Features 5
2.1 Introduction . . . 5
2.2 Background Reading . . . 6
2.2.1 Historical Anthropometry . . . 6
2.2.2 Modern Anthropometry . . . 8
2.3 Traits and Terms . . . 11
2.3.1 Traits . . . 11
2.3.2 Terms . . . 13
2.3.2.1 Discrete Metrics . . . 14
2.3.2.2 Value Metrics . . . 15
2.3.3 Semantic Biometric Terms and Traits . . . 15
2.4 Semantic Annotation . . . 15
2.5 Dataset Statistics . . . 19
2.5.1 Overall Data Composition . . . 20
2.5.2 Dataset Distributions . . . 22
2.5.3 Internal Correlations . . . 28
2.5.3.1 Subject Annotations Auto-Correlation and Self Annotation Auto-Correlation . . . 30
2.5.3.2 Self Annotations vs Ascribed Annotations . . . 36
2.6 Conclusions . . . 37

3 Semantic Biometric Fusion 38
3.1 Introduction . . . 38
3.2 Biometric Signatures . . . 40
3.2.1 Face . . . 41
3.2.2 Gait . . . 43
3.3 Biometric Fusion . . . 44

3.3.1 Feature Level . . . 45
3.3.1.1 Examples . . . 46
3.3.2 Score Level . . . 47
3.3.2.1 Density-Based Score Fusion . . . 47
3.3.2.2 Classifier Score Fusion . . . 48
3.3.2.3 Transformation-Based Score Fusion . . . 48
3.3.3 Decision and Rank Level . . . 48
3.3.4 Soft Biometric Fusion . . . 49
3.4 Semantic Fusion . . . 50
3.4.1 Semantic Features . . . 50
3.4.2 Automatic Visual Features . . . 51
3.4.2.1 HIDDB Dataset . . . 52
3.4.2.2 TunnelDB Dataset . . . 54
3.5 Semantic Recognition Experiments . . . 61
3.5.1 Semantic Features Significance . . . 61
3.5.1.1 ANOVA . . . 62
3.5.1.2 Pearson's r . . . 64
3.5.2 Semantic Significance Validation . . . 66
3.5.3 Fusion Experiments . . . 67
3.5.3.1 Approach . . . 68
3.5.3.2 Results . . . 71
3.6 Conclusions . . . 77

4 Content-Based Analysis **78**
4.1 Introduction . . . 78
4.2 Latent Semantic Analysis . . . 79
4.2.1 History . . . 79
4.2.2 The Singular Value Decomposition . . . 81
4.2.3 Using the Singular Value Decomposition . . . 84
4.2.4 An Example: LSA using the SVD . . . 86
4.2.4.1 Cars, Trees and the Sun . . . 86
4.2.4.2 Example Retrieval . . . 88
4.3 Semantic Retrieval Experiments . . . 90
4.3.1 Southampton Large (A) HumanID Database (HIDDB) Gait Retrieval . . . 95
4.3.1.1 Results . . . 95
4.3.2 Southampton Multibiometric Tunnel Database (TunnelDB) Gait Retrieval . . . 98
4.3.2.1 Results . . . 98
4.3.3 TunnelDB Face Retrieval . . . 101
4.3.3.1 Results . . . 101
4.3.3.2 Compound Queries . . . 104
4.4 Conclusions . . . 105

5 Feature Significance **106**
5.1 Introduction . . . 106
5.2 Vote Gathering Procedure . . . 107

5.2.1 Correlation Analysis . . . 107
5.2.2 ANOVA and Pearson's r Ordering . . . 107
5.2.3 Retrieval Capability . . . 108
5.3 Ordering approach: Majority Voting . . . 108
5.4 Final Trait Ordering . . . 109
5.5 Conclusions . . . 113

6 Future Work **114**
6.1 Introduction . . . 114
6.2 Semantic Terms . . . 114
6.3 Practical Applications . . . 115
6.4 Trait and Term Validity . . . 115
6.5 Ground Truths . . . 116
6.6 Fusion Approaches . . . 117
6.7 Content Based Information Retrieval (CBIR) Refinement . . . 117

7 Conclusion **118**

Declaration of Authorship

I, Sina Samangooei, declare that this thesis titled, "Semantic Biometrics" and the work presented in it are my own. 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

Writing a thesis takes a long time.
Without the assistance, helpful conversations, patience and existence of certain individuals, this thesis would have taken substantially longer. I would like to thank the people without whom I could not have hoped to understand exactly what semantic biometrics are.

First and foremost I would like to thank my supervisor Prof. Mark S. Nixon for the opportunity to explore a novel problem in an interesting field and for allowing me the space to work without looking over my shoulder. His helpful guidance and refreshing pragmatism at key points throughout this research have guaranteed I have not lost my way.

The nature of this work required the careful gathering of *a lot* of human ascribed data. To the hundreds of individuals who sat through the annotation process and those who were kind enough to be annotated I extend my warmest thanks. Without their time and generosity there would be nothing to work with and nothing to analyse.

To all my friends and colleagues I extend my unending appreciation. Their fellowship, humour and confidence in me have made this project possible. My thanks go to my friends John D. Bustard and Darren A. Golbourn for the philosophical and mathematical chats, they kept my mind forever engaged. Thanks also to Jade Richards, her tireless efforts and infinite patience gave me access to key background research describing the information practise of the modern UK Police force. I'd also like to thank Jon Hare for the discussions regarding the SVD and for a technical proof reading of the final document. Helena Baser has been of great assistance. Her stylistic and grammatical corrections have been the deciding factor in this document's readability. This work would not look as it does now without her patient proof reading and helpful corrections.

Finally, I'd like to thank my parents Fazlollah and Tahereh. Their love and support are what made all this possible.

Chapter 1

Context and Contributions

In today's security conscious climate there is an increasing interest in efficient identification of humans. When close contact and subject co-operation are assured, biometric techniques using DNA, iris signature and fingerprint recognition [58] have been shown to address this need effectively. However, there is an increasing interest in human recognition when contact and subject co-operation are not assured. This is demonstrated by the recent large scale uptake of surveillance technologies such as Closed Circuit Television (CCTV), with 4 million CCTV cameras in operation in the UK in 2006 [7]. Non-contact biometrics such as gait [89], face [111] and ear [48] address the need for identification at a distance whilst automatic surveillance analysis techniques [22] [47] attempt to address the need for the analysis of large[1] video data-sets generated automatically by CCTV surveillance systems. The principal aim of this thesis is to show that human ascribed semantic descriptions of individuals witnessed at a distance can be used to improve identification and aid the retrieval of these individuals in large surveillance datasets.

The human ability to identify individuals has been shown to be consistently effective at a distance, under varying weather conditions, light conditions [116] and behavioural configurations (e.g. walking, running, various emotional states); situations which automated techniques often find challenging. Humans can easily perceive and express higher level *semantic* concepts [68] such as Sex, Race, Bulk etc. and use them for description and identification. However, human recognition has various issues itself which can impede accurate description ability, recall and subsequently recognition.

[1] 3.6GBytes of video data per hour per camera calculated using 25.5 frames per second using 704 × 576 Common Intermediate Format (4CIF) images compressed using MPEG4 (`http://www.info4security.com/story.asp?storyCode=3093501`)

For the purposes of improved identification, the potential strengths offered by automated techniques compared to human descriptions are distinct and indeed complementary. Furthermore, relating human descriptions to automatic features extracted from video sources empowers the efficient manipulation and exploration of large surveillance datasets by humans. To these ends, we explore the relationship between semantic descriptions and primary biometric sources for applications in both biometric fusion and in Content Based Information Retrieval (CBIR).

The largest portion of this work, presented in Chapter 2, is devoted to defining a set of physical traits and associated semantic terms. We concentrate on physical attributes which can be easily perceived from a distance. Using the combined results of work originating in cognitive psychology, eye witness analysis and existing practical applications we choose and justify a set of physical traits describable by a set of associated semantic terms. We also outline the development of a web based interface designed to facilitate the efficient and effective annotation of terms to traits against arbitrary biometric sources. The decisions made in the system's construction are justified against psychological considerations. We begin the exploration of our semantic annotations by discussing the content of the datasets gathered. We provide exact figures with regards to the number of individuals annotated, the number of annotators and the number of terms gathered. We also present a correlation analysis where internal structures found between semantic annotations gathered are discussed.

Using the annotations gathered against the Southampton Large (A) HumanID Database (HIDDB) and Southampton Multibiometric Tunnel Database (TunnelDB) datasets, we explore the recognition capabilities of the semantically described traits in Chapter 3. One of our main goals is the exploration of retrieval capabilities of the semantic annotations in combination with other existing automatic biometrics. To this end we provide an overview of current research in face and gait biometrics as well as biometric fusion, including an exploration of soft biometrics. After this general overview, we outline the six specific biometric signatures across the two datasets which we use in the identification experiments in Chapter 3, as well as the the retrieval experiments in Chapter 4.

Once these features are outlined, the semantic traits are explored with the goal of ordering them with regards to some metric of worth, as well as gauging their recognition capability. Firstly, we use Analysis Of Variance (ANOVA) to outline an order of significance with regards to a feature's ability to separate disparate groups. Secondly, we present a similar experiment using Pearson's product-moment correlation coefficient (Pearson's r), exploring the stability of annotations ascribed to different traits across several annotators. We use these two orderings to perform a feature subset selection to achieve a high Correct Classification Rate (CCR) and Equal Error Rate (EER) using smaller

feature subsets. Following this work, we explore the retrieval capability of the semantic annotations when compared to existing biometric techniques, both in isolation and in fusion. We perform a set of exhaustive Leave-one-Out (LoO) classification experiments and employ two simple fusion schemes: min-max normalised feature fusion and transformation score fusion. We show that, in isolation, semantic traits in the HIDDB and TunnelDB can achieve an EER of 14.66% and 15.3%. When combined with any of the more powerful visual signatures, semantic features are shown to universally perform better than the more powerful biometrics. This ranges from a small improvement of 0.01% in feature fusion with the Average Face features of the TunnelDB up to a more impressive improvement of 3.89% in score fusion with the Projected Gait Signature of the TunnelDB.

In Chapter 4 we present another application of the semantic biometric traits, this time as used in CBIR of surveillance footage. We introduce a form of Latent Semantic Analysis (LSA) which uses the Singular Value Decomposition (SVD) in a conceptually similar way to the Principal Components Analysis (PCA). The chosen approach has the ability to perform semantic retrieval of unlabelled documents, given a training set of annotated examples. Retrieval performance for each physical trait is discussed across all six biometric signatures, and for comparable reasons to the recognition results, some traits can be retrieved successfully whilst others fail entirely. We also outline how this approach could feasibly be used to annotate surveillance video with regards to the humans they contain and also how LSA techniques could be used to improve unannotated biometric identification.

In Chapter 5 we combine the notions of worth ascribed to our traits by each of the preceding chapters. We utilise two vote combination techniques and combine 15 different ordering schemes and attempt to understand which traits are most suitable for the description of individuals. We discover that traits related to global attributes of individuals portray higher significance than more granular traits. We also note that between whole body descriptions, those describing some notion of general bulk surpass specific descriptions of Limbs or body parts. This confirms findings in the existing eye witness literature.

Finally, in Chapter 6 we discuss future research directions. Firstly, to make more concrete judgements on semantics as a biometric, we recommend larger semantic datasets be collected. We also recommend an exploration into the correlation between semantic annotations and some concrete ground truth statistics of individual height, weight and appearance. In turn this will allow more concrete statements to be made with regards to the accuracy of self annotations as compared to ascribed annotations. Given the success of semantic annotation of physical traits in both retrieval and identification, an

exploration into semantic description of behaviour and action is recommended. Such semantic descriptions of dynamic aspects of human movement are more likely to compliment dynamic features of gait recognition, allowing further advantage to be taken of existing gait biometrics. In this way this new approach can be further extended.

Several papers are based on this work, they are listed chronologically below.

1. S. Samangooei and M. S. Nixon, Semantic Attributes in Gait Biometrics. *At MMKM'07: Multimedia Knowledge Management Workshop*, 2007

2. S. Samangooei, B. Guo and M. S. Nixon, The Use of Semantic Human Description as a Soft Biometric. *In BTAS'08: Proceedings of the IEEE Biometrics: Theory, Applications and Systems*, 2008

3. J. S. Hare, S. Samangooei, P. H. Lewis and M. S. Nixon. Semantic spaces revisisted: investigating the performaces of auto-annotation and semantic retrieval using semantic spaces. *In Proc. CIVR*, 2008

4. R. D. Seely, S. Samangooei, L. Middleton, J. N. Carter and M. S. Nixon, The University of Southampton Multi-Biometric Tunnel and introducing a novel 3D gait dataset. *In BTAS'08: Proceedings of the IEEE Biometrics: Theory, Applications and Systems*, 2008

5. S. Samangooei and M. S. Nixon, Performing Content-based Retrieval of Humans using Gait Biometrics. *In SAMT'08: Proceedings of Semantic and Digital Media Technologies*, 2008

6. S. Samangooei and M. S. Nixon, Performing content-based retrieval of humans using gait biometrics. *Multimedia: Tools and Applications*, 2009

7. S. Samangooei, J. D. Bustard, R. D. Seely, M. S. Nixon, J. N. Carter, *Multibiometrics for Human Identification*, Chapter 6, On Acquisition and Analysis of a Dataset Comprising of gait, ear and semantic data. To be published.

Chapter 2

Semantic Features

2.1 Introduction

The description of humans based on their physical features has been explored for several purposes including medicine [107], biometric fusion [51], eyewitness analysis [67] and human identification [49]. Descriptions gathered vary in levels of visual granularity and include both features that can be measured visibly and those that are only measurable using specialised tools. The principal aim of this thesis is to show that semantic descriptions of individuals witnessed at a distance can be used in to improve identification and aid the retrieval of individuals. To these ends, we must firstly explore the semantic terms people use to describe one another. Once these terms are outlined, the second task becomes the collection of a set of manually ascribed annotations against these terms. In isolation these terms allow the exploration of semantic descriptions as a tool for identification. To explore their capabilities in biometric fusion and automatic retrieval, these annotations must be collected against a set of individuals in an existing biometric dataset.

In this chapter we develop a set of key semantic terms people use to describe one another at a distance. Once outlined, we introduce a set of semantic annotations made using these terms gathered against two existing biometric datasets. In Section 2.2 we start with an overview of human description, from early anthropometry, to modern usage in police evidence forms and in soft biometrics. In Section 2.3 we outline a set of key physiological traits noticeable at a distance and explore a set of semantic terms usable for their description. Once identified, we give the details of the procedures used to gather two new semantic biometric datasets in Section 2.4. These datasets are comprised of annotations of several subjects, each described by several distinct annotators across two multibiometric datasets. The exact contents of the semantic annotation datasets are

examined in Section 2.5 where we also perform correlation analysis, briefly exploring the underlying structures and other interesting facets of the gathered data.

2.2 Background Reading

In this section we provide an outline of the use of anthropometric measurements for purposes of human identification.

2.2.1 Historical Anthropometry

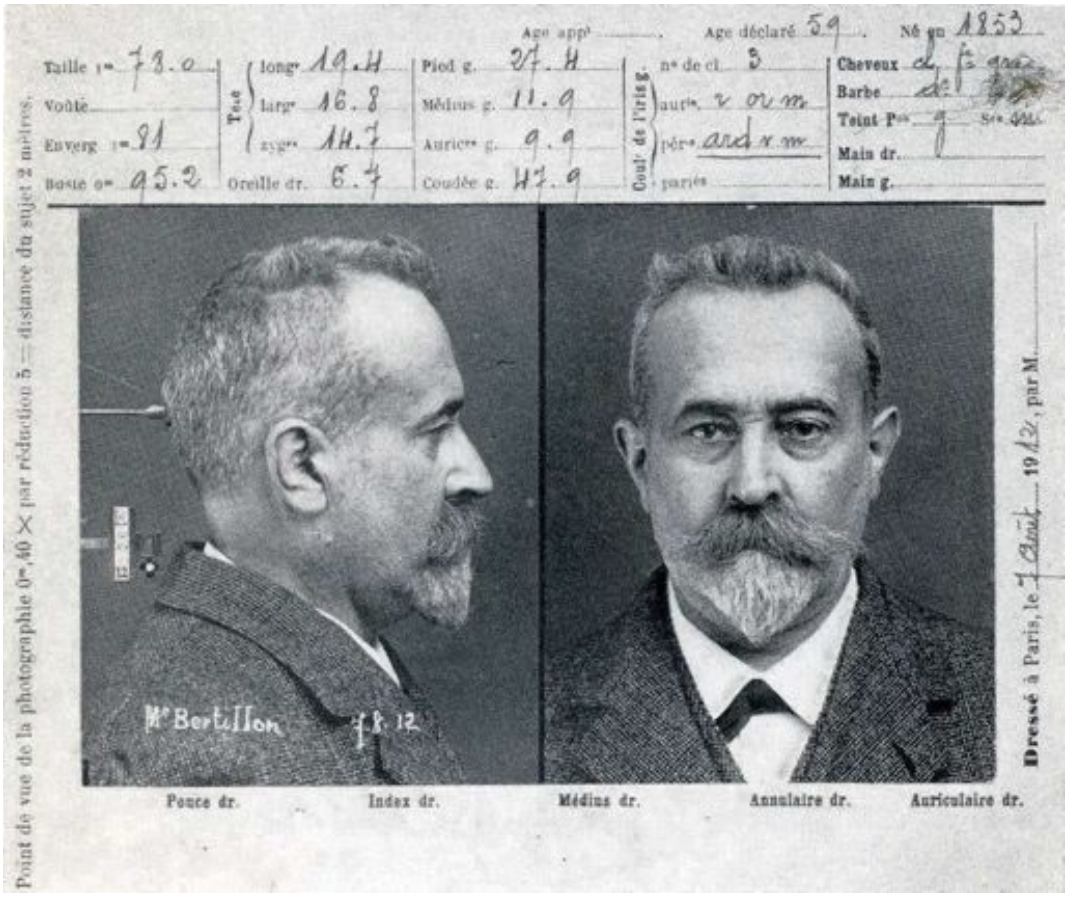


FIGURE 2.1: Example identity slate produced from *Bertillonage*. This particular slate portrays Alphonse Bertillon himself. Taken from Rhodes [102]

One of the first attempts to systematically describe people for identification based on their physiological traits was the anthropometric system developed by Alphonse Bertillon [12] in 1879. By 1809 France had abandoned early methods of criminal identification such as branding. However, no systematic method of identification was outlined as an alternative, which meant the verification of repeat offenders or confirmation of the identity of criminals was a near impossible task. Long descriptions in prose were held

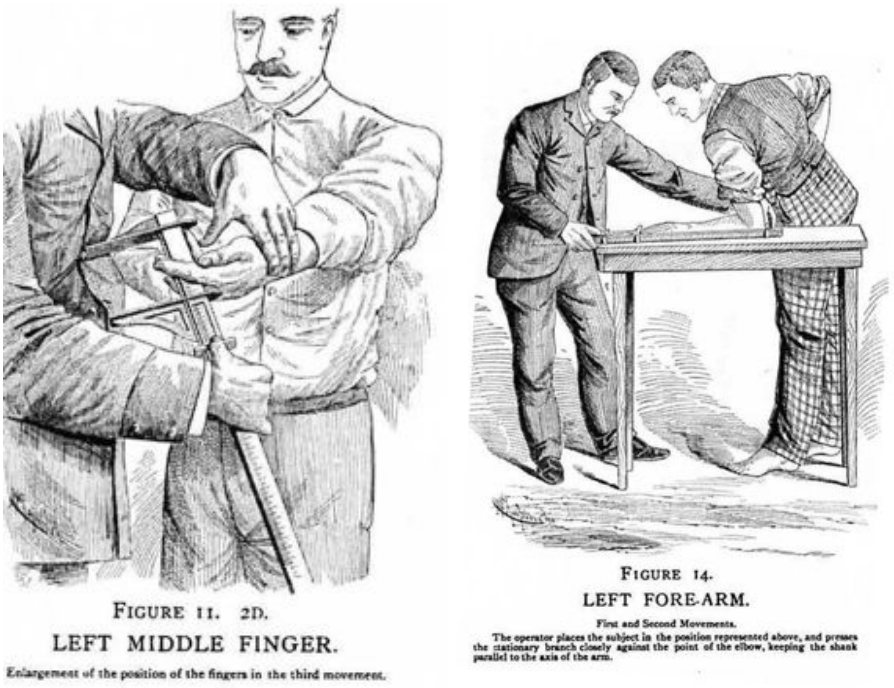


FIGURE 2.2: Two example diagrams taken from Bertillon's [12] instructional manual designed as a reference manual for police gathering Bertillonage measurements.

including semantic terms such as "Large" or "Average" to describe height and limbs. However, these descriptions proved inadequate due to subjectivity as well as to disproportionate numbers of "Average" height and "Brown" haired individuals in a given population. This, coupled with an uncontrolled lexicon, resulted in many descriptions which added nothing to identification process whatsoever. By 1840, the photography of criminals was introduced. However, the photographic techniques themselves were not standardised and, though useful for *confirmation* of identity, a photograph is of little use in *discovery* of identity given that any existing photograph collection had to be searched manually. In this landscape, Alphonse Bertillon worked as a clerk in the departments of the Prefecture of Police in 1879 making him a firsthand witness to the failings of the police identification and cataloguing system. He was therefore in an ideal position to apply his father's anthropological work to the development of a more systematic method of identifying people.

His system of **anthropometrics**, eponymously *Bertillonage*, outlined the tools and techniques for the careful measurement of:

- 10 physiological features including Length/Width of head, Length of middle and little fingers and the dimensions of the Feet, Arm, Right Ear and standing Height
- descriptions of the dimensions of the nose, eye and hair colour
- the description and location of notable scars, tattoos and other marks

The method for gathering these features was rigorously outlined in Bertillon's manual [12] along with a set of descriptive diagrams (see Fig. 2.2). The measurements for a given individual were held on separate slides along with standardised photographs of the individual. The metrics of the system were chosen primarily to be simple simple so that they could be gathered accurately. This meant measurements were taken by a trained individual, though not necessarily a skilled individual. To this end, features were chosen to allow easy identification of points to begin and to end measurement on the body. The success of *Bertillonage* came from its ability to geometrically reduce the probability of type 1 errors[1]. Though two individuals may have very similar heights, the chance of the same two having similar measurements for all the other 13 features is very unlikely. Furthermore, *Bertillonage* inherently allowed for efficient discovery of an individual's existing measurement card and therefore their identity. Cards were held in drawers where each drawer was allocated to specific range combination of each metric in a given order. This meant that once new measurements of an unidentified individual were taken the identity of the individual could be easily ascertained[2].

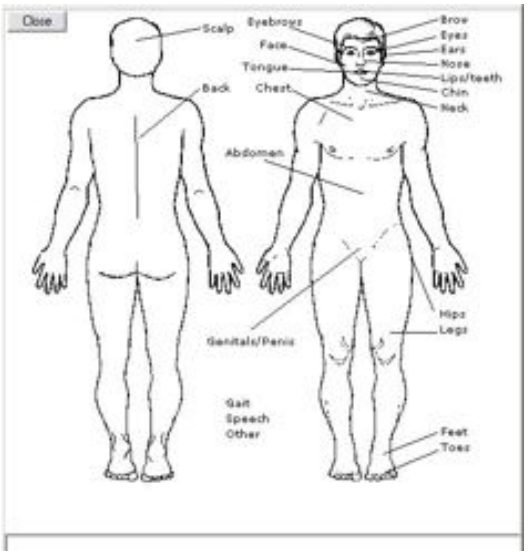
Achieving great success and popularity in France, *Bertillonage* went on to see application in the United States as well as Great Britain in the late 19th century [95]. Difficulties in cases such as Will West vs. William West [92] lead to the system being superseded by more rigorous forms of identification such as fingerprint analysis and more recently biometric analysis. In spirit, all these systems attempt to reduce the identity of an individual to a representative and measurable set of classification metrics, though none directly use descriptions of the human body as a whole.

2.2.2 Modern Anthropometry

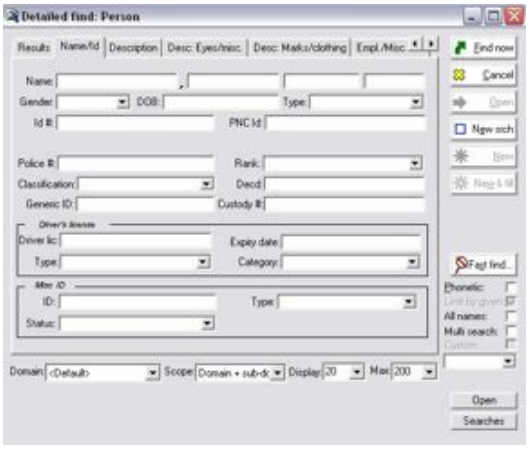
Police Records

An example of a modern use of anthropometric descriptions, both numeric and semantic, is the information repositories held by separate UK police constabularies: individually refereed to as the Records Management System (RMS). Such systems are employed to store information pertinent to criminal investigations in a given constabulary, including: vehicle description and registration information; property information and, most importantly, anthropometric suspect descriptions. The interface to any individual RMS supports semantic and vague descriptions of anthropometric features, a level of description regularly expected from witness reports and suspect descriptions as noted by Police Officers. All records in an RMS are manually added after being translated to match a

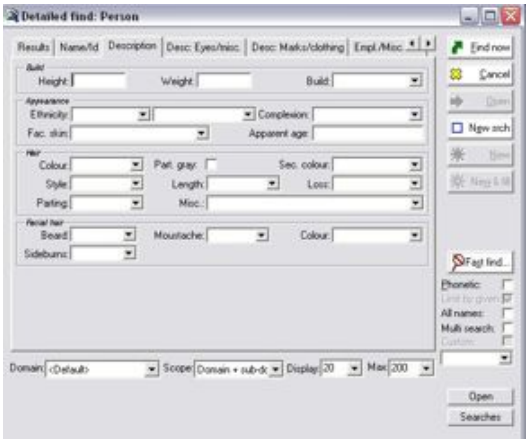
[1] error by coincidence
[2] given that the individual had previously been measured and stored



(a) Body part naming convention diagram



(b) Query Global features (e.g. Gender)



(c) Query whole body descriptions

(d) Query other features

FIGURE 2.3: Example screens from the querying system of the Hampshire Constabulary Records Management System (RMS). These images and associated descriptions are provided care of PCSO Jade Richards

prescribed lexicon for any given piece of information described. This facilitates semantic querying of the details held on the RMS and therefore anthropometric querying against the dataset. To query this system, users must use a bespoke search engine as seen in Fig. 2.3. Each field with a drop down box represents a controlled set of keywords. With regards to anthropometric descriptions of people the RMS can be queried against such features as:

- global information such as Sex and Name
- ethnic information including description of Nationality, Ethnicity and Skin Colour
- body shape information including Weight, Height and Build

- notable marks, Scars and Tattoos

Furthermore, the system holds sets of descriptions not related explicitly to the human body description, but which could none the less prove useful. These include:

- non-visual information, such as accent
- non-physical information, including address, employment and marital status
- non-permanent information, such as clothing

Most of the fields relating to humans in an RMS are categorical. For example: build can only be prescribed using keywords such as "Broad", "Slight" and "Proportionate", while Ethnic descriptions are restricted to those found in the UK census[3]. Users of the system cannot construct arbitrary queries; they are instead asked to translate search description into the closest matching terms in the subsets provided, for example a description of "young black boy" is searched for by setting ethnicity to "Black", setting gender to "Male" and choosing a relatively young apparent age. Though incredibly useful for police investigations, these systems are by their nature non-automatic. This guarantees a certain reliability and quality, but is undoubtedly expensive and also prone to human error. Furthermore, no two constabularies share the same RMS nor is there a standard for the attributes held or the keywords used to describe them. This limits the clear benefits gained from these anthropometric descriptions of individuals.

Biometric Anthropometry

In research, a recent use of anthropomorphic traits to aid primary biometric schemes was suggested by Wayman [128] in the form of filtering by Age or Gender. One of the few explorations into this approach was performed later by Nandakumar et al. [88], who used methods for automatic extraction of *soft biometric* values and fusion methods (see Section 3.3.4) on these features with primary biometrics using a Bayesian framework. Their experiments show an improvement of around 1-2% when combining ethnicity and gender traits with fingerprint signals. Other related approaches such as Zewail et al. [136] use iris colour (a soft biometric) with automatic fingerprint and iris signatures using a weighted average scheme and a Parzen Classifier. These approaches used automatically extracted soft biometrics from existing video or image signals. In behaviour analysis, several model based techniques [2] attempt the automatic extraction of individual body components as a source of behavioural information. Though the information about the

[3] `http://www.statistics.gov.uk/about/Classifications/ns_ethnic_classification.asp`

individual components is not used directly, these techniques provide some insight into the level of granularity at which body features are taken to be discernible at a distance.

In the surveillance and biometrics community, approaches that use human body descriptions do not attempt to formally outline exactly how humans identify each other. This results in ad hoc choices of descriptions with no general justification. Furthermore, apart from obviously semantic descriptions such as gender and ethnicity, most anthropometric data is inherently numerical. Little to no consideration is given to the improved identification of individuals using prose or semantic descriptions one might often find in witness descriptions. In the next section of this chapter we attempt to bridge this gap. We provide a more complete analysis of potential physiological traits humans may notice at a distance. Once outlined, we explore the associated semantic terms used in their descriptions. We offer clear justifications for the choice of these traits with respect to psychological considerations as well as practical eyewitness analysis. In doing so we outline the ground work for the analysis of semantic witness descriptions in identification and retrieval.

2.3 Traits and Terms

In this section we introduce a set of anthropometric traits and associated semantic terms suitable for the description of humans at a distance. The traits selected for description are justified on their psychological merits and an appropriate constrained set of semantic terms are outlined for each trait. The datasets discussed in future sections are collected against these traits.

2.3.1 Traits

To match the advantages of automatic surveillance media, one of our primary concerns is to choose traits that are discernible by humans at a distance. To do so, we must determine which traits humans are able to *consistently* and *accurately* notice in each other and describe at a distance. The traits we discuss are grouped by similar levels of meaning, namely:

- global traits (Sex, Ethnicity etc.)
- build features that describe the target's perceived somatotype [80] (Height, Weight etc.)

- head features, an area of the body humans pay great attention to if it is visible [44] (Hair Colour, Beards etc.).

With regards to global attributes, three independent traits - Age, Race and Sex - are agreed to be of primary significance in cognitive psychology with respect to human description. For gait, humans have been shown to successfully perceive such categories using generated point light experiments [68, 119] and in other adverse viewing conditions involving limited visual cues [116].

In the eyewitness testimony research community there is a relatively well formed notion of which features witnesses are most likely to recall when describing individuals [129]. Koppen and Lochun [67] provide an investigation into witness descriptions in archival crime reports. Unsurprisingly, the most accurate and highly mentioned traits were Sex (95% of the respondents mentioned this and achieved 100% accuracy), Height (70% mention 52% accuracy), Race (64% mention 60% accuracy) and Skin Colour (56% mention, accuracy not discussed). Detailed head and face traits such as Eye Shape and Nose Shape are not mentioned as often and when they are mentioned, they appear to be inaccurate. More prominent head traits such as Hair Colour and Length are mentioned more consistently, a result also noted by Yarmey and Yarmey [135]. Descriptive features which are visually prominent yet less permanent (e.g. clothing) often vary with time and are of less interest than other more permanent physical traits.

Traits regarding build are of particular interest in our investigation having a clear relationship with gait while still being reliably recalled by eyewitnesses at a distance. Few studies thus far have attempted to explore build in any amount of detail beyond passing mention of Height and Weight. MacLeod et al. [79] performed a unique analysis on whole body descriptions using bipolar scales to define traits. There were two phases in their approach towards developing a set of descriptive build traits.

Firstly a broad range of useful descriptive traits was outlined with a series of experiments where a mixture of moving and stationary subjects were presented to a group of annotators who were given unlimited time to describe the individuals. A total of 1238 descriptors were extracted, of which 1041 were descriptions of overall physique and the others were descriptions of motion. These descriptors were grouped together (where synonymous) and a set of 23 traits generated, each formulated as a bipolar five-point scale.

Secondly the reliability and descriptive capability of these traits was gauged. Annotators were asked to watch video footage of subjects walking at a regular pace around a room and rate them using the 23 traits identified. The annotators were then split into two groups randomly from which two mean values were extracted for each subject

for each trait. Pearson's product-moment correlation coefficient (Pearson's r) was calculated between the sets of means and was used as an estimate of the reliability for each trait. Principal Components Analysis (PCA) was also used to group traits which represented similar underlying concepts. The 13 most reliable terms, the most representative of the principal components, have been incorporated into the final trait set described later.

Jain et al. [57] outline a set of key characteristics which determine a physical trait's suitability for use in biometric identification. These include: Universality, Distinctiveness, Permanence and Collectability

The choice of our physiological traits keeps these tenets in mind. Our semantic descriptions are universal in that we have chosen factors which everyone has. We have selected a set of subjects who appeared to be semantically distinct in order to confirm that these semantic attributes can be used in the best case. The descriptions are relatively permanent: overall Skin Colour naturally changes with tanning, but our description of Skin Colour has racial overtones and these are perceived to be more constant. Our attributes are easily collectible and have been specifically selected for being easily discernible at a distance by humans. However much care has been taken over procedure and definition to ensure consistency of acquisition (see Section 2.4). The final set of traits chosen can be seen at the end of this subsection in Table 2.1.

2.3.2 Terms

Having outlined the considerations made in choosing the physical traits which should be collected, the next question is how these traits should be represented. One option for their representation in our scheme is a free text description for each trait. The analysis of such data would require lexical analysis to correlate words used by different annotators. Though interesting in itself, this study is beyond the scope of this thesis. Following the example of existing soft biometric techniques, a mixture of semantic categorical metrics (e.g. Ethnicity) and value metrics (e.g. Height) could be used to to represent the traits. Humans are generally less accurate when making value judgements when compared to category judgements. Therefore we compromise by formulating all traits with sets of mutually exclusive semantic terms. This approach avoids the inaccuracies of value judgments, being more representative of the categorical nature of human cognition [80, 118, 119]. Simultaneously this approach avoids the complex synonymic analysis that would be required to correlate two descriptions if free text descriptions were gathered. With categorical metrics there is an inherent risk that none of the categories fit, either because the information is unclear or due to the presence of a boundary case where

any annotation whatsoever may feel disingenuous. For this purpose each trait is given the extra term "Unsure", allowing the user to make the ambiguity known. For reasons covered in Section 2.4 the "Unsure" annotation is also the default option for any given trait on the annotation user interface.

What remains is the selection of semantic terms which best represent the many words that could potentially be used to describe a particular trait. This task can be logically separated by considering those traits which are intuitively describable using discrete metrics and those intuitively requiring value metrics.

2.3.2.1 Discrete Metrics

Discrete metrics are those traits not describable intuitively or commonly by numerical values. Sex is the most clear cut and it splits into Male and Female.

Age is another of the primary categories used by humans during cognition. Although based on a value metric, it has been noted in the field of human developmental biology [8] that there are several key developmental stages in a human's life. The categorical terms chosen for age in our system are synthesised from these stages. We specifically take note of the higher number of categories required to describe early life when compared to later life.

Ethnicity is also of primary significance and intuitively categorical, however it is perhaps the most difficult trait for which to find a limited set of terms. There is a large corpus of work [3, 35, 101] exploring ethnic classification, each outlining different ethnic terms. These range from the use of 3 to 200, with none necessarily convergent. Our ethnic terms encompass the three categories mentioned most often and an extra two categories (Indian and Middle Eastern) matching the United Kingdom (UK) census[4].

The colours which appear throughout the human anatomy can be described by values extracted from a continuous space. Methods such as reflection spectrophotometery can be used to extract exact values of colour but are clearly inappropriate to provide terms usable by humans. Human perception and description of colour is often categorically described [43], however, Skin Colour remains a complex area of discussion, partially due to controversy about race, but also due to inherent skin colour variability due to exposure to sun. To allow agreement, Skin descriptions cannot be too detailed. The approach chosen to define skin colour is the Identity Code (IC)[5] system, using primarily racial cues to describe skin colour. Similar problems occur with Hair Colour description; our

[4] `http://www.statistics.gov.uk/about/Classifications/ns_ethnic_classification.asp` Ethnic classification

[5] `http://www.mpa.gov.uk/committees/eodb/2005/050110/08.htm` UK police IC code

descriptions avoid these issues using categories mentioned in literature [33] and existing human description methodologies [49].

2.3.2.2 Value Metrics

For other traits representable with intuitive value metrics (Lengths, Sizes etc.) bipolar scales with intermediate categories (ranging from 5 to 7) representing concepts from *Small* to *Large* are used as semantic terms. This approach closely matches human categorical perception. Annotations obtained from such approaches have been shown to correlate with measured numerical values [23]. Note that our value metrics avoid any notion of "political correctness" aiming to reduce annotator confusion.

2.3.3 Semantic Biometric Terms and Traits

Using a combination of the studies in cognitive science, witness descriptions and the work by MacLeod et al. [79] outlined in Section 2.3 we outline the set of traits we have chosen to investigate in this thesis. Following this, in Section 2.3.2 we described a strategy for the description of these traits through a set of categorical semantic descriptions. Table 2.1 shows the corpus of physiological traits and associated semantic terms generated by this investigation and used in the following sections and chapters.

2.4 Semantic Annotation

In this section we describe the process undertaken to gather a novel dataset of semantic annotations of individuals in an existing biometric dataset. We outline the design of the data entry system created to allow the assignment of manual annotations of physical attributes to individuals. Using this system, individuals in the Southampton Large (A) HumanID Database (HIDDB) and the new Southampton Multibiometric Tunnel Database (TunnelDB) datasets were annotated against recordings taken of the individuals in lab conditions. The original purpose of these recordings was the analysis of subject gait biometrics and, in the case of TunnelDB, their face and ear biometrics. We discuss the composition of these datasets in greater detail in Section 2.5, here we concentrate on the procedure undertaken to assign annotations.

Two systems were developed to gather annotations: The PHP based Gait Annotation system (GAnn), and later, the Python/Pylons based Python Gait Annotation system (PyGAnn). The collection interface was initially developed in GAnn, written in HTML and CSS for the bespoke system. This web application was designed for the

TABLE 2.1: Physical traits and associated semantic terms

Body	
Trait	**Term**
0. Arm Length	(0.1) Very Short (0.2) Short (0.3) Average (0.4) Long (0.5) Very Long
1. Arm Thickness	(1.1) Very Thin (1.2) Thin (1.3) Average (1.4) Thick (1.5) Very Thick
2. Chest	(2.1) Very Slim (2.2) Slim (2.3) Average (2.4) Large (2.5) Very Large
3. Figure	(3.1) Very Small (3.2) Small (3.3) Average (3.4) Large (3.5) Very Large
4. Height	(4.1) Very Short (4.2) Short (4.3) Average (4.4) Tall (4.5) Very Tall
5. Hips	(5.1) Very Narrow (5.2) Narrow (5.3) Average (5.4) Broad (5.5) Very Broad
6. Leg Length	(6.1) Very Short (6.2) Short (6.3) Average (6.4) Long (6.5) Very Long
7. Leg Direction	(7.1) Very Bowed (7.2) Bowed (7.3) Straight (7.4) Knock Kneed (7.5) Very Knock Kneed
8. Leg Thickness	(8.1) Very Thin (8.2) Thin (8.3) Average (8.4) Thick (8.5) Very Thick
9. Muscle Build	(9.1) Very Lean (9.2) Lean (9.3) Average (9.4) Muscly (9.5) Very Muscly
10. Proportions	(10.1) Average (10.2) Unusual
11. Shoulder Shape	(11.1) Very Rounded (11.2) Rounded (11.3) Average (11.4) Square (11.5) Very Square

Global	
Trait	**Term**
12. Weight	(12.1) Very Thin (12.2) Thin (12.3) Average (12.4) Big (12.5) Very Big
13. Age	(13.1) Infant (13.2) Pre Adolescence (13.3) Adolescence (13.4) Young Adult (13.5) Adult (13.6) Middle Aged (13.7) Senior
14. Ethnicity	(14.1) European (14.2) Middle Eastern (14.3) Indian/Pakistan (14.4) Far Eastern (14.5) Black (14.6) Mixed (14.7) Other
15. Sex	(15.1) Female (15.2) Male

Head	
Trait	**Term**
16. Skin Colour	(16.1) White (16.2) Tanned (16.3) Oriental (16.4) Black
17. Facial Hair Colour	(17.1) None (17.2) Black (17.3) Brown (17.4) Red (17.5) Blond (17.6) Grey
18. Facial Hair Length	(18.1) None (18.2) Stubble (18.3) Moustache (18.4) Goatee (18.5) Full Beard
19. Hair Colour	(19.1) Black (19.2) Brown (19.3) Red (19.4) Blond (19.5) Grey (19.6) Dyed
20. Hair Length	(20.1) None (20.2) Shaven (20.3) Short (20.4) Medium (20.5) Long
21. Neck Length	(21.1) Very Short (21.2) Short (21.3) Average (21.4) Long (21.5) Very Long
22. Neck Thickness	(22.1) Very Thin (22.2) Thin (22.3) Average (22.4) Thick (22.5) Very Thick

initial experiments used to extract annotations with the existing HIDDB. Later, as part of the TunnelDB data collection process, PyGAnn was developed to provide an

integrated interface for the dual purposes of leading a subject through the tunnel multi-biometric data acquisition process [112] and secondly gathering annotations from the user, including both self annotations and annotations of previous subjects gathered.

