Abstract

This thesis describes research into the score-level fusion process in multimodal biometrics. The emphasis of the research is on the fusion of face and voice biometrics in the two recognition modes of verification and open-set identification.

The growing interest in the use of multiple modalities in biometrics is due to its potential capabilities for eradicating certain important limitations of unimodal biometrics. One of the factors important to the accuracy of a multimodal biometric system is the choice of the technique deployed for data fusion. To address this issue, investigations are carried out into the relative performance of several statistical data fusion techniques for combining the score information in both unimodal and multimodal biometrics (i.e. speaker and/ or face verification).

Another important issue associated with any multimodal technique is that of variations in the biometric data. Such variations are reflected in the corresponding biometric scores, and can thereby adversely influence the overall effectiveness of multimodal biometric recognition. To address this problem, different methods are proposed and investigated.

The first approach is based on estimating the relative quality aspects of the test scores and then passing them on into the fusion process either as features or weights. The approach provides the possibility of tackling the data variations based on adjusting the weights for each of the modalities involved according to its relative quality.

Another approach considered for tackling the effects of data variations is based on the use of score normalisation mechanisms. Whilst score normalisation has been widely used in voice biometrics, its effectiveness in other biometrics has not been previously investigated. This method is shown to considerably improve the accuracy of multimodal biometrics by appropriately correcting the scores from degraded modalities prior to the fusion process.

The investigations in this work are also extended to the combination of score normalisation with relative quality estimation. The experimental results show that, such a combination is more effective than the use of only one of these techniques with the fusion process.

The thesis presents a thorough description of the research undertaken, details the experimental results and provides a comprehensive analysis of them.

Acknowledgments

Taking this opportunity I would like to express my gratitude to my very helpful principal supervisor Dr. Aladdin Ariyaeeinia for his supervision, guidance and support. This excellent support and guidance have indeed helped me find a proper direction in my work. His valuable advice and technical guidance have encouraged me to understand research methodology. I thank and appreciate him so much for his knowledge contribution; unconstrained guidance and encouragements throughout this research that let me reach this level. Also I would like to express my appreciation to my second supervisor Dr Lily Meng for her support, and her knowledge contribution in the field of classification using Support Vector Machines. I would also like to thank my colleague Dr. Amit Malegaonkar for the provision of the speech recognition scores and his frequent discussions with me which have helped strengthen my knowledge in this domain. I would like to thank the University of Hertfordshire, Faculty of Engineering and Information Sciences for providing a superb academic environment. Many thanks to all research students I have met within the Faculty for providing me with a warm and friendly environment. I would also like to thank IDIAP Research Institute, in particular, Dr Norman Poh for providing the face and speech verification scores. I also wish to express my gratitude to the Aristotle University of Thessaloniki and the University of Vigo and, in particular, Professor Ioannis Pitas, Professor Carmen Garcia Mateo, Professor Jose Luis Alba, Mr Stefanos Zafeiriou, and Mr Daniel Gonzalez for their support and provision of the face recognition scores.

Also, I would like to express my thanks and deep appreciations for King Faisal University in Saudi Arabia for their financial support.

Last, but not least, I would like to express my gratitude and appreciation to all of my family (specially my wife) for their support, encouragement, faith, understanding and help. Without them, this work would not have been possible.

Table of Contents

Abstract i
Acknowledgments ii
Table of Contents iii
Glossary of Important Terms vii
List of Tables ix
List of Figures xi
Chapter 1 1
Introduction 1
1.1. Motivation behind multimodal biometrics 4
1.2. Aims and Objectives 6
1.3 Thesis Organisation 7
Chapter 2 9
Literature Review 9
2.1 Introduction 9
2.2 Fusion Levels 9
A. Pre-mapping fusion I: Fusion at the sensor level 10
B. Pre-mapping fusion II: Fusion at the feature level 10
C. Post-mapping fusion I: Fusion at the matching score level 10
D. Post-mapping fusion II: Fusion at the decision level 11
2.3 Previous research into multimodal biometrics 11
2.3.1 Fusion based on fixed rules 12
2.3.1.1 Fixed rules 12
2.3.1.2 Recent work on biometric fusion using fixed rules 14
2.3.2 Fusion based on trained rules 15
2.3.2.1 Trained rules 16
2.3.2.2 Recent work on biometric fusion using trained rules 19
2.3.3 Comparison of fixed rules and trained rules 21
2.3.4 Quality-Based Fusion 25
2.4 Summary 28

Chapter 3 29
Fusion Techniques for Multimodal Biometrics 29
3.1 Introduction 29
3.2 Range-Normalisation Techniques 29
3.2.1 Min-Max Normalisation (MM) 30
3.2.2 Z-score Normalisation (ZS) 30
3.3 Evaluation Criteria for identity recognition 31
3.3.1 Evaluation Criterion for identity verification 32
3.3.2 Evaluation Criterion for identification 33
3.4 Effective fusion techniques 34
3.4.1 Weighted Average Fusion 34
3.4.1.1 Brute Force Search (BFS) 34
3.4.1.2 Matcher Weighting using False Acceptance Rate and False Rejection Rate (MW – FAR/FRR) 35
3.4.1.3 Matcher Weighting based on Equal Error Rate (MW - EER) 36
3.4.2 Fisher Linear Discriminant (FLD) 36
3.4.2.1. Fisher Linear Discriminant for the Data from Two Classes 37
3.4.3 Quadratic Discriminant Analysis (QDA) 37
3.4.4 Logistic Regression (LR) 38
3.4.5 Support Vector Machines (SVM) 39
3.4.5.1 Linear SVM for linearly separable data 39
3.4.5.2 Linear SVM for non-linearly separable data 42
3.4.5.3 Non-linear SVM 43
3.5 Summary 44
Chapter 4 46
Score-level Fusion in Biometric Verification 46
4.1 Introduction 46
4.2 Speech and Face data 46
4.2.1 Classifiers and features 47
4.3 Experimental investigations and discussions 48
4.3.1 Score fusion in unimodal biometrics based on multiple matching algorithms 48

