

UNIVERSITY OF SOUTHAMPTON

FACULTY OF PHYSICAL AND APPLIED SCIENCES

Electronics and Computer Science

**Unconstrained Human Identification Using Comparative Facial Soft
Biometrics**

by

Nawaf Yousef Almudhahka

Thesis for the degree of Doctor of Philosophy

November 2, 2017

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF PHYSICAL AND APPLIED SCIENCES

Electronics and Computer Science

Doctor of Philosophy

UNCONSTRAINED HUMAN IDENTIFICATION USING COMPARATIVE FACIAL
SOFT BIOMETRICS

by Nawaf Yousef Almudhahka

The recent growth in CCTV systems and the challenges of automatically identifying humans under the adverse visual conditions of surveillance have increased the interest in soft biometrics, which are physical and behavioural attributes that are used to semantically describe people. Soft biometrics enable human identification under the challenging conditions of surveillance where it is impossible to acquire traditional biometrics such as iris and fingerprint. The existing work on facial soft biometrics is focused on categorical attributes, while comparative attributes have received very little attention, although they have demonstrated a better accuracy. Thus, it is still unknown whether comparative soft biometrics can scale to large and more realistic databases. Also, the automatic retrieval of comparative facial soft biometrics from images needs to be investigated.

The purpose of this thesis is to explore human identification and verification in large and realistic databases via comparative facial soft biometrics using the Labelled Faces in the Wild (LFW) database. A novel set of comparative facial soft biometrics is introduced, and a thorough analysis that assesses attribute significance and discriminative power is presented. Also, a set of identification and verification experiments was conducted to evaluate the comparative facial soft biometrics. Moreover, this thesis proposes MIURank, a novel fully unsupervised ranking algorithm that is based on mutual information.

The experiments demonstrate that a correct match can be found in the top 71 retrieved subjects from a database of 4038 subjects by comparing an unknown subject to ten subjects only. Additionally, the experiments reveal that face retrieval by verbal descriptions in a database of images can yield a correct match in the top 15 retrieved subjects from a database of 430 subjects. Furthermore, the performance analysis of the MIURank algorithm shows that it can result in a ranking accuracy that is comparable to the maximum likelihood estimator of Bradley-Terry and the state-of-the-art SerialRank algorithm. By these analyses and developments, it is now possible not only to use human labels for recognition, but also to derive them by computer vision.

Contents

Declaration of Authorship	xv
Acknowledgements	xvii
1 Context and Contributions	1
1.1 Context	1
1.2 Contributions	3
1.3 Publications	3
1.4 Thesis Outline	4
2 On Soft Biometrics	7
2.1 Human Identification Using Soft Biometrics	7
2.2 Literature Review	9
2.2.1 Facial Identification	10
2.2.1.1 Face Recognition in Psychology	10
2.2.1.2 Human Descriptions in Eyewitness Statements	11
2.2.2 Soft Biometrics	13
2.2.2.1 Global Soft Biometrics	13
2.2.2.2 Body and Clothing Soft Biometrics	14
2.2.3 Facial Soft Biometrics	16
2.3 Performance Evaluation in Biometric Systems	19
2.3.1 The Receiver Operating Characteristic (ROC)	19
2.3.2 The Cumulative Match Characteristic (CMC)	21
2.4 Summary	22
3 Comparative Facial Soft Biometrics	23
3.1 Relative Attributes	24
3.2 Defining Facial Attributes and Comparative Labels	26
3.3 Dataset and Label Acquisition via Crowdsourcing	26
3.4 Dataset Analysis	30
3.4.1 Dataset Distribution	30
3.4.2 Attribute Significance	31
3.4.2.1 Attribute Discriminative Power	31
3.4.2.2 Attribute Semantic Stability	35
3.4.3 Attribute Correlations	36
3.5 Experiments	38
3.5.1 Identification Using Facial Comparisons	38
3.5.2 Label Compression	39

3.5.3	Verification Using Facial Comparisons	40
3.5.4	Elo's K and Identification Performance	41
3.6	Summary	42
4	Ranking of Soft Biometrics	45
4.1	Introduction	45
4.2	Ranking Using Mutual Information	47
4.2.1	Algorithm Formulation	48
4.2.2	Example: Ranking Tennis Players	51
4.3	Experiments	53
4.3.1	Synthetic Dataset	53
4.3.1.1	Robustness to Noise	54
4.3.1.2	Tolerance for Missing Pairwise Comparisons	55
4.3.1.3	Effect of Dataset Size	57
4.3.2	Real Datasets	58
4.3.2.1	English Premier League	58
4.3.2.2	Human Identification Using Comparative Soft Biometrics	59
4.4	Conclusions	61
5	Unconstrained Identification Using Comparative Facial Soft Biometrics	63
5.1	Enhanced Comparative Facial Soft Biometrics	64
5.2	Dataset and Label Acquisition via Crowdsourcing	65
5.2.1	The LFW Database	65
5.2.2	Crowdsourcing of Comparative Labels	67
5.3	Dataset Analysis	68
5.3.1	Dataset Distribution	68
5.3.2	Attribute Significance	70
5.3.2.1	Attribute Discriminative Power	71
5.3.2.2	Attribute Semantic Stability	72
5.3.3	Attribute Correlations	75
5.4	Experiments	76
5.4.1	Unconstrained Identification Using Facial Comparisons	76
5.4.2	Unconstrained Verification Using Facial Comparisons	79
5.4.3	Attribute Contribution in Identification	80
5.4.4	Attribute Contribution in Verification	82
5.5	Conclusions	83
6	Automatic Biometric Signatures	85
6.1	Extracting Visual Features from Face Images	86
6.1.1	Facial Landmarks Detection and Parts Segmentation	86
6.1.2	Generating Visual Features	88
6.2	Retrieval of Biometric Signatures	89
6.2.1	Relative Rating of Visual Features (REL)	89
6.2.1.1	Multiple Linear Regression (LR)	89
6.2.1.2	Regression Trees (RT)	90
6.2.1.3	Support Vector Machines for Regression (SVR)	90

6.2.1.4	Analysis and Evaluation	91
6.2.2	Estimation of Comparative Labels (ECL)	96
6.2.2.1	Analysis and Evaluation	96
6.3	Experiments	101
6.3.1	Identity Retrieval by Semantic Descriptions	102
6.3.2	Verification Using Automatic Biometric Signatures	104
6.4	Conclusions	106
7	Conclusions and Future Work	109
7.1	Conclusions	109
7.2	Future Work	111
7.2.1	Soft Biometric Identification	111
7.2.2	Descriptions from Memory	111
7.2.3	Ranking of Soft Biometrics	111
7.2.4	Automatic Retrieval of Biometric Signatures	111
	Bibliography	113

List of Figures

1.1	CCTV footage released in connection with a robbery incident that took place in Edinburgh 2016 [1].	1
2.1	Face identification using semantic descriptions.	7
2.2	Example categorical soft biometric attributes for a sample from the BioT database.	8
2.3	Example comparative soft biometric attributes for two subjects from the BioT database. The labels express the presence of an attribute in subject A as compared to subject B	9
2.4	Example Receiver Operating Characteristic (ROC) curve.	20
2.5	Example Cumulative Match Characteristic (CMC) curve.	21
3.1	Age of samples from the LFW database expressed using: (a) categorical labels; and (b) comparative labels.	24
3.2	(a) The layout of the University of Southampton's Biometric Tunnel (BioT) [2]; (b) Example front image captured by the face camera in the tunnel, with the subject's face detected using the Viola-Jones detector [3].	28
3.3	Sample face images from the BioT dataset where the effect of some visual conditions can be noted: (a) occlusion of eyebrows; (b) pose variability; and (c) video motion blur.	28
3.4	Example question from the crowdsourcing job launched to collect comparative labels for the BioT dataset.	29
3.5	Distribution of the collected comparative labels for the BioT dataset.	30
3.6	Box plot for scores of the BioT attributes.	31
3.7	Distribution of the BioT attributes.	32
3.8	Ranking of selected attributes for the BioT dataset using the Elo rating system. For each attribute, the top image represents the top ranked subject, while the bottom image shows the least ranked subject. The comparative labels that correspond to each attribute are listed in Table 3.1.	33
3.9	Discriminative power of the BioT attributes.	35
3.10	Semantic stability of the BioT attributes.	36
3.11	Correlations between the BioT attributes using Pearson's r	37
3.12	Illustration for identification using comparative facial soft biometrics.	38
3.13	Identification performance on the BioT dataset using $c = \{5, 10, 15, 20\}$ comparisons.	39
3.14	Effect of compressing comparative labels on identification performance using $c = \{5, 10, 15, 20\}$ comparisons.	40
3.15	Verification performance of the BioT dataset using $c = \{5, 10, 15\}$ comparisons.	41

3.16	Effect of the score adjustment parameter, K , on identification performance (x-axis labels are displayed in steps of powers of 4).	42
4.1	Relation matrix for the AUSOpen 2016 subset.	52
4.2	Effect of uniform randomly distributed noise on the binary relation matrix, R .	54
4.3	Robustness to noisy comparisons.	55
4.4	Tolerance for missing comparisons.	56
4.5	Effect of dataset size on ranking accuracy.	57
4.6	Percentage of ranking upsets with the EPL dataset.	59
4.7	Identification performance of MIURank compared with the best outcome of the Elo rating system.	60
5.1	Sample face images from the LFW database that show variations in: (a) pose; (b) illumination (c); resolution; and (d) facial expressions.	66
5.2	Example question from the crowdsourced job launched to collect comparative labels for the LFW-V1 dataset.	66
5.3	Distribution of crowdsourced comparative labels for the LFW-V1 dataset.	68
5.4	Box plot for the scores of the attributes.	69
5.5	Relationship between normalised scores and ranks of the attributes.	70
5.6	Distribution of the attributes.	71
5.7	Visualization of the ranking of selected attributes for the LFW-V1 dataset using MIURank. For each attribute, the upper image represents the top ranked subject, while the lower image shows the least ranked subject. The comparative labels that correspond to each attribute are listed in Table 5.1.	72
5.8	The top five similarities among the subjects of the LFW-V1 dataset.	73
5.9	Discriminative power of the attributes.	74
5.10	Semantic stability of the attributes.	74
5.11	Correlations between the attributes using Pearson's r .	75
5.12	Identification performance with the LFW-V1 dataset using $c = \{10, 15, 20\}$ comparisons.	77
5.13	Identification performance in the LFW-V1 dataset with MIURank and the Elo rating system using $c = 10$ comparisons.	77
5.14	Effect of number of comparisons, c , on identification performance.	78
5.15	Compression of search range in the LFW-V1 dataset with different number of comparisons, c , and with probability $p = \{0.9, 0.95, 0.99\}$.	79
5.16	Verification performance of the comparative facial soft biometrics with the LFW-V1 dataset using $c = \{10, 15, 20\}$ comparisons.	80
5.17	Error curves for the comparative facial soft biometrics in terms of FPR and FNR for $c = \{10, 15, 20\}$ comparisons: (a) EER=12.14%; (b) EER=8.52%; and (c) EER=7.71%.	81
5.18	Identification performance with attributes' subsets ranked according to discriminative power based on: ANOVA, entropy and mutual information.	81
5.19	Verification performance with attributes' subsets ranked according to discriminative power based on: ANOVA, entropy and mutual information.	83
6.1	Automatic retrieval of biometric signatures from face images.	86
6.2	Localization of facial landmarks using a face alignment framework [4] for a sample face image from the LFW-MS4 dataset.	86

6.3	Segmented face parts with the corresponding attributes indices as described in Table 5.1.	87
6.4	Correspondence between semantic and visual scores resulting from the REL approach.	94
6.5	Visual correspondence among scores resulting from the REL approach. . .	95
6.6	Correspondence between semantic and visual scores resulting from the ECL approach.	97
6.7	Visual correspondence among scores resulting from the ECL approach. . .	98
6.8	Examples for the outcomes of ranking of: (a) <i>age</i> , (b) <i>gender</i> , (c) <i>eyebrow thickness</i> and (d) <i>nose width</i> on the LFW-MS4 dataset. For each of the shown attributes, the lower and upper rows represent the least and top ranked subjects respectively, while the columns represent the different methods of ranking the attribute.	99
6.9	The LFW-MS4 subjects with the weakest and the strongest overall presence of attributes.	100
6.10	Gender clustering based on relative gender scores.	100
6.11	Illustration for subject retrieval from a visual database using semantic descriptions.	102
6.12	Retrieval performance resulting from automatically generated biometric signatures.	103
6.13	Compression achieved in search range with probability, p , of finding a correct match within range.	104
6.14	Illustration for face verification using automatic biometric signatures. . .	104
6.15	Verification performance of automatically estimated biometric signatures: (a) Receiver Operating Characteristic (ROC) curve; (b) Error curves with the REL approach; and (c) Error curves with the ECL approach.	105
6.16	Genuine and imposter distributions of automatically estimated biometric signatures.	106
6.17	Example pairs of samples from the LFW-MS4 dataset: (a) - (c) are true positives, and (d) - (f) are false positives.	106

List of Tables

3.1	Soft biometric attributes.	27
3.2	Relation inference rules.	29
3.3	Crowdsourcing job statistics for the BioT dataset.	29
3.4	Discriminative power of the BioT attributes.	34
4.1	Example subset of players from AUSOpen 2016.	51
4.2	Performance vectors for AUSOpen 2016 subset.	52
4.3	Predicted scores and ranks for the AUSOpen 2016 subset.	53
5.1	Enhanced soft biometric set.	65
5.2	Crowdsourcing job statistics for the LFW-V1 dataset.	67
5.3	Discriminative power of the attributes.	73
6.1	Mean absolute error for the predicted scores.	93
6.2	Mean level of concordance for the predicted ranks.	93
6.3	Average prediction accuracy of comparative labels.	97
6.4	Summary for the correspondence analysis.	101

Declaration of Authorship

I, **Nawaf Yousef Almudhahka** , declare that the thesis entitled *Unconstrained Human Identification Using Comparative Facial Soft Biometrics* and the work presented in the thesis are both my own, and have been generated by me as the result of my own original research. I confirm that:

- this work was done wholly or mainly while in candidature for a research degree at this University;
- where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
- where I have consulted the published work of others, this is always clearly attributed;
- where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
- I have acknowledged all main sources of help;
- where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
- parts of this work have been published as: [5], [6], [7], [8] and [9]

Signed:.....

Date:.....

Acknowledgements

To my supervisors, Professor Mark Nixon and Dr Jonathon Hare, I would like to express my greatest gratitude for your persistent guidance, care and valuable advices throughout my PhD journey. Your enlightening feedback and continuous encouragement have always helped me to overcome the research difficulties. I am very pleased that I had the opportunity to do academic research on this interesting topic under your supervision.

To my father, I would like to express my deep gratitude and appreciation to you for being always beside me and supporting my dreams. To my mother, I am sincerely thankful to you for surrounding me with your love and care, and encouraging me towards the achievement. To my dear wife and life partner, I highly appreciate your support throughout this journey, and I am deeply thankful to you.

Finally, many thanks to the Public Authority for Applied Education and Training (PAAET) in the State of Kuwait for financially supporting my PhD scholarship and providing me with the convenient atmosphere for achievement during my PhD journey.

Chapter 1

Context and Contributions

1.1 Context

Soft biometrics refers to physical and behavioural attributes that are used to describe individuals verbally and that, accordingly, can be used to identify people. In criminal investigations, the first challenge that is faced by law enforcement agencies is the identification of suspects, which requires collecting eyewitness descriptions for the suspects' physical traits such as face, body and clothing [10]. In such situations, soft biometrics become vital, as they provide means of semantically describing suspects, and thus they can enable them to be identified from a database of mugshots or CCTV footage [11].



Figure 1.1: CCTV footage released in connection with a robbery incident that took place in Edinburgh 2016 [1].

Traditionally, humans have been automatically identified using hard biometrics such as fingerprint, DNA and iris. However, hard biometrics require individuals' cooperation to be acquired. Furthermore, they cannot be captured at a distance or under the adverse visual conditions of surveillance such as variations in illumination and resolution, as can

be seen in Figure 1.1. Soft biometrics can address human identification in challenging surveillance environments, as they can be captured at distances without subjects' involvement [12]. In addition, soft biometrics have excellent viewpoint invariance capabilities, which are attributed to the comprehension of the semantic space [13]. Taken together, these advantages promote the role of soft biometrics for unconstrained human identification and highlight their potential for bridging the semantic gap between humans and machines.

Driven by the increasing realisation of the importance of surveillance systems for crime prevention and security of societies, soft biometrics have been attracting considerable attention in the research community, and attributes that are extracted from face, body and clothing have been explored. Obviously, most of the existing studies in soft biometrics have focused on facial attributes, which can be attributed to the prominence and informative richness of the human face, as compared to other distance biometrics such as body, gait and clothing [11, 14]. However, the research in facial soft biometrics has focused on exploring identification using categorical attributes, which represent physical traits in absolute form (e.g., subject *A* has a *thick* eyebrow), while it has been demonstrated that estimating attributes in a comparative format (e.g., subject *A* has a *more thick* eyebrow than subject *B*) results in better recognition accuracy [15]. Nevertheless, there is an apparent knowledge gap in the literature concerning comparative facial soft biometrics that results from the following: (1) very little is known about identification using comparative facial soft biometrics, as they have been explored using a small constrained database with the labels collected from a limited group of annotators; (2) no single study exists that explores identification and verification via comparative facial soft biometrics in large unconstrained databases such as the Labeled Faces in the Wild (LFW) database; and (3) no studies have been found that investigate automatic retrieval of comparative facial attributes from images for biometric purposes.

In light of the inadequacies of the previous studies in comparative facial soft biometrics, this thesis aims to explore human identification via comparative facial soft biometrics in large unconstrained databases, and to investigate the automatic retrieval of comparative facial soft biometrics from images. The thesis extends the current limited knowledge about comparative facial soft biometrics by proposing a novel set of relative facial attributes, in addition to analysing the significance and discriminative power of facial attributes. More importantly, the thesis explores the reliability of comparative facial soft biometrics for identification and verification in large unconstrained databases using the well-known LFW database. Also, this thesis intends to determine the extent to which the semantic gap between humans and machines can be bridged, with regards to the estimation of relative facial attributes, and explores the automatic retrieval of comparative facial soft biometrics from images. Finally, the thesis examines face retrieval by verbal descriptions from a database of images, which simulates an operational scenario that is of high importance in real life.

1.2 Contributions

The main contributions of this thesis are:

- Exploration of human identification and verification via comparative facial soft biometrics in large unconstrained databases using the LFW database, and proposing a novel set of comparative facial soft biometrics that emphasises the relative importance of facial features in human face recognition.
- Introduction of a scheme for analysing the significance, semantic stability and discriminative power of comparative facial soft biometrics, in addition to assessing the contribution of attributes in identification and verification.
- Presentation of a novel fully unsupervised and parameterless algorithm for ranking from pairwise comparisons based on mutual information (MIURank), which exploits the solid science of information theory, added to the strong intuition of relative performance, to efficiently rank comparative soft biometric attributes.
- Investigation of the automatic retrieval of comparative facial soft biometrics from images and analysing the semantic gap between humans and machines with respect to comparative facial attributes. This includes the proposal of two novel approaches: the Relative Rating (REL) and the Estimation of Comparative Labels (ECL), for retrieving biometric signatures from images. Also, the thesis provisions a framework for face retrieval by semantic descriptions in a database of images.

1.3 Publications

The papers associated with this work are:

1. Nawaf Y. Almodhahka, Mark S. Nixon, Jonathon S. Hare. Human face identification via comparative soft biometrics. In *Identity, Security and Behaviour Analysis (ISBA), 2016 IEEE 2nd International Conference on*, pages 1-6. IEEE, 2016.
2. Nawaf Y. Almodhahka, Mark S. Nixon, Jonathon S. Hare. Unconstrained human identification using comparative facial soft biometrics. In *Biometrics Theory, Applications and Systems (BTAS), 2016 IEEE 8th International Conference on*, pages 1-6. IEEE, 2016.
3. Mark S. Nixon, Bingchen H. Guo, Sarah V. Stevenage, Emad S. Jaha, Nawaf Almodhahka, and Daniel Martinho-Corbishley. Towards automated eyewitness descriptions: describing the face, body and clothing for recognition. *Visual Cognition*, pages 1-15, 2016.

4. Nawaf Y. Almodhahka, Mark S. Nixon, Jonathon S. Hare. Automatic semantic face recognition. In *Automatic Face and Gesture Recognition (FG), 2017 IEEE 12th International Conference on*, pages 180-185. IEEE, 2017.
5. Nawaf Y. Almodhahka, Mark S. Nixon, Jonathon S. Hare. Mutual information for unsupervised ranking from pairwise comparisons. (In preparation)
6. Nawaf Y. Almodhahka, Mark S. Nixon, Jonathon S. Hare. Semantic face signatures: recognizing and retrieving faces by verbal descriptions. *IEEE Transactions on Information Forensics and Security*, 2017. (In press)
7. Nawaf Y. Almodhahka, Mark S. Nixon, Jonathon S. Hare. Comparative Face Soft Biometrics for Human Identification. *Surveillance in Action*. Springer International Publishing AG. (To be published)

1.4 Thesis Outline

This thesis is structured as follows:

Chapter 2 introduces soft biometrics, highlights their advantages over the traditional hard biometrics for human identification, and establishes their importance for society. The chapter outlines the different groups of soft biometrics and their formats (i.e. categorical and comparative). In addition, the chapter includes a literature review that surveys the relevant work that has been published so far concerning soft biometrics and highlights the inadequacies of previous studies. Finally, the chapter summarises the biometric performance evaluation measures that are used throughout the subsequent chapters of this thesis to report the experimental results.

Chapter 3 provides an overview of relative attributes and explains how they can be exploited to achieve soft biometric identification. Also, the chapter presents a novel set of soft biometric attributes with the corresponding comparative labels and describes how crowdsourced comparative labels can be utilised to generate biometric signatures, using the Elo rating system. The chapter explores constrained human face identification and verification via comparative soft biometrics using the University of Southampton's Biometric Tunnel (BioT) database. Moreover, it presents an evaluation of attribute significance and discriminative power. The chapter concludes by outlining the implications of the attribute analysis and the experimental results. In addition, it highlights the potential areas of enhancement for the soft biometric identification process in light of the analysis and experimental outcomes.

Chapter 4 discusses the ranking problem in the context of comparative soft biometrics in more depth and gives an overview of the existing algorithms for ranking from pairwise

comparisons. Furthermore, the chapter highlights the complexities resulted from training or tuning the existing algorithms. Subsequently, the chapter proposes MIURank, which is a novel fully unsupervised algorithm for ranking from pairwise comparisons using mutual information. In addition, the chapter presents experiments that evaluate MIURank with respect to well-known ranking algorithms using synthetic and real datasets, and outlines the findings of the experiments along with their implications.

Chapter 5 investigates human face identification and verification in large unconstrained databases, which reflects the realistic scenarios of identification. First, the chapter presents the LFW-V1 dataset that is used to explore unconstrained identification, and proposes an enhanced set of soft biometric attributes in light of the findings of the constrained identification analysis, which are presented in chapter 3. Furthermore, the chapter presents experiments that assess the reliability of comparative facial soft biometrics for identification and verification in large unconstrained databases. Lastly, the chapter presents experiments that evaluate the impact of attributes on identification and verification.

Chapter 6 explores the automatic retrieval of biometric signatures from images. The chapter examines the prediction of relative attributes, which compose the biometric signatures, using different visual descriptors and regression models. Also, the chapter proposes two approaches for generating automatic biometrics signatures: the ECL approach, which estimates comparative labels from differential visual features; and the REL approach, which is based on ranking from visual features. Using LFW-MS4, which is a multi-sample dataset extracted from the LFW database, the chapter analyses the attributes' correspondence between the semantic and visual spaces. Moreover, the chapter presents a retrieval experiment that simulates identifying a subject by verbal descriptions in a database of images.

Chapter 7 summarises the main conclusions, highlights the implications of the findings on research in comparative soft biometrics, and proposes directions for future work.

Chapter 2

On Soft Biometrics

The role of soft biometrics in human identification is recognised by a growing body of literature, and attributes have been proposed from various physical traits to cope with the challenging visual conditions of surveillance. This chapter provides an overview of soft biometric attributes, and highlight their importance for human identification and verification. Furthermore, the chapter presents a literature review that explores human face recognition from a psychological perspective, and explains the significance of attributes in eyewitness descriptions. The literature review also covers the work that has been undertaken so far with respect to soft biometric identification and verification, in addition to the automatic estimation of attributes from images. The chapter outlines the major knowledge gaps in the existing work, and accordingly describes the aim of this thesis. Finally, the chapter explains the performance measurement tools in biometric systems, which are used in the following chapters of this thesis to report the experimental results.

2.1 Human Identification Using Soft Biometrics

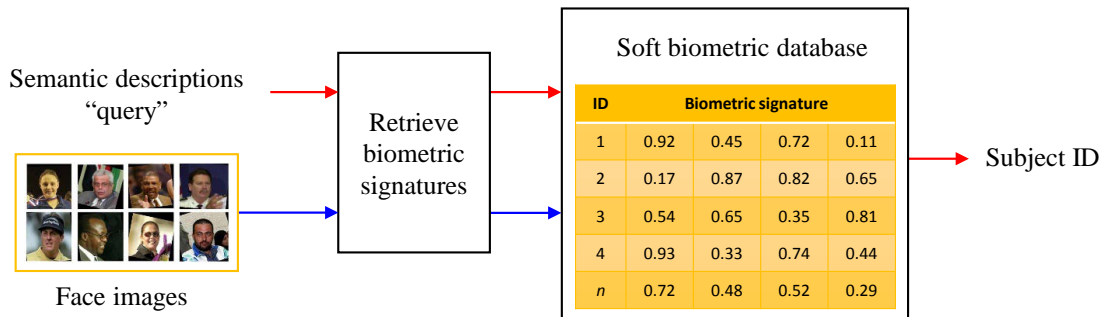


Figure 2.1: Face identification using semantic descriptions.

The increased awareness of the role of surveillance systems for public security and crime prevention has driven vast deployments of CCTV networks around the world [16, 17, 18, 19]. For example, in the UK alone, the number of CCTV cameras deployed in cities and town centres was estimated between 1.85 [20] and 5.9 million [21]. This growth in CCTV systems has increased the reliance on surveillance data for suspect identification, which is the first challenge faced by law enforcement agencies in criminal investigations [18]. As a result, the need for identifying suspects from imagery databases (i.e. mugshots or CCTV footage) has motivated research in human identification using semantic descriptions based on eyewitnesses statements with a view to enable searching a database of subjects through verbal descriptions [12, 13, 11]. These semantic descriptions are constructed from soft biometrics, which refer to physical and behavioural attributes that can be used to identify people. Whereas human identification has been traditionally based on hard biometrics such as iris, DNA and fingerprint, which require individuals' cooperation, soft biometrics, on the other-hand, can be acquired at a distance while enjoying more robustness to the challenging visual conditions of surveillance such as occlusion of features, viewpoint variance, low resolution and changes in illumination [12, 15, 11, 22]. Therefore, soft biometrics can play a significant role in criminal investigations, which require identity retrieval of suspects by verbal descriptions from an imagery database. Figure 2.1 illustrates semantic identification in soft biometric databases.

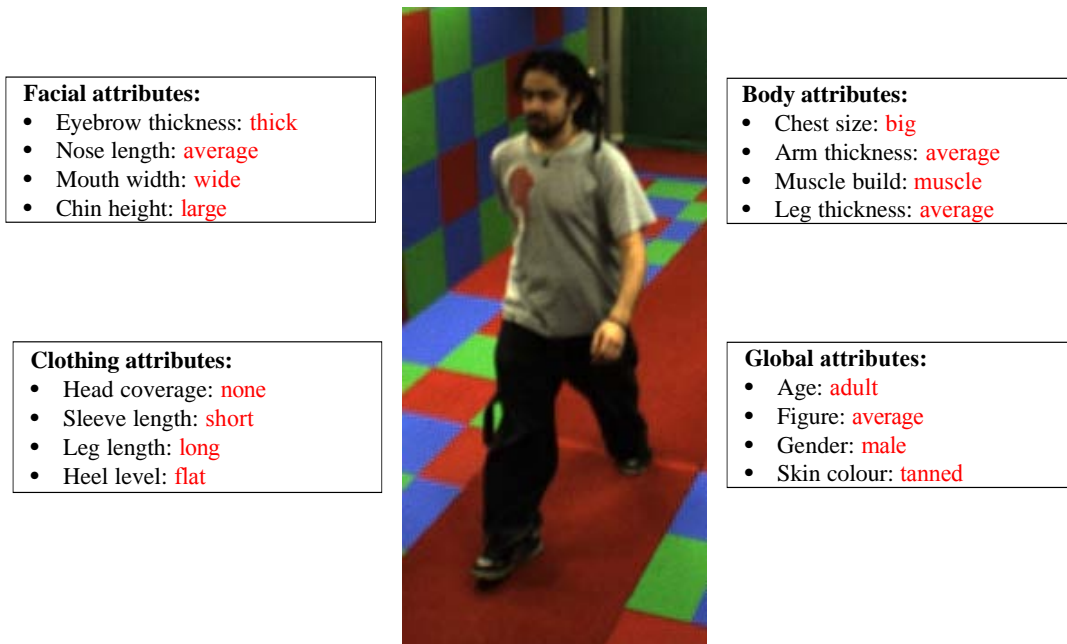


Figure 2.2: Example categorical soft biometric attributes for a sample from the BioT database.

In general, soft biometrics can be categorised based on two criteria: group and format. In terms of the group, soft biometrics may fall under global, facial, body, or clothing

attributes. In terms of format, soft biometrics can be classified as either categorical, where a person's attributes are assigned to specific classes (e.g., *square* versus *round* jaw), or comparative, where attributes of a person are classified relative to another person (e.g., subject *A* has a *more rounded* jaw than subject *B*) [23]. Comparative soft biometrics are discussed further in more details in the next chapter. This taxonomy of soft biometrics (i.e. group and format) will be followed throughout this chapter to enable a simple tracking of the existing work, and to highlight potential knowledge gaps in the field. Figures 2.2 and 2.3 show example soft biometric attributes in categorical and comparative formats, respectively.

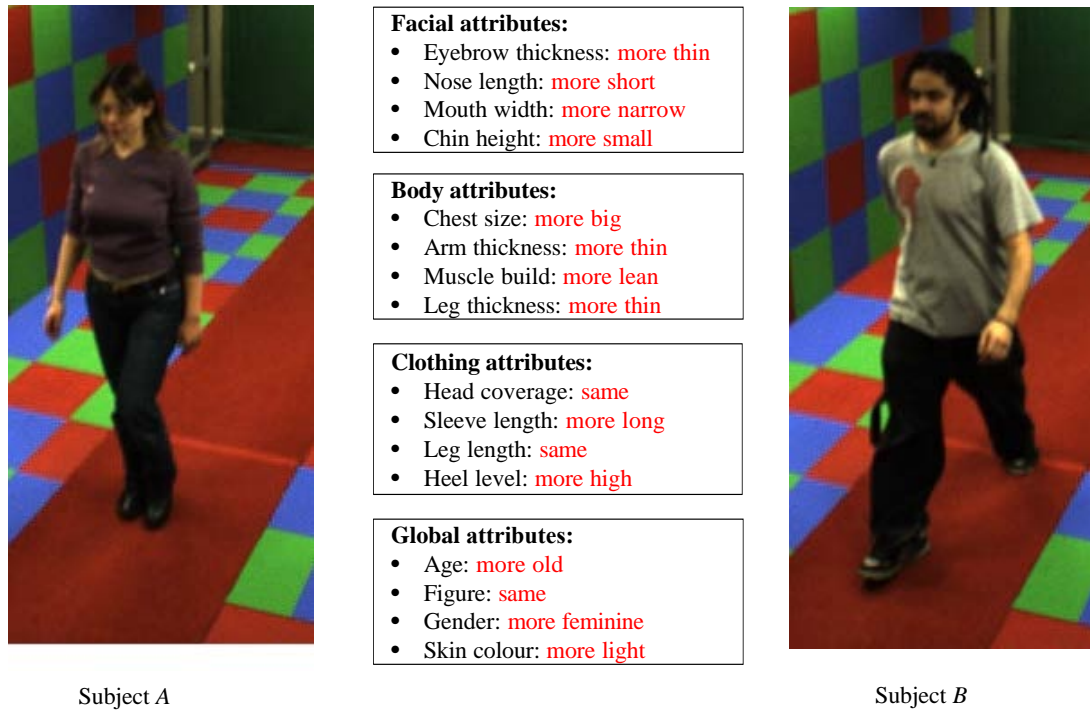


Figure 2.3: Example comparative soft biometric attributes for two subjects from the BioT database. The labels express the presence of an attribute in subject *A* as compared to subject *B*.

2.2 Literature Review

The purpose of this literature review is to show the main directions that have been taken so far in studying soft biometrics and to outline what has been already addressed by existing studies on soft biometrics. The literature review starts with background information about face recognition in psychology, and present highlights on research into eyewitness descriptions in criminal investigations. Then, a review of the research in soft biometrics is presented with an emphasis on facial attributes, as it is the main focus of this thesis. Throughout the literature review, the existing work is positioned with respect to the two-dimensional taxonomy of soft biometrics that was presented in the

previous section (i.e. group and format). Thus, from a group perspective, the literature review will address the existing work as non-facial (i.e. global, body and clothing), versus facial soft biometrics. For each group, the existing work will be surveyed in a chronological order for both categorical and comparative soft biometrics.

2.2.1 Facial Identification

2.2.1.1 Face Recognition in Psychology

Understanding face recognition from a psychological perspective is vital for studying facial identification using soft biometrics, as human identification accuracy is significantly affected by behavioural and physiological factors such as memory and encoding [24, 25]. In addition, human face recognition has substantial implications on automatic face recognition [26, 27], which has a wide range of real-life applications. The existing psychology literature on face recognition by humans is extensive and focuses particularly on understanding how the human visual and neural systems process the facial features and recognise faces. One of the key studies that has addressed face recognition in humans is that of Bruce and Young [28], who provisioned a theoretical model for understanding how humans recognize faces that is based on a set of distinguishable codes for face recognition, which are: label pictorial, structural, identity specific semantic, visually derived semantic and name. Bruce and Young's study suggested that regular face recognition involves the structural and identity-specific semantic codes (e.g., a familiar person's occupation and friends), while the other codes are used in lower levels for face recognition and perception. In [29], Hancock et al. outlined the factors that affect humans accuracy of recognizing unfamiliar faces. It has been found that changes in viewpoint and illumination significantly affect the ability of humans to recognize unfamiliar faces from images. Also, the spatial relationship between the facial components (i.e. configuration) has a high impact on recognition accuracy. Furthermore, the study emphasised the effect of face distinctiveness on recognition accuracy, and highlighted the role that can be played by the machines in aiding face recognition performance by humans.

A detailed overview of the human face recognition in a psychological and neural contexts was offered by O'Toole [30]. Her study provided some insights for addressing the problem of automatic face recognition. O'Toole stressed on humans' ability to identify faces with the aid of semantic categories such as age, gender and race, which increases recognition accuracy. Moreover, the study pointed to the significance of both the observer's and the observed person's race on the recognition accuracy. Sinha et al. [26] outlined the key findings from previous experimental studies of face recognition in humans that can aid face recognition in computational systems. Their study highlighted the impact of spatial resolution on recognition performance, in addition to highlighting the holistic

processing of facial features by the human visual system. Also, the study showed that pigmentation cues (i.e. texture or colour) have at least the same importance as the shape cues in face recognition. Tsao and Livingstone [31] highlighted the differences between computer vision algorithms and humans in face detection and recognition. The study emphasised on the independence of face detection and recognition stages in the human neural system, and outlined the differences in processing faces as compared with other objects in the neural system. Also, the study stressed the norm-based coding nature of face processing in humans, and the interpretation of faces in terms of their distinctiveness. Finally, the study highlighted the holistic processing of faces in the human neural system and stressed on the effect of the whole face on the processing of the individual facial components.

Several studies have investigated the significance of facial features in recognition performance. Davies et al. [32] assessed the saliency of different facial features, where their experiments were based on manipulating faces using the Photofit Kit and monitoring the identification performance of different alternatives of samples. They have found that the forehead and eyes have the highest saliency (i.e. their changes are most likely to be noticed), while the chin and nose have the least saliency. Haig [33] explored the relative importance of facial features in face recognition by manipulating the primary facial features of target samples and evaluating the recognition performance of the observers accordingly. The study found that eye-eyebrow, followed by mouth and nose, have the highest importance for face recognition. Sadr et al. [34] investigated the role of eyes and eyebrows in face recognition and found that eyebrows have a significant role in face recognition that is at least similar to that of the eyes. Furthermore, their experiments reported that the absence of eyebrows results in a larger degradation in the face recognition performance as compared to the absence of eyes.

In summary, the outcomes of these studies highlight the role of semantic bindings in human face recognition and show how categorisation of facial or personal characteristics can affect the human face recognition performance. In addition, they outline the relative importance of the different facial parts and show that the eye-eyebrow region is the most important in face recognition. The implications of these studies will be noted in the definition of facial soft biometrics in this thesis as shown later in chapters 3 and 5.

2.2.1.2 Human Descriptions in Eyewitness Statements

There is a large volume of published studies that have explored the reliability of verbal descriptions in eyewitness statements for suspects' identification in criminal investigations. Kuehn [35] evaluated the effectiveness of verbal descriptions provided by victims of violent crimes to the police and examined the descriptions' completeness using a random sample from the police records. The examined descriptions included the suspects' physical traits which are: age, sex, height, weight, build, skin colour, hair and eye. The

results of the study showed that more than 85% of the eyewitnesses were able to provide at least six attributes in their descriptions, where sex could be identified in 93% of the descriptions, while eye colour was the least to be recognised in the sample. The results also revealed that eyewitnesses cannot recall discrete traits of the suspects, but rather they have a general impression about the suspects.

In an analysis of the content of 139 eyewitnesses' descriptions, Sporer [10] found that 31% of the witnesses reported clothing attributes, 29.6% described facial features, 22% specified global features (i.e. age and gender) in addition to movement descriptions, and the remaining descriptors used other features (e.g., smell, accent and accessories). Sporer's analysis of the facial descriptions showed that most of the eyewitnesses described the upper half of the face, and more specifically, the hair of the suspect. Also, the study pointed out that although the hair was the most mentioned in eyewitnesses descriptions, it is less reliable for locating the suspects as compared with the inner facial features, since hairstyles can be easily changed.