PyGAnn was built on a modern web development framework called Pylons [1]. Development in Pylons follows Model, View, Controller (MVC) oriented design practise as well as making extensive use of Web Server Gateway Interface (WSGI), a web framework standard used to promote a common ground for web application development. These factors mean future maintenance of the TunnelDB interface is made easier as is the integration of the user interface with the existing Python based Southampton tunnel backend [112]. Furthermore, modern database interface methodologies such as Object-Relational Mapping (ORM) are well supported in Pylons. This heavily relieves the data manipulation burden inherent with the co-ordinated use of semantic annotations with the related subjects and their biometric data samples.

Collection Interface

Both systems were used to collect semantic annotations using the web interface initially designed for the GAnn web application (See Fig. 2.4). This interface allows annotators to view all samples of an arbitrary biometric gathered from a subject as many times as they require. Annotators were asked to describe subjects by selecting semantic terms for each physical trait. They were instructed to label *every* trait for *every* subject and that each trait should be completed with the annotator's own notions of what the trait *meant*. Guidelines were provided to avoid common confusions, for example that rough overlapping boundaries for different age terms and height of an individual should be assigned absolutely compared to perceived global "Average", while traits such as Arm Length could be annotated in comparison to the subject's overall physique.

To attain an upper limit for the capabilities of semantic data we strive to assure our data is of optimal quality. The annotation gathering process was designed carefully to avoid (and allow the future study of) inherent weaknesses and inaccuracies present in human generated descriptions. The error factors that the system was designed to deal with include:

- **Memory [27] -** Passage of time may affect a witness' recall of a subject's traits. Memory is affected by variety of factors e.g. the construction and utterance of featural descriptions rather than more accurate (but indescribable) holistic descriptions. Such attempts often alter memory to match the featural descriptions.

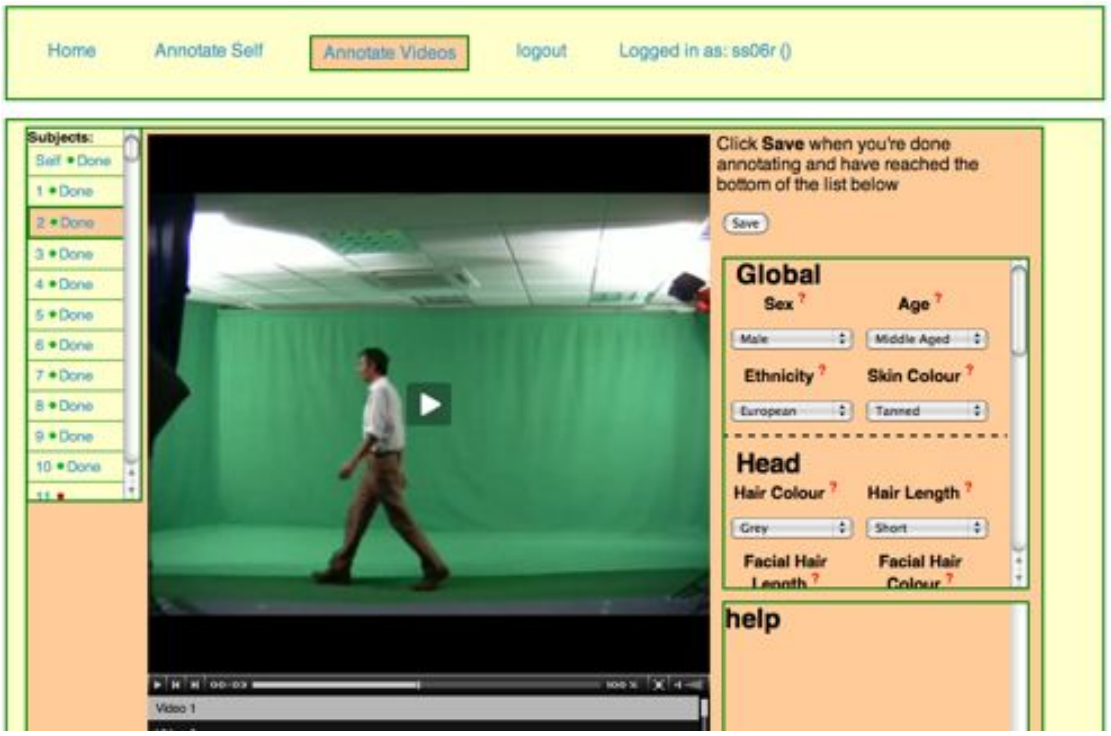


FIGURE 2.4: Example of GAnn interface

- **Defaulting [76] -** Features may be left out of descriptions in free recall. This is often not because the witness failed to remember the feature, but rather that the feature has some default value. Race may be omitted if the crime occurs in a racially homogenous area, Sex may be omitted if suspects are traditionally Male.

- **Observer Variables [32, 93] -** A person's own physical features, namely their self perception and mental state, may affect recall of physical variables. For example, tall people have a skewed ability to recognise other tall people but will have less ability when it comes to the description of shorter individuals, not knowing whether they are average or very short.

- **Anchoring [18] -** When a person is asked a question and is initially presented with some default value or even seemingly unrelated information, the replies given are often weighted around those initial values. This is especially likely when people are asked for answers which have some natural ordering (e.g. measures of magnitude)

We have designed our semantic data gathering procedure to account for all these factors. Memory issues are addressed by allowing annotators to view videos of subjects as many times as they please, also allowing them to repeat a particular video if necessary. Defaulting is avoided by explicitly asking individuals for each trait outlined in Table 2.1, this means that even values for apparently *obvious* traits are filled in and captured. This

style of interrogative description, where constrained responses are explicitly requested, is more complete than free-form narrative recall but may suffer from inaccuracy, though not to a significant degree [135]. Subject variables can never be completely removed so instead we allow the study of differing physical traits across various annotators. Users are asked to self annotate based on self perception, also certain subjects being annotated themselves provided annotations of other individuals (See Section 2.5). This allows for some concept of the annotator's own appearance to be taken into consideration when studying their descriptions of other subjects. Anchoring can occur at various points of the data capture process. We have accounted for anchoring of terms gathered for individual traits by setting the default term of a trait to a neutral "Unsure" rather than any concept of "Average". Another potential source of anchoring is that attributed by the order subjects are presented to an annotator. Seeing a string of relatively tall individuals may unfairly weight the perception of an averaged sized individual as short. We attempt to account for this by randomising the order of subjects presented to different annotators so that, overall, the descriptions reflect some notion of the true description.

In order to efficiently involve these annotations in future analysis, they are numerically represented. The exact representation scheme depends on how the data is to be used and is discussed in further detail in Chapter 3 and Chapter 4 where the annotations are formatted for use in two distinct experiments. In the final section of this chapter, we outline some statistics of the gathered datasets including their content and some structures inherent in the semantic data in isolation.

2.5 Dataset Statistics

In this section we discuss the composition of the semantic annotations gathered and the biometric datasets they were gathered against. Furthermore, in Section 2.5.3 we present some evidence for the validity of the datasets gathered by exploring their internal structure. By showing the inherent structure and correlation between annotations as well as those between annotator self annotations and the annotations they were given, we show some initial evidence that the data gathered has some regularity and thus merit. Further evidence is then presented in future chapters where the gathered data's abilities with regards to identification and retrieval, both in isolation and in combination with other biometrics, is explored.

		HIDDB	TunnelDB	Totals
Terms	Observed	20976	58023	78999
	Self	1659	4957	6616
	Of Annotators	0	31874	31874
	Total	22635	62980	85615
Partial Descriptions	Observed	334	956	1290
	Self	10	77	87
	Of Annotators	0	544	544
	Total	344	1033	1377
Complete Descriptions	Observed	625	1685	2310
	Self	63	149	212
	Of Annotators	0	904	904
	Total	688	1834	2522
Individuals Described	Observed	115	71	186
	Self	73	226	299
	Of Annotators	0	43	43
	Total	188	226	414

TABLE 2.2: Table summarising composition of the annotations gathered against two biometric datasets

2.5.1 Overall Data Composition

Southampton Large (A) HumanID Database (HIDDB) contains between 6 and 20 sample videos of 115 individual subjects each taken from a front-parallel viewpoint to extract side-on 2D gait information. The new Southampton Multibiometric Tunnel Database (TunnelDB) contains biometric samples of 227 subjects for which 10 gait sample videos from between 8 to 12 viewpoints are taken simultaneously and stored to extract 3D gait information. TunnelDB also contains high resolution frontal videos to extract face information and high resolution still images taken to extract ear biometrics. There are roughly 10 such sets of information gathered for each subject in TunnelDB

The GAnn annotation system used to collect data against the HIDDB was designed to allow annotation by anonymous annotators across the internet, though in reality the primary source of annotations came from two separate sessions involving a class of psychology students. In the first session, all the students were asked to annotate the same group of subjects, while in the second session 4 equally sized groups of subjects were allocated between the students.

The PyGAnn annotation system used to collect data against the TunnelDB was designed to gather annotations as part of the collection of an individual's multibiometric signature. After performing the experiment annotators were asked to annotate themselves and a group of 15 subjects. Due to various time constraints some annotators annotated fewer subjects but all annotators captured provided a self annotation. We selected 4 groups of 15 subjects to be annotated by progressively few annotators, aiming to maximise the number of annotators describing the same subjects while simultaneously annotating the maximum spread of subjects.

Table 2.2 shows a summary of the data collected. In this table *Terms* refers to individual semantic terms collected to describe physiological traits. *Descriptions* refer to a set of terms used to describe an individual. Here *Partial Descriptions* are those which contain terms for only a subset of the physiological traits outlined in Table 2.1, where *Complete Descriptions* contain terms for the full set of traits. Finally, *Individuals* denotes a count of the number of distinct subjects annotated, not counting repeat annotations made by separate annotators. In each of these sections, *Observed* is a count of annotations made by an annotator to an individual subject, *Self* is a count of self annotations and *Of Annotators* makes a note of annotations ascribed to annotators when they themselves were subjects. Each of these sets of annotations are explored in more detail in the following sections.

Overall, across both datasets, 85615 descriptive semantic terms were collected. Of these 6616 were self annotation terms and 78999 were ascribed to individuals by annotators. This results in 2522 complete descriptions of individual subjects within which there were 212 complete self descriptions and 2310 complete descriptions ascribed to individuals. Here, a complete description is defined as a group of terms describing all 23 physical traits of an individual subject.

In future sections, the annotations gathered are discussed in three ways:

- **Self Annotations -** Annotations an individual gave to themselves.
- **Subject Annotations -** Annotations given by an individual to a subject
- **Ascribed Annotations -** A subset of subjects in TunnelDB were in fact annotators. The annotations of these annotators are referred to as ascribed annotations

2.5.2 Dataset Distributions

In the respective datasets a total of 414 individuals were described. In this section we explore the distribution of the annotations describing individuals as well as the distribution of the self annotations gathered. In Fig. 2.5 and Fig. 2.6 we show the distributions of the annotations gathered from the TunnelDB and the HIDDB respectively. Following these graphs, in Table 2.4 and Table 2.3 we show a significance analysis of the difference between self annotations and ascribed annotations of the two datasets.

Trait Distribution Comparison

In Fig. 2.5 and Fig. 2.6 we show the normalised distribution of self and subject annotations for all traits in both datasets. An aspect of note is the distribution of measures of physical length including Height, Leg Length and Arm Length. For both datasets ascribed lengths tend towards long and average annotations meaning annotators avoid the use of the term short. This is in contrast to measurements of thickness or bulk such as Figure, Weight, Chest and Arm/Leg Thickness which display a more normal distribution. From these graphs we can also see different terms for traits such as Proportions were not used. It is possible that such traits were not perceived or the trait itself was not understood by either group of annotators, with most subjects described as having normal Proportions. Alternatively, the subjects collected may indeed portray inherently "Normal" proportions. Leg Direction seemed to enjoy similar term patterns in both datasets, a relatively unexpected result as the HIDDB did not provide the viewpoints one would expect to be necessary to make such judgements. The results for the major global features seem weighted towards Young Adult as Age; White as Ethnicity and Male as Sex. This distribution is to be expected from the datasets as both contain many subjects from the Engineering departments of the University of Southampton, UK.

Overall, we note that self annotations taken in both systems used semantic terms in ratios comparable to those used in the ascribed annotations, as well as ratios comparable to each other. This is evidence towards the idea that individuals do not wholly believe themselves to be an average; rather individuals can reasonably describe themselves as others might see them, using the full set of semantic terms others might use.

Cross-Dataset Distribution Comparison

In Table 2.4 and Table 2.3 we explore the differences in the distribution from self annotations and ascribed annotations of the two datasets. We note small disparities between the self annotations of HIDDB when compared to those of TunnelDB, though these are

FIGURE 2.5: Normalised annotation distributions of ascribed annotations of the TunnelDB dataset

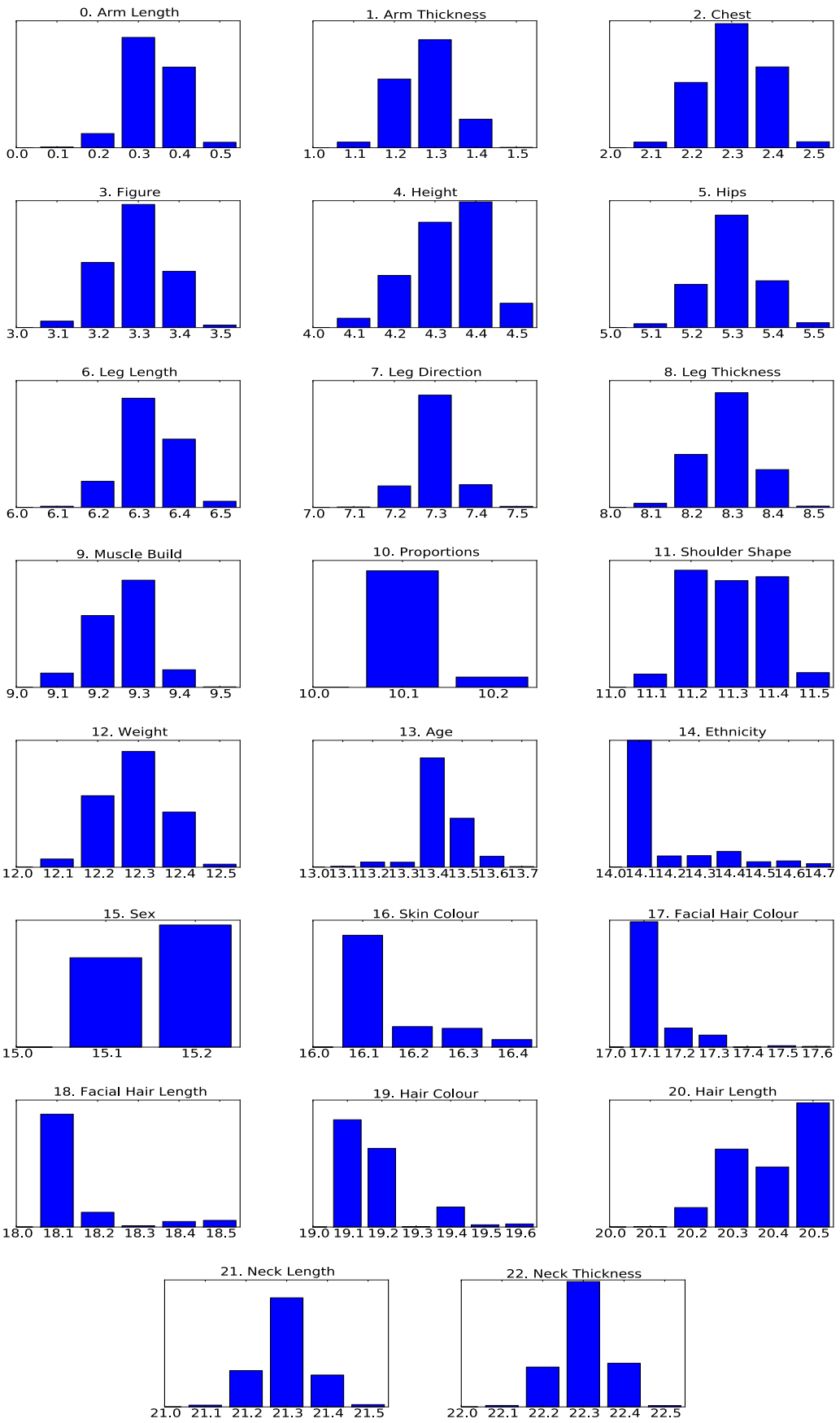


FIGURE 2.6: Normalised annotation distributions of ascribed annotations of the HIDDB dataset

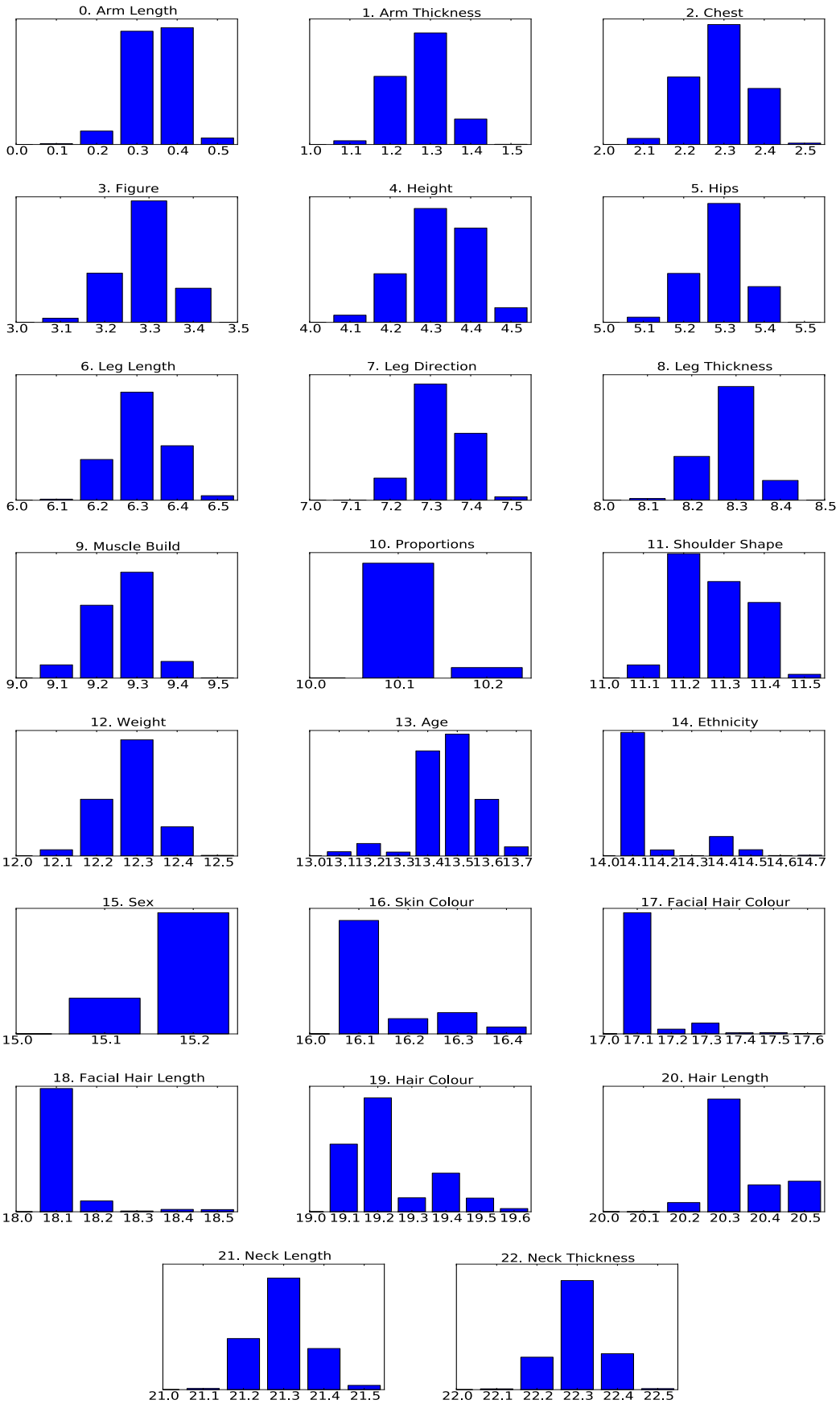


FIGURE 2.7: Normalised annotation distributions of self annotations of the TunnelDB dataset

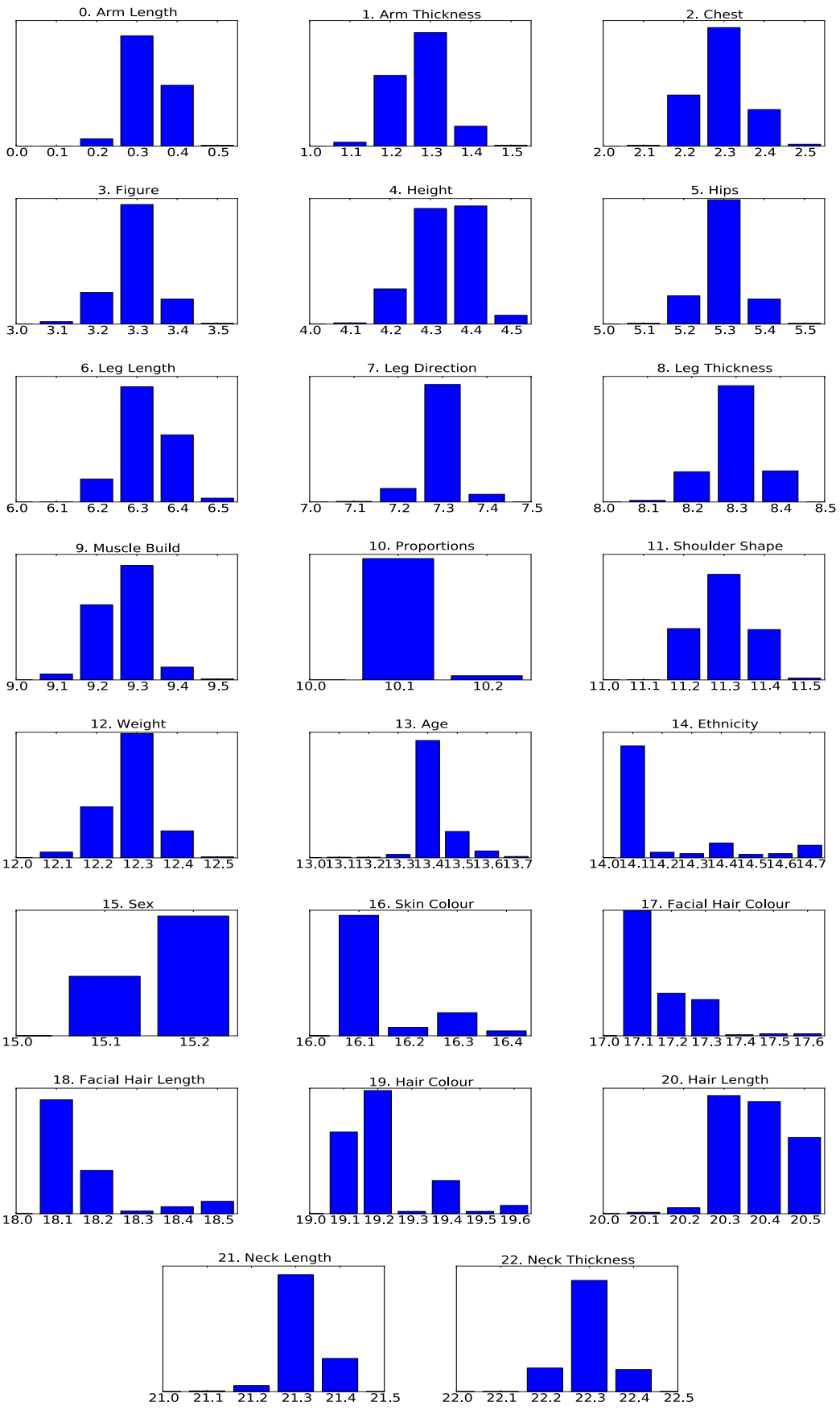
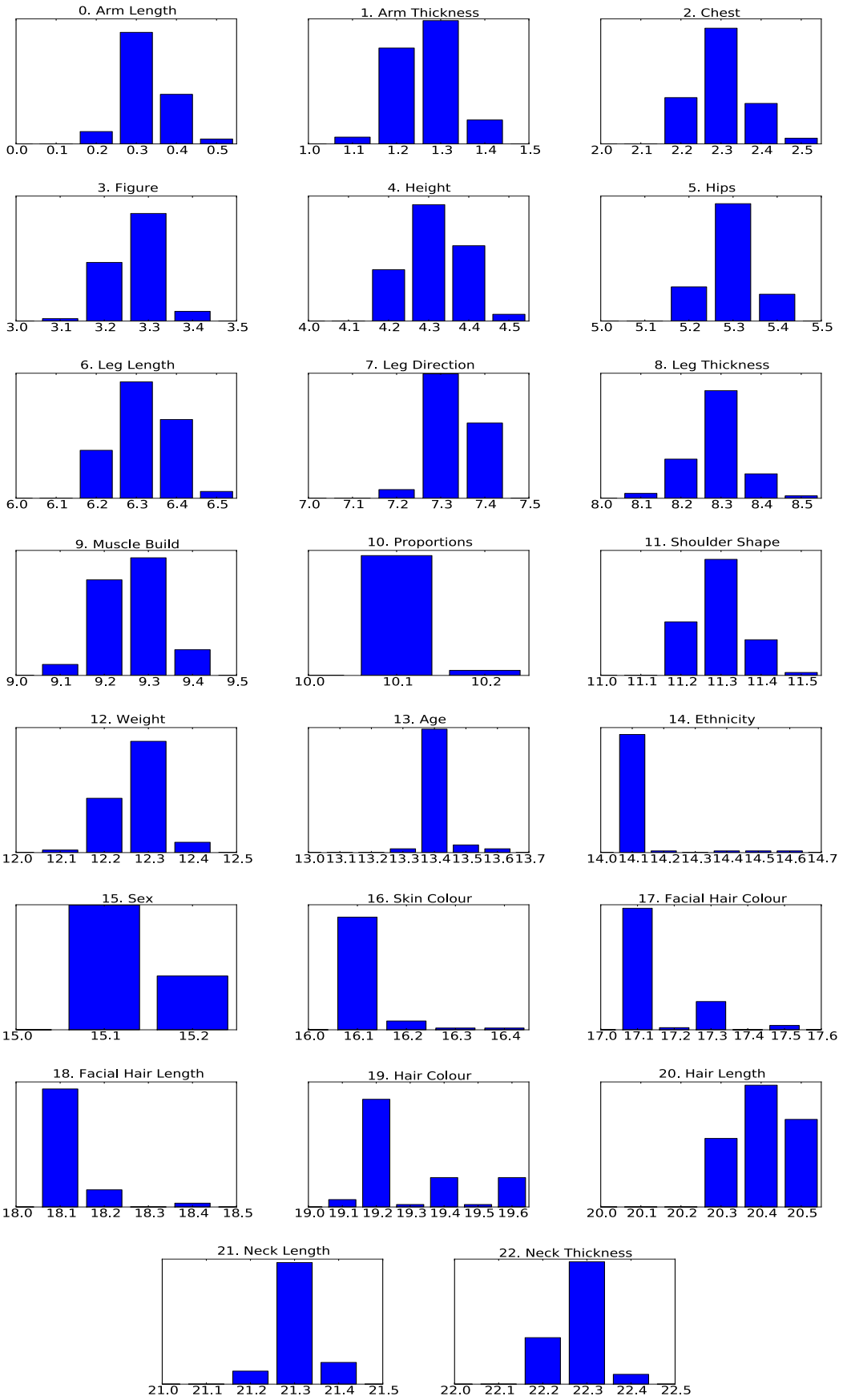


FIGURE 2.8: Normalised annotation distributions of self annotations of the HIDDB dataset



Trait	p-value
Ethnicity	0.62
Hair Colour	0.70
Hair Length	0.84
Facial Hair Length	0.84
Age	0.90
Shoulder Shape	0.91
Sex	0.92
Leg Direction	0.92

Trait	p-value
Chest	0.95
Facial Hair Colour	0.95
Leg Length	0.95
Hips	0.95
Height	0.96
Weight	0.97
Arm Length	0.97
Leg Thickness	0.98

Trait	p-value
Figure	0.98
Skin Colour	0.99
Muscle Build	1.00
Neck Thickness	1.00
Proportions	1.00
Arm Thickness	1.00
Neck Length	1.00

TABLE 2.3: The p-value of the difference in ascribed annotations between the TunnelDB and HIDDB dataset. Here we note no differences are significant to $p \leq 0.1$

Trait	p-value
Hair Colour	0.66
Facial Hair Length	0.66
Skin Colour	0.79
Sex	0.80
Facial Hair Colour	0.86
Ethnicity	0.87
Hair Length	0.92
Figure	0.93

Trait	p-value
Leg Direction	0.93
Height	0.93
Neck Thickness	0.95
Weight	0.96
Chest	0.97
Leg Thickness	0.97
Age	0.97
Neck Length	0.97

Trait	p-value
Arm Thickness	0.97
Leg Length	0.98
Shoulder Shape	0.99
Arm Length	0.99
Muscle Build	0.99
Hips	0.99
Proportions	1.00

TABLE 2.4: The p-value of the difference in self annotations between the TunnelDB and HIDDB dataset. Again we note no differences are significant to $p \leq 0.1$

mostly insignificant differences with large p-values. The p-values in these tables represent the probability of a shared distribution having created the annotation distributions across the HIDDB and TunnelDB datasets. Two extremely similar distributions will produce p-values close to 1.0 while completely dissimilar distributions will produce p-values close to 0. A more detailed explanation of Analysis Of Variance (ANOVA) can be seen in Section 3.5.1.1.

From the graphs and the relatively high p-values of ascribed annotations, we note that the individuals annotated were overall similarly distributed in appearance. More precisely, disparate groups of annotators described the different individuals in the different datasets using similar annotations. Some traits enjoy higher disparity between the datasets and therefore lower p-values; namely Ethnicity and associated attributes of Hair Colour. A special effort was made in the collection of TunnelDB to include individuals of different ethnic backgrounds in order to analyse ethnicity as a co-variate of gait; this may explain the apparent higher degree of ethnic disparity reported by annotators of the TunnelDB. Individuals with beards were specifically chosen to be annotated in the TunnelDB due to a lack of such individuals in the HIDDB. This was performed to test the ability of the facial hair related traits to some degree.

With regards to self annotations across the two datasets, both from the graphs and the relatively lower p-values in Table 2.4, we note a disparity in the ratio of self annotation of Sex. However, the graphs and p-values show comparatively similar distributions in other traits.

There were key differences in how the groups of annotators ascribed descriptions to the two datasets. Firstly, in TunnelDB annotators saw their own samples for purposes of self annotation; the annotators of HIDDB only had self perception on which to base their self annotation responses. Furthermore, although HIDDB was originally intended to be gathered from anonymous participants across the internet, in reality most of the HIDDB annotations were gathered from the attendants of a female dominated 3rd year Psychology course at the university of Southampton in two different years. This second detail explains the higher usage of Sex Female in self annotations recorded in the HIDDB dataset. The slight visible differences in the Hair Length distributions could also be attributed to a secondary effect of the difference in Sex distributions. However, other distribution differences in metrics such as Figure, Height and Arm Thickness are shown to be non-significant using a one-way ANOVA (See Table 2.4). This result is surprising as it might be expected that a group of young, primarily female individuals would present different annotation distributions in such areas as Height and Figure.

However, as the annotators in the HIDDB were not themselves annotated by other people, commenting on exactly what has caused the similarity between the two sets of self annotations lies beyond the scope of this dataset and this thesis. To measure such effects, a direct comparison of self annotations against third party annotations or some ground truth measurements must be made. Such ground truths would include numerical measurements of Weight, Height and Hair Length. If the ground truths are significantly disparate between the two datasets, then there would be an argument for a shift in perception on the part of the annotators in the HIDDB. It would show that female annotators have self-normalised and, if they themselves were annotated by others against the whole population, they would be attributed different annotations. If however the ground truth sets were not significantly different one could argue that the individuals annotating the HIDDB were in fact similarly distributed in appearance to those who annotated the TunnelDB. This would then explain the similarity in self annotation terms measured. There is some argument for this second notion in the correlations explored in Section 2.5.3.2, though even then we cannot make any conclusion with complete certainty.

2.5.3 Internal Correlations

Having outlined the overall content and distributions of the gathered datasets in the previous sections, in this section we explore notable correlations found between the various semantic annotations gathered. The goal of this section is to highlight internal structures inherent in the datasets gathered, some of which are supported by previous

studies, therefore confirming the data's validity. In this section we explore the correlation between relevant pairings of self, subject and ascribed annotations (See Section 2.5.1).

Though interesting for its own merits, these correlations could also have some useful practical applications. For example, by knowing the correlation between traits, estimated terms for missing traits could be inferred. This would result in more accurate results for a given incomplete semantic query, though such query competition could also be achieved through related techniques discussed in Chapter 4. In this section we also explore in greater detail the correlation between especially notable traits, such as Sex and Ethnicity when compared to other physical characteristics.

The following sections present correlation matrices containing the Pearson's r between each term; represented graphically. Colours closer to red represent correlation coefficients closer to 1.0 and thus a positive correlation, while colours closer to blue represent correlation coefficients closer to -1.0 and thus a negative correlation. Pale green represents 0 correlation.

We calculate the correlation coefficient between two terms using individual annotator responses of individual subjects. The calculation of Pearson's r is shown in Equation 2.1.

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}, \tag{2.1}$$

Here X and Y represent two semantic terms. In this experiment each semantic term is set to 1 if the annotation contains the term and 0 if the annotation does not. X_i and Y_i are the value ascribed to an individual in a single annotation, where there exist n annotations. Note that if $(X_i - \bar{X})(Y_i - \bar{Y}) > 0$ then X_i and Y_i lie on the same side of their respective means. In the binary case, where X and Y can only take the values 0 or 1, this denotes simultaneous annotation. Therefore, Pearson's r when applied to these semantic annotations is positive if X_i and Y_i are simultaneously present in an annotation. Furthermore, a higher correlation simultaneously represents how far an appearance of X or Y is from the mean, as well as the frequency of simultaneous appearances of X and Y across all n annotations.

In the graphs below, each solid pixel represents the Pearson's r between two terms. For clarity, individual terms are only labelled in the 6 subgraphs below each major graph, with only whole term groups representing traits labelled in the larger graphs for each group pairing. We present the larger graphs to show the general trait trends across the whole feature set, and the more detailed graphs for more in-depth analysis of specific example term pairings.

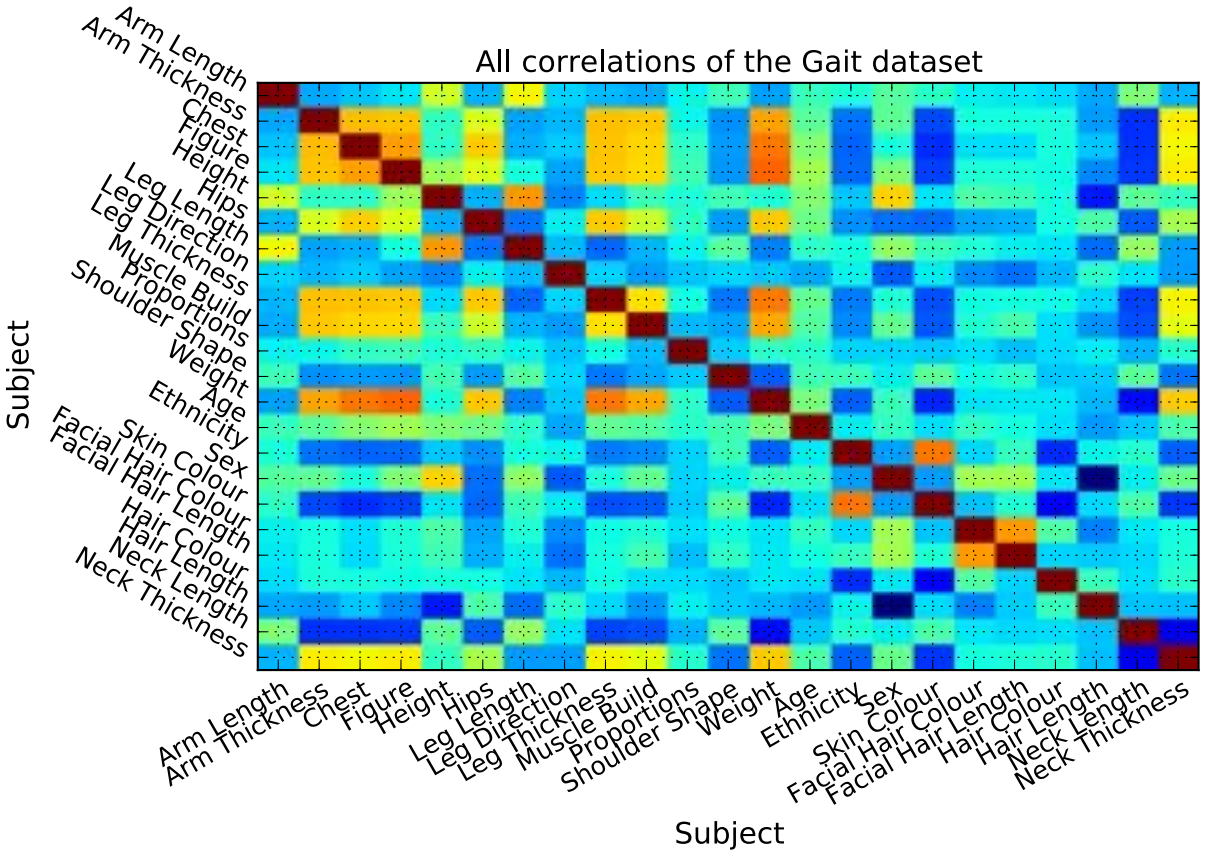
2.5.3.1 Subject Annotations Auto-Correlation and Self Annotation Auto-Correlation

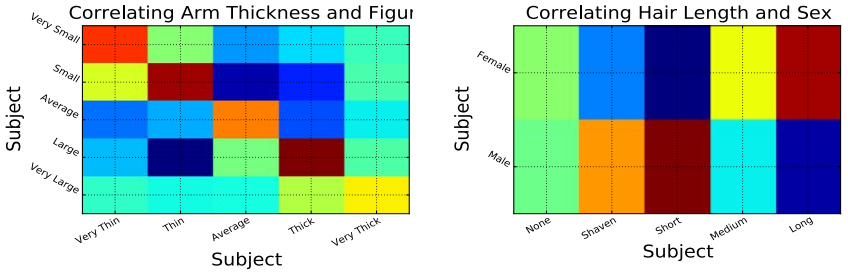
In Fig. 2.9 and Fig. 2.10 we explore the correlations between subject annotation auto-correlation, representing how often individual trait and term pairings were used by annotators. Due to its nature, in the identity of the graph we achieve a perfect correlation. This is a trivial result meaning simply that a term appeared with itself every time it was used in an annotation. More informative correlations can be seen firstly between traits 0 to 12. These are build traits whose terms describe overall thickness and length of the body, as well as extremities. We note that Figure and Weight are highly correlated. In turn they are both correlated with Arm Thickness, Leg Thickness and Chest annotations. Correlation can also be noted between Height and Leg Length, each also portraying correlations with Arm Length. We also notice some inverse correlations. In Neck Length against Neck Thickness we see signs of thinner necks being correlated with longer necks, bulky necks with shorter necks and so on. This inverse correlation can also be noted in both Neck Length and Neck Thickness compared to other traits of bulk and length respectively, though it should be noted that these inverse relationships are not as significant. There seems to exist two groups of traits whose terms correlate in ascending order. Namely traits denoting some notion of bulk or girth (represented by Weight, Figure etc.) and those denoting some notion of length or longness (represented by Height and appendage lengths).

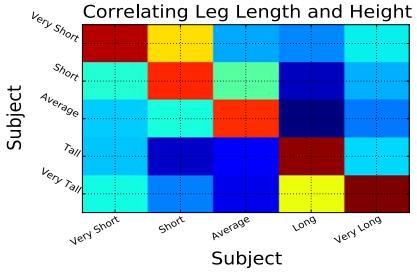
Another informative set of correlations can be noted between the global and head traits. Again both datasets show clear correlations between annotated Skin Colour and Ethnicity. This is to be expected as skin colour is a major contributor to the description of ethnicity. We also note a correlation between skin colour and hair colour; this was expected due to physiological and anthropological reasons. With regards to Sex we observe a high correlation with Females and longer hair and Males and shorter hair. Alternative fashion trends notwithstanding, within our datasets Hair Length seems to be a reasonable distinction between the genders.

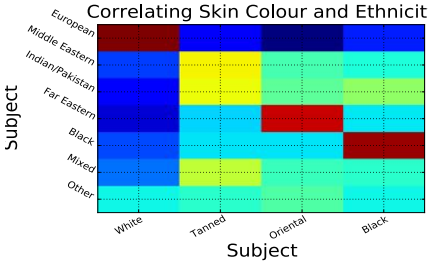
The rest of pairings show little to no correlation, bar a few outliers, which is to be expected. We find basically no correlation between most build features and global features for example. Though we estimate that ethnicity can dictate stature to some extent, either our dataset was too small, or within stereotypes variance is too high to show correlation in our results. An outlier of note is the strong correlation between younger Ages and shorter Heights. Upon further inspection these proved to be the height annotations ascribed to the children present in the respective datasets, a result to be expected with human height often achieving stability in the adolescent years.