4.3.1.1 Score-level fusion results based on MM range-normalisation 49
4.3.1.2 Score-level fusion results based on ZS range-normalisation 51
4.3.2 Multimodal fusion 53
4.4 Summary 59
Chapter 5 60
Multimodal Authentication using Qualitative Support Vector Machines 60
5.1 Introduction 60
5.2 Proposed approach 61
5.2.1. Estimation of the quality aspects for the development data samples 62
5.2.2. Estimation of the quality aspects for the test data samples 63
5.2.3. Methods of passing the quality aspects to SVM 64
5.2.3.1. Relative quality aspects as independent features (RQ-IF) 64
5.2.3.2. Modality specific fusion of relative quality aspects (RQ-MSF) 65
5.3. Experimental Investigation 68
5.3.1. Speech and Face data 68
5.3.2. Testing with Fusion 68
5.3.3. Results and Discussions 69
5.4 Summary 74
Chapter 6 75
Enhancement of Multimodal Biometric Accuracy 75
6.1 Introduction 75
6.2. Motivation and proposed approach 76
6.3. Experimental investigations and results 81
6.3.1. Fusion under Clean Data Conditions 81
6.3.2. Fusion under Varied Data Quality Conditions 84
6.3.2.1. Fusion under clean face data and degraded voice data 84
6.3.2.2. Fusion under degraded face data and clean voice data 85
6.3.3. Fusion under degraded Data Conditions 87
6.4 Summary 90
Chapter 7 92
Combined Approach to Enhancing Multimodal Biometric Accuracy 92

7.1 Introduction 92
7.2. Proposed approach 92
7.3. Experimental investigations and results 93
7.3.1. Fusion under Clean Data Conditions 94
7.3.2 Fusion under Varied Data Conditions 97
7.3.3 Fusion under Degraded Data Conditions 100
7.4 Summary 105
Chapter 8 106
Conclusions and Future work 106
8.1 Summary and conclusions 106
8.2 Suggestions for future work 111
8.2.1 Range-Normalisation 111
8.2.2 Fusion techniques 112
8.2.2.1 Quality estimation 112
8.2.2.2 Unconstrained cohort normalisation at feature level 112
8.2.2.3 Unconstrained cohort normalisation for other types of biometrics 113
8.2.2.4 Unconstrained fusion techniques 113
References 114
APPENDIX A. Publications 124

Glossary of Important Terms

ATM	Automatic Telling Machines
PIN	Personal Identification Number
BFS	Brute Force Search
MW – FAR/FRR	Matcher Weighting based on False Acceptance Rate and False Rejection Rate
MW – EER	Matcher Weighting based on Equal Error Rate
FLD	Fisher Linear Discriminant
QDA	Quadratic Discriminant Analysis
LR	Logistic Regression
SVM	Support Vector Machine
Poly SVM	Polynomial Support Vector Machine
RBF SVM	Radial Basis Function Support Vector Machine
SLT	Statistical Learning Theory
SRM	Structural Risk Minimisation
ERM	Empirical Risk Minimisation
LPC	linear predictive coding
AMR	Arithmetic Mean Rule
SNR	Signal to Noise Ratio
PSNR	Peak Signal to Noise Ratio
MLP	Multi-Layer Perceptrons
MM	Min-Max Normalisation
ZS	Z-Score Normalisation
FAR	False Acceptance Rate
FRR	False Rejection Rate
EER	Equal Error Rate
ROC	Receiver Operating Characteristics
DET	Detection Error Tradeoff
LPCC	Linear Predictive Cepstral Coefficients
MFCC	Mel Frequency Cepstral Coefficients
GMM	Gaussian Mixture Model
FH	Face Histogram
DCTs	Discrete Cosine Transform (s indicates the use of small image)
DCTb	Discrete Cosine Transform (b indicates the use of bigger image)
LFCC	Linear Filter-bank Cepstral Coefficient
PAC	Phase Auto-Correlation
SSC	Spectral Subband Centroid
LP	Lausanne Protocol
RI	Relative Improvement
SVM-RQM	Support Vector Machine with Relative Quality Measurement
RQ-IF	Relative Quality aspects as Independent Features
RQ-MSF	Modality Specific Fusion of Relative Quality aspects
CN	Cohort Normalisation
UCN	Unconstrained Cohort Normalisation

UBM	Universal Background Model
NFS	Normalised Face Score
NVS	Normalised Voice Score
OSI	Open Set Identification
IER	Identification Error Rate
OSI-EER	Open Set Identification-Equal Error Rate
SVM-UCN	Support Vector Machine with Unconstrained Cohort Normalisation
RQ-IF-UCN	Relative Quality aspects as Independent Features with Unconstrained Cohort Normalisation
RQ-MSF-UCN	Modality Specific Fusion of Relative Quality aspects with Unconstrained Cohort Normalisation
CI	Confidence Interval

List of Tables

No.	Caption	Page No.
4.1	Lausanne Protocols for the XM2VTS database.	46
4.2	Baseline EERs computed using the unimodal verification scores in various cases. The best performance in each of the face and voice modalities is shown in italics.	47
4.3	Unimodal face verification results in terms of EER (%), based on score-level fusion with MM range-normalisation.	48
4.4	Unimodal voice verification results in terms of EER (%), based on Score-level fusion with MM range-normalisation.	48
4.5	Relative improvements (RI) for the unimodal face features using various fusion schemes based on MM range-normalisation.	50
4.6	Relative improvements (RI) for the unimodal voice features using various fusion schemes based on MM range-normalisation.	50
4.7	Unimodal face verification results in terms of EER (%), based on score-level fusion with ZS range-normalisation.	51
4.8	Unimodal voice verification results in terms of EER (%), based on score-level fusion with ZS range-normalisation.	51
4.9	Relative improvements (RI) for the unimodal face features using various fusion schemes based on ZS range-normalisation.	51
4.10	Relative improvements (RI) for the unimodal voice features using various fusion schemes based on ZS range-normalisation.	52
4.11	Bimodal verification results in terms of EER (%), based on Score-level fusion with MM range-normalisation.	53
4.12	Relative improvements (RI) for various fusion schemes based on MM range-normalisation.	53
4.13	Bimodal verification results in terms of EER (%), based on Score-level fusion with ZS range-normalisation.	54
4.14	Relative improvements (RI) for various fusion schemes based on ZS range-normalisation.	54
5.1	Baseline EERs computed using the unimodal verification scores in various cases. The best performance in each of the face and voice modalities is shown in italics.	68
5.2	Bi-modal authentication results in terms of EER (%), with and without relative quality learning.	69
5.3	Relative improvements for the bi-modal authentication with and without relative quality learning.	69
6.1	Effectiveness of UCN in Multimodal verification based on clean biometric data.	81
6.2	Experimental results for open-set identification based on clean biometric data.	82
6.3	Performance of UCN in biometric verification based on mixed-quality data (clean face data and degraded voice data).	83
6.4	Experimental results for open-set identification based on mixed-quality data (clean face data and degraded voice data).	84