Koppen et al. [36] assessed the completeness and accuracy of eyewitnesses' descriptions using 2299 statements of 1313 eyewitnesses for 582 different robbers from official police records, and investigated the factors that affect the accuracy and completeness of the statements. The findings that emerged from their study revealed that the eyewitnesses tended to provide more information about general descriptions (e.g., age and gender) as compared to facial features. In addition, the study showed that the completeness of the descriptions did not necessarily imply their accuracy, thus, although the information provided by the eyewitness was slight, it tended to be accurate. In [14], Burton et al. explored subjects' ability to identify target people from surveillance videos. They found that familiarity with target faces has a substantial impact on the accuracy of identification. Thus, face recognition performance with familiar targets is much better than it is with unfamiliar ones. Furthermore, the study revealed that even with the poor quality of surveillance data, the face has a significantly higher impact on identification accuracy compared with the gait and body.

A significant analysis was presented by Meissner et al. [37], which outlined the psychological factors that affect eyewitness descriptions as follows: (1) encoding-based, which affects a person's perception such as illumination, distance and stress; (2) person variables, which are age, gender and race; and (3) the effect of inaccurate information from co-witnesses or investigators. Lee et al. [38] have conducted a detailed examination of the impact of a feature-based approach in suspect identification. Their experiments have shown that using a subjective feature-based approach for retrieving suspects from a database of mugshots is more efficient and accurate than presenting mugshots for an eyewitness in arbitrary order. Their experiments have also revealed that the feature-based approach to identify suspects is effective for recognising faces in realistic conditions.

Overall, these studies reveal that the accuracy and completeness of eyewitnesses' descriptions are determined by multiple factors such as the spatial and temporal settings of the incident, in addition to the eyewitness personal variables (e.g., age and gender). Furthermore, the findings of these studies stress the tendency of eyewitnesses to describe general physical characteristics such as age and gender (i.e. global soft biometrics) in their statements, whereas facial features were less likely to be described. Also, the feature-based approach of identifying faces has been investigated and found to have a better impact on the efficiency and accuracy of suspect identification. Taken all together, these outcomes imply that global soft biometrics are essential for identification. In addition, the findings highlight the inadequacy of facial features for verbal descriptions as compared with other physical features. This suggests the introduction of more effective semantic facial attributes, which is one of the objectives of this thesis.

2.2.2 Soft Biometrics

2.2.2.1 Global Soft Biometrics

Global soft biometrics include information such as age, gender, weight (or figure), skin colour and ethnicity [12, 39]. As global soft biometrics lack for distinctiveness and cannot be used alone to differentiate between two people [40], they have been used in the literature as an ancillary information that could result in significant improvement in the identification performance when augmented with traditional biometrics such as fingerprint and face images. Also, global soft biometrics have been used for clustering databases in law enforcement agencies in order to achieve more efficient searches [41]. The earliest work that has explicitly mentioned soft biometrics is Wayman's study [42, 11], which proposed the use of global soft biometrics such as age, gender and ethnicity, to enhance the search in a database of traditional hard biometrics that consists of fingerprints and face images of 160 subjects. The global soft biometrics were used to enrich the information contained in biometric signatures. The study concluded that these soft biometrics limit the entries to be searched and improves the efficiency of biometric systems in terms of searching time. Jain et al. [40] proposed the utilisation of global soft biometrics (i.e. gender, height, weight, ethnicity and age) as an auxiliary information that is used to augment fingerprint identifiers. Their experiments using a database of 263 subjects have shown that soft biometrics can significantly improve the recognition performance when augmented with traditional biometrics such as the fingerprint.

Samangoeei et al. [12] proposed a set of soft biometric attributes that outlines a description of the physical characteristics of humans. Among those, age, race, sex and skin colour were included alongside groups of body soft biometrics. They evaluated the reliability of these soft biometrics using Analysis of Variance (ANOVA) and found that race, sex, and skin colour have the highest discriminative power as compared to the other soft

biometrics (i.e. body and head shape). In addition, they performed identification experiments utilising soft biometrics in isolation as well as in fusion with gait features using the Soton Gait Database (SGDB) [43]. Their experiments showed that soft biometrics are significantly effective for identification, and can improve identification performance of gait signatures when fused with them. In [39], Tome et al. used age, ethnicity and sex, in addition to other soft biometrics as ancillary information that can be augmented with face hard biometrics to improve the performance of a face recognition system. Their experiments revealed that the fusing soft biometrics with traditional face recognition information results in improving identification performance especially at a distance.

Whereas the previously mentioned studies used categorical soft biometrics, Reid and Nixon [23] were the first to introduce comparative soft biometrics, which are based on visual comparisons between individuals. Their study included the definition of a novel soft biometric set that consists of body attributes with corresponding comparative labels. Their soft biometric bundle also included global attributes, with age expressed in comparative format (i.e. *younger* versus *older*), while maintaining other global soft biometrics (i.e. weight, ethnicity and skin colour) in categorical format. Using the SGDB database, their innovative approach of utilising comparative soft biometrics showed more robustness and better recognition accuracy as compared with categorical soft biometrics. Later, Martinho-Corbishley et al. [44] extended the use of comparative global soft biometrics by including figure and gender (i.e. *feminine* versus *masculine*). Their experiments showed that relative gender scores exhibit a highly binary distribution between *feminine* and *masculine* attributes. Recent work by Martinho-Corbishley et al. [45] studied the categorization of gender from surveillance footage as a soft biometric attribute. This can be achieved by encountering ambiguity and uncertainty to explore super fine-grained taxonomies of gender from a body that goes beyond binary male-female classification.

In a nutshell, it can be noted from these studies that improving traditional biometric systems such as fingerprint, gait or face was the motivating factor behind the exploration of global soft biometrics. Also, it can be noted that most of the existing work is focused on categorical global soft biometrics, whereas comparative soft biometrics have not been explored adequately, although they have shown a better impact on identification performance [23, 15].

2.2.2.2 Body and Clothing Soft Biometrics

There are relatively few studies in the area of body soft biometrics; Samangoeei et al. [12] outlined the first approach for identification using categorical body soft biometrics [11]. The study included a set of categorical soft biometrics that describes human body and head shape in addition to global soft biometrics. Using the SGDB database,

their experiments revealed the identification capabilities inherited in body soft biometrics. In addition, their experiments showed that body soft biometrics could improve gait recognition performance when fused with gait signatures. Later, Reid and Nixon [23] explored human identification through the body soft biometrics defined by Samangooei et al. in [12] using the same dataset. However, they presented the soft labels in an innovative format, where a soft biometric attribute is expressed as a comparison between two individuals (e.g., subject *A* has *much longer* arm than subject *B*), which is utilised to construct a biometric signature for each subject using the Elo rating system [46]. The comparative body soft biometrics introduced by Reid and Nixon significantly outperformed the classical categorical soft biometrics used in [12] achieving a retrieval rate of 92% at rank-1, while the categorical soft biometrics achieved a retrieval rate of 48% at the same rank.

Tome et al. [39] explored the utilisation of body and global soft biometrics as ancillary information that can be easily extracted at a distance to improve face recognition performance. In their study, 23 categorical soft biometrics were defined (13 body, 3 global and 7 head), and feature fusion was utilised in performing face recognition experiments using the BioT database [2]. The performance of soft biometrics at varying distances was evaluated, and the results showed that soft biometrics are capable of improving face recognition performance particularly at large distances. Martinho-Corbishley et al. [44] conducted the first crowdsourcing of comparative soft biometrics, where eight comparative body soft biometric attributes were defined and comparisons between subjects of the BioT database were crowdsourced. In this study, human comparisons were interpreted using RankSVM [47] to generate relative measures that were used to construct biometric signatures. The experimental results revealed the potential of the approach for human identification using comparative body soft biometrics even in challenging situations.

Besides body soft biometrics, clothing attributes have been exploited as soft biometrics for human identification first by Jaha and Nixon in [48], where a set of semantic clothing attributes (21 categorical and 16 comparative) was defined in a way that covers clothing from head to foot. Using subjects from the SGDB database, identification experiments were performed, and the results revealed that clothing soft biometrics could enrich identification performance in fusion with body soft biometrics. Also, the experiments indicated the potential of clothing soft biometrics for re-identification. Later, using a subset of the clothing soft biometrics that has been introduced in [48] and the same dataset, Jaha and Nixon discovered the utilisation of clothing soft biometrics for person re-identification [49]. Also, the study analysed the correlations between multiple viewpoints and chose the soft biometrics that are highly correlated to perform subject retrieval using clothing soft biometrics. Their experiments showed that clothing attributes could be utilised to achieve a better identification performance when multiple viewpoints are involved.

2.2.3 Facial Soft Biometrics

Due to its richness of features and details, the human face is considered as the most informative source of attributes for identification at a short distance, as compared to other soft biometrics such as the body and clothing [13, 11, 39]. Also, human face recognition has demonstrated great robustness for challenging visual conditions such as low resolution and pose variability [50]. Therefore, a great deal of previous research into soft biometrics has focused on facial attributes, either to improve the performance of traditional face recognition systems [51, 52], or to perform identification exclusively based on facial attributes [53, 22, 54, 23, 55].

The earliest exploration of semantic facial attributes emerged in [22], where face verification using the LFW database [56] was examined via attribute classifiers that were trained to recognise the presence or absence of a feature (i.e. categorical soft biometrics). The attributes covered 65 describable visual traits such as eyebrow shape, nose size and eye width. The approach resulted in lowering the error rates by 23.92% compared to the state-of-the-art reported at that time on the LFW database. In [57], the authors studied the use of facial marks such as scars, moles and freckles, as categorical soft biometrics to improve face recognition performance using the FERET database [58]. Their experiments demonstrated that augmenting facial marks (as soft biometrics) with traditional facial hard biometrics could result in improving the recognition accuracy.

A key study in facial soft biometrics is that of Reid and Nixon [59], which was the first study to investigate human face identification using comparative soft biometrics. In their study, 27 comparative facial attributes were defined, and annotations were collected for subjects from the SGDB database. Their experiments have shown that comparative facial soft biometrics results in significantly more accurate descriptions as compared with categorical facial soft biometrics and demonstrated an increase of 25.6% in rank-1 identification rate.

The first study that explored the interaction between automatically extracted soft biometrics and human generated soft biometrics is that of Klare et al. [54], which proposed a method for using categorical facial attributes to perform identification in criminal investigations. Aiming to capture all persistent facial features, a set of 46 facial attributes was defined, and an SVM regressor [60] was trained to estimate the attributes automatically from face images. Identification experiments were performed using the FERET database with all the possible combinations of the probe-gallery (i.e. human vs. machine). Identification using an automatic probe in an automatic gallery resulted in the best recognition accuracy as compared with the other three identification scenarios in which human annotations are used (i.e. for a probe, gallery, or both). In [55], Song et al. proposed a methodology for constructing biometric signatures that are based on the relationships between categorical facial attributes. Their method has demonstrated its effectiveness for face verification using the LFW and the PubFig [22] databases.

The study by Tome et al. [51] considered shape, orientation and size of facial attributes as soft biometrics that can be utilised for forensic face recognition. The study proposed an approach to automatically convert a set of facial landmarks to a set of facial attributes (shape and size), which can be in continuous or discrete values (i.e. categorical). These features were used to generate statistics that aid forensic examiners in carrying morphological comparisons of facial images. Also, they were used to improve the performance of traditional face recognition systems. Using the ATVS [39] and MORPH [61] databases, the experiments revealed that the facial soft biometrics could improve the accuracy of traditional face recognition systems.

Estimation of facial attributes from images for non-biometric objectives has also been an area of interest for several researchers. Liu et al. [62] proposed a deep learning approach for estimating binary facial attributes in the wild for 40 facial attributes, where the attributes were all binary, representing the presence of a visual feature (e.g., bushy eyebrow). The experiments were conducted using the CelebA dataset, which consists of 10000 subjects from the CelebFaces dataset [63], and the LFW database. The proposed approach showed an improvement over existing methods by 8% and 13% for the CelebFaces and LFW databases respectively. Ehrlich et al. [64] explored the estimation of categorical soft biometrics through a multi-task learning approach, in which a model was trained to learn a shared feature representation. They have used the CelebA datasets with its 40 facial attributes [62]. Their approach showed a significant improvement in classification accuracy over the state-of-the-art. A more recent work by Samangouei et al. [52] has investigated the use of categorical facial attributes for active authentication on mobile devices using the MOBIO [65] and AA01 [66] unconstrained mobile datasets. Their approach has highlighted the reliability of binary facial attributes for face verification on mobile devices, and demonstrated that improvement can be gained from fusing the scores of low-level features with the attribute-based approach.

The first study that explored the estimation of relative facial attributes was conducted by Parikh and Grauman in [67], where a RankSVM [47] with similarity constraints was trained to predict the relative strength of a facial attribute from pairs of face images. Using a subset of 800 subjects from the PubFig and 11 relative facial attributes, their approach of using relative attributes showed better accuracy in predicting unseen categories as compared with the categorical (binary) approach. In [68], Sandeep et al. proposed a method for ranking facial attributes from images that was based on learning local parts that are shared among categories rather than learning the global appearance, which was used in [67]. Using a subset of 2000 images from the LFW database, their experiments have revealed that the part-based approach can yield significant improvement in relative attribute prediction accuracy. Recent work by Yu and Grauman [69] addressed the situation in which a facial attribute is perceived with equal strength (i.e. indistinguishable) by human observers. They proposed a learning model that infers the situations in which a facial attribute is indistinguishable. Using a subset of 1000 subjects

from the LFW database [56] and 10 relative facial attributes, their proposed method resulted in a better accuracy of predicting a noticeable difference in attribute pairs.

In general, it can be seen from the literature review presented in this chapter that the use of facial soft biometrics for human face identification has been studied using relatively constrained databases [13, 54, 39], whereas the real identification scenarios involve larger populations with a significant demographic diversity, in addition to more challenging visual conditions of surveillance such as high variability in illumination, facial expressions, resolution and pose. With the exception of the work of Kumar et al. [22], no study has so far addressed the use of facial soft biometrics for subject identification in a large unconstrained database such as the LFW database. In addition, although comparative soft biometrics have shown better identification accuracy as compared to categorical soft biometrics [23, 13], it is still not known whether comparative soft biometrics can scale for large and unconstrained databases (e.g., LFW), as the studies in comparative soft biometric were performed using a small and relatively constrained database [2].

The literature review also shows that a considerable volume of work has explored the use of global soft biometrics (e.g., age, gender and ethnicity) as ancillary information to improve the performance of traditional face recognition systems [40, 39, 51, 70], while more attention needs to be given to semantic facial attributes, as the real identification scenarios involve eyewitness statements (i.e. semantic descriptions) of suspects' faces [13, 12]. Also, concerning the automatic estimation of facial soft biometrics, it can be seen from the literature review that although work has been dedicated to exploring the estimation of categorical facial soft biometrics for human recognition purposes [22, 54, 71, 51], as well as for attribute classification [62, 64], the studies of estimating relative facial attributes have not dealt with human face identification or verification problems. Thus, the research in relative facial attributes was mainly focused on ranking and predicting similarities at attribute level [67, 68, 69]. Therefore, there is a need to explore the automatic estimation of comparative facial soft biometrics from images and discover its impact on the identification and verification performance.

Altogether, the findings from this literature review highlight the importance of exploring human identification using comparative facial soft biometrics in large unconstrained databases. In addition, the findings show the need to investigate the automatic estimation of comparative soft biometrics and the extent to which the semantic gap between human and machine vision can be bridged. Accordingly, this thesis aims to address the inadequacies of previous studies and to explores unconstrained human identification using comparative facial soft biometrics, in addition to the automatic retrieval of comparative facial soft biometrics.

2.3 Performance Evaluation in Biometric Systems

The objective of a biometric system is to achieve recognition of humans based on their physical or behavioural characteristics [72]. These characteristics may include the human face, iris, gait and fingerprint. Typically, a biometric system operates in one of two modes: verification or identification. In the verification mode, the objective is to validate a claimed identity. Thus, the system compares the probe biometric signature against the biometric signatures labelled with the claimed identity in the database, and eventually produce an output as match versus non-match. On the other hand, in the identification mode, the objective is to recognise the identity of an unknown subject by comparing its biometric signature against the biometric signatures of each subject in the database.

Two measures are typically used for performance reporting in biometric systems: (1) the Receiver Operating Characteristic (ROC) curve, which is used to report verification performance; and (2), the Cumulative Match Characteristic (CMC) curve, which is used to report identification performance [73]. The following sections explain performance measurement in each of these tools.

2.3.1 The Receiver Operating Characteristic (ROC)

When performing verification in a biometric system by comparing the biometric signature of the probe with the biometric signature of the claimed subject in the database, four types of outcomes are possible:

- False positive (FP), where the system predicts two biometric signatures as being for the same subject, while they actually belong to different subjects.
- False negative (FN), where the system predicts two biometrics signatures as being for different subjects, while they actually belong to the same subject.
- True positive (TP), where the system successfully matches two biometric signatures that actually belong to the same subject.
- True negative (TN), where the system reports a mismatch between two biometric signatures that actually belong to different subjects.

For a system with n subjects and s samples per subject, the total number of samples will be $n_{total} = n.s$. In the verification performance evaluation process, each sample is compared against all the remaining $n_{total} - 1$ samples in an "all-to-all" manner [73], and two types of scores are generated: the genuine match score, which results from comparing the probe sample with the other samples that related to the same subject; and the imposter match score, which is generated from comparing the probe sample

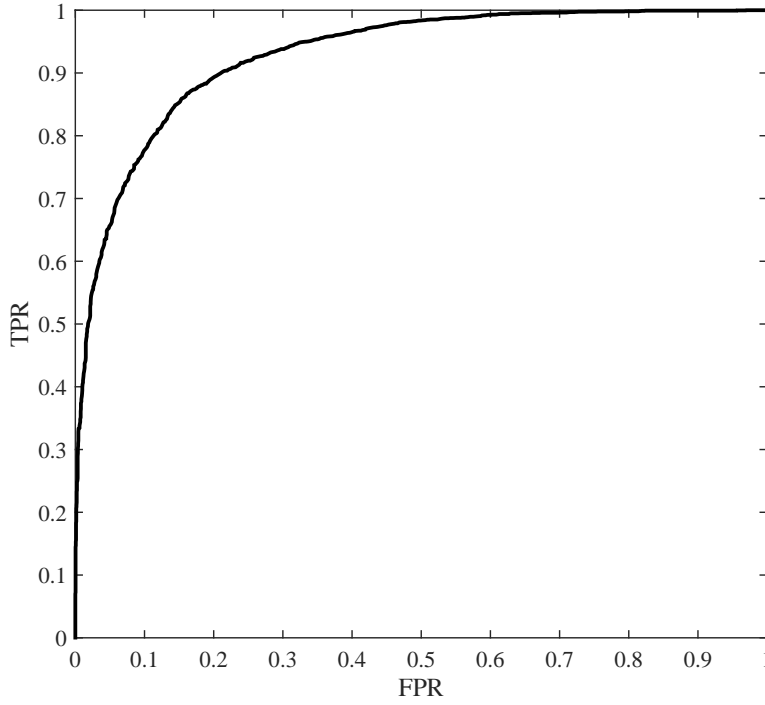


Figure 2.4: Example Receiver Operating Characteristic (ROC) curve.

with the other samples related to different subjects. The genuine and imposter scores distributions, G and I , are generated correspondingly. Then, the True Positive Rate (TPR) and False Positive Rate (FPR) are computed by varying a threshold in the distributions G and I . As a result, the ROC curve is plotted as the relationship between TPR and FPR. Figure 2.4 shows an example ROC curve.

Besides the rates used to represent the ROC curve, several performance metrics can be derived from the four outcomes of the verification process, which can be listed as follows:

- Accuracy, which is the proportion of correct classifications to the total classifications, and which is calculated as follows:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \quad (2.1)$$

- Precision, which can be defined as the proportion of true positives to the total positively predicted samples (i.e. correctly classified samples), and which is found as:

$$P = \frac{TP}{TP + FP} \quad (2.2)$$

- Recall, which can be defined as the proportion of positives samples that are correctly predicted, and which is computed as:

$$R = \frac{TP}{TP + FN} \quad (2.3)$$

- F_1 score, which is a summary measure that embeds both precision and recall. F_1 score is calculated as the harmonic mean of precision, P and recall, R , as follows:

$$F_1 = \frac{2}{\frac{1}{R} + \frac{1}{P}} \quad (2.4)$$

- Equal Error Rate (EER), which is the rate at which the False Positive Rate (FPR) equals the False Negative Rate (FNR). The lesser is the EER; the more accurate is the system.

2.3.2 The Cumulative Match Characteristic (CMC)

As explained at the beginning of this section, the identification process involves comparing a probe biometric signature of an unknown subject with the biometric signature of each subject in the database, while the aim is to find the closest match to the probe in the database. The distance between the probe and each of the subjects in the database is measured using a distance metric such as the Euclidean, Pearson, cosine and city block. The subjects are then sorted with respect to their distance to the probe, from lowest (i.e. the most similar) to highest, and the rank at which a true match exists is found.

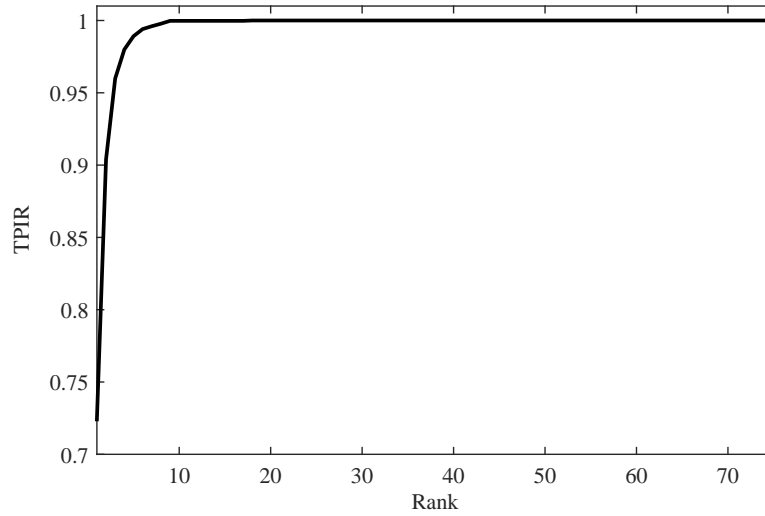


Figure 2.5: Example Cumulative Match Characteristic (CMC) curve.

As a result of identification testing, the True Positive Identification Rate (TPIR) [74], which is the probability of finding a correct match at the top k ranks is computed as follows:

$$TPIR_k = \frac{\omega_k}{n} \quad (2.5)$$

where $1 \leq k \leq n$ is the rank index, n is the total number of subjects, and ω_k is the total number of correct matches that were found at rank k or lower ranks. The CMC curve represents the value of the TPIR at different ranks. In other words, it shows the probability that a correct match is found at a particular rank. Figure 2.5 presents an example CMC curve.

2.4 Summary

This chapter has provided an overview on soft biometrics and highlighted their vital role in criminal investigations. Also, the chapter has proposed a taxonomy, which is based on group and format, for enabling the description and review of the existing work related to soft biometrics. The literature review presented in this chapter has highlighted the implications of face recognition and eyewitness descriptions in psychology on identification using soft biometrics, such as holistic versus component based processing of faces, and the relative importance of facial features. Furthermore, the literature review has summarised the key studies on soft biometric identification and verification with an emphasis on the face, while also highlighting the inadequacies of the existing studies. Then, the literature review has concluded by describing the implications of the knowledge gaps in existing work on the formulation of the objectives of this thesis, which will be noted in the next chapters. Finally, a description was provided for the fundamental performance evaluation tools in biometric systems, which will be utilised throughout this thesis to report the experimental results.

Chapter 3

Comparative Facial Soft Biometrics

The aim of this chapter is to investigate human face identification in constrained databases via comparative facial soft biometrics using a novel set of attributes. The chapter extends the limited knowledge of the single existing work on comparative facial soft biometrics [59] in five aspects. First, it proposes a novel soft biometric set with an emphasis on eyebrows, which have demonstrated an essential role in face recognition by humans, as mentioned in the literature review in chapter 2. Second, it involves the crowdsourcing of comparative labels, which provides a better assessment of the robustness of the approach as compared with the limited annotations in the previous work. Third, it assesses the significance and discriminative power of facial attributes. Fourth, it examines the effect of label levels compression on identification performance. Last, it evaluates the verification potential of comparative soft biometric attributes, which has not been addressed before.

The chapter starts by explaining relative attributes, how they relate to soft biometrics, and how the ranking of attributes is achieved. Then, the chapter describes the basis of selecting the novel comparative facial soft biometrics and introduces the BioT dataset, which has been chosen to explore human identification using the proposed soft biometrics. Furthermore, the chapter outlines the acquisition of comparative labels via crowdsourcing and shows statistical analysis of the attributes' data. Also, the effectiveness of the proposed soft biometric attributes is evaluated through identification and verification experiments using the BioT dataset. Finally, the main findings from the experiments are summarised, and their implications on identification using comparative facial soft biometrics are highlighted.

3.1 Relative Attributes

The literature review that was presented in the previous chapter has shown that soft biometric attributes can be expressed in two different formats: categorical and comparative (or relative) [22, 13]. With the categorical format, an attribute can be expressed either by assigning it to a particular class that represents its strength using an absolute label (e.g., subject *A* has a *wide* mouth), or assigning it to a binary class that represents its presence (e.g., *wearing* spectacles versus *not wearing* spectacles). On the other hand, with the comparative format, an object's attribute is described relative to another object (e.g., subject *A* has a *more wide* mouth than subject *B*).

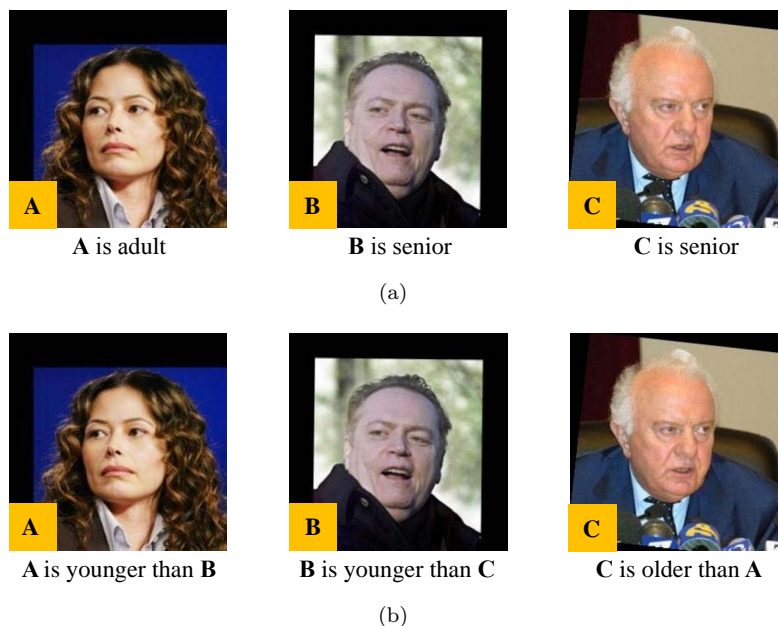


Figure 3.1: Age of samples from the LFW database expressed using: (a) categorical labels; and (b) comparative labels.

As mentioned earlier in chapter 2, the research in soft biometrics has largely been focused on categorical descriptions of facial attributes [54, 12, 70, 39, 51]. However, it has been found that describing attributes in a relative (or comparative) format has several advantages [67]. First, it makes richer semantics for humans (e.g., subject *A* is *more thin* than subject *B*). Second, it enables comparisons with a reference object (e.g., subject *A* is *taller* than Bernie Ecclestone). Third, it improves interactive learning and makes searching based on an attribute more efficient (e.g., search for a *younger* subject). Besides these advantages of relative attributes, the use of comparative soft biometrics for human identification has demonstrated the superiority of relative attributes as compared to the categorical attributes [13, 23, 15]. Figure 3.1 illustrates the descriptive enrichment that can be gained by using relative attributes.

Comparative soft biometrics aim to create a biometric signature for each subject that embeds the individual's physical attributes, and consequently, allows the subject to be

uniquely identified in a database. This biometric signature is a vector that is composed of scores representing the strengths of the soft biometric attributes of the subject. These scores are inferred from pairwise comparisons between the subject and other subjects in the database using a ranking algorithm. One popular ranking algorithm that has been used in prior work on comparative soft biometrics is the Elo rating system [23, 59, 13, 15], which is a well-known algorithm for ranking chess players [75, 46]. Also, RankSVM [47], which is a formulation that learns a ranking function from example pairwise comparisons to infer the scores of attributes, has been used in some studies on comparative soft biometrics [48, 44] to predict biometric signatures. Throughout this thesis, the term "ranking" will be used to implicitly refer to score generation, as ranking involves score inference from pairwise comparisons. More details about ranking are provided in chapter 4.

In this chapter, the Elo rating system is used to generate the biometric signatures, as its applicability and effectiveness for comparative soft biometrics have been already demonstrated [23, 59]. In addition, the Elo rating does not require training as is the case with RankSVM [47], which has been also proposed for ranking comparative soft biometrics [48, 44]. The ranking process in the Elo rating system starts by initialising the scores of all players in a tournament to an initial default score. Then, for a game between two players, A and B , with the initial scores, R_A and R_B , correspondingly, the expected scores, E_A and E_B , are calculated as:

$$E_A = \left[1 + 10^{\frac{(R_B - R_A)}{400}} \right]^{-1} \quad (3.1)$$

$$E_B = \left[1 + 10^{\frac{(R_A - R_B)}{400}} \right]^{-1} \quad (3.2)$$

Subsequently, based on the game outcome (i.e. loss, win, or draw), the new scores, \bar{R}_A and \bar{R}_B , for players A and B respectively, are calculated as follows:

$$\bar{R}_A = R_A + K(S_A - E_A) \quad (3.3)$$

$$\bar{R}_B = R_B + K(S_B - E_B) \quad (3.4)$$

where S_A and S_B are the game scores that are set depending on the game outcome as: 0 for loss, 1 for win, and 0.5 for draw, while K is the score adjustment parameter that determines sensitivity of players' scores update, and thus its value can have a high impact on the ranking outcomes as we will see later in section 3.5.4. The value of K was selected through cross validation by varying it between 21 logarithmically spaced values that range between 2^{-10} and 2^{10} .

The Elo rating system can be used for ranking facial soft biometrics in a similar way to chess players ranking. Thus, by considering the subjects of a dataset as players in a chess tournament, and assuming that a comparison between two subjects, A and B , for a particular facial attribute, X , is a game between two players. Subsequently, this comparison can result in one of three possible outcomes for each of the two subjects as: "*Less X* ", "*More X* ", or "*Same X* ", for example: "Subject A has a *more thick* eyebrow than subject B ". Accordingly, the scores of the compared subjects are updated based on the comparison outcome using Equations 3.1 to 3.4, and the ranking of the subjects with respect to the evaluated attribute is updated correspondingly.

3.2 Defining Facial Attributes and Comparative Labels

The human face has the richest information content for identification at short distances as compared to the other groups of attributes such as body and clothing [13, 76, 13, 11, 39]. Furthermore, global soft biometrics such as age and gender can be inferred more accurately from a human face as compared to body or gait [11, 77]. However, a facial attribute is required to be understandable, memorable, and describable, to contribute efficiently as a soft biometric attribute in semantically describing and distinguishing people for identification. These aspects have governed the selection of facial features to define the soft biometric set used in this thesis, which covers the major facial components (i.e. eyes, eyebrows, nose and mouth), with an emphasis on eyebrows due to their pivotal role in human face recognition [34, 26].

The soft biometric set (the BioT attribute set hereafter), which is studied and analysed throughout this chapter, consists of 24 comparative attributes (20 facial and 4 global) as shown in Table 3.1. Each attribute is associated with a comparative label that is based on five-point bipolar scale, which ranges from -2 to 2, where for an attribute X , -2 is associated with the "*Much less X* " label, and 2 is associated with the "*Much more X* " label. For a comparison between subjects A and B , the value of the comparative label is normalised and set as the comparison score of subject A , S_A , while the comparison score of subject B , S_B , is set as $1 - S_A$. These values are used for estimating the scores of the two subjects based on Equations 3.3 and 3.4, and deducing their ranks correspondingly.

3.3 Dataset and Label Acquisition via Crowdsourcing

The comparative facial soft biometrics proposed in the previous section were investigated with a dataset from the University of Southampton's Multi-Biometric Tunnel (BioT) [2], which is an environment that was designed to study relatively constrained human identification, and to collect non-contact biometrics (e.g., gait, face and body). To a large degree, the tunnel simulates surveillance in indoor environments of high throughputs

No.	Attribute	Labels
1	Chin height	[Much Smaller, More Small, Same, More Large, Much Larger]
2	Cheek size	[Much Smaller, More Small, Same, More Large, Much Larger]
3	Cheek shape	[Much Flatter, More Flat, Same, More Prominent, Much Prominent]
4	Eyebrow length	[Much Shorter, More Short, Same, More Long, Much Longer]
5	Eyebrow thickness	[Much Thinner, More Thin, Same, More Thick, Much Thicker]
6	Eye-to-eyebrow distance	[Much Smaller, More Small, Same, More Large, Much Larger]
7	Eye shape	[Much Tilted Inward, More Tilted Inward, Same, More Tilted Outward, Much Tilted Outward]
8	Eye size	[Much Smaller, More Small, Same, More Large, Much Larger]
9	Face length	[Much Shorter, More Short, Same, More Long, Much Longer]
10	Face width	[Much Narrower, More Narrow, Same, More Wide, Much Wider]
11	Face shape	[Much Ovoid, More Oval, Same, More Round, Much Rounder]
12	Forehead hair	[Much Less Hair, Less Hair, Same, More Hair, Much More Hair]
13	Inter eyebrow distance	[Much Closer, More Close, Same, More Wide, Much Wider]
14	Inter pupil distance	[Much Closer, More Close, Same, More Wide, Much Wider]
15	Jaw shape	[Much Chiseler, More Chiseled, Same, More Lantern-Shaped, Much Lantern-Shaped]
16	Lips thickness	[Much Thinner, More Thin, Same, More Thick, Much Thicker]
17	Mouth width	[Much Narrower, More Narrow, Same, More Wide, Much Wider]
18	Nose length	[Much Shorter, More Short, Same, More Long, Much Longer]
19	Nose-to-mouth	[Much Shorter, More Short, Same, More Long, Much Longer]
20	Nose width	[Much Narrower, More Narrow, Same, More Wide, Much Wider]
21	Age	[Much Younger, More Young, Same, More Old, Much Older]
22	Figure	[Much Thinner, More Thin, Same, More Thick, Much Thicker]
23	Gender	[Much Feminine, More Feminine, Same, More Masculine, Much Masculine]
24	Skin colour	[Much Lighter, More Light, Same, More Dark, Much Darker]

Table 3.1: Soft biometric attributes.

such as the walkways of shopping malls, airports and enterprises. The tunnel consists of multiple cameras to capture images of subjects as they walk through it. Figure 3.2a shows the placement of cameras in the tunnel.

The dataset used to explore the comparative facial soft biometrics proposed in this chapter (the BioT dataset hereafter) was constructed by extracting front face images for 100 different subjects from the BioT database, and locating the subjects' faces using the Viola-Jones face detector [3] as shown in Figure 3.2b. Then, similarly to the approach used in [78], all face images were in scale by translating the inter-pupil distance to 70 pixels. This normalisation is meant to ensure consistent comparisons between the subjects. Samples of the processed face images from the BioT dataset are shown in Figure 3.3. The subjects of the BioT dataset are mostly university staff or students, and the dataset is gender-balanced. Moreover, most of the males are Caucasians, and most of the females are Chinese.

Crowdsourcing is an effective way of collecting labels for a dataset from human annotators, as it involves the contribution of annotators with a great diversity in abilities, backgrounds, cultures and perceptions, which enriches the collected labels [44]. Accordingly, the collection of comparative labels for the BioT dataset, a job that consists of 4950 comparisons (i.e. the number of all possible pairs for a dataset of 100 subjects) was crowdsourced, whereby each comparison between a pair of subjects included the attributes of the BioT attribute set that are listed in Table 3.1. Figure 3.4 shows an example crowdsourced comparison.

The job was launched using the CrowdFlower* platform and resulted in the collection

*<https://www.crowdfunder.com>

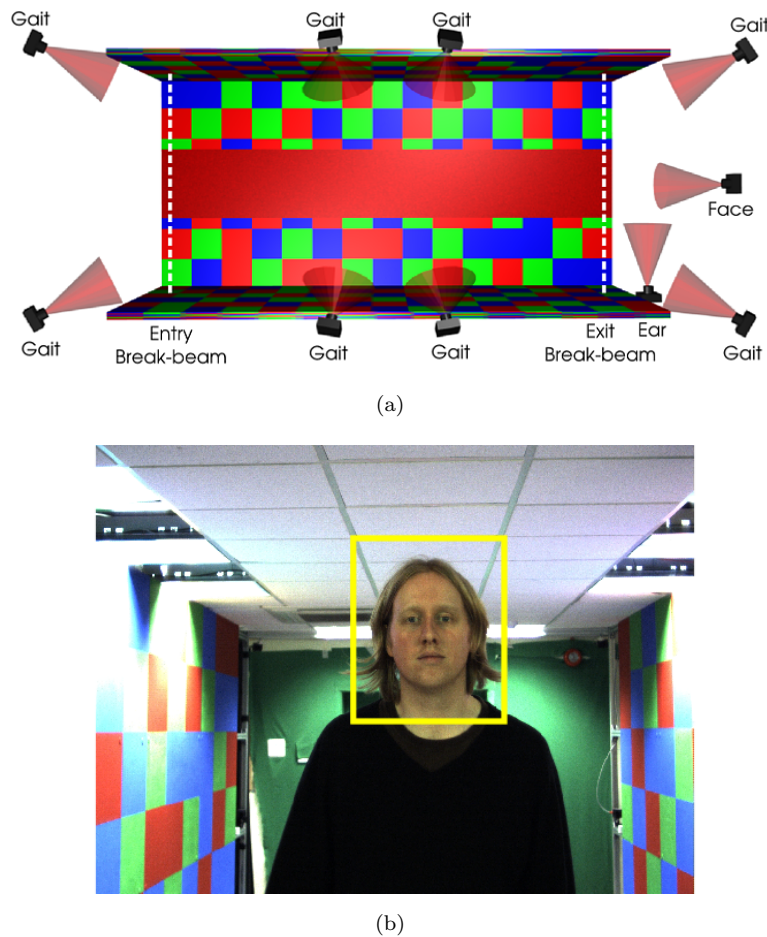


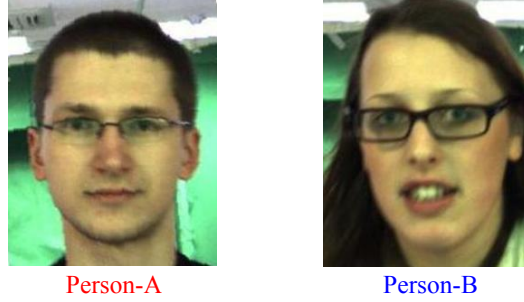
Figure 3.2: (a) The layout of the University of Southampton's Biometric Tunnel (BioT) [2]; (b) Example front image captured by the face camera in the tunnel, with the subject's face detected using the Viola-Jones detector [3].