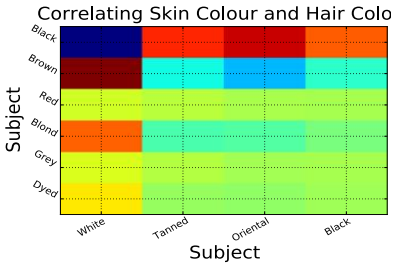
FIGURE 2.9: Term Correlations of annotations ascribed by individuals in TunnelDB











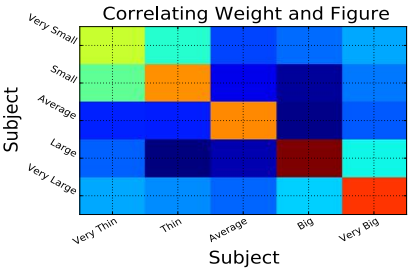
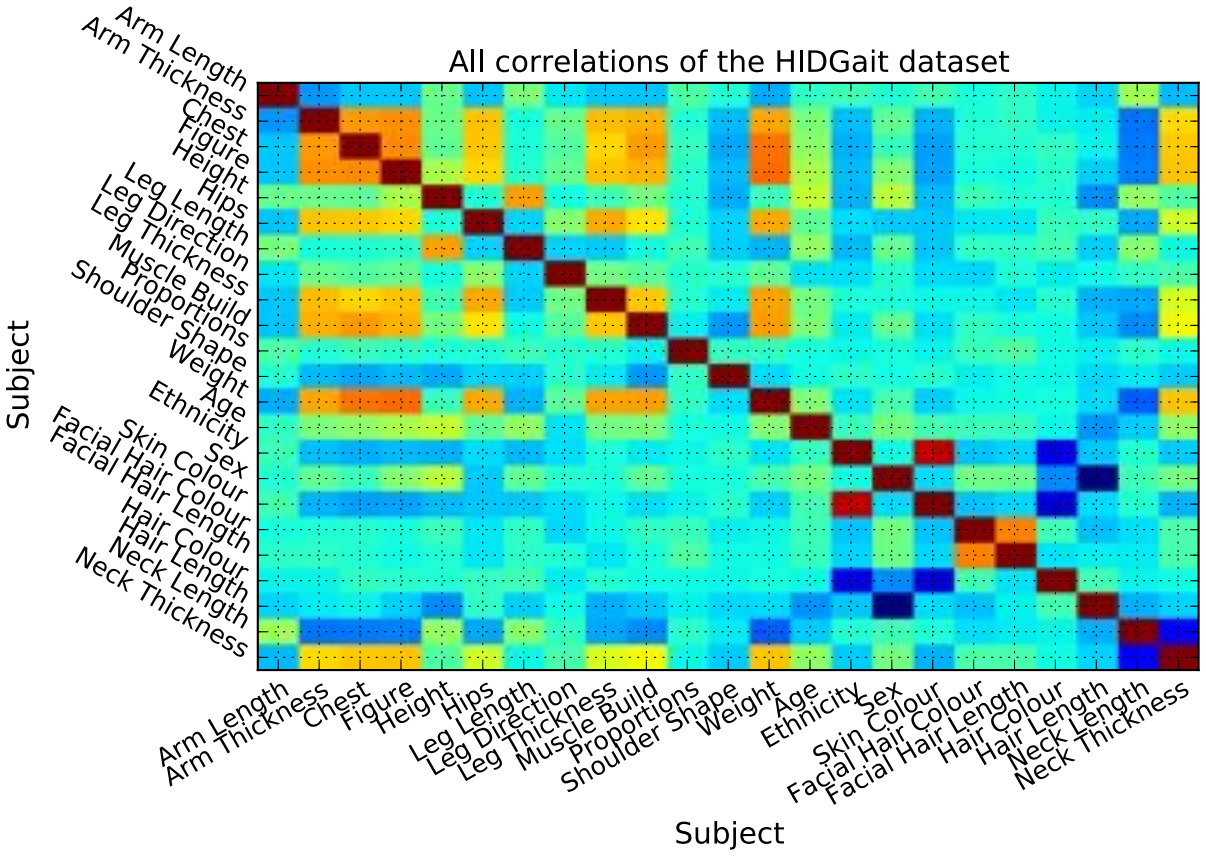
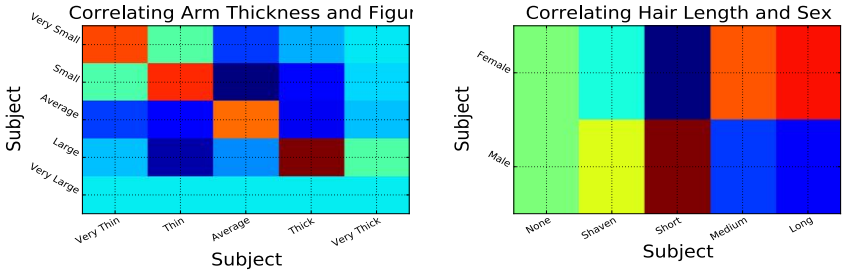
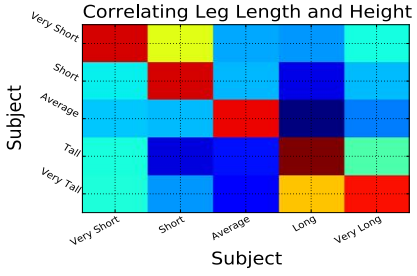
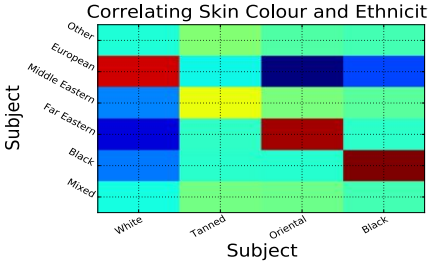


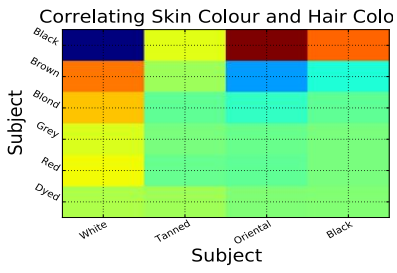
FIGURE 2.10: Term Correlations of annotations ascribed by individuals in HIDDB











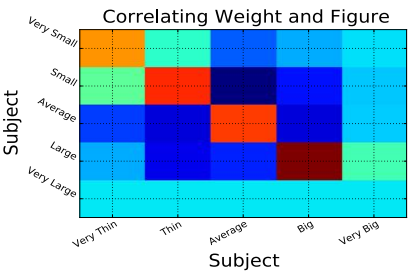
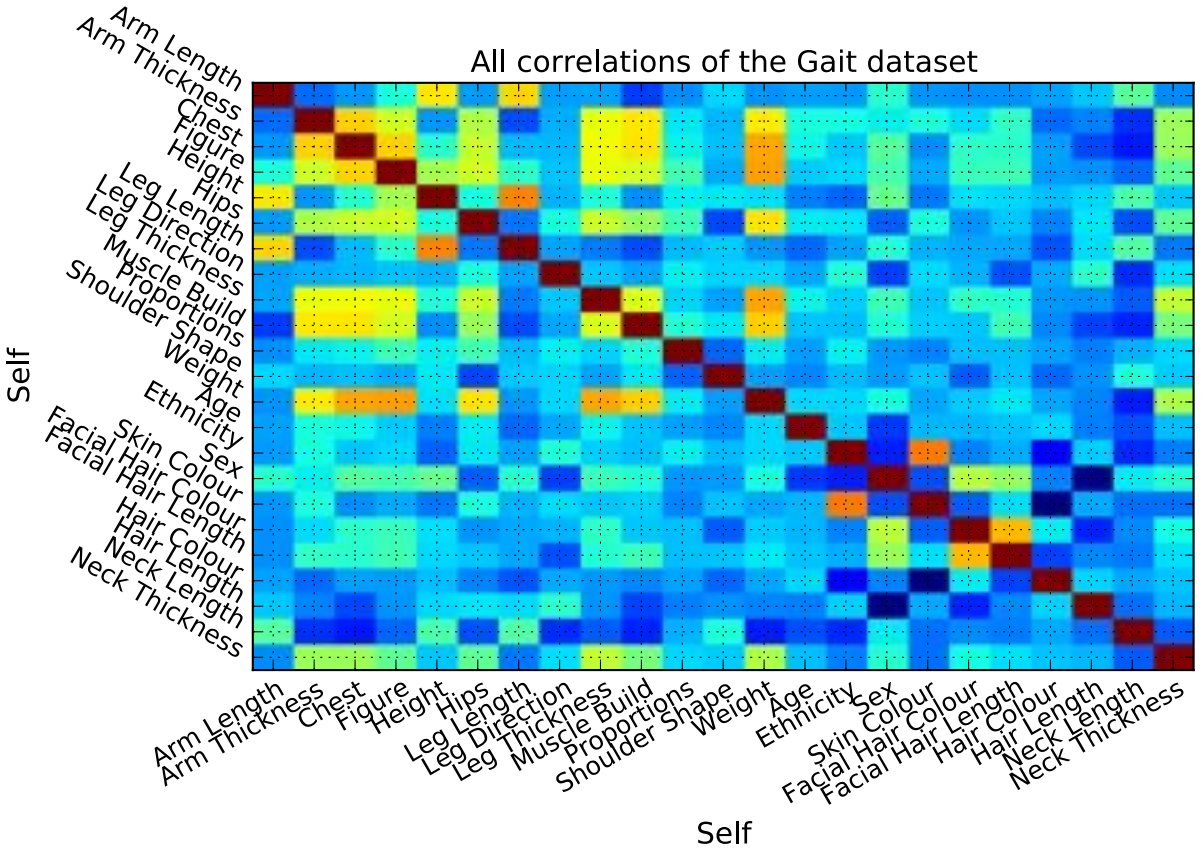
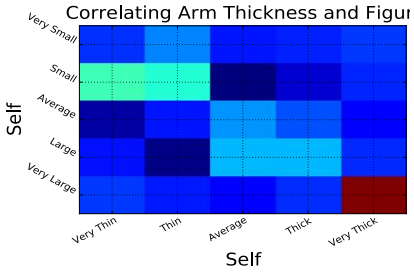
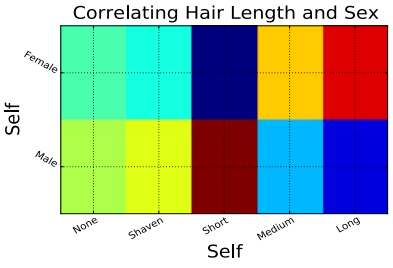
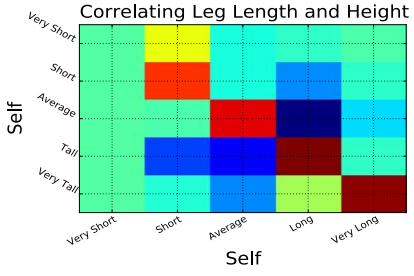


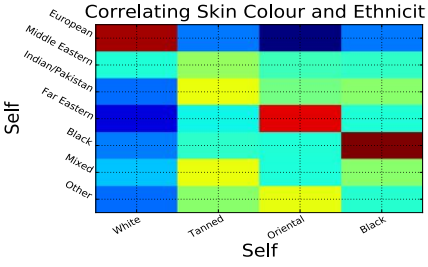
FIGURE 2.11: Term Correlations of self annotations in TunnelDB

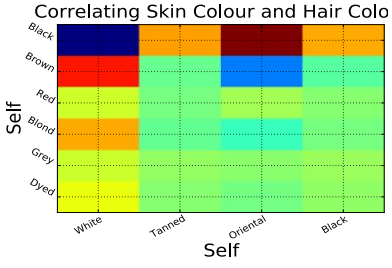












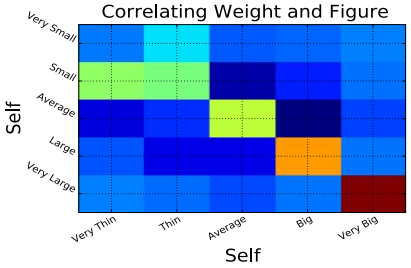
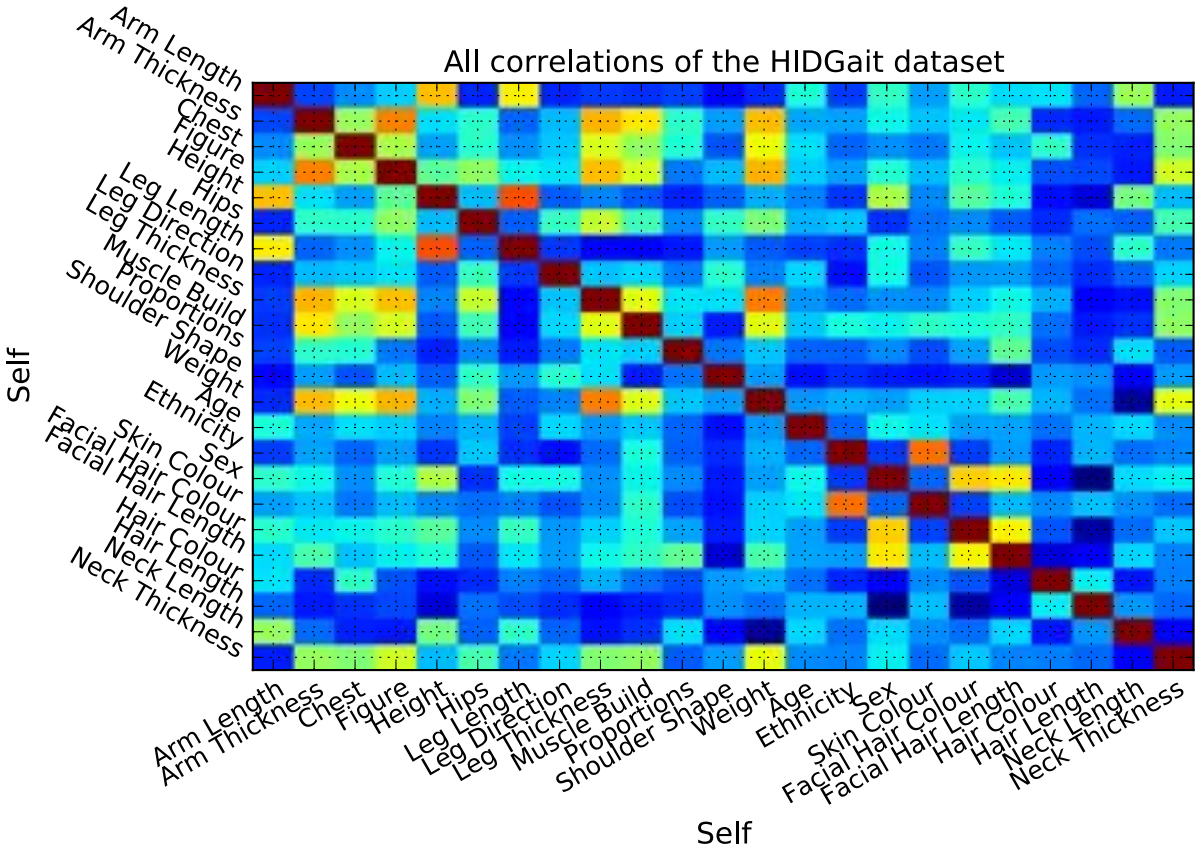
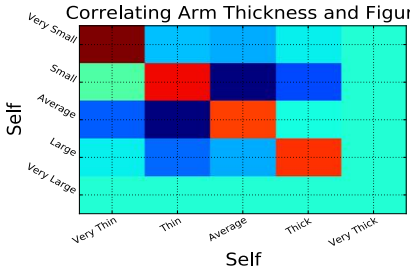
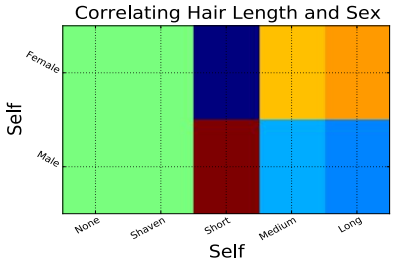
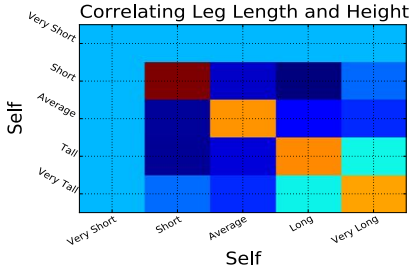


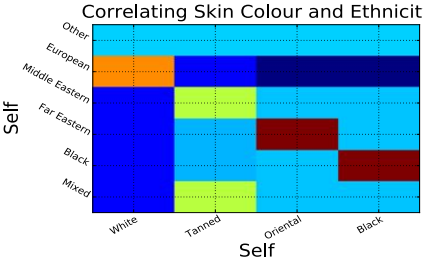
FIGURE 2.12: Term Correlations of self annotations in HIDDB

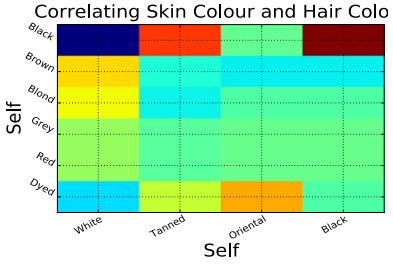












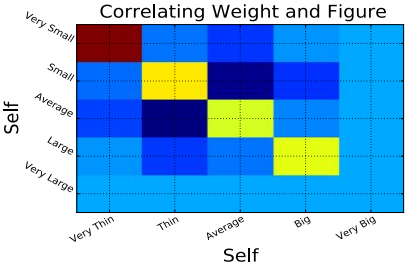
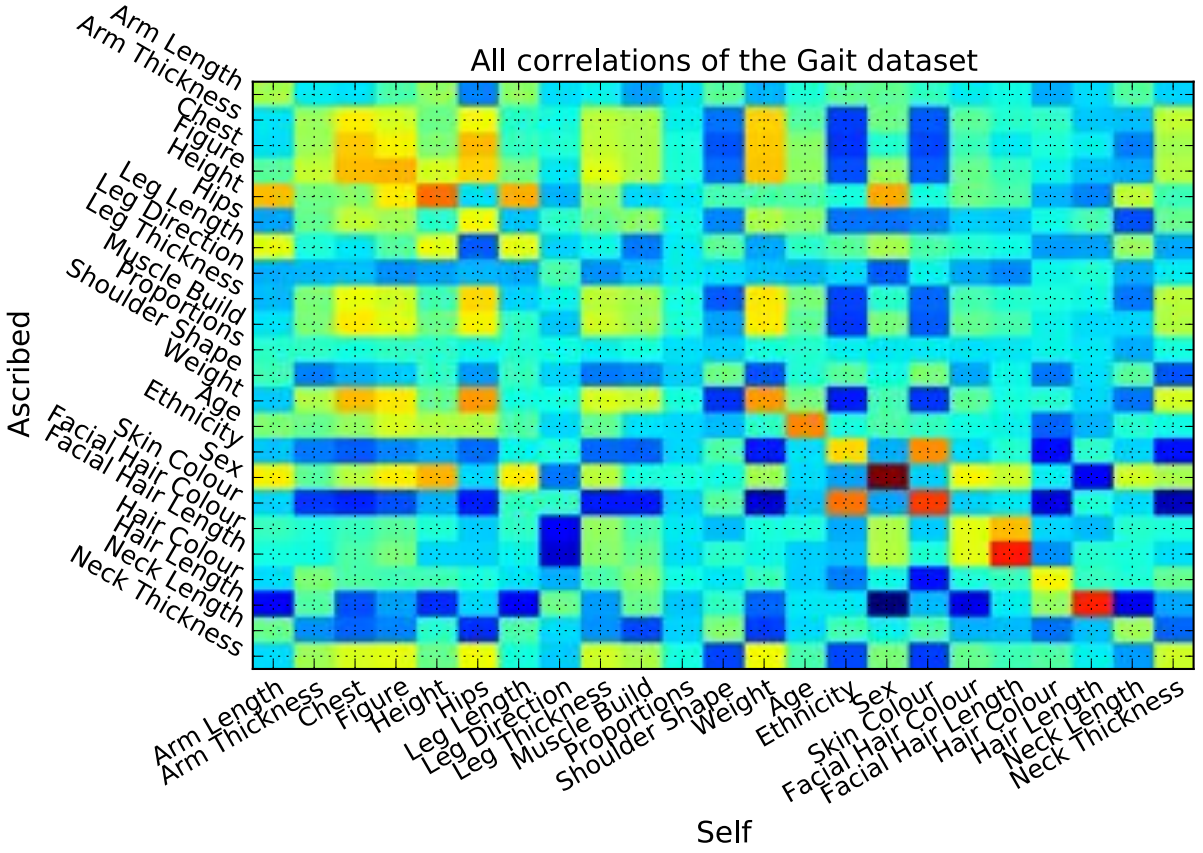
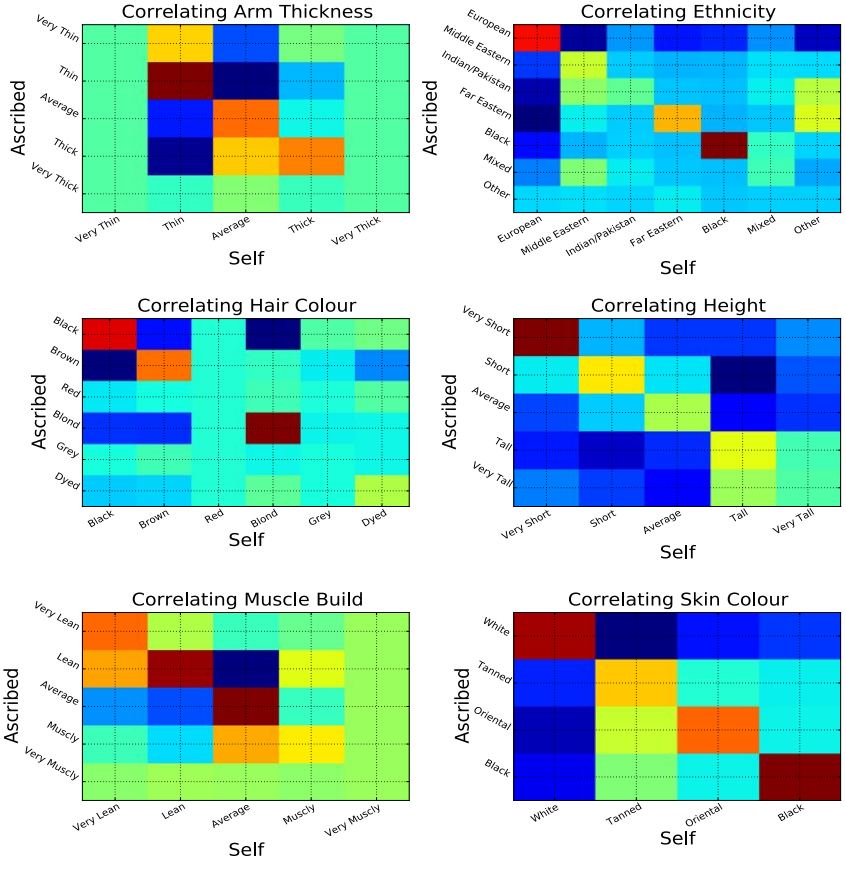


FIGURE 2.13: Term Correlations of annotations ascribed TO individuals against their Self Annotations in TunnelDB





In Fig. 2.11 and Fig. 2.12 we see the auto-correlations of self annotations. The correlations in self annotations are very similar to those found between ascribed annotations and many of the same statements with regards to build and global features can be made as above. This shows that in describing themselves that annotators are as consistent as they are when describing other people. This corresponds well with the similarity in annotations distributions noticed in Section 2.5.2.

The correlations noted between self auto-correlations and ascribed auto-correlations can be broadly interpreted in two ways. One possibility is that these correlations exist inherently in the human population. In this case annotators may be acting on the existence of some natural correlation between these traits. With regards to the build traits, naturally bulky individuals may often have bulkier legs and arms, shorter individuals have shorter arms and legs. With regards to global traits, white people have pale skin, black people have dark skin and so on. Though this may be intuitive, one can easily imagine contrary situations, such as pregnant women, who could potentially have large stomachs and waists, but average or thin legs and arms.

Another possibility is that annotators are making holistic decisions which effect their annotation of sets of traits in unison. In this case annotators make some categorical decision, holistically considering all the attributes of an individual; they then proceed to assign tags to a set of traits in unison based on this decision. For example, people may only notice two variables with regards to build, namely some notion of bulk coupled with some notion of lengths; or may notice some overall concept of ethnicity. Upon viewing an individual and making this decision, the annotator proceeds to choose terms for individual's traits which coincide with these decisions. A small european girl may be denoted as having white skin and thin arms regardless of actual perceived dimensions of her arms or the relative tone of her skin. Trying to understand which is the case would require a data outside the scope of the current dataset. As described at the end of Section 2.5.2, to understand the reason for such correlations a direct comparison of ascribed annotations against some ground truth measurements must be made. Such ground truths would help us understand whether these correlations exist inherently in the population or whether human perception is ignoring parts of the feature descriptions in favour of decisions made against holistic features.

2.5.3.2 Self Annotations vs Ascribed Annotations

In Fig. 2.13 we examine the correlations between annotator self annotations and the annotations those annotators were ascribed. All participants in the gathering of TunnelDB were requested to make self annotations. Therefore, all annotations made on subjects

in TunnelDB can be compared with their self annotations. Unlike the previous pairing, the identity of this matrix is of clear interest. High correlation in the identity means the same terms were used in self annotation and ascribed annotation, low correlation means the opposite. The diagrams clearly show less correlation in build features than in global features. This could show that although annotators can accurately gauge the population's response to their Age, Sex and Ethnicity they have more trouble understanding how their physical appearance will be gauged by the population as a whole. Some physical descriptions are also clearly better than others. While knowledge of ones own Limb and Bulk descriptions is lacking, we seem to have a better idea of our own Height.

2.6 Conclusions

In this chapter we have introduced our approach to semantic physical description analysis. We have chosen a set of physical traits which are consistently and accurately discernible at a distance. The traits were justified in the context of cognitive psychology and eyewitness analysis. For each trait, a further set of semantic categorical terms were outlined and justified.

To discover the potential of semantic terms in biometrics identification and retrieval, a set of annotations against two existing biometric datasets has been gathered. We have designed a purpose-built system for the annotation of biometric signals using the physical traits taking into considerations and counteracting possible points of weakness in human descriptive ability. The content of the dataset gathered has been summarised. Finally, we have presented an exploration of the annotations gathered in two ways. Firstly, we explored the distribution of the annotations gathered, highlighting notable patterns found in these distributions. Secondly we explored the internal correlations between ascribed annotations and self annotations. Through analysis of this correlation we uncovered structures inherent in the data gathered, providing some evidence for the validity of the data gathered.

The following chapters further analyse semantic annotations through their practical application in two distinct scenarios. In doing so we show the capability of the semantic description of physiological traits for purposes of recognition and multimedia retrieval.

Chapter 3

Semantic Biometric Fusion

3.1 Introduction

The identification of humans is an important task, essential for controlling access to resources or locations, as well as identification in surveillance scenarios. The identification task can be expressed as a multi-class classification problem: the identity (*class*) of an individual (*probe-element*) must be ascertained based on its similarity to some set of individuals (*reference identities* or *gallery-elements*). The effectiveness of a given classification system can be measured by its ability to separate elements of the reference identity set based on their inherent distinguishing attributes. We can also measure a given system's False Positive (FP) and False Negative (FN) classifications for given thresholds. This classification process is the main goal of Biometrics [58]; the science of establishing probe-element class membership through the analysis of inherent human physiological, chemical or behavioural modalities. These modalities must hold certain characteristics [57], namely: Universality, Distinctiveness, Permanence and Collectability in order to be applicable to the identification problem for large populations.

The process by which biometrics achieve identification starts with the capture of a biometric signal. Depending on the given situation, some sensor is used to translate a given modality into some computable format. For example, this may take the form of a audio signals capturing a voice, 2D images capturing a face, relief information for fingerprints, video information for gait etc. Once the signal is captured various stages of pre-processing are undertaken which attempt to extract the useful information regarding the individual from any irrelevant background information in the modality; this process is called enrolment. Once enrolled, the useful signal information is represented in some comparable numerical format called a *feature vector*. Various techniques can be used

which attempt the classification of this feature vector, and hence the identification of the probe-element that the feature vector represents.

Several human physiological modalities have been identified which are suitable for use in such a process, each with associated techniques for the extraction of usable features. These modalities include, but are not limited to: iris, face, ear, speech, signature, DNA, fingerprint and gait. They each have their weaknesses and strengths: DNA, fingerprint and iris are noted for their accuracy, but require suspect co-operation and contact; gait is effective at range, but is affected by uncontrollable covariates such as mood or clothing. As a consequence none are considered (or expected) to accurately identify in all situations; there is no panacea in biometrics. As outlined by Jain et al. [55, 57], biometrics in isolation may have the following limitations:

Noise: Sensed data from a particular modality may be noisy or distorted. An example in surveillance is Closed Circuit Television (CCTV) where, although the amount of CCTV cameras installed in public locations has increased, the quality of the recorded data remains poor due to low resolution. This means the biometric features extracted from CCTV and other surveillance sources are usually of poor quality, and their signatures are therefore susceptible to noise.

Non-universality or unavailability: It is unreasonable to assume that every modality can be extracted from every member of a population. In some cases the quality of a modality may be too low, for example, dry cracked finger tips mean unusable fingerprints. Also the signal may not be collectable, for example in cases of a mugging, CCTV footage may be available showing gait and posture but not iris and fingerprint.

Intra-Class variation: These variations describe the different signals which could be extracted from two recordings of the same subject. This is particularly a problem with techniques such as gait, which is not mood invariant. Related to this is spoofing where the Intra-Class variation can be increased immensely for purposes of deception, for example in voice recognition.

Inter-Class similarities: When datasets are large it becomes more likely that separate individuals will share similar biometric signals simply due to limited range of the feature space and subsequent overlaps.

The overarching issue is that when noise is high or populations are large, the *intra-class* variance increases and as this approaches the *inter-class* variance, it becomes more difficult to correctly classify individuals i.e. FP classifications increase. To tackle some of the issues present in individual biometrics techniques (*uni-modal biometrics*)

many approaches have combined multiple biometrics (*multi-modal biometrics*), i.e. the production of a single classification from multiple sources. Biometric fusion has received much interest in the last decade, with several approaches taken towards choosing the level of fusion (data, feature, score etc.) as well as several implemented fusion scenarios (multiple sensors, multiple classifiers, multiple biometrics features etc.). Examples of such approaches include the combination of: 3 separate gait signatures [4], 3 face and 2 voice signatures [16] and face and gait signatures [61] all with promising results.

In this chapter we introduce the use of semantic annotations as a biometric modality, both in isolation and in fusion with 2 primary biometric modalities across six different biometric signatures. In Section 3.2, we outline some background of the two existing modalities used in our experiments, namely Face and Gait biometrics. In Section 3.3 we explore biometric fusion in general, summarising techniques and discussing some previous work. In Section 3.4 we present the feature vectors used in our experiments. These include our semantic features constructed from annotations described in Chapter 2 and our six automatic features extracted from our two modalities available in the HIDDB and TunnelDB datasets against which we have a collection semantic annotations. In Section 3.5, we outline a set of experiments which highlight the ability of our semantic features to function as biometric modalities and also outline the most important physical traits with regards to identification.

3.2 Biometric Signatures

In the previous chapter, we outlined the collection of a novel dataset of semantic annotations associated with individuals stored in existing biometric datasets. To explore the abilities of semantic annotations as compared against, as well as in combination with, existing biometric techniques we must first outline those techniques and the modalities they analyse. Gait and face biometrics have been chosen for the comparison and fusion tasks. Both biometric modalities are non contact and therefore amongst the few which are usable across larger distances. This factor complements the situations in which witness descriptions are necessary, making these biometrics related to the semantic annotations gathered, a topic we explore further in Chapter 4. In this section we present a brief history of the fields of gait and face biometrics. This overview provides the tools necessary to understand how we can compare the use face and gait signatures with semantic biometrics, as well as how we can use these techniques effectively in fusion.

3.2.1 Face

Face recognition has been called the holy grail of artificial vision and biometrics [100]. It is research area is extremely active, primarily motivated by the inherent advantages of the face as a modality, including:

- Its inherent non-contact nature. Face biometrics are non-contact and therefore non-obtrusive. This results in higher levels of public acceptability and also makes face biometrics potentially acquirable without subject co-operation in surveillance scenarios.
- The prevalence of large face datasets. With the rise of cheaper digital cameras and large police mugshot datasets the collection of face signals is both easier and more prolific with several standard datasets of faces available [82, 83, 99][1] for analysis and comparison of techniques.
- Its relationship with the human ability to recognise each other. In the interest of semantic biometrics, faces share a direct analogy with a major component of human recognition [44]. This fact itself may be an underlying motivator for the numerous research efforts focusing on face as a biometric.

Interest in the use of the human face as a tool for identification can be dated to Galton [34] in 1890. Later, initial experiments with automatic computerised face recognition can be dated to the first large-scale computers in 1964 [14]. This work appeared over 40 years ago and since then face recognition has been applied to a wide variety problems attracting an extremely broad range of researchers, from biometric analysis to commercial applications. The COMPENDEX reports over 900 works published under the controlled term "Face Recognition" in 2009 alone and over the past decade there has been mention of 15 conferences dedicated to facial recognition [139]. Therefore this summary does not hope to provide a complete dissemination of the field. Instead, our aim is an overview of the aspects of face recognition of interest for our purposes: namely the process of face detection and simple techniques for face recognition with faces gathered from the TunnelDB. Reviews of the research can be found in [53, 133, 139].

Broadly speaking, face recognition can be separated into the two major tasks inherent in any biometric system. Firstly, the face in a given image or video must be extracted. This process must take into consideration such factors as: pose; occlusion; facial expression; image orientation; lighting and the removal of background information [133]. This portion of the task is critical, if these covariates are completely accounted for, face

[1] Several others found at: `http://www.face-rec.org/databases/`

recognition is made trivial. A popular approach to the task of face localisation in simple scenarios is the Harr-cascade as proposed by Viola and Jones [125]. This technique uses an integral image and a set of Harr wavelet features to make quick decisions about whether a patch of pixels contains a given object which it has been trained against. The technique is very powerful and incorporates resilience to occlusion, changes in lighting and scale. However, this approach is sensitive to pose, orientation and extreme occlusion of faces when compared to the training set meaning that for more complex system other approaches must be taken. However, for the TunnelDB this is sufficient as direction of gaze and lighting are controlled variables.

Once face registration is achieved, a set of features must be extracted from the detected probe face for comparison with the same features extracted from faces in the gallery set. The bulk of modern face recognition research is made up of the various kinds of features which can be gathered from a given face. Broadly speaking, these approaches can be broken down into holistic matching and structural matching techniques [139].

Holistic approaches treat the whole face as raw input, often utilising statistical methods to deal with registration errors. One of the most widely used representations of the face is the eigen-face implemented by Turk and Pentland [122], an approach based on PCA which finds a low dimensional space in which new faces can be projected and compared, excluding dimensions which likely represent error due to mis-registration. This is achieved by finding the main directions of change in a set of training faces. Other methods include: the use of Linear Discriminant Analysis (LDA) which more explicitly attempts to find subspaces which best separate individuals [117]; and also the reformulation of the 2 class Support Vector Machines (SVM) problem to the k class face recognition problem [98]. All these approaches amount to the application of some mathematical transform applied to detected faces aiming to increase the inter-class variance while decreasing the intra-class. With adequate registering holistic approaches work very well. However they tend to perform badly under difficult covariates including changes in lighting, pose and facial expression between the probe and the gallery. This is the case because, by definition, they rely on the overall per-pixel similarity of faces.

Structural approaches locate some local features such as eyes, nose, mouth and chin and measure a set of characteristics for each. By comparing the characteristic and local statistics of these features, identity can be discovered. Some of the earliest approaches in face recognition approached the problem structurally, attempting to measure features such as the width of head and distance between eyes [63] and discovering the geometry of local features [62]. More recent structural approaches include the notably successful Elastic Bunch Graph Matching (EBGM) system by Wiskott et al. [130]. This approach uses a Gabor wavelet transform to discover a set of feature points called jets.

Graphs of jets connected by distance edges can be compared. Another kind of structural approach attempts the estimation of the parameters of a 3D model of a face given a 2D face image [13]. Such approaches attempt to explicitly account for pose, lighting and occlusion once the 3D model is estimated. However, as with all structural approaches, they are inherently reliant on the discovery of local features for parameter estimation, the robust discovery of which is still an open question.

As the face gathering process in the TunnelDB is strictly controlled, the quality of the faces detected are high, pose is practically guaranteed and lighting is controlled. Therefore we successfully employ a simple holistic technique in our usage of the face biometrics from the TunnelDB to gauge a baseline. This technique is described in more detail in Section 3.4.2.2

3.2.2 Gait

In the medical, psychological and biometric community, automatic gait recognition has enjoyed considerable attention in recent years. Psychological significance in human identification has been demonstrated by various experiments [60, 119]; it is clear that the way a person walks and their overall structure hold a significant amount of information used by humans when identifying each other. Like the face, human gait recognition portrays several attractive advantages as a biometric:

- It is unobtrusive, meaning people are more likely to accept gait analysis over other, more accurate, yet more invasive biometrics such as finger print recognition or iris scans.
- It is difficult to conceal. Unlike the human face which can easily be covered by masks, the alteration of gait is difficult. To do so takes considerable effort which is often detrimental to active movement such as running.
- It is one of the few biometrics which has been shown to identify individuals effectively at large distances and low resolutions. However this flexibility also gives rise to various challenges in the use of gait as a biometric. Gait is (in part) a behavioural biometric and as such is affected by a large variety of co-variates including mood, fatigue, clothing etc. all of which can result in large within-subject (intra-class) variance.

Over the past 20 years there has been a considerable amount of work dedicated to effective automatic analysis of gait. Marker-less machine vision techniques have been employed in order to match the capabilities of human gait perception [90]. Broadly

speaking, these techniques can be separated into model-based techniques and holistic statistical techniques.

The latter approaches tend to analyse the human silhouette and its temporal variation, making few assumptions as to how humans tend to move. An early example of such an approach was performed by Little and Boyd [77] who successfully extracted optic flow "blobs" between frames of a gait video which they use to fit an ellipsoids to describe predominant axes of motion. Murase and Sakai [87] analyse gait videos by projecting each frame's silhouettes into the eigenspace separately and using the trajectory formed by all of an individual's separate frames in the eigenspace as their signature. Combining each frame silhouette and averaging by number of frames, or simply average silhouette [38, 78, 123], is the most popular holistic approach. It produces relatively promising results and is comparatively simple to implement and as such is often used as a baseline algorithm.

Model based techniques start with some assumption of how humans move or a model for human body structure, usually restricted to one view point, though some tackle the problem in 3D. Values for model parameters are estimated which most faithfully represent the sensed video data. An elegant early approach by Niyogi and Adelson [91] stacked individual silhouettes in an x-y-time (XYT) space, fitting a helix to the distinctive pattern caused by human legs at individual XT slices. The helix perimeters are used to define the parameters for a five-part stick model. Another, more recent approach by BenAbdelkader et al. [6] uses a structural model and attempts to gather evidence for subject height and cadence.

Model based techniques make several assumptions and explicitly extract certain information from subject videos. Though this would be useful for specific structural semantic terms (Height, Arm/Leg dimensions etc.), the model could feasibly ignore global semantic terms (Sex, Ethnicity etc.) evidence for which could exist in the holistic information [75]. Subsequently we choose the simple yet powerful average silhouette operation for our automatic gait signature both for purposes of simplicity and to increase the likelihood of correlation with global semantic terms. These holistic average silhouettes are extracted from subjects in both HIDDB and TunnelDB. The different datasets collect gait in significantly different ways, therefore the specifics of how these signatures are generated are discussed in more detail in Section 3.4.2.1 and Section 3.4.2.2

3.3 Biometric Fusion

A key problem with any biometric system is intra-class variance caused either by noise or by lack of distinguishing capability of the biometric trait or algorithm. An approach

to addressing this problem is combining data captured from multiple biometric modalities or multiple sets of the same biometric modality. Such *multi-modal biometric systems* [104, 105] can be shown to have less than or equal error rates when compared to a uni-modal system, as shown with some theoretical rigour by Hong et al. [46]. The benefits of multi-modal biometrics systems are somewhat more intuitive; it can be expected that with more information regarding an individual's various traits, a better picture of the identity of the individual can be created. Multiple traits also improve a system's resilience to spoofing attacks, if an impostor is to pass a mutli-biometric system they are required to steal impressions of multiple traits, thus increasing difficulty. The issue of non-universality is also addressed; if a trait on an individual is of poor quality or non existent, the ability to use another trait for validation is desirable. With these benefits in mind it is clear that independent biometric traits are desirable in multi-modal systems. Indeed it is shown by Kuncheva et al. [69] that it is not only desirable to have statistically independent classifiers, but that it is desirable that the classifiers are negatively dependent, i.e. classifiers which commit errors on different objects.

The principle of fusion of multiple biometric signals can be approached in several ways [57]. These may include multiple sensors (e.g. several finger print scanners), multiple units (e.g. using multiple fingers, using both eyes) and multiple traits (e.g. fingerprint and hand, gait and face). In relation to the task of biometric fusion with semantic information, one approach is to define the semantic information as another biometric trait and treat this as a *multiple biometric traits* scenario. There exist four stages at which the fusion of multiple biometric traits could be approached: at sensor level, at the feature level, at the score level or at the decision level. The general consensus is that the lower levels contain richer information about the source traits and as a consequence improve fusion results. Viable approaches along with existing example applications are outlined below. Note that sensor level fusion is specifically ignored as here it requires compatible sensor level signals, which semantic features and the chosen automatic signatures do not share.

3.3.1 Feature Level

In feature level fusion, feature sets from multiple sources (samples, algorithms, modalities etc.) are consolidated into a single feature set after some normalisation scheme is applied. Feature level fusion occurs at the lowest level at which it is still feasible to combine semantic features with automatic features. As a consequence feature level fusion has the capacity to hold the richest information of all fusion levels discussed. Feature level fusion also allows for the exploration of correlation between components of automatic signatures and semantic features. This means correlated features can be removed due to

redundancy, or their correlation can provide useful insight into the relationships between the different sources; this is discussed in further detail in Chapter 4.