6.5	Performance of UCN in biometric verification based on mixed-quality data (degraded face data and clean voice data).	85
6.6	Experimental results for open-set identification based on mixed-quality data (degraded face data and clean voice data).	85
6.7	Effectiveness of UCN in multimodal verification based on degraded data.	87
6.8	Experimental results for open-set identification based on degraded biometric data.	87
7.1	Effectiveness of combining qualitative linear SVM with UCN based on clean biometric data.	95
7.2	Experimental results for open-set identification based on clean biometric data.	96
7.3	Performance of UCN and quality learning in biometric verification based on mixed-quality data.	97
7.4	Experimental results for open-set identification based on mixed-quality data.	98
7.5	Effectiveness of UCN and quality learning in verification based on degraded data.	101
7.6	Experimental results for open-set identification based on degraded data.	102

List of Figures

NO.	Caption	Page No.
1.1	General scheme of a biometric system.	2
2.1	Fusion levels in multimodal biometric fusion.	9
3.1	Effects of Min-Max and Z-score on the distributions of original face and voice scores.	30
3.2	Possible separating hyper-planes.	39
3.3	Optimal separating hyper-pane.	40
3.4	Transformation to higher dimension space (Asano(2004)).	42
4.1	Comparison of RIs for the unimodal and multimodal verification experiments based on ZS range-normalisation.	56
4.2	Relative performance of fused biometrics (based on LR) and individual modalities (face and voice).	57
5.1	Proposed Scheme of SVM-RQM using quality aspect as separate features in the development stage.	64
5.2	Proposed Scheme of SVM-RQM using quality aspect as separate features in the test stage.	64
5.3	Proposed Scheme of SVM-RQM using quality aspect as weights at the development stage.	66
5.4	Proposed Scheme of SVM-RQM using quality aspect as weights at the test stage.	66
5.5	DET plots for SVM fusion with and without relative quality learning in bi-modal fusion together with the baseline performers.	71
5.6	Relative improvements for the bi-modal authentication with and without relative quality learning.	72
6.1	Unconstrained cohort normalisation of scores in multimodal biometric fusion.	79
6.2	DET plots for the verification experiments with degraded data.	88
6.3	DET plots for the verification process in the second stage of open-set identification experiments with degraded data.	89
7.1	Unconstrained cohort normalisation with relative quality learning of scores in multimodal biometric fusion.	94
7.2	Comparison of EERs for various fusion methods with the baseline ERR for face modality based on varied quality data. Recognition mode: Verification.	97
7.3	Comparison of OSI-EERs for various fusion methods with the baseline OSI-ERR for face modality based on varied quality data. Recognition mode: Verification process in the second stage of open-set identification.	99
7.4	Comparison of EERs for various fusing configurations with the baseline ERR for face modality based on degraded data. Recognition mode: Verification.	101
7.5	DET plots showing the effects of qualitative SVM and UCN on the verification process in the second stage of open-set	102

	identification experiments with degraded data.	
7.6	Comparison of OSI-EERs for various fusing configurations with the baseline OSI-ERR for face modality based on degraded data. Recognition mode: Verification process in the second stage of open-set identification.	103

Chapter 1

Introduction

The automatic verification of the identities of individuals is becoming an increasingly important requirement in a variety of applications, especially, those involving automatic access control. Examples of such applications are teleshopping, telebanking, physical access control, and the withdrawal of money from automatic telling machines (ATMs).

Traditionally, passwords, personal cards, PIN-numbers and keys have been used in this context. However, security can easily be breached in these systems when a card or key is lost or stolen or when a password is compromised. Furthermore, difficult passwords may be hard to remember by a legitimate user and simple passwords are easy to guess by an impostor. The use of biometrics offers an alternative means of identification which helps avoid the problems associated with conventional methods.

The word biometrics is defined as the recognition of an individual by checking the measurements of certain physical characteristics or personal traits against a database. Recognition could be by measurement of features in any of the three biometric categories: intrinsic; extrinsic; and hybrid. Intrinsic biometrics identifies the individual's generic make-up (e.g. fingerprint or iris patterns). Extrinsic biometrics involves the individual's learnt behaviour (e.g. signature or keystrokes). Finally, hybrid biometrics is based on a combination of the individual's physical characteristics and personal traits (e.g. voice characteristics).

A critical question is what biological (physical characteristics/ personal traits) measurements qualify to be a biometric. Any human trait can be considered as a biometric characteristic as long as it satisfies the following requirements[1, 2]:

- Universality: each person should have the selected biometric identifier.
- Distinctiveness: any two persons should be sufficiently different in terms of the selected biometric identifier.
- Permanence: the biometric identifier should be sufficiently invariant over a given period of time.
- Collectability: the biometric identifier should be measurable quantitatively.

In real life applications, there are a number of additional factors which should be considered:

- Performance: which includes accuracy, speed and resource requirements;
- Acceptability: the willingness of people to accept the biometric identifier in their daily lives;
- Circumvention: it should be sufficiently robust to withstand various fraudulent practices.

Biometric Systems

A simple biometric system consists of four basic components (Figure 1.1)[3]:

- Sensor module: this component is for acquiring the biometric data;
- Feature extraction module: the data obtained from the sensor is used to compute a set of feature vectors;
- Matching module: the feature vectors generated via the previous component are checked against those in the template;
- Decision making module: to accept or reject the claimed identity or to establish a user's identity.

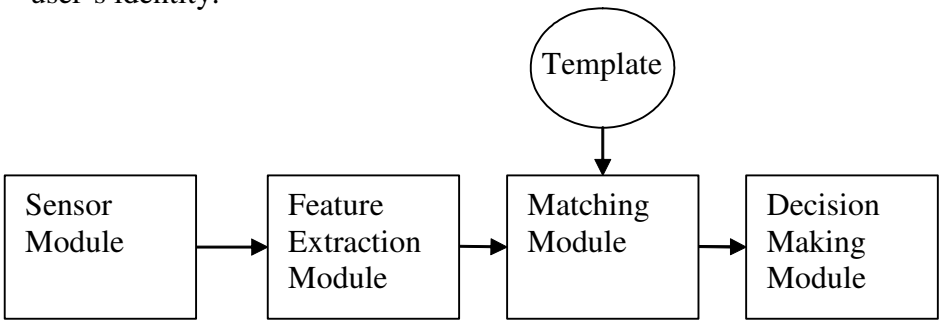


Figure 1.1: General scheme of a biometric system.