Figure 3.3: Sample face images from the BioT dataset where the effect of some visual conditions can be noted: (a) occlusion of eyebrows; (b) pose variability; and (c) video motion blur.

of 37968 attribute comparisons, which were used to infer more comparisons. Inferring a relation between two subjects A and B can be achieved according to the rules listed in Table 3.2, whenever relations between each of them and a common target subject G exists. More insights on the crowdsourcing of the BioT dataset are provided in Table 3.3.

Figure 3.5 provides more details about the collected comparative labels through the crowdsourcing of the BioT dataset.



The eyebrow of **Person-A** relative to that of **Person-B** is:

- ☐ Much Thinner
- ☐ More Thin
- ☐ Same
- ☐ More Thick
- ☐ Much Thicker
- ☐ Don't know

Figure 3.4: Example question from the crowdsourcing job launched to collect comparative labels for the BioT dataset.

(A,G)	(B,G)	$\text{inf}(A,B)$
=	=	=
>	<	>
<	>	<
>	=	>
<	=	<
>	>	N/A
<	<	N/A

Table 3.2: Relation inference rules.

	Collected	Inferred	Total
Attribute comparisons	37968	78969	116937
Subjects' comparisons	3522	1428	4950
Average number of annotations per comparison	2.22	N/A	N/A
Number of annotators	3073	N/A	N/A

Table 3.3: Crowdsourcing job statistics for the BioT dataset.

As can be seen from Figure 3.5, most of the attributes have the tendency to distribute uniformly among the "*Less*", "*Same*" and "*More*" labels, with the exception to *eyebrow thickness*, which might be affected by the application of makeup by females, who comprise 50% of the dataset. Also, Figure 3.5 shows that most of the annotations are within the middle three labels (i.e. "*Less*", "*Same*" and "*More*"), while the "*Much less*" and "*Much more*" are infrequent. Finally, it can be noted that the all the attributes are

clear and understandable by the annotators; thus, the highest percentage of the "Don't know" label is less than 5%.

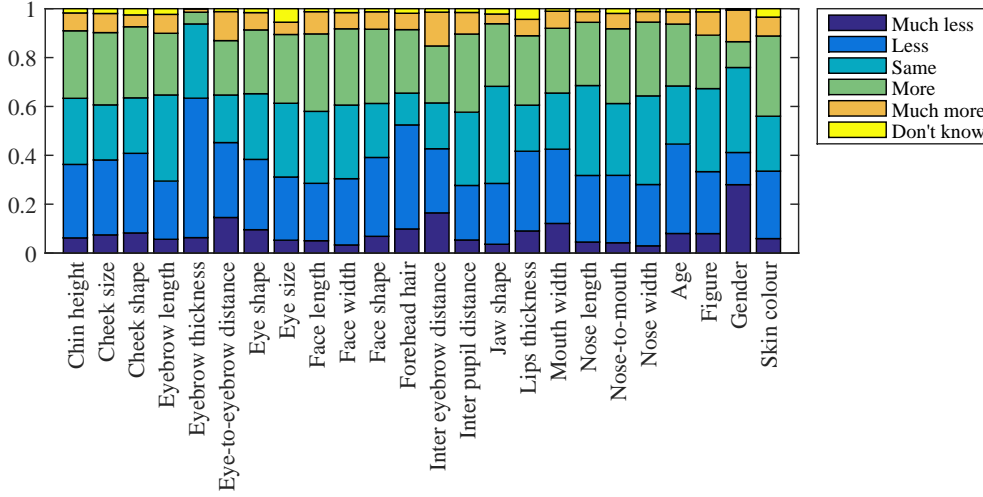


Figure 3.5: Distribution of the collected comparative labels for the BioT dataset.

3.4 Dataset Analysis

The previous section provided insights on the acquisition of comparative labels through the crowdsourcing of subjects' comparisons. Also, it presented a statistical summary on the collected labels for the BioT dataset. However, to better summarise the distribution of the attributes, and to assess their significance in distinguishing people, the labels should be analysed in the form of scores; this is because it is the representation used in constructing the biometric signatures, as explained in section 3.1, and thus performing attribute analysis using the scores provides better insights on their significance and contribution in identification. Therefore, all the analysis in this section was performed with the scores of the attributes, which were generated from the crowdsourced comparisons using the Elo rating system.

3.4.1 Dataset Distribution

The distribution of attribute data for the BioT dataset is summarised using a box plot as shown in Figure 3.6, where it can be seen that most of the attributes have no or a negligible number of outliers. Furthermore, Figure 3.7 shows the distribution of each attribute along with the p -value (probability value) resulting from the Anderson-Darling test for normality [79], where p -value < 0.05 implies that the attribute data is not normally distributed, and p -value ≥ 0.05 indicates that the data follows the normal distribution. From Figure 3.7, it can be seen that most of the attributes are following the normal distribution. The outcomes of attribute ranking using the Elo rating system are

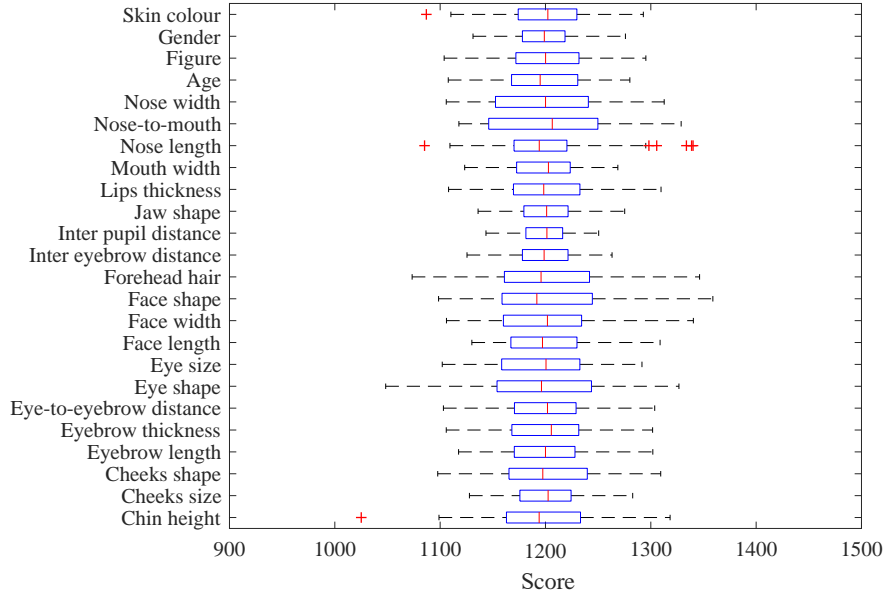


Figure 3.6: Box plot for scores of the BioT attributes.

demonstrated in Figure 3.8, which shows the least and top-ranked subjects for selected attributes.

3.4.2 Attribute Significance

The evaluation of attribute significance is important to understand its strength as a semantic descriptor and to explore its contribution in distinguishing humans and identifying them correspondingly. Two types of analysis are used to assess attribute significance: discriminative power, and semantic stability.

3.4.2.1 Attribute Discriminative Power

Discriminative power analysis allows attributes to be ranked with respect to their capabilities in distinguishing subjects and contribution in the identification accordingly. Three different methods are used in this chapter to determine the attributes' discriminative power: Analysis of Variance (ANOVA), entropy and mutual information.

1. Analysis of Variance (ANOVA)

The one-way ANOVA test [80] is based on the F statistic, which is variance between groups divided by variance within groups. The F statistic is calculated as follows:

$$F = \frac{\sum_{i=1}^k [(\bar{X}_i - \bar{X}_G)/(k-1)]}{\sum_{ij} [(X_{ij} - \bar{X}_i)/(n-k)]} \quad (3.5)$$

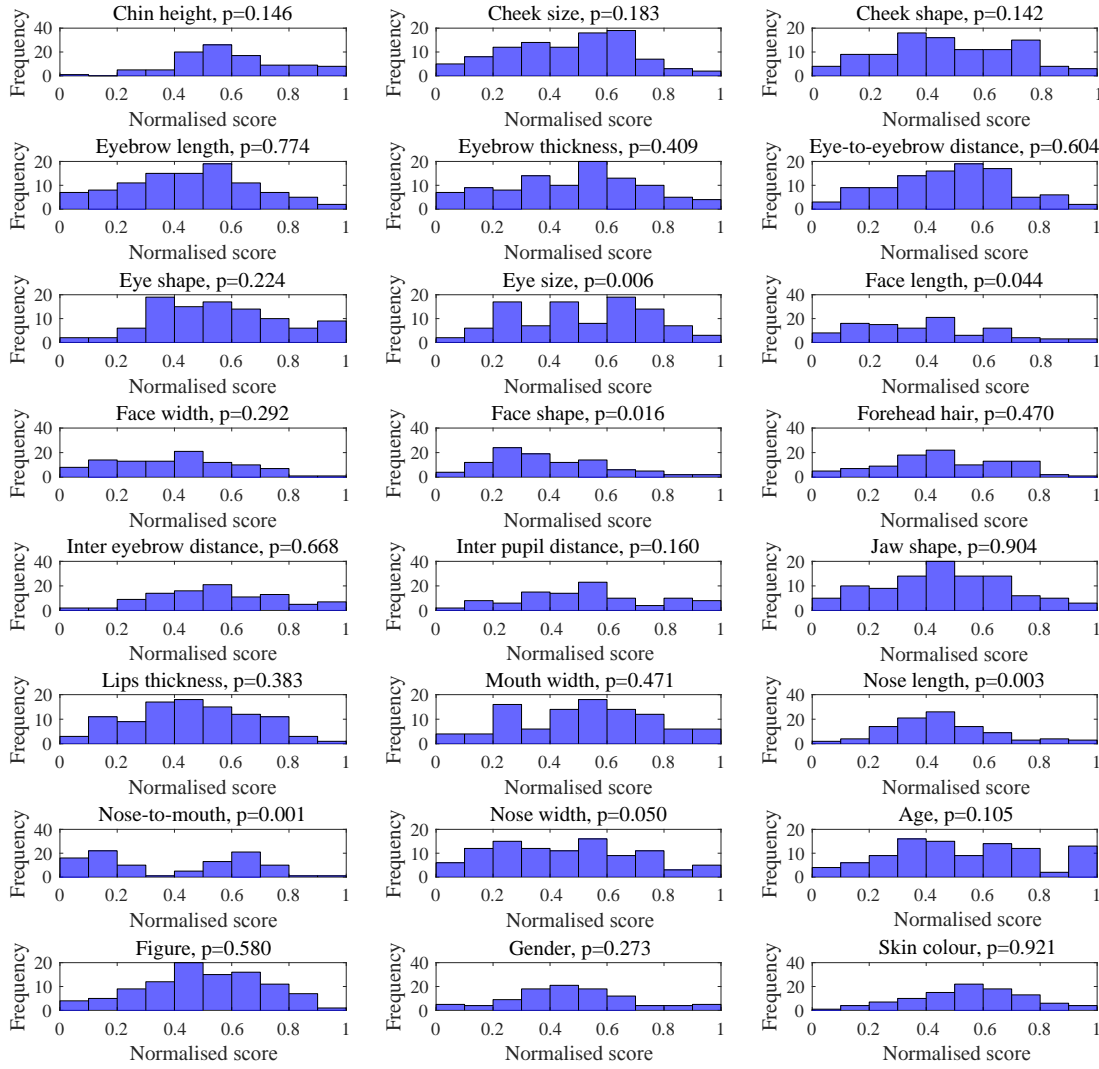


Figure 3.7: Distribution of the BioT attributes.

where k is the number of groups (i.e. samples), n is the total number of items in the k groups, \bar{X}_i is the mean of the i^{th} group, \bar{X}_G is the grand mean of all the items in the k groups, X_{ij} is the value of the j^{th} item in the i^{th} group, and n_i is the number of items in the i^{th} group. The one-way ANOVA can be used to evaluate attributes' discriminative power by measuring the variances between and within groups for each attribute and finding the F statistic correspondingly. To calculate the F statistic for an attribute, two samples of the subjects' scores are created by using the Elo rating system with two mutually exclusive subsets of the comparative labels. The higher is the value of the F statistic, the greater is the discriminative power of the attribute.

2. Entropy

Entropy [54] is an information theoretic measure that represents the average amount of information contained in a random variable, X , and it is calculated as follows:

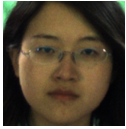

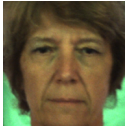

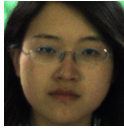




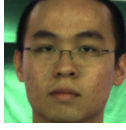

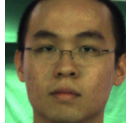



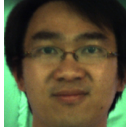
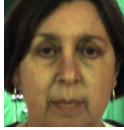







Cheek size	Eyebrow length	Eye-to-eyebrow dist	Eye size	Face width	Forehead hair
					
					
Inter pupil dist	Lips thickness	Nose length	Nose width	Figure	Skin colour
					
					

Figure 3.8: Ranking of selected attributes for the BioT dataset using the Elo rating system. For each attribute, the top image represents the top ranked subject, while the bottom image shows the least ranked subject. The comparative labels that correspond to each attribute are listed in Table 3.1.

$$H(X) = - \sum_{x \in X} p(x) \log_2 [p(x)] \quad (3.6)$$

where X is a discrete random variable, and $p(x)$ is the probability distribution function of X . In the context of soft biometric attributes, it is assumed that the ranks of the subjects with respect to an attribute, which are deduced from their scores, are discrete random variables. Accordingly, entropy can be used to measure the information contained in each attribute, providing us with an indicator of the impact of each attribute in discriminating subjects. The higher is the entropy of an attribute, the more information it contains, thus the higher its discriminative power.

3. Mutual Information

Mutual information [81] is another information theoretic measure that reveals the amount of information gained about a random variable, X , by observing another random variable, Y , (or vice versa). Mutual information is computed as follows:

No.	Attribute	ANOVA (F-statistic)	Entropy	Mutual information
1	Chin height	2.5707	3.9290	5.6986
2	Cheek size	2.4076	3.8260	5.5274
3	Cheek shape	2.5451	3.8313	5.5274
4	Eyebrow length	4.3502	4.0549	5.7939
5	Eyebrow thickness	3.4918	3.8920	5.5686
6	Eye-to-eyebrow distance	3.1519	3.9897	5.5910
7	Eye shape	2.6672	3.7869	5.5198
8	Eye size	3.6522	3.9145	5.5949
9	Face length	3.6416	3.9499	5.6684
10	Face width	3.6522	3.8688	5.6149
11	Face shape	2.2323	3.8313	5.4633
12	Forehead hair	2.8198	3.8599	5.5815
13	Inter eyebrow distance	3.9248	3.9696	5.7137
14	Inter pupil distance	4.3247	3.9588	5.7559
15	Jaw shape	2.3650	3.8260	5.5198
16	Lips thickness	3.2130	4.0020	5.7500
17	Mouth width	2.9400	3.8781	5.7149
18	Nose length	3.2552	3.8715	5.5854
19	Nose-to-mouth	3.7004	3.9856	5.8186
20	Nose width	3.3566	4.0160	5.7270
21	Age	6.2338	3.9613	5.7737
22	Figure	3.4776	3.8754	5.7113
23	Gender	6.4123	3.9604	5.6474
24	Skin colour	3.9836	4.0331	5.8500

Table 3.4: Discriminative power of the BioT attributes.

$$I(X, Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \log_2 \left[\frac{p(x, y)}{p(x)p(y)} \right] \quad (3.7)$$

where X and Y are discrete random variables, while $p(x)$ and $p(y)$ are probability distribution functions for the random variables X and Y respectively, and $p(x, y)$ is the joint probability distribution function. In the context of soft biometric attributes, X represents the attribute ranking, and Y represents the subjects' labels. The higher is the mutual information between the attribute ranking and the subject labels; the higher is the assumed discriminative power of the attribute.

The results of discriminative power analysis for the BioT attributes are listed in Table 3.4, they were used to rank the attributes with respect to their discriminative power, and accordingly, generate a rank score for each attribute that reflects the attribute significance, and enable comparing the outcomes of the three method (i.e. ANOVA, entropy, and mutual information). It is important not to confuse the rank score in the discriminative power analysis, which reflects attribute significance, with the ranking and scoring of soft biometric attributes from comparative labels, which aims to create biometric signatures for human identification. As there is a total of 24 attributes, the rank score ranges between 1 and 24, such that the attribute with the highest discriminative power

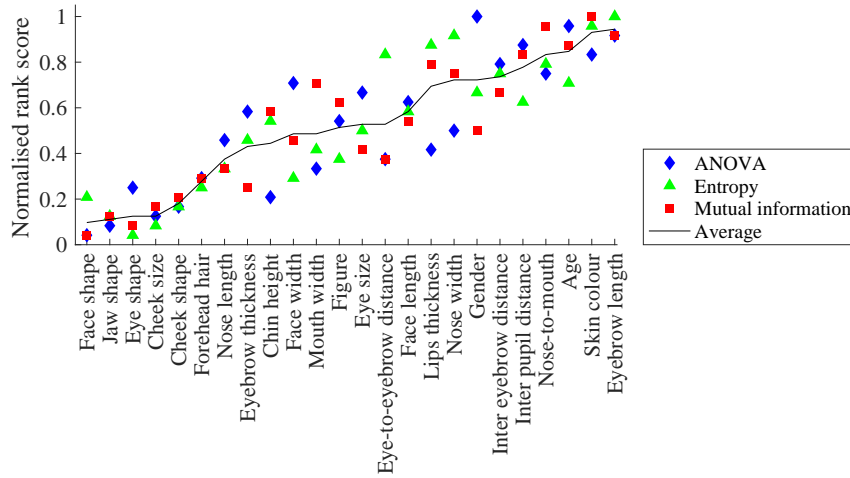


Figure 3.9: Discriminative power of the BioT attributes.

was assigned a rank score of 24, while the attribute with the lowest discriminative power was assigned a rank score of 1. The rank scores were scaled (i.e. normalised) between 0 and 1. Figure 3.9 shows the normalised rank score for each attribute based on each of the three methods. Also, it shows the average normalised rank score of the three methods. The results in Figure 3.9 reveal several interesting findings. Firstly, the shape-based attributes, which are: *eye shape*, *face shape*, *jaw shape* and *cheek shape*, have significantly low discriminative power. Secondly, the global soft biometrics (e.g., *age* and *skin colour*) have relatively high discriminative power, which supports Samangooei's et al. findings in [12]. Thirdly, the nose attributes have notable discriminative power, which might be due to the expression invariance capabilities of nose [82]. Finally, the eye and eyebrow region attributes show significant discriminative power, especially *eyebrow length*, and hence, this further supports the emphasis made on eyebrows in this thesis based on their role in human face recognition [34, 26]. The analysis in Figure 3.9 reveals the consistency among the evaluations of the different discriminative power assessment methods (i.e. ANOVA, entropy and mutual information). Thus, there is more tendency for agreement among the three evaluations when used with the BioT dataset, and this can be attributed to the relatively small size of the BioT dataset, which makes the differences among the three methods marginal.

3.4.2.2 Attribute Semantic Stability

Semantic stability can be defined as the consistency of an attribute ranking among different annotators, which is crucial for assessing the attribute effectiveness and robustness. The semantic stability of the attributes was evaluated by creating two different galleries, each of which consisted of the biometric signatures of all the subjects in the dataset, where each biometric signature was composed of the scores of the 24 soft biometric attributes (listed in Tables 3.1). The scores in each gallery were inferred using the

Elo rating system based on two mutually exclusive subsets of comparative labels, which represent two different groups of annotators. Then, the semantic stability was measured for each attribute based on the two galleries as the Pearson's correlation between the subjects' scores. The results of the semantic stability analysis are shown in Figure 3.10.

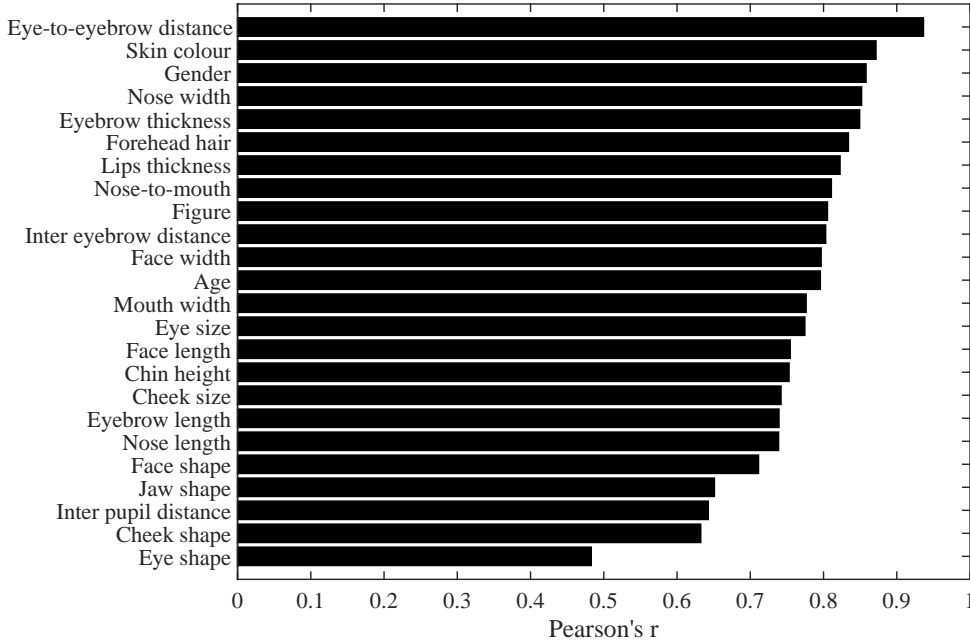


Figure 3.10: Semantic stability of the BioT attributes.

Figure 3.10 shows that the shape-based attributes such as *eye shape* and *cheek shape* have the lowest stability, which agrees with the findings of the discriminative power analysis in section 3.4.2.1 about the low discriminative capabilities of the shape-based attributes. Also, *inter pupil distance* has low stability, perhaps because all the facial images in the BioT dataset have been normalised to a fixed *inter pupil distance*, which might cause the annotators to fail to recognise the difference between subjects. On the other hand, some eyebrow attributes (*eye-to-eyebrow distance* and *eyebrow thickness*), in addition to global soft biometrics (e.g., *skin colour* and *gender*), show the highest stability, which supports the findings of Sadr et al. in [34] and Sinha et al. in [26] on the role of eyebrows in human face recognition. In addition, it agrees with results of Samangooei et al. in [12] regarding the significance of global soft biometrics. In general, all the p -values resulting from the semantic stability analysis are very close to zero, which indicates that all the attributes are statistically significant, regardless of the strength of the correlations (i.e. semantic stability) between the two galleries.

3.4.3 Attribute Correlations

Assessing the correlations between the attributes is useful to discover the significant relationships between them, which can be used to predict the strength of an attribute

based on the other, as in the case of partial occlusion of a face. In addition, analysing correlations can reveal the extent of independence of each attribute, which implies its unique informative value in identification. The correlation analysis in this thesis was conducted using the Pearson's correlation coefficient, r , with the scores estimated using the Elo rating system, whereas $r = 0$ represents no correlation, while $|r| = 1$ represents a perfect correlation. The correlations between the BioT attributes are shown in Figure 3.11.

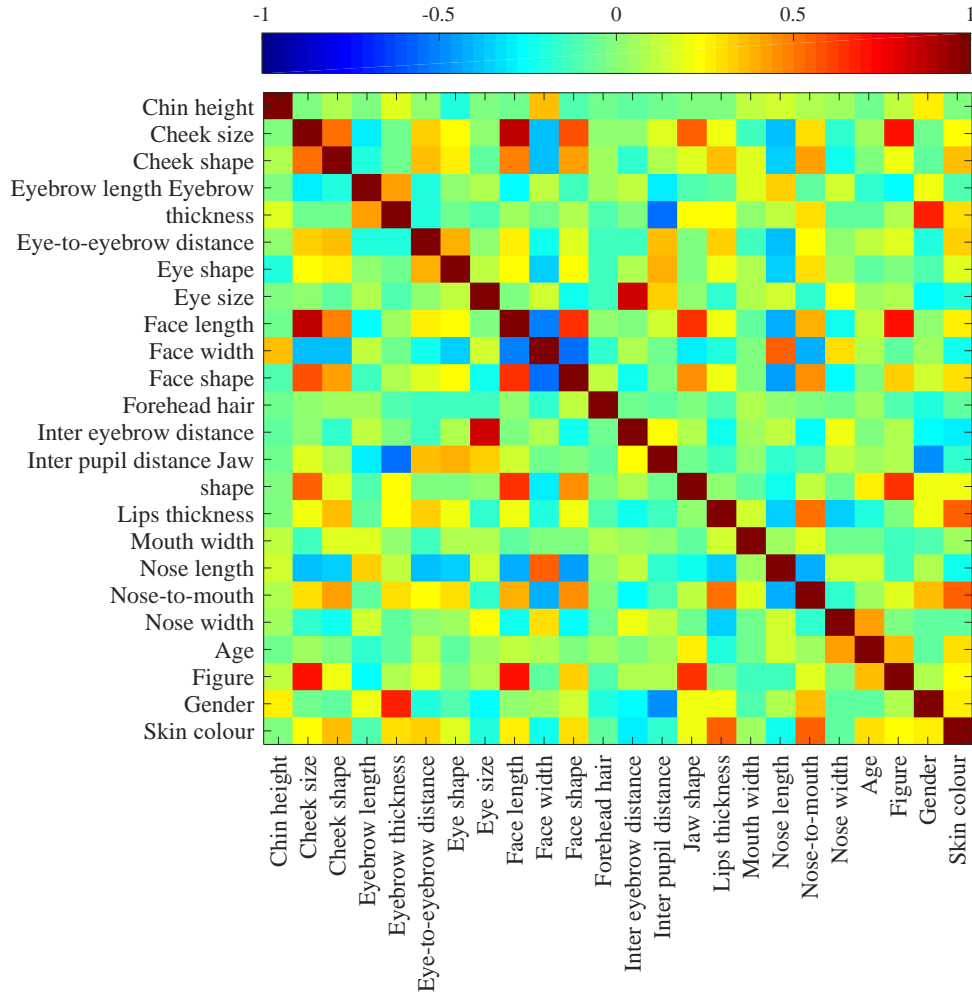


Figure 3.11: Correlations between the BioT attributes using Pearson's r .

The correlations map in Figure 3.11 shows a number of significant correlations between the BioT attributes, from which, the most interesting are those between facial and global attributes such as: (1) the strong positive correlation between *gender* and *eyebrow thickness*, which implies the association between femininity and thin eyebrows; (2) the negative correlation between *gender* and *inter pupil distance*; and (3) the positive correlation between *skin colour* and *lip thickness*. In general, most of the BioT attributes have weak or no correlations between them, which indicate their independence and the potential contribution of each attribute in the identification performance.

3.5 Experiments

3.5.1 Identification Using Facial Comparisons

The purpose of the identification experiment is to assess the effectiveness of the proposed attributes (listed in Table 3.1) for identification in small and relatively constrained databases using the BioT dataset. The experiment simulates a realistic scenario that aims to retrieve the identity of an unknown subject (or probe) from a soft biometric database using verbal descriptions for the probe (i.e. eyewitness statement). Figure 3.12 illustrates the identification process. The experiment followed a Leave One Out Cross Validation (LOOCV), in which each of the 100 subjects in the dataset was chosen as a probe, and comparisons between the probe and c other randomly selected subjects were removed from the dataset, leaving the remaining $m - c$ for gallery construction, where m is the number of all pairwise comparisons in the dataset. The c comparisons were used with the Elo rating system to generate a score for the probe with respect to each attribute, which construct in aggregation the biometric signature of the probe, while the remaining $m - c$ were used to construct the biometric signatures of the 100 subjects in the gallery.

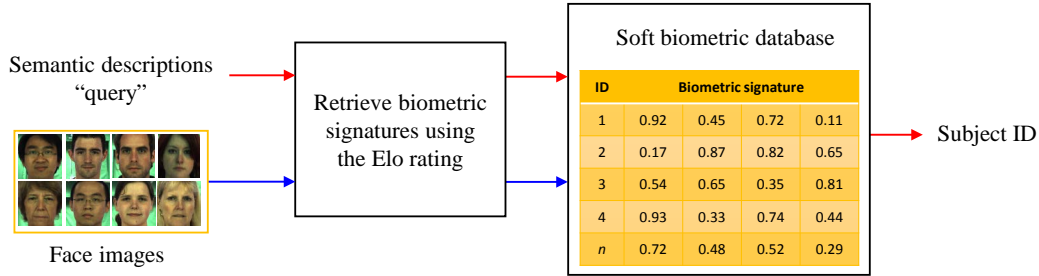


Figure 3.12: Illustration for identification using comparative facial soft biometrics.

The identification of unknown subjects was performed by computing the Euclidean distance, d_E , between the biometric signature of the probe and the biometric signature of each subject in the gallery as follows:

$$d_E = \sqrt{\sum_{i=1}^T (X(i) - Y(i))^2} \quad (3.8)$$

where X is a vector that represents the biometric signature of the probe, Y is a vector that represents the biometric signature of the subject in the gallery that is compared with the probe, and $T = 24$ is the number of soft biometric attributes composing the biometric signatures. The subjects were sorted in ascending order according to their distances with the probe and the rank of the correct match was used to report the

identification performance via a CMC curve. The experiment was repeated 30 times and the arithmetic mean of the ranks resulting for each subject over the 30 trials was considered as the experiment outcome. Figure 3.13 shows a CMC curve representing the identification performance of the BioT attributes using different numbers of comparisons.

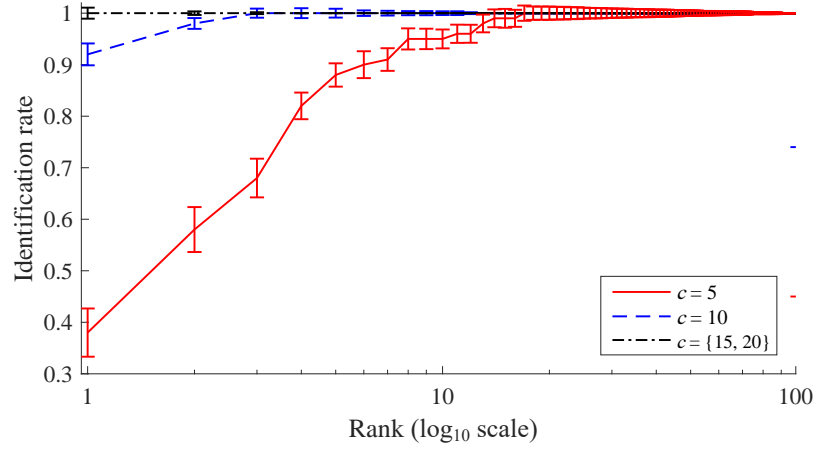


Figure 3.13: Identification performance on the BioT dataset using $c = \{5, 10, 15, 20\}$ comparisons.

As Figure 3.13 shows, a rank-1 identification rate of 93% is achieved using 10 pairwise comparisons only, and a correct match is always guaranteed at rank-1 when 15 or more comparisons are used for the identification. These results exceed the performance achieved in Reid and Nixon study [59], which is the only existing work on comparative facial soft biometrics, in three respects: (1) whereas it required 20 comparisons to achieve a rank-1 recognition rate of 100% in [59], the attributes proposed in this chapter required only 15 comparisons to achieve the same identification rate; (2) this experiment uses fewer attributes than what have been used in [59] (24 as compared to 27 in [59]); and (3) the labelling in [59] is based on frontal and side views of the faces, while the labels used in this experiment were collected only for the frontal view of the faces, which limits the information available to annotators; hence, the dataset used in this experiment is considered to be a more challenging dataset as compared with the SGDB dataset [43] that was used in [59]. Overall, the outcomes of this experiment demonstrate the performance gains that result from the proposed soft biometric set, in addition to the effectiveness of crowdsourcing as a channel for collecting comparative facial labels.

3.5.2 Label Compression

Section 3.2 stated that the comparative labels associated with the attributes (listed in Table 3.1) were defined based on a 5-point bipolar scale that ranges from -2 to 2, where -2 is associated with the "Much less" label and 2 is associated with the "Much more" label. Interestingly, the identification experiments have shown that compressing the levels of the comparative labels from 5 to 3 can improve the identification performance.

Figure 3.14 shows the effect of label compression with different number of subject comparisons.

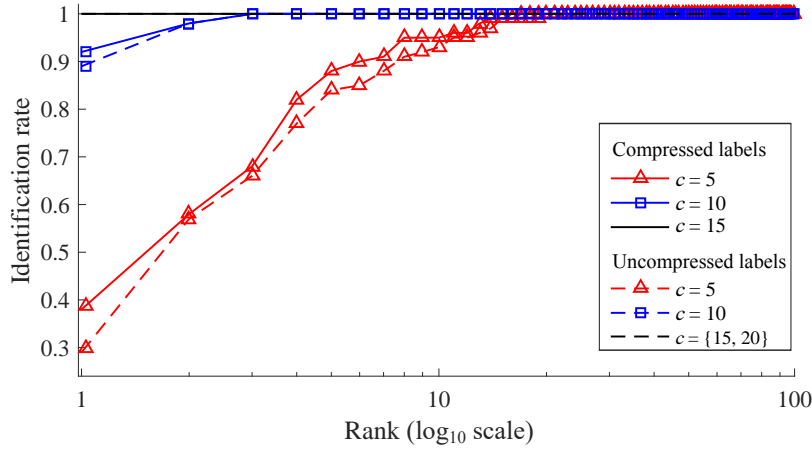


Figure 3.14: Effect of compressing comparative labels on identification performance using $c = \{5, 10, 15, 20\}$ comparisons.

It can be observed from Figure 3.14 that the effect of label compression is most significant with a smaller number of subject comparisons. The performance improvement resulting from label compression can be attributed to the insignificance of levels -2 and 2 in distinguishing the differences between subjects, and this can be noted in Figure 3.4, where the "Much more" and "Much less" labels were used significantly less than the middle three labels (i.e. "Less", "Same" and "More"). Also, this gain can be attributed to the effect of introducing categorical descriptions to the comparative basis, which is antithetical to the nature of comparative descriptions. In summary, reducing the levels of comparative labels improves performance considerably.

3.5.3 Verification Using Facial Comparisons

So far, comparative soft biometrics have been explored with an emphasis on human identification, in which a database of subjects is searched using a semantic description for an unknown subject (probe) to find the closest match [12, 59, 23, 44]. However, it will be interesting to explore the verification performance of comparative soft biometrics; thus it can reveal the level of agreement of two different semantic descriptions (i.e. eyewitness statements) for the same subject, and thus it indicates the robustness of soft biometrics. Therefore, a verification experiment that is based on two galleries was designed, where each gallery in the experiment consisted of the biometrics signature of the 100 subjects in the BioT dataset. The biometric signatures were generated using comparisons between each subject and c other randomly selected subjects from the dataset. It is important to mention that the comparisons used to construct the biometric signatures of the subjects in the two galleries are mutually exclusive.

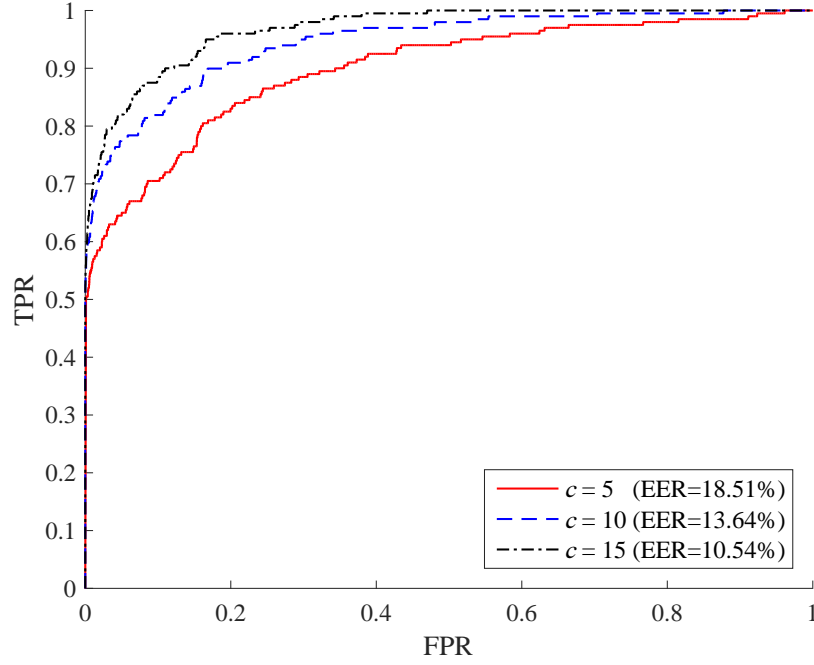


Figure 3.15: Verification performance of the BioT dataset using $c = \{5, 10, 15\}$ comparisons.

The verification performance of the BioT attributes is reported through the ROC curve shown in Figure 3.15. The ROC curve shows that the verification performance improves as the number of subject comparisons increases. Using the BioT database, Tome et al. [39] achieved an EER of 13.54% by utilising an automatic face detection and recognition system, with the fusion of 23 categorical soft biometrics (13 body, 3 global and 7 head); whereas, the comparative facial soft biometrics proposed in this chapter resulted (without fusion with face hard biometrics) in a slightly higher EER of 13.64% using 10 subject comparisons, which decreases to 10.54% outperforming the approach in [39] with only five more subject comparisons. Overall, the results of this experiment demonstrate the verification capability of the proposed comparative soft biometrics and show that the attributes can outperform the performance of automatic face recognition when fused with categorical soft biometrics on the same database.

3.5.4 Elo's K and Identification Performance

As mentioned in section 3.1, ranking using the Elo rating system involves a score adjustment parameter, K , that determines the sensitivity of score update as in Equations 3.3 and 3.4. The literature has not specified a default value for the parameter K , and this requires tuning it to select its best value through cross validation. In the experiments presented in this chapter, the value of K was tuned within a set of 21 logarithmically spaced values (i.e. steps in powers of 2) that range from 2^{-10} to 2^{10} , and the mean identification performance was determined with each value of K as shown in Figure 3.16,

which demonstrates how the impact of the parameter K on the identification performance can be significant. For example, the mean rank of the retrieved match improves by 61.53% as a result of a single step change in the K value from 4 to 8, as can be observed in Figure 3.16.