Feature level fusion presents several challenges. On a practical level, most Commercial Off-The-Shelf (COTS) biometric implementations do not openly provide access to feature vectors they use, though this is not an issue when all information is open, as in most research level systems. More importantly, it cannot be guaranteed that feature sets are compatible. Finger print minutia generate feature sets of varying length, incompatible with the constant length feature vectors produced by iris analysis. Also, as noted by Ross and Govindarajan [103], the simple concatenation of feature vectors may result in the *curse of dimensionality* [120] problem, damaging the identification capability rather than improving it. This can be avoided through careful selection of feature components which affect matching performance most favourably. Normalisation may also be necessary in feature fusion as features being fused exhibit significant differences in their range and form (i.e. their distributions). Several strategies have been proposed to tackle feature normalisation (min-max, median etc.).

3.3.1.1 Examples

Due to perceived difficulties of incompatible feature sets in feature fusion, most techniques in the past used score fusion. We present some notable examples of multi-modal feature fusion as this compliments semantic fusion, though there are a few examples in multi-sample [86] and multi-algorithm [28, 132] scenarios.

Ross and Govindarajan [103] present an extensive discussion on feature level fusion. In their approach, hand and face feature vectors are concatenated and subjugated to a feature subset selection using PCA. Euclidian distance and threshold absolute distance of the concatenated, dimensionally reduced vectors are combined using a score fusion technique. The authors show some improvement when feature fused scores are themselves fused with simple match scores, success attributed primarily to removal of redundant features. They argue that this itself is justification for biometrics vendors to make feature level information available. Feng et al. [30] present another example, applying Independent Component Analysis (ICA) and PCA on both face and palm print feature vectors, combining them using a feature concatenation. Other examples can be found by Chibelushi et al. [20] who combine voice and lip shape features, reducing dimensionality using PCA and Son and Lee [115] who combine face and iris features, reducing dimensionality using Direct Linear Discriminant Analysis (DLDA).

3.3.2 Score Level

Scores are generated from classifiers as a measure of how well a probe element matches a given class. Score fusion attempts to combine the scores from multiple classifiers to improve recognition. Kittler et al. [64] present a theoretical framework for score-based approaches for consolidating evidence from multiple classifiers. Classification scores provide the richest input pattern information that is still readily available from most COTS biometric matchers. These factors make score fusion the most popular and well-explored fusion strategy in the literature.

Scores generated by different classifiers are likely to be incompatible in their raw form. One issue is orientation; some classifiers produce a distance score, where small values denote relevance, while other classifiers produce a similarity score, where large values denote relevance. There is also no guarantee that scores have a common distribution and range. These factors produce complications which can been approached in three main ways [58, 106]: density-based, transformation-based and classifier-based schemes.

3.3.2.1 Density-Based Score Fusion

This approach starts by formulating the classification problem using conditional probability. For a set of scores generated by R classifiers, scores held in vector $\mathbf{s} = \{s_1, ..., s_R\}$ such that s_j is the score generated from the j^{th} classifier, we define a classification as:

$$\text{Assign } \mathbf{s} \rightarrow \omega_i, \text{ if} \tag{3.1}$$

$$P(\omega_i|\mathbf{s}) > P(\omega_j|\mathbf{s}), i \neq j \tag{3.2}$$

Where $\boldsymbol{\omega} = \{\omega_1, ..., \omega_N\}$ and ω_i is the i^{th} class. This formulation of the *posterior* probability can be calculated using the probability density of the score set given a class label class using Bayes theorem:

$$P(\omega_i|\mathbf{s}) = \frac{P(\mathbf{s}|\omega_i)P(\omega_i)}{P(\mathbf{s})} \tag{3.3}$$

Where $P(\omega_i)$ is the probability of observing a class, $P(\mathbf{s})$ is the probability of observing a given score. The class conditional probability $P(\mathbf{s}|\omega_i)$ is the only unknown and is estimated using parametric [114] or non-parametric [52] techniques. Parametric techniques assume an underlying function for the density function, (e.g. a Gaussian Distribution) and attempt to calculate its parameters. However assigning such limitations may be inappropriate in common multi-biometric score distributions which have large tails and

have multiple modes. Alternatively non-parametric approaches assume no model and are essentially data-driven, using training examples to estimate underlying probability densities. Approaches such as Parzen-Window presented by Jain et al. [52] may estimate densities inaccurately due to finite training data.

3.3.2.2 Classifier Score Fusion

Classification approaches treat each score s_j from each classifier as the dimensions in an R dimensional space, resulting in a feature vector $\mathbf{s} = \{s_1, ..., s_R\}$. Similar to density-based schemes, classifier approaches require a large amount of correct classification examples in the training phase to accurately estimate the parameters of the classifier. The benefit lies in no prior requirement to transform the scores into some common domain (as in transformation based schemes) and no need to estimate complex probability distributions (as in density based approaches). Several classifiers have been used for this approach score fusion techniques, including: the use of a HyperBF network [16], K Nearest Neighbours (KNN), decision trees and logistic regression [124] and several examples using SVM [5, 19, 31, 110].

3.3.2.3 Transformation-Based Score Fusion

Density and classification based schemes require large numbers of training examples, even if independence is assumed and a product of marginal densities is calculated as opposed to the joint-density function [64]. In the scenarios such as that of semantic annotations, gathering many training examples may not be viable. Therefore, rather than using probabilistic frameworks or classifiers to learn the underlying structure of generated scores, another approach is to combine the scores directly using simple fusion operators (sum, product, min-max etc.) and guarantee meaningful results by normalising and orientating scores from each classifier. A variety of normalisation schemes can be employed [106], many of which have been shown to have merit, as discussed in the comprehensive set of normalisation experiments discussed by Jain et al. [52].

3.3.3 Decision and Rank Level

Many COTS biometric matchers do not provide access to scores or features, subsequently the final decision to accept/reject (in the verification case) or rank (in the classification case) of a certain candidate is the only information available. Decision level fusion techniques take advantage of this data, attempting to combine the ranks or final reject/accept decisions given to each class by each matcher.

In the classification case, ranks of classes can be combined in a variety of ways as outlined by Ho et al. [45]. Approaches essentially separate into those performing *class set reduction* and those performing *class set reordering*. Class set reduction attempts to reduce the number of classes in the output list while keeping the true class present, whereas class set reordering attempts to improve the rank of the true class in the output list.

In the validation scenario, the final decision can be combined in a variety of strategies [106]. The most simple approach is the "AND" and "OR" rules, though they come with high False Reject Rates and False Accept Rates respectively [21]. More forgiving approaches use matcher decisions as votes, the most common of which is majority voting [59, 71], though majority voting has limits [70] and more successful results have been reported if votes from stronger and weaker classifiers are weighted appropriately. More elaborate approaches have also been attempted, Xu et al. [131] report improvement in handwriting recognition using Bayesian Decision Fusion. Firstly, the conditional probabilities $P(c_j|w_k)$ (i.e. classification to a particular class c_j given a true class w_k) is calculated using the decisions of a training set. Bayes rule can then be used to calculate $P(w_k|\mathbf{c})$ where $\mathbf{c} = c_1, ..., c_J$, i.e. the probability of true class given a set of decisions. This calculation can be simplified if independence is assumed between matcher decisions.

It has been argued [104] that decision based techniques are coarse, losing rich information by not taking into account the detail held in the features or the scores. However, pure decision technique bypasses incompatibilities between classifiers, ignoring normalisation issues present in score techniques or possible feature space incompatibles and "the curse of dimensionality" [54].

3.3.4 Soft Biometric Fusion

One of the few efforts made towards the incorporation of physical traits held in a format comprehensible, and often collectable, by humans has been the use of so called *soft biometrics* [58, 106] as ancillary data in a process called *soft biometric fusion*. As briefly mentioned in Section 2.1 a few efforts [56, 88, 136] have been made to incorporate attributes such as Gender, Ethnicity, Height and Weight as a source of information alongside primary biometric sources.

The Bayesian framework recommended by Jain et al. [56] approaches soft biometric fusion in a similar way to density estimation in score fusion. Let $\mathbf{x} = [x_1, ..., x_{R_p}]$ be a set of features provided by R_p primary biometrics and $\mathbf{y} = [y_1, ..., y_{R_s}]$ features (such as Gender, Ethnicity etc.) provided by R_s soft biometrics. Independence is assumed between primary and soft features and the probability of a class assignment given an

observation $P(\omega_i|\mathbf{x},\mathbf{y})$ (of both soft and primary features) is calculated using the underlying probability densities of an observation given a class $P(\mathbf{x},\mathbf{y}|\omega_i)$ by following Bayes rule (see Section 3.3.2). These densities are calculated using a set of training examples for primary biometric features whereas, for the soft features, the accuracy of the underlying estimator is used as the parameters for the probability density. Jain et al. [56] discuss the possibility that the calculation of $P(\omega_i|\mathbf{x},\mathbf{y})$ could be dominated by the soft features due to their high variance. This problem is solved through the use of scaling factors used to reduce the effect of soft biometric traits.

Justifications for the use of soft biometrics [58, 106] often cite the benefits provided by information obtained at negligible extra cost to the user and COTS biometric implementation. As such, soft biometric approaches are often discussed along side automatic approaches, extracting the soft biometric information from existing primary sources. However, in using existing automatic feature extraction techniques, we argue that there is potential for the extracted ancillary data to be a reiteration, i.e. information already present in the primary biometric signature. Alternatively, by incorporating human understanding (in forms such as our semantic annotations), we actively enrich the biometric signature with a novel source [69], distinct from information extracted automatically.

3.4 Semantic Fusion

In this section we describe the structure and source of the data used in our biometric and fusion experiments. We describe the process undertaken to represent semantic labels numerically in a manner suitable for classification and fusion. We also describe the automatic visual features extracted from our two datasets, namely the extraction of gait signatures from the HIDDB and the TunnelDB as well as the face signatures extracted from the latter.

3.4.1 Semantic Features

To allow for the analysis of semantic data, we must first numerically represent terms ascribed to traits. There are two strategies which we have explored to represent semantic features. Firstly, as most of our semantic traits are described using terms which lie on some sort of continuum of size (big to small), a logical approach is the assignment of normalised values between 0 and 1 to each trait. In this scheme low values would be assigned to annotations such as Small or Thin, higher values given to annotations such as Large or Fat for each trait. This also preserves the implicit order of the values. For example, Small and Very Small would be values which lie on the same side of the number

line from Average as the centre. Our first experiments with semantic labels used this approach with some success [109] as did our correlation analysis of all traits against all other traits presented in Section 2.5.3.

However though meaningful for traits with ordered value centred terms, this scheme is unnatural for clearly categorical attributes like Sex or Ethnicity which have no concept of order. Choosing an arbitrary order artificially relates two terms and pushes others apart. More subtly, choosing an arbitrary equal separation of the number range between the chosen terms of even value orientated traits may be misleading. For example, the distinction between calling an individual Average or Small may be little; many annotators may be fickle with regards to the fact. However if an individual ascribes Very Thin or Very Large this might be a rare annotation that in turn carries more meaning. In a simple scheme its value may be simply twice the distance from the Average when compared to Small, though in reality the distinction in the annotators mind may be larger. Therefore, another scheme has been explored centred around a binary notation. In this scheme each term rather than each trait is represented. An individual's annotation of a given subject is represented by setting assigned terms to 1 and setting non assigned terms to 0. Though we lose the notion of explicit ordering of value based terms, we open the exploration and correlation of feasibly disparate terms. Also, with careful analysis, the order relationship of terms such as Small compared to Big can still be detected and therefore exploited using this encoding technique (See Section 2.5.3 and Chapter 4)

Following this scheme, for each annotation assigned to each subject, a *semantic feature vector* is generated. This is a 137 dimensional feature vector per annotation attributed to each subject, one dimension per term of each trait. Each feature vector is directly comparable to those of another annotator using any given distance metric; often the Euclidian distance metric is used in our experiments though a cosine metric also has meaning in this context. A unique annotation for a given subject across a set of annotators can also be represented by averaging the responses to each term from each annotator. Such an approach is useful when constructing gallery and probe sets of annotators as well as subjects in the results sections. In Section 3.5 this annotation representation scheme is used to explore the ability of our semantic datasets both in isolation and in fusion with other biometric signatures.

3.4.2 Automatic Visual Features

To give context to the ability of semantic annotations and also to explore their ability in fusion with primary biometric signatures we must first outline those signatures. In

this section we explore the automatic biometric features extracted from the datasets for which we have collected semantic annotations.

3.4.2.1 HIDDB Dataset

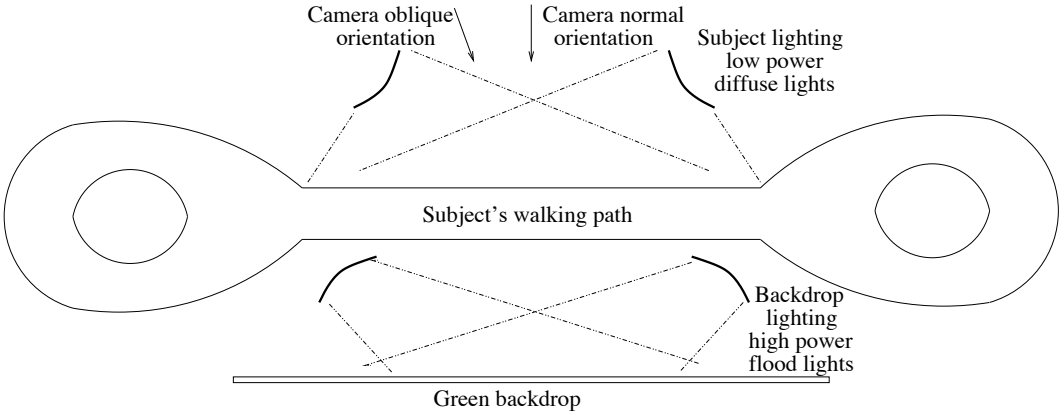
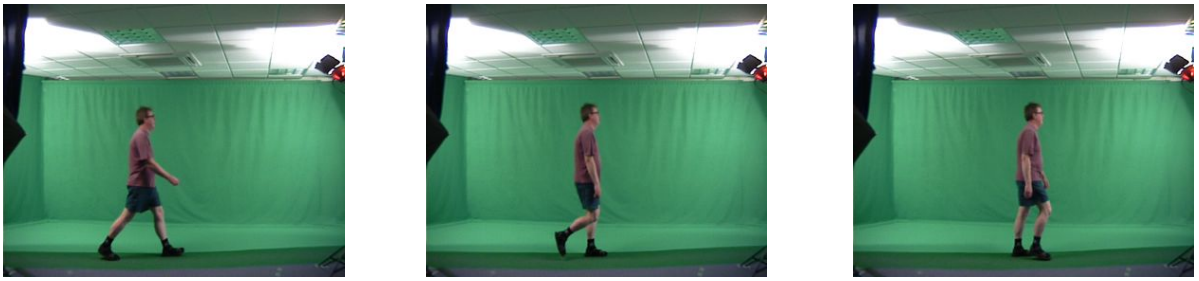


FIGURE 3.1: The configuration of the laboratory portion of the HIDDB from Shutler et al. [113]

Here we describe the features automatically extracted from HIDDB (See Section 2.5.1). Two automatic gait features were extracted from this dataset and used in the experiments in this work: the **Average Silhouette** gait signature and the newly developed **Average Colour Silhouette**. The configuration of the biometric dataset collection environment itself is shown in Fig. 3.1 and discussed in greater detail by Shutler et al. [113]. Subjects collected in the HIDDB walked continuously around the track shown in Fig. 3.1. During their walk the subjects were filmed continuously from two different viewpoints, but in our experiments we use only the "Normal" viewpoint described in the diagram, here called the fronto-parallel viewpoint. As they walk the subjects were captured against a chroma-keyed background allowing for easy background subtraction. A single sample is classified as one complete traversal across the central area of the walking path with the subject walking from right to left or left to right. In practise this amounted to between 6 and 20 samples for each subject in the HIDDB

Standard Average Gait Signature

For each gait sample, firstly the subject is extracted from each frame with a median background subtraction and the frame is transformed into a binary silhouette image (Fig. 3.2(b)). In this image the largest set of connected pixels is taken as the subject. This results in a set of binary silhouettes, one for each frame. At this point each



(a) Individual subject captured by a single high definition camera as they walk

(b) Each frame of a given gait cycle walk has the background subtracted. The subject is the largest connected region

(c) The connected region is isolated and a mask created. This is used directly for the greyscale signature

(d) The connected region is isolated and a mask created. This is used to extract the individual for the colour signature

(e) The masks are averaged across a single gait cycle to create the Average Silhouette Signature

(f) The colour silhouettes are averaged across a single gait cycle to create the Average Colour Silhouette Signature

FIGURE 3.2: Subfigures (a)-(f) showing the silhouette signature generation from the HIDDB

frame is height normalised such that the individual's height is set to 64 pixels and their width is normalised in proportion (Fig. 3.2(c)). This process retains the aspect ratio but purposefully loses absolute height information as to allow the sample to be taken from an arbitrary distance to the camera, making the signatures distance invariant. The silhouette is then centred by its center of mass on a 64x64 final frame. The gait signature of a particular sample is the averaged summation of all these binary silhouettes across one gait cycle (Fig. 3.2(e)). For simplicity the gait signature's intensity values are used directly as the feature vector, although there have been several attempts made to explore a subset of significant features in such feature vectors, using ANOVA or PCA [123] and also mutual information [37].

Colour Average Gait Signature

We formulat another colour gait signature which is likely to correlate with semantic features such as Ethnicity and Skin Colour. The binary silhouettes extracted during the first stage of the standard average gait signatures are used to mask the original full colour videos on a frame by frame basis (Fig. 3.2(d)). From these masked colour images the subject is extracted and once again height normalised and centred to a 64x64 image for each frame. A colour signature is generated from the averaged summation of all these images across the same gait cycle as the standard average gait signature (Fig. 3.2(f)).

These two techniques result in two automatic feature vectors of size 4096 (64x64) and 12288 (64x64x3) (See Table 3.1) respectively which describe each sample video of each of the 115 subjects. This complete set of automatically and semantically observed subjects is manipulated in Section 3.5

3.4.2.2 TunnelDB Dataset

Here we describe the visual features automatically extracted from subjects in TunnelDB [84, 112] (see Section 2.5.1). Two automatic gait features were extracted from this dataset and used in the experiments in this work: the **Projected Gait (Normalised)** signature and the **Projected Non-Normalised Gait** signature. Furthermore, two face features were extracted from another portion of this dataset: the newly developed **Average Face** signature and the related **Average Face Histogram** signature. The configuration of the biometric tunnel itself is shown in Fig. 3.3. Subjects collected in the TunnelDB walk through an entry beam on a straight red path towards the exit beam and therefore towards a face camera. During a single walk (a sample), the subject is simultaneously captured by the gait cameras and the face camera. Upon reaching the exit beam, a single flash camera is used to photograph the right ear.

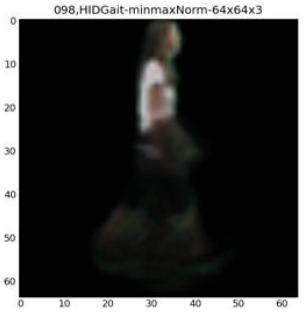
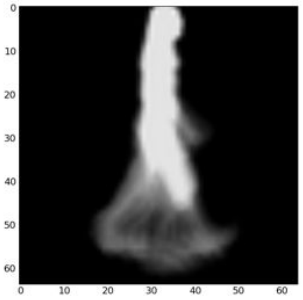
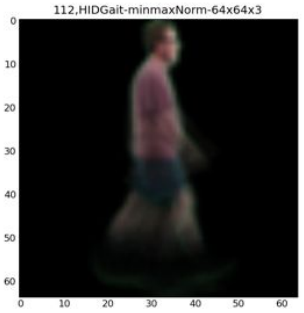
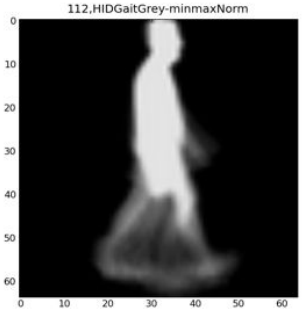
	Average Color Silhouette	Average Silhouette
Subject 098		
Subject 112		

TABLE 3.1: HIDDB Signature Examples

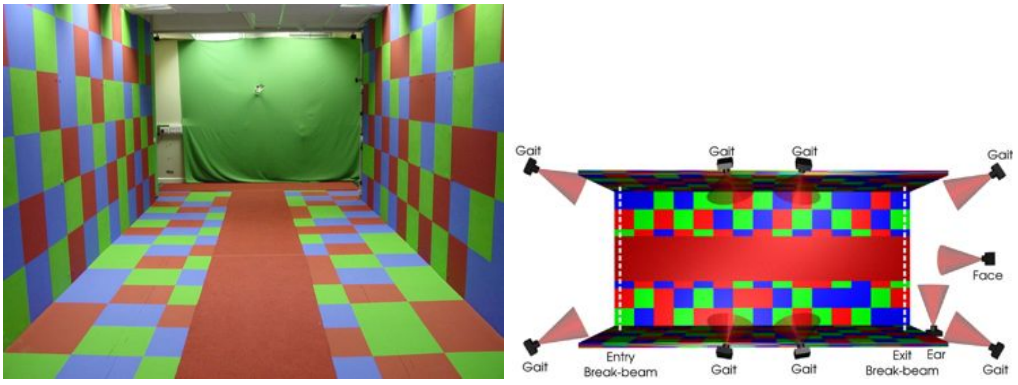
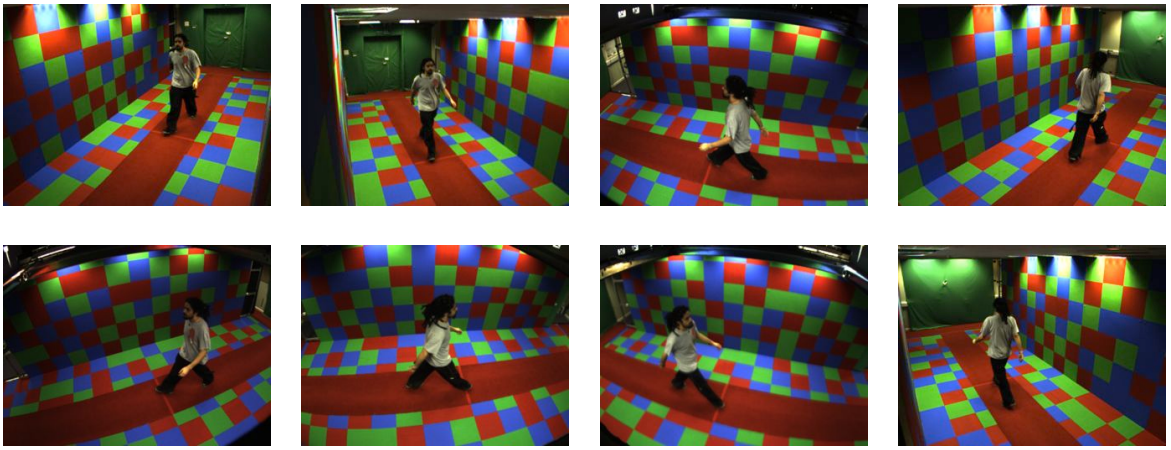


FIGURE 3.3: The configuration of the biometric tunnel used to gather TunnelDB

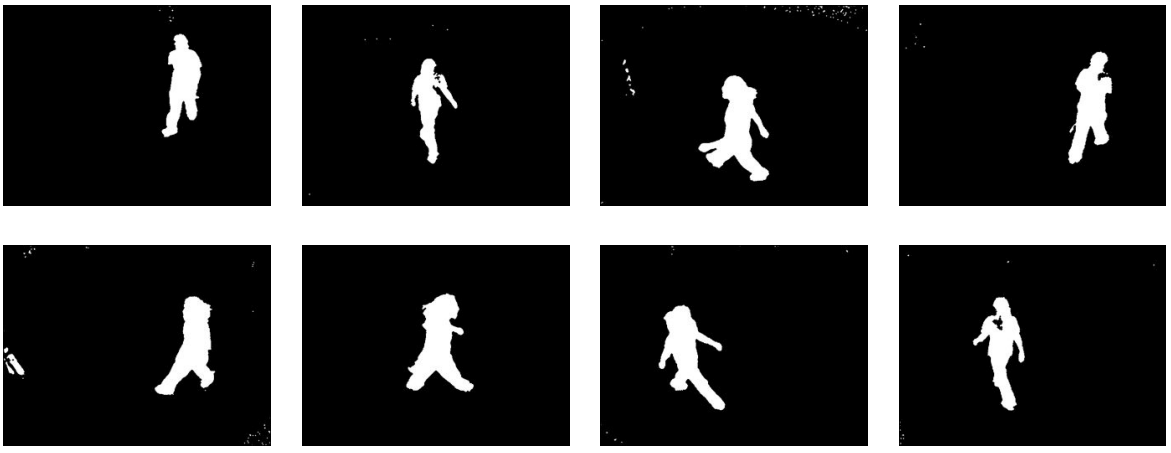
Projected Gait (Normalised and Non-Normalised)

One of the main contributions of the TunnelDB is its novel dataset of 3D gait signatures. Subjects in the dataset are synchronously captured by 8^2 and later 12^3 cameras. These cameras are used in combination to produce a 3D model of a given subject's walk and can therefore produce gait signatures of a subject from several novel viewpoints. This results in several applications including self contained security checks (e.g. airport identity verification) as well as viewpoint reproduction to replicate signatures extracted

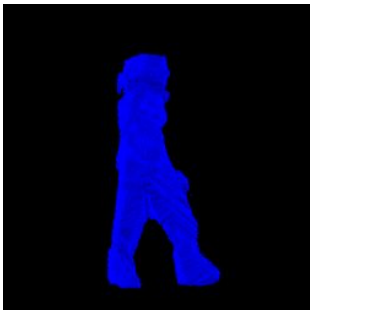
[2] Until July 2007
[3] Current configuration



(a) Individual subject captured by 8 synchronized cameras simultaneously



(b) Each camera is background subtracted. The largest connected area is the subject's silhouette



(c) Using each silhouette, volumetric carving is used to construct a model

(d) The model of each frame is used to capture a silhouette generated from a novel perspective

(e) These silhouettes are combined to produce a single signature

FIGURE 3.4: Subfigures (a)-(e) showing the signature generation from the TunnelDB

from arbitrary cameras (e.g. CCTV). There are several stages involved in producing a 3D model from videos collected by this system. Before the subject walks through the tunnel, a snapshot of the tunnel background is taken for each camera. For each image taken of the subject as they walk through the tunnel (Fig. 3.4(a)), this background is subtracted resulting in a silhouette per frame of the subject's walk for each camera (Fig. 3.4(b)). These silhouettes are used as the basis for a volumetric carving technique [112]. This process can be intuitively understood by picturing a 3D scene where all volumetric-pixels (or voxels) are potentially those of the subject at a given frame. Given the knowledge of the exact calibration information of each camera it is possible to project a cone representing a given camera's silhouette of the subject into this 3D scene. By "keeping" voxels in the scene covered by the projection of most or all of the camera's silhouettes while "removing" those voxels covered by few or none of the camera's silhouettes, it is possible to carve a 3D representation of a given subject at a given frame (Fig. 3.4(c)). This process is demonstrated visually in the 2D case in Fig. 3.5 and produces static 3D models of a human. Using the generated 3D model, gait signatures can be created of a subject from novel viewpoints. This involves producing a model for each frame of an individual gait cycle of a subject (Fig. 3.4(d)) and then combining each of these frames to form an average silhouette from the given perspective (Fig. 3.4(e)). This process is described in more detail by Seely et al. [112].

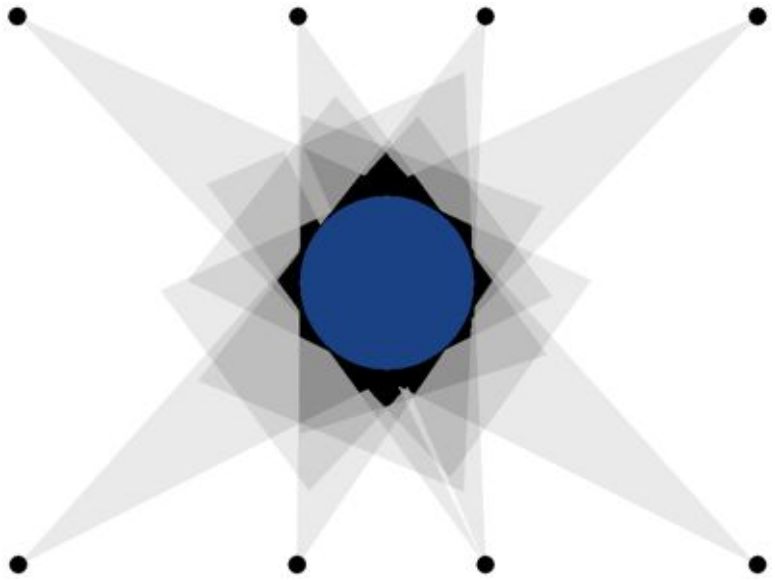
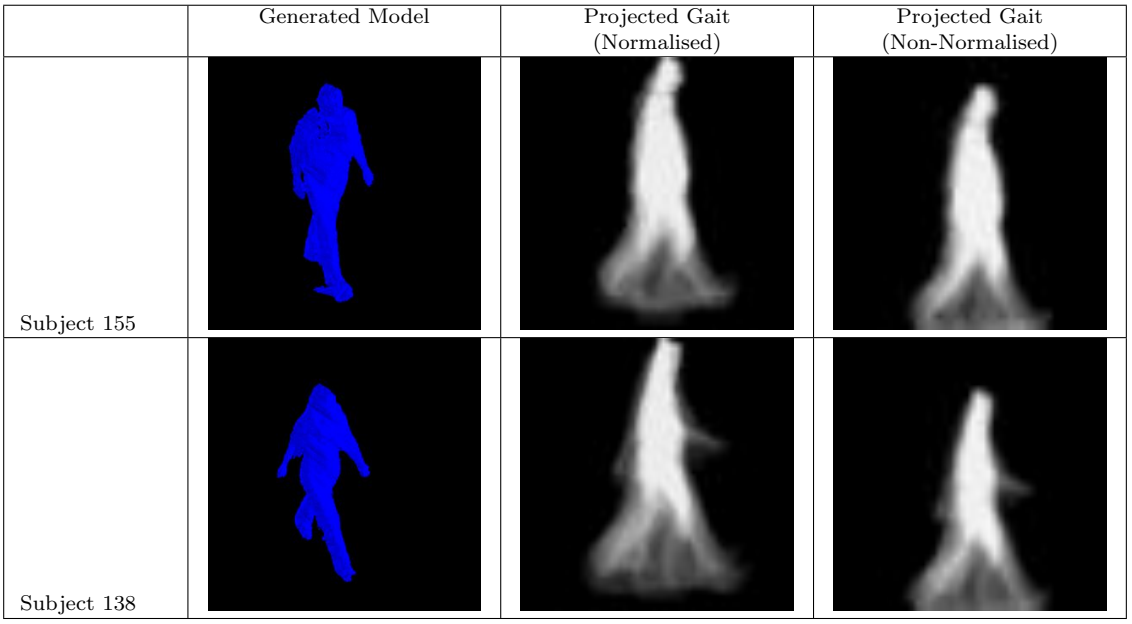


FIGURE 3.5: An example of volumetric carving in the 2D case. Here the actual object being perceived is the circle in blue while the regenerated object found from the volumetric carving of each camera is shown in black.

To complement the features generated from the HIDDB we chose to generate 2D signatures from a fronto-parrallel perspective from the 3D models gathered in TunnelDB. For each 3D model generated for each frame, the virtual camera is placed parallel to

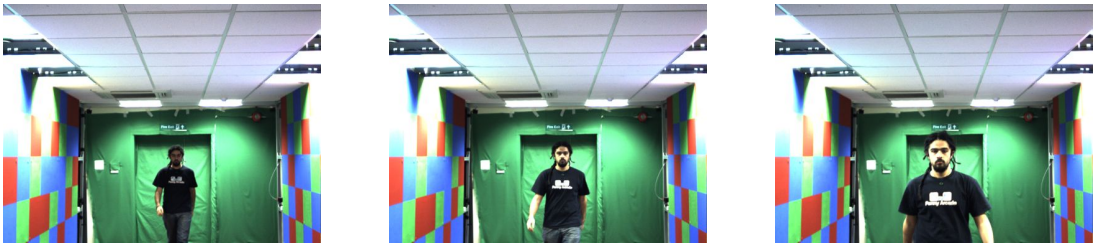
the direction of walk and a single 2D frame is projected. Each frame is then treated in one of two ways. Firstly, we can normalise each frame, making the signature distance invariant. We call this the **Projected Gait (Normalised)** signature. However, as each of these projections is made from a synthetic camera viewpoint whose exact 3D position is known, the scale information of the subject from the camera need not be removed, and has been shown in Section 4.3 to contain important information about the subject's identity. Therefore we also generate and analyse a second signature, called **Projected Gait (Non-Normalised)**. With or without this normalisation step, these frames are scaled to a (64x64) image and averaged across a single gait cycle producing a gait signature for a given sample of a given subject in a similar manner to those generated in Section 3.4.2.1. This results in two sets of 4096 (64x64) signatures (see Fig. 3.8) for each subject generated from a camera perspective dependent on an automatically generated 3D model.

FIGURE 3.6: TunnelDB Signature Examples



Average Face and Average Face Histograms

While gait images are taken, a single higher definition camera at the end of the tunnel captures a 1600x1200 high resolution face images at 27 frames per second. In the tunnel scenario, direction of gaze is guaranteed by instruction to subject as well as by their walking direction. Lighting and other environmental variables are also controlled. This means that many of the difficulties inherent in face detection and face recognition discussed in Section 3.2.1 can be ignored. Also, as the background of the tunnel is known it can be easily removed, isolating the subject in the scene which is further simplified by



(a) Individual subject captured by a single high definition face camera as they walk

(b) Each frame of the walk has the background subtracted. The largest connected region is taken as being the subject

(c) The connected region is isolated by using as the largest connected region above as a mask on the original frame

(d) The Viola-Jones faces detector implemented in the OpenCV library is used to locate the face in the isolated region

(e) Each face is normalised to 32x32 coloured image. A single average face is generated and used as the signature

FIGURE 3.7: Subfigures (a)-(e) showing the signature generation from the TunnelDB

the presence of only one subject in the tunnel for a given sample. With this knowledge, some preprocessing is undertaken to aid the localisation of a face in each frame. Firstly, for each frame, the background is subtracted (Fig. 3.7(b)) and a bounding box is drawn around the area of the image containing the largest bulk of pixels distinct from the background (Fig. 3.7(c)). The face is assumed to appear in the upper portion of the located bulk of pixels. This preprocessing lowers the area within which to look for a face from 1600x1200 to between 200x100 on earlier frames and 500x200 on later frames. At this point a more powerful face detector is used to find the exact location of possible faces (Fig. 3.7(d)). We use the Viola-Jones face detector [126] implemented in the OpenCV library [15]. This allows the final narrowing down of a bounding box drawn around the most likely location of a face in the background-subtracted image. The location of

this face is used to provide further clues as to the face location in neighbouring frames, further reducing the calculation time and the probability of error. This results in a set of localised faces, one per frame captured per sample per individual. Background information is also ignored from each face frame, increasing the information gathered per face. The number of frames per sample can be significantly affected by the height of the participant. Notably, the face of the younger participants was below the camera's view towards the end of the walk.

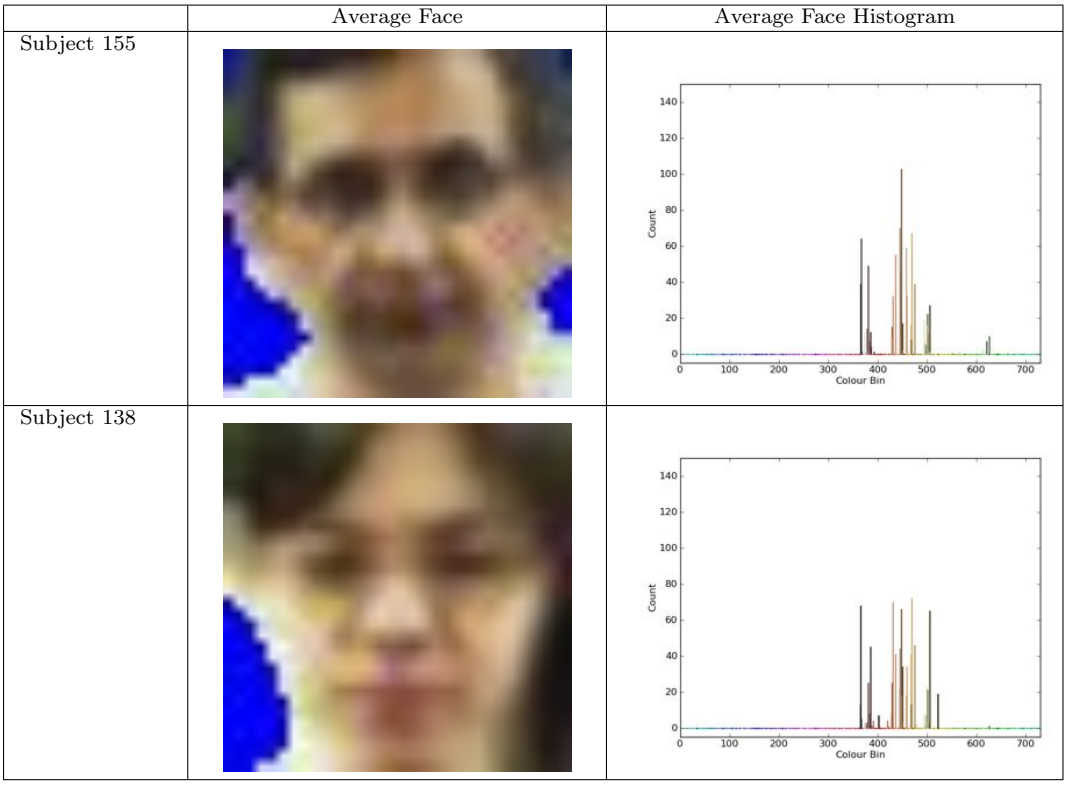


FIGURE 3.8: TunnelDB Average Face and Average Face Histogram Examples

To account for blinking, changes in expression and other sources of noise, the first signature we generate from these faces is an average of the faces localised using this technique. Firstly, each face is height normalised to a common size of 32x32, maintaining the aspect ratio and thus preserving key characteristics of the face, while allowing the comparison of face images taken at different points of the subject's walk. At this point, the same pixel across each frame of each face sample is summed. If a pixel contains no information in a given frame (i.e. if it is a background pixel) it is ignored. This means that key structure around the edges of the face are maintained. For example, if a subject has short or tied back hair, there will exist completely blank pixels around the neck area in the final signature. The summed values are then divided by the number of frames composing the summed pixel's value. This results in a single colour **Average Face Signature** per sample of each subject of size 3072 (32x32x3).

A second signature, related to the first, is also generated. Any particular average face signature inherently maintains some notion of the structure of the face. This includes the edges of the face, but also nose, lips and hairline. Some of our semantic traits such as Sex and Hair Length are likely to correlate with this structure and it may therefore prove useful. Others however may need only to correlate with absolute colours of a given average face but may incorrectly correlate with less meaningful structure. We construct a second face signature called the **Average Face Histogram**. This histogram is constructed by binning the RGB colour space into a simple (9x9x9) space; the colour space is discretised into 729 bins. Each bin in the histogram is assigned a count of the pixels in the Average Face which fall into the colour range represented by that bin. This results in a signature of size 729 which directly encodes colour while ignoring predominant structure. Example Average Face and Average Face Histograms can be seen in Table 3.8.

3.5 Semantic Recognition Experiments

In this section we outline a set of experiments used to explore the semantic annotation data we have designed and collected. In the first section we explore the relative significance of the various semantic traits. Firstly using ANOVA and secondly using Pearson's r we explore which traits are best in terms of their ability to distinguish individuals across separate annotators. Once the best semantic traits are outlined with respect to identity separation, we explore their ability in comparison to gait and face biometric signals from the two datasets. Finally, we outline a tactic of feature fusion, exploring the possible benefits of semantic annotations used in conjunction with standard biometrics.

3.5.1 Semantic Features Significance

The utility of any given trait can be explored in many ways. In Section 3.5.3 we explore each trait's identification capability while in Chapter 4 we explore the retrieval capability of individual traits. In this section we use a trait's ability to separate groups and its consistency across separate annotators to measure its usefulness. Using ANOVA and Pearson's r we investigate each trait and thus provide another set of metrics with which to gauge their relative worth. These metrics are also used as the basis for a feature set selection, allowing the maintenance of a low EER with a reduced subset of semantic traits.