In general, a biometric recognition system involves two stages of operation. The first of these is the enrolment. There are two general processes in this stage. The first is acquisition of the user's biometric data, by means of a biometric reader appropriate to the data sought. The second concerns storage of the biometric data for each user in a reference database. This can be in a variety of forms including a template or a statistical model generated using the raw data. Whichever method is used, the stored data is labelled according to user identity to facilitate subsequent authentication.

The second stage of operation is termed testing. In this stage, the test biometric data obtained from the user is checked against the reference database for the purpose of recognition. A biometric recognition system can operate in one of the two modes of verification (also referred to as authentication in this thesis) and identification. In the verification mode, the user also makes an identity claim. In this case the test data is compared only against the reference data (e.g. template, statistical model) associated with the claimed identity. The result of this comparison is used to accept or reject the identity claim. In identification, the test data is compared against the data for all the registered individuals to determine the identity of the user (p. 2117)[4]. Thus, verification and identification are two distinct issues having their own inherent complexities.

Although this thesis is mainly focused on biometrics-based verification, Chapters 6 and 7 discuss both biometrics-based verification and Open Set Identification (OSI).

Biometric System Errors

Since this thesis discusses both biometric-based verification and Open-Set Identification, a brief discussion on the errors occurring in the verification process as well as those occurring in the identification process is presented below.

A biometric recognition system may make two types of errors [1] : 1) False Acceptance (FA), occurring when the system accepts an impostor, and 2) False Rejection (FR), taking place when the system rejects a client. Other errors that may occur in a biometric system are Failure To Capture (FTC) and Failure To Enroll (FTE). The FTC occurs when the device is not able to locate a biometric signal of sufficient quality (e.g. an

extremely faint fingerprint) whilst the FTE takes place when a user is not able to enroll in the recognition system. The Equal Error Rate (EER: i.e. when FAR=FRR) is used in this thesis as the performance measure of an identity verification method. The task of Open Set Identification consists of two component processes of identification and verification. For the verification in Open Set Identification, the system performance is measured in terms of Open Set Identification Equal Error Rates (OSI-EER: i.e. when OSI-FAR=OSI-FRR). The identification performance, however, is measured in terms of Identification Error Rate (IER). This occurs when an individual in a database is incorrectly identified. More details about performance evaluation in biometric systems are given in Section 3.3.

1.1. Motivation behind multimodal biometrics

Despite considerable advances in recent years, there are still serious challenges in obtaining reliable authentication through unimodal biometric systems. These are due to a variety of reasons. For instance, there are problems with enrolment due to the non-universal nature of relevant biometric traits. Non-universality means the possibility that a subset of users do not possess the biometric trait being acquired. Equally troublesome is biometric spoofing. Biometric spoofing means that it is possible for unimodal systems to be fooled, e.g. through the use of contact lenses with copied patterns for iris recognition. Moreover, the environmental noise effects on the data acquisition process can lead to deficient accuracy which may disable systems, virtually from inception [5]. Speaker verification, for instance, degrades rapidly in noisy environments. Similarly, the effectiveness of face verification depends strongly on lighting conditions and on variations in the subject's pose before the camera. Some of the limitations imposed by unimodal biometrics systems can be overcome by using multiple biometric modalities. Multiple evidence provision through multimodal biometric data acquisition may focus on multiple samples of a single biometric trait, designated as multi-sample biometrics. It may also focus on samples of multiple biometric types. This is termed multimodal biometrics. Higher accuracy and greater resistance to spoofing are basic advantages of multimodal biometrics over unimodal biometrics. Multimodal biometrics involves the use of complementary information as well as making it difficult for an intruder to spoof simultaneously the multiple biometric traits of a registered user. In addition, the problem

of non-universality is largely overcome, since multiple traits can ensure sufficient population coverage. Because of these advantages of multimodal biometrics, a multimodal biometric system is preferred over single modality even though the storage requirements, processing time and the computational demands of a multimodal biometric system are much higher.

The fusion of the complementary information in multimodal biometric data has been a research area of considerable interest, as it plays a critical role in overcoming certain important limitations of unimodal systems. The efforts in this area are mainly focused on fusing the information obtained from a variety of independent modalities. For instance, a popular approach is to combine face and speech modalities to achieve a more reliable recognition of individuals. Through such an approach, separate information from different modalities is used to provide complementary evidence about the identity of the users. In such scenarios, fusion is normally at the score level. This is because the individual modalities provide different raw data types, and involve different classification methods for discrimination. To date, a number of score-level fusion techniques have been developed for this task [6]. These range from the use of different weighting schemes that assign weights to the information streams according to their information content, to support vector machines which use the principle of obtaining the best possible boundary for classification, according to the training data. Despite these developments, the literature lacks a thorough comparison of various fusion methods for multimodal biometrics.

The purpose of the present work is to examine whether the performance of a biometric system can be improved by integrating complementary information which comes primarily from different modalities (multimodality). Another issue of concern in this thesis is the effect of data variation on the recognition performance of biometric systems. Such variations are reflected in the corresponding biometric scores, and thereby can adversely influence the overall effectiveness of biometric recognition. Therefore, an important requirement for the effective operation of a multimodal biometric system in practice is minimisation of the effects of variations in the data from the individual modalities deployed. This would allow maximisation of the recognition accuracy in the presence of variation (e.g. due to contamination) in some or all types of biometric data

involved. However, this is a challenging requirement as the data variation can be due to a variety of reasons, and can have different characteristics. Another aspect of difficulty in multimodal biometrics is the lack of information about the relative variation in the different types of biometric data.

The term data variation, as used in this thesis, is now subdivided into two types. These are, variation in each data type arising from uncontrolled operating conditions, and variation in the relative degradation of data. The former variation can be due to operating in uncontrolled conditions (e.g. poor illumination of a user's face in face recognition, background noise in voice biometrics, etc.), or user generated (e.g. uncharacteristic sounds from speakers, carelessness in using the sensor for providing fingerprint samples, etc.)[1]. The variation in the relative degradation of data is due to the fact that in multimodal biometrics different data types are normally obtained through independent sensors and data capturing apparatus. Therefore, any data variation of the former type (discussed above) may in fact result in variation in the relative degradation (or goodness) of different biometric data deployed. Since, in practice, it may not be possible to fully compensate for the degradation in all biometric data types involved, the relative degradation of data appears as another important consideration in multimodal biometrics. This thesis reports a number of contributions to increasing the accuracy of multimodal biometrics in the presence of variation. These are based on investigating methods of tackling the effects of data degradation and estimating the relative quality of different biometric data.