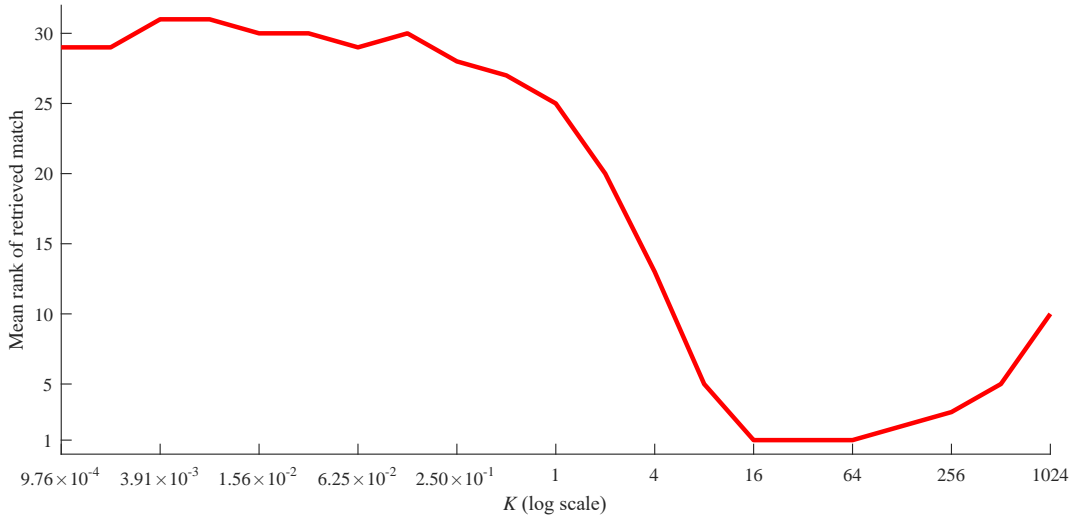


Figure 3.16: Effect of the score adjustment parameter, K , on identification performance (x-axis labels are displayed in steps of powers of 4).

The process of tuning the parameter K can be extremely costly when using the Elo rating system with large datasets, as the pairwise nature of samples in identification using comparative soft biometrics imposes an exponential growth in space and time complexities with the dataset size. The overhead of parameter tuning during the ranking of attributes will be avoided if the ranking algorithm is fully unsupervised. This motivated the proposal of a novel parameterless and fully unsupervised ranking algorithm (MIURank), which is presented in chapter 4, as an effective and efficient way of ranking attributes.

3.6 Summary

This chapter investigated human identification using comparative facial soft biometrics in small constrained databases and extends our knowledge of relative facial attributes. A novel set of comparative facial soft biometrics was proposed with an emphasis on eyebrows due to their important role in human face recognition. Also, the acquisition of comparative labels via crowdsourcing was described, and its outcomes were outlined and discussed. Furthermore, the ranking of attributes, which composes biometric signatures, was explained. The attribute analysis presented in this chapter has revealed the statistical significance of all the proposed comparative facial soft biometrics and resulted in ranking the attributes based on their discriminative power as well as semantic

stability. The identification experiment has demonstrated the performance advantage that is resulted from using the proposed attributes as compared with the existing ones. Thus, using 15 subject comparisons only, a correct match will be returned at rank-1. Moreover, the experiments have highlighted the impact of compressing the labels, in addition to the complexities associated with parameter tuning in the Elo rating system towards reaching a higher identification accuracy. Additionally, the verification experiment has indicated the robustness of the attributes by achieving an EER of 13.64% using 10 subject comparisons. In conclusion, the findings of this chapter significantly enrich our knowledge of identification via comparative facial soft biometrics in small constrained databases and provide more insights to extend the study further towards larger and more realistic databases.

Chapter 4

Ranking of Soft Biometrics

Ranking of soft biometrics is essential for creating biometric signatures that enable identification and retrieval of subjects in a database. In chapter 3, a scheme that is based on the Elo rating system was demonstrated for generating biometrics signatures from comparative facial soft labels. The experiments have revealed the reliability of the scheme for human face identification. However, the experiments have also highlighted the need to tune the score adjustment parameter, K , of the Elo rating system through cross validation to optimise the impact of the resulting ranking on the identification accuracy. Tuning the K parameter can result in substantial time and space complexities, specially when dealing with large datasets. Therefore, this chapter is dedicated to present a novel fully unsupervised ranking algorithm that infers ranks from pairwise comparisons without any training or parameter tuning, and thus it yields a more efficient ranking of comparative soft biometrics.

The chapter starts by highlighting the importance of ranking and provides a synopsis on the existing algorithms for ranking from pairwise comparisons. Furthermore, the chapter outlines the key issues with the existing methods of ranking, which also affect the ranking of soft biometrics attributes. Then, the chapter presents a novel mutual information based ranking system, Mutual Information for Unsupervised Ranking (MIURank), which is a parameterless algorithm for ranking from pairwise comparisons, and explains its theoretical background. The performance of MIURank is evaluated relative to a collection of well-known ranking algorithms through synthetic and real datasets. Also, the experimental results are analysed, and the implications of the findings are summarised by the end of this chapter.

4.1 Introduction

Humans are often interested in evaluating a set of objects relative to a certain criterion and producing a corresponding ranking for these objects. Accordingly, the ranking

is needed for a wide range of real-life applications such as sports, products evaluations, document retrieval, movies rating and soft biometric identification (as explained in chapter 3). Given pairwise comparisons between n items of a dataset (e.g., Person A is *older* than Person B or Product X is *more preferred* than Product Y). The objective of a ranking algorithm is to sort these items according to their relative strengths with respect to a particular criterion of interest as can be inferred from the pairwise comparisons. However, in real datasets, only a subset of all the possible pairwise comparisons is available. Moreover, pairwise comparisons can be corrupted by noise (e.g., human subjectivity in comparative labelling). These issues can be noted in the crowdsourcing of comparative labels, which was described in section 3.4.2.2, where the semantic stability analysis revealed the discrepancies in the annotators' judgements of the attributes, and not all the possible pairwise comparisons were acquired through the crowdsourcing. Accordingly, an effective ranking algorithm should be robust to corrupted or missing comparisons. Another critical aspect to be considered is the computational efficiency of ranking algorithms, in particular with the massive global interest in real-time applications and big data, which stimulate the need for ranking models of low computational complexity.

As a classical machine learning problem, ranking can be achieved through a wide variety of approaches [83, 84, 85, 86, 87, 88, 75, 47, 89, 90, 91, 92]. Therefore, ranking algorithms may be classified into four categories [85]: score-based, spectral, maximum likelihood estimation and learning to rank. Score-based methods produce scores from pairwise comparisons and sort items according to their scores. An example of a well-known score-based ranking algorithms is the Elo rating system [75] that was originally introduced for ranking chess players, and has been later extended to a wider range of applications such as soft biometric identification as it was shown in chapter 3. The Elo rating assumes normally distributed skill levels for players and updates scores based on the difference between the expected and actual outcomes of a game. TrueSkill [91] is another popular score-based ranking algorithm, with a formulation similar to Elo, but it considers the changes in a player's skill and updates the confidence in the players' skill over time. Both the Elo rating and TrueSkill systems are unsupervised, but they require parameters tuning (e.g., the score adjustment parameter, K , in the Elo system) to fit different datasets, which boosts their computational costs. Also, simpler and more computationally efficient score-based methods that rely on points difference (e.g., comparisons won minus comparisons lost) were introduced in [93] and [92].

The learning to rank algorithms, which represent the second category of ranking algorithms, are based on inferring a scoring function from training data that can be used to infer ranks for new (unseen) data. Examples for well known learning to rank formulations are: RankSVM [47], RankBoost [89] and Unified Robust Learning to Rank (URLR) [90]. This type of algorithms involves supervised learning since they require example pairwise comparisons with features to infer scoring functions. Learning to rank

algorithms also need parameter tuning besides the training required to build ranking models. In the context of soft biometric identification, RankSVM has been used in investigating comparative soft biometrics for body [44], and demonstrated the capability of producing accurate biometric signatures. Furthermore, RankSVM has been utilised for ranking clothing attributes for subject identification and retrieval [48].

Another approach for ranking from pairwise comparisons is through the estimation of scores based on maximum likelihood as in the probabilistic model of Bradley-Terry [86]. Also, some spectral formulations, in which the theory of linear maps is applied to matrices of relationships, have been proposed to solve ranking problems. PageRank [88] is an example spectral approach in which the relationships between the items to be ranked are considered as a graph, and a random walk is assumed to measure the likelihood of certain node to be visited. In addition, SerialRank [85] is a very recent spectral approach that infers the ranking through spectral seriation for a similarity matrix constructed from pairwise comparisons.

This literature review reveals the diversity of the approaches that have been taken so far for addressing ranking, and shows that most of the existing algorithms require learning to some extent. Also, it shows that the ranking of soft biometrics attributes has been so far based on algorithms that require parameter tuning to optimise identification performance as it is the case with the Elo rating [15, 23, 13, 59], and RankSVM [48, 49, 44], which also needs pairwise comparisons for training. Taken together, this motivates the proposal of a fully unsupervised algorithm that can be used to rank comparative soft biometrics without the need for training or parameter tuning, while having a competitive robustness against noise and missing comparisons.

4.2 Ranking Using Mutual Information

In the previous section, it was mentioned that the score-based approaches achieve ranking through generating a score for each item in a dataset that represents the strength of this item relative to the rest of the items. This chapter presents a novel score-based ranking algorithm that is established on an intuitive concept that is measuring an item's performance with respect to virtual super and inferior performers. The relative measurement of an item's performance is achieved by computing the similarity between the item and each of the virtual performers (super and inferior) using mutual information [94], which can be defined as the information gained about a random variable by observing another random variable. The virtual super and inferior performers can be considered as two anchor points to assess an item's performance. The virtual super performer is an imaginary item that is assumed to have the top rank (i.e. maximum score) in a given dataset; while the virtual inferior performer is an imaginary item that is assumed to have the lowest rank in that dataset. This intuition is also felicitous with the concept

of ranking a soft biometric attribute of a subject relative to the subjects that have the strongest and the weakest appearances of that attribute respectively.

In the context of information theory, the mutual information between an actual item and the virtual super performer indicates the amount of information gained about the virtual super performer by observing the performance of the actual item, thus, the more is the information revealed about the virtual super performer by the item, the higher is the item's score, which is reflected on its rank in the dataset. Likewise, the mutual information between an actual item and the virtual inferior performer represents the information gained about the virtual inferior performer by observing the performance of the actual item. This intuition enables a proper positioning for the actual item being evaluated with respect to both virtual performers and reflects its rank in the dataset. Obviously, the closer an item is to the super performer, the further it is from the inferior performer (and vice versa).

Besides the solid intuition of employing mutual information for ranking. The selection of mutual information for measuring performance and generating scores in the proposed approach is also motivated by the following:

1. Mutual information has strong scientific foundations, which are derived from its information-theoretic roots [94];
2. Mutual information embeds more and better generality for measuring the statistical dependence between two variables as compared to correlation [95]; and
3. Mutual information is simple to calculate, and hence, computationally efficient [81].

4.2.1 Algorithm Formulation

Given a dataset, D , that consists of n items, an $n \times n$ relation matrix, R , that represents all the possible pairwise comparisons in D can be defined, where an entry $R(i, j)$ represents a pairwise comparison between items, i and j , as follows:

$$R(i, j) = \begin{cases} -1 & \text{if } j \text{ is stronger than } i \\ 0 & \text{if } i \text{ and } j \text{ have similar strength, or} \\ & \text{no relation exists between } i \text{ and } j \\ 1 & \text{if } i \text{ is stronger than } j \end{cases} \quad (4.1)$$

where $i, j \in [1, n]$, and $i \neq j$. This results in the matrix R being antisymmetric since $R(i, j) = -R(j, i) \forall i \neq j$. Also, the diagonal entries of R are set to zeros, as these entries represent cyclic relations (i.e. between an item and itself). From the relation matrix R , two vectors are derived for each item $1 \leq i \leq n$ that represent the item's performance as follows: (1) the wins vector, w_i , which contains the indices of the comparisons won by

the item i ; and (2) the losses vector, l_i , which consists of the indices of the comparisons lost by the item i . The length of each performance vector is n , and the i^{th} entry of a performance vector for the item i is ignored in the computations as it represent a cyclic relation. The remaining $n - 1$ entries of w_i and l_i are set as follows:

$$w_i(j) = \begin{cases} j & \text{if } R(i, j) = 1 \\ 0 & \text{otherwise} \end{cases} \quad (4.2)$$

and,

$$l_i(j) = \begin{cases} -j & \text{if } R(i, j) = -1 \\ 0 & \text{otherwise} \end{cases} \quad (4.3)$$

where $1 \leq j \leq n$ is the index of the counterpart item in the pairwise comparison $R(i, j)$. So the winner item in a pairwise comparison gets the loser's index as a label stored its wins vector, while the loser item gets the index of the winner's as a label stored in its losses vector with a negative sign. Obviously, the virtual super performer, s , has the wins vector $w_s = \langle 1, 2, \dots, n \rangle$, as it is assumed to be the winner in all its pairwise comparisons. Similarly, the virtual inferior performer, f , has the losses vector $l_f = \langle -1, -2, \dots, -n \rangle$, as it is assumed to be the loser item in all its pairwise comparisons.

Recall that the objective is to rank the n items of the dataset D by generating scores that reflect the information gained from each item about the virtual super and inferior performers. As mentioned earlier, the information revealed by an item's performance about the virtual performers is measured based on mutual information, which is denoted as I and is calculated between two discrete random variables, X and Y , the as follows:

$$I(X, Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \log_2 \left[\frac{p(x, y)}{p(x)p(y)} \right] \quad (4.4)$$

where $p(x)$ and $p(y)$ are the probability distribution functions for the discrete random variables X and Y , correspondingly, while $p(x, y)$ is the joint probability distribution function for the discrete random variables X and Y . In the context of this thesis, it is assumed that the performance vectors for any item in a dataset are random variables. As a result, Equation 4.4 can be used to compute the mutual information between the wins vector for an item k , w_k , and the wins vector of the virtual super performer s , w_s , as follows:

$$I(w_k, w_s) = \sum_{i \in n|k} \sum_{j \in n|k} p(w_k, w_s) \log_2 \left[\frac{p(w_k, w_s)}{p(w_k)p(w_s)} \right] \quad (4.5)$$

such that $p(w_k)$, $p(w_s)$ are probability distribution functions that are calculated as follows:

$$p(w_k(i)) = \frac{c(w_k(i))}{n} \quad (4.6)$$

$$p(w_s(j)) = \frac{c(w_s(j))}{n} \quad (4.7)$$

and $p(w_k, w_s)$ is a joint probability distribution function that is calculated as:

$$p(w_k(i), w_s(j)) = \frac{c(w_k(i), w_s(j))}{n} \quad (4.8)$$

where $c(x)$ is the number of occurrences of label x , and $c(x, y)$ is the number of co-occurrences of labels x and y together. Equivalently, the mutual information between the losses vector of item k , l_k , and the virtual inferior performer f , l_f , can be measured as follows:

$$I(l_k, l_f) = \sum_{i \in n} \sum_{j \in n} p(l_k, l_f) \log_2 \left[\frac{p(l_k, l_f)}{p(l_k)p(l_f)} \right] \quad (4.9)$$

where $p(l_k)$ and $p(l_f)$ are probability distribution functions of the losses vector of the item and that of the virtual inferior performer respectively, while $p(l_k, l_f)$ is their joint probability distribution function, which can be calculated in the same way as Equations 4.6, 4.7, and 4.8. Then, the score of item k , $t(k)$, is found as the difference between the two mutual information values computed based on Equations 4.5 and 4.9 as follows:

$$t(k) = I(w_k, w_s) - I(l_k, l_f) \quad (4.10)$$

Algorithm 1 describes the MIURank procedure for ranking in detail.

In summary, in light of the algorithm description and with reference to the objectives that were set in section 4.2 for the proposed ranking algorithm, MIURank can be described as:

1. Intuitive, as the approach has information-theoretic basis and it is centred on the principle of ranking relative to two ideal performers (the best and the worst);
2. Unsupervised and parameterless, in the sense that MIURank does not involve learning, and has no parameters;
3. Efficient, as mutual information has low computational requirements [81]; and

4. Simple, since MIURank algorithm can be easily implemented and used without underlying constraints or assumptions.

Algorithm 1 Using Mutual Information for Unsupervised Ranking

```

1: Let  $R$  denote the relation matrix of the dataset  $D$  that consists of  $n$  items;  $1 \leq i \leq n$ 
   and  $1 \leq j \leq n$  are the items' indices in  $R$ ;  $w_i$  and  $l_i$  are performance vectors
   initialized to 0 for each of the  $n$  items; and  $t$  is the scores vector. For the performance
   vectors of any item, the element that represent a cyclic relation is ignored when
   computing the score in line-14.
2: for  $i \leftarrow 1; i \leq n; i \leftarrow i + 1$  do
3:   for  $j \leftarrow i + 1; j \leq n; j \leftarrow j + 1$  do
4:     if  $R(i, j) = 1$  then
5:        $w_i(j) \leftarrow j$ 
6:        $l_j(i) \leftarrow -i$ 
7:     else if  $R(i, j) = -1$  then
8:        $w_j(j) \leftarrow i$ 
9:        $l_i(i) \leftarrow -j$ 
10:    end if
11:  end for
12: end for
13: for  $k \leftarrow 1; k \leq n; k \leftarrow k + 1$  do
14:    $t(k) \leftarrow I(w_k, w_s) - I(l_k, l_f)$  {(c.f. Equation 4.10)}
15: end for
16: Sort  $t$  and derive the ranking  $\hat{r}$ .

```

4.2.2 Example: Ranking Tennis Players

The following example demonstrates how the MIURank algorithm can be used for ranking a set of items through a real dataset, which consists of the results for the top 5 players in men's single tournaments of the 2016 Australian Open Tennis Championship (AUSOpen 2016). In AUSOpen 2016, the players can only play a small number of games against each other, and yet this can influence their ranks. This example shows how this can be achieved using MIURank. The pairwise comparisons between the five players were derived from the match results [96] and the ground truth ranking of the players is based on the official ATP ranking system as shown in Table 4.1.

Index	Ground truth rank	Player
1	2	Andy Murray
2	1	Novak Djokovic
3	3	Roger Federer
4	5	Rafael Nadal
5	4	Stan Wawrinka

Table 4.1: Example subset of players from AUSOpen 2016.

The ranking process in the MIURank algorithm starts by constructing a relation matrix, R , based on the results of the games between the five players as shown in Figure 4.1, where each row i in R represents a player, while each column j represents his opponent, and an entry $R(i, j)$ represents the outcome of the game between players i and j , which is set according to Equation 4.1. Then, two performance vectors (wins and losses), are allocated to each player, whereby the length of a performance vector is set to the number of players in the dataset (i.e. five elements) and all the elements of the performance vectors are set for each player based on matrix R with the opponents' indices (as listed in Table 4.1).

$$R = \begin{bmatrix} 0 & -1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Figure 4.1: Relation matrix for the AUSOpen 2016 subset.

As mentioned previously in section 4.2.1, the element of a performance vector that corresponds to a cyclic relation between a player and himself is ignored in the ranking, and thus it is set as 'x' (i.e. *don't care*). For instance, Djokovic, who was indexed as 2 in the dataset, has won against Murray and Federer, while he did not play against Nadal or Wawrinka. So Djokovic's wins vector becomes: $w_2 = \langle 1, x, 3, 0, 0 \rangle$, while his losses vector becomes: $l_2 = \langle 0, x, 0, 0, 0 \rangle$. Table 4.2 shows the performance vectors for all the five players in this example.

Murray		Djokovic		Federer		Nadal		Wawrinka	
w_1	l_1	w_2	l_2	w_3	l_3	w_4	l_4	w_5	l_5
x	x	1	0	0	0	0	0	0	0
0	-2	x	x	0	-2	0	0	0	0
0	0	3	0	x	x	0	0	0	0
0	0	0	0	0	0	x	x	0	0
0	0	0	0	0	0	0	0	x	x

Table 4.2: Performance vectors for AUSOpen 2016 subset.

After the performance vectors are set for all the players, two virtual performers are introduced for performance benchmarking, the first is a super virtual player s with the performance vector $w_s = \langle 1, 2, 3, 4, 5 \rangle$, who theoretically has won all his four games, and the second is an inferior virtual player f with the performance vector $l_f = \langle -1, -2, -3, -4, -5 \rangle$, who has lost all his four games. Then, for each actual player $1 \leq k \leq 5$, the mutual information between each of the two performance vectors and the corresponding virtual player, $I(w_k, w_s)$ and $I(l_k, l_f)$, is calculated using Equations 4.5 and 4.9, respectively. Finally, the score, $t(k)$, for the player is determined as the difference between $I(w_k, w_s)$ and $I(l_k, l_f)$.

Player	Predicted score	Predicted rank	Ground truth rank
Andy Murray	-0.722	4	2
Novak Djokovic	1.37	1	1
Roger Federer	-0.722	5	3
Rafael Nadal	0	2	4
Stan Wawrinka	0	3	5

Table 4.3: Predicted scores and ranks for the AUSOpen 2016 subset.

Based on the relation matrix, R , and the ground truth ranking of the AUSOpen 2016 subset, the scores of the five players along with the corresponding ranking were predicted using the MIURank algorithm as shown in Table 4.3. It is important to mention that the aim of this example is to demonstrate how the MIURank algorithm works. Thus, it is not intended for performance evaluation, since the subset used in this example is very small in terms of numbers of players, and the ranking results are influenced by only two games.

4.3 Experiments

The performance of MIURank was assessed using synthetic and real-world datasets. A selection of both popular and modern algorithms were used for comparison: (1) the Elo rating; (2) the row-sum ranking (RS); (3) SerialRank (SR), which is a very recent state-of-the-art ranking algorithm; (4) PageRank (PR); and (5) Bradley-Terry maximum likelihood estimator (BTL), which is a well-known baseline for evaluating ranking algorithms. This collection was selected to cover three different categories of ranking algorithms: (1) score-based, through the Elo rating and row-sum; (2) spectral, represented by SerialRank and PageRank; and (3) maximum likelihood estimators via Bradley-Terry. The learning to rank algorithms (e.g., RankSVM) were not included in the assessment as they involve training and parameter tuning, while the objective is to compare MIURank with algorithms that involve minimum training or parameter tuning.

4.3.1 Synthetic Dataset

The synthetic dataset consists of 100 items with their corresponding ranks. The ranks are randomly sampled integers without replacement from a uniform distribution that ranges between 1 and 100. These composed the ground truth ranking, r . A binary antisymmetric matrix, R , of all the pairwise comparisons was generated. Each entry in the binary matrix, $R(i, j)$, represents a pairwise comparison between two items, i and j , where $R(i, j) = -1$ if $r(i) < r(j)$, and $R(i, j) = 1$ if $r(i) > r(j)$. Three experiments were designed to assess MIURank's performance. The first investigates robustness against noisy comparisons; the second explores tolerance for missing comparisons; and the third

assesses the effect of dataset size on the performance of a ranking algorithm. The Kendall's τ coefficient was used to measure the correlation between the ground truth and the predicted ranking in the three experiments.

4.3.1.1 Robustness to Noise

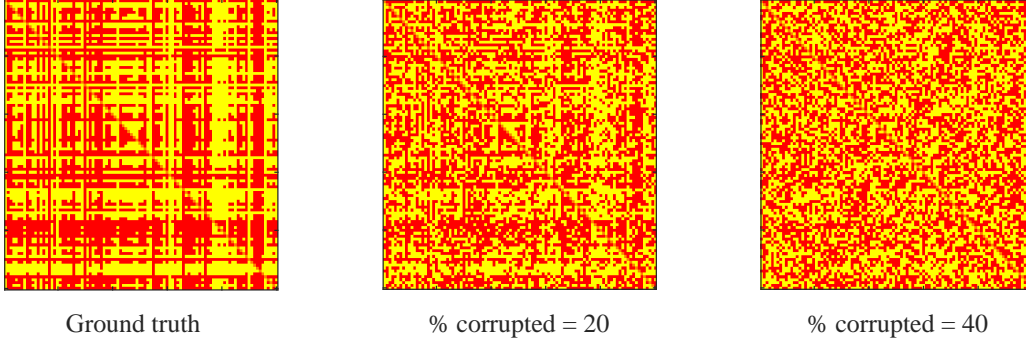


Figure 4.2: Effect of uniform randomly distributed noise on the binary relation matrix, R .

The experiment starts by randomly selecting some comparisons from the binary matrix, R , of the synthetic dataset and inverting their sign (positive to negative or vice versa) to create a partially noisy version of R as illustrated in Figure 4.2. The sign inversion is applied with each noisy comparison and its symmetrical counterpart in R . Then, each of the six ranking algorithms was used with the noisy relation matrix to retrieve a predicted ranking, \hat{r} . This procedure was performed while varying the proportion of noisy comparisons between 0 and 40% of the total comparisons in R . The experiment was repeated 50 times, and the arithmetic mean of the 50 trials was considered as the experiment outcome in terms of the Kendall's τ correlation coefficient between the ground truth ranking, r , and the ranking retrieved at each level of noise for each of the six algorithms. The Kendall's τ correlation coefficient was calculated as follows:

$$\tau = \frac{\sum_{i=1}^n \sum_{j=i+1}^n [\text{sgn}(r(i) - r(j)) \times \text{sgn}(\hat{r}(i) - \hat{r}(j))]}{\frac{n(n-1)}{2}} \quad (4.11)$$

where \hat{r} is the retrieved ranking, and n is the number of items in the dataset.

The results of this experiment are shown in Figure 4.3. The performance of MIURank was compared against: (1) the best performance of the Elo rating system (ELO), selected from those for the score adjustment parameter, K , varying between 4 logarithmically spaced values $\{4^{-1}, 4^0, 4^1, 4^2\}$; (2) the performance of row-sum (RS); (3) the performance of SerialRank (SR); (4) the performance of PageRank (PR); and (5) the performance of the Bradley-Terry model (BTL). The results that are shown in Figure 4.3, reveal

the competitiveness of MIURank as compared with Elo and Bradley-Terry in terms of its robustness to noise. Furthermore, although SerialRank slightly outperforms MIURank for the noise levels that are less than 33%, MIURank is more robust for the noise levels that are higher than 33%, where SerialRank has a steeper decline in its performance as the proportion of noisy comparisons is closer to 40%. The results also demonstrate that despite it is a parameterless algorithm that doesn't involve tuning for optimum performance, MIURank can result in at least a similar accuracy to the maximum likelihood estimator of the Bradley-Terry model and the Elo rating, which are parameter-based algorithms. Consequently, this indicates the computational efficiency of MIURank. Moreover, the results demonstrate the appropriateness of mutual information as a basis of ranking in addition to its robustness as an unsupervised framework for ranking based on pairwise comparisons.

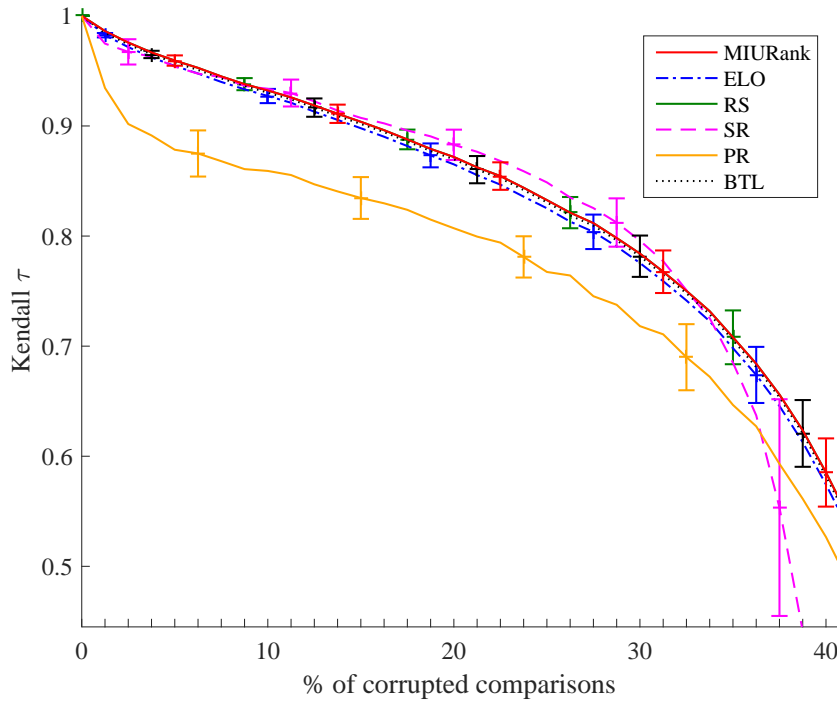


Figure 4.3: Robustness to noisy comparisons.

4.3.1.2 Tolerance for Missing Pairwise Comparisons

The 100 items of the synthetic dataset can result in a total of 4950 pairwise comparisons that can be evenly divided into 99 rounds, where each round involves each of the 100 items in exactly one pairwise comparison. As in the noise robustness analysis, the effect of missing comparisons was assessed with the six algorithms (i.e. MIURank, rowsum (RS), Elo (ELO), SerialRank (SR), PageRank (PR) and Bradley-Terry (BTL)) by randomly selecting a number of rounds (i.e. a subset of pairwise comparisons), and running each of the six ranking algorithms with the randomly selected subset. The

proportion of the randomly selected rounds was allowed to vary between 10% and 100% of the total rounds. The effect of missing comparisons was assessed using the Kendall's τ correlation between the ranking retrieved from each of the six algorithms, and the ground truth ranking. The score adjustment parameter of Elo, K , was allowed to vary between 4 logarithmically spaced values $\{4^{-1}, 4^0, 4^1, 4^2\}$. This experiment was repeated 50 times, and the arithmetic mean of the trials was considered as the experiment outcome. Figure 4.4 presents the results of this experiment.

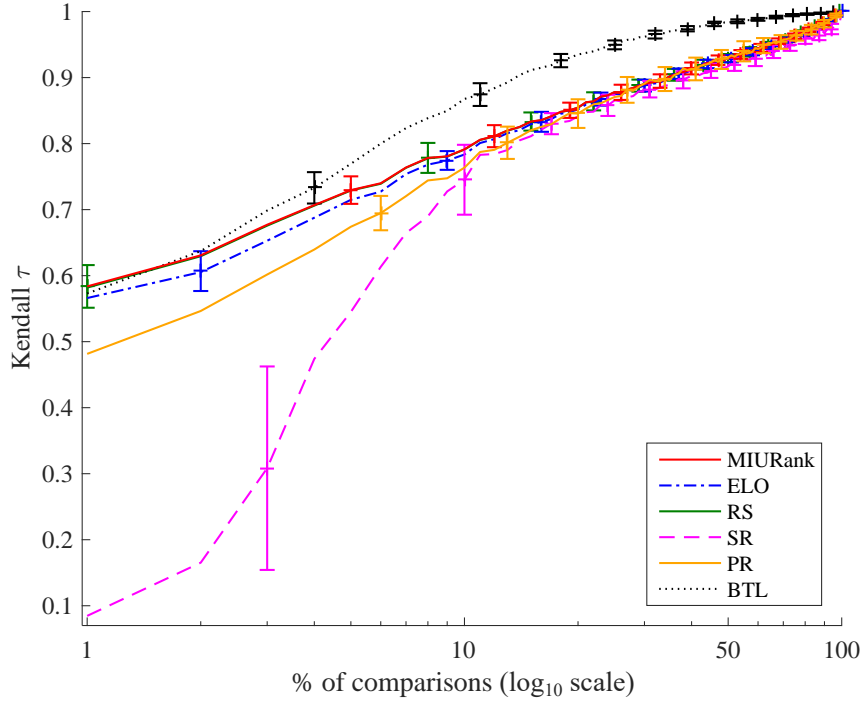


Figure 4.4: Tolerance for missing comparisons.

The results in Figure 4.4 demonstrate that MIURank outperforms SerialRank in its tolerance for missing comparisons. Surprisingly, SerialRank has a significantly low tolerance for missing pairwise comparisons when the proportion of the missing pairwise comparisons is high, and this remains an interesting question for future work investigating this behaviour by SerialRank. In addition, the results show that MIURank has a slightly better tolerance for missing pairwise comparisons than Elo. Although the Bradley-Terry model converges faster to the ground truth with less pairwise comparisons than MIURank, its superiority is at the expense of the computational complexity resulting from the requisite parameter tuning. Finally, the similarity in the performance between MIURank and row-sum in this experiment, in addition to the previous experiment (noise robustness), highlight the intuition behind MIURank, which is the difference between superiority and inferiority. Thus, it is equivalent to the main idea of row-sum, which is the difference between won and lost points.

4.3.1.3 Effect of Dataset Size

The purpose of this experiment is to explore the impact of dataset size on ranking algorithm performance in terms of its accuracy and consistency of outcomes. For this, synthetic datasets of 9 logarithmically spaced values $\{2^3, \dots, 2^{11}\}$ were created with 25% of the pairwise comparisons corrupted in the same way mentioned previously in this section. The noise level was set as 25% as it is the middle point between the ideal case and the random chance (i.e. no noise and 50% noised). As in the previous experiments, the accuracy of each algorithm was measured based on the Kendall's τ correlation between the predicted ranking and the ground truth ranking.

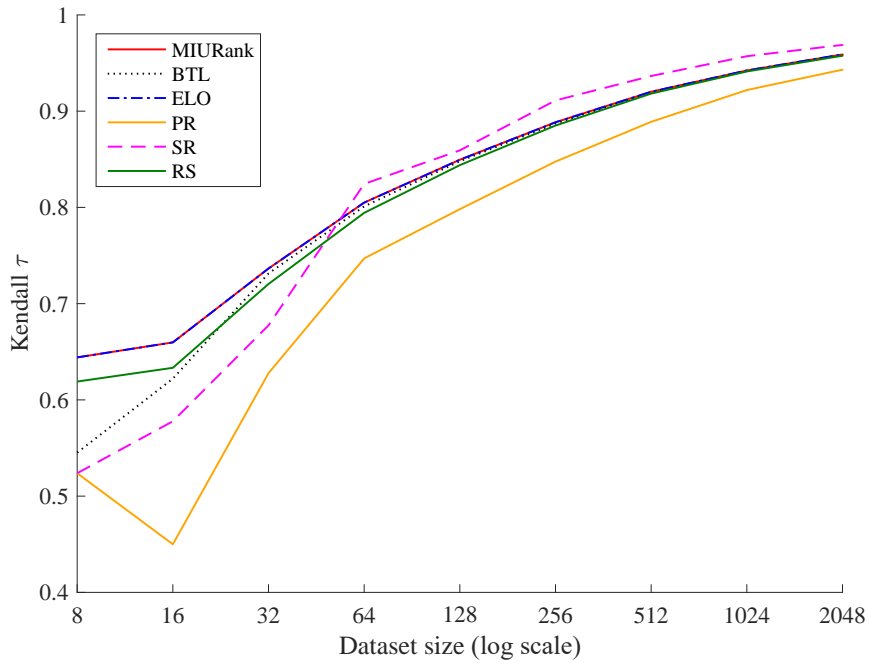


Figure 4.5: Effect of dataset size on ranking accuracy.

The results of this experiment (shown in Figure 4.5) reveal that both MIURank and Elo have the best outcomes for small datasets of (i.e. 8 to 32 items). Moreover, although Bradley-Terry has lower accuracy with small datasets, its accuracy significantly improves with the growth of dataset size. An interesting finding from this experiment is the strong association between dataset size and the accuracy of SerialRank. Thus, the accuracy of SerialRank outperforms the other five algorithms for larger datasets (greater than 128), while it rapidly deteriorates for smaller datasets (64 and less). In conclusion, the results of this experiment show that MIURank has a relatively consistent performance regardless of the dataset size, and hence, it reveals MIURank scalability.

4.3.2 Real Datasets

4.3.2.1 English Premier League

Examining the performance of a ranking algorithm using sports data enables the evaluation of its effectiveness in real life applications. Therefore, the results of English Football Premier League (EPL) teams for season 2014-2015 [97] were chosen to evaluate the performance of MIURank and compare it with the other five algorithms (i.e. row-sum (RS), Elo (ELO), SerialRank (SR), PageRank (PR) and Bradley-Terry (BTL)). The ground truth for this experiment is the EPL final ranking (based on the points of the 20 EPL teams). Since each of the teams has played against each of the other 19 teams twice (i.e. home and away matches), two pairwise comparisons can be produced between each pair of teams, which results in a total of 380 pairwise comparisons for the entire tournament. The two pairwise comparisons for each pair of teams were then averaged, and the value of the average pairwise comparison was mapped to -1, 0, or 1 for loss, tie, or win, correspondingly. These resulted in 190 pairwise comparisons that were used to retrieve rankings based on the six algorithms. The performance evaluation in this experiment is based on the percentage of upsets in top k ranks, U_k , where a ranking upset is the case in which a low ranked team beats a highly ranked team, and it is calculated as follows:

$$U_k = 1 - \frac{\sum_{i=1}^k \sum_{j=i+1}^k \left| \text{sgn}(r(i) - r(j)) + \text{sgn}(\hat{r}(i) - \hat{r}(j)) \right|}{n(n-1)} \quad (4.12)$$

where r is ground truth ranking, \hat{r} is the predicted ranking, and n is the number of items in the dataset. It is important to mention that ranking upsets occur as a result of form, which is not included in this (post hoc) analysis.

The score adjustment parameter of Elo, K , was allowed to vary between the values: $\{4^{-1}, 4^0, 4^1, 4^2\}$ and the best outcome of the Elo rating system, ELO, was considered for evaluation. Figure 4.6 shows the percentage of upsets at different ranks with each of the six algorithms in addition to the official EPL ranking that is based on the sum of points (1 point for a draw and 3 points for a win). The results show that MIURank is more accurate than the other methods in its predicted ranking. Thus, the ranking generated based on MIURank is 100% accurate in predicting the top ten teams in the EPL dataset, and has a roughly equivalent accuracy to Elo for the lowest ten ranks. The Elo rating system comes second in terms of its low percentage of upsets, while Bradley-Terry has a significantly high percentage of upsets for the top five positions in the ranking. The new SerialRank approach has the same performance as MIURank up to rank-6 and then follows the upsets of the official ranking before diverging after rank-10. Note that Elo appears better than MIURank between rank-12 to 14. It is of note that this is the best performance and Elo involves tuning. Finally, whereas row-sum yields in a performance

that is equal to MIURank in the experiments performed with the synthetic dataset, the results of this experiment highlight the preponderance of MIURank and reveals its effectiveness for ranking real datasets.