3.5.1.1 ANOVA

In statistics there is a notion of significance for a given set of measurements. Generally an experiment can be described in terms of a set of groups separated by some difference in experimental conditions. Often it is of great importance to judge whether a given experimental condition *significantly* separates or maintains group distributions. If the groups are not significantly separated, one would argue that the experimental conditions made no difference, this is called the null hypothesis (H_0). If they are significant, this can be used as evidence to reject the null hypothesis and thus support a hypothesis H_1 with regards to different outcomes given different experimental conditions. To measure the significance of a single experimental variable in isolation, one can use the one-way Analysis Of Variance (ANOVA). This process calculates a statistic called the F-ratio:

$$\text{F-ratio} = \frac{\text{total between-group variance}}{\text{total within-group variance}}, \tag{3.4}$$

$$= \frac{\sum_i n_i(\bar{X}_i - \bar{X})^2/(K-1)}{\sum_{ij}(X_{ij} - \bar{X}_i)^2/(N-K)}, \tag{3.5}$$

In Equation 3.4, X_{ij} represents the sample value for the j^{th} sample of the i^{th} group. In turn, $\bar{X}_i$ is the mean of the i^{th} group's samples and $\bar{X}$ is the mean across all samples. K represents the number of group while N represents the total number of samples. Therefore, the F-ratio is a ratio of the within group variance against the between group variance weighted by the degrees of freedom $K-1$,$N-K$. The values of this statistic will be large if the between group variability is large when compared to the within group variability, which in turn is unlikely to happen if the null hypothesis is true. Put another way, this is a measure used to discover whether the groups are all the result of the same distribution, and therefore whether the effects which supposedly separate the groups actually do so significantly.

In the case of human identity, the separate groups are the different individuals to be observed and the different experimental variables are the various physical traits on which they can be semantically described. If a given trait is significant in terms of ANOVA, i.e. has a higher F-ratio, then it could be said to be more successful at separating individuals and therefore a more useful measure of identity.

In Table 3.2 we show the ordering and associated F-ratios of the physical traits as described by the semantic terms we have proposed. The ordering is a result of the non-self annotations given to subjects in the TunnelDB and HIDDB. Of note is the

TABLE 3.2: Ordering of the semantic traits by their F-ratios given the respective datasets

TunnelDB Ordering		**HIDDB Ordering**	
Feature	F-ratio $df = (59, 2630)$	Feature	F-ratio $df = (50, 817)$
Sex	675.11	Sex	383.70
Hair Length	210.16	Skin Colour	149.44
Facial Hair Length	155.87	Ethnicity	96.10
Skin Colour	131.14	Hair Length	79.05
Age	77.82	Age	57.02
Weight	67.32	Hair Colour	52.18
Height	63.58	Facial Hair Length	25.72
Hair Colour	58.67	Height	25.14
Figure	51.28	Weight	20.75
Chest	46.09	Figure	20.69
Ethnicity	42.19	Chest	18.32
Leg Thickness	32.28	Neck Length	15.57
Facial Hair Colour	31.55	Neck Thickness	14.73
Hips	31.25	Arm Thickness	13.90
Neck Thickness	28.50	Leg Length	13.68
Arm Thickness	28.12	Muscle Build	12.85
Muscle Build	27.38	Leg Thickness	11.61
Leg Length	25.49	Hips	10.55
Neck Length	18.67	Arm Length	5.74
Shoulder Shape	14.58	Facial Hair Colour	5.61
Arm Length	11.26	Leg Direction	3.25
Leg Direction	8.09	Proportions	2.77
Proportions	4.17	Shoulder Shape	2.54

comparable top and bottom halves of the two sets, containing roughly similar features though not in exactly the same order. We also note that global features such as Sex, Age and Ethnicity are more separating, while descriptions of physical features are less

so. Proportions and Leg Direction were quite uninformative in both datasets, showing their weakness or ambiguity as traits. The discrepancies between the capabilities of the traits across the two datasets are likely to be a reflection of the different viewpoints available to the two sets of annotators as well as a reflection of the contents of the datasets themselves. The ability of Facial Hair in the HIDDB for example is notably lower than in TunnelDB but this is more likely a reflection on the lower resolution of the face and facial details in the HIDDB videos compared to the facial videos available in the TunnelDB. It should be noted that the exact distribution which the F-ratios form is affected by differing degrees of freedom of a given dataset. As a result of this, the magnitude of the F-ratios cannot be compared directly but must instead be compared through their p-values extracted from the F cumulative probability distribution. Due to the tiny p-values ($p \ll 10^{-9}$) associated with the many degrees of freedom in these datasets it is meaningless to extract these values in this case and therefore impossible to directly compare the two datasets using ANOVA. However, statements regarding the relative power of traits within a given dataset remain valid. With the exploration of Pearson's r as used for feature ordering we can more rigorously compare features across the two datasets.

3.5.1.2 Pearson's r

In their paper investigating whole body descriptions, MacLeod et al. [79] used Pearson's r to to discover the stability of a given feature. In their method, the annotators of a given subject are randomly split into two groups and whose descriptions are averaged, producing two descriptions for each subject. By producing 100 such random groupings, 100 pairs of annotations are gathered per subject per annotation. By finding the correlation coefficient (See Section 2.5.3) of each semantic trait given by these random groupings we can find the semantic traits which are most correlated. This can be interpreted as those semantic traits which are most stable, or put another way: most commonly agreed upon by disparate groups of annotators.

In Table 3.3 we show the ordering and associated Pearson correlation coefficients of the physical traits. It should be stated that for the degrees of freedom in these datasets, all these correlation coefficients were significant ($p \ll 0.01$). For the most part the information presented here is as expected, agreeing with the ordering of the ANOVA F-ratios. This is to be expected, as some of the group separating ability of a given trait is undoubtedly related to its stability across several annotators. If this were not the case there would be no group separation as each individual annotation could be taken as a potential sample of any given group. We see a fairly large correlation in Sex in both datasets, showing that this is a feature for which there exists little ambiguity. We can

Table 3.3: Ordering of the semantic traits by their correlation coefficients given the respective datasets

TunnelDB Ordering		**HIDDB Ordering**	
Feature	Pearson's **r**	Feature	Pearson's **r**
Sex	0.99	Sex	0.99
Skin Colour	0.98	Skin Colour	0.97
Facial Hair Length	0.98	Age	0.95
Hair Length	0.98	Hair Colour	0.95
Age	0.93	Hair Length	0.95
Hair Colour	0.91	Ethnicity	0.94
Weight	0.91	Height	0.90
Ethnicity	0.91	Facial Hair Length	0.88
Height	0.91	Figure	0.88
Chest	0.89	Weight	0.87
Facial Hair Colour	0.89	Chest	0.84
Figure	0.88	Leg Thickness	0.80
Hips	0.85	Muscle Build	0.80
Leg Thickness	0.81	Leg Length	0.76
Neck Thickness	0.81	Arm Thickness	0.75
Leg Length	0.81	Neck Length	0.74
Arm Thickness	0.79	Hips	0.74
Muscle Build	0.77	Facial Hair Colour	0.67
Neck Length	0.68	Neck Thickness	0.63
Arm Length	0.67	Arm Length	0.56
Shoulder Shape	0.64	Leg Direction	0.36
Leg Direction	0.50	Proportions	0.34
Proportions	0.39	Shoulder Shape	0.33

reach similar conclusions for other global features in both datasets including Age, Skin Colour and Ethnicity showing them all to be reliable features. As suggested by ANOVA, Facial Hair Colour and Length is more consistently agreed upon and thus more stable

in the TunnelDB than in the the HIDDB. This, along with descriptions of the Neck and Shoulders, are probably aided in the TunnelDB by the multiple perspectives offered by the gait cameras and importantly, the face camera.

3.5.2 Semantic Significance Validation

To test whether the feature ordering recommended by Pearson's r and ANOVA are meaningful the orderings were used in a set of *Leave-one-Out (LoO)* classification tests [65]. Each test involves a set of LoO classifications using a feature vector comprised of a subset of the best traits in the order outlined by the significance tests above. The first annotation feature vector is constructed using the best trait in isolation, the next appends the second best trait and so on until progressively an annotation feature vector containing all traits is tested. For each test, the Equal Error Rate (EER) was calculated using an Receiver Operator Characteristic (ROC) as well as KNN classification with $k = 1$. An exhaustive LoO strategy was utilised with regards to annotators. A *probe set* was constructed containing a single annotator's annotations on all the subjects they had seen. The remaining annotators were used to construct a *gallery set*. This gallery set was the single, averaged description of each subject by all the annotators, minus that of the single annotator separated for the probe set. Therefore, an individual annotator was never compared to their own responses.

The Euclidean distance was then calculated between each subject in the probe set and gallery set, resulting in a distance matrix. Such a matrix was then calculated for each annotator left out as the probe. These distance matrices were used to calculate a Correct Classification Rate (CCR) using a KNN scheme as well as an ROC used to calculate an EER. These two numbers were calculated for each set of features recommended by the ANOVA and Pearson's r ordering. Furthermore, these numbers were calculated for the *reverse* ordering recommended by these schemes. The results for these tests can be seen in Fig. 3.9 to Fig. 3.12.

The results validate the feature significance ordering prescribed by both ANOVA and Pearson's r. When compared to their reverse ordering, both schemes show significantly faster improvements of CCR and EER. Both orderings also show that after roughly 50% of the more important traits are considered, the optimal recognition rates, measured both using CCR and EER, can be achieved. In both feature orderings, the first 50% of the features are the global and head traits, including: Sex, Skin Colour, Hair descriptions and Age. The general body traits such as Leg and Arm descriptions come later in both orderings and are shown here to also be less able in the reverse ordered classification experiment.

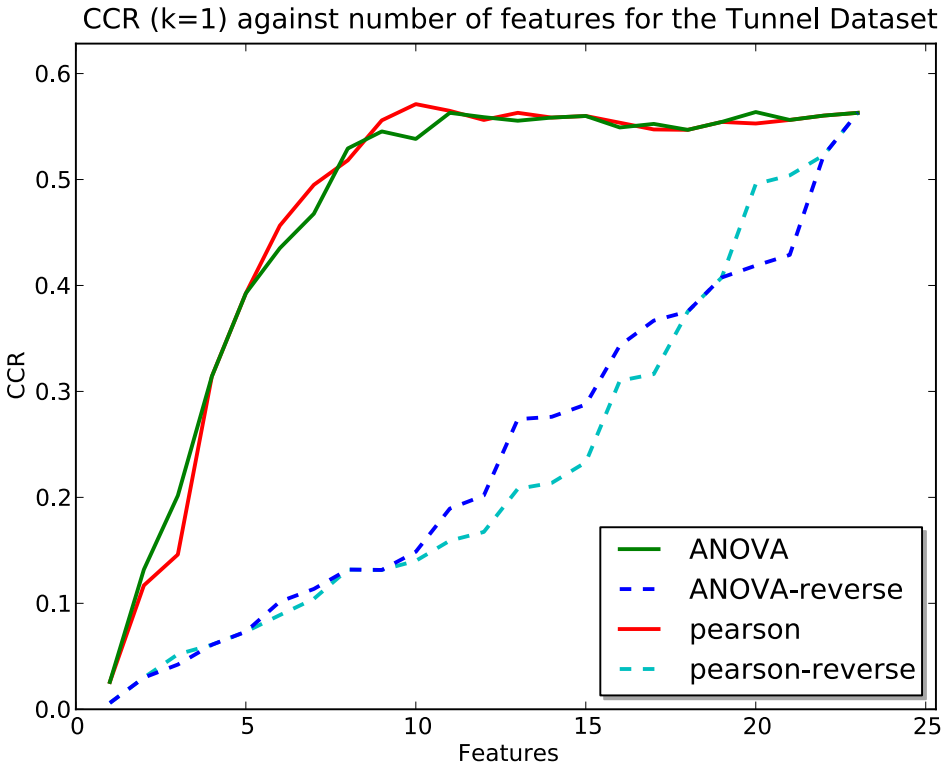


FIGURE 3.9: CCR against number of features used in the Tunnel Dataset. CCR calculated using KNN (with k=1) and a LoO classification test. The graph compares the use of features in order of significance recommended by ANOVA, Pearson's r and in the reverse order.

There exist some visible fluctuations in the results achieved by the two ordering techniques across the two datasets. We check whether these differences are significant by comparing the CCR and EER of the two ordering techniques using a one-way ANOVA. It can be shown that these deviations are not significant ($p >> 0.01$) and so there is no major difference between the two approaches to ordering feature significance. The orderings of traits produced by the two approaches are both useful pieces of evidence in discovering which human trait is most useful when semantically described. This is discussed in more detail in Chapter 5

3.5.3 Fusion Experiments

In this section we explore the ability of the semantic annotations gathered against existing visual biometrics. We also implement a simple feature fusion and score fusion strategy to show the ability of the gathered semantic annotations in fusion with existing biometric signals. Both the ability of visual biometrics and fusion experiments are performed against the sources of visual features outlined in Section 3.4.2, namely

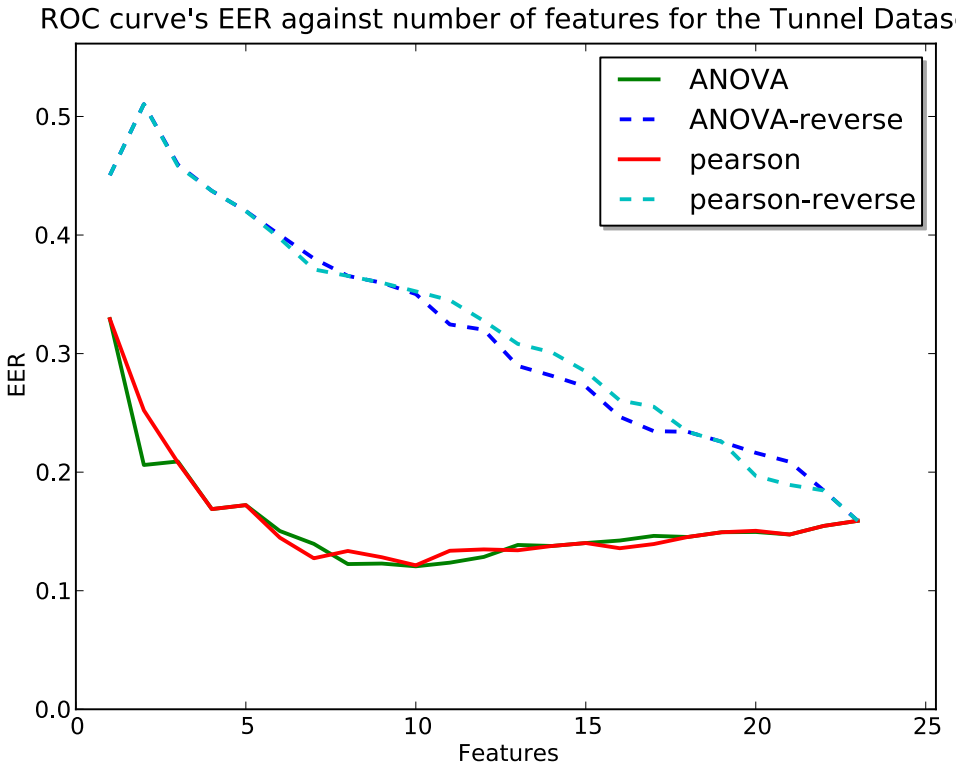


FIGURE 3.10: EER against number of features used in the Tunnel Dataset. EER calculated by plotting an ROC, finding the threshold resulting in an equal number of FPs and FNs. The graph compares the use of features in order of significance recommended by ANOVA, Pearson's r and in the reverse order.

the Average Silhouette and Colour Average Silhouette from the HIDDB and the two Projected Gait and two Average Face signatures from the TunnelDB.

3.5.3.1 Approach

In following section several performance ROC curves are shown. Individual graphs depict the annotations of a dataset, a single visual feature of the dataset and the fusion of semantic annotations with this visual feature. The ROC curves are generated from a LoO classification scheme.

Unimodal Biometrics

This scheme is firstly used to gauge the performance of unfused signatures. For the visual signatures, a single sample from a single subject is separated as the probe set and compared to the rest of the samples of all the other subjects as the gallery set. For annotations a similar strategy to Section 3.5.2 is undertaken where each annotator is

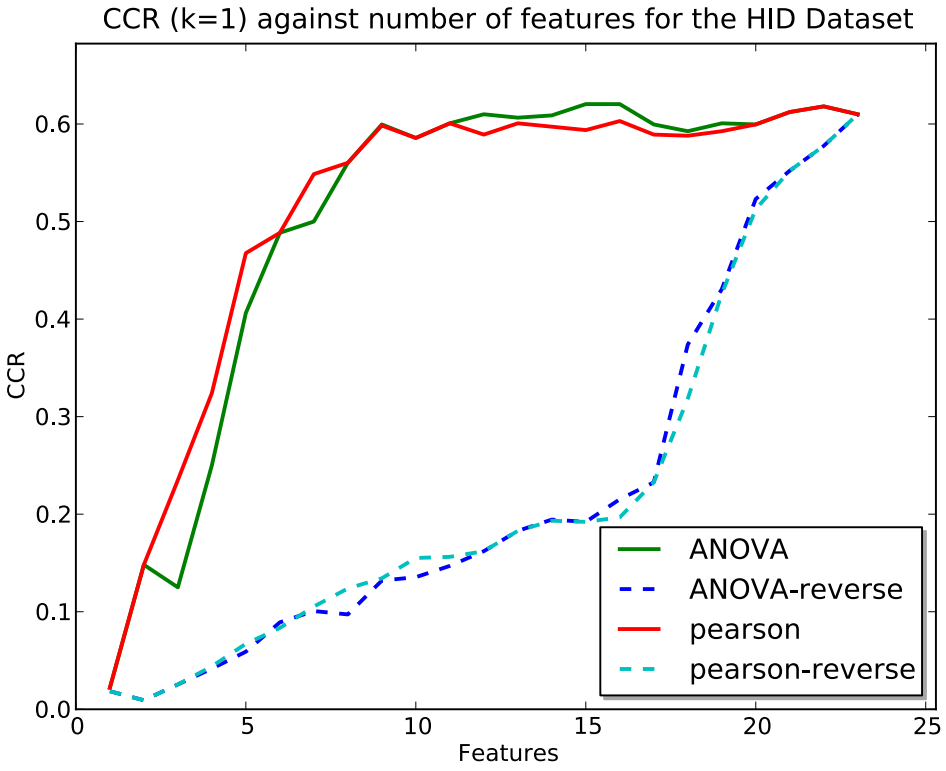


FIGURE 3.11: CCR against number of features used in the HID Dataset. CCR calculated using KNN (with k=1) and a LoO classification test. The graph compares the use of features in order of significance recommended by ANOVA, Pearson's r and in the reverse order.

left out as the probe set and compared to the averaged response of each other annotator for a given subject.

Semantic Biometrics Fusion

Once these unimodal biometrics are examined, two fusion strategies are employed to test semantic features in fusion with existing biometrics.

Feature Fusion - The first fusion strategy undertaken is a simple normalised feature fusion. This approach assumes independence between the automatic and semantic data sources and concatenates the two feature domains. We also assume that the semantic annotations generated for a particular subject's sample would have been generated identically across all samples of that subject. This is reasonable as annotators were given access to all sample videos of a subject when making annotations. To fuse annotations firstly the annotations are averaged so there exists a single consensus annotation per subject. Then **all** visual samples of each subject are extended with the semantic features of that subject. To make such a concatenation valid, a simple min-max normalisation is

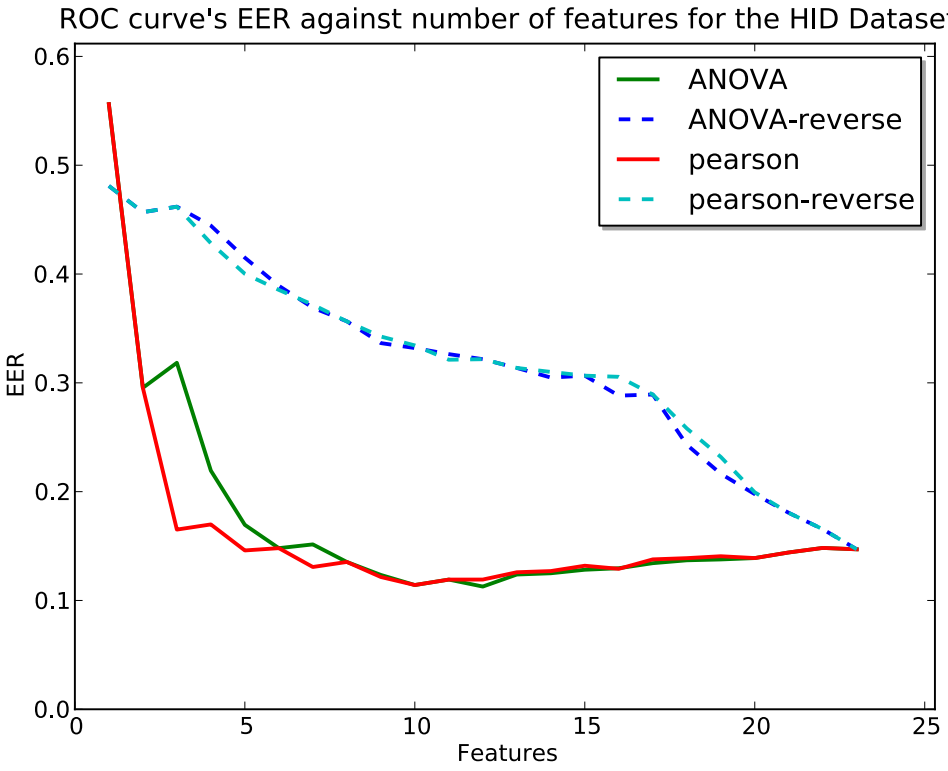


FIGURE 3.12: EER against number of features used in the HID Dataset. EER calculated by plotting an ROC, finding the threshold resulting in an equal number of FPs and FNs. The graph compares the use of features in order of significance recommended by ANOVA, Pearson's r and in the reverse order.

employed on the visual features meaning that each component of a given feature vector maintains its relative magnitude with relation to other features in the feature vector, but now lies within the range of 0 and 1.

Score Fusion - The second fusion strategy undertaken is a transformation-based score fusion (See Section 3.3.2.3). In this scheme, a distance matrix is separately generated using the automatic biometric features and semantic features in isolation. The two distances are then normalised. The normalisation factor for the annotations is taken as being the number of traits, as even when represented by trait, the maximum two annotations could be away from each other is if they disagreed on every physical trait. The normalisation factor for the visual features was calculated as the largest distance two signatures could be away from each other, namely if each pixel disagreed between two samples. In the case of the HIDDB's Average Gait Signature, the features being compared are 4096 normalised pixels, therefore the maximum Euclidian distance two samples could theoretically be from each other is if all pixels disagreed, therefore 64 ($\sqrt{\sum_{i=1}^{n=4096} 1^2} = \sqrt{4096} = 64$). Upon normalisation, a simple sum-rule was used, fusing the two signatures. It is also possible to fuse these scores with different weightings

applied to annotations and visual signatures. In doing so we can find the weighting of each signature which produces optimal classification results, as well as an improved understanding of which signature is more important for purposes of accurate classification.

To perform a LoO classification on these fused results, a probe and gallery set must be carefully generated. The probe set is constructed from a single visual sample fused with its semantic annotations as attributed by a single annotator. The gallery set contains the remaining samples fused with the averaged annotations of the remaining annotators. This is repeated with all combinations of all annotators and samples. We use these combinations to perform an exhaustive LoO test, generating an ROC and calculating an EER for the visual features fused with annotations. We also perform this test using the visual features and annotations in isolation, finding the non-fused EERs. In the next section, we present these results for each visual feature in both datasets. Finally, we also present a set of results depicting the effect of varying the weighting between visual and annotation signatures on the EER in score fusion.

3.5.3.2 Results

Fig. 3.13 to Fig. 3.18 shows the results of LoO experiments for both fused and non-fused features across both datasets. These results show that universally annotation features in isolation perform less effectively than automatic visual features from all datasets. However, regardless of the relative weakness of annotations, the fusion of annotations and visual features out-performs visual features in isolation in all feature sets in both datasets. This is the case both in feature fusion and in score fusion, though there is some discrepancy between the two results. The extent to which semantic annotations aid automatic visual features varies depending on the dataset and fusion scheme. In this case, Projected Gait signatures from the TunnelDB are assisted most, with an increased EER of roughly 3.89%, while Average Face signatures are aided least with a small improvement of 0.01% in feature fusion.

Fig. 3.19 presents the results of variable weightings between visual and annotation signatures in score fusion. We note that improved EERs can be achieved through exploration of appropriate weightings between the signatures. We also note that most signatures achieve an optimal EER below a 50-50 split between the two signatures, instead achieving an optimal score with a weighting between 0.2 to 0.4 for the visual signatures and a corresponding weighting of 0.8 to 0.6 for annotation signatures. We note an improvement of between 0.15% and 2.16% when selecting these weightings over the standard

FIGURE 3.13: ROC for HIDDB annotations with Average Gait Signatures

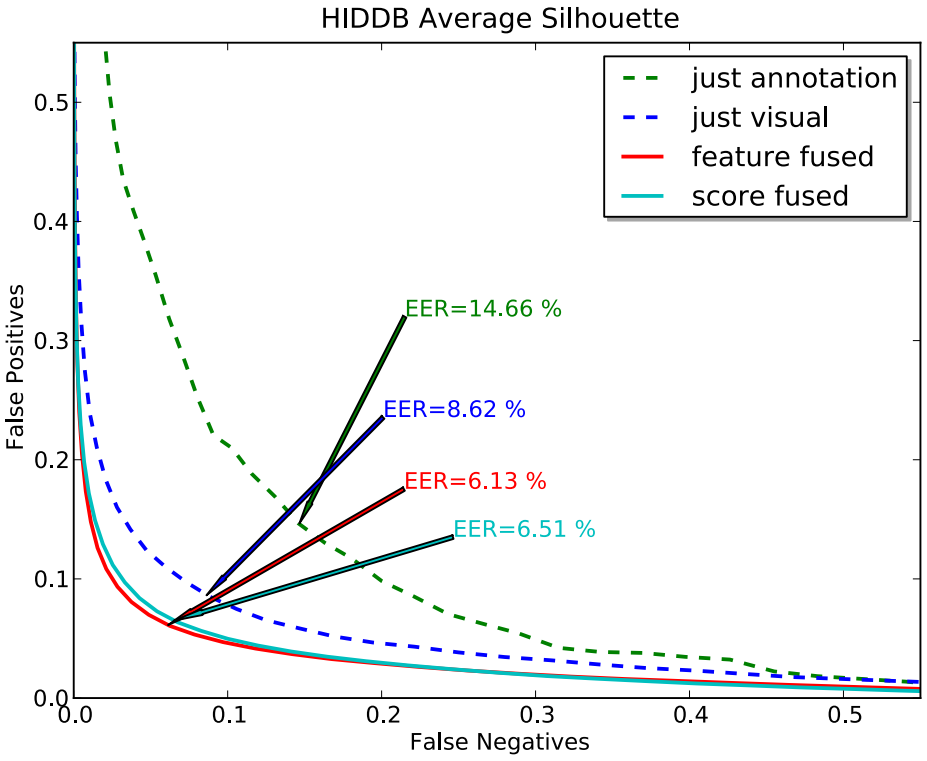


FIGURE 3.14: ROC for HIDDB annotations with Average Colour Signatures

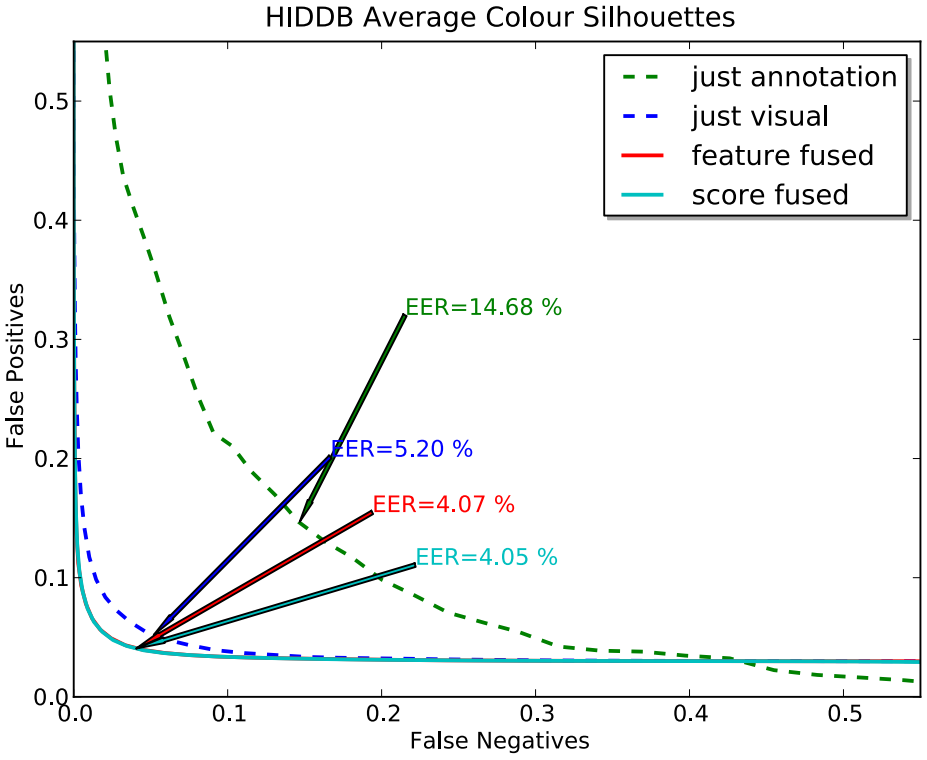


FIGURE 3.15: ROC for TunnelDB annotations with Projected Gait signatures

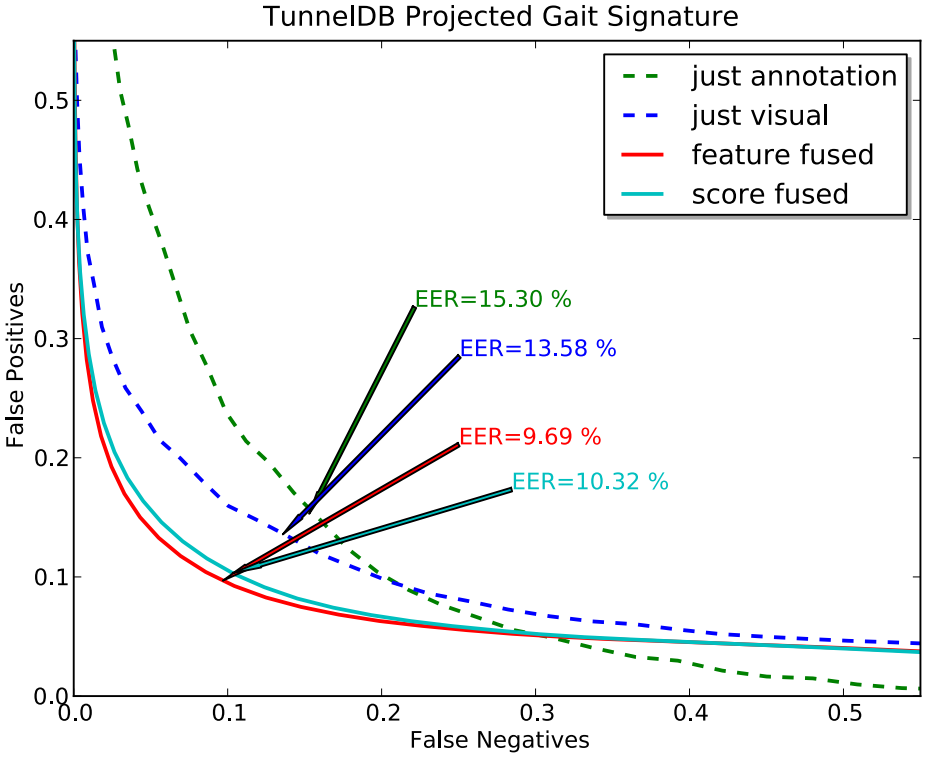


FIGURE 3.16: ROC for TunnelDB annotations with Projected Non-Normalised Gait signatures

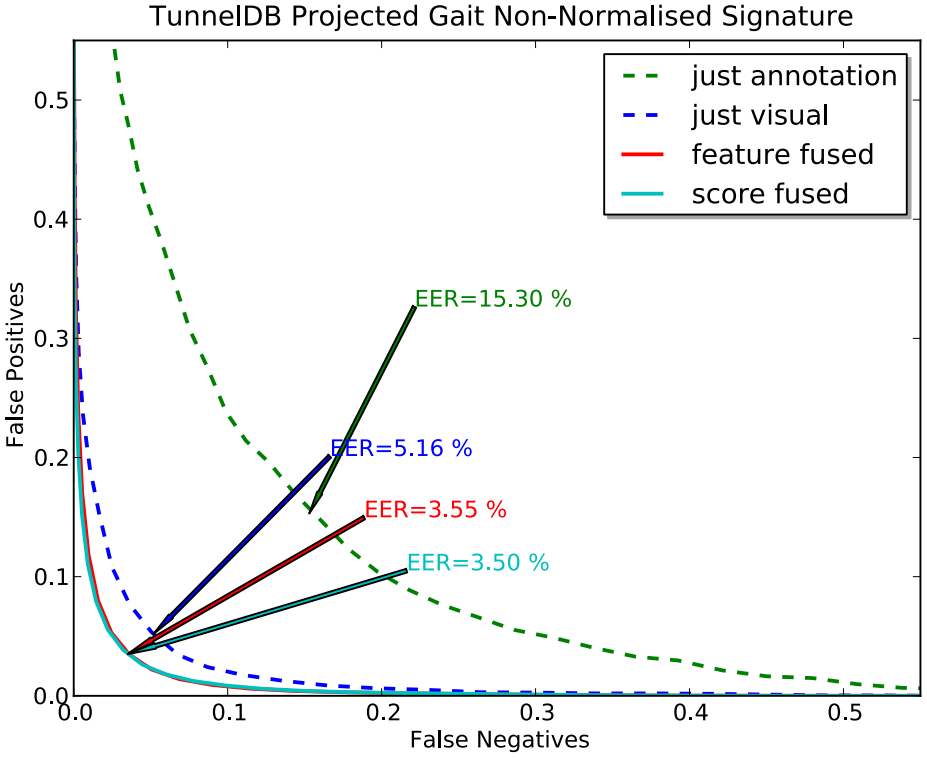


FIGURE 3.17: ROC for TunnelDB annotations with Average Face signatures

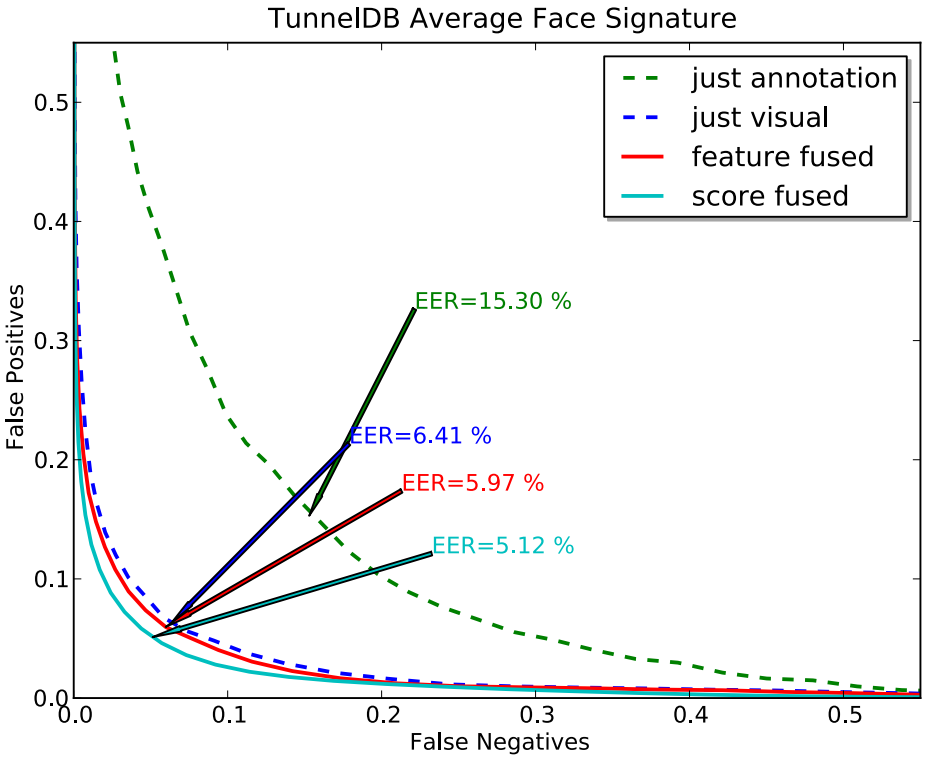
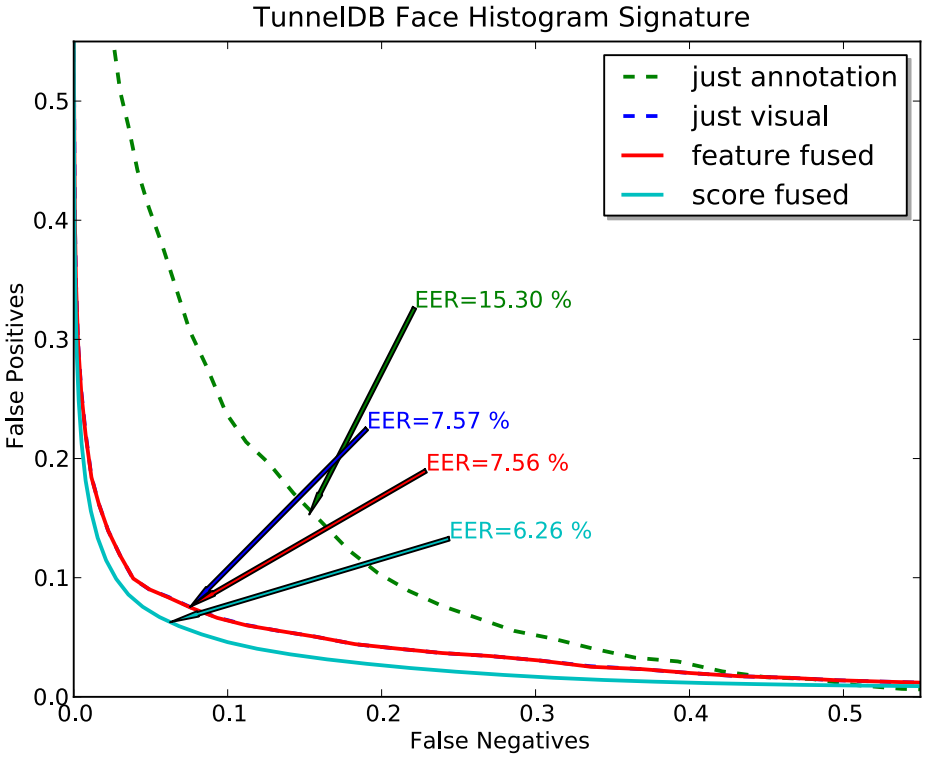
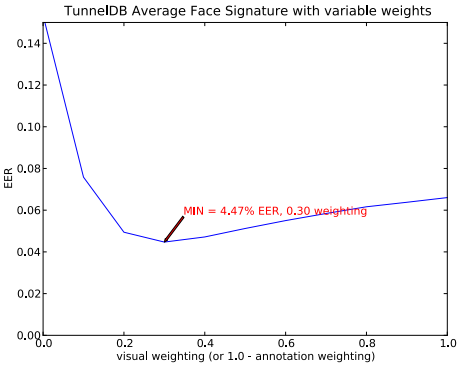
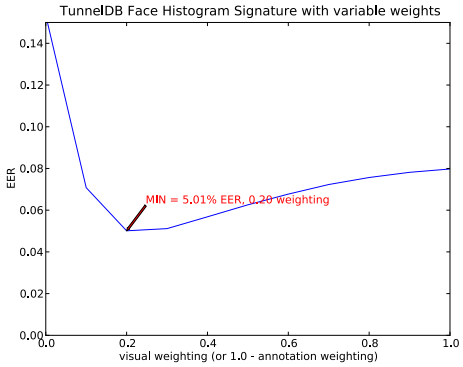


FIGURE 3.18: ROC for TunnelDB annotations with Average Face Histogram signatures

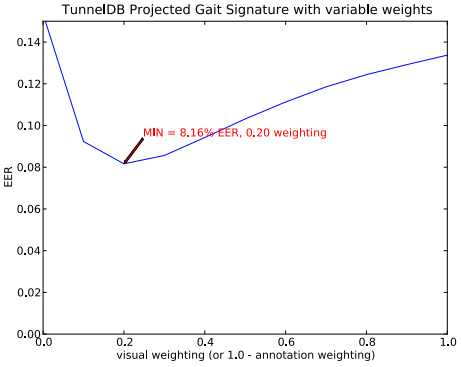




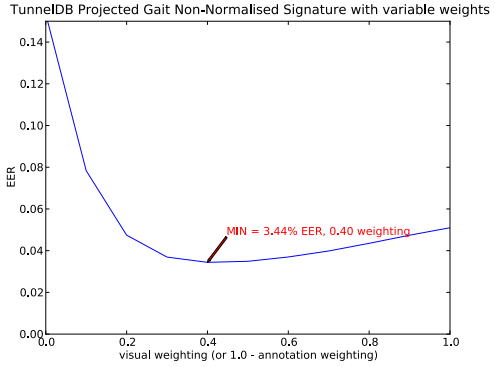
(a) Average Face



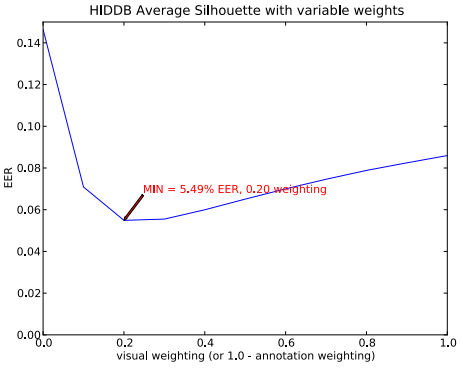
(b) Average Face Histogram



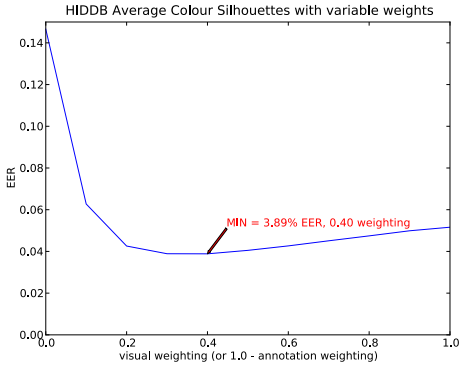
(c) Projected Gait



(d) Projected Non-Norm Gait



(e) HIDDB Average Silhouette

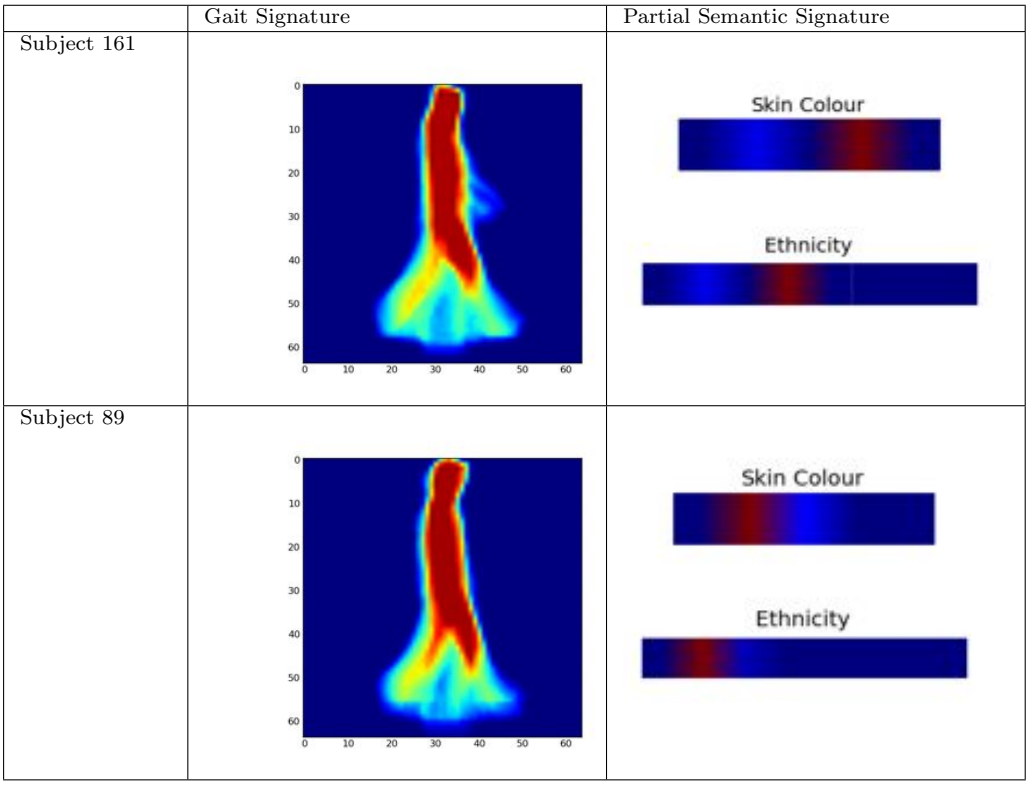


(f) HIDDB Average Colour Silhouette

FIGURE 3.19: Subfigures (a)-(f) showing the effect of alternative weightings between annotations and visual signatures in score fusion

50-50 weighting. These results show that though annotation signatures are comparatively weaker than visual signatures in isolation, better results are achieved by weighting annotation signatures higher than visual signatures in score fusion.