1.2. Aims and Objectives

The main aim of this work is to investigate the effectiveness of fusion techniques for multimodal biometrics, with the following specific objectives:

- A review of the existing approaches.
- Investigations into effective fusion methods for selected types of biometrics (i.e. face and voice). These involve
    - fusion of different types of biometrics (voice and face), and

- fusion of complementary information in unimodal biometrics (voice/ face).

- Identification of the main issues and challenges in the use of fusion methods for multimodal biometrics. This involves
  - the effect of variation in relative degradation of data on the multimodal biometrics accuracy (i.e. face and voice).
  - the effect of variation arising due to uncontrolled operating conditions on the recognition performance of biometric systems.

1.3 Thesis Organisation

The thesis is organised into eight chapters. An overview of these chapters is presented below.

- Chapter 1 introduces the topic of multimodal biometric systems and gives the motivations for and outline of this PhD thesis.

- Chapter 2 describes different architectures for information integration, and presents a review of previous investigations into multimodal biometrics and details the motivations for this thesis based on the previous works.
- Chapter 3 identifies, based on the outcomes of the investigations carried out in the previous chapter, the more effective fusion methods and describes their principles in detail. The Chapter also introduces the most effective and widely used supporting techniques for multimodal biometrics reallocation and evaluation.
- Chapter 4 presents a thorough experimental investigation, based on two types of biometrics (i.e. face and voice), into the effectiveness of various fusion approaches in both unimodal and multimodal biometrics. The scope of the investigation includes the use of verification scores obtained from different types of features extracted from biometric data. The Chapter presents the experimental results together with an analysis of them.

- Chapter 5 studies the application of relative quality-based score level fusion to reduce the effects of relative degradation in multimodal fusion. The Chapter describes the experimental investigation and discusses the results.
- Chapter 6 presents an investigation into the effects, on the accuracy of multimodal biometrics, of introducing appropriate normalisation into the score level fusion process. The experimental investigations involve the two recognition modes of verification and open-set identification, both in clean, mixed-quality, and in degraded data conditions. The Chapter presents the motivation for, and the potential advantages of, the proposed approach and details the experimental study.
- Chapter 7 presents a qualitative fusion method using score normalisation to enhance the accuracy of multimodal biometrics. The Chapter introduces the motivation for the proposed approach and presents the experimental results together with an analysis of them.
- Chapter 8 presents a summary of the work carried out and its important conclusions. The latter part of the Chapter presents suggestions for future work.

Chapter 2

Literature Review

2.1 Introduction

This chapter focuses on various fusion methods for multimodal biometrics. Section 2.2 describes the categorisation of multimodal biometric systems into four architectures in accordance with the strategies used for information integration. Section 2.3 gives a summary of the main investigations carried out to date in the field of multimodal biometrics.

2.2 Fusion Levels

The literature shows that four possible levels of fusion are used for integrating data from two or more biometric systems[7, 8]. These are the sensor level, the feature level, the matching score level, and the decision level. The sensor level and the feature level are referred to as pre-mapping fusion while the matching score level and the decision level are referred to as post-mapping fusion [9]. In pre-mapping fusion, the data is integrated before any use of classifiers, while in post-mapping fusion, the data is integrated after mapping into matching score/ decision space. Figure 1 shows the four possible fusion levels.

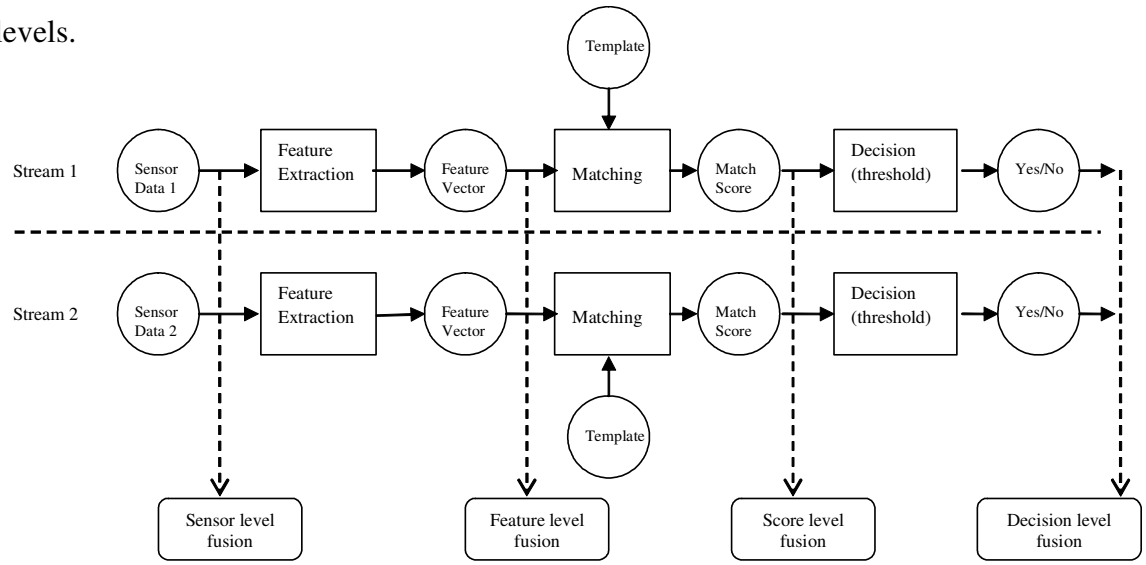


Figure 2.1: Fusion levels in multimodal biometric fusion.

A. Pre-mapping fusion I: Fusion at the sensor level

The raw data, acquired from sensing the same biometric characteristic with two or more sensors, is combined (Figure 1). An example of the sensor fusion level is sensing a speech signal simultaneously with two different microphones. Although fusion at such a level is expected to enhance the biometric recognition accuracy [7, 10], it can not be used for multimodal biometrics because of the incompatibility of data from different modalities [7].

B. Pre-mapping fusion II: Fusion at the feature level

Fusion at this level, as shown in Figure 1, can be applied to the extraction of different features from the same modality or different multimodalities [7]. An example of a unimodal system is the fusion of instantaneous and transitional spectral information for speaker recognition. On the other hand, concatenating the feature vectors extracted from face and fingerprint modalities is an example of a multimodal system. It is stated in [7, 10] that fusion at the feature level is expected to perform better in comparison with fusion at the score level and decision level. The main reason is that the feature level contains richer information about the raw biometric data. However, such a fusion type is not always feasible [7, 10]. For example, in many cases the given features might not be compatible due to differences in the nature of modalities. Also such concatenation may lead to a feature vector with a very high dimensionality. This increases the computational load. It is reported that a significantly more complex classifier design might be needed to operate on the concatenated data set at the feature level space[7].