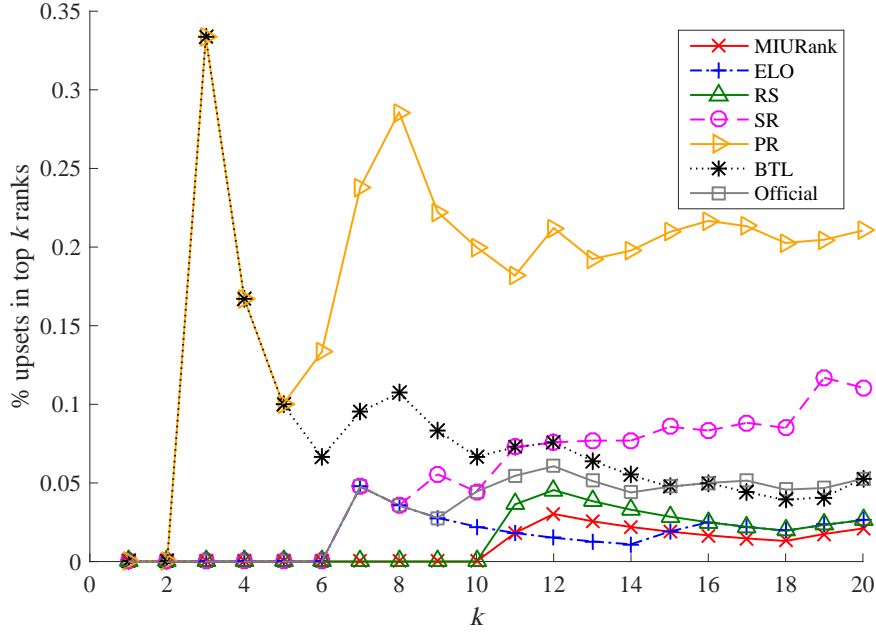


Figure 4.6: Percentage of ranking upsets with the EPL dataset.

4.3.2.2 Human Identification Using Comparative Soft Biometrics

As mentioned at the beginning of this chapter, the proposal of MIURank was motivated by the need to rank comparative facial soft biometrics efficiently via a fully unsupervised scheme. Whereas the Elo rating system demonstrated a significant contribution in the soft biometric identification experiments in chapter 3, its use is accompanied by the overhead of tuning the K parameter as it was shown in section 3.5.4. This experiment aims to assess the performance of MIURank in ranking the comparative soft biometric attributes (listed in Table 3.1), and to explore its impact on identification using the BioT dataset, which consists of 100 subjects. The experimental design that was followed here is similar to that explained in section 3.5.1. The experiment was conducted in two versions, the first uses the Elo rating system for ranking the attributes with the optimum K value (as highlighted in 3.5.4), and the second uses MIURank. The number of subjects comparisons used in generating a probe biometrics signatures is ten, as it is the average size of an ideal identity parade [98]. The ten subjects used for comparison with each probe were randomly selected, and to ensure consistency in the experiment inputs for the two ranking algorithms, the same randomly selected ten subjects were used for both the Elo and MIURank trials. The experiment was repeated 30 times, and the arithmetic mean of the ranks resulting for each subject is considered as the experiment outcome for each trial (i.e. Elo and MIURank).

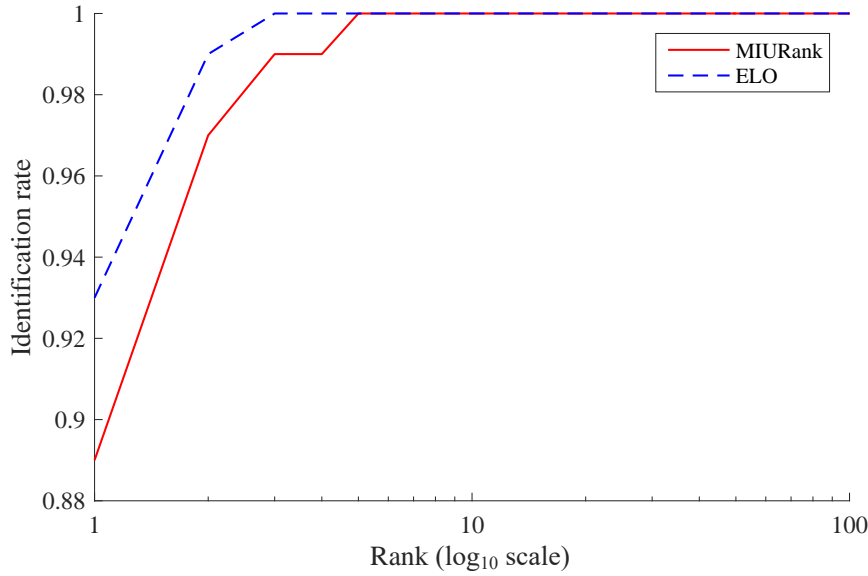


Figure 4.7: Identification performance of MIURank compared with the best outcome of the Elo rating system.

Figure 4.7 shows that there is a slight advantage for the Elo rating system between rank-1 and 4, while the performance of both algorithms converges starting from rank-5. Taking into consideration that MIURank is a fully unsupervised algorithm, while the identification performance is highly sensitive to the K parameter value in the Elo rating system (refer to section 3.5.4), the results demonstrate the reliability of MIURank for ranking soft biometric attributes and its potential for efficiently ranking larger datasets, which will be investigated in chapter 5.

Computational Cost Analysis

In the context of ranking soft biometric attributes from pairwise comparisons, it is of interest to analyse the computational cost of MIURank and compare it to that of the Elo rating system, as this analysis provides better insights on the efficiency that can be expected from either algorithm. Assuming that addition and subtraction operations are of cost c_1 , multiplication and division are of cost c_2 , and exponential and logarithmic functions are of cost c_3 , such that $c_1 < c_2 < c_3$ [99]. From the description provided in section 3.1, and for a dataset of n items, the Elo rating system requires a computational cost of $8c_1 + 6c_2 + 2c_3$ per comparison. On the other hand, MIURank needs $n(c_1 + 2c_2 + c_3)$ computations per item. As a result, the computational cost for ranking a dataset of n items can be approximated as $2n^2(4c_1 + 3c_2 + c_3)$ for the Elo rating system, and $n^2(c_1 + 2c_2 + c_3)$ for MIURank. This demonstrates that MIURank is more computationally efficient than the Elo rating system. Additionally, the Elo rating system requires iterative processing (i.e. the new score is updated based on the old score), which restricts the potential of vectorisation towards more efficient computations. Also, it has been shown that the Elo rating system needs cross validation for selecting the K

factor, while MIURank is fully parameterless. In light of the aforementioned aspects, this analysis clearly demonstrate that MIURank is more computationally efficient than the Elo rating system.

4.4 Conclusions

This chapter proposed a novel mutual information based ranking system, Mutual Information for Unsupervised Ranking (MIURank), which can be used for ranking attributes in soft biometric identification. As the motivation for proposing MIURank was to achieve the ranking of soft biometric attributes in a fully unsupervised way, it was shown that this could be accomplished by exploiting mutual information to infer scores for attributes from pairwise comparisons. The strong intuition on which MIURank is centralised was explained in this chapter besides the information theoretic foundations of the score generation process.

The investigation of MIURank effectiveness and accuracy was conducted using synthetic and real datasets relative to a collection of well-known ranking algorithms. The experiments have shown that MIURank could outperform the maximum likelihood estimator of the Bradley-Terry model in its robustness to noisy comparisons. Also, the experiments have revealed the reliability of MIURank for ranking soft biometrics. Thus, it can reach a performance level that is slightly lower than the best of the Elo rating system with the BioT dataset. The results of the experiments conducted throughout this chapter reveal that MIURank can achieve robustness and efficiency without any learning or parameters. In summary, these findings demonstrate the potential of MIURank for ranking soft biometrics in larger datasets, which will be investigated in the next chapter. Moreover, the findings highlight the power of information theory in addressing ranking as a machine learning problem.

Chapter 5

Unconstrained Identification Using Comparative Facial Soft Biometrics

In chapter 3, human identification via comparative facial soft biometrics was explored using the small and relatively constrained BioT dataset. Further, the experimental results have demonstrated the effectiveness of the proposed attributes as compared to the existing work on similar databases. However, the analysis and experiments presented in chapter 3 have revealed three key issues. First, the shape-based attributes have, in general, low discriminative power and semantic stability. Second, the level of comparative labels can be compressed without negatively affecting the identification performance. Third, the use of the Elo system for ranking attributes has highlighted the overhead of the cross validation that is required for tuning the score adjustment parameter, K . Accordingly, the following enhancements can be applied to address the issues raised by the findings from chapter 3:

1. Substituting the shape-based attributes, which have shown low discriminative power in the BioT dataset, with new attributes;
2. Reducing the levels of the comparative labels from five to three; and
3. Addressing the overhead of parameter tuning in the ranking process by the novel fully unsupervised MIURank algorithm, which was introduced in chapter 4.

Despite addressing the three issues mentioned above, there is still a major knowledge gap that needs to be filled. Thus, the experiments in chapter 3 were performed using the BioT dataset, which is small and relatively constrained, while the real surveillance scenarios involve larger and more challenging databases. This knowledge gap was also

identified in the literature review that was presented in chapter 2, which showed the limitations of the existing work in comparative soft biometrics, as it is restricted to small and constrained dataset. Therefore, the aim of this chapter is to explore unconstrained human identification through comparative facial soft biometrics, where the term "unconstrained" refers to the nature of the visual conditions in which facial images have been acquired such as variance in pose, facial expressions, illumination and resolution. Moreover, the chapter explores unconstrained face verification using comparative facial soft biometrics. To enable such exploration, the analyses and experiments in this chapter use Labelled Faces in the Wild (LFW) [56], which is a well-known database for studying unconstrained face recognition, and it is described in detail later in this chapter.

The chapter introduces an enhanced set of comparative facial soft biometrics, which is proposed based on the findings from the attribute analysis conducted in chapter 3. Then, a description for the LFW dataset is provided, and the acquisition of comparative labels via crowdsourcing is described. Furthermore, the chapter presents statistical insights and significance analysis for the attributes. Finally, the identification and verification experiments are explained, and their outcomes are discussed.

5.1 Enhanced Comparative Facial Soft Biometrics

In section 3.2, the guidelines that governed the selection of comparative facial soft biometrics in this thesis were highlighted as follows: (1) maintaining a coverage of the major facial parts; (2) emphasizing the substantial role of eyebrows in human face recognition [34, 26]; and (3) the effectiveness of attributes as semantic descriptors (i.e. understandability, memorability and describability). These guidelines have been followed in chapter 3 to define the comparative facial soft biometrics for studying constrained human identification using the BioT dataset (listed in Table 3.1). Furthermore, the results of the identification and verification experiments in section 3.5 have demonstrated the effectiveness of the proposed comparative facial soft biometrics, and the performance improvement they can achieve as compared with the existing work on comparative facial soft biometrics. However, the analyses in section 3.4 have indicated that the attributes significantly vary in their discriminative power and semantic stability. Thus, while eyebrows, nose and global attributes show relatively high discriminative power, the shape-based attributes (e.g., *jaw shape* and *face shape*) have the least discriminative power, in addition to low semantic stability.

Based on the outcomes from the analysis presented in section 3.4, the soft biometric set (listed in Table 3.1) was revised and enhanced by substituting the five least significant attributes, which are: *cheek shape*, *eye shape*, *jaw shape*, *face shape* and *cheek size*, with the following attributes: *eyebrow hair colour*, *eyebrow shape*, *nose septum*, *spectacles* and *facial hair*. As can be noted, the new attributes signify the eyebrows and nose

No.	Attribute	Labels
1	Chin height	[More Small, Same, More Large]
2	<i>Eyebrow hair colour</i>	[More Light, Same, More Dark]
3	Eyebrow length	[More Short, Same, More Long]
4	<i>Eyebrow shape</i>	[More Low, Same, More Raised]
5	Eyebrow thickness	[More Thin, Same, More Thick]
6	Eye-to-eyebrow distance	[More Small, Same, More Large]
7	Eye size	[More Small, Same, More Large]
8	Face length	[More Short, Same, More Long]
9	Face width	[More Narrow, Same, More Wide]
10	<i>Facial hair</i>	[Less Hair, Same, More Hair]
11	Forehead hair	[Less Hair, Same, More Hair]
12	Inter eyebrow distance	[More Small, Same, More Large]
13	Inter pupil distance	[More Small, Same, More Large]
14	Lips thickness	[More Thin, Same, More Thick]
15	Mouth width	[More Narrow, Same, More Wide]
16	Nose length	[More Short, Same, More Long]
17	<i>Nose septum</i>	[More Short, Same, More Long]
18	Nose-mouth distance	[More Short, Same, More Long]
19	Nose width	[More Narrow, Same, More Wide]
20	<i>Spectacles</i>	[Less Covered, Same, More Covered]
21	Age	[More Young, Same, More Old]
22	Figure	[More Thin, Same, More Thick]
23	Gender	[More Feminine, Same, More Masculine]
24	Skin colour	[More Light, Same, More Dark]

Table 5.1: Enhanced soft biometric set.

weights in discriminating subjects, which have been shown through the analyses in section 3.4. Also, they include binary-like features (*spectacles* and *facial hair*), which were included to further discover the performance of binary attributes when used in a comparative format. Table 5.1 shows the enhanced soft biometric set, where we can also see that the comparative labels are expressed based on a 3-point bipolar scale (compared with the 5-point bipolar scale in Table 3.1), as it was demonstrated in section 3.5.2 that label compression can significantly improve identification performance.

5.2 Dataset and Label Acquisition via Crowdsourcing

5.2.1 The LFW Database

Labelled Faces in the Wild (LFW) [56] is a popular database that is used for unconstrained face recognition; it consists of more than 13000 facial images extracted from the web. The images of LFW have significant variations in pose, lighting, resolution and expressions, which make them suitable to study unconstrained face recognition. The LFW database is composed of two subsets: View 1, which is dedicated to training and model selection; and View 2, which is dedicated to performance analysis. The training



Figure 5.1: Sample face images from the LFW database that show variations in: (a) pose; (b) illumination (c); resolution; and (d) facial expressions.



The eyebrow horizontal length of Person-A relative to that of Person-B is:

- ☐ More Short
- ☐ Same
- ☐ More Long
- ☐ Don't know

Figure 5.2: Example question from the crowdsourced job launched to collect comparative labels for the LFW-V1 dataset.

subset of View 1 consists of 9525 sample face images for 4038 subjects; some of these subjects have one sample in the database, while the others have two or more samples.

To study unconstrained identification using comparative facial soft biometrics, a dataset (LFW-V1 hereafter), which includes the 4038 subjects of the training subset of View 1 from the LFW database, was created by selecting one sample face image for each subject, and applying random selection whenever multiple samples exist for a subject. The selected images were all aligned using deep funnelling [100], which is an approach that incorporates unsupervised joint alignment with unsupervised feature learning to align face images and reduces the effect of pose variability correspondingly. Also, all the images in the dataset were normalised to an inter-pupil distance of 50 pixels, which was similar to a procedure applied earlier with the BioT dataset in chapter 3. Figure 5.1 shows sample face images from the LFW database.

5.2.2 Crowdsourcing of Comparative Labels

The number of pairwise comparisons that result from a set of n items is $n(n-1)/2$, accordingly, the 4038 subjects in the LFW-V1 dataset result in 8.15 million pairwise comparisons, which is a massive number that is infeasible to be crowdsourced. Therefore, a graph that models pairwise relations between the 4038 subjects has been designed using a topology that ensures the involvement of each subject in at least four pairwise comparisons. The graph resulted in 10065 pairwise comparisons that were crowdsourced in the same way described in section 3.3; thus, each crowdsourced comparison consists of 24 questions targeting the comparative labelling of the attributes that are listed in Table 5.1.

As explained earlier, each attribute is labelled based on a 3-point bipolar scale that represents the difference between the subject-pair being compared. Figure 5.2 shows an example crowdsourced question. The crowdsourcing of the LFW-V1 dataset comparisons resulted in the collection of 241560 comparative labels. These comparative labels were used to infer more comparisons following the inference rules illustrated in Table 3.2. More statistics about the crowdsourcing of the LFW-V1 dataset are shown in Table 5.2.

	Collected	Inferred	Total
Attribute comparisons	241560	132879504	133121064
Subject comparisons	10065	5536646	5546711
Average number of comparisons per subject	4.98	1371.1	N/A
Number of annotators (contributors)	9901	N/A	N/A

Table 5.2: Crowdsourcing job statistics for the LFW-V1 dataset.

Figure 5.3 shows the distribution of the collected comparative labels via the crowdsourcing of the LFW-V1 dataset. It can be noted from Figure 5.3 that the attributes of the enhanced soft biometric set, which is proposed in this chapter, exhibit a more uniform distribution among the three label levels as compared with the distribution of the BioT labels that is shown in Figure 3.5. However, the ambiguity of the comparative facial soft biometrics (i.e. the percentage of "Don't know" labels) seems to be higher in the LFW-V1 dataset as compared with the BioT dataset, which can be attributed to the challenging conditions of the samples in the LFW database. Also, the *inter pupil distance* and *nose septum* have the highest ambiguity, which might be due to face normalisation to a fixed inter-pupil distance and the vagueness of the *nose septum* attribute respectively. Moreover, it can be observed that *facial hair*, *spectacles* and *gender* have high percentage of the "Same" label, which might be due to their binary nature. In general, although the distribution in Figure 5.3 gives some insights about the collected labels, its data does not necessarily reflect the attribute significance and informative value, which will be addressed in detail throughout the next section.

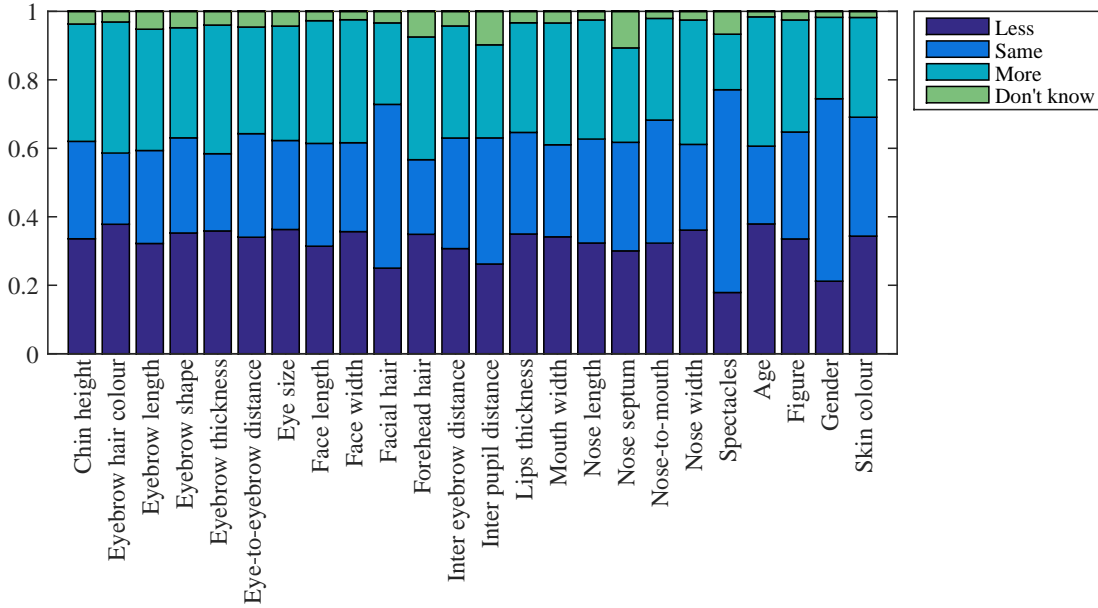


Figure 5.3: Distribution of crowdsourced comparative labels for the LFW-V1 dataset.

5.3 Dataset Analysis

This section aims to describe the attributes' data in their relative form in addition to assessing their significance and discriminative power. The analysis presented in this section is similar to those that have been applied to the BioT dataset in section 3.4. Thus, the crowdsourced comparative labels (Table 5.2 and Figure 5.3) were used to infer more comparisons through relation inference (listed in Table 3.2), and accordingly, they were used to deduce scores and ranks. The only difference between the BioT dataset experiments in section 3.4 and this analysis is in the ranking of the attributes. Thus, while the Elo rating system was used for ranking with the BioT dataset in chapter 3, MIURank was used here with the LFW-V1 dataset for the same purpose. As the LFW-V1 dataset is significantly more diverse, realistic and larger than the BioT dataset, it is assumed that the outcomes of the LFW-V1 dataset are more reflective of the real surveillance databases than the outcomes of the BioT dataset.

5.3.1 Dataset Distribution

Figure 5.4 shows the distribution of the attributes in the LFW-V1 dataset summarised using box plot. The most emphatic aspect from Figure 5.4 is the strong presence of outliers in most of the attributes, whereas the BioT attributes in Figure 3.6 have very few outliers. Furthermore, the variations of the LFW attributes differ greatly as compared to the BioT attributes. These two findings are likely to be related to the fact that the LFW-V1 dataset is much larger and more diverse than the BioT dataset.

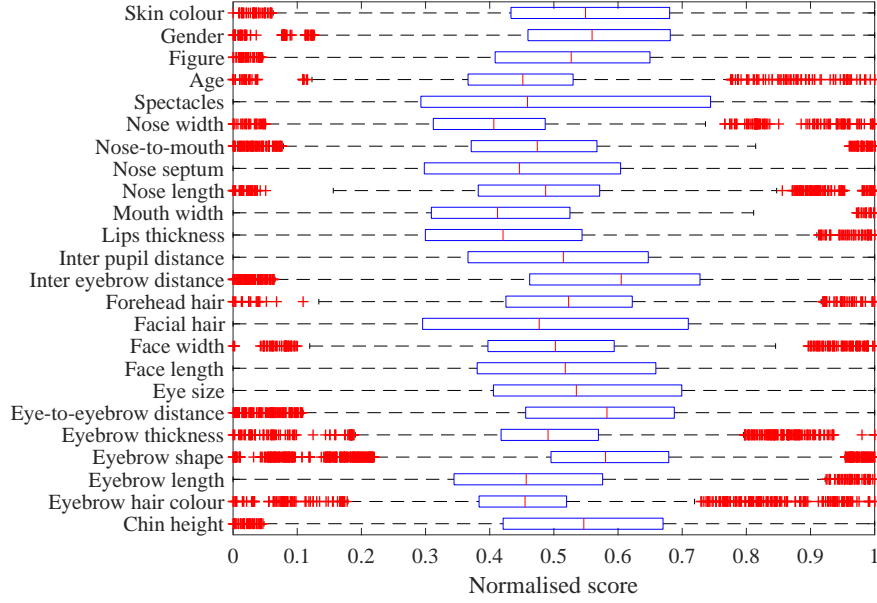


Figure 5.4: Box plot for the scores of the attributes.

The relationship between the scores and the corresponding ranks of the subjects is presented in Figure 5.5, which indicates that the higher is the variation of an attribute, the greater is the tendency of the relationship between its scores and ranks to linearity. Thus, the relationship between scores and ranks in *facial hair* and *spectacles*, which are binary-like attributes with the highest variation, has more tendencies towards linearity. On the other hand, the relationship between scores and ranks for the attributes with a low variation such as *age*, *eyebrow hair colour* and *eyebrow thickness*, has more tendency to form a threshold-like sigmoid function.

When examining the attributes' data for normality using the Anderson-Darling test [79], the resulting p -values for all the attributes are very close to zero, which indicates that the attributes are not normally distributed. The non-normality of the attributes might be due to the strong presence of outliers as illustrated in Figure 5.4. Additional highlights on the distribution of attributes' data are presented in Figure 5.6, which shows histogram representation for the attributes' data, where it can be noted that some attributes have more tendencies to uniformity (e.g., *facial hair* and *spectacles*), and some attributes have notable skewness (e.g., *eye-to-eyebrow distance* and *nose width*).

Examples of the outcomes from the ranking of attributes are visually illustrated in Table 5.7, which presents the top and least ranked subject for selected attributes. Also, Figure 5.8 shows the top five most similar subjects in the dataset, where the similarity between the subjects' biometric signatures was computed using the cosine distance measure as in Equation 5.1.

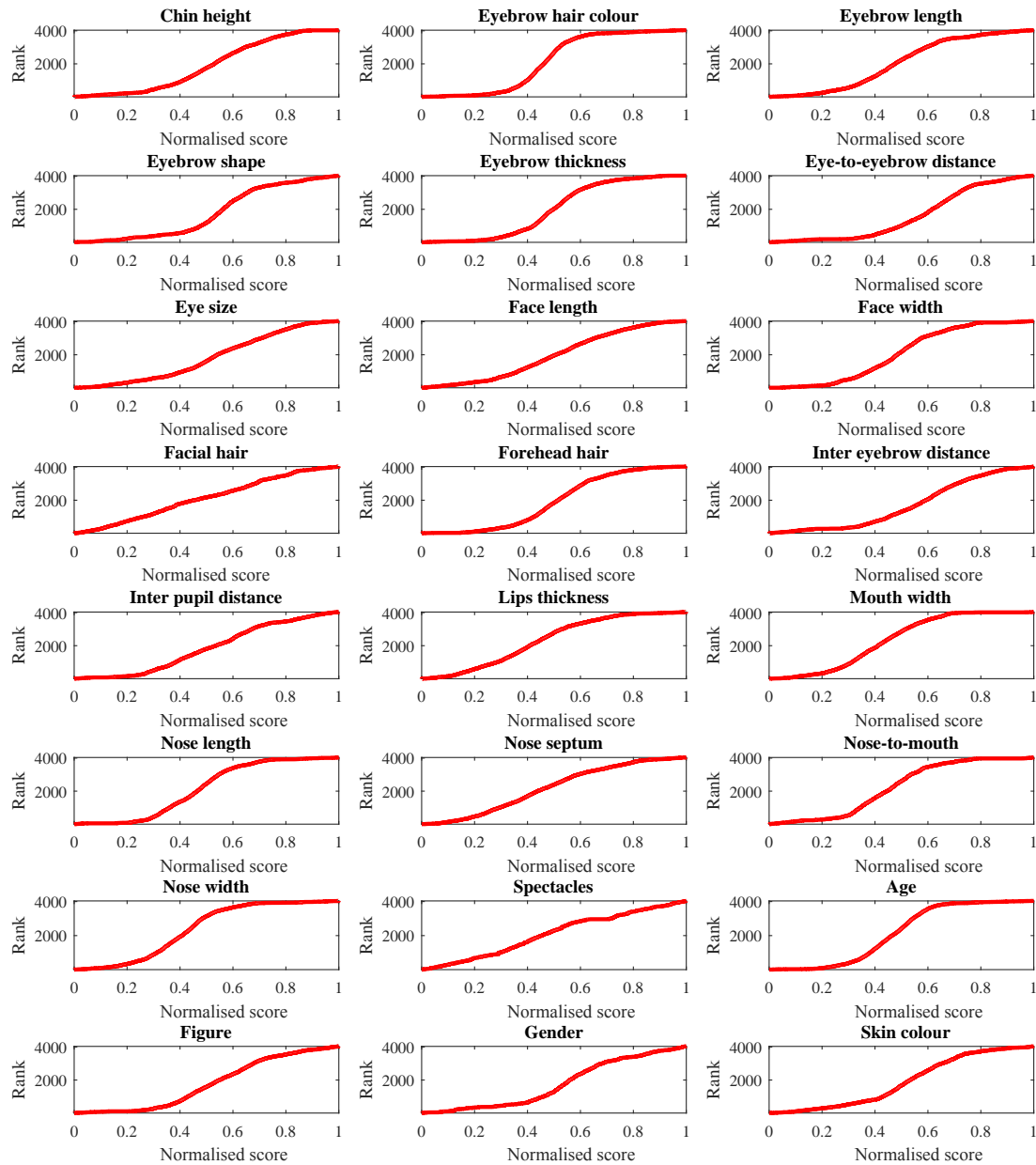


Figure 5.5: Relationship between normalised scores and ranks of the attributes.

5.3.2 Attribute Significance

Similarly to the analysis that was presented in section 3.4, this section presents assessments for: (1) attribute discriminative power, which reveals the capability of an attribute in distinguishing subjects; and (2) attribute semantic stability, which refers to the consistency of the attribute ranking among different annotators.

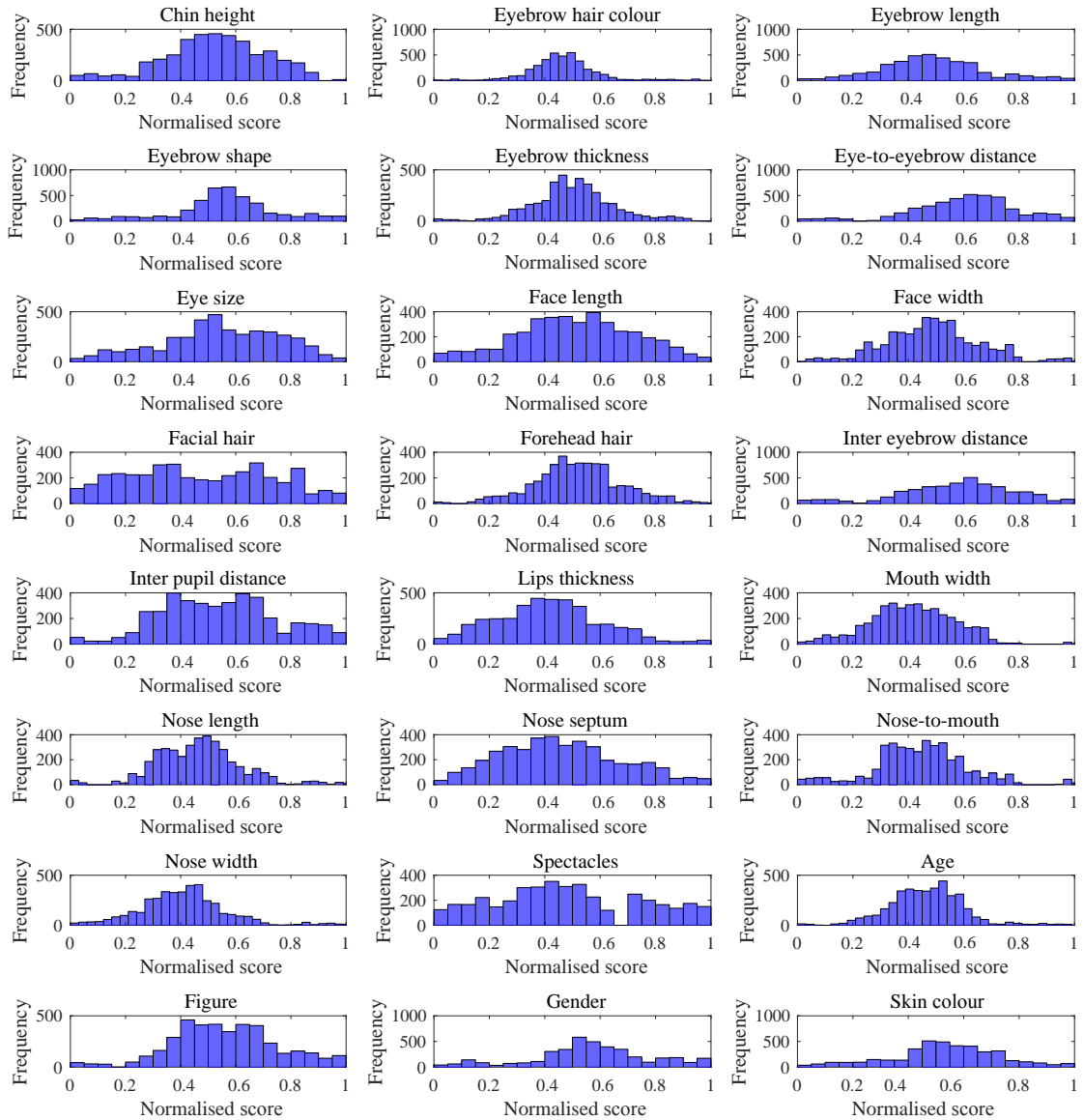


Figure 5.6: Distribution of the attributes.

5.3.2.1 Attribute Discriminative Power

Table 5.3 presents the outcomes of the discriminative power of each attribute measured with the three methods described in section 3.4.2.1 earlier, which are: Analysis of Variance (ANOVA), entropy and mutual information. Also, Figure 5.9 shows the discriminative power of the attributes represented in terms of the normalised rank score, which was introduced and explained in section 3.4.2.1. The results of the discriminative power analysis in Figures 5.9 show that the three methods appear to differ in their evaluations of the attributes, and hence, this enriches the analysis. The results show that the binary-like attributes, which are: *facial hair*, *spectacles* and *gender*; have the highest discriminative power. Also, the results demonstrate the significance of face


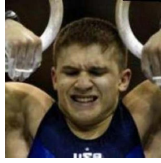



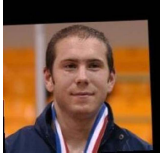
Eye-brow hair colour	Eye-brow shape	Eye-to- eyebrow dist	Face length	Facial hair	Inter eye- brow dist
					
					
Lips thickness	Nose length	Nose-to- mouth	Spectacles	Figure	Skin colour
					
					

Figure 5.7: Visualization of the ranking of selected attributes for the LFW-V1 dataset using MIURank. For each attribute, the upper image represents the top ranked subject, while the lower image shows the least ranked subject. The comparative labels that correspond to each attribute are listed in Table 5.1.

measurement attributes in general such as: *face width*, *face length* and *inter pupil distance*. Interestingly, *eyebrow hair colour* has a relatively low discriminative power by the three methods, which might indicate its low effectiveness as a soft biometric attribute. However, the significance of *eyebrow hair colour* will be further investigated with the semantic stability analysis in the next section.

5.3.2.2 Attribute Semantic Stability

Following the same procedure in section 3.4.2.2, the semantic stability analysis was performed with the LFW-V1 dataset by assessing the Pearson's correlation for each attribute between two galleries that were constructed using two mutually exclusive subsets of comparisons. The outcomes from the semantic stability analysis are shown in Figure 5.10 and provide another perspective, along with the discriminative power analysis,



Figure 5.8: The top five similarities among the subjects of the LFW-V1 dataset.

No.	Attribute	ANOVA (F-statistic)	Entropy	Mutual information
1	Chin height	0.0576	7.2442	10.4512
2	Eyebrow hair colour	0.0025	6.1789	8.3484
3	Eyebrow length	0.0997	6.9784	10.3989
4	Eyebrow shape	0.5514	7.4050	9.5648
5	Eyebrow thickness	1.1210	6.6298	9.9985
6	Eye-to-eyebrow distance	0.1230	7.2554	10.3062
7	Eye size	0.0003	6.8863	10.5556
8	Face length	0.2737	7.0862	10.6832
9	Face width	1.2384	7.1719	10.2233
10	Facial hair	0.0093	7.4314	10.6333
11	Forehead hair	0.0050	6.4189	10.4673
12	Inter eyebrow distance	1.1984	7.3168	9.5488
13	Inter pupil distance	0.0142	7.2880	10.7213
14	Lips thickness	0.0534	7.1201	9.9349
15	Mouth width	0.9685	6.9304	10.2721
16	Nose length	0.0568	7.1931	10.3775
17	Nose septum	0.5042	7.1746	9.2605
18	Nose-mouth distance	0.2695	7.3166	10.3469
19	Nose width	0.0007	6.6187	10.5559
20	Spectacles	0.0097	7.4360	10.7279
21	Age	4.5699	5.7867	8.9142
22	Figure	0.0010	7.1437	10.5144
23	Gender	0.2192	7.3704	10.3508
24	Skin colour	0.0017	7.2080	10.0677

Table 5.3: Discriminative power of the attributes.

to assess the significance of the attributes in describing and identifying the subjects of the LFW-V1 dataset.

The results of the semantic stability analysis (shown in Figure 5.10) provide additional evidence of the superiority of the binary-like facial attributes, which are: *facial hair* and *spectacles*. On the other hand, the results show the inconsistency of *age* estimations

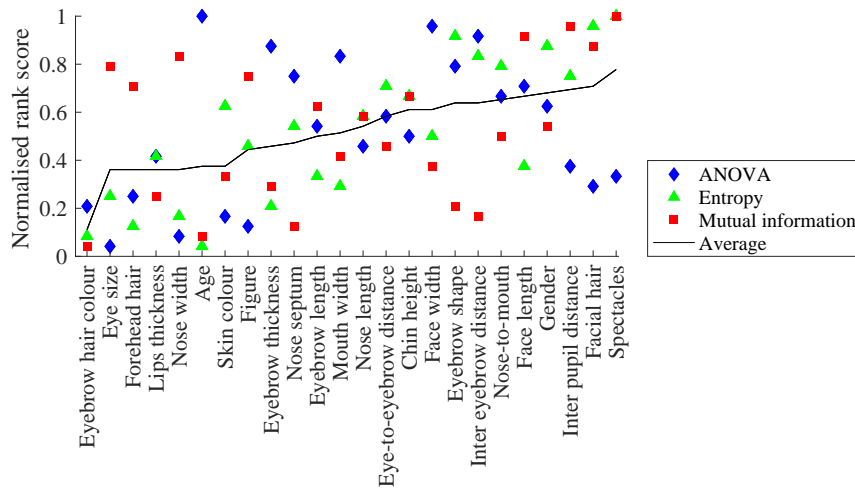


Figure 5.9: Discriminative power of the attributes.

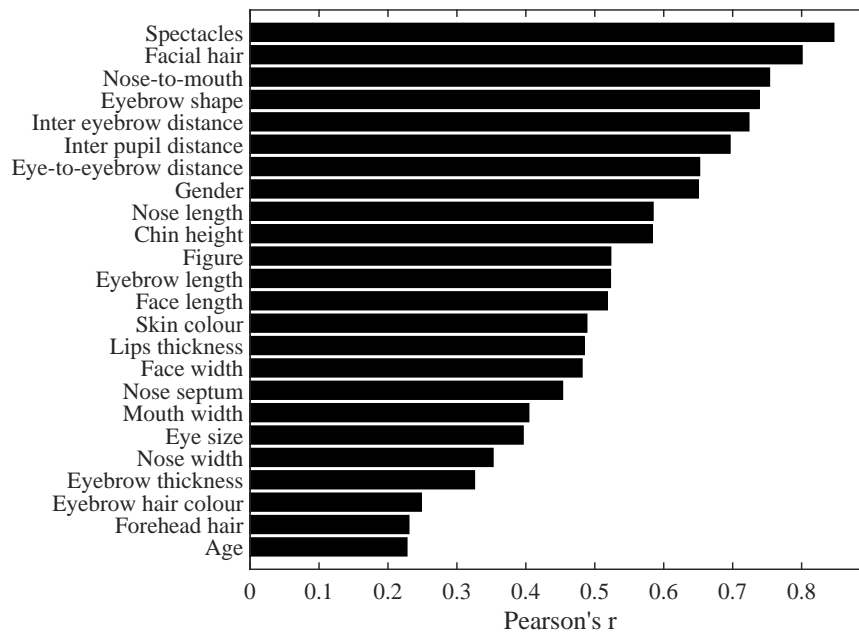


Figure 5.10: Semantic stability of the attributes.

among annotators, which supports the findings of [77] regarding the inaccuracy of humans in estimating age from facial images. The attribute that comes second in terms of low semantic stability is *forehead hair*, which might be due to the high variability of hairstyles, which confuses the annotators and may affect the accuracy of their judgments. Furthermore, the analysis highlights the low stability of *eyebrow hair colour*, which reveals the discrepancies in annotators' evaluation of this attribute, and explains its low discriminative power that was identified in the previous section. The primary finding from this analysis is that the p -values yielded from assessing the semantic stability using Pearson's correlation coefficient are all close to zero, which demonstrate the statistical significance of all the attributes. Moreover, it can be seen in Figures 5.9

and 5.10 that there is notable association between the attributes' discriminative power and their semantic stability.