FIGURE 3.20: An example of a corrected TunnelDB misclassification



It has been shown that if two classifiers are negatively dependent, improved classification is to be expected from fusion [69]. Therefore, upon closer inspection of specific cases, it becomes clear how semantic annotations are achieving these improved results. For example, in Fig. 3.20 we see a common misclassification made using the reprojected Gait signatures from the TunnelDB. Here subject 89 was misclassified as 161. The gait signatures of both subjects were visually similar, due to the fact that both subjects are small females. However, the subjects did have dissimilarities, the foremost being Ethnicity: 161 is Asian and 89 is European. On average across most annotators this feature was agreed upon and therefore a correct classification was made in both fusion strategies. Inherently semantic notions such as ethnicity are not encoded explicitly in grayscale projected gait signatures, however we explore how correlation of some visual features can help the automatic estimation semantic annotations against those features in the following chapter.

3.6 Conclusions

We have introduced the use of semantic human descriptions as a soft biometric. We outlined the procedure undertaken to transform a set of semantic annotations for use as a biometric signature. To explore the semantic feature's ability in fusion and also to provide a point of comparison, we outlined a set of automatic visual biometric signatures across two datasets. Using ANOVA and Pearson's r we explored the notion of the most important semantic traits, our results confirming prominent traits of previous studies. Finally, we have shown that semantic traits have inherent identification capability and also that they can successfully improve identification results of a primary biometric when combined in both score and feature fusion.

The following chapter uses semantic annotations for the related, though separate, task of retrieval. We show that with some simple mathematical models, the same semantic terms used in this section of recognition can be used for semantic query and Content Based Information Retrieval (CBIR).

Chapter 4

Content-Based Analysis

4.1 Introduction

In the previous chapters we have highlighted the capability of semantic features for purposes of recognition. In this chapter we explore a logical extension of this recognition ability, namely whether features which can be used for recognition can also facilitate retrieval. Towards efficient human usage of large collections of surveillance data, media items should be traversable through a semantic query and therefore meaningfully *semantically transcoded* or *annotated*. The desire for such searching ability was shown in the example of the Hampshire Police RMS in Section 2.2.2. However, surveillance datasets suffer from issues presented by the multimedia semantic gap [138], a requirements gap between semantic queries which users readily express and which systems cannot answer.

Semantic descriptions have been discussed [47, 127] as an open area of interest in surveillance. This includes a mapping between behaviours and the semantic concepts which encapsulate them. Although some efforts in the past have attempted to bridge this gap [127] for behavioural descriptions, no attention as of yet has been devoted to semantic appearance descriptions. As discussed in Chapter 2, semantic descriptions are a natural way to describe individuals. Their use is abundant in character description in fiction and non-fictional narrative, with a few key words such as *slender* or *stout* helping readers understand what characters look like, therefore understanding plot in a richer context. In a more practical capacity, stable physical descriptions are of key importance in eyewitness crime reports, a scenario where human descriptions are paramount as high detail images of assailants are, generally, unavailable. Such features are challenging to extract and analyse automatically and yet are readily discernible from surveillance videos by humans. Unfortunately, the manual annotation of videos is a laborious [22, 47] process, too slow for effective use in real time CCTV footage and vulnerable to various

sources of human error (subject variables, anchoring etc.). Automatic analysis of the way people walk [89] (their gait) and analysis of their face are efficient and effective approaches towards collecting human features at a distance. Yet automatic techniques such as face and gait do not necessarily generate signatures which are immediately comprehensible by humans.

In this chapter, we explore how Latent Semantic Analysis (LSA) techniques can be used to associate the semantic physical descriptions presented in Chapter 2 with automatically extracted visual features such as gait and face. We show the resulting retrieval of unannotated surveillance footage based on semantic queries. Furthermore, we outline the possibility of automatically inferred semantic variables (automatic annotation) in the application of improved recognition of unannotated individuals. In doing so, we outline a novel application for semantic physical descriptions.

The rest of this chapter is organised in the following way. In Section 4.2 we describe LSA using the Singular Value Decomposition (SVD), the technique explored to bridge the gap between semantic physical descriptions and gait signatures. We proceed to the methodology by which semantic retrieval and also improved recognition of unannotated samples can be achieved through LSA, presenting an example of this approach. Once the methodology is outlined, we discuss the process by which the annotations and automatic signatures outlined in the previous chapters are used in tests to explore semantic retrieval (See Section 4.3). Finally, in Section 4.4, we discuss the final results and what has been achieved in this chapter.

4.2 Latent Semantic Analysis

4.2.1 History

The analysis of documents, especially those containing text, has been an active area of interest throughout human history. This is exemplified by tables of contents in books and the Dewey decimal system. Indeed during the 1960s, corresponding to the beginning of the information age, major research efforts were focused around the automatic computational indexing of large, primarily textual, corpuses. Furthermore, the idea of the representation of documents as vectors of weightings of terms is not new, with works such as Salton et al. [108] presenting an early discussion on this topic in the mid 1970s. Although useful for mathematical conceptualisation of the problem, vector representations alone are, in essence, direct lexical comparisons of terms in documents to ascertain similarity to other documents and to queries. In this approach *synonyms* (i.e. different words sharing the same meaning) and *polysemes* (i.e. single words having

different meanings in different contexts) cause major problems with regards to incorrect classification and subsequently retrieval rate deterioration.

In their seminal work Deerwester et al. [24] present a well received extension to the notion of documents as vectors of terms, attempting to address the problem of synonyms and polysemes. Briefly (see Section 4.2.2 for a more in-depth discussion), their approach, dubbed LSA, uses SVD to extract a set of linearly independent latent concepts from a term-document matrix $\mathbf{O}$, represented by the set of eigenvectors forming an orthonormal basis for $\mathbf{O}^T\mathbf{O}$ and $\mathbf{O}\mathbf{O}^T$ (the document and term co-occurence matrices respectively, see Section 4.2.2). Their argument is that documents in the corpus and their associated terms are in fact artefacts of this set of generative underlying concepts. Furthermore, Deerwester et al. [24] argue that by selecting only the eigenvectors with the k largest eigenvalues to represent this underlying set of concepts, improved retrieval rates can be achieved by projecting documents into this space prior to comparison. The commonly cited argument is that this is due to eigenvectors with smaller eigenvalues representing so called "obscuring noise" in the model [81], however a more rigorous explanation has been suggested by Papadimitriou et al. [94] (see Section 4.2.2 for more details).

Given its conceptually satisfactory model and relative success, as compared to simple lexical comparison, LSA (also known as Latent Semantic Indexing (LSI)) has been applied to various applications in different fields. An early review by Berry et al. [10] showed a great deal of interest in LSA from the text information retrieval community in the early 1990s. An initial work by Dumais [25] presents the application of LSA to querying of automatically indexed bibliographic citations. Their results show a 30% improvement when compared to simple lexical matching. This initial work also explores the benefits of weighting terms according to their appearance in a document, their appearance in the whole corpus or various other weighting techniques. Another notable use in text retrieval is presented by Landauer and Littman [72] who use LSA (under the name Cross Language Latent Semantic Indexing (CL-LSI)) to achieve the automatic translation of French documents to English using a training matrix containing contextual usage of a corpus of French and English words.

Following the success enjoyed by LSA in the text retrieval domain it is of no surprise that many attempts have been made to duplicate the performance in other problem areas, namely that of the automatic retrieval of images or CBIR. This is an active field of research with several interesting approaches [17, 36, 39–42, 96, 97, 137], a few of which we shall summarise here.

Initial research [17, 36, 96, 137] in achieving CBIR using LSA concentrated around improving query by example by utilising semantic annotations in conjunction with so

called "visual terms" to construct the concept space. The first mention of such an approach was presented by Pecenovic [96] in 1997. The author concentrates on a query by example backed by a relevance feedback approach where image features and their semantic features are compared wholesale after being projected in a rank reduced concept space. Efforts are made to quantise continuous visual features in the form of binned colour, texture and block correlations alongside semantic features. Another notable work presented later by Grosky and Zhao [36], Zhao and Grosky [137] concatenates 15 semantic category features (which they call *category bits*) with global and local colour histograms. In their experiments they construct a concept space with images represented fully by their visual and semantic components. The semantic components of their query documents are artificially set to 0s, both before and after projection into the concept space. Using this approach, they present improved classification results when compared to a concept space constructed using visual components alone. This work exposes the positive effect semantic components have on weighting the non-semantic components of the concept space.

Later, several attempts [39–42, 85, 97] were made to go beyond simple usage of LSA to improve image to image comparison and instead use it to automatically prescribe or retrieve unannotated images using text annotations. These approaches generally construct a concept space using a standard LSA performed on a fully observed (i.e. both semantically and visually) training matrix. Documents containing no annotations can be projected into the concept space with their annotations terms set to 0. Retrieval queries are constructed as pseudo documents such that visual features are set to zero values and appropriate semantic features are set to non zero values. Retrieval is now a simple matter of comparing the cosine distance of the projected retrieval query with the projected unannotated documents. Automatic annotation is also possible if the projected documents are compared to the position of semantic terms in the concept space (see Section 4.2.3 for more details of this approach).

4.2.2 The Singular Value Decomposition

In this section we go into further detail with regards to the usage of the SVD for LSA. We start by constructing an $n \times m$ occurrence matrix $\mathbf{O}$ whose values represent the *presence* of n terms in m documents (columns represent documents and rows represent terms). In our scenario, documents represent individual samples of individual subjects. Semantic features and automatic biometric features are considered to be terms. The "occurrence" of an individual visual feature signifies the magnitude of that portion of the feature vector while the "occurrence" of a semantic term signifies its semantic relevance to the subject in the video given an individual annotation or the average annotation of a set of

individuals. Our goal is the production of a rank reduced factorisation of the observation matrix consisting of a two orthogonal basis matrices. These matrices are the term matrix $\mathbf{T}$ which can represent the space of terms and the document matrix $\mathbf{D}$ for representing the space of documents, such that:

$$\mathbf{O} \approx \mathbf{TD}. \tag{4.1}$$

The row vectors in $\mathbf{T}$ and $\mathbf{D}$ represent the location of individual terms and documents whereas their columns represent two related sets of orthogonal bases. The orthogonal vectors making up the columns of $\mathbf{T}$ are in fact weightings against a set of terms, therefore they can be thought of as a set of basis documents. The vectors making up the columns of $\mathbf{D}$ are in fact weightings against a set of documents, therefore they can be thought of as a set of basis terms.

Once these matrices are calculated, novel terms and novel documents can be projected into the appropriate space and compared with other terms and documents. Benefits to retrieval, recognition and annotation arise when certain basis documents and terms are discarded, i.e. rank reduced spaces are used for projection.

In practice, these orthogonal sets of document and term bases held in $\mathbf{T}$ and $\mathbf{D}$ in have been calculated in a variety of ways in existing LSA research [81]. These methods include: the QR factorisation [11, 50], the ULV low-rank orthogonal decomposition [9] and the semi-discrete decomposition (SDD) [66]. Whilst these methods are viable options, the most popular and common approach by far is to calculate $\mathbf{T}$ and $\mathbf{D}$ using the Singular Value Decomposition (SVD) which is defined as:

$$\mathbf{O} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \tag{4.2}$$

Such that $\mathbf{T} = \mathbf{U}$ and $\mathbf{D} = \mathbf{\Sigma}\mathbf{V}^T$. The rows of $\mathbf{U}$ represent positions of the terms of $\mathbf{O}$ while its columns represent the orthogonal dimensions used to represent these terms; the aforementioned basis documents or *eigen-documents*. The rows of $\mathbf{V}$ represent the position of the documents of $\mathbf{O}$ while its columns represent the orthogonal dimensions used to represent these documents, the aforementioned basis terms or *eigen-terms*. The diagonal entries of $\mathbf{\Sigma}$ are equal to the singular values of $\mathbf{O}$. The columns of $\mathbf{U}$ and $\mathbf{V}$ are, respectively, *left-* and *right-singular* vectors for the corresponding singular values in $\mathbf{\Sigma}$. The singular values of any $n \times m$ matrix $\mathbf{O}$ are defined as values $\{\sigma_1, .., \sigma_r\}$ such that :

$$\mathbf{O}\mathbf{v_i} = \sigma_i \mathbf{u_i}, \tag{4.3}$$

and

$$\mathbf{O}^T\mathbf{u_i} = \sigma_i \mathbf{v_i} \tag{4.4}$$

Where $\mathbf{v}_i$ and $\mathbf{u}_i$ are defined as the right and left singular vectors respectively.

In can be shown that $\mathbf{v}_i$ and $\mathbf{u}_i$ are in fact the *eigenvectors* with corresponding *eigenvalues* $\{\lambda_1 = \sigma_1^2, .., \lambda_r = \sigma_r^2\}$ of the square symmetric matrices $\mathbf{O}^T\mathbf{O}$ and $\mathbf{O}\mathbf{O}^T$ respectively, referred to as the *co-occurrence* matrices. The matrix $\mathbf{U}$ contains all the eigenvectors of $\mathbf{O}\mathbf{O}^T$ as its rows while $\mathbf{V}$ contains all the eigenvectors of $\mathbf{O}^T\mathbf{O}$ its rows and $\mathbf{\Sigma}$ contains all the eigenvalues along its diagonal. Subsequently:

$$\mathbf{O}^T\mathbf{O} = \mathbf{V}\mathbf{\Sigma}^T\mathbf{U}^T\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T = \mathbf{V}\mathbf{\Sigma}^T\mathbf{\Sigma}\mathbf{V}^T, \tag{4.5}$$

$$\mathbf{O}\mathbf{O}^T = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T\mathbf{V}\mathbf{\Sigma}^T\mathbf{U}^T = \mathbf{U}\mathbf{\Sigma}\mathbf{\Sigma}^T\mathbf{U}^T. \tag{4.6}$$

To intuitively appreciate the importance of SVD in this context and the eigenvector matrices $\mathbf{V}$ and $\mathbf{U}$ for information retrieval purposes, consider the meaning of the respective co-occurrence matrices.

$$\mathbf{T_{co}} = \mathbf{O}\mathbf{O}^T, \tag{4.7}$$

$$\mathbf{D_{co}} = \mathbf{O}^T\mathbf{O}. \tag{4.8}$$

The magnitude of the values in $\mathbf{T_{co}}$ relate to how often a particular term appears with every other term throughout all documents, therefore some concept of the "relatedness" of terms. The values in $\mathbf{D_{co}}$ relate to how many terms every document shares with every other document, therefore the "relatedness" of documents. By definition, the matrix of eigenvectors $\mathbf{U}$ and $\mathbf{V}$ of the two matrices $\mathbf{T_{co}}$ and $\mathbf{D_{co}}$ form two bases for the co-occurrence spaces, i.e. the combination of terms (or documents) which the entire space of term co-occurrence can be projected into without information loss. The eigenvectors or bases of these two matrices subsequently represent the principal ways in which terms and documents co-occur. The more highly weighted directions represent main ways in which documents and terms co-occur. This could be thought of as the main *concepts* of the set of documents and term in $\mathbf{O}$

Therefore, having attained these bases, the improved representation of document similarity is achieved by using only eigenvectors of $\mathbf{U}$ and $\mathbf{V}$ corresponding to the k highest eigen values:

$$\mathbf{O}_k = \mathbf{U}_k\mathbf{\Sigma}_k\mathbf{V^T}_k \tag{4.9}$$

By selecting an appropriate value for k we can guarantee that minimal information is lost according to Eckart and Young [26]:

Theorem 4.1. *"Among all $n \times m$ matrices C of rank at most k, O_k is the one that minimises $||O - C||_F^2 = \Sigma_{i,j}(O_{i,j} - C_{i,j})$"*

This is the theorem often cited by Berry et al. [10] for the improved performance gained by choosing only k largest eigenvectors $\mathbf{U}_k$ and $\mathbf{V}_k$ as part of LSA. However, this only explains why LSA *does not deteriorate too much* from the true answer and not why a notable improvement in performance is achieved. Only after a decade of interest in the technique was a convincing argument suggested by Papadimitriou et al. [94] discussing the cause of the success of this approach which chooses only the components of $\mathbf{U}$ and $\mathbf{V}$ related to the largest eigenvalues as concepts. In summary, under the assumption that each term belongs to one and only one concept and furthermore that each document also contains only one concept, it can be shown that the eigenvectors of the k largest eigenvalues have a highest probability of being the sole eigenvectors necessary for representing each concept. Though the assumptions are somewhat restrictive, they guarantee that if these top k eigenvectors alone are chosen to represent $\mathbf{O}$, two documents will be projected onto some scalar multiple of an eigenvector in $\mathbf{V}_k$ if they are from the same underlying concept and onto some orthogonal pair otherwise. Papadimitriou et al. [94] go on to show that this statement holds under small perturbations of $\mathbf{O}$. Subsequently if two vectors are projected into the reduced concept space and their similarity measured using a cosine metric[1] the LSA procedure is likely to force similar documents close to each other and dissimilar documents further apart.

4.2.3 Using the Singular Value Decomposition

With these insights, our task becomes the generation of an observation matrix for a set of subjects comprising of semantic terms and visual features in feature fusion. Once this matrix is generated, several tasks can be performed and improved by exploiting the projection of partially observed vectors into the eigenspace represented by either $\mathbf{T}$ or $\mathbf{D}$.

Assume we have two subject collections, a fully annotated training collection and a test collection, lacking semantic annotations. A matrix $\mathbf{O}_{train}$ is constructed such that training documents are held in its columns. Both the visual and the semantic terms are fully observed for each training document, i.e. a term is set to a non-zero value encoding its existence or relevance to a particular video. Using the process described in Section 4.2.2 we can obtain $\mathbf{T}_{train}$ and $\mathbf{D}_{train}$ for the training matrix $\mathbf{O}_{train}$ using the SVD. In turn a matrix $\mathbf{O}_{test}$ is constructed such that test documents are held in

[1] This explanation for the success of LSA also explains why the cosine distance metric is needed. By using the angle between two vectors as a metric for similarity, the scalar multiplier of each concept is ignored and only the relation with the concept itself is considered.

its columns. However, in $\mathbf{O}_{test}$ only visual features of documents are observed while semantic features are set to 0.

Content-Based Retrieval

The retrieval of unannotated documents in $\mathbf{O}_{test}$ against some semantic query is one scenario aided through LSA. This task is thought to be common given the prevalence of CCTV surveillance video. Operators may want to answer questions such as "Which video contains a person of this given description?". This task would be impossible given an unannotated surveillance video database, but is made possible given an existing training set of surveillance videos. In this scenario both the unannotated document being retrieved and the query retrieving it are considered *partially observed* documents; while the documents in $\mathbf{O}_{test}$ lack semantic descriptions, the query document lacks all its visual components and all but some of its semantic components. By carefully projecting these matrices into the semantic space it is possible gauge their relative positions in the space of concepts and therefore compare a semantically unobserved document in $\mathbf{O}_{test}$ to a semantic query.

A new partially observed document matrix $\mathbf{O}_{test}$ is constructed holding all documents to be retrieved with semantic terms set to zero. Similarly a partially observed document matrix $\mathbf{O}_{query}$ is constructed for the query where all visual and non-relevant semantic terms are set to zero while relevant semantic terms are given a non-zero value [2]. These matrices are now projected in the latent space in following manner:

$$\mathbf{D}_{test} = \mathbf{T}_{train}^{T}\mathbf{O}_{test}, \tag{4.10}$$

$$\mathbf{D}_{query} = \mathbf{T}_{train}^{T}\mathbf{O}_{query}. \tag{4.11}$$

Projected test documents held in $\mathbf{D}_{test}$ are ordered according to their cosine distance to query documents in $\mathbf{D}_{query}$ for retrieval. We explore the ability of semantic retrieval against our datasets in Section 4.3.

Semantically Mediated Identification

Another area involving the individuals represented by documents in $\mathbf{O}_{test}$ which gains benefit through LSA of a training set is the improved performance in a biometric identification task. The document concept space discovered by LSA from $\mathbf{O}_{train}$ is in fact a set of basis vectors in the space of features, therefore the concept vectors themselves can regarded as a series of weightings against which both visual and semantic features

[2] usually 1.0, but weighting corresponds to importance of a term in a query

gain or lose significance. In our case where the concept space was trained not only on visual features in isolation, but in fusion with semantic features, it is reasonable to assume that these weightings were effected by the co-occurrence and therefore underlying relationship between visual features and semantic features. One of the main arguments of this thesis is that semantic features represent some readily comprehensible underlying space in which humans are separable. Therefore we postulate that improved retrieval rates of subjects given a set of completely unannotated samples can be achieved by first projecting these *partially observed* documents into the concept space. In doing so the visual features of the samples will be weighted according to their relevance to the semantic features and so identification must improve.

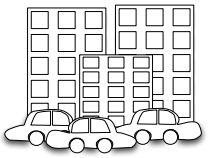
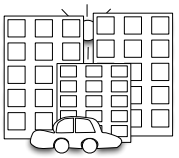
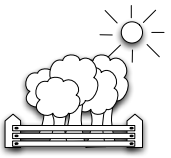
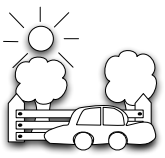
4.2.4 An Example: LSA using the SVD

Having discussed the specifics of the SVD as used in LSA, what remains is to present an example. In the following section, LSA and CBIR are performed on an very small, synthetic dataset. The goal is understanding of the less obvious elements of the process and its benefits, more specifically how the representation of documents and terms in a space made of orthogonal basis vectors can serve to separate dissimilar documents and terms.

4.2.4.1 Cars, Trees and the Sun

Imagine four pictures, each containing combinations of cars, buildings, trees and the sun. From each picture, or document, one can automatically extract visual features describing visible components of the image (assuming a sufficiently powerful computer vision algorithm). For example: cars can be described as being a car shape and two visible circles (wheels) while the sun can be described as a single visible circle surrounded by beams of light. There also exist visible features of the images which cannot be easily extracted used computer vision techniques. For these purposes semantic attributes have also been manually ascribed to the images. Specifically the terms “sunny day”, “nature”, “man made” and “driving” have been ascribed to each, though this process is incomplete and mistakes have been made. More importantly, these terms are also those people are likely to use to search for these images. An example of such a set of images can be seen in Table 4.1 where the rows represent terms, the columns represent images and the data entries represent the number of times a given term or feature occurs in a given image.

TABLE 4.1: Feature-by-Document matrix for countryside/city scene pictures with corresponding frequencies of of a given feature

	 City1	 City2	 Country 1	 Country 2
 buildingShape	3.0	3.0	0.0	0.0
 carShape	3.0	1.0	0.0	1.0
 circleShape	5.0	2.0	1.0	3.0
 sunshineShape	0.0	1.0	8.0	8.0
 treeShape	0.0	0.0	5.0	2.0
sunny day	0.0	0.2	1.0	1.0
nature	0.0	0.0	1.0	1.0
driving	0.8	0.4	0.0	0.0
man made	1.0	1.0	0.0	0.0

$$\mathbf{U} = \begin{pmatrix} -0.6 & 0.4 & -0.4 & 0.5 \\ -0.3 & -0.7 & 0.4 & 0.5 \\ 0.6 & -0.2 & -0.6 & 0.5 \\ 0.3 & 0.5 & 0.6 & 0.5 \end{pmatrix}, \tag{4.12}$$

$$\mathbf{\Sigma} = \begin{pmatrix} 1.1 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.3 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.2 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 \end{pmatrix}, \tag{4.13}$$

$$\mathbf{V}^T = \begin{pmatrix} -0.3 & -0.2 & -0.2 & 0.8 & 0.4 & 0.1 & 0.1 & -0.1 & -0.1 \\ -0.4 & 0.4 & 0.8 & 0.2 & -0.0 & 0.1 & 0.1 & 0.0 & -0.1 \\ -0.1 & -0.2 & -0.1 & 0.4 & -0.9 & 0.1 & 0.0 & -0.1 & -0.0 \\ -0.0 & 0.0 & 0.0 & 0.0 & 0.0 & -0.7 & 0.0 & -0.7 & -0.0 \end{pmatrix} \tag{4.14}$$

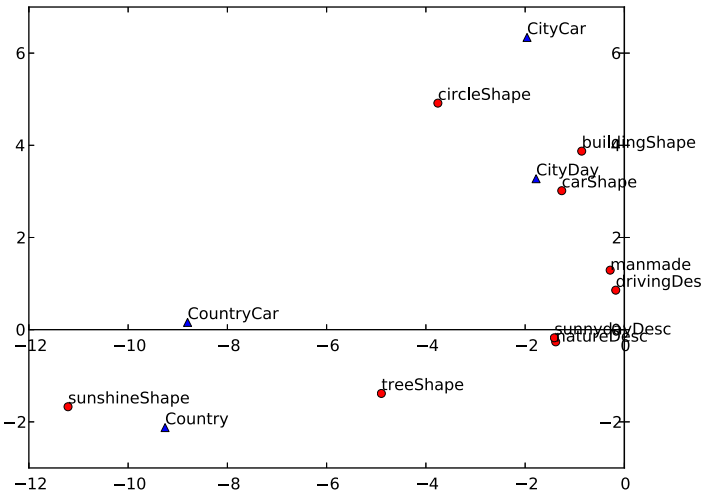


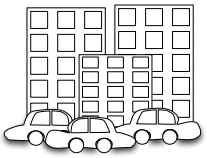
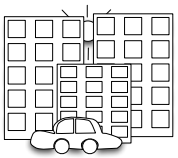
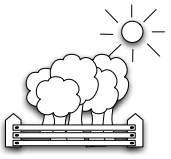
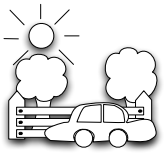
FIGURE 4.1: A 2D rank-2 LSA vector space for the countryside/city scene pictures

Using the documents outlined in Table 4.1 we can calculate the Singular Value Decomposition (SVD) shown in Equation 4.12. In this scenario, the rows of the $\mathbf{U}$ matrix represent the positions of the 4 images according to 4 left-singular vectors where the rows of $\mathbf{V}$ (i.e the columns of $\mathbf{V}^T$) represent the positions of the 9 terms with regards the 4 right-singular vectors. By weighting $\mathbf{U}$ and $\mathbf{V}$ by the eigenvalues and choosing only the first 2 highest eigenvalues we can visualise the position of the documents and terms in a 2 dimensional *concept space* (See Fig. 4.1). From this plot we can firstly see a clear separation of city scenes and country scenes along the x-axis (the largest eigenvector) and a separation of images containing cars and images not containing cars in the y-axis (the 2nd largest eigenvector). Furthermore we can see that visual shapes and descriptions regarding natural scenes lie in the general direction of the countryside images.

4.2.4.2 Example Retrieval

Given that some of the features in this example are semantic, it is possible to perform semantic retrieval. Assuming this ability, let us attempt the retrieval of the image in our example most relevant to the notion of "a drive in the countryside". Given this query, the desired image is "Country2", having both a car and evidence of a natural scene. We achieve this by formulating a novel document $\mathbf{d}_{test}$ which contains only the terms "driving" and "nature". By projecting this document into the eigen-term concept space represented by $\mathbf{V}$, we can find its position in the 2D concept space;. By measuring the cosine distance between this projection and the projected position of all existing documents we can order the documents according to their relevance to the query. We

TABLE 4.2: Cosine distances of projected query to projected documents. Larger values for cosine distances mean closer documents as $\cos(1.0) = 0°$ and $\cos(0.0) = 90°$ whereas small values for the euclidean distances mean closer documents.

Documents	**Query Distances** $driving = 1.0, nature = 1.0$	
	LSA (cosine distance)	Lexical (euclidian distance)
 City1	0.76	0.08
 City2	0.90	0.07
 Country1	0.88	0.07
 Country2	0.94	0.08

start with the query document and project it into the feature concept space $\mathbf{V}^T$ weighted by the singular values in $\mathbf{\Sigma}$:

$$\mathbf{d}_{test} = \begin{pmatrix} 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 1.0 & 1.0 & 0.0 \end{pmatrix}, \tag{4.15}$$

$$\mathbf{d}_{test} = \mathbf{d}_{projected}\mathbf{\Sigma}\mathbf{V}^T, \tag{4.16}$$

$$\mathbf{d}_{projected} = \mathbf{d}_{test}(\mathbf{\Sigma}\mathbf{V}^T)^T \tag{4.17}$$

To allow comparison to this query, the original images are also projected in this way. A distance metric, in this case the cosine distance metric, is now used to order documents by their distance to our novel query document. Closer documents are more relevant to the query.

The distances from projected query to all other documents can be seen in Table 4.2 where the correct image is indeed "Country2". Furthermore, other images are ordered correctly, with "Country1" and "City2" coming before the least similar image "City"; containing least features related to nature (i.e. sunshine and trees). It should be noted that due to incorrect annotations in the source data, the desired image does not directly contain the label "driving" and instead simply shares a similarity to the underlying *Latent Concept* which generated driving. If we were to rely upon direct lexical comparison instead of this LSA approach "Country2" would be as close to the query as all the other images, each containing either "nature" or "driving" but not both. This shows the power of the LSA approach over direct lexical comparison.

4.3 Semantic Retrieval Experiments

In this section we explore the retrieval capability of the semantic features introduced in Section 2.3. This is done by applying LSA to the feature fused annotations gathered against the HIDDB and TunnelDB datasets with all the biometric signatures covered in Section 3. We firstly outline our experimental procedure, defining how a training matrix is constructed in order to learn a latent semantic space and then how tests are performed with this semantic space. We then outline the retrieval ability of our technique with all semantic annotations and visual features collected. For each set of experiments we analyse results and discuss their meaning and implications.

Experimental Procedure

For each set of biometric signatures in each of the biometric datasets, along with their associated annotations, a training matrix $\mathbf{O}_{train}$ is constructed comprising of some of the subjects in the dataset. This matrix's visual features and semantic features are fully observed. A second test matrix $\mathbf{O}_{test}$ is also constructed using the rest of the subjects such that visual features are observed while semantic features are unobserved. The retrieval task attempts to order the documents in $\mathbf{O}_{test}$ against a set semantic queries o_{query}. This ordering is then assessed for quality against the set of semantic features omitted from the test set.

The documents in the training stage are the samples (and associated semantic annotations) of a randomly selected set of half of the *subjects* in the datasets. The test documents are the other subjects with their semantic terms set to zero. Importantly, this means that subjects in the test set do not appear at all in the training set, unlike the recognition tests in Section 3.5. This means that successful retrieval can be convincingly

attributed to similarity of underlying semantic concepts rather than trivial matching of identity. Ideally, every combination of subjects would be used for training and testing purposes, but this would be unfeasible given the computational complexity of the training matrix analysis process. Instead, we simulate this by generating 20 such random training-test sets, generating the associated matrix decompositions $\mathbf{U}_{train}$,$\mathbf{\Sigma}_{train}$ and $\mathbf{V}^T_{train}$ for each. The test documents are projected into the training space through the process described in Section 4.2.3.

Once the test matrix is projected, they are compared and ordered against projected queries. In these tests we measure the retrieval ability of each semantic term in isolation (e.g. *Sex Male*, *Height Tall* etc.). All test documents are retrieved, but only a few are *relevant*. The *relevance* of retrieval is assessed by considering the annotations of the test set which are known, but thus far not included. For example, a sample retrieved under the query "Age: Senior" is relevant if most annotators ascribed the term Senior to the trait Age of the subject represented in the sample.

Measuring Performance

To measure the performance of any given query, a variant on the standard mean Average Precision (mAP) metric is calculated. For a set of document received by a query, *precision* is defined as the number of correct documents retrieved divided by the total number of document retrieved:

$$P(r) = \frac{|\text{relevant}(r) \cap \text{retrieved}(r)|}{|\text{retrieved}(r)|} \tag{4.18}$$

Where r is the rank along the set of retrieved documents, relevant(r) is the set of relevant documents in the first r returned, and retrieved(r) is the set of r documents returned. If precision is 1.0 it means that all the documents retrieved are relevant. With systems used by humans, it is not only important that the correct results are retrieved with a good ratio to incorrect results (e.g. EER), but also that the correct results appear earlier in the ordering. Another measure called *average precision* can help gauge this by finding the average precision value found at every rank at which each a relevant document is retrieved.

$$AveP(R) = \frac{\sum_{r \in R}^{R} P(r)}{|R|} \tag{4.19}$$

Where R is the set of ranks of retrieved documents amongst all documents returned. The mAP is the average of average precisions over a set of experiments, in our case across 20 random orderings to construct different training sets.

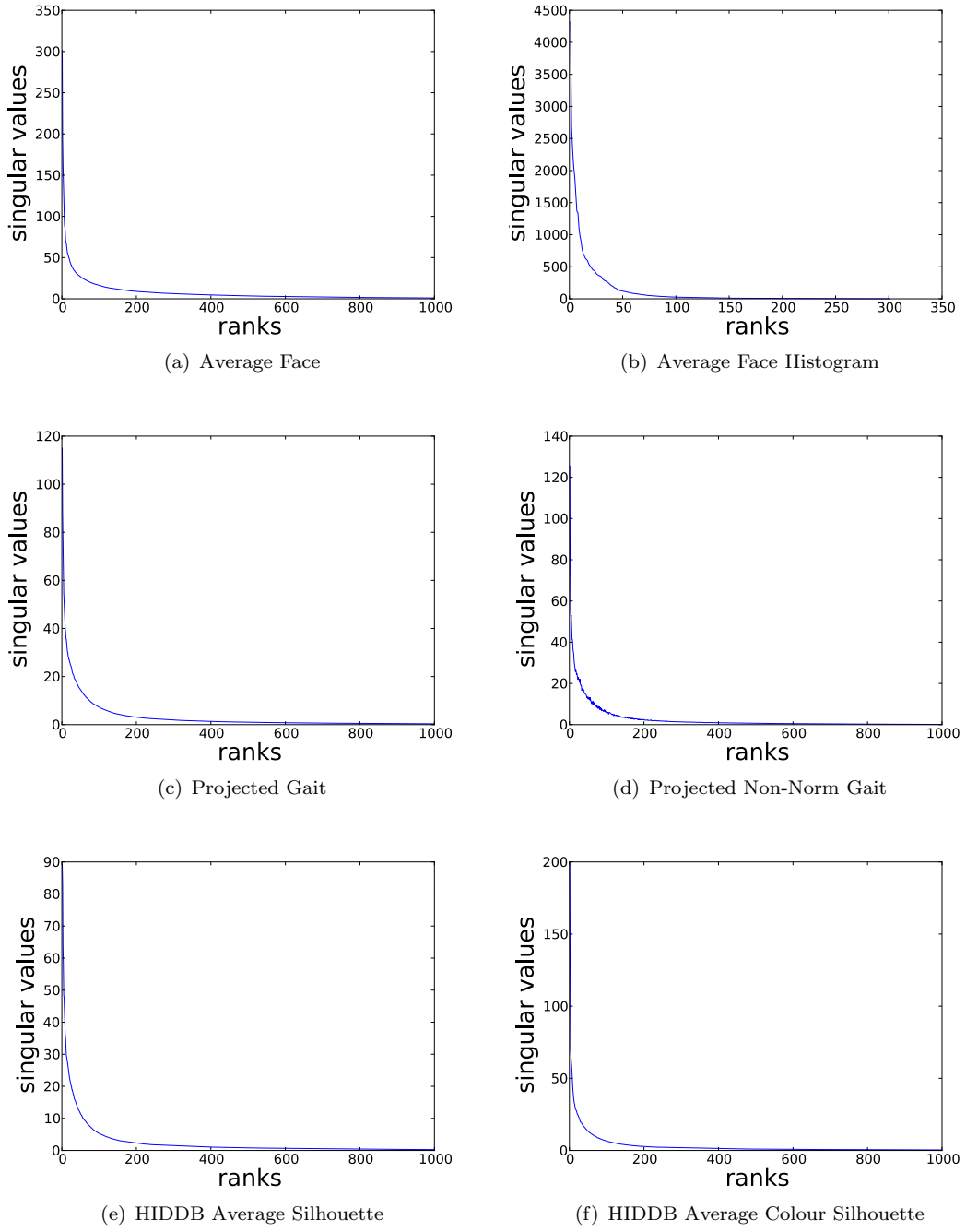
Furthermore, rather than simply measuring absolute mAP for a given term we show the improvement of the mAP of each semantic term as compared to the mAP of a random ordering for each query. We call this the improved mean Average Precision (i-mAP). To generate the random mAP we generate 100 completely random orderings for each semantic query and average their mAP. We generate the i-mAP in order to account for the situation where all subjects were annotated with the same term. In this case the absolute mAP would be high but meaningless because of random ordering's mAP would be equally high. Therefore, in this situation the i-mAP would be low giving us a better idea of which terms successfully retrieved relevant documents. We present the i-mAP of each physiological trait as a sum of the i-mAP of its semantic terms. These results give some idea of which traits our approach is most capable of performing queries against, while gathering these numbers for all 6 biometric features tells us which visual features are most effective for each trait.

We now present these results grouped by related biometric features. Namely, the two gait signatures of the two datasets are grouped and the two face signatures of the TunnelDB are presented together. Each set of i-mAPs are shown on the same graphical scale making the separate graphs visually comparable. Along with each pairing an ANOVA is performed comparing the i-mAP scores of each of the traits in the two related signatures. This allows a clear analysis of the ability of the related features against one another.

Exploration of Singular Values

As explained in Section 4.2.2 improved performance is gained using the SVD for LSI when projection is performed using only the singular vectors with high singular values, i.e. a rank reduced version of $\mathbf{U}$ and $\mathbf{V}$. What remains is the selection of this rank. In Fig. 4.2 we show some example distributions of singular values in our 6 datasets. It can be seen that in most datasets roughly 70% of the variance can be represented using only it's largest 100 or so singular values and associated singular vectors. In Fig. 4.3 we show how the summed i-mAP of each trait is affected by the selection of ranks. We note that extremely small ranks result in erratic results while large ranks seem to introduce

error and damage retrieval results. With these results in mind we choose to perform the experiments in this section with a rank 100, though it is has been shown that better results can be expected through careful selection of rank using a validation set [42].