C. Post-mapping fusion I: Fusion at the matching score level

At this level, it is possible to combine scores obtained from the same biometric characteristic or different ones. Such scores are obtained, for example, on the basis of the proximity of feature vectors to their corresponding reference material (Figure 1). The overall score is then sent to the decision module [4]. Currently, this appears to be the most useful fusion level because of its good performance and simplicity [11, 12] This fusion level can be divided into two categories: combination and classification. In the former approach, a scalar fused score is obtained by normalising the input matching

scores into the same range and then combining such normalised scores. In the latter approach, the input matching scores are considered as input features for a second level pattern classification problem between the two classes of client and the Impostor [13].

D. Post-mapping fusion II: Fusion at the decision level

In this approach, as shown in Figure 1, a separate decision is taken for each biometric type at a very late stage. This seriously limits the basis for enhancing the system accuracy through the fusion process. Thus, fusion at such a level is the least powerful [14].

2.3 Previous research into multimodal biometrics

This section provides a review of the outcomes of the investigations carried out to date in the area of multimodal biometric fusion. Due to the advantages offered by the score level fusion, the discussions are focused on this type of fusion. In literature, the score level fusion techniques are divided into two main categories of fixed rules (rule-based) and trained rules (learning-based)[15, 16]. The fixed rules are also referred to as the non-parametric rules while the trained rules are referred to as the parametric rules [17]. The main reason for categorising the fusion techniques in this way is that trained rules require sample outputs from the individual modalities to train the pattern classifiers. In other words, they use development data to calculate some required parameters. These parameters are then used to appropriately fuse the score data in the test phase. Examples of the trained rules are Weighted Sum rule and Weighted Product rule (Section 2.3.2). On the other hand, fixed rules are applied directly to fuse the given test scores for different modalities. In other words, the contribution of each modality is fixed a priori. Examples of fixed rules are AND rule, OR rule, Maximum, Minimum and Majority voting (Section 2.3.1). The next three sections discuss the previous research into fixed rules, trained rules, and a comparison between them.

2.3.1 Fusion based on fixed rules

This section provides a review of the recent work in the area of fixed-rule based biometric score fusion. The emphasis of the discussions is on the most popular methods in this category.

The discussions include both decision level fusion techniques (e.g. AND, rule, OR rule, Majority voting rule) and score level fusion techniques (e.g. Maximum rule, Minimum rule, Sum rule, Product rule, Mean rule). The next section has a brief description of such rules.

2.3.1.1 Fixed rules

A. AND fusion

In AND fusion, the outputs of different classifiers are thresholded. An acceptance decision is reached only when all the classifiers agree [18, 19].

B. OR fusion

In OR fusion, again the outputs of different classifiers are compared to a preset threshold. A positive decision is made as soon as one of the classifiers makes an acceptance decision [18, 19].

C. Majority voting rule

In Majority voting rule, the outputs of different classifiers are thresholded. In this case, reaching a decision is based on having the majority of the classifiers declaring the same decision [20-22]. To prevent ties, for a two class classification task, the number of classifiers must be odd and greater than two. The number of votes determines the security level of the system: the more the votes, the higher the security level.

For the fixed and trained rules presented in this thesis, f is the fused score, x_m is the score of the m^{th} matcher, m=1,2,…, M

D. Maximum rule

Maximum rule method selects the score having the largest value amongst the modalities involved. It is defined mathematically as [23, 24]:

$$f = \max\left(x_1, x_2, ..., x_M\right) \tag{2.1}$$

E. Minimum rule

In the Minimum rule, the match-score x_m represents the distance score. Minimum rule method chooses the score having the least value of the modalities involved. It is defined as [23, 24]:

$$f = \min\left(x_1, x_2, ..., x_M\right) \tag{2.2}$$

F. Sum rule

In Sum rule, the fused score is computed by adding the scores for all modalities involved. The computation here is defined as [23, 24]:

$$f = \sum_{m=1}^{M} x_m \tag{2.3}$$

G. Product rule

In Product rule, the fused score is calculated by multiplying the scores for all modalities involved. It is mathematically defined as [25]:

$$f = \prod_{m=1}^{M} x_m \tag{2.4}$$

H. Arithmetic Mean Rule

In the Arithmetic Mean rule, the fused score is obtained by first adding the scores for all modalities, and then dividing the result by the number of modalities involved. It is also known as the simple mean rule. Mathematically, the Arithmetic Mean rule is defined as [26]:

$$f = (\sum_{m=1}^{M} x_m) / M \tag{2.5}$$

2.3.1.2 Recent work on biometric fusion using fixed rules

This section contains a summary of the recent work on biometric fusion based on fixed rules.

The use of hybrid biometric person authentication based on face and voice features has been explored in a study presented in [27]. Although a simple logical AND scheme is used for the purposes of fusion, the experimental results have confirmed that a multimodal approach is better than any single modality.

The combination of scores for noisy speech and clean handwriting is the subject of investigations in [28]. The chosen fusion method in this study is the Sum rule. The results show that fusion of more than one modality could lead to better results compared with the use of only one modality.

Another study [29]considers two different fusion methods of integrating the scores or decisions for face, speech and lip movement. The methods considered are the Sum rule (score level) and Majority voting (decision level). Majority voting requires the agreement of two traits out of the three, although, for a higher security level, the system can demand the agreement of all three traits. It has been found that the combined system could provide more security than each of the individual systems involved.

The combination of face and gait cues for identification purposes has been studied in [30] and in this case the fusion process takes place at either the score level or the decision level. Four different score-level fusion methods and one decision-level fusion method are empirically compared in that study. These are the Product rule, Sum rule, Maximum rule, Minimum rule and Majority voting rule. The Product rule has shown the best performance out of all the fusion methods considered. The Minimum and Maximum rules demonstrate poor performance because of the high degree of overlap of the distribution of client and impostor scores. That has proved them to be less robust than Sum and Product rules.

An automatic person identification system is proposed in [31]. This system is based on the integration of the scores for clean face and fingerprint by simply multiplying the

scores obtained from the two. The results have shown that the integrated system could overcome even the best modality involved in that study.