5.3.3 Attribute Correlations

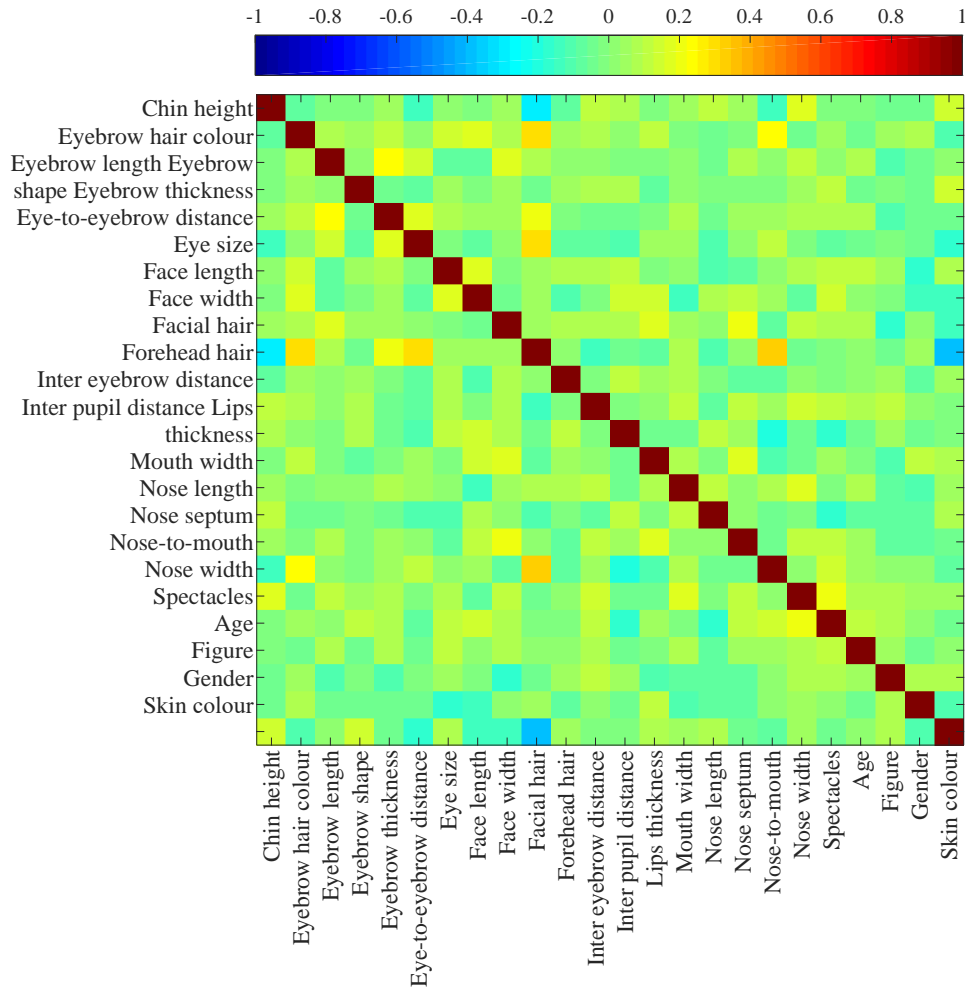


Figure 5.11: Correlations between the attributes using Pearson's r .

The Pearson's correlation coefficient, r , is used in this analysis to explore any associations (or dependencies) between the attributes. The correlations map is shown in Figure 5.11, where $|r| = 1$ represents a perfect correlation, while $r = 0$ indicates no correlation. As can be seen from Figure 5.11, all the correlations between the attributes are either insignificant or very weak, except for the interesting negative moderate correlation between *facial hair* and *skin colour*, which might be related to the preferences of some ethnicities. Another notable weak negative correlation can be observed between *chin height* and *facial hair*, which can be attribute to the interference caused by beard (as *facial hair*) on *chin height* estimations. Overall, the correlations between the attributes

reveal their independence and the potential unique contribution of each attribute in identification, which will be further discovered in section 5.4.3.

5.4 Experiments

5.4.1 Unconstrained Identification Using Facial Comparisons

The identification experiment in this section simulates a realistic scenario in which a soft biometric database is searched to identify an unknown subject (probe) using a semantic description (eyewitness statement) for the subject's face (refer to Figure 3.12). The identification performance evaluation followed 6-fold cross validation, where the 4038 subjects of the LFW-V1 dataset were randomly divided into six equal folds, and each fold was used for testing while the remaining five folds were used for training. For each subject in the test set, a probe biometric signature was generated from the scores of the 24 attributes (listed in Table 5.1). The scores were computed using MIURank and based on comparisons between the probe and c other randomly selected subjects from the training folds. The remaining comparisons, excluding those used for generating the probe biometric signature, were used to generate a biometric signature for each subject, which makes up the gallery (i.e. the soft biometric database). Then, the distance, d_c , between the probe and each subject in the gallery was calculated using the cosine distance measure as follows:

$$d_c = 1 - \frac{\sum_{i=1}^T X(i)Y(i)}{\sqrt{\sum_{i=1}^T (X(i))^2} \sqrt{\sum_{i=1}^T (Y(i))^2}} \quad (5.1)$$

where X is the biometric signature of the probe, Y is the biometric signature of the gallery subject that is compared to the probe, and $T = 24$ is the number of attributes composing the biometric signature. The subjects were sorted in ascending order based on their distances with the probe, and the rank of the correct match was used to report the identification performance via a CMC curve. This cross validation was run over the six folds and repeated until the harmonic mean of identification rates among all the ranks converged. Figure 5.12 shows the CMC curve resulting from this experiment.

As the results in Figure 5.12 show, a rank-10 identification rate of 95.17% can be achieved with the comparative facial soft biometrics using ten subject comparisons only, which is the ideal size of an identity parade [98]. The identification rate reaches 100% at rank-71, which implies that a correct match to the suspect is always guaranteed in the top 1.76% returned subjects (i.e. top 71 retrieved subjects out of the total 4038 subjects). When 15 comparisons are used in identification, a correct match will be guaranteed in the top

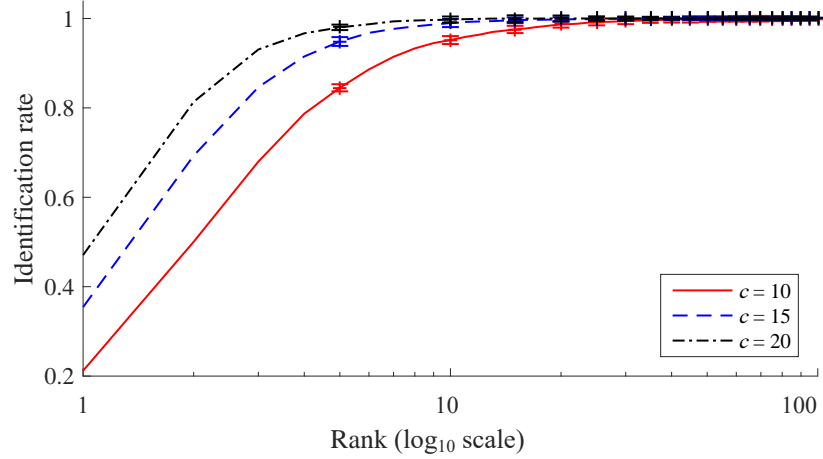


Figure 5.12: Identification performance with the LFW-V1 dataset using $c = \{10, 15, 20\}$ comparisons.

33 returned subjects; increasing the comparisons to 20 means that a correct match will always be found in the top 13 returned subjects (i.e. the top 0.3% retrieved subjects).

The performance of MIURank as a fully unsupervised algorithm that can be used to rank comparative soft biometrics has been demonstrated using the BioT dataset in chapter 4. However, it will be interesting to compare the impact of MIURank and the Elo rating system on identification performance using the LFW-V1 dataset, which is much larger and more challenging than the BioT dataset. Figure 5.13 shows the identification performance resulted from the biometric signatures generated using: (1) MIURank; and (2) the Elo rating system. The results reveal that MIURank outperforms the best outcome from the Elo rating system. These results highlight the effectiveness and the parameterless advantage of MIURank. Also, they demonstrate the scalability of MIURank for large datasets.

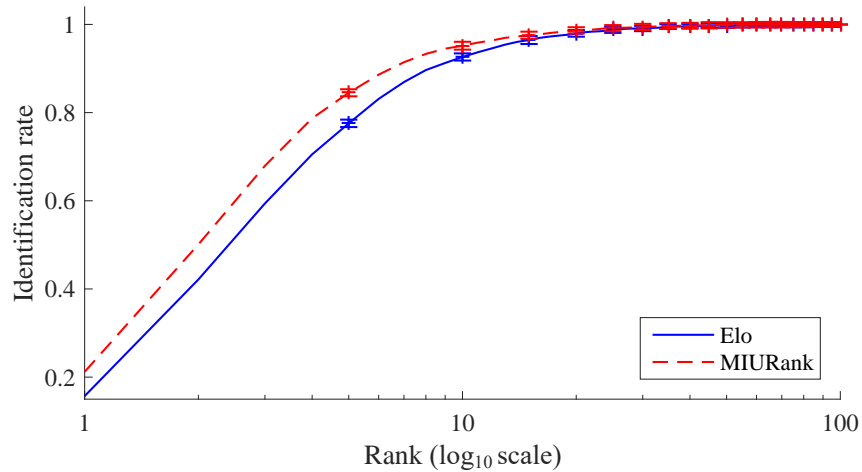


Figure 5.13: Identification performance in the LFW-V1 dataset with MIURank and the Elo rating system using $c = 10$ comparisons.

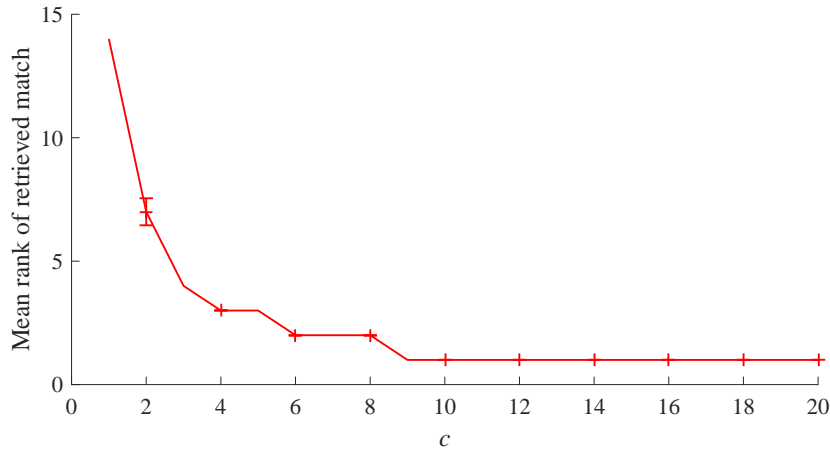


Figure 5.14: Effect of number of comparisons, c , on identification performance.

Figure 5.14 provides more insights into the relationship between the number of subject comparisons used for generating the probe biometric signatures, c , and the identification performance (in terms of the mean retrieved rank). It can be observed that the identification performance converges starting at nine comparisons. Also, the results show that the confidence level in the identification performance increases with the number of comparisons.

The identification performance can also be assessed from another perspective, which is the compression of the search range. Thus, narrowing down the search range becomes vital for the efficiency of identification in large databases. In addition, when eyewitness descriptions are not sufficiently accurate, search range compression can lead to filtering out a long list of suspects, making criminal investigations more efficient. Figure 5.15 shows the compression of the search range that can be achieved in the LFW-V1 dataset using the comparative facial soft biometrics. It shows that by using two comparisons only, the search range can be narrowed down to 12.88% of the total dataset with probability $p = 0.99$ that a correct match with the suspect will be found. Also, a correct match can be found in the top 3.44% returned subjects with probability $p = 0.9$ using two comparisons only. Furthermore, the results in Figure 5.15 indicate that the compression of the search range at different probabilities $p = \{0.9, 0.95, 0.99\}$ start to significantly converge from eight comparisons. Interestingly, this convergence agrees with the statement of the Police and Criminal Evidence Act [98] that an ideal identity parade consists of 8 to 12 persons. The results also suggest that the impact of the number of comparisons on the identification performance is more apparent when fewer comparisons are used in identification.

At the time of writing this thesis, the only work that has studied human identification using facial soft biometrics for both probe and gallery in a relatively large database is that of Klare et al. [54]. By comparing the identification performance obtained from this experiment with the results of Klare et al. in [54], which achieved a rank-1 accuracy

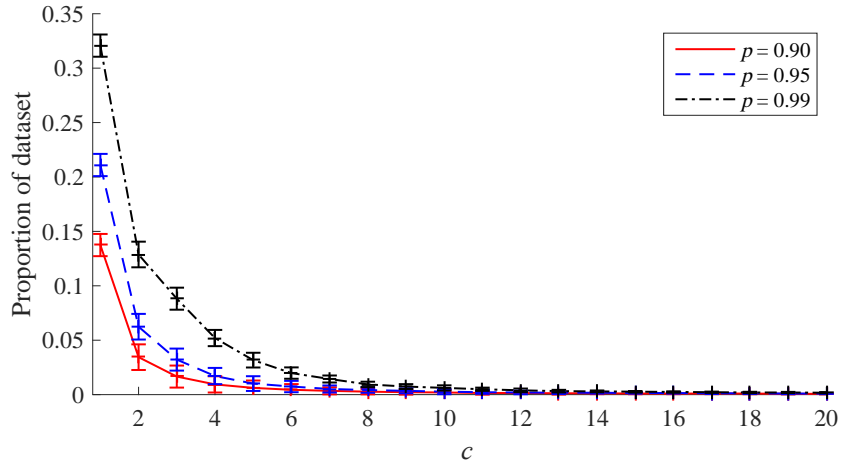


Figure 5.15: Compression of search range in the LFW-V1 dataset with different number of comparisons, c , and with probability $p = \{0.9, 0.95, 0.99\}$.

of 22.5% using 46 attributes (19 binary and 27 categorical) with 1196 subjects from the FERET database, there is an evident advantage to using comparative soft biometrics as compared to categorical soft biometrics. Thus, the comparative labels have resulted in a better identification performance, which is 30.2% at rank-1, with a larger and more challenging database, using 24 attributes and ten subject comparisons only. This advantage is attributed to the use of comparative labels.

5.4.2 Unconstrained Verification Using Facial Comparisons

Exploring unconstrained human verification using comparative facial soft biometrics has never been addressed previously, although it is essential to assess the performance of the comparative facial soft biometrics in verifying identities of people. Thus, it enables the exploration of the extent to which different eyewitnesses' descriptions of a suspect agree, which is a critical measure for the reliability of comparative facial soft biometrics. This experiment simulates a scenario in which two eyewitness descriptions for the same suspect are available, and the aim is to measure the accuracy of comparative facial soft biometrics when used to construct these eyewitness descriptions. To simulate such a scenario, two biometric signatures for each subject in the dataset were generated using MIURank from $c = \{10, 15, 20\}$ comparisons, such that the comparisons used to create both samples are mutually exclusive. The verification performance was then reported using the ROC curve that is shown in Figure 5.16, where it can be seen that the accuracy improves as the number of comparisons used in the verification increases.

Figure 5.17 shows the verification performance in terms of FPR and FNR curves, where the intersection of the two curves corresponds to the EER. With ten comparisons, the comparative facial soft biometrics can achieve an EER of 12.14%, which decreases to 8.52% when five more comparisons are used. The only known work that has studied

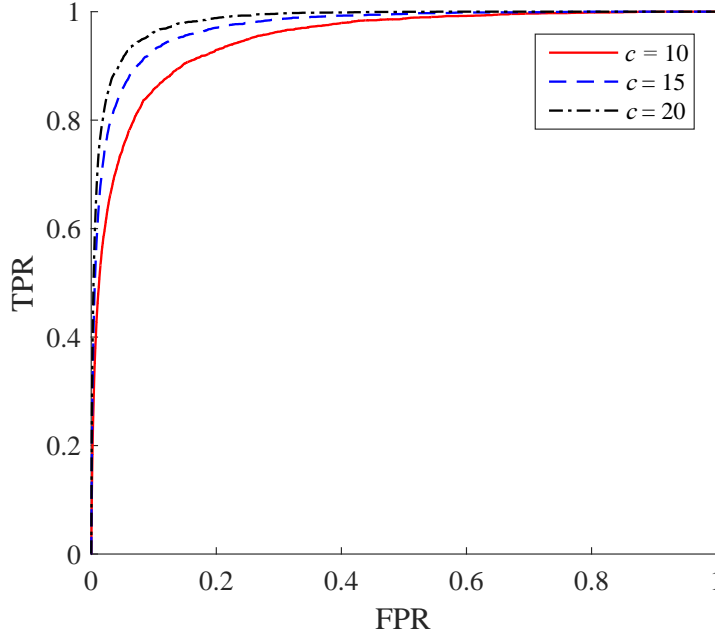


Figure 5.16: Verification performance of the comparative facial soft biometrics with the LFW-V1 dataset using $c = \{10, 15, 20\}$ comparisons.

attributes for face verification with the LFW database is that of Kumar et al. [22], in which 65 binary (categorical) attributes were used to train classifiers to predict the presence or absence of attributes, and the approach has achieved an EER of 14.83%. It is important to point here that the approach of Kumar et al. followed the standard LFW testing protocol [101], which involves training with View 1 and reporting performance with View 2. In addition, it used attribute classifiers that were trained with low level (hard) features, thus, it can predict the attributes of unseen subjects, while in this thesis, the View 1 training subset (LFW-V1) is used to explore identification through comparative soft biometrics for subjects that are already enrolled in the database (i.e. previously seen). Overall, the verification performance reveals the accuracy of comparative facial soft biometrics for matching eyewitness descriptions and highlights their effectiveness for unconstrained human face verification in general.

5.4.3 Attribute Contribution in Identification

The purpose of this experiment is to examine the impact of the number of attributes used in identification (i.e. feature set size) on the identification performance, and to assess the effectiveness of the three discriminative power assessment methods that were introduced in section 3.4.2.1 in measuring the importance of the attributes for identification. This experiment followed almost the same procedure that is described in section 5.4.1 with ten subject comparisons, the only difference is that while identification in section 5.4.1 was tested using all the 24 attributes (listed in Table 5.1), it was performed via three

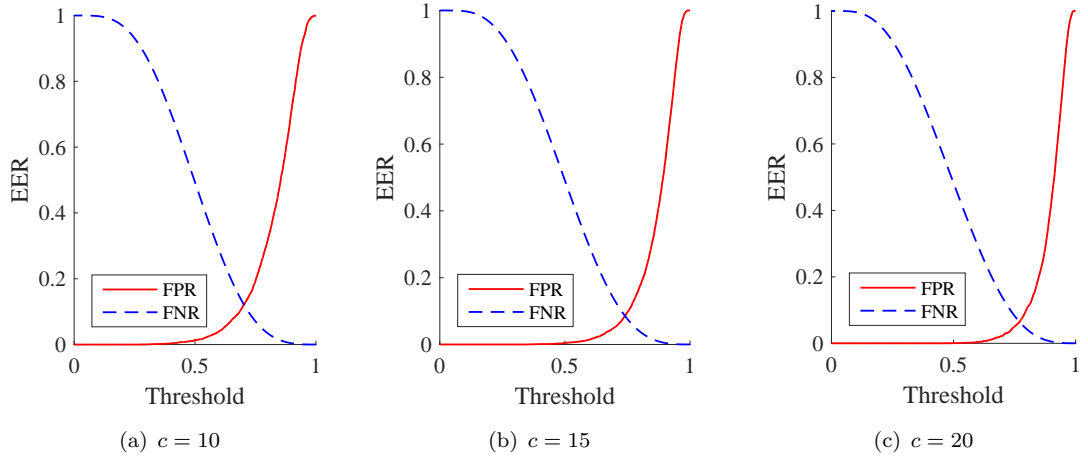


Figure 5.17: Error curves for the comparative facial soft biometrics in terms of FPR and FNR for $c = \{10, 15, 20\}$ comparisons: (a) $EER = 12.14\%$; (b) $EER = 8.52\%$; and (c) $EER = 7.71\%$.

parallel trials using three attribute sets that correspond to ANOVA, entropy and mutual information. In each of the three attribute sets, the 24 attributes were sorted in descending order according to their discriminative power. Then, identification was performed using the top k attributes, allowing k to vary between 2 and 24. The results of the experiment are summarised in terms of the mean rank of retrieved match as shown in Figure 5.18.

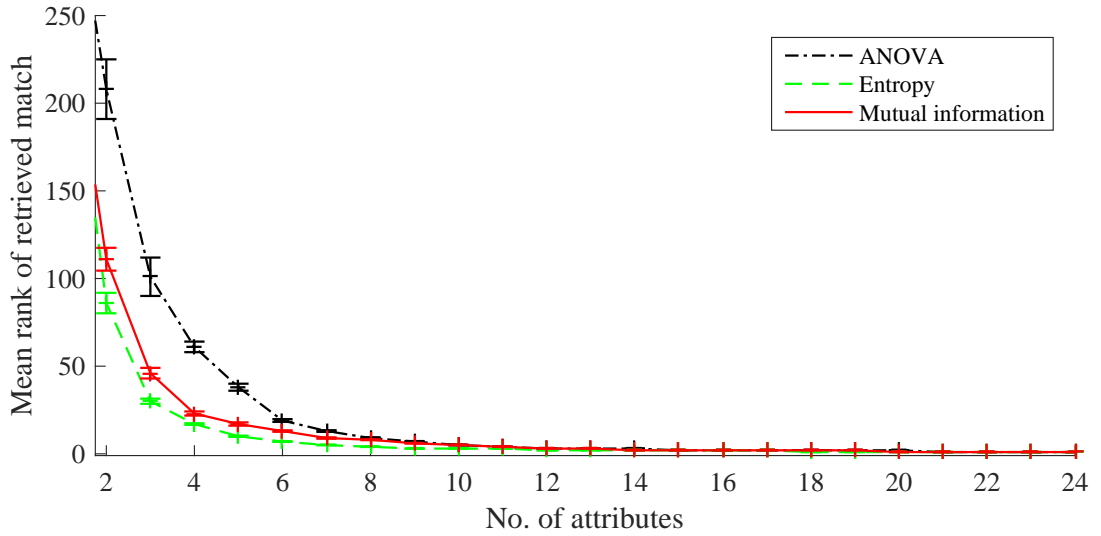


Figure 5.18: Identification performance with attributes' subsets ranked according to discriminative power based on: ANOVA, entropy and mutual information.

The results in Figure 5.18 show that five attributes are enough to yield a correct match at rank-10 on average and 21 attributes are enough to yield a correct match at rank-1 on average. Also, the results demonstrate the informative value of the attributes and the significant enhancement that can be achieved in identification accuracy with slight

increases in the number of attributes used (from 2 to 10 attributes). Moreover, the results reveal several important aspects. First, the attribute subsets generated based on entropy yield the best identification performance as compared with the other subsets (ANOVA and mutual information), which implies the robustness of entropy for assessing the discriminative power of comparative facial soft biometrics. Furthermore, starting from the 9th attribute, the impact of the mutual information attribute set on the identification performance gradually converges with that of entropy. Second, ANOVA seems to be significantly less robust for assessing the discriminative power of the attributes as compared to the other two methods (mutual information and entropy). The low robustness of ANOVA might be due to the tendency of the attribute distribution towards non-normality and the existence of outliers in the attributes (see section 5.3.1), which might affect the outcomes of ANOVA, as it assumes the normality of data [102]. Third, the differences among the three methods with respect to the identification performance start to diminish from the 11th attribute. Finally, the results reveal that the lowest three attributes in discriminative power based on the three methods are redundant in the identification performance, where the most common attributes with low discriminative power are *age* and *eyebrow hair colour* (refer to Figure 5.9).

5.4.4 Attribute Contribution in Verification

The attributes' significance in face verification was explored by following a similar experimental design to that outlined in section 5.4.2. In addition, three attribute sets were created, where each set consists of the 24 attributes ranked in descending order according to the discriminative power computed using ANOVA, entropy and mutual information respectively. Then, verification was performed with the three attribute sets using the top k attributes, allowing k to vary between 2 and 24. The experiment was performed using ten subject comparisons, and the verification performance is reported in terms of the EER.

Figure 5.19 shows the effect of attribute set size on verification performance. It can be seen from the results that ranking of attributes based on entropy yields the most robust feature subsets as compared to ANOVA and mutual information. Also, the results show that the three methods of evaluating the discriminative power of attributes are significantly different in their impact on the verification performance. Mutual information comes second in terms of robustness of features subsets, while ANOVA yields the least robust feature subsets. Again, the low robustness of ANOVA might be due to the non-normality of the attributes in addition to the dense presence of outliers (refer to section 5.3.1), which influence the outcomes of ANOVA. Also, the least three attributes in terms of discriminative power based on the three methods are redundant to the verification performance. The most interesting finding to emerge from Figure 5.19 is the consistency of the EER behaviour with respect to change in the number of attributes

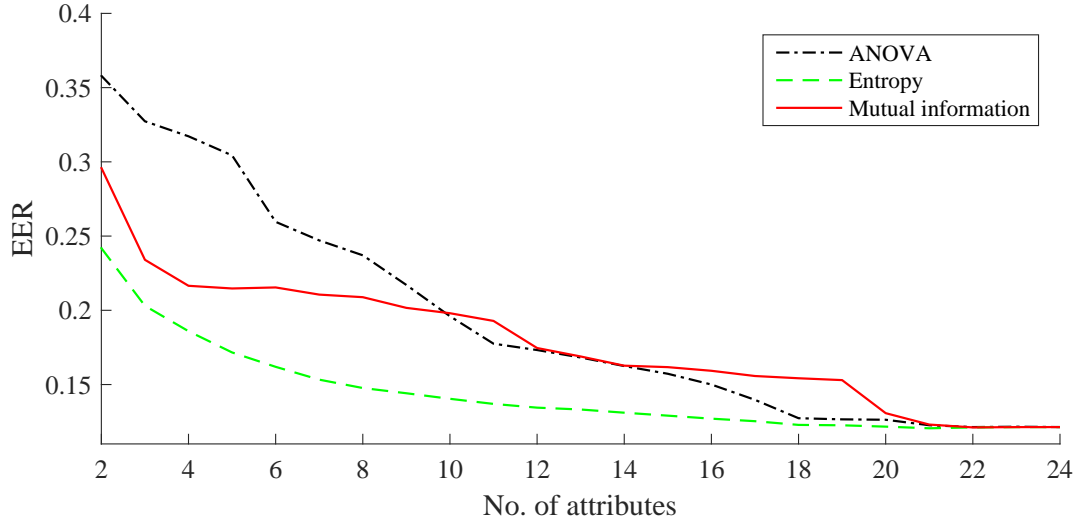


Figure 5.19: Verification performance with attributes' subsets ranked according to discriminative power based on: ANOVA, entropy and mutual information.

in case of entropy, whereas there are significant fluctuations in the case of ANOVA and mutual information.

Overall, the findings from sections 5.4.3 and 5.4.4 indicate that entropy is an effective method to assess the significance of the comparative facial soft biometrics in large unconstrained datasets, seeing that the resulted attribute subsets through entropy show the least reductions in both identification and verification performance as compared with ANOVA and mutual information. In addition, the behaviour of the entropy attribute subsets seems more robust and consistent relative to the ANOVA and mutual information subsets. The findings also suggest that ANOVA should be used carefully while considering the normality of attributes. Finally, the findings indicate that the most powerful 21 attributes in the LFW-V1 dataset are adequate to achieve the same performance as the 24 attributes, which implies the informative richness of the comparative soft biometrics proposed in this chapter.

5.5 Conclusions

In this chapter, the aim was to explore unconstrained human identification and verification through comparative facial soft biometrics in large and challenging databases. To achieve this, a large dataset that consists of 4038 subjects was extracted from the LFW database, and an enhanced set of comparative facial soft biometrics, which reflects the implications of the BioT experiments, was introduced. The acquisition of comparative labels was performed using crowdsourcing, and the statistical analysis of the generated biometrics signatures have revealed the challenging nature of the attribute data.

The identification experiment has shown that comparative facial soft biometrics can achieve a rank-10 recognition accuracy of 95.17% using ten comparisons only and a correct match can always be found in the top 1.76% returned subjects. Also, the verification experiment has shown that an EER of 12.14% can be achieved using the comparative facial soft biometrics. Furthermore, the assessment of attribute contribution and significance has indicated that the achieved identification and verification performance can be still maintained while the number of attributes used for identification is reduced from 24 to 21.

In general, the findings of this chapter imply the reliability of comparative facial soft biometrics for identification and verification in challenging and more realistic databases. They also demonstrate the scalability of comparative facial soft biometrics for relatively large databases. Altogether, the results pave the way for examining unconstrained human identification using other comparative soft biometrics such as body and clothing. The next chapter will investigate the extent to which the semantic gap between humans and machines can be bridged, in addition to the automatic retrieval of biometric signatures.

Chapter 6

Automatic Biometric Signatures

The results of the identification and verification experiments that were presented in the chapter 5 have demonstrated the effectiveness of comparative facial soft biometrics for subject identification in large unconstrained databases. The results have also revealed the potential embedded in the semantic space for human face identification and verification using comparative soft biometrics. Whereas the experiments of the previous chapters have addressed identification and verification in soft biometric databases that were constructed from human generated labels (i.e. crowdsourced comparisons), identification in criminal investigations also involves searching databases of different modalities such as mugshots or CCTV footage [16] to identify an unknown suspect based on verbal descriptions from eyewitnesses. Accordingly, this requires a framework for automatically generating biometric signatures from face images, and creating a new visual space that narrows the semantic gap between humans and machines with respect to relative facial attributes.

The objective of this chapter is to investigate the automatic retrieval of biometric signatures from imagery databases and to assess the potential of searching a database of automatically retrieved biometric signatures (visual database hereafter) using verbal descriptions. The chapter introduces a framework for extracting visual features from face images and proposes two different approaches for automatically retrieving biometric signatures from images as shown in Figure 6.1. The first approach is based on retrieving scores of relative attributes from visual features (REL), while the second approach suggests the automatic estimation of comparative labels from visual features (ECL) and producing biometric signatures from them using the MIURank algorithm. Moreover, the chapter analyses the accuracy of the automatically estimated soft biometric attributes, and assesses the correspondence between the semantically and visually generated attributes. Also, the chapter presents identification experiments that use both approaches (REL and ECL) and explores the impact of the visually estimated attributes on face verification.

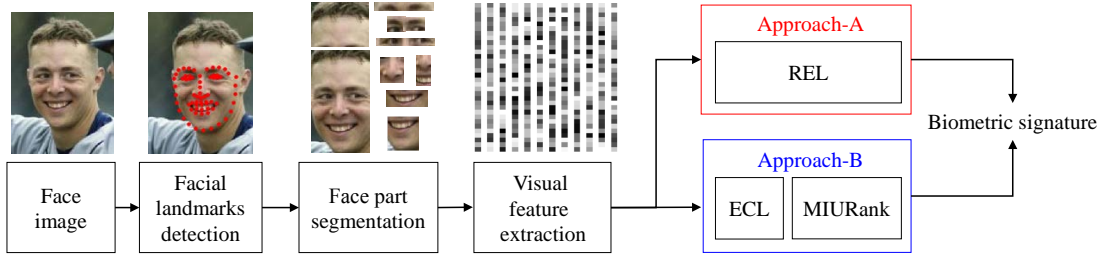


Figure 6.1: Automatic retrieval of biometric signatures from face images.

The experiments in this chapter use the LFW-MS4 dataset that consists of 430 subjects from the LFW-V1 dataset with 1720 sample images. The basis of selection for the subjects of LFW-MS4 is that they are represented by at least four sample images in the LFW-V1. Similarly to the previous approach in section 5.2, the samples of LFW-MS4 were all aligned using deep funnelling [100] and normalised to an inter-pupil distance of 50 pixels. From the 1720 samples, four galleries (databases) were constructed by randomly assigning a sample of each subject to one of the galleries, which resulted in 430 samples per gallery. Also, the biometric signatures of the LFW-MS4 subjects, which are used as ground truth scores in this chapter, were constructed by extracting all the relevant human-annotated comparisons between the 430 subjects of LFW-MS4 and using MIURank as explained earlier in section 5.3.

6.1 Extracting Visual Features from Face Images

6.1.1 Facial Landmarks Detection and Parts Segmentation

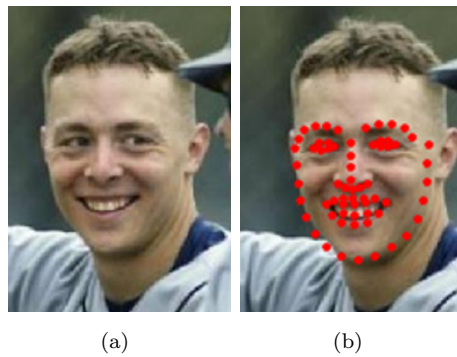


Figure 6.2: Localization of facial landmarks using a face alignment framework [4] for a sample face image from the LFW-MS4 dataset.

Extracting visual features from a face image requires locating landmarks (key points) that enable segmenting the face parts, which correspond to the soft biometric attributes listed in Table 5.1. Finding landmarks' locations on face images can be automatically achieved by using a deformable model that interprets face images and estimates the



Figure 6.3: Segmented face parts with the corresponding attributes indices as described in Table 5.1.

locations of the most important facial landmarks such as nose base and eye pupils. The most well known deformable models that have been used to estimate face landmarks are the Active Shape Models (ASMs) [103, 104], Active Appearance Models (AAMs) [105, 106] and the Constrained Local Models (CLMs) [107, 108]. These models are based on flexible templates and exploit global shape constraints derived from training data to locate a set of facial landmarks on a new face image. Given a statistical shape model that consists of the points making up the shape obtained from training images, an ASM iteratively fits a shape model to a new image by increasing the match between the image and the model. The AAMs combine both shape and texture to match a model to an image, and have demonstrated more robustness as compared to the ASMs [109]. The CLMs follow a part-based approach in which a face image is sampled into a set of regions and the corresponding response maps, which represent the likelihood of having a particular landmark point at each pixel, are generated. Then, the parameters of the shape model are tuned to find the set of points that optimised the cost of the response maps.

In this chapter, a more recent and efficiently trained framework for face alignment in-the-wild [4] was used to locate facial landmarks and, accordingly, enable the segmentation of face parts. This framework combines the AAM approach with response maps that are generated by trained local detectors. This approach has demonstrated its capability of handling face alignment in-the-wild. In addition, it has shown a high efficiency. This face alignment framework was utilised to locate 66 points (shown in Figure 6.2) on the LFW-MS4 dataset samples, which guided the segmentation of the face parts that correspond to the soft biometric attributes as shown in Figure 6.3. The segmented face parts for each sample image in the LFW-MS4 dataset were used to generate visual features, as explained in the next section.

6.1.2 Generating Visual Features

Visual features were generated from the segmented face parts images (shown in Figure 6.3) with the aim of training models for predicting the attributes as will come later in this chapter. Here, four different categories of visual features were investigated for estimating comparative facial soft biometrics, which are: *shape*, *spatial*, *texture*, and *colour*. From each of these categories, a visual descriptor was selected by virtues of popularity and sophistication to conduct the analysis and experiments as follows:

- **Histogram of Oriented Gradients (HOG)** [110] is a *shape* descriptor chosen because it encodes the details of a shape in terms of local distribution of edge orientations. The HOG features are extracted by computing the gradients of the image, then dividing the image into cells of 8×8 pixels, and encoding the edge orientations of each 2×2 cells (or block) into a 9-bin histogram. The blocks are allowed to overlap by one cell, and the histograms resulting from the blocks are concatenated to form a visual features vector*.
- **GIST** [111] is a *spatial* descriptor selected because it can provide a low-dimensional characterization for a scene. The extraction of GIST features involves intensity normalisation for the image and processing it through a series of Gabor filters in four scales and eight orientations per scale that yield 32 orientation maps. Each orientation map is divided into a 4×4 grid, and the average intensity is calculated for each block in the grid to form a vector of 512 features. The parameters used to generate the GIST features are the implementation defaults**.
- **Uniform Local Binary Pattern (ULBP)** [112]. The ULBP features are used because they are now one of the most popular approaches to represent the local *texture* information from a grey scale image. A local binary pattern is generated for each pixel in the image by thresholding its 8-connected neighbours. Then, the number of 0 to 1 (or vice versa) transactions is counted in the resulting binary pattern and mapped to one of 59 possible uniform patterns. Accordingly, a histogram of 59 bins is generated for the image, which encodes the texture information in terms of the frequency of each uniform pattern. The ULBP features are invariant to rotation and monotonic changes in illumination. Given that there are many new variants of LBP [113, 114, 115], the analysis here is sufficient to show the suitability of texture discrimination by a standard form of LBP though there are advanced versions which might improve performance further.
- **Deformation and Viewpoint Invariant Colour Histograms (DVICH)** [116] are *colour* descriptors that are based on utilising gradients in different colour channels to weight pixels, which results in features that are invariant to local affine

*The parameters are MATLAB R2015b package defaults.

**The MATLAB code is available on: <http://people.csail.mit.edu/torralba/code/spatialenvelope>

transformations. The colour space sampling is performed for each of the three colour channels (red, green and blue) by mapping the pixel values, which are weighted by the gradients, to five bins*. This process results in a histogram of 5^3 bins, which represent the DVICH features of the image.

6.2 Retrieval of Biometric Signatures

The previous section explained the extraction and generation of visual features that are used for retrieving biometric signatures from face images. In this section, two different approaches are presented for ranking the soft biometrics by predicting the scores of the subjects, which are used to infer biometric signatures from images as described in Figure 6.1. The first, REL, is a pointwise approach that predicts scores directly from visual features; the second, ECL, is a pairwise approach that predicts comparative labels, which are used respectively to generate biometric signatures through the MIURank algorithm that was presented in chapter 4. The following sections detail each approach and evaluate its accuracy in predicting soft biometric attributes.

6.2.1 Relative Rating of Visual Features (REL)

This approach addresses the ranking of soft biometric attributes from a face image (i.e. visual features) as a pointwise problem, and makes use of regression to predict scores that construct biometric signatures. The exploration of the REL approach involves the examination of three different regression techniques, which are: (1) Multi Linear Regression (LR); (2) Regression Trees (RT); and (3) Support Vector Machines for Regression (SVR). These three regression techniques are described in the following sections.