(a) Average Face

(b) Average Face Histogram

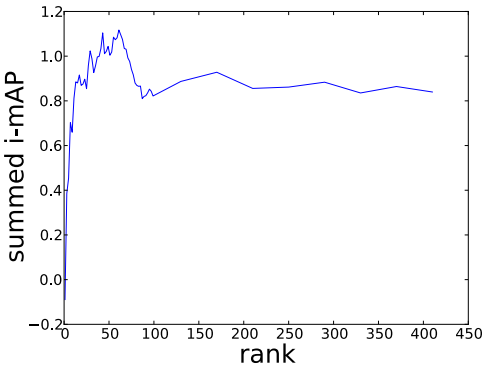
(c) Projected Gait

(d) Projected Non-Norm Gait

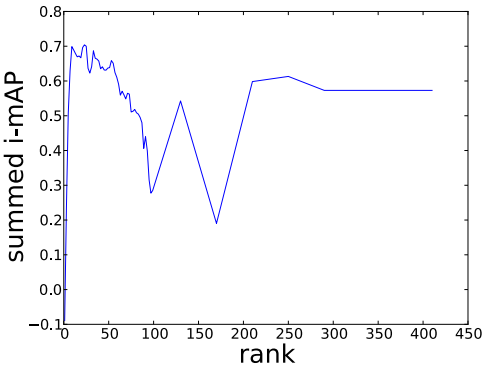
(e) HIDDB Average Silhouette

(f) HIDDB Average Colour Silhouette

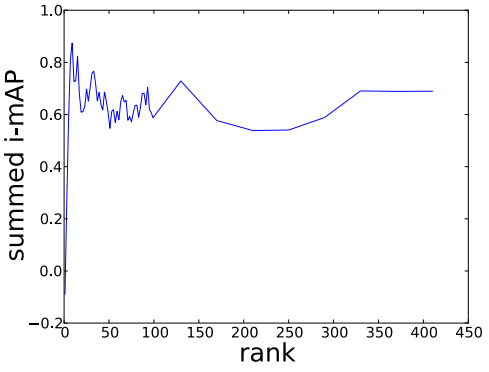
FIGURE 4.2: Subfigures (a)-(f) showing the singular values for the first 1000 singular vectors of each dataset



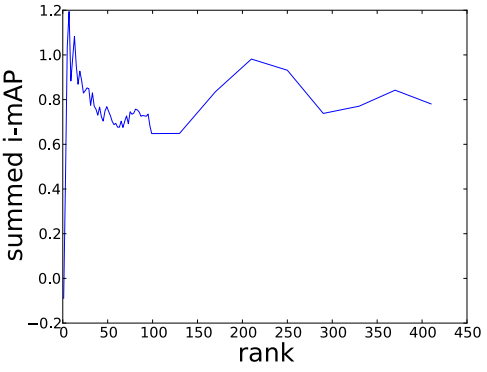
(a) Average Face



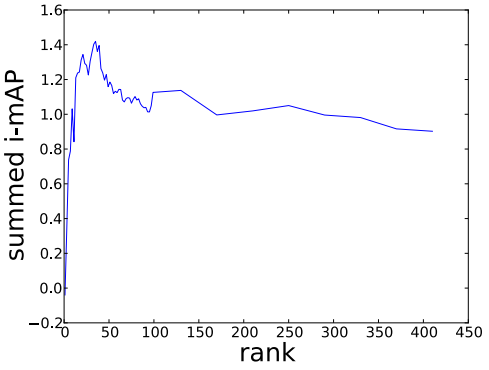
(b) Average Face Histogram



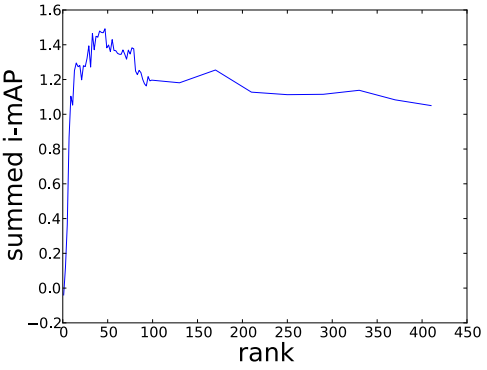
(c) Projected Gait



(d) Projected Non-Norm Gait



(e) HIDDB Average Silhouette



(f) HIDDB Average Colour Silhouette

FIGURE 4.3: Subfigures (a)-(f) showing summed i-mAP compared to selected singular values

4.3.1 HIDDB Gait Retrieval

In this section we present the i-mAP for each semantic trait against the Average Silhouette and Average Colour Silhouette gait signatures gathered from the Southampton Large (A) HumanID Database (HIDDB). For this test, all 115 subjects of the dataset were used. 50 subjects were annotated by at least 5 annotators where the rest were annotated by at least 1 annotator. Table 4.3 presents some example generated query signatures and the related results while Fig. 4.4 shows the i-mAP of each trait across the 20 random training configurations.

4.3.1.1 Results

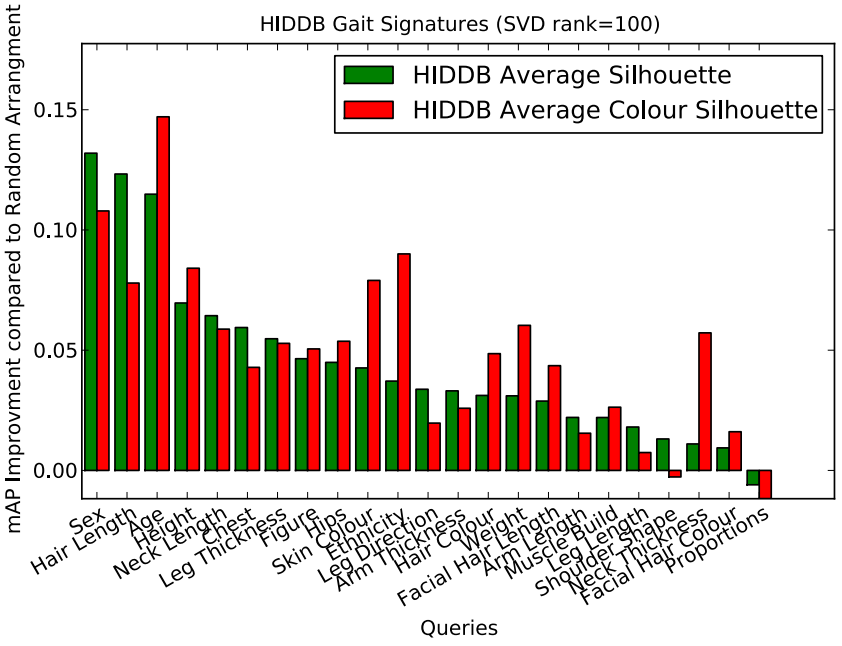


FIGURE 4.4: The mean average precision improvement for each semantic trait. Each trait's mAP is the average summed difference of its associated semantic terms

Here we use $\text{AvgSil}_{\text{i-mAP}}$ to denote the Average Silhouette i-mAP and $\text{AvgCol}_{\text{i-mAP}}$ to denote the Average Colour Silhouette i-mAP. The results show some merit and produce both success and failure, as expected. It has been shown in previous work for example that *Sex* ($\text{AvgSil}_{\text{i-mAP}} = 0.13$ and $\text{AvgCol}_{\text{i-mAP}} = 0.11$) for Average Colour Silhouette) is decipherable from average silhouettes alone [75], achieved by analysing the separate parts of the human silhouette. Some physical metrics such as *Height* ($\text{AvgSil}_{\text{i-mAP}} = 0.07$ and $\text{AvgCol}_{\text{i-mAP}} = 0.08$), *Figure* ($\text{AvgSil}_{\text{i-mAP}} = 0.05$ and $\text{AvgCol}_{\text{i-mAP}} = 0.059$) and *Neck Length* ($\text{AvgSil}_{\text{i-mAP}} = 0.06$ and $\text{AvgCol}_{\text{i-mAP}} = 0.06$) were also relatively successful, as

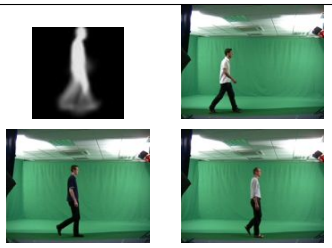
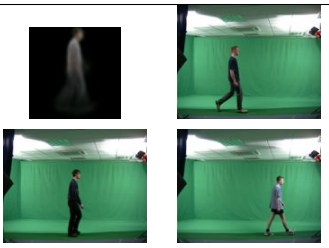
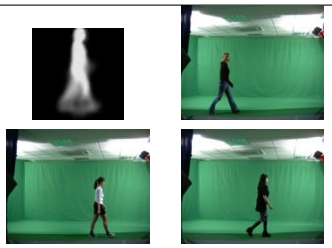
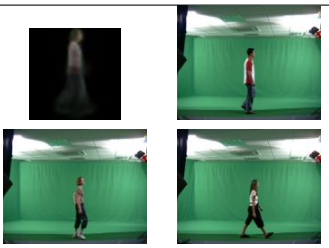
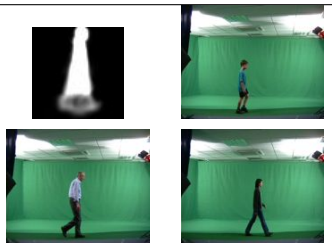
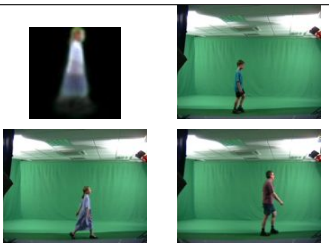
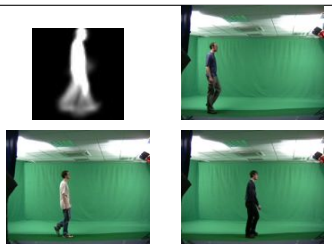
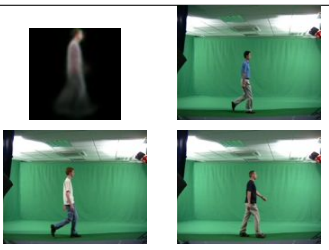
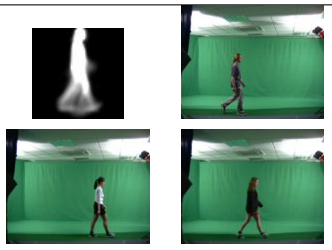
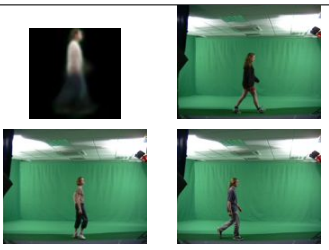
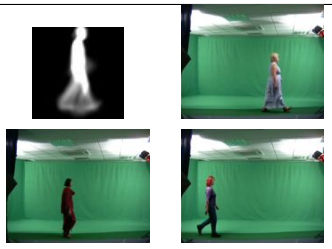
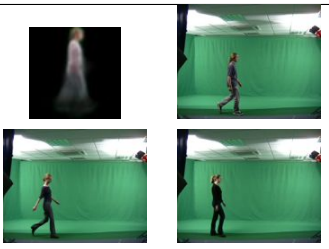
Query	HIDGaitGrey-minmaxNorm	HIDGait-minmaxNorm-64x64x3
Sex: Male		
Sex: Female		
Age: Pre+Adolescence		
Height: Tall		
Hair Length: Long		
Hair Colour: Blond		

TABLE 4.3: Some Example Retrieval Results. The first image in each set is the image generated for a semantic query as part of the method explained in Section 4.2.2. The next 3 images are video keyframes of the 3 top ranked subjects from a particular experiment.

Significant Features	
Trait	**p-value**
Ethnicity	$p \ll 0.0001$
Neck Thickness	0.0001
Hair Length	0.0001
Skin Colour	0.0006
Shoulder Shape	0.0120
Weight	0.0302
Sex	0.0321

Insignificant Features	
Trait	**p-value**
Hair Colour	0.0741
Leg Length	0.2853
Proportions	0.3740
Arm Length	0.4845
Facial Hair Colour	0.4847
Neck Length	0.5123
Muscle Build	0.5424
Figure	0.6849

Insignificant Features	
Trait	**p-value**
Leg Direction	0.0517
Hips	0.1454
Chest	0.2329
Age	0.2677
Facial Hair Length	0.3137
Height	0.3577
Arm Thickness	0.3814
Leg Thickness	0.8449

TABLE 4.4: The i-mAP p-values treating HIDGaitGrey-minmaxNorm and HIDGait-minmaxNorm-64x64x3 signatures as seperate classes for each physiological trait. Here we use the significance value of $p \leq 0.05$

was expected, because the average silhouette maintains a linear representation of these values in the overall intensity of pixels.

In Table 4.3 we see example orderings provided by our scheme and an anecdotal comparison of the ability of colour signatures against monochrome silhouettes. The examples aid to show the potential merits and pitfalls of using the different signatures. Both configurations perform well with *Sex*, though for our example *Sex Female* query, colour signatures incorrectly correlate light coloured clothing with gender. The colour of clothing is ignored by the standard average silhouettes as the whole body silhouette of the individual is used and the internal detail ignored. The average colour signature has a similar problem with the example *Age* query. The opposite performance is evident for queries which inherently correlate with colour. In Table 4.3 we see that for the *Hair Colour* the average colour silhouette achieves more favourable results, correctly finding a correlation with light shades in the head area with blond hair (as can be seen on the automatically generated *Hair Colour* query signature)

Fig. 4.4 shows the relative merits of the two approaches. Table 4.4 shows the significance of these differences across all random training set selections; the significance is calculated using a one-way ANOVA (See Section 3.5.1.1). It can be seen that whilst performing relatively poorly in both configurations, *Hair Colour* ($p = 0.0006$); *Ethnicity* ($p \ll 0.0003$) and *Skin Colour* ($p = 0.0006$) perform significantly better when colour average silhouettes are used. It should be noted however that, for *Sex* ($p = 0.0321$) and *Hair Length* ($p = 0.0001$), all mAPs are significantly lower on the average colour silhouettes. This result was expected as the colour signature allows for misleading correlations with clothing, a failure which can be seen in the example query projections of *Sex Female* and *Hair Length Long* in Table 4.3 both showing correlation with light coloured clothing. Such errors cannot be avoided easily using the holistic colour signatures; they could potentially be avoided by considering only pertinent regions such as the head area.

4.3.2 TunnelDB Gait Retrieval

In this section we present the i-mAP for each semantic trait against the normalised and non-normalised projected gait signatures of the TunnelDB. To increase the sample set size against which queries are made and to allow the growth of the training set all subjects collected in the TunnelDB are used. This means that as well as using the 60 subjects which had self annotations, we incorporate subjects who are only self annotated. Given the healthy correlation of self annotations with ascribed annotations shown in Section 2.5.3 we still expect acceptable results. The total number of subject in this set is 227, 60 of whom are annotated by other subjects while the rest are only self annotated. This results in training sets of roughly 110 individual subjects. Table 4.5 presents some example generated query signatures and the related results while Fig. 4.5 shows the i-mAP of each trait across the 20 random training configurations.

4.3.2.1 Results

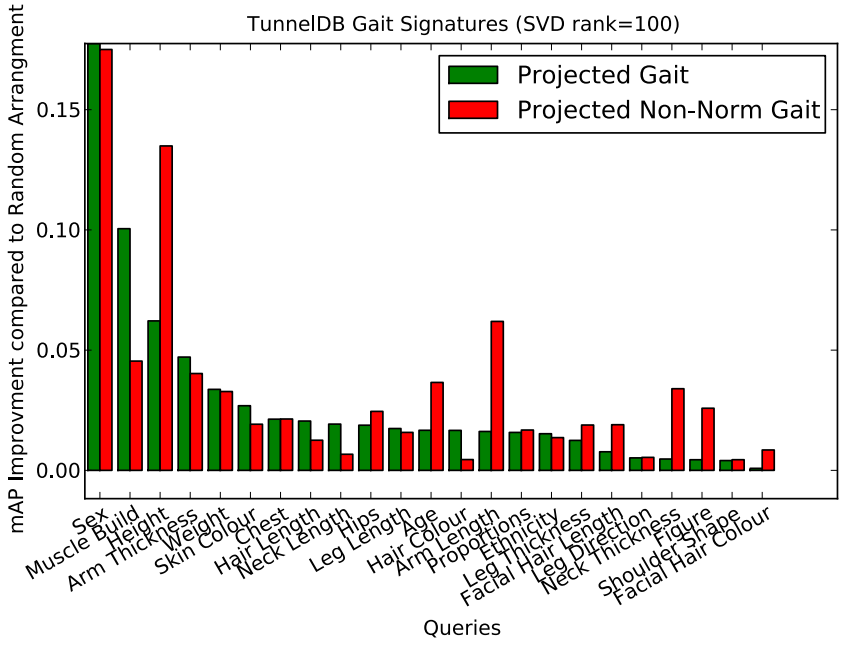


FIGURE 4.5: The mean average precision improvement for each semantic trait. Each trait's mAP is the average summed difference of its associated semantic terms

Here we use $\text{GaitNorm}_{\text{i-mAP}}$ to denote the Projected Gait i-mAP and $\text{GaitNonNorm}_{\text{i-mAP}}$ to denote the Non-Normalised Projected Gait i-mAP. In this configuration, both signatures show promise in some features and fail in others. Both display ability in Sex ($\text{GaitNorm}_{\text{i-mAP}} = 0.178$ and $\text{GaitNonNorm}_{\text{i-mAP}} = 0.175$) and Height ($\text{GaitNorm}_{\text{i-mAP}} =$

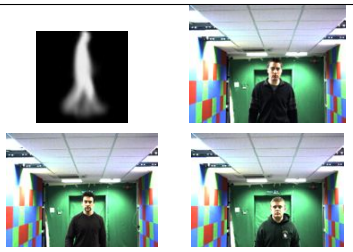
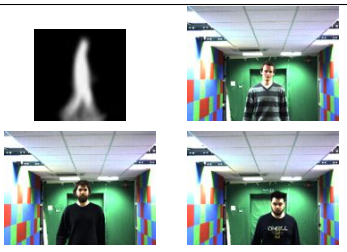
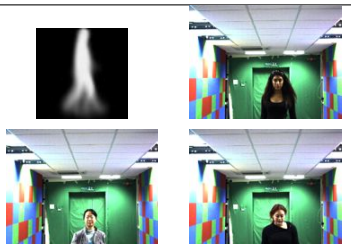
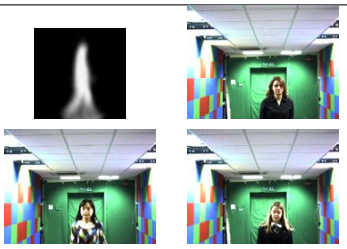
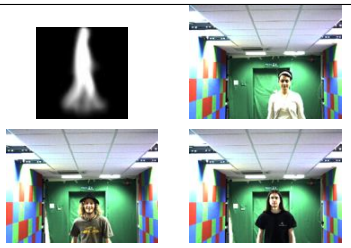
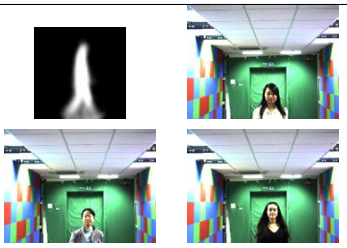
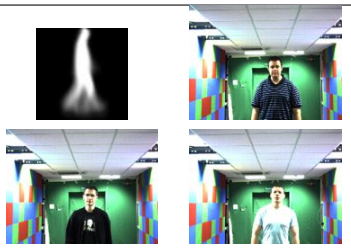
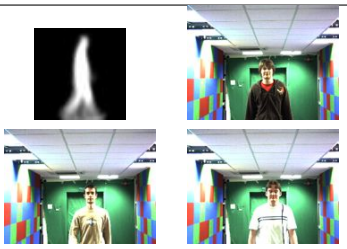
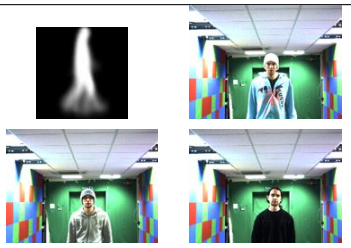
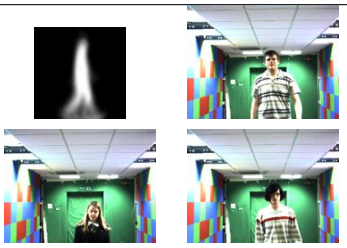
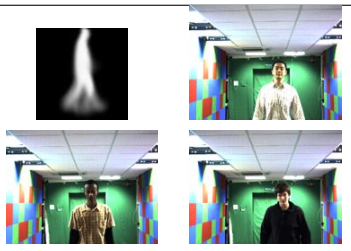
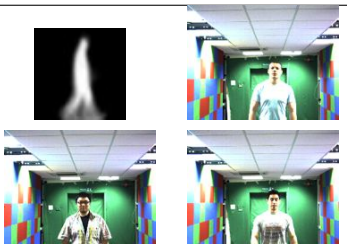
Query	Gait-minmaxNorm	GaitNonNorm-minmaxNorm
Sex: Male		
Sex: Female		
Height: Short		
Height: Tall		
Hair Length: Long		
Hair Length: Short		

TABLE 4.5: Some Example Retrieval Results. The first image in each set is the image generated for a semantic query as part of the method explained in Section 4.2.2. The next 3 images are video keyframes of the 3 top ranked subjects from a particular experiment.

Significant Features	
Trait	**p-value**
Arm Length	$p \ll 0.0001$
Neck Thickness	$p \ll 0.0001$
Height	0.0017
Figure	0.0078
Muscle Build	0.0209
Neck Length	0.0235
Hair Colour	0.0299
Facial Hair Length	0.0366
Facial Hair Colour	0.0428
Hair Length	0.0480

Insignificant Features	
Trait	**p-value**
Age	0.1102
Skin Colour	0.1944
Sex	0.6360
Weight	0.8778
Shoulder Shape	0.9166
Chest	0.9869

Insignificant Features	
Trait	**p-value**
Leg Thickness	0.1156
Hips	0.5615
Arm Thickness	0.5852
Ethnicity	0.5932
Leg Length	0.7585
Proportions	0.9255
Leg Direction	0.9680

TABLE 4.6: The i-mAP p-values treating Gait-minmaxNorm and GaitNonNorm-minmaxNorm signatures as seperate classes for each physiological trait. Here we use the significance value of $p \leq 0.05$

0.101 and $\text{GaitNonNorm}_{\text{i-mAP}} = 0.135$), a result we have come to expect from gait signatures. Also some ability can be seen in bulk features such as Weight ($\text{GaitNorm}_{\text{i-mAP}} = 0.034$ and $\text{GaitNonNorm}_{\text{i-mAP}} = 0.033$), Arm Thickness ($\text{GaitNorm}_{\text{i-mAP}} = 0.047$ and $\text{GaitNonNorm}_{\text{i-mAP}} = 0.040$) and Neck Thickness ($\text{GaitNorm}_{\text{i-mAP}} = 0.019$ and $\text{GaitNonNorm}_{\text{i-mAP}} = 0.034$), though the features in this dataset achieve significantly worse results than the HIDDB gait signatures. Unexpectedly the Hair Length ($\text{GaitNorm}_{\text{i-mAP}} = 0.017$ and $\text{GaitNonNorm}_{\text{i-mAP}} = 0.013$) was not retrieved at all effectively, regardless of its known ability as portrayed by the HIDDB gait signatures. This is arguably related to the lesser quality of the projected gait signatures in TunnelDB with regards to upper body visual features. The signatures are of a generated viewpoint of the volumetric carved individual. This volumetric model has been shown to lose some details of the upper body and therefore lose features in the head region where hair length features could be discovered. Such features have been shown to be important in discovering identity [123]; therefore Seely et al. [112] compensated for the lack of these with upper body features by involving multiple novel viewpoints in feature fusion in their recognition experiments. This approach could also be used to aid our retrieval scenario. Also it goes without saying that colour based features such as Skin Colour, Ethnicity and Hair Colour completely fail due to the lack of colour in these signatures. This goes further towards showing that latent attributes of race are not efficiently encoded in gait alone, rather skin pigmentation is by far the best signifier of race.

We note that there are significant benefits gained by using non normalised gait signatures when compared to normalised. Several features such as Height ($p = 0.0017$) and Figure ($p = 0.0078$) which describe the shape of the individual are retrieved significantly more efficiently with non-normalised gait signatures. This is expected as, while the normalised gait signature keep only minimal information in latent aspects of the gait with regards to body shape, the non-normalised gait signatures more directly encode features relating to body shape. This can be seen anecdotally in the example signatures returned

in Table 4.5. It should be noted here that the video still representing the example of any given subject is taken at the same point in the individual's walk, making their height with reference to the top of the door frame in the background a meaningful comparison.

4.3.3 TunnelDB Face Retrieval

In this section we present the i-mAP for each semantic trait against the Average Face and Average Face Histograms of the TunnelDB. As with the TunnelDB gait signatures, the whole sample set is used including some samples with only self annotations. Table 4.7 presents some example generated query signatures and the related results while Fig. 4.6 shows the i-mAP of each trait across the 20 random training configurations.

4.3.3.1 Results

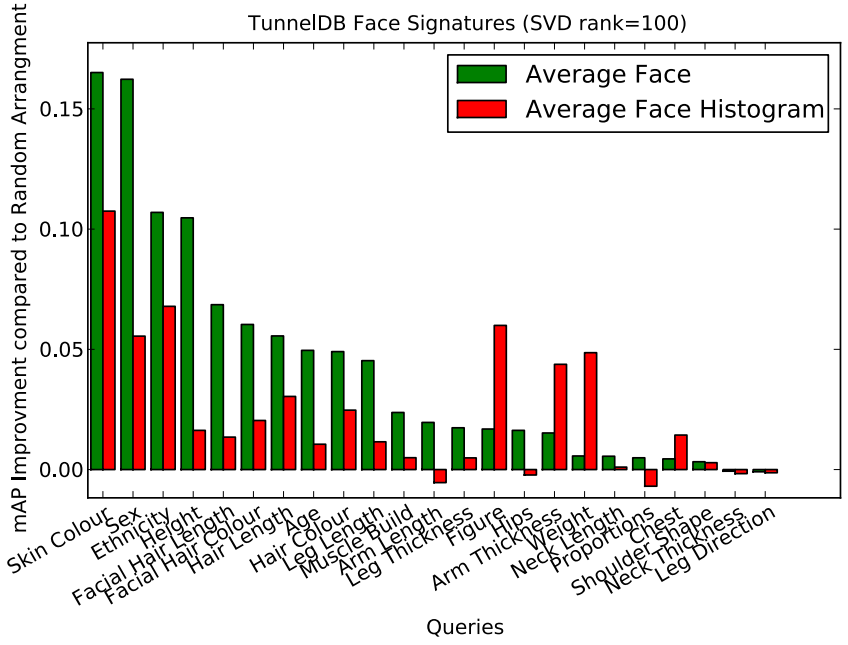


FIGURE 4.6: The mean average precision improvement for each semantic trait. Each trait's mAP is the average summed difference of its associated semantic terms

Here we use $\text{AvgFace}_{\text{i-mAP}}$ to denote the Average Face i-mAP and $\text{AvgFaceHist}_{\text{i-mAP}}$ to denote the Average Face Histogram i-mAP. Both approaches display ability in many features which most gait signatures found challenging. We see high i-mAP values for traits such as Skin Colour ($\text{AvgFace}_{\text{i-mAP}} = 0.165$ and $\text{AvgFaceHist}_{\text{i-mAP}} = 0.107$), Ethnicity ($\text{AvgFace}_{\text{i-mAP}} = 0.107$ and $\text{AvgFaceHist}_{\text{i-mAP}} = 0.068$), Hair Colour ($\text{AvgFace}_{\text{i-mAP}} = 0.049$ and $\text{AvgFaceHist}_{\text{i-mAP}} = 0.025$) and Facial Hair Colour ($\text{AvgFace}_{\text{i-mAP}} = 0.060$

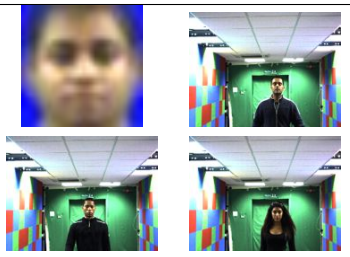
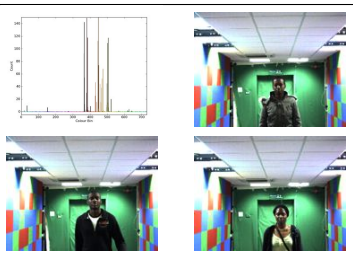
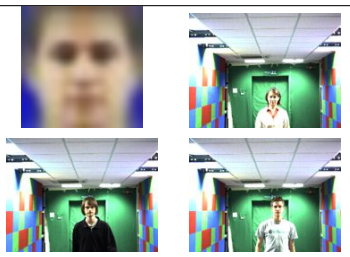
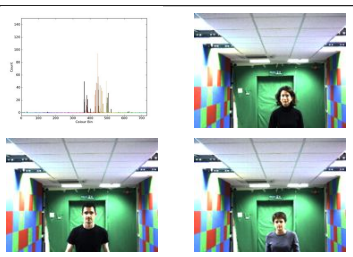
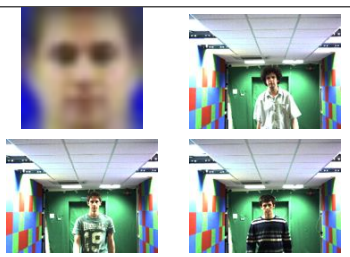
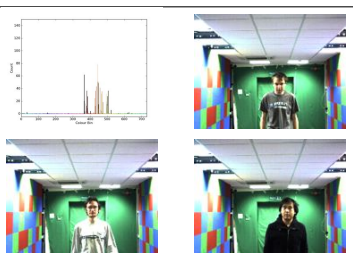
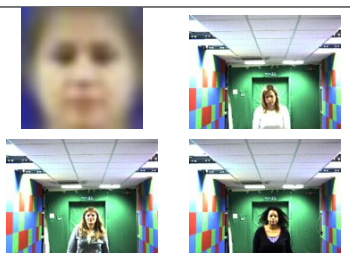
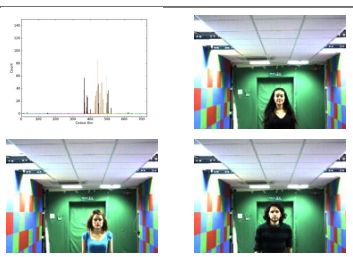
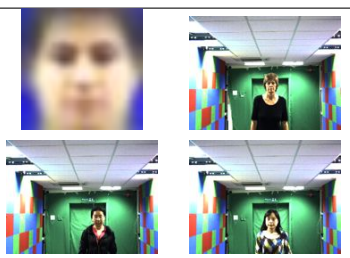
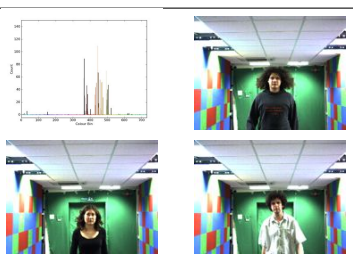
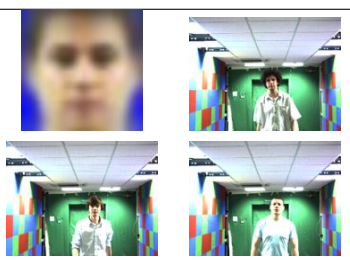
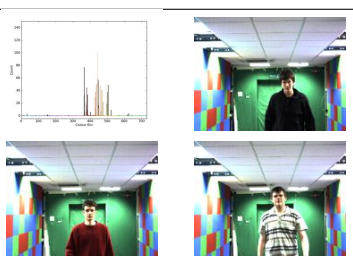
Query	Face-minmaxNorm	FaceHist
Skin Colour: Black		
Skin Colour: White		
Sex: Male		
Sex: Female		
Figure: Small		
Height: Tall		

TABLE 4.7: Some Example Retrieval Results. The first image in each set is the image generated for a semantic query as part of the method explained in Section 4.2.2. The next 3 images are video keyframes of the 3 top ranked subjects from a particular experiment.

Significant Features	
Trait	**p-value**
Sex	$p \ll 0.0001$
Facial Hair Length	$p \ll 0.0001$
Skin Colour	0.0001
Hair Colour	0.0001
Hair Length	0.0002
Arm Length	0.0002
Facial Hair Colour	0.0002
Height	0.0004
Ethnicity	0.0040
Muscle Build	0.0050
Age	0.0067
Hips	0.0077
Leg Length	0.0124
Weight	0.0137
Leg Thickness	0.0162
Proportions	0.0315
Figure	0.0328

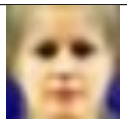
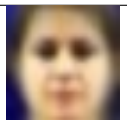
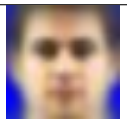
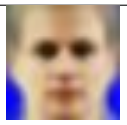
Insignificant Features	
Trait	**p-value**
Neck Length	0.3566
Neck Thickness	0.8228
Shoulder Shape	0.9549

Insignificant Features	
Trait	**p-value**
Arm Thickness	0.0889
Chest	0.3371
Leg Direction	0.9152

TABLE 4.8: The i-mAP p-values treating Face-minmaxNorm and FaceHist signatures as seperate classes for each physiological trait. Here we use the significance value of $p \leq 0.05$

and $\text{AvgFaceHist}_{\text{i-mAP}} = 0.020$). We also see an ability for the face signature to distinguish Sex ($\text{AvgFace}_{\text{i-mAP}} = 0.162$ and $\text{AvgFaceHist}_{\text{i-mAP}} = 0.055$), a result which has been noted by several other approaches in the past [134]. A surprisingly highly rated feature is Height ($\text{AvgFace}_{\text{i-mAP}} = 0.105$ and $\text{AvgFaceHist}_{\text{i-mAP}} = 0.016$) given the fact that only face information is being analysed. However, we can see that only the average face signature and not the histogram can accurately gauge height. Through inspection of the generated query face for *Height: Tall*, it is clear that a miss correlation was measured between Height and Male Sex. This problem can be addressed with more training information, namely more women of varying heights.

It is to be expected that for several features the Average Face signature performs significantly better than the Average face histogram. The Average face holds some information with regards to the positioning of key face components which clearly holds substantial information with regards to many of our semantic features. More surprisingly however, colour histograms significantly beat the average face signatures on features regarding overall bulk such as Figure ($p = 0.0328$) and Weight ($p = 0.0137$). Through close inspection of the generated query signatures, it is clear that the overall area of the colour histogram is smaller for small bulks, and larger for large bulks. This gives colour histograms a better chance of correlation than overall face signatures, which can get confused through the correlation of misleading facial details that hold less information with regards to bulk than does overall pixel intensity.

Query	Projected Query
Hair Colour: Blond AND Sex: Female	
Sex: Female AND Hair Colour: Brown	
Sex: Male AND Hair Colour: Brown	
Hair Colour: Blond AND Sex: Male	
Skin Colour: Oriental AND Sex: Male	

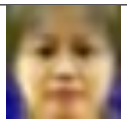
Query	Projected Query
Skin Colour: Oriental AND Sex: Female	
Skin Colour: Oriental AND Hair Colour: Blond AND Sex: Female	
Skin Colour: Black AND Sex: Male	
Skin Colour: Black AND Sex: Female	
Hair Colour: Blond AND Skin Colour: Black AND Sex: Male	

TABLE 4.9: Some example average face signatures generated by projecting selected semantic queries into a semantic visual space of the average face signature.

4.3.3.2 Compound Queries

Given that the features of the human face are more immediately comprehensible than those of gait, we take this opportunity to present some of the other capabilities of our LSA approach, namely in novel compound queries. The SVD approach provides the unique ability to represent the position of each feature separately in the concept space. By projecting individual semantic features into the space it is possible to generate the visual signatures most accurately representing those queries; this is the technique which has been presented so far. The next logical step is compound queries. By projecting documents with multiple semantic attributes set to different weightings it is possible to retrieve against multiple semantic terms efficiently; generating some very interesting novel query images in the process. This can include semantic feature combinations that were never actually recorded in combination, rather their effect on pixels has been

measured separately and can therefore be mixed. A few interesting examples of such compound queries are presented in Table 4.9.

Note specifically example of "Skin Colour Black AND Hair Colour Blond". Obviously no blond haired black individual was ever annotated, but the correlation of "Hair Colour Blond" with light upper pixels and the correlation with "Skin Colour Black" with an overall darker complexion was measured, and therefore can be combined. This is another example of the potential power of LSA.

4.4 Conclusions

We have introduced the use of semantic human descriptions as queries in CBIR against human gait and face biometric signatures from two datasets. Using an LSA technique we construct an ordered list of un-annotated subjects against a set of semantic queries based on the similarity of their biometric signatures through their projection into a trained linear algebraic semantic space.

Our results confirm those of previous work with regards to certain semantic traits, such as Sex, in their ability to correlate with gait and face biometrics. We also note the capability of retrieval using other traits, previously unexplored, such as Ethnicity, Age and some build attributes. We go on to demonstrate the potential capabilities of our semantic dataset and analysis technique through the construction of novel compound query signatures, showing generated query samples constructed from with previously unseen feature combinations.

This chapter goes further towards exploiting the capabilities of the semantic biometric signatures designed and gathered in this thesis. By showing that the collected semantic datasets portray some merit in CBIR, we argue for their power as biometric signatures in general. In the next chapter we extend this argument through by combining the analysis of the previous 3 chapters. By ordering the semantic terms with regards to their combined effectiveness against a variety of metrics, we present a final ordered list of semantic biometric traits and terms, giving us further insight into their individual scope and usefulness.

Chapter 5

Feature Significance

5.1 Introduction

Throughout this thesis we have explored the ability of using semantic human descriptions in a variety of scenarios. We started by exploring internal correlations found both within self annotations and ascribed annotations, as well as the correlation between them. Next we attempted to explore the ability of the semantic features to separate individuals as well as the stability of the semantic terms used to describe the given traits of an individual subject. Later in Chapter 3 we explored a more practical rating, showing the performance gain noticed when adding given annotations in a biometric LoO classification scenario. Finally in Chapter 4 we explored the value of each trait when incorporated in a semantic Content Based Information Retrieval (CBIR) across 6 biometric signatures from 2 datasets.

In this chapter, we collate this work. By using two simple majority voting schemes and treating each ordering strategy highlighted in the thesis as a separate classifier we provide an overall rating for each physiological trait and its associated semantic terms. By definition, this final rating will incorporate various aspects of what makes a trait powerful with regards to its ability to be described semantically. In Section 5.2 we briefly summarise the ordering schemes to be used as well as what meaning they add to final ordering. Once these ordering schemes are outlined in Section 5.3 we will discuss how their orderings are to be combined into a final trait ordering. We present this final ordering and discuss its implications.

5.2 Vote Gathering Procedure

In this section we will outline the approaches taken to transform each chapter's contribution into a ranked list of physiological traits and explain what meaning each ranked list incorporates into the final trait ordering.

5.2.1 Correlation Analysis

In Section 2.5.3 we presented a set of correlation matrices showing the the Pearson's r correlation coefficient for each trait against each other trait. We looked at how self annotations and ascribed annotations of traits vary between themselves and each other. For each of these sets of correlation coefficients we take the summation of the absolute correlation score of a given trait against each other trait as a measure of its worth. In this approach we take negative correlations as having as much meaning as positive correlations by taking the absolute. By summing absolute correlation coefficients we obtain some idea of how consistent a trait's annotations are.

Here we assume that at least some physiological traits should reliably vary together and that traits which do so consistently between annotators are more stable and thus more susceptible to effective semantic description. An inconsistent or erratic trait would show less correlation than one annotated consistently with respect to the changes of other traits. Each set of correlations, namely: subject autocorrelations; self autocorrelations and ascribed-vs-self correlations for each dataset were taken as different classifiers.

5.2.2 ANOVA and Pearson's r Ordering

In Section 3.5.1 we presented two feature subset selection schemes. These schemes attempted to order the significance of the traits by judging their ability to distinguish individuals and by judging the stability of the annotations ascribed across separate subject groups. By incorporating orderings obtained through such analysis we inherently incorporate ability to separate individuals and stability of terms into the final ordering. Also, given that these schemes were initially used to provide an ordering of the traits no further analysis needs be performed to obtain an ordering. Both ordering schemes applied to both datasets are treated as a separate classifiers.

5.2.3 Retrieval Capability

In Section 4.3 we presented a set of retrieval experiments aimed at gauging the ability of each trait to be used in the process of Content Based Information Retrieval (CBIR) of unannotated images using LSA. This ability was gauged across 6 biometric features across 2 datasets, measuring the improvement of mAP results as compared to the mAP of completely random orderings, the improved mean Average Precision (i-mAP). By using these scores as a metric to order traits, we include in the final ordering some notion of a trait's ability to be used in retrieval tasks in general, and therefore its ability as a query-able feature in CBIR.

5.3 Ordering approach: Majority Voting

The order a trait exists in a particular scheme can be thought of as that scheme's vote for that rating of that trait. By taking a sum of all the positions of each trait we can obtain an ordering of all traits taking into consideration information attained across all the mentioned ordering schemes. This approach to vote collation is a form of majority voting. That being the case, there are several known problems with such majority voting [121]. The main issue which affects our approach is that of differing goals. There is no pretence that all our ordering schemes work towards similar goals, nor that there is a single ordering which is best suited to all scenarios. In some scenarios, the accuracy of annotations may be paramount and therefore our two sets involving Pearson's product-moment correlation coefficient (Pearson's r) would be the most telling notifier of a trait's importance. In other scenarios the general accuracy of the trait is irrelevant and only its ability to recognise and separate individuals is important and therefore the ANOVA or i-mAP orderings should be considered. For these disparate purposes in application the appropriate ordering scheme should be selected and used in isolation. In this section we show an average and aim to satisfy all criteria simultaneously to some degree, though in doing so we may in fact achieve an ordering which satisfies no given criteria.