There have been some experimental studies on the fusion of face and gait for a single camera case [32]. On this occasion two different scenarios are used in order to fuse the scores for face and gait. The first scenario involves the use of a gait classifier [33] as a filter in order to pass a smaller number of candidates to the more accurate face classifier [34]. In the second scenario the matching scores for the two considered modalities are directly combined. This is based on the Sum rule, Minimum rule and Product rule. Although both scenarios have shown overall systems performance improvement, the second scenario is preferred if the requirement from the fusion is that of accuracy as against computational speed.

Based on the above studies, it can be concluded that fusion techniques based on fixed rules have some advantages. These include the fact that such techniques usually show better performance when compared to single modalities involved (at least in the above-mentioned studies). Secondly, they are very simple techniques to implement. However, in another study [35], it has been indicated that the advantage of obtaining better performance based on the fixed rules might not be held when the ensembles of involved modalities are not exhibiting similar performance [35]. Unfortunately, one of the main problems that is related to multimodal biometrics fusion is that individual biometrics often show significantly different performance [35, 36]. Thus, using fixed rules for multimodal biometric fusion might degrade the performance of the fused system compared to the performance of the best individual modality involved. Hence, trained rules are introduced for multimodal biometric fusion as an alternative approach to the fixed rules.

2.3.2 Fusion based on trained rules

A description of the important multimodal biometric fusion methods based on trained rules is presented in this section. These are Weighted Sum rule, Weighted Product rule, Fisher Linear Discriminant, Quadratic Discriminant Analysis, Logistic Regression, Support Vector Machine, Multi-Layer Perceptrons and Bayesian classifier. This is followed by a summary of the recent work in this field.

2.3.2.1 Trained rules

A. Weighted Sum rule

This is also known as the weighted average rule (see Section 3.4.1). In this technique, the fused score is obtained through a two-stage task. Firstly, each score is multiplied by the corresponding weight of its modality. Secondly, the multiplication results are added together in order to produce the fused score. This is mathematically represented as:

$$f = \sum_{m=1}^{M} w_m x_m \tag{2.6}$$

where f is the fused score, M is the number of matching streams, x_m is the match score from the m^{th} matcher and w_m is the corresponding weight (obtained on some development data) in the interval of 0 to 1, with the condition

$$\sum_{m=1}^{M} w_m = 1 \tag{2.7}$$

There are several sub-classes of this scheme, which differ primarily in the method used for the estimation of weight values (e.g. Brute force Search, Matcher Weighting, User Weighting [12]). More details about Weighted Sum rule are given in Section 3.4.1.

B. Weighted Product rule

Like the Weighted Sum rule, the Weighted Product rule is also a two-stage task. However, the Weighted Product rule differs from the Weighted Sum rule in the second stage. In this case, the products of the weight and score from each modality are multiplied instead of being added. Mathematically, the scores from M modalities are combined as follows [26, 37]:

$$f = \prod_{m=1}^{M} w_m \, x_m \tag{2.8}$$

where f is the fused score, M is the number of matching streams, x_m is the match score from the m^{th} matcher and w_m is the corresponding weight. It should be noted that w_m is computed on development data.

C. Fisher Linear Discriminant

Fisher Linear Discriminant (FLD) is a simple linear projection of the input vector. It is defined as:

$$h(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + b \tag{2.9}$$

where *T* indicates the transpose operation and the estimated output $h(\mathbf{x})$ is a function of the input vector **x** as well as the parameters w and *b*. Such parameters are obtained through an appropriate training procedure (see Section 3.4.2)

D. Quadratic Discriminant Analysis

Quadratic Discriminant Analysis (QDA) is similar to FLD but is based on forming a boundary between two classes using a quadratic equation. More details about this technique are given in Section 3.4.3.

E. Logistic Regression

The Logistic Regression method classifies the data based on using two functions: logistic regression function (2.10) and logit transformation (2.11) as follows:

$$E(Y \mid \mathbf{x}) = \frac{e^{g(\mathbf{x})}}{1 + e^{g(\mathbf{x})}} \tag{2.10}$$

where, $E(Y \mid \mathbf{x})$ is the conditional probability for the binary output variable Y where the *M*-dimensional input vector **x** exists, with:

$$g(\mathbf{x}) = w_0 + w_1 \cdot \mathbf{x}_1 + ... + w_M \cdot \mathbf{x}_M \tag{2.11}$$

where w_m is the weight for the m^{th} modality. Such weights are calculated during the development stage. More details about LR can be found in Section 3.4.4.

F. Support Vector Machine

Support Vector Machine (SVM) is based on the principle of Structural Risk Minimisation (SRM) which aims to find the optimal separating hyper-plane that has the largest margin to the closest data points in the two classes being separated. SVM is simply defined for the linearly separable data as:

$$\mathrm{w}^T \cdot \mathbf{x}_i + b = 0 \tag{2.12}$$

where w is a weight (coefficient) vector, **x** is an input vector consisting of the scores for different modalities and b is a bias estimated on the development set. This definition could be then generalised for non-linearly separable data. This is achieved through some non-linear functions (e.g. Radial Basis Function Support Vector Machine, Polynomial kernel function). More details about SVM are given in Section 3.4.5.

G. Multi-Layer Perceptrons

A Multi-Layer Perceptrons (MLP) is a particular architecture of artificial neural networks [38, 39]. The MLP architecture should have an input layer, hidden layer(s) and an output layer. With no hidden layer, the perceptron can only perform linear tasks. Classification in MLP is achieved by processing the input scores through successive layers of "neurons". For a two-class problem (c=1,2), an example of a MLP with one hidden layer can be written mathematically as:

$$y_c = f_o\left(\sum_{j=0}^{M_H} w_{cj}\, f_h\left(\sum_{i=0}^{M_I} w_{ji}\, x_i\right)\right) \tag{2.13}$$

where y_c is the c^{th} output, w_{ji} are the input-to-hidden weights, w_{cj} are the hidden-to-output weights, x_i is the i^{th} input, M_I and M_H are the number of input and hidden nodes respectively, and f_h and f_o are the sigmoid activation functions for the hidden and output layers respectively. The weights in MLP are calculated using the development data through an appropriate training procedure [38, 39].

H. Bayesian classifier

The Bayesian classifier is a simple classification method. In the Bayesian classifier, the classification requires the estimation of many conditional distributions [40, 41]. This in turn requires large amounts of training data. The Bayesian classifier, in the case of a two-class problem (C for clients and I for impostors, or C_i, i=1,2), can be described as

follows: let **x** be *M*-dimensional input vector. The a posteriori probability $p(C_i/\mathbf{x})$ can be computed as:

$$p(C_i/\mathbf{x}) = \frac{p(\mathbf{x}/C_i)\cdot p(C_i)}{p(\mathbf{x})} \tag{2.14}$$

where $p(C_i)$ and $p(\mathbf{x})$ are the a priori probabilities of C_i and **x** respectively, and $p(\mathbf{x}/C_i)$ is the conditional probability of **x**, given C_i. Since $p(\mathbf{x})$ does not depend on the class index, the maximum a posteriori decision only depends on the numerator of the right-hand side of equation (2.14) [41].

$$MAP = \max_i p(\mathbf{x}/C_i)\cdot p(C_i) \tag{2.15}$$

where $p(C_i)$ is computed on the development data.