6.2.1.1 Multiple Linear Regression (LR)

The aim of multiple linear regression is to model a linear relationship between a set of independent variables (features) and the corresponding dependent variables (labels), and thus enable the prediction of the label for a new (unseen) sample based on its features. Given a training set of n samples that consists of a feature space, $x_i \in \mathbb{R}^m$, $1 \leq i \leq n$, with the corresponding dependent variables, $y \in \mathbb{R}^n$, a linear model that represents the relationship between the features and the corresponding labels can be fitted as follows:

$$y = \beta x + \epsilon \quad (6.1)$$

*The MATLAB code is available on: <http://www.cs.umd.edu/~domke/histograms/>

where ϵ is the random error term, $x = \langle 1, x_1, \dots, x_m \rangle$ is the feature vector, and $\beta = \langle \beta_0, \beta_1, \dots, \beta_m \rangle$ is a vector of weight coefficients that is estimated based on the following objective function:

$$\min \left(\sum_{i=1}^n (y_i - f(x_i))^2 \right) \quad (6.2)$$

where y_i is the value of the dependent variable (label) for the sample i , and $f(x_i) = \beta x_i$ is the fitted value for the sample i . Accordingly, the response, \hat{y}_j , for a new (unseen) sample, j , can be predicted from its features based on the following equation:

$$\hat{y}_j = \beta x_j \quad (6.3)$$

6.2.1.2 Regression Trees (RT)

Regression can also be achieved via decision trees [117], which are a supervised model that learns simple decision rules from a training feature space, $x_i \in \mathbb{R}^m$, $1 \leq i \leq n$, and its associated labels, $y \in \mathbb{R}^n$, to predict the value of a new (unseen) target variable from its features. Regression trees are created by a recursive partitioning process starting with a region R that contains the whole feature space. Then, in each iteration, and for each feature $1 \leq j \leq m$, a split point, s , is selected based on the following objective:

$$\min \left(\sum_{x_{ij} \leq s} (y - y_i)^2 \right) + \min \left(\sum_{x_{ij} > s} (y - y_i)^2 \right) \quad (6.4)$$

where the sample i is in the region, and y is the fitted response value for the region. This process is repeated recursively with the child nodes until the maximum number of splits, which has a default of $m - 1$, is reached, or each region has one sample in it*. For a new (unseen) sample, k , the predicted response, \hat{y}_k , is found by following the decision rules set by the regression tree with the feature vector of the new sample, x_k .

6.2.1.3 Support Vector Machines for Regression (SVR)

Support Vector Machines for Regression (SVR) is a machine learning tool that is effective in high dimensional spaces and enjoys high versatility by using various kernel methods for adapting to different feature spaces [60]. Given a training set of n samples that consists of the feature space, $x_k \in \mathbb{R}^m$, $1 \leq k \leq n$, and the corresponding labels, $y \in \mathbb{R}^n$, the objective of the SVR model is to find the coefficients, α_k and α_k^* for each training sample, k , that minimise the following loss function:

*The parameters are MATLAB R2015b package defaults.

$$\begin{aligned}
L(\alpha) = & \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) K(x_i, x_j) \\
& + \epsilon \sum_{i=1}^n (\alpha_i + \alpha_i^*) - \sum_{i=1}^n y_i (\alpha_i - \alpha_i^*)
\end{aligned} \tag{6.5}$$

subject to

$$\begin{aligned}
& \sum_{k=1}^n (\alpha_k - \alpha_k^*) = 0 \\
0 \leq \alpha_k \leq C \quad \forall \quad 1 \leq k \leq n \\
0 \leq \alpha_k^* \leq C \quad \forall \quad 1 \leq k \leq n
\end{aligned} \tag{6.6}$$

where C determines the penalty of misclassification out of the margin ϵ , while $K(x_i, x_j)$ is the radial basis kernel function [118] that is defined for two samples x_i and x_j as:

$$K(x_i, x_j) = \exp \left(- \frac{\|x_i - x_j\|^2}{\sigma^2} \right) \tag{6.7}$$

where σ is the scaling factor in the kernel function $K(x_i, x_j)$. Accordingly, the label of a new sample x can be predicted from its features based on the following function:

$$f(x) = \sum_{k=1}^n (\alpha_k - \alpha_k^*) K(x_k, x) + b \tag{6.8}$$

where b is the bias term that is used to fit the training data optimally.

6.2.1.4 Analysis and Evaluation

The objective of this analysis is to explore the accuracy of the proposed REL approach in ranking the comparative soft biometric attributes (shown in Table 5.1). Furthermore, this analysis aims to determine the correspondence of the scores estimated from visual features (visual space) with those inferred from the crowdsourced comparative labels using the MIURank algorithm (semantic space). The analysis discovers the prediction accuracy of each of the three regression models introduced in this section along with each of the visual descriptors listed in section 6.1.2. The process follows the flow outlined in Figure 6.1 with the REL approach. The ground truth scores used in this analysis are those estimated from the crowdsourced comparative labels for the LFW-MS4 dataset

using the MIURank algorithm as explained in section 4.3.2.2. The experimental design is based on a 10-fold cross validation and applied to each of the soft biometric attributes. The 430 subjects of LFW-MS4 were randomly divided into ten folds, where each fold consists of 43 subjects. It is important to emphasise that the subjects in the test fold and the training folds were mutually exclusive (i.e. non-overlapping sets). The regression model was trained using the visual features of the training subjects with their four samples, in addition to the corresponding normalised scores deduced using MIURank from crowdsourced comparative labels. Finally, the scores of the retrieved attribute were inferred for each sample of each test subject using the trained regressors and the four visual descriptors introduced in section 6.1.2, while the performance is reported as the grand average of the outcomes across the ten test folds and the four samples of each test subject.

The accuracy of the scores that were predicted from visual features and their correspondence with the semantically generated ground truth are reported using the following two measures:

- **Mean Absolute Error (MAE).** The mean absolute error for an attribute's predictions is computed for each test fold as the absolute difference between the ground truth scores, y , and the corresponding predicted scores from visual features, \hat{y} , for the subjects of the test fold as follows:

$$\text{MAE} = \frac{\sum_{i=1}^l |y_i - \hat{y}_i|}{l} \quad (6.9)$$

where $1 \leq i \leq l$ is the test subject index, and l is the number of subjects in the test fold. This measure was calculated for each of the ten test folds and reported as the average among the ten folds.

- **Level of concordance.** The purpose of this measure is to assess the accuracy of the evaluated approach in ranking test subjects according to attribute strength (i.e. score). Given a ground truth rank, r , and a predicted rank, \hat{r} , for a pair of subjects, i and j . The pair is considered concordant if and only if the following condition is satisfied:

$$r_i > r_j \quad \text{and} \quad \hat{r}_i > \hat{r}_j \quad (6.10)$$

or

$$r_i < r_j \quad \text{and} \quad \hat{r}_i < \hat{r}_j \quad (6.11)$$

whereas all the other possible relations between the ground truth rank and the predicted rank for a subject pair are considered discordant. For a test subject,

the level of concordance can be defined as the ratio between the concordant pairs and all the pairs that involve the test subject with the subjects of the training folds. As compared with other ranking evaluation measures such as Kendall's τ and Spearman's ρ , the level of concordance gives more emphasis on the testing set, as it is concerned only with rank changes among pairs that involve a test subject. On the other hand, Kendall's τ and Spearman's ρ are more generic and give no prominence to the pairs that include test subjects. The level of concordance was determined for each sample of each test subject in the LFW-MS4 dataset and reported as the average of the ten folds.

		Technique		
		LR	RT	SVR
Descriptor	HOG	19.0%	20.8%	14.7%
	GIST	20.8%	21.3%	14.8%
	ULBP	15.1%	21.4%	14.6%
	DVICH	22.0%	20.4%	14.8%

Table 6.1: Mean absolute error for the predicted scores.

		Technique		
		LR	RT	SVR
Descriptor	HOG	68.0%	66.8%	74.9%
	GIST	67.3%	67.7%	75.1%
	ULBP	74.3%	67.4%	74.8%
	DVICH	66.6%	67.7%	74.6%

Table 6.2: Mean level of concordance for the predicted ranks.

The outcomes of the prediction analysis are shown in Tables 6.1 and 6.2 averaged for the soft biometric attributes (listed in Table 5.1). The two tables show that the SVR yielded the lowest MAE and the highest level of concordance when used with each of the four visual descriptors. It also shows that the GIST descriptor resulted in the best overall performance in terms of level of concordance, and accordingly, reveals a higher visual-semantic correspondence as compared with the other descriptors. Therefore, throughout this chapter, the SVR model with the GIST features were used together to conduct the analysis and experiments for the REL approach, which is based on ranking from visual features. The misclassification penalty, C , of the SVR model was chosen through cross validation on the training set by varying C between 21 logarithmically spaced values that range between 2^{-10} and 2^{10} , while the scaling factor was set as $\sigma = 1$ *.

Figure 6.4 shows the accuracy of the REL approach in predicting scores of the individual attributes in terms of MAE. The results show that *spectacles* and *facial hair* have the highest MAE, which might indicate the unreliability of transforming binary-like facial attributes (e.g., *spectacles* and *facial hair*) to the relative continuous space based on visual features. On the other hand, the results show that *eyebrow hair colour* has the

*The parameter is MATLAB R2015b package default.

least MAE, which is probably due to the precision of machines in extracting colour information relative to the higher level interpretation of colour by humans [119]. Also, the result shows that *nose width* has the second lowest MAE, which may be due to its expression invariance capabilities [82].

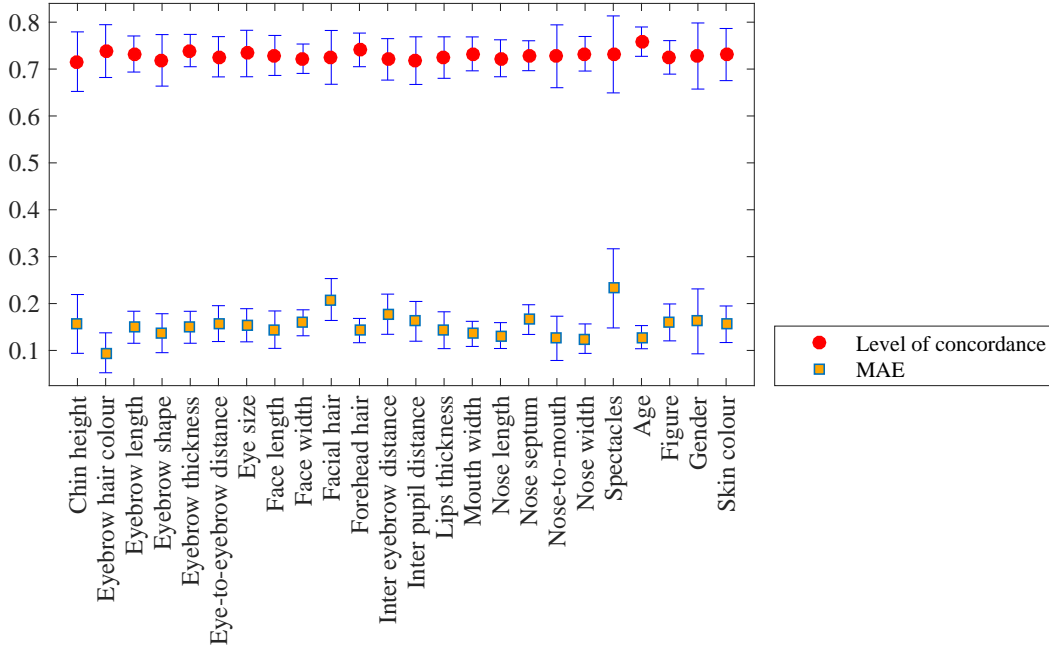


Figure 6.4: Correspondence between semantic and visual scores resulting from the REL approach.

The correspondence between the predicted visual attributes and the ground truth semantic attributes in terms of the level of concordance can be observed in Figure 6.4, from which it can be seen that *age* and *eyebrow thickness* have the highest level of concordance as compared with the other attributes. Interestingly, *age* also has a relatively low MAE, and this is likely to be related to the efficiency of machines in learning relative age estimates from face images. Also, *spectacles*, which has the least accuracy in terms of MAE, shows the least level of concordance in Figure 6.4. Another interesting finding from the level of concordance results is the high variance associated with the binary-like features such as *spectacles*, *gender* and *facial hair*, which reveals the uncertainty associated with the estimations of binary-like attributes in the relative continuous space.

The accuracy of the REL approach can also be assessed from an intra-space perspective by determining the correspondence between multiple samples of each subject in the visual space. In this analysis, four galleries were constructed in association with the four samples of the LFW-MS4 dataset and using the automatically predicted biometric signatures to assess the visual correspondence of the attributes. Then, the visual correspondence was evaluated between all the possible pairs of the four galleries using the

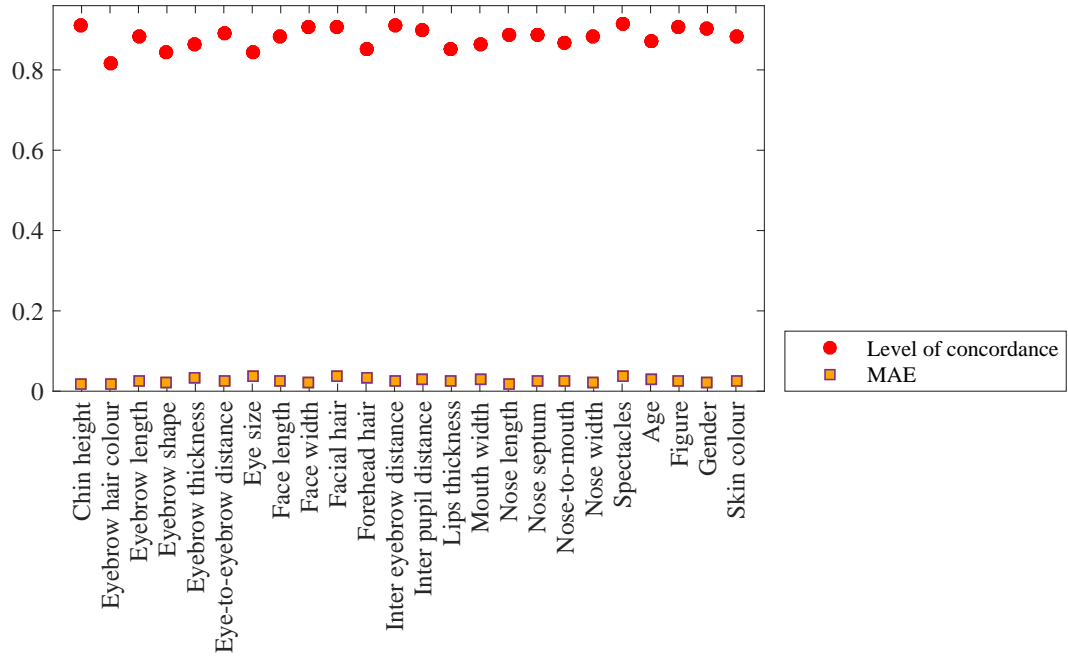


Figure 6.5: Visual correspondence among scores resulting from the REL approach.

same measures applied with the semantic correspondence: MAE and level of concordance. Figure 6.5 shows the results of the visual correspondence analysis, where it can be noted that the ranking of the attributes in terms of their visual correspondence is not consistent between the MAE and level of concordance. The most important aspect of Figure 6.5 is the presence of outliers with the binary-like facial attributes (i.e. *facial hair* and *spectacles*), in addition to *eye size*. With respect to ranking accuracy, the level of concordance results in Figure 6.5 imply that the binary-like attributes, which are *spectacles*, *facial hair* and *gender*, have the highest visual correspondence. Interestingly, this contradicts with the semantic correspondence of these attributes (shown in Figure 6.4), which demonstrates that the binary-like attributes have a relatively low concordance. The high visual correspondence of the binary-like attributes might be due to the accuracy of estimating their scores in the same space (i.e. visual space). This high correspondence of the binary-like attributes has also been noted when the semantic stability was assessed for the attributes in the semantic space in section 5.3.2.2, where the binary-like features showed a relatively high stability. In summary, the results of the visual correspondence demonstrate the consistency of the REL approach in estimating scores among different samples.

6.2.2 Estimation of Comparative Labels (ECL)

The second approach proposed in this chapter for automatically retrieving biometric signatures is based on the estimation of comparative labels from visual features (ECL), and correspondingly, using the estimated labels for ranking soft biometrics through MIU-Rank, which composes biometric signatures as illustrated in Figure 6.1. The estimation of comparative labels from visual features is achieved based on an intuitive concept, which is associating an attribute comparison between two subjects with the difference of the visual feature vectors of the two subject (i.e. differential features). The estimation of comparative labels from differential visual features is explored in binary relative format, in which only the "Less" and "More" semantic labels are considered, while the "Same" semantic label is ignored because it is ineffective in the MIURank ranking process as explained previously in section 4.2. The estimation of comparative labels from differential visual features was achieved using regression, where the regressor was trained using a set of differential visual features and the corresponding numeric comparative labels (-1 for "Less" and 1 for "More"). Then, a comparative label, l_c , between a new subject, i , in the test fold and another subject, j , in the training fold can be estimated from the corresponding differential visual features, and using the regression model based on the sign of the estimated dependent variable, \hat{y}_c , as follows:

$$l_c(i, j) = \begin{cases} -1 & \text{if } \hat{y}_c < 0 \\ 1 & \text{if } \hat{y}_c \geq 0 \end{cases} \quad (6.12)$$

The main reason behind the selection of regression for estimating comparative labels is that it embeds the labels' order (i.e. -1 for "Less" and 1 for "More"). As a result, it can also be extended to estimate comparative labels with more levels (e.g., 3 or 5 levels).

6.2.2.1 Analysis and Evaluation

The ECL approach was investigated using the regression models described earlier in section 6.2.1 and the visual descriptors listed in section 6.1.2. The evaluation used the 430 subjects of the LFW-MS4 dataset and followed a 10-fold cross validation. To enable a consistent comparison with the performance of the REL approach, the subject to fold assignment used in this analysis was similar to that of the REL approach, which is described in section 6.2.1.4. In each of the ten test iterations, the subjects of the nine training folds were used to train the regression model by deriving all the comparisons (comparative labels) between them along with the corresponding differential visual features extracted from the four samples available for each of the training subject. Then, a comparative label, l_c , between a subject in the test fold and other subjects in the training fold was estimated from the corresponding differential visual features based on Equation 6.12.

The process was repeated with each of the ten folds and includes each of the 24 attributes listed in Table 5.1. The evaluation covered all the possible combinations of regressors and visual descriptors. Table 6.3 shows the prediction accuracy averaged over the ten folds and the 24 attributes. The results reveal that multiple linear regression (LR) and the GIST differential visual features combination produces the best accuracy as compared with the other combinations of models and visual descriptors. Accordingly, this combination was used for further analysing the ECL approach and conducting the retrieval experiments throughout this chapter.

		Technique		
		LR	RT	SVR
Descriptor	HOG	78.77%	56.03%	54.44%
	GIST	79.65%	55.92%	68.27%
	ULBP	58.35%	53.90%	51.45%
	DVICH	56.91%	58.25%	55.77%

Table 6.3: Average prediction accuracy of comparative labels.

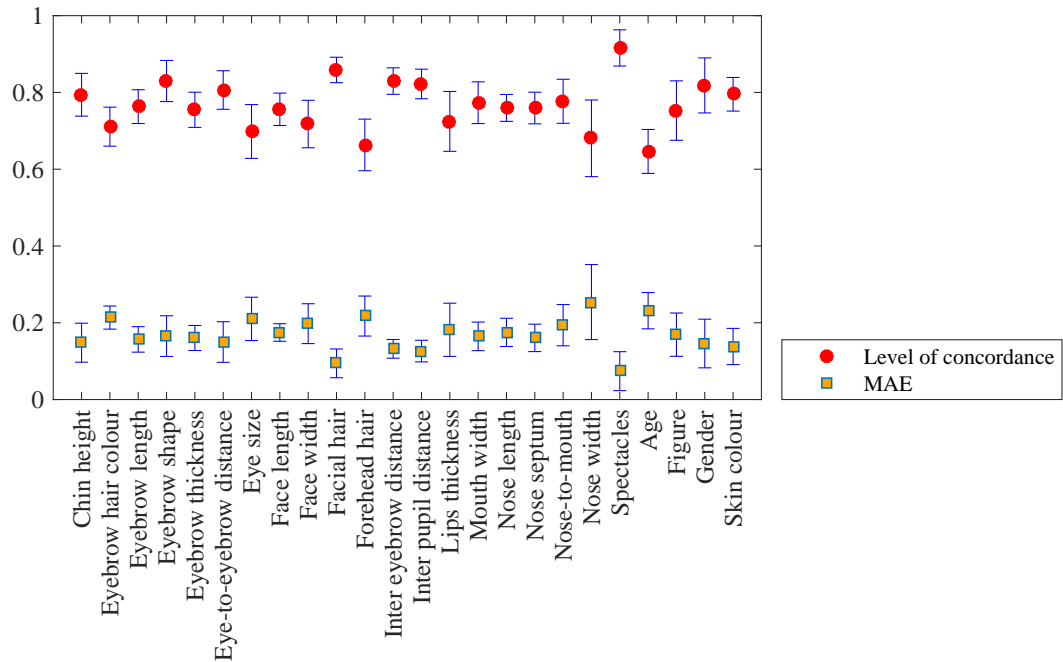


Figure 6.6: Correspondence between semantic and visual scores resulting from the ECL approach.

Similar to what was undertaken with the REL approach in section 6.2.1.4, the correspondence between semantic and automatic comparative labels is assessed using MAE and level of concordance that are shown in Figure 6.6. The correspondence analysis reveals several interesting findings. First, there is a level of agreement in the attributes' rankings according to MAE and level of concordance. Second, the binary-like attributes (*spectacles*, *facial hair* and *gender*) show the highest correspondence with the semantic space, and this might be attributed to the accuracy of the model in classifying binary

classes as compared with the other attributes, which are more of multi-class nature. Third, the attributes that are the least correspondence with the semantic space (i.e. *age*, *forehead hair* and *eyebrow hair colour*) have also been found to have low semantic stability in the analysis presented in section 5.3.2.2.

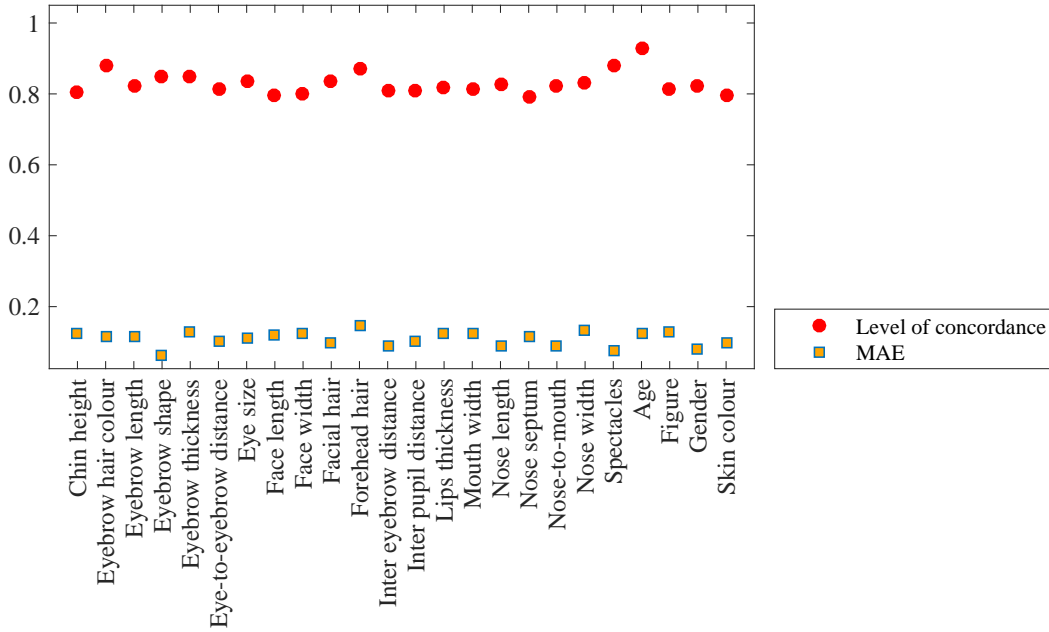


Figure 6.7: Visual correspondence among scores resulting from the ECL approach.

The visual correspondence of the ECL approach predictions was assessed in the same way as that was used with the REL approach in section 6.2.1.4. The results in Figure 6.7 show that the attributes significantly vary in their visual correspondence. However, *eyebrow shape* and *spectacles* seem to have a high visual correspondence. What is surprising from the results of visual correspondence is that while *age* predictions have low accuracy in terms of MAE, it has the highest level of concordance. This also applies to *forehead hair*, which has a high MAE and a low level of concordance. These disagreements between the two measures can be attributed to the sensitivity of MAE for outliers, and accordingly, these inconsistencies leverage the impact of the level of concordance measure, as it is a rank-based measure that has better robustness against outliers [120, 121].

By comparing the findings from the correspondence analysis of the REL approach in section 6.2.1.4 with the finding of the ECL approach, it can be noted that the two approaches significantly differ in their evaluations for the attributes' correspondence. This can be attributed to the difference between the two approaches in addressing the generation of biometric signatures. Thus, while the REL approach ranks an attribute based on the visual features of the corresponding subject, the ECL approach only estimates

the comparative label using differential visual features of a pair of subjects, and delegates the ranking to the MIURank algorithm. Therefore, the noted association between the outcomes of the correspondence analysis of the ECL approach, and the semantic stability analysis of the LFW dataset in section 5.3.2.2 might be due to the impact of MIURank.



Figure 6.8: Examples for the outcomes of ranking of: (a) *age*, (b) *gender*, (c) *eyebrow thickness* and (d) *nose width* on the LFW-MS4 dataset. For each of the shown attributes, the lower and upper rows represent the least and top ranked subjects respectively, while the columns represent the different methods of ranking the attribute.

Figure 6.8 shows examples for the outcomes of ranking some attributes using the three different methods: (1) MIURank in the semantic space, which forms the ground truth for the experiments and analysis in this chapter; (2) ECL; and (3) REL. Further, Figure 6.9 presents the subjects that have the weakest and strongest presence of overall attributes based on the three ranking methods, where the strength of the attributes is calculated as the median value of the scores of the attributes (i.e. the median of the biometric signature of a subject).

Where subjects sorting by gender becomes a key aspect for increasing search efficiency in databases of law enforcement agencies [41], the correspondence between relative and binary gender was also assessed by clustering the subjects of the LFW-MS4 dataset

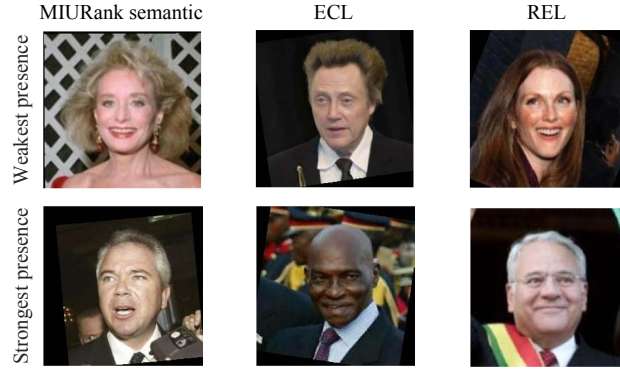


Figure 6.9: The LFW-MS4 subjects with the weakest and the strongest overall presence of attributes.

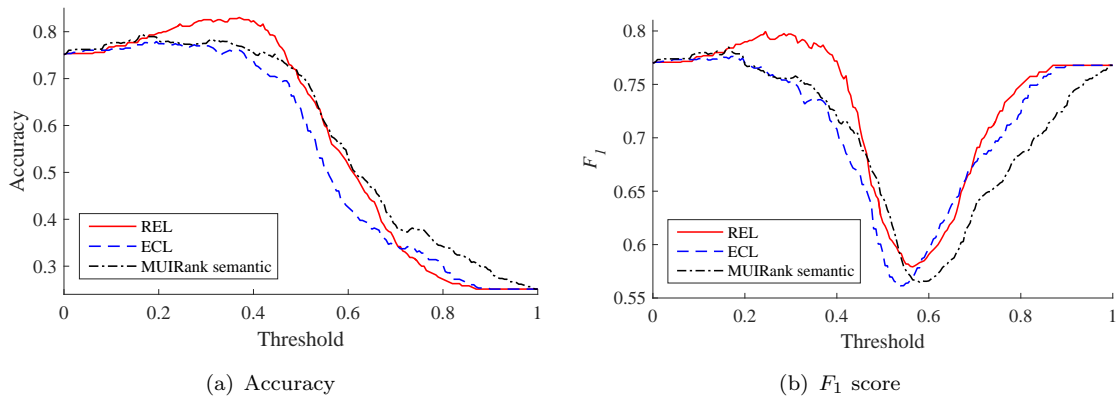


Figure 6.10: Gender clustering based on relative gender scores.

based on relative gender to *feminine* and *masculine*, and determining the accuracy of the resulting clusters with respect to the binary ground truth that was collected from human annotators via crowdsourcing. The clustering of the subjects by their relative gender is based on a threshold value $0 \leq T \leq 1$, such that the subjects with relative gender scores below T are classified as *feminine*, while the subjects with scores greater than or equal to T are classified as *masculine*. The accuracy of such categorization was evaluated with the scores resulting from both the REL and ECL approaches, in addition to the scores generated from crowdsourced labels (semantic MIURank). The results are shown in terms of accuracy and F_1 score in Figure 6.10, where it can be seen that the performance in terms of the F_1 score of the threshold-based clustering of relative gender reaches its maximum value of 79.94% at the threshold value $T = 0.245$ with the REL approach; this is mostly due to the formation of the LFW-MS4 dataset, which consists of 24.8% females and 75.2% males. Moreover, the results in Figure 6.10 shows that the relative gender scores that were predicted by the REL approach yield the highest overall correspondence with the binary gender. This might indicate the effectiveness of the SVR model in the REL approach in learning features that are significant for binary gender clustering.

	semantic-visual		visual	
	REL	ECL	REL	ECL
MAE	0.1524±0.0284	0.1681±0.0413	0.0265±0.0006	0.1094±0.0201
Concordance	0.7299±0.0009	0.7670±0.0635	0.9083±0.0105	0.8303±0.0327

Table 6.4: Summary for the correspondence analysis.

To summarise the findings from this section, Table 6.4 provides an overview of the correspondence analysis of the REL and ECL approaches. The results of semantic-visual correspondence show that although the REL predictions are more accurate in terms of MAE, the ECL predictions have a higher level of concordance with the semantic space. With regards to visual correspondence, the results reveal that the REL approach yielded more accurate predictions than the ECL approach. Moreover, it can be noted in Table 6.4 that the attributes estimates based on the ECL approach significantly vary in their correspondence, while the estimates of the REL approach tend more towards the average, which can be linked to the regression towards the mean [122]. The impact of each approach on face retrieval and verification is assessed through experiments in the next section.

6.3 Experiments

This section explores human face identification and verification using automatically retrieved biometric signatures that have been deduced from face images. The prediction of automatic biometric signatures was performed using the ECL and the REL approaches, which were introduced in section 6.2. The retrieval experiments simulate a realistic scenario in which a database of face images (e.g., mugshots or CCTV footage) is searched to identify an unknown subject based on verbal descriptions (i.e. eyewitness statements) for the subject's face. The face image of each subject in the dataset is represented by an automatically generated biometric signature as illustrated in Figure 6.1 and explained in the previous sections. As each subject has four different face image samples, this results in the construction of four visual databases, in which each subject is represented by its automatically generated biometric signature. Since the automatic biometric signatures can be retrieved using two different approaches (ECL and REL), two versions of each visual database were used in the experiments correspondingly, DB_{ECL_i} and DB_{REL_i} where $1 \leq i \leq 4$. The identification and verification experiments were performed using the LFW-MS4 dataset and followed a 10-fold cross validation approach as previously explained in section 6.2.1.4, where the test subjects are all new (unseen).

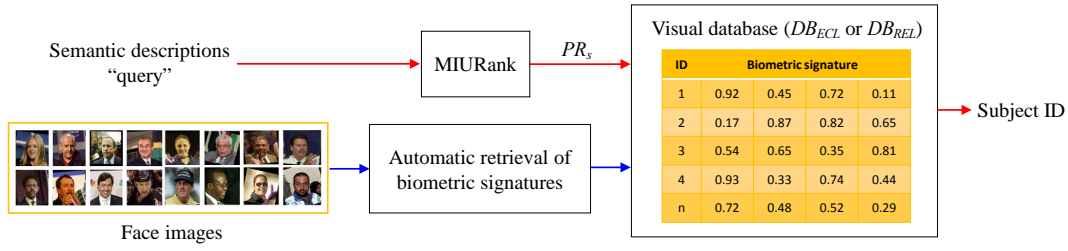


Figure 6.11: Illustration for subject retrieval from a visual database using semantic descriptions.

6.3.1 Identity Retrieval by Semantic Descriptions

In this experiment, the objective is to assess the accuracy of retrieving the identity of an unknown subject (probe) from a visual database (gallery) using verbal descriptions for the subject's face, as illustrated in Figure 6.11. For each subject in the test fold, a probe in the form of a biometric signature, PR_s , which represents a semantic description for an unknown subject, was generated from the crowdsourced comparative labels between the probe and ten randomly selected counterpart subjects from the training set. These comparative labels were used by the MIURank algorithm for generating the scores that compose the biometric signature. As explained previously in section 5.4.1, the number of subject comparisons (counterparts) that were used to construct the probe PR_s , was set as ten, since it is the average size of the ideal identity parade [15, 123].

Retrieval was performed by measuring the L_1 distance between the probe biometric signature, PR_s and the automatic biometric signature of each subject in the visual databases DB_{ECL} and DB_{REL} , which were constructed based on the ECL and REL approaches, respectively. The returned subjects were ranked according to their similarity with the probe subject, and the rank of the correct match was reported for performance evaluation. The experiment was repeated 100 times with each of the four visual databases (in both versions: ECL and REL). In each of the 100 trials, ten randomly selected subject were chosen from the training set to generate the probe PR_s . Finally, for each test subject, the mode rank among the 100 trials and the four visual databases was used for the performance reporting via the CMC curve that is shown in Figure 6.12.

It can be observed from Figure 6.12 that the retrieval accuracy from a visual database, which was constructed using the ECL approach, DB_{ECL} , outperforms the accuracy resulting when retrieving from a visual database that is constructed using the REL approach, DB_{REL} . Also, the results show that the retrieval accuracy in DB_{ECL} drastically increases exceeding 90% starting from rank-4, and reaching 100% at rank-15, which implies that a correct match can always be found in the top 3.49% subjects returned from the DB_{ECL} visual database. On the other hand, by observing the retrieval performance with the DB_{REL} visual database, it can be seen that although rank-1 accuracy is low, the retrieval performance significantly improves exceeding 90% at rank-18 and reaching

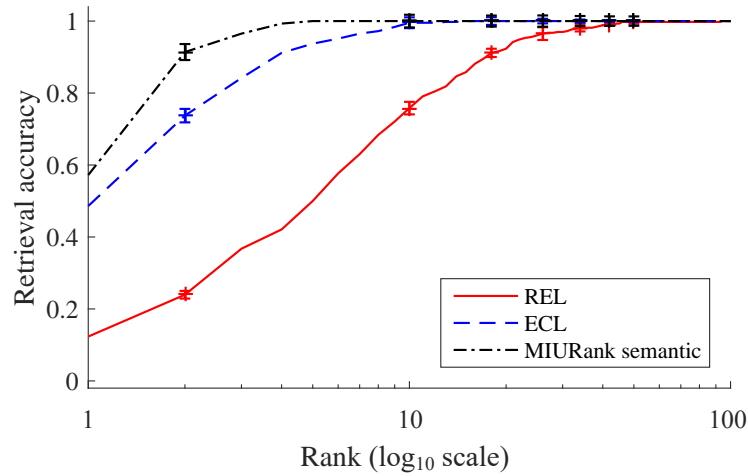


Figure 6.12: Retrieval performance resulting from automatically generated biometric signatures.

100% at rank-46, and thus, a correct response to a subject query will be found in the top 10.23% returned subjects. Figure 6.13 shows the retrieval performance in terms of the compression that can be achieved in the search range for the LFW-MS4 dataset, which consists of 430 subjects.

Although the ECL approach resulted in a better retrieval accuracy than the REL approach, several aspects need to be considered when comparing ECL with REL. First, the contribution of ECL in generating biometric signatures is limited to the estimation of binary comparative labels from differential visual features that are used by MIURank for ranking, while the REL approach implicitly involves the ranking of attributes. Second, the ECL is a pairwise approach, and thus, its time and space complexities exponentially increase with the dataset size, whereas the REL is a pointwise approach, which can be more efficient with larger datasets as compared with the ECL approach.

The only known work that has investigated subject retrieval by verbal descriptions for facial attributes is that of Klare et al. [54], which proposed a method for extracting categorical attributes from face images to build a visual database; it achieved a rank-1 retrieval rate of 23% using 46 facial attributes with the relatively constrained FERET database [58]. By comparing the outcomes from the semantic retrieval experiments in this chapter with Klare et al.'s results, the advantage of using comparative soft biometrics can be realised. Thus, using less attributes (24 as compared to 46 in [54]) and a more challenging dataset, which is the LFW-MS4, the comparative facial soft biometrics resulted in a better retrieval accuracy (48.6% at rank-1 using the ECL approach) as compared with the categorical attributes in [54]. In [124], Martinho-Corbishley and Nixon explored the retrieval of relative body soft biometrics using SoBiR [125], which is small and constrained data extracted from the BioT dataset, and achieved rank-10 retrieval rate of 26%. Overall, the results of the semantic retrieval experiment using both approaches (REL and ECL) highlight the reliability of comparative facial soft biometrics

for automatic estimation from face images, and reveal the informative value embedded in relative facial attributes.

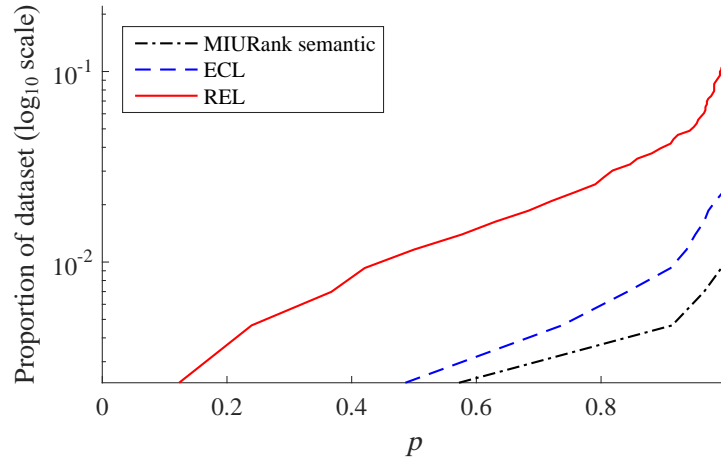


Figure 6.13: Compression achieved in search range with probability, p , of finding a correct match within range.