Another major problem with majority voting is the assumption of equal weighting. In modern democratic systems all participants are equal, therefore any passion or fervour in the argument of a given individual is rightfully ignored when counting votes [121]. However, in our scenario, all classifiers can also provide estimates of confidence, as represented by the various scores each scheme produces. Therefore, the position in the ranks of any given trait is not all that matters, rather the confidence of a particular positioning of a given annotation should also be taken into consideration. For example, it was shown in Section 4.3 that the projected gait signatures, while proficient in detection

of Sex, were notably worse at retrieval by features such as Hair Length; a feature which the HIDDB Gait Signatures were notably more capable. To take such differing ability into consideration we also present a normalised weighted voting scheme where each classifier is given a single vote which is divided equally between its own ranks. The sum of these weights are then used to order the traits. It should be noted that this approach does not take into consideration the underlying distribution of any given scheme. This is especially a problem for the F-ratio of the ANOVA ordering which are known to follow the F-distribution. However, for purposes of simplicity a linear distribution is assumed.

Majority voting and weighted majority voting schemes are similar to decision fusion and transformation based score fusion techniques. These are discussed in more detail in Chapter 3.

5.4 Final Trait Ordering

TABLE 5.1: Majority voting for Majority set

Key	Experiment
A0	ANOVA TunnelDB Experiment
A1	ANOVA HIDDB Experiment
B0	Pearson's r TunnelDB Experiment
B1	Pearson's r HIDDB Experiment
C0	i-mAP HIDDB-GaitAverageColour Experiment
C1	i-mAP TunnelDB-ProjectedGaitNonNorm Experiment
C2	i-mAP HIDDB-GaitAverageGrey Experiment
C3	i-mAP TunnelDB-ProjectedGait Experiment
C4	i-mAP TunnelDB-FaceHist Experiment
C5	i-mAP TunnelDB-AverageFace Experiment
D0	Correlation HIDDB-ascrbed Experiment
D1	Correlation TunnelDB-selfVSascribed Experiment
D2	Correlation HIDDB-self Experiment
D3	Correlation TunnelDB-self Experiment
D4	Correlation TunnelDB-ascrbed Experiment

In Table 5.2 and Table 5.3 we present the trait orderings for the Majority and Weighted Majority voting approaches respectively. The tables show the final ordering of the features and the scores used to achieve those orderings. Colour in the tables represent a comparable normalised scale generated from the range of scores possible in each classification scheme. Green cells represent more significant scores whilst red cells represent scores of lower significance. In Table 5.1 shows the mapping between the codes used in the tables and the schemes which generated the orderings and normalised scores.

TABLE 5.2: Voting results for Majority approach

Feature Ordering	T	A0	A1	B0	B1	C0	C1	C2	C3	C4	C5	D0	D1	D2	D3	D4
Sex	42	0	0	0	0	1	0	0	0	3	1	11	5	6	7	8
Weight	84	5	8	6	9	6	7	14	4	4	16	1	1	1	2	0
Height	85	6	7	8	6	3	1	3	2	9	3	8	6	5	9	9
Figure	99	8	9	11	8	11	8	7	20	2	13	0	0	0	1	1
Skin Colour	101	3	1	1	1	4	11	9	5	0	0	12	9	19	16	10
Chest	115	9	10	9	10	14	10	5	6	10	19	2	2	7	0	2
Hair Length	135	1	3	3	4	5	17	1	7	6	6	15	20	15	17	15
Age	136	4	4	4	2	0	5	2	11	13	7	9	18	22	21	14
Arm Thickness	136	15	13	16	14	16	4	12	3	5	15	3	11	3	3	3
Ethnicity	138	10	2	7	5	2	16	10	15	1	2	13	10	17	15	13
Leg Thickness	145	11	16	13	11	10	13	6	16	15	12	7	4	2	5	4
Muscle Build	145	16	15	17	12	15	3	17	1	14	10	4	8	4	4	5
Hips	161	13	17	12	16	9	9	8	9	20	14	6	3	12	6	7
Facial Hair Length	167	2	6	2	7	13	12	15	17	11	4	20	19	8	12	19
Hair Colour	187	7	5	5	3	12	21	13	12	7	8	18	17	20	18	21
Neck Thickness	188	14	12	14	18	8	6	20	19	19	21	5	7	11	8	6
Neck Length	191	18	11	18	15	7	19	4	8	17	17	10	12	14	10	11
Leg Length	208	17	14	15	13	20	15	18	10	12	9	14	15	13	11	12
Facial Hair Colour	221	12	19	10	17	18	18	21	22	8	5	17	13	10	13	18
Arm Length	228	20	18	19	19	19	2	16	13	21	11	16	14	9	14	17
Leg Direction	286	21	20	21	20	17	20	11	18	18	22	19	21	18	20	20
Shoulder Shape	293	19	22	20	22	21	22	19	21	16	20	21	16	16	22	16
Proportions	304	22	21	22	21	22	14	22	14	22	18	22	22	21	19	22

The tables give us several insights into the physiological traits we have outlined and analysed in this thesis. First and foremost, it is clear that Sex is of key importance for all classifiers in both voting schemes. This shows that as a feature it is stable, easy to comprehend, performs well in retrieval tests and effectively separates the population. Indeed, intuitively it is clear that given any variable to separate individuals in the human population, Sex is a key feature, a point also agreed upon in the literature [67].

Other global features such as Skin Colour, Ethnicity and Age are voted relatively highly in both schemes; again showing these factors to be of key importance and value when describing individuals. It should be noted that in the simple majority voting scheme these secondary global features perform worse than some overall body shape variables such as Weight and Figure. Here the increased reliability of the weighted voting scheme can be seen. The reason global features such as Ethnicity and Skin Colour perform worse in the simple Majority Voting scheme is that they do not correlate with other

TABLE 5.3: Voting results for Weighted Majority approach

Feature Ordering	**T**	A0	A1	B0	B1	C0	C1	C2	C3	C4	C5	D0	D1	D2	D3	D4
Sex	2.09	0.37	0.37	0.05	0.06	0.09	0.23	0.13	0.27	0.11	0.16	0.04	0.06	0.05	0.05	0.05
Skin Colour	1.06	0.07	0.14	0.05	0.05	0.07	0.02	0.04	0.04	0.21	0.17	0.04	0.05	0.03	0.04	0.04
Height	0.92	0.03	0.02	0.05	0.05	0.07	0.17	0.07	0.09	0.03	0.1	0.04	0.05	0.05	0.05	0.05
Ethnicity	0.81	0.02	0.09	0.05	0.05	0.08	0.02	0.04	0.02	0.13	0.11	0.04	0.05	0.03	0.04	0.04
Hair Length	0.8	0.11	0.08	0.05	0.05	0.07	0.02	0.12	0.03	0.06	0.06	0.03	0.03	0.03	0.03	0.03
Weight	0.76	0.04	0.02	0.05	0.05	0.05	0.04	0.03	0.05	0.09	0.01	0.07	0.07	0.06	0.06	0.07
Age	0.75	0.04	0.05	0.05	0.05	0.13	0.05	0.11	0.03	0.02	0.05	0.04	0.03	0.03	0.03	0.04
Figure	0.74	0.03	0.02	0.05	0.05	0.04	0.03	0.04	0.01	0.12	0.02	0.07	0.07	0.06	0.06	0.07
Muscle Build	0.68	0.01	0.01	0.04	0.05	0.02	0.06	0.02	0.15	0.01	0.02	0.06	0.05	0.06	0.06	0.06
Arm Thick-ness	0.66	0.02	0.01	0.04	0.04	0.02	0.05	0.03	0.07	0.08	0.02	0.06	0.04	0.06	0.06	0.06
Chest	0.65	0.03	0.02	0.05	0.05	0.04	0.03	0.06	0.03	0.03	0.0	0.07	0.07	0.05	0.06	0.06
Leg Thick-ness	0.59	0.02	0.01	0.04	0.05	0.05	0.02	0.05	0.02	0.01	0.02	0.06	0.06	0.06	0.06	0.06
Facial Hair Length	0.58	0.08	0.02	0.05	0.05	0.04	0.02	0.03	0.01	0.03	0.07	0.03	0.03	0.05	0.04	0.03
Hips	0.56	0.02	0.01	0.04	0.04	0.05	0.03	0.04	0.03	-0.0	0.02	0.06	0.06	0.04	0.06	0.06
Hair Colour	0.52	0.03	0.05	0.05	0.05	0.04	0.01	0.03	0.02	0.05	0.05	0.03	0.03	0.03	0.03	0.02
Neck Thickness	0.48	0.02	0.01	0.04	0.04	0.05	0.04	0.01	0.01	-0.0	-0.0	0.06	0.05	0.04	0.05	0.06
Neck Length	0.46	0.01	0.01	0.04	0.04	0.05	0.01	0.06	0.03	0.0	0.01	0.04	0.04	0.04	0.04	0.04
Leg Length	0.44	0.01	0.01	0.04	0.04	0.01	0.02	0.02	0.03	0.02	0.05	0.03	0.04	0.04	0.04	0.04
Facial Hair Colour	0.44	0.02	0.01	0.05	0.04	0.01	0.01	0.01	0.0	0.04	0.06	0.03	0.04	0.05	0.04	0.03
Arm Length	0.42	0.01	0.01	0.04	0.03	0.01	0.08	0.02	0.02	-0.01	0.02	0.03	0.04	0.05	0.04	0.03
Leg Direc-tion	0.25	0.0	0.0	0.03	0.02	0.02	0.01	0.03	0.01	-0.0	-0.0	0.03	0.02	0.03	0.03	0.02
Shoulder Shape	0.23	0.01	0.0	0.03	0.02	-0.0	0.01	0.01	0.01	0.01	0.0	0.02	0.03	0.03	0.02	0.03
Proportions	0.16	0.0	0.0	0.02	0.02	-0.01	0.02	-0.01	0.02	-0.01	0.0	0.02	0.01	0.03	0.03	0.02

features particularly; a result achieved from "D" range experiments. This is because, apart from Skin Colour correlating with Ethnicity and possibly Hair Colour, there are fewer correlation between Ethnic appearance and other features. This is correct and to be expected; one can expect to find variations in Height, Build, Age and Sex regardless of Ethnicity. Any correlation with such features would require much larger datasets. Correlating ability is therefore a factor which makes Ethnicity and Skin Colour rank lower, therefore decreasing their overall scores if correlation metrics are given an equal weighting. However, by noticing that these features perform exceedingly well in some "C" retrieval experiments and equally so in the "B" Pearson's r stability experiments, we can better estimate their worth as traits.

Build features portray some ability, and depending on the voting scheme they can be shown to be better or worse than some global features. Both orderings show that

some build features, especially the more global features such as Height, Weight and Figure demonstrate real potential and usefulness as physiological traits to be described semantically. However it is also clear that, although useful in some retrieval experiments, low level build features such as Neck, Arm and Leg descriptions are less useful than more general notions of build, or apparently any other feature except for those seemingly useless descriptions of Shoulder and Proportion. Indeed, lower level build features were shown to have clear correlations with Weight and Height in Section 2.5.3. This may signify that, unless specially trained or lead to do so, humans are unlikely to have an accurate or useable opinion on specific areas of another person's body, opting instead to use Leg and Arm descriptions as synonyms for more general descriptions of bulk.

Another explanation may be that there simply was no opportunity for annotators to accurately describe limbs. It should be noted that such specific body features along with descriptions of Facial Hair perform well in the "B" Pearson's r experiments. This shows that they are stable, and yet they perform poorly in "C" retrieval and "A" separation analysis orderings. This along with the distribution seen in Section 2.5.2 may show that there simply were not enough individuals with particularly noteworthy limbs or facial hair. Given the subjective nature of the annotations gathered it is impossible to gauge whether individuals were simply not noticing the variance in limbs that existed, or whether the variance in limbs in the dataset was not high enough to be noticed. Another experiment beyond the scope of this thesis will need to be performed to investigate this matter further.

Overall, we see a preference of less precise features over specific features for purposes of semantic description. The generally understandable and immediately recognisable global features of Sex, Ethnicity and Age along with the global descriptions of bulk seem to be powerful in retrieval, accurately annotated and stable. This is in contrast with the more specific and more unclear features of limb description, shoulder shape and proportions which perform unequivocally worse. There are evolutionary arguments one can make to put these results in context. Indeed, there are several potential benefits to be gained from being able to make accurate and speedy judgments with regards to any given individual. Important questions can be answered, such as: is this person Male or Female, and therefore are they a potential mate; are they young or old and therefore can a status judgment be made; are they of my people or are they strangers and finally, are they generally bigger or generally smaller in build and therefore, is there danger? These decisions must be made quickly and accurately, a skill which our results seem to confirm.

5.5 Conclusions

In this chapter we have presented a collation of the analysis of the previous chapters. We present a combination of the various ranking schemes highlighted in Chapters 2 through 3. Using two combination techniques, one taking raw ranking and another incorporating a metric of confidence, we present a final ordering of our semantic traits. The ordering highlights the power of less precise features over specific features, with global features out ranking build features, and less specific build features out ranking granular ones.

The following chapter concludes this thesis by presenting discussions of possible future research directions.

Chapter 6

Future Work

6.1 Introduction

In this thesis we have explored semantic descriptions for a set of physical traits and shown their utility in biometric fusion and information retrieval. In this section we discuss future directions of this research and some of the open questions.

6.2 Semantic Terms

In this work we have outlined a set of descriptions useful for the purposes of human description at a distance using semantic terms. However, this set is by no means exhaustive and certain features visible at a distance may not be represented. Subsequently efforts should be made towards the expansion of the corpus of semantic traits identified thus far to include other traits defining other physical appearances, and furthermore, subject actions and environments.

Physical appearance traits such as clothing, piercings or distinguishing marks have yet to be explored. Although such features are easily altered and changed over larger time scale, they are often mentioned by witnesses of crime and help forge an annotators perception of a subject. Clothes are also mentioned in the Police RMS investigated in Chapter 2.

Semantic descriptions of actions could also be investigated. Action descriptions may include perceived mood, subject goals and social roles. These topics are difficult to explore automatically in the general case, but are readily mentioned by humans semantically, for example the concept of a *suspicious* action. Action features also complement dynamic

aspects of gait, rather than the static aspects captured by physical descriptions studied thus far.

A subject's location and environment undoubtedly affect perception of subject features, but also define the concept of outliers and "Unusual" behaviour. Questions such as "Is this subject acting inappropriately?" or "Does this person look out of place?" are inherently related to the environments within which the subject is observed. An exploration into these semantic attributes, supported by this initial work, will facilitate the involvement of human knowledge in biometric systems and also help bridge the semantic gap.

6.3 Practical Applications

Throughout this work we have concentrated on semantic description of traits regularly collected by police in witness statements. Our work has shown the practical application of these traits and associated terms in combination with biometric signatures both for identification and retrieval. This analysis can in turn feed back into police procedure and evidence analysis. By better understanding which features have potential for high reliability across a population, police investigations and witness questioning can be performed with more rigour. Furthermore, through the understanding of which biometrics perform well with which semantic traits, surveillance strategies and querying systems can be improved by recording appropriate details for the appropriate semantic features and also cater for human semantic queries.

6.4 Trait and Term Validity

In this work we have attempted to suggest justifications for the traits and terms outlined and we have also shown that if a subset of our terms are available, improvements in recognition can be achieved and retrieval can be facilitated. Our research has gone to great lengths to use a set of features that are consistently available in real world scenarios, and that when available are accurate.

Our annotation gathering process was specifically designed to be interrogative to avoiding defaulting issues. Therefore, a future study must be formulated to further explore the validity of the semantic terms used. This can be readily achieved by gathering semantic annotations from more individuals spanning a larger set of subjects being annotated. In doing so we can better understand the discriminatory ability of our chosen traits and terms.

Furthermore our current research does not explicitly explore how common or prevalent our chosen terms actually are in day to day human description. A study is required to gauge whether our physical features are adequate to encapsulate descriptions given in real world scenarios. This could be achieved through experiments performed in reverse of those presented by MacLeod et al. [79], i.e. full text descriptions of individuals could be gathered and our feature set used to encapsulate their descriptions. The utility of our features would be measured against their ability to define descriptions given and also on how many features in our set are actually consistently used.

In this thesis we prescribed a subset of semantic terms designed to encompass the broader range of terms which could be used to describe an individual physically. Though useful for an initial investigation, this approach could be broadened to incorporate a larger set of terms through the construction of ontologies or the use of subsections of existing ontologies of terms such as WordNet [29] or CYC [73]. The structures of such ontologies allow the analysis of a large set of terms and specify their interconnected structure. Ontologies allow the explicit consideration of inherently difficult aspects of dealing with a larger corpus of terms such as synonyms (e.g. describing a Large individual as Huge, Massive or Built) as well as terms describing multiple traits simultaneously (e.g. Gangly, describing an individual who is simultaneously Tall, Thin and awkwardly built). Natural language descriptions of physical traits could be subsequently analysed more efficiently and incorporated readily in the retrieval and identification applications outlined in this thesis.

6.5 Ground Truths

Throughout this work, while answering questions such as the retrieval and recognition capability of the traits, we could only indirectly ascertain the accuracy of any given annotation. This resulted in certain avenues of analysis being left untouched. In Chapter 2.5.3.1 we could not fully understand the reason for the strong annotation correlations between related descriptions of Weight and between related descriptions of Height. Similarly in Chapter 2.5.2 it was impossible to understand why self annotation distributions across two datasets were so similar. To explore this aspect of the problem, a ground truth must be gathered of exact subject measurements. This includes measurements of Ethnic Origin, Height, Weight, Limb Lengths, Hair Length and exact colour measurements of Skin and Hair. Once these measurements are known their relationship to semantic annotations can be analysed and therefore the accuracy of annotations and variance of semantic annotations can be understood.

6.6 Fusion Approaches

The fusion approaches used in our current research were relatively naive in their assumptions, primarily due to the lack of large training sets required for more complex density based approaches. An open area of study is the optimal fusion technique for the new semantic biometric we have outlined. We propose that more semantic data should be gathered so more rigorous density estimation based score fusion strategies can be effectively investigated.

We also suggest the fusion of annotator self descriptions, or descriptions others have given of the annotator, with the annotators response. Our system currently holds such information, aiming to take into account subject variables. Improvements in both recognition and retrieval could be achieved if annotations are normalised according to the annotator themselves.

6.7 CBIR Refinement

Several interesting avenues of research were opened with the retrieval experiments undertaken. Firstly, the LSI approach chosen is by no means the only approach available with regards to exploration of the correlation between semantic and visual spaces. Probabilistic Latent Semantic Analysis (PLSA) uses a Bayesian model to calculate the conditional probability of terms and documents belonging to underlying latent classes, estimated using an iterative Expectation Maximisation (EM) method. The successful use of PLSA in the past [85] for automatic image annotation as well as semantic behavioural inference [74], with some teams reporting improvements compared to SVD based LSI, warrant an investigation of the use PLSA in biometric CBIR.

Chapter 7

Conclusion

Semantic descriptions are a natural way humans use to describe one another. In this thesis, by formalising and collecting a set of semantic descriptions, we have shown their use in biometrics and surveillance scenarios. In Chapter 2 we outline our set of semantic descriptions and describe a novel dataset of semantic annotations gathered against two existing biometric datasets. In Chapter 3, we explore the use of semantic annotations as a soft biometric. We show their application in identification scenarios both in isolation and in fusion where we achieve better results than existing biometric signatures in isolation. In Chapter 4 we show the application of semantic annotations in a surveillance retrieval scenario using LSI techniques. Finally, in Chapter 5 we explore which semantic descriptions are most significant in the context of biometrics and retrieval tasks.

Bibliography

[1] Pylons python web framework. http://www.pylonshq.com.

[2] J. K. Aggarwal and Q. Cai. Human motion analysis: A review. *Computer Vision and Image Understanding*, 73(3):428–440, 1999.

[3] G. Barbujani. Human races: Classifying people vs understanding diversity. *Current Genomics*, 6:215–226(12), Jun. 2005.

[4] A. I. Bazin, L. Middleton, and M. S. Nixon. Probabilistic Fusion of Gait Features for Biometric Verification. In *Proc. Fusion*, 2005.

[5] S. Ben-Yacoub, Y. Abdeljaoued, and E. Mayoraz. Fusion of face and speech data for person identity verification. *IEEE Trans. NN*, 10(5):1065–1074, Sep. 1999. ISSN 1045-9227. doi: 10.1109/72.788647.

[6] C. BenAbdelkader, R. Cutler, and L. Davis. Stride and cadence as a biometric in automatic person identification and verification. In *Proc. IEEE FG*, pages 372–377, May 2002.

[7] J. Bennetto. Big brother britain 2006: "We are waking up to a surveillance society all around us". *The Independent*, 2006.

[8] L. E. Berk. *Development through the lifespan.* Allyn & Bacon, third edition, 1999.

[9] M. W. Berry and R. D. Fierro. Low-rank orthogonal decompositions for information retrieval applications. *Numerical Linear Algebra with Applications*, 3(4): 301–327, 1996.

[10] M. W. Berry, S. T. Dumais, G. W. O'Brien, and M. W. Berry. Using linear algebra for intelligent information retrieval. *SIAM Review*, 37:573–595, 1995.

[11] M. W. Berry, Z. Drmac, and E. R. Jessup. Matrices, vector spaces, and information retrieval. *SIAM Rev.*, 41(2):335–362, 1999.

[12] A. Bertillon. *Instructions For Taking Descriptions For The Identification Of Criminals And Others, By Means Of Anthropometric Indications*. American Bertillon Prison Bureau, 1889.

[13] V. Blanz and T. Vetter. Face recognition based on fitting a 3d morphable model. *IEEE Trans. PAMI*, 25(9):1063–1074, 2003.

[14] W. W. Bledsoe. The model method in facial recognition. Technical Report PRI 15, Panoramic Research, Inc., Palo Alto, California, 1964.

[15] G. Bradski and A. Kaehler. *Learning OpenCV: Computer Vision with the OpenCV Library*. O'Reilly, Cambridge, MA, 2008.

[16] R. Brunelli and D. Falavigna. Person identification using multiple cues. *IEEE Trans. PAMI*, 17:955–966, 1995.

[17] M. L. Cascia, S. Sethi, and S. Sclaroff. Combining textual and visual cues for content-based image retrieval on the world wide web. In *Proc. IEEE CBAIVL*, pages 24–28, 1998.

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

[19] M.-C. Cheung, M.-W. Mak, and S.-Y. Kung. A two-level fusion approach to multimodal biometric verification. *ICASSP*, 5:v/485–v/488 Vol. 5, Mar. 2005. ISSN 1520-6149. doi: 10.1109/ICASSP.2005.1416346.

[20] C. Chibelushi, J. Mason, and F. Deravi. Feature-level data fusion for bimodal person recognition. In *Proc. ICIPA*, volume 1, pages 399–403 vol.1, Jul 1997.

[21] J. Daugman. Biometric decision landscapes. Tech Report, 2000.

[22] A. Davies and S. Velastin. A Progress Review of Intelligent CCTV Surveillance Systems. In *Proc. IEEE IDAACS*, pages 417–423, Sept. 2005.

[23] R. M. Dawes. Suppose We Measured Height With Rating Scales Instead of Rulers. *App. Psych. Meas.*, 1(2):267–273, 1977.

[24] S. C. Deerwester, S. T. Dumais, T. K., G. W. Furnas, and R. A. Harshman. Indexing by latent semantic analysis. *J. of the American Society of Information Science*, 41(6):391–407, 1990.

[25] S. I. Dumais. Improving the retrieval of information from external sources. behavior research methods. *Instruments and Computers*, pages 229–236, 1991.

[26] C. Eckart and G. Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, Sep. 1936.

[27] H. D. Ellis. *Practical aspects of facial memory*, section 2, pages 12–37. Eyewitness Testimony: Psychological perspectives. Cambridge University Press, 1984.

[28] Y. Fang, T. Tan, and Y. Wang. Fusion of global and local features for face verification. In *Proc. ICPR*, pages 382–385, 2002.

[29] C. Fellbaum, editor. *WordNet: an electronic lexical database*. MIT Press, 1998.

[30] G. Feng, K. Dong, D. Hu, and D. Zhang. When faces are combined with palmprints: A novel biometric fusion strategy. In *Proc. ICBA*, pages 701–707, 2004.

[31] J. Fierrez-Aguilar, J. Ortega-Garcia, D. Garcia-Romero, and J. Gonzalez-Rodriguez. A comparative evaluation of fusion strategies for multimodal biometric verification. In *Proc. AVBPA*, pages 1056–1056, 2003.

[32] R. H. Flin and J. W. Shepherd. Tall stories: Eyewitnesses' ability to estimate height and weight characteristics. *Human Learning*, 5, 1986.

[33] P. Frost. European hair and eye color a case of frequency-dependent sexual selection? *Evolution and Human Behavior*, 27:85–103, 2006.

[34] F. Galton. Composite portraits made by combining those of many different persons into a single figure. *Journal of the Anthropological Institute*, 1879.

[35] S. J. Gould. The Geometer of Race. *Discover*, pages 65–69, 1994.

[36] W. Grosky and R. Zhao. Negotiating the semantic gap: From feature maps to semantic landscapes. In *Proc. SOFSEM*, pages 33–52, 2001.

[37] B. Guo and M. S. Nixon. Gait feature subset selection by mutual information. *IEEE Trans. SMC(A)*, 39(1):36–46, 2009.

[38] J. Han and B. Bhanu. Statistical feature fusion for gait-based human recognition. In *Proc. IEEE CVPR*, volume 2, pages II–842–II–847 Vol.2, Jun. 2004.

[39] J. S. Hare and P. H. Lewis. Saliency-based models of image content and their application to auto-annotation by semantic propagation. In *Proc. ESWC*, 2005.

[40] J. S. Hare and P. H. Lewis. On image retrieval using salient regions with vector-spaces and latent semantics. In *Proc. CIVR*, pages 540–549, 2005.

[41] J. S. Hare, P. H. Lewis, P. G. B. Enser, and C. J. Sandom. A Linear-Algebraic Technique with an Application in Semantic Image Retrieval. In *Proc. CIVR*, pages 31–40, 2006.

[42] J. S. Hare, S. Samangooei, P. H. Lewis, and M. S. Nixon. Semantic spaces revisited: investigating the performance of auto-annotation and semantic retrieval using semantic spaces. In *Proc. CIVR*, pages 359–368, New York, NY, USA, 2008. ACM.

[43] G. A. Harrison. Differences In Human Pigmentation: Measurements, Geographic Variation and Causes. *Investigative Dermatology*, 60:418–426, 1973.

[44] J. Hewig, R. H. Trippe, H. Hecht, T. Straube, and W. H. R. Miltner. Gender Differences for Specific Body Regions When Looking at Men and Women. *J. of Nonverbal Behavior*, 32(2):67–78, 2008.

[45] T. K. Ho, J. J. Hull, and S. N. Srihari. Decision Combination in Multiple Classifier Systems. *IEEE Trans. PAMI*, 16(1):66–75, 1994.

[46] L. Hong, A. Jain, and S. Pankanti. Can Multibiometrics Improve Performance. In *Proc. AUTOID*, 1999.

[47] W. Hu, T. Tan, L. Wang, and S. Maybank. A survey on visual surveillance of object motion and behaviors. *IEEE Trans. SMC(A)*, 34(3):334–352, 2004.

[48] D. J. Hurley, B. Arbab-Zavar, and M. S. Nixon. *The Ear as a Biometric*, chapter 7. Handbook of Biometrics. Springer, 2008.

[49] Interpol. Disaster Victim Identification Form (Yellow), 2008.

[50] J. Conroy and D. P. O'Leary. Text summarization via hidden markov models and pivoted qr matrix decomposition. Technical report, Univ. of MD Comp. Sci., 2001.

[51] A. Jain, S. Dass, and K. Nandakumar. Can soft biometric traits assist user recognition. In *Proc. SPIE*, 2004.

[52] A. Jain, K. Nandakumar, and A. Ross. Score normalization in multimodal biometric systems. *Pattern Recognition*, 38:2270–2285, Dec 2005.

[53] A. K. Jain and S. Z. Li. *Handbook of Face Recognition*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2005.

[54] A. K. Jain and A. Ross. Multibiometric systems. *Commun. ACM*, 47(1):34–40, 2004.

[55] A. K. Jain, R. Bolle, and S. Pankanti. *Biometrics: Personal Identification in Networked Society*, chapter 1 Introduction To Biometrics, pages 1–37. Kluwer Academic Publishers, 1999.

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

[57] A. K. Jain, A. Ross, and S. Prabhakar. An Introduction to Biometric Recognition. *IEEE Trans. CSVT*, 14:4–19, 2004.

[58] A. K. Jain, P. Flynn, and A. A. Ross. *Handbook of Biometrics*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2007.

[59] L. Jimenez, A. Morales-Morell, and A. Creus. Classification of hyperdimensional data based on feature and decision fusion approaches using projection pursuit, majority voting, and neural networks. *IEEE Trans. GRS*, 37(3):1360–1366, May 1999.

[60] G. Johansson. Visual perception of biological motion and a model for its analysis. *Percept. Phychophys.*, 14(2):201–211, 1973.

[61] A. Kale, A. Roychowdhury, and R. Chellappa. Fusion of gait and face for human identification. In *Proc. IEEE ICASSP*, volume 5, pages 901–904, 2004.

[62] T. Kanade. *Computer Recognition of Human Faces*, 47, 1977.

[63] M. D. Kelly. *Visual identification of people by computer*. PhD thesis, Stanford University, 1971.

[64] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas. On combining classifiers. *IEEE Trans. PAMI*, 20(3):226–239, 1998.

[65] R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proc. IJCAI*, pages 1137–1145, 1995.

[66] T. G. Kolda and D. P. O'Leary. A semidiscrete matrix decomposition for latent semantic indexing information retrieval. *ACM Trans. Inf. Syst.*, 16(4):322–346, 1998.

[67] P. V. Koppen and S. K. Lochun. Portraying perpetrators; the validity of offender descriptions by witnesses. *Law and Human Behavior*, 21(6):662–685, 1997.

[68] L. T. Kozlowski and J. E. Cutting. Recognizing the sex of a walker from a dynamic point-light display. *Perception & Psychophysics*, 21:575–580, 1977.

[69] L. I. Kuncheva, C. J. Whitaker, C. A. Shipp, and R. P. W. Duin. Is independence good for combining classifiers? In *ICPR*, volume 15, 2000.

[70] L. I. Kuncheva, C. J. Whitaker, C. A. Shipp, and R. P. W. Duin. Limits on the majority vote accuracy in classifier fusion. *Pattern Analysis & Applications*, 6: 22–31, Apr 2003.

[71] L. Lam and S. Suen. Application of majority voting to pattern recognition: an analysis of its behavior and performance. *IEEE Trans. SMC(A)*, 27(5):553–568, Sep 1997.

[72] T. Landauer and M. Littman. Fully automatic cross-language document retrieval using latent semantic indexing. In *Proc. Annual Conference of the UW Centre for the New OED.*, pages 31–38, 1990.

[73] D. B. Lenat and R. V. Guha. *Building Large Knowledge-Based Systems; Representation and Inference in the Cyc Project*. Addison-Wesley Longman Publishing Co., Inc., 1989.

[74] J. Li, S. Gong, and T. Xiang. Global behaviour inference using probabilistic latent semantic analysis. In *British Machine Vision Conference*, 2008.

[75] X. Li, S. Maybank, S. Yan, D. Tao, and D. Xu. Gait components and their application to gender recognition. *IEEE Trans. SMC(C)*, 38(2):145–155, Mar. 2008.

[76] R. Lindsay, R. Martin, and L. Webber. Default values in eyewitness descriptions. *Law and Human Behavior*, 18(5):527–541, 1994.

[77] J. Little and J. Boyd. Describing motion for recognition. In *Proc. ISCV*, page 5A Motion II, 1995.

[78] Z. Liu and S. Sarkar. Simplest representation yet for gait recognition: averaged silhouette. In *Proc. ICPR*, volume 4, pages 211–214 Vol.4, Aug. 2004.

[79] M. D. MacLeod, J. N. Frowley, and J. W. Shepherd. *Whole body information: Its relevance to eyewitnesses*, chapter 6. Adult Eyewitness Testimony. Cambridge University Press, 1994.

[80] C. N. Macrae and G. V. Bodenhausen. Social Cognition: Thinking Categorically about Others. *Ann. Review of Psych.*, 51(1):93–120, 2000.

[81] D. I. Martin and M. W. Berry. Mathematical Foundations Behind Latent Semantic Analysis. In T. K. Landauer, D. S. McNamara, S. Dennis, and W. Kintsch, editors, *Handbook of Latent Semantic Analysis*, chapter 2. Lawrence Erlbaum Associates, 2007.

[82] A. M. Martinez. The ar face database. *CVC Technical Report*, 24, 1998.

[83] K. Messer, J. Matas, J. Kittler, and K. Jonsson. XM2VTSDB: The extended M2VTS database. In *AVBPA*, pages 72–77, 1999.

[84] L. Middleton, D. K. Wagg, A. I. Bazin, J. N. Carter, and M. S. Nixon. A smart environment for biometric capture. In *IEEE Conference on Automation Science and Engineering*, 2006.

[85] F. Monay and D. Gatica-Perez. On image auto-annotation with latent space models. In *Proc. Multimedia*, pages 275–278, 2003.

[86] Y. Moon, H. Yeung, K. Chan, and S. Chan. Template synthesis and image mosaicking for fingerprint registration: an experimental study. *ICASSP*, 5:409–412, 2004.

[87] H. Murase and R. Sakai. Moving object recognition in eigenspace representation: gait analysis and lip reading. *Pattern Recogn. Lett.*, 17(2):155–162, 1996.

[88] K. Nandakumar, S. C. Dass, and A. K. Jain. Soft biometric traits for personal recognition systems. In *Proc. ICBA*, pages 731–738, 2004.

[89] M. Nixon and J. Carter. Automatic recognition by gait. *Proceedings of the IEEE*, 94(11):2013–2024, Nov. 2006.

[90] M. S. Nixon and J. N. Carter. Automatic recognition by gait. *Proceedings of the IEEE*, 94(11):2013–2024, Nov. 2006.

[91] S. Niyogi and E. Adelson. Analyzing and recognizing walking figures in XYT. In *Proc. CVPR*, pages 469–474, Jun 1994.

[92] R. D. Olsen. A Fingerprint Fable: The Will and William West Case. *Identification News*, 37(11), 1987.

[93] A. J. O'Toole. Psychological and Neural Perspectives on Human Face Recognition. In *Handbook of Face Recognition*. Springer-Verlag, 2004.

[94] C. H. Papadimitriou, P. Raghavan, H. Tamaki, and S. Vempala. Latent semantic indexing: A probabilistic analysis. *Computer and System Sciences*, 61:217–235, 1998.

[95] G. Pavlich. The subjects of criminal identification. *Punishment Society*, 11(2): 171–190, 2009.

[96] Z. Pecenovic. Image retrieval using latent semantic indexing. Master's thesis, AudioVisual Communications Lab, Ecole Polytechnique, Federale de Lausanne, Switzerland, 1997.

[97] T. Pham, N. E. Maillot, J. Lim, and J. Chevallet. Latent semantic fusion model for image retrieval and annotation. In *Proc. CIKM*, pages 439–444, New York, NY, USA, 2007. ACM.

[98] P. J. Phillips. Support vector machines applied to face recognition. In *Proc. NIPS II*, pages 803–809, Cambridge, MA, USA, 1999. MIT Press. ISBN 0-262-11245-0.

[99] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss. The feret evaluation methodology for face-recognition algorithms. *IEEE Trans. PAMI*, 22:1090–1104, 2000.

[100] N. Pinto, J. J. Dicarlo, and D. D. Cox. Establishing Good Benchmarks and Baselines for Face Recognition. In *Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition*, 2008.

[101] J. G. Ponterotto and B. Mallinckrodt. Introduction to the special section on racial and ethnic identity in counseling psychology: Conceptual and methodological challenges and proposed solutions. *J. of Counselling Psych.*, 54(3):219–223, Jul. 2007.

[102] H. T. F. Rhodes. *Alphonse Bertillon, father of scientific detection.* George G. Harrap & Co. LTD, 1956.

[103] A. Ross and R. Govindarajan. Feature Level Fusion Using Hand and Face Biometrics. In *Proc. SPIE*, pages 196–204, 2005.

[104] A. Ross and A. Jain. Information fusion in biometrics. *Pattern Recogn. Lett.*, 24 (13):2115–2125, 2003.

[105] A. Ross, K. Nandakumar, and A. K. Jain. *Introduction to Multibiometrics*, chapter 14. Handbook of Biometrics. Springer, 2008.

[106] A. A. Ross, K. Nandakumar, and A. K. Jain. *Handbook of Multibiometrics (International Series on Biometrics).* Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[107] C. Rosse and J. L. V. Mejino. A reference ontology for biomedical informatics: the foundational model of anatomy. *J. of Biomed. Informatics*, 36(6):478–500, 2003.

[108] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620, 1975.

[109] S. Samangooei, B. Guo, and M. S. Nixon. The use of semantic human description as a soft biometric. In *Proc. IEEE BTAS*, 2008.

[110] C. Sanderson and K. K. Paliwal. Information Fusion and Person Verification Using Speech & Face Information. Technical Report 33, IDIAP, Sep. 2002.

[111] M. Savvides, J. Heo, and S. W. Park. *Face Recognition*, chapter 3. Handbook of Biometrics. Springer, 2008.

[112] R. D. Seely, S. Samangooei, L. Middleton, J. N. Carter, and M. S. Nixon. The University of Southampton Multi-Biometric Tunnel and introducing a novel 3D gait dataset. In *Proc. IEEE BTAS*, Sep. 2008.

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

[114] R. Snelick, M. Indovina, J. Yen, and A. Mink. Multimodal biometrics: Issues in design and testing. In *Proc. ICMI*, pages 68–72, 2003.

[115] B. Son and Y. Lee. Biometric authentication system using reduced joint feature vector of iris and face. In *Proc. AVBPA*, pages 513–522, 2005.

[116] S. V. Stevenage, M. S. Nixon, and K. Vince. Visual analysis of gait as a cue to identity. *App. Cog. Psych.*, 13(6):513–526, 1999.

[117] D. Swets and J. Weng. Using Discriminant Eigenfeatures for Image Retrieval. *IEEE Trans. PAMI*, 18:831–836, 1996.

[118] H. Tajfel. Social Psychology of Intergroup Relations. *Ann. Rev. of Psych.*, 33: 1–39, 1982.

[119] N. F. Troje, J. Sadr, and K. Nakayama. Axes vs averages: High-level representations of dynamic point-light forms. *Vis. Cog.*, 14:119–122, 2006.

[120] G. Trunk. A Problem of Dimensionality: A Simple Example. *IEEE Trans. PAMI*, 1979.

[121] G. Tullock. Problems of majority voting. *Journal of Political Economy*, 67:571, 1959.

[122] M. Turk and A. Pentland. Eigenfaces for recognition. *J. Cognitive Neuroscience*, 3(1):71–86, 1991.

[123] G. Veres, L. Gordon, J. Carter, and M. Nixon. What image information is important in silhouette-based gait recognition? In *Proc. IEEE CVPR*, volume 2, pages II–776–II–782 Vol.2, Jun. 2004.

[124] P. Verlinde and G. Cholet. Comparing decision fusion paradigms using k-nn based classifiers, decision trees and logistic regression in a multi-modal identity verification application. In *Proc. AVBPA*, pages 188–193, 1999.

[125] P. Viola and M. Jones. Rapid Object Detection using a Boosted Cascade of Simple Features. *IEEE CSC CVPR*, 1:511, 2001.

[126] P. Viola and M. Jones. Robust real-time object detection. *J. of Computer Vision*, 2002.

[127] B. Vrusias, D. Makris, J.-P. Renno, N. Newbold, K. Ahmad, and G. Jones. A framework for ontology enriched semantic annotation of cctv video. In *Proc. WIAMIS*, pages 5 – 5, Jun. 2007.

[128] J. L. Wayman. Benchmarking Large-Scale Biometric System: Issues and Feasibility. In *Proc. CTST*, 1997.

[129] G. L. Wells and E. A. Olson. Eyewitness testimony. *Ann. Rev. of Psych.*, 54: 277–295, 2003.

[130] L. Wiskott, J. M. Fellous, N. Krüger, and C. von der Malsburg. Face recognition by elastic bunch graph matching. *IEEE Trans. PAMI*, 19:775–779, 1999.

[131] L. Xu, A. Krzyzak, and C. Suen. Methods of combining multiple classifiers and their applications to handwriting recognition. *IEEE Trans. SMC*, 22(3):418–435, May/Jun. 1992.

[132] J. Yang, J.-y. Yang, D. Zhang, and J.-f. Lu. Feature fusion: parallel strategy vs. serial strategy. *Pattern Recognition*, 36:1369–1381, Jun 2003.

[133] M. H. Yang, D. J. Kriegman, and N. Ahuja. Detecting faces in images: A survey. *IEEE Trans. PAMI*, 24:34–58, 2002.

[134] Z. Yang, M. Li, and H. Ai. An experimental study on automatic face gender classification. *Pattern Recognition*, 3:1099–1102, 2006.

[135] A. D. Yarmey and M. J. Yarmey. Eyewitness recall and duration estimates in field settings. *J. of App. Soc. Psych.*, 27(4):330–344, 1997.

[136] R. Zewail, A. Elsafi, M. Saeb, and N. Hamdy. Soft and hard biometrics fusion for improved identity verification. *MWSCAS*, 1:225–8, 2004.

[137] R. Zhao and W. Grosky. From features to semantics: some preliminary results. In *Proc. ICME*, volume 2, pages 679–682 vol.2, 2000.

[138] R. Zhao and W. Grosky. Bridging the Semantic Gap in Image Retrieval. *IEEE Trans. Multimeda*, 4:189–200, 2002.

[139] W. Zhao, R. Chellappa, A. Rosenfeld, and P. J. Phillips. Face recognition: A literature survey. *ACM Computing Surveys*, pages 399–458, 2003.