6.3.2 Verification Using Automatic Biometric Signatures

Whereas the objective of face identification is to recognise an unknown subject by comparing it to a database of known subjects using facial attributes, the objective of face verification is to validate a claimed identity against a database of known subjects (as explained in section 2.3). This experiment aims to assess the verification accuracy of automatic biometric signatures and evaluate the performance of automatic comparative facial soft biometrics. Here, it is important to highlight that the aim is not to promote the automatic biometric signatures as a competitor to traditional automatic face recognition, which has reached near perfection accuracies [126], but the aim is to position automatic comparative facial soft biometrics with respect to similar work that has used automatic categorical facial attributes for verification [51, 71].

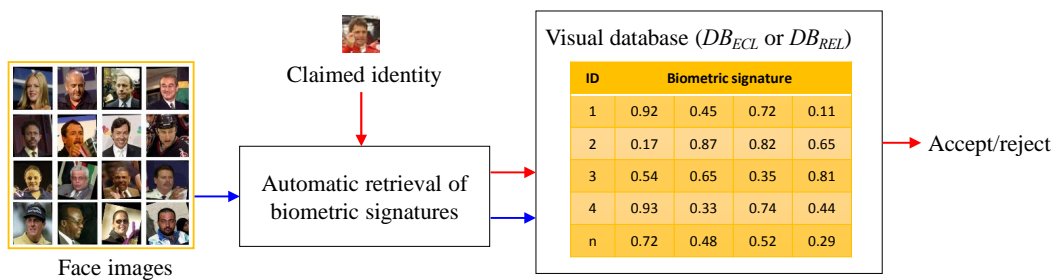


Figure 6.14: Illustration for face verification using automatic biometric signatures.

In this experiment, the accuracy of automatic biometric signatures for face verification was evaluated with both the ECL and REL approaches through the two visual databases that were constructed using these approaches (DB_{ECL} and DB_{REL}). The four samples of each visual database were used for assessing the verification performance. Figure 6.15 shows the verification performance of the automatically estimated biometric signatures through the ROC curve and the error curves. The resulting areas under the curve (AUC) are 96.29% and 91.02% for the REL and ECL approaches, respectively. It can be seen from Figure 6.15 that the automatic biometric signatures achieved an EER of 9.07% and 16.24% with the REL and ECL approaches respectively. Also, it is apparent from the verification results that while the ECL approach demonstrated better retrieval accuracy than the REL approach in the semantic identification experiment, the REL revealed a better verification performance than the ECL. This might be explained by the distribution of match scores resulting from each approach [127].

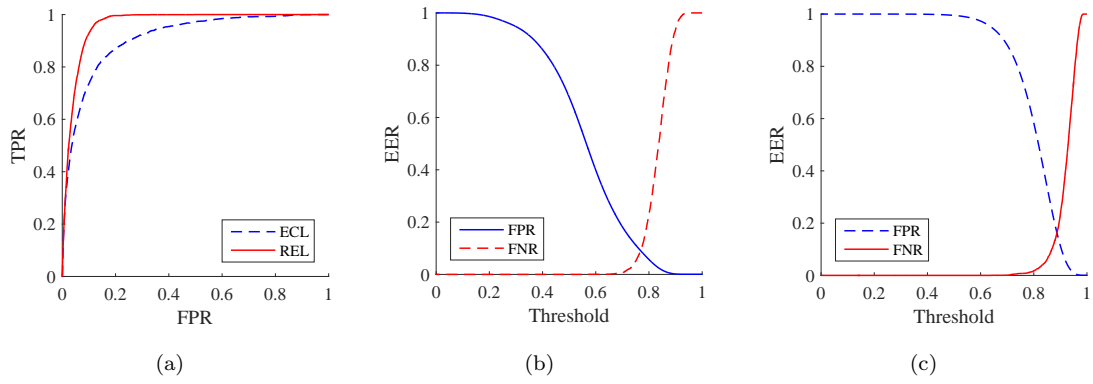
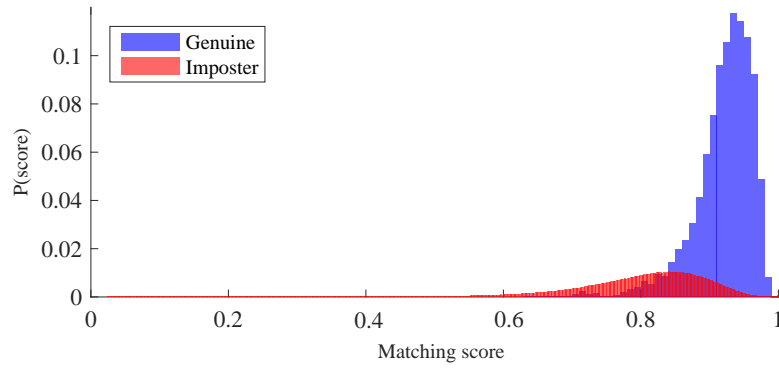
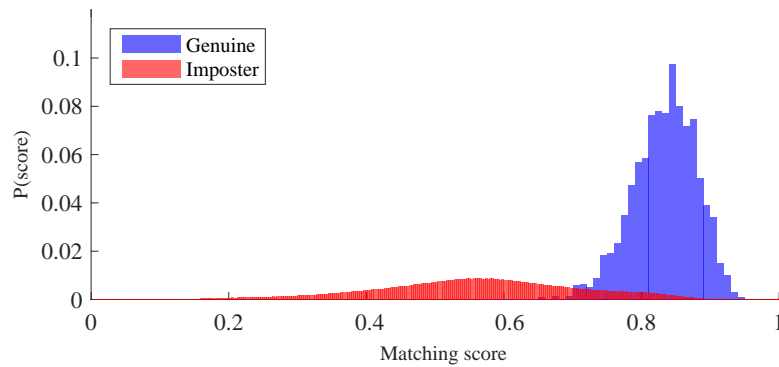


Figure 6.15: Verification performance of automatically estimated biometric signatures: (a) Receiver Operating Characteristic (ROC) curve; (b) Error curves with the REL approach; and (c) Error curves with the ECL approach.

Figure 6.16 shows the genuine and imposter distributions resulting from each of the approaches, and examples for true positive (TP) and false positive (FP) matches are shown in Figure 6.17. The face recognition approach proposed by Tome et al. in [51] used shape and size features of facial regions as soft biometrics, and achieved EERs of 3.06% and 12.27% for the relatively constrained ATVS [39] and MORPH [61] databases, respectively. In [71], categorical facial attributes were proposed for active authentication on mobile devices and achieved an EER of 14% with the MOBIO dataset [65], which is composed of video footage acquired by mobile cameras for 152 subjects. While Jaha and Nixon [119] achieved an EER of 10.2% with the 14 automatically estimated clothing attributes, using 128 subjects from the constrained SGDB database. In general, the results of this experiment indicate the latent benefits in comparative facial soft biometric for face verification and suggest the utilisation of the visual space resulting from the attribute in conducting face recognition.



(a) ECL



(b) REL

Figure 6.16: Genuine and imposter distributions of automatically estimated biometric signatures.



Figure 6.17: Example pairs of samples from the LFW-MS4 dataset: (a) - (c) are true positives, and (d) - (f) are false positives.

6.4 Conclusions

This chapter explored the automatic retrieval of biometric signatures from face images and investigated semantic identity retrieval as well as face verification using automatic

biometric signatures. The chapter presented a framework for estimating biometric signatures from face images, and proposed two different approaches for automatically generating soft biometric attributes. The first approach, REL, retrieves the ranks of attributes from visual features, and the second approach, ECL, estimates comparative labels from images pairs while utilising the MIURank algorithm for ranking for attributes. Using the unconstrained LFW-MS4 dataset, which consists of 430 subjects with 1720 samples, the outcomes from the analysis conducted in this chapter suggest that different facial attributes exhibit different levels of correspondence between semantic and visual spaces, depending on the approach used in retrieving attributes.

The results of the semantic retrieval experiment reveal that the proposed framework can yield a correct match in the top 3.49% and 10.23% returned subjects with the ECL and REL approaches respectively, using 24 attributes only. Furthermore, the automatically retrieved biometric signatures demonstrated a significant verification accuracy, as the results of the verification experiment showed that they can result in an EER of 9.07% and 16.24% with the ECL and REL approaches, respectively. Given the challenging visual conditions of the LFW-MS4 dataset and the relatively low number of attributes that compose the biometric signatures, these results reveal the reliability of the comparative facial soft biometric attributes for automatic estimation and the effectiveness of the proposed framework for identity retrieval by verbal descriptions in a database of images.

Chapter 7

Conclusions and Future Work

7.1 Conclusions

In this thesis, the aim was to explore human identification through comparative facial soft biometrics in large unconstrained databases. Towards this aim, a novel set of soft biometric attributes has been presented, and a large dataset of the well known unconstrained Labelled Faces in the Wild (LFW) database was used for the identification and verification experiments. The analyses have shown the significance of all the soft biometric attributes and revealed their contribution to the identification as well as verification performance. The identification experiments have demonstrated the effectiveness and scalability of comparative facial soft biometrics. Thus a rank-10 identification rate of 95.17% has been achieved using ten subject comparisons only. Furthermore, the proposed soft biometric attributes have shown that a correct match can always be retrieved in the top 1.76% returned subjects from the dataset that consists of 4038 subjects. Additionally, the comparative soft biometrics have revealed a significant verification accuracy. Thus, an EER of 12.14% was achieved using biometric signatures generated from ten subjects comparisons only. Altogether, these experimental results highlight the efficacy and the scalability of comparative facial soft biometrics for identification and verification in large unconstrained databases.

A major contribution of this thesis was MIURank, which was a novel fully unsupervised algorithm for ranking from pairwise comparisons based on mutual information. The motivation behind MIURank was to rank soft biometric attributes from comparative labels in a parameterless, yet effective, method. With its strong information-theoretic background and based on an intuitive concept that is ranking with respect to virtual superior and inferior performers, MIURank has shown a ranking accuracy that is comparable to the maximum likelihood estimator of the Bradley-Terry model with both synthetic and real datasets. Furthermore, MIURank has demonstrated that it can yield biometric signatures that have at least the same performance impact as those generated using the

Elo rating system. These results show the power of the information-theoretic foundations of MIURank and demonstrate that it is possible to achieve ranking from pairwise comparisons through a fully unsupervised and computationally efficient method.

Another aim of this thesis was to investigate the automatic retrieval of biometric signatures from face images. Two approaches have been proposed for estimating ranks of attributes, which compose biometric signatures, from face images: (1) the ECL approach, which predicts binary comparative labels from differential visual features, and which utilises MIURank for ranking; and (2) the REL approach, which predicts ranks from visual features. The correspondence of the visually inferred ranks has been assessed for the attributes under each of the two approaches. The predictions of the REL approach have demonstrated higher visual correspondence, while the predictions of the ECL approaches has revealed greater semantic correspondence.

Experiments have been conducted using the LFW-MS4 dataset, which consists of 430 subjects and 1720 samples, to simulate identity retrieval from an imagery database that is constructed from automatic biometric signatures. Using biometric signatures that were estimated using the ECL and REL approaches, respectively, the retrieval by semantic descriptions experiment has shown that a correct match can be found in the top 3.49% returned subjects. In summary, this result demonstrates the possibility of face retrieval by semantic descriptions, and, accordingly, highlights the extent to which comparative facial soft biometrics can bridge the semantic gap between machines and humans.

The automatically retrieved biometric signatures have shown promising potential for face verification. Thus, the experiments have demonstrated that EERs of 9.07% and 16.24% can result when face verification is performed using biometric signatures that are estimated using the REL and ECL approaches, respectively. These findings demonstrate the latent potential of the visual space that is deduced from comparative soft biometric attributes for performing face verification solely based on attributes.

In conclusion, the findings of this thesis significantly add to a growing body of literature on soft biometrics by showing the effectiveness and scalability of comparative facial attributes for large realistic databases. Also, the findings of the thesis show that comparative facial soft biometrics can significantly contribute to bridging the semantic gap between humans and machines, allowing identity retrieval by semantic descriptions from a database of images. Finally, the thesis extends our knowledge of the capabilities of information theory for addressing ranking as a classical machine learning problem.

7.2 Future Work

7.2.1 Soft Biometric Identification

Although LFW is an unconstrained dataset that reflects the realistic and challenging visual conditions of face images, it would be interesting to assess comparative facial soft biometrics for identification in surveillance databases that involve more adverse visual conditions. Furthermore, it is recommended that further research is undertaken to explore the factors that affect the accuracy and completeness of semantic descriptions for relative facial features. Also, future work could investigate the fusion of facial attributes with other groups of soft biometrics such as body and clothing in more comprehensive datasets, in addition to studying multi-viewpoint re-identification using comparative facial soft biometrics.

7.2.2 Descriptions from Memory

The comparisons collected for studying comparative facial soft biometrics in this thesis, as well as in the existing work on comparative attributes, have been collected while subjects' images were presented for annotators. However, real life scenarios will involve describing a suspect's face from an eyewitness's memory to retrieve the suspects' identity from a database. Therefore, it would be interesting to determine the accuracy of humans in recalling facial attributes of unknown subjects from memory relative to presented subjects and exploring the effectiveness of these soft biometric comparisons for identity retrieval.

7.2.3 Ranking of Soft Biometrics

Mutual information has demonstrated a great effectiveness for ranking from pairwise comparisons through the MIURank algorithm, which has revealed a ranking accuracy that is comparable to existing well-known algorithms. Furthermore, MIURank has shown an accuracy of ranking soft biometric attributes that is similar to the Elo rating system. Therefore, a natural progression for the MIURank algorithm is to investigate the possibility of exploiting mutual information for ranking attributes based on visual features.

7.2.4 Automatic Retrieval of Biometric Signatures

The automatic retrieval of biometric signatures from face images has been explored using both the ECL and REL approaches based on binary comparative labels, which were used to infer ranks of attributes. Nevertheless, more studies are required to examine

the reliability of the ECL and REL approaches with multi-polar comparative labels that involve more than two levels of comparisons. Furthermore, whereas the automatic biometric signatures have shown promising verification accuracy in the visual space, it would be interesting to explore vision-based verification on a larger scale. In addition, another possible area of future research would be to investigate the retrieval of biometric signatures from images using deep learning techniques and assess their implications on retrieval accuracy.

Where there is an increasing interest in sorting subjects in law enforcement agencies' databases by age and gender, future trials should assess the impact of categorising age and gender from scores, which are estimated from the visual space. Also, future work could determine the potential of utilising relative attributes in improving the accuracy of age and gender predictors in the categorical space. Finally, studying the scalability of automatic biometric signatures would be a fruitful area for future work.

Bibliography

- [1] BBC. *CCTV released in connection with robbery in Edinburgh*, 2016 (accessed January 16, 2017). <http://www.bbc.co.uk/news/uk-scotland-edinburgh-east-fife-38451363>.
- [2] Richard D. Seely, Sina Samangooei, Middleton Lee, John N. Carter, and Mark S. Nixon. The University of Southampton multi-biometric tunnel and introducing a novel 3d gait dataset. In *Biometrics Theory, Applications and Systems (BTAS), 2008 IEEE 2nd International Conference on*, pages 1–6. IEEE, 2008.
- [3] Paul Viola and Michael J. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, 2004.
- [4] Akshay Asthana, Stefanos Zafeiriou, Georgios Tzimiropoulos, Shiyang Cheng, and Maja Pantic. From pixels to response maps: Discriminative image filtering for face alignment in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(6):1312–1320, 2015.
- [5] Nawaf Y. Almudhahka, Mark S. Nixon, and Jonathon S. Hare. Human face identification via comparative soft biometrics. In *Identity, Security and Behavior Analysis (ISBA), 2016 IEEE International Conference on*, pages 1–6. IEEE, 2016.
- [6] Nawaf Y. Almudhahka, Mark S. Nixon, and Jonathon S. Hare. Unconstrained human identification using comparative facial soft biometrics. In *Biometrics Theory, Applications and Systems (BTAS), 2016 IEEE 8th International Conference on*, pages 1–6. IEEE, 2016.
- [7] Mark S. Nixon, Bingchen H. Guo, Sarah V. Stevenage, Emad S. Jaha, Nawaf Almudhahka, and Daniel Martinho-Corbishley. Towards automated eyewitness descriptions: describing the face, body and clothing for recognition. *Visual Cognition*, pages 1–15, 2016.
- [8] Nawaf Y. Almudhahka, Mark S. Nixon, and Jonathon S. Hare. Automatic semantic face recognition. In *Automatic Face and Gesture Recognition (FG), 2017 IEEE 12th International Conference on*, pages 180–185. IEEE, 2017.

- [9] Nawaf Yousef Almodhahka, Mark S. Nixon, and Jonathon S. Hare. Semantic face signatures: Recognizing and retrieving faces by verbal descriptions. *IEEE Transactions on Information Forensics and Security*, 2017.
- [10] S. L. Sporer. An archival analysis of person descriptions. In *Biennial Meeting of the American Psychology-Law Society in San Diego, California*, 1992.
- [11] Mark S. Nixon, Paulo L. Correia, Kamal Nasrollahi, Thomas B. Moeslund, Abdenour Hadid, and Massimo Tistarelli. On soft biometrics. *Pattern Recognition Letters*, 68:218–230, 2015.
- [12] Sina Samangooei, Baofeng Guo, and Mark S. Nixon. The use of semantic human description as a soft biometric. In *Biometrics Theory, Applications and Systems (BTAS), 2016 IEEE 2nd International Conference on*, pages 1–7. IEEE, 2008.
- [13] Daniel Reid, Sina Samangooei, Cunjian Chen, Mark Nixon, and Arun Ross. Soft biometrics for surveillance: an overview. *Machine Learning: Theory and Applications*, pages 327–352, 2013.
- [14] A. Mike Burton, Stephen Wilson, Michelle Cowan, and Vicki Bruce. Face recognition in poor-quality video: Evidence from security surveillance. *Psychological Science*, 10(3):243–248, 1999.
- [15] Daniel A. Reid, Mark S. Nixon, and Sarah V. Stevenage. Soft biometrics; human identification using comparative descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(6):1216–1228, 2014.
- [16] Josh P. Davis and Tim Valentine. Human verification of identity from photographic images. *Forensic Facial Identification: Theory and Practice of Identification from Eyewitnesses, Composites and CCTV*, pages 209–238, 2015.
- [17] T. A. M. Crawford and Karen Evans. *Crime Prevention and Community Safety*. Oxford University Press, 2016.
- [18] Tim Valentine and Josh P. Davis. Forensic facial identification. *Forensic Facial Identification: Theory and Practice of Identification from Eyewitnesses, Composites and CCTV*, pages 323–347, 2015.
- [19] Richard Kemp and David White. Face identification 3. *An Introduction to Applied Cognitive Psychology*, page 39, 2016.
- [20] Graeme Gerrard and Richard Thompson. Two million cameras in the UK. *CCTV Image*, 42(10):e2, 2011.
- [21] David Barrett. One surveillance camera for every 11 people in Britain, says CCTV survey. *The Telegraph*, 10, 2013.

- [22] Neeraj Kumar, Alexander C. Berg, Peter N. Belhumeur, and Shree K. Nayar. Attribute and simile classifiers for face verification. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 365–372. IEEE, 2009.
- [23] Daniel A. Reid and Mark S. Nixon. Using comparative human descriptions for soft biometrics. In *Biometrics (IJCB), 2011 International Joint Conference on*, pages 1–6. IEEE, 2011.
- [24] Hannah Ryder, Harriet M. J. Smith, and Heather D. Flowe. Estimator variables and memory for faces. *Forensic Facial Identification: Theory and Practice of Identification from Eyewitnesses, Composites and CCTV*, pages 159–183, 2015.
- [25] Rachel Robbins and Elinor McKone. No face-like processing for objects-of-expertise in three behavioural tasks. *Cognition*, 103(1):34–79, 2007.
- [26] Pawan Sinha, Benjamin Balas, Yuri Ostrovsky, and Richard Russell. Face recognition by humans: Nineteen results all computer vision researchers should know about. *Proceedings of the IEEE*, 94(11):1948–1962, 2006.
- [27] P Jonathon Phillips and Alice J. O’Toole. Comparison of human and computer performance across face recognition experiments. *Image and Vision Computing*, 32(1):74–85, 2014.
- [28] Vicki Bruce and Andy Young. Understanding face recognition. *British Journal of Psychology*, 77(3):305–327, 1986.
- [29] Peter J. B. Hancock, Vicki Bruce, and A. Mike Burton. Recognition of unfamiliar faces. *Trends in Cognitive Sciences*, 4(9):330–337, 2000.
- [30] Alice J. O’Toole. Psychological and neural perspectives on human face recognition. In *Handbook of Face Recognition*, pages 349–369. Springer, 2005.
- [31] Doris Y. Tsao and Margaret S. Livingstone. Mechanisms of face perception. *Annu. Rev. Neurosci.*, 31:411–437, 2008.
- [32] Graham Davies, Hadyn Ellis, and John Shepherd. Cue saliency in faces as assessed by the ‘Photofit’ technique. *Perception*, 6(3):263–269, 1977.
- [33] Nigel D. Haig. Exploring recognition with interchanged facial features. *Perception*, 15(3):235–247, 1986.
- [34] Javid Sadr, Izzat Jarudi, and Pawan Sinha. The role of eyebrows in face recognition. *Perception*, 32(3):285–293, 2003.
- [35] Lowell L. Kuehn. Looking down a gun barrel: Person perception and violent crime. *Perceptual and Motor Skills*, 39(3):1159–1164, 1974.
- [36] Peter J. Van Koppen and Shara K. Lochun. Portraying perpetrators: The validity of offender descriptions by witnesses. *Law and Human Behavior*, 21(6):661, 1997.

- [37] Christian A. Meissner, Siegfried L. Sporer, and Jonathan W. Schooler. Person descriptions as eyewitness evidence. *Handbook of Eyewitness Psychology: Memory for People*, pages 1–34, 2013.
- [38] Eric Lee, Thomas Whalen, John Sakalauskas, Glen Baigent, Chandra Bisesar, Andrew McCarthy, Glenda Reid, and Cynthia Wotton. Suspect identification by facial features. *Ergonomics*, 47(7):719–747, 2004.
- [39] Pedro Tome, Julian Fierrez, S.en Vera-Rodriguez, and Mark S. Nixon. Soft biometrics and their application in person recognition at a distance. *IEEE Transactions on Information Forensics and Security*, 9(3):464–475, 2014.
- [40] Anil K. Jain, Sarat C. Dass, and Karthik Nandakumar. Soft biometric traits for personal recognition systems. In *Biometric Authentication*, pages 731–738. Springer, 2004.
- [41] Victoria Z. Lawson and Jennifer E. Dysart. Searching for suspects. *Forensic Facial Identification: Theory and Practice of Identification from Eyewitnesses, Composites and CCTV*, pages 71–92, 2015.
- [42] J. L. Wayman. Large-scale civilian biometric systems-issues and feasibility. In *Proceedings of Card Tech/Secur Tech ID*, volume 732, 1997.
- [43] Jamie D. Shutler, Michael G. Grant, Mark S. Nixon, and John N. Carter. On a large sequence-based human gait database. In *Applications and Science in Soft Computing*, pages 339–346. Springer, 2004.
- [44] Daniel Martinho-Corbishley, Mark S. Nixon, and John N. Carter. Soft biometric recognition from comparative crowdsourced annotations. *IET Biometrics*, pages 1–16, 2016.
- [45] Daniel Martinho-Corbishley, Mark S. Nixon, and John N. Carter. On categorising gender in surveillance imagery. In *Biometrics Theory, Applications and Systems (BTAS), 2016 IEEE 8th International Conference on*, pages 1–6. IEEE, 2016.
- [46] Arpad E. Elo. *The rating of chessplayers, past and present*. Arco Pub., 1978.
- [47] Thorsten Joachims. Optimizing search engines using clickthrough data. In *Proceedings of The Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 133–142. ACM, 2002.
- [48] Emad Sami Jaha and Mark S. Nixon. Soft biometrics for subject identification using clothing attributes. In *Biometrics (IJCB), 2014 IEEE International Joint Conference on*, pages 1–6. IEEE, 2014.
- [49] Emad Sami Jaha and Mark S. Nixon. Viewpoint invariant subject retrieval via soft clothing biometrics. In *Biometrics (ICB), 2015 International Conference on*, pages 73–78. IEEE, 2015.

- [50] Alice O'Toole and P Jonathon Phillips. Evaluating automatic face recognition systems with human benchmarks. *Forensic Facial Identification: Theory and Practice of Identification from Eyewitnesses, Composites and CCTV*, page 263, 2015.
- [51] Pedro Tome, Ruben Vera-Rodriguez, Julian Fierrez, and Javier Ortega-Garcia. Facial soft biometric features for forensic face recognition. *Forensic Science International*, 257:271–284, 2015.
- [52] Pouya Samangouei, Vishal M. Patel, and Rama Chellappa. Facial attributes for active authentication on mobile devices. *Image and Vision Computing*, 58:181–192, 2017.
- [53] Olasimbo Ayodeji Arigbabu, Sharifah Mumtazah Ahmad, Wan Azizun Adnan, and Salman Yussof. Recent advances in facial soft biometrics. *The Visual Computer: International Journal of Computer Graphics*, 31(5):513–525, 2015.
- [54] Brendan F. Klare, Scott Klum, Joshua C. Klontz, Emma Taborsky, Tayfun Akgul, and Anil K. Jain. Suspect identification based on descriptive facial attributes. In *Biometrics (IJCB), 2014 IEEE International Joint Conference on*, pages 1–8. IEEE, 2014.
- [55] Fengyi Song, Xiaoyang Tan, and Songcan Chen. Exploiting relationship between attributes for improved face verification. *Computer Vision and Image Understanding*, 122:143–154, 2014.
- [56] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, Technical Report 07-49, University of Massachusetts, Amherst, 2007.
- [57] Unsang Park and Anil K. Jain. Face matching and retrieval using soft biometrics. *IEEE Transactions on Information Forensics and Security*, 5(3):406–415, 2010.
- [58] P Jonathon Phillips, Hyeonjoon Moon, Syed A. Rizvi, and Patrick J. Rauss. The feret evaluation methodology for face-recognition algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(10):1090–1104, 2000.
- [59] Daniel A. Reid and Mark S. Nixon. Human identification using facial comparative descriptions. In *Biometrics (ICB), 2013 International Conference on*, pages 1–7. IEEE, 2013.
- [60] Alex J. Smola and Bernhard Schölkopf. A tutorial on support vector regression. *Statistics and Computing*, 14(3):199–222, 2004.
- [61] Karl Ricanek and Tamirat Tesafaye. Morph: A longitudinal image database of normal adult age-progression. In *Automatic Face and Gesture Recognition (FG), 2006 IEEE 7th International Conference on*, pages 341–345. IEEE, 2006.

- [62] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3730–3738, 2015.
- [63] Yi Sun, Yuheng Chen, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation by joint identification-verification. In *Advances in Neural Information Processing Systems*, pages 1988–1996, 2014.
- [64] Max Ehrlich, Timothy J. Shields, Timur Almaev, and Mohamed R. Amer. Facial attributes classification using multi-task representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 47–55, 2016.
- [65] Christopher McCool, Sebastien Marcel, Abdenour Hadid, Matti Pietikäinen, Pavel Matejka, Jan Cernocký, Norman Poh, Josef Kittler, Anthony Larcher, Christophe Levy, et al. Bi-modal person recognition on a mobile phone: using mobile phone data. In *Multimedia and Expo Workshops (ICMEW), 2012 IEEE International Conference on*, pages 635–640. IEEE, 2012.
- [66] Mohammed E. Fathy, Vishal M. Patel, and Rama Chellappa. Face-based active authentication on mobile devices. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 1687–1691. IEEE, 2015.
- [67] Devi Parikh and Kristen Grauman. Relative attributes. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 503–510. IEEE, 2011.
- [68] Ramachandruni N. Sandeep, Yashaswi Verma, and C. V. Jawahar. Relative parts: Distinctive parts for learning relative attributes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3614–3621, 2014.
- [69] Aron Yu and Kristen Grauman. Just noticeable differences in visual attributes. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2416–2424, 2015.
- [70] Anil K. Jain and Unsang Park. Facial marks: Soft biometric for face recognition. In *Image Processing (ICIP), 2009 16th IEEE International Conference on*, pages 37–40. IEEE, 2009.
- [71] Pouya Samangouei, Vishal M. Patel, and Rama Chellappa. Facial attributes for active authentication on mobile devices. *Image and Vision Computing*, 58:181–192, 2017.
- [72] Anil K. Jain, Arun Ross, and Salil Prabhakar. An introduction to biometric recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 14(1):4–20, 2004.

- [73] Brian DeCann and Arun Ross. Relating roc and cmc curves via the biometric menagerie. In *Biometrics: Theory, Applications and Systems (BTAS), 2013 IEEE Sixth International Conference on*, pages 1–8. IEEE, 2013.
- [74] Patrick J. Grother, George W. Quinn, and P Jonathon Phillips. Report on the evaluation of 2d still-image face recognition algorithms. *NIST Interagency Report*, 7709:106, 2010.
- [75] Mark E. Glickman. A comprehensive guide to chess ratings. *American Chess Journal*, 3:59–102, 1995.
- [76] Wenyi Zhao, Rama Chellappa, P Jonathon Phillips, and Azriel Rosenfeld. Face recognition: A literature survey. *ACM Computing Surveys (CSUR)*, 35(4):399–458, 2003.
- [77] Hu Han and Anil K. Jain. Age, gender and race estimation from unconstrained face images. *Dept. Comput. Sci. Eng., Michigan State Univ., East Lansing, MI, USA, MSU Tech. Rep.(MSU-CSE-14-5)*, 2014.
- [78] Brendan Klare and Anil K. Jain. On a taxonomy of facial features. In *Biometrics: Theory Applications and Systems (BTAS), 2010 Fourth IEEE International Conference on*, pages 1–8. IEEE, 2010.
- [79] Theodore W. Anderson and Donald A. Darling. A test of goodness of fit. *Journal of the American Statistical Association*, 49(268):765–769, 1954.
- [80] Galina V. Veres, Layla Gordon, John N. Carter, and Mark S. Nixon. What image information is important in silhouette-based gait recognition? In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 2, pages II–II. IEEE, 2004.
- [81] Baofeng Guo and Mark S. Nixon. Gait feature subset selection by mutual information. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 39(1):36–46, 2009.
- [82] Niv Zehngut, Felix Juefei-Xu, Rishabh Bardia, Dipan K. Pal, Chandrasekhar Bhagavatula, and Marios Savvides. Investigating the feasibility of image-based nose biometrics. In *Image Processing (ICIP), 2015 IEEE International Conference on*, pages 522–526. IEEE, 2015.
- [83] Sahand Negahban, Sewoong Oh, and Devavrat Shah. Iterative ranking from pairwise comparisons. In *Advances in Neural Information Processing Systems*, pages 2474–2482, 2012.
- [84] Hyokun Yun, Parameswaran Raman, and S. Vishwanathan. Ranking via robust binary classification. In *Advances in Neural Information Processing Systems*, pages 2582–2590, 2014.

- [85] Fajwel Fogel, Alexandre d’Aspremont, and Milan Vojnovic. Serialrank: Spectral ranking using seriation. In *Advances in Neural Information Processing Systems*, pages 900–908, 2014.
- [86] David R. Hunter. MM algorithms for generalized Bradley-Terry models. *Annals of Statistics*, pages 384–406, 2004.
- [87] Xi Chen, Paul N. Bennett, Kevyn Collins-Thompson, and Eric Horvitz. Pairwise ranking aggregation in a crowdsourced setting. In *Proceedings of The Sixth ACM International Conference on Web Search and Data Mining*, pages 193–202. ACM, 2013.
- [88] Arvind Arasu, Jasmine Novak, Andrew Tomkins, and John Tomlin. Pagerank computation and the structure of the web: Experiments and algorithms. In *Proceedings of the Eleventh International World Wide Web Conference, Poster Track*, pages 107–117, 2002.
- [89] Yoav Freund, Raj Iyer, Robert E. Schapire, and Yoram Singer. An efficient boosting algorithm for combining preferences. *Journal of Machine Learning Research*, 4(Nov):933–969, 2003.
- [90] Yanwei Fu, Timothy M. Hospedales, Tao Xiang, Jiechao Xiong, Shaogang Gong, Yizhou Wang, and Yuan Yao. Robust subjective visual property prediction from crowdsourced pairwise labels. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(3):563–577, 2016.
- [91] Ralf Herbrich, Tom Minka, and Thore Graepel. Trueskill: a bayesian skill rating system. In *Advances in Neural Information Processing Systems*, pages 569–576, 2007.
- [92] Fabian Wauthier, Michael Jordan, and Nebojsa Jojic. Efficient ranking from pairwise comparisons. In *International Conference on Machine Learning*, pages 109–117, 2013.
- [93] Peter J. Huber et al. Pairwise comparison and ranking: optimum properties of the row sum procedure. *The Annals of Mathematical Statistics*, 34(2):511–520, 1963.
- [94] Claude E. Shannon. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1):3–55, 2001.
- [95] Ralf Steuer, Jürgen Kurths, Carsten O. Daub, Janko Weise, and Joachim Selbig. The mutual information: detecting and evaluating dependencies between variables. *Bioinformatics*, 18(suppl_2):S231–S240, 2002.
- [96] Tennis Australia Limited (TA). *Australian Open Tennis Championships 2016*, 2016 (Accessed May 5, 2016). "http://www.ausopen.com/en_AU/players/index.html.

- [97] The Football Association Premier League Limited. *Premier League 2014/15*, 2017 (Accessed June 23, 2017). "<https://www.premierleague.com/tables?co=1&se=27&mw=-1&ha=-1>".
- [98] Michael Zander. *The Police and Criminal Evidence Act 1984*. Sweet & Maxwell, 2013.
- [99] Richard Brent and Paul Zimmermann. *Modern Computer Arithmetic*. Cambridge University Press, New York, NY, USA, 2010.
- [100] Gary Huang, Marwan Mattar, Honglak Lee, and Erik G. Learned-Miller. Learning to align from scratch. In *Advances in Neural Information Processing Systems*, pages 764–772, 2012.
- [101] Gary B. Huang and Erik Learned-Miller. Labeled faces in the wild: Updates and new reporting procedures. *Dept. Comput. Sci., Univ. Massachusetts Amherst, Amherst, MA, USA, Tech. Rep*, pages 14–003, 2014.
- [102] P. K. Ito. 7 Robustness of ANOVA and MANOVA test procedures. *Handbook of Statistics*, 1:199–236, 1980.
- [103] Timothy F. Cootes, Christopher J. Taylor, David H. Cooper, and Jim Graham. Active shape models-their training and application. *Computer Vision and Image Understanding*, 61(1):38–59, 1995.
- [104] Andreas Lanitis, Christopher J. Taylor, and Timothy F. Cootes. Automatic interpretation and coding of face images using flexible models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):743–756, 1997.
- [105] Timothy F. Cootes, Gareth J. Edwards, and Christopher J. Taylor. Active appearance models. In *European Conference on Computer Vision*, pages 484–498. Springer, 1998.
- [106] Gareth J. Edwards, Christopher J. Taylor, and Timothy F. Cootes. Interpreting face images using active appearance models. In *Automatic Face and Gesture Recognition (FG), 1998 IEEE 3rd International Conference on*, pages 300–305. IEEE, 1998.
- [107] David Cristinacce and Tim Cootes. Automatic feature localisation with constrained local models. *Pattern Recognition*, 41(10):3054–3067, 2008.
- [108] Jason M. Saragih, Simon Lucey, and Jeffrey F Cohn. Deformable model fitting by regularized landmark mean-shift. *International Journal of Computer Vision*, 91(2):200–215, 2011.
- [109] Timothy F. Cootes, Gareth J. Edwards, and Christopher J. Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):681–685, 2001.

- [110] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005.
- [111] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001.
- [112] Timo Ojala, Matti Pietikainen, and Topi Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):971–987, 2002.
- [113] Jie Chen, Vili Kellokumpu, Guoying Zhao, and Matti Pietikäinen. RLBP: Robust Local Binary Pattern. In *BMVC*, 2013.
- [114] Li Liu, Songyang Lao, Paul W. Fieguth, Yulan Guo, Xiaogang Wang, and Matti Pietikäinen. Median robust extended local binary pattern for texture classification. *IEEE Transactions on Image Processing*, 25(3):1368–1381, 2016.
- [115] Ningning Zhou, A. G. Constantinides, Guofang Huang, and Shaobai Zhang. Face recognition based on an improved center symmetric local binary pattern. *Neural Computing and Applications*, pages 1–7, 2017.
- [116] Justin Domke and Yiannis Aloimonos. Deformation and viewpoint invariant color histograms. In *BMVC*, pages 509–518, 2006.
- [117] Leo Breiman, Jerome Friedman, Charles J. Stone, and Richard A. Olshen. *Classification and regression trees*. CRC press, 1984.
- [118] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.
- [119] Emad Sami Jaha and Mark S. Nixon. From clothing to identity: Manual and automatic soft biometrics. *IEEE Transactions on Information Forensics and Security*, 11(10):2377–2390, 2016.
- [120] Mokhtar Bin Abdullah. On a robust correlation coefficient. *The Statistician*, pages 455–460, 1990.
- [121] Rudy A. Gideon and Robert A. Hollister. A rank correlation coefficient resistant to outliers. *Journal of the American Statistical Association*, 82(398):656–666, 1987.
- [122] Brian S. Everitt. *The Cambridge dictionary of statistics*. Cambridge University Press, 2006.
- [123] Michael D. A. Freeman. *The Police and Criminal Evidence Act 1984*. Sweet & Maxwell, 1985.

- [124] Daniel Martinho-Corbishley, Mark S. Nixon, and John N. Carter. Retrieving relative soft biometrics for semantic identification. In *Pattern Recognition (ICPR), 2016 23rd International Conference on*, pages 3067–3072. IEEE, 2016.
- [125] Daniel Martinho-Corbishley, Mark S. Nixon, and John N. Carter. Soft biometric retrieval to describe and identify surveillance images. In *Identity, Security and Behavior Analysis (ISBA), 2016 IEEE International Conference on*, pages 1–6. IEEE, 2016.
- [126] Erik Learned-Miller, Gary B. Huang, Aruni RoyChowdhury, Haoxiang Li, and Gang Hua. Labeled faces in the wild: A survey. In *Advances in Face Detection and Facial Image Analysis*, pages 189–248. Springer, 2016.
- [127] Brian DeCann and Arun Ross. Can a poor verification system be a good identification system? a preliminary study. In *Information Forensics and Security (WIFS), 2012 IEEE International Workshop on*, pages 31–36. IEEE, 2012.