2.3.2.2 Recent work on biometric fusion using trained rules

A review of recent work carried out on biometric fusion based on trained rules is presented below.

A multimodal biometric user-identification system is proposed in [42]. The system is based on combining the scores for hand geometry, palm and fingerprint. The Weighted Sum rule is used as the fusion technique. Results of this experiment show that the Weighted Sum rule provides better performance than even the best individual modality.

In another study [43] the integration of fingerprint, face and hand geometry at the score level is explored. That study demonstrates that user independent Weighted Linear combination of similarity scores can be enhanced by using either user dependent weights or user dependent decision thresholds. Weights and thresholds are computed by exhaustive search on the development data. It has been found that using thresholds improves the performance by about 2%, whilst the use of weights improves it by about 3%.

In [44], the face and speaker identification techniques are tested on data collected in uncontrolled environments using inexpensive sound and image capture hardware. Despite the fact that the system performance can be harmed under these circumstances, it has been proved that using a combination of biometric modalities can improve the

robustness and accuracy of the person identification task. Simple Brute Force Search is used for the fusion process in that study too. It has been shown that through this approach the fused biometrics outperforms the two individual modalities involved.

Another multimodal person verification system (based on facial profile views and features extracted from speech), which integrates the scores obtained from the two biometrics using the Weighted Sum rule, is considered in [45]. The results of a study on that system have shown that the improvement gained through the integration of scores is particularly noticeable when operating under noisy conditions.

A demonstration biologin system, with bimodal (face and speech) authentication system, is proposed in [46]. The scores for face and speech modalities are combined using a Weighted Sum rule. The experimental results show that the fusion has led to better performance compared to the individual modalities involved.

In another study [4] the integration of face, fingerprint and hand geometry at the score level is explored. The results of the study show that using Weighted Sum rule as the fusion process has led to considerable improvement.

A multimodal identity verification system based on the integration of the scores for face image and text independent speech data of a person [47] has used Multi-Layered Perceptron and Weighted Average for the fusion purpose. In a study of that system it has been found that the text independent speaker verification algorithm is more robust compared to the face verification algorithm. Nevertheless, fusion of these two modalities has led to considerable improvement.

Two different speaker verification algorithms have been discussed in another study which considers a robust person verification system based on speech and facial images [48]. These are a text independent method using a second order statistical measure and a text dependent method based on hidden Markov modeling. As a unimodal verification system, text dependent has shown the best performance compared to face and text independent modules. Support Vector Machine is used to integrate the scores for the different recognition modules and it has been found that the combination of different modalities outperforms even the best individual modality involved. Results have also shown that the combination of the two modules with the lowest performance (face and

text independent) leads to better performance than the best single modality (text dependent).

In another study [49] a trained supervisor based on Bayesian statistics has been used to combine scores from face and speaker voice, using the modulus of complex Gabor responses [50] as a face feature, and representing speech using LPC (linear predictive coding) features [51]. The results of that study have shown that the proposed system outperforms the aggregation of the individual modalities by averaging.

Audio-visual person verification based on frontal face image and speech has been considered in [52]. In doing this, Linear Weighted and SVM are used as fusion techniques and the results have shown that the performance of the system is increased by combining the two modalities. The Linear Weighted classifier has outperformed the Linear SVM, but the SVM is demonstrated to have possessed an advantage in combining potentially any number of modalities at the same computational cost with very good fusion results (1999, p. 8)[52].

A study which proposes an adaptive multimodal person verification system based on speech and face images has found that the system adapts to noise present in the speech signal by modifying the parameters of the fusion method [53]. A set of parameters for different Peak Signal to Noise Ratio (PSNR) of the speech signal is calculated a priori during the development stage. Then, during the test stage, the estimation of the PSNR of the given speech signal takes place and parameters most closely corresponding to that PSNR are used by Linear and SVM fusion methods. The results have demonstrated that the adaptive system significantly outperforms the non-adaptive one.

2.3.3 Comparison of fixed rules and trained rules

As indicated earlier, the main aim of this chapter is to provide a direct comparison between fixed rules and trained rules in the field of multimodal biometric fusion. However, any direct comparison between the results of the above studies would be meaningless. This is due to the fact that the work reviewed above is based on using different databases and/or different experimental setup (e.g. modalities and performance measures). Therefore, this section aims to compare the effectiveness of fixed rules and trained rules based on studies involving the same database (same population and size).

In [12], five different score fusion methods (Matcher Weighting, User Weighting, Sum rule, Minimum rule, Maximum rule) are used to combine the scores obtained from fingerprint and face data. Different range-normalisation techniques are also introduced: Min-Max, Z-score, Tanh and Quadric-Line-Quadric [12]. The study has shown that trained rules, particularly through User Weighting scheme, lead to the best performance, whilst Minimum rule (fixed rule) leads to the worst.

In another study[26, 54], the combination of scores or decisions obtained from face and speaker voice has been considered. The fusion methods used are the Weighted Sum rule, Weighted Product rule, Maximum rule, Minimum rule, Majority voting rule and Median rule. Results in this case show that the Weighted Sum rule has outperformed the other fusion methods. The robustness of Weighted Sum rule in term of errors made by individual classifiers is shown to account for the better performance.

The integration of speaker voice and frontal face image by using trained-rule and fixed-rule fusion methods is considered in [16]. The study has shown that the advantages of trained rules depend strongly on the quality and size of the development set. It has also been found that the performance of both fixed and trained rules is affected by the correlation between the outputs of the different features. For example, fixed rules have performed well for modalities exhibiting a similar correlation while it has shown that trained rules should handle modalities exhibiting different correlation more effectively [35, 55]. However, it has not been completely clear under which conditions of performance imbalance trained rules can significantly outperform fixed rules.

Two other studies which have provided a direct comparison between the rule-based fusion with learning-based (or trained) fusion are presented in [15, 56]. In both studies, Sum rule and Radial Basis Function Support Vector Machine (RBF SVM) are used to combine the scores for face, fingerprint and online signature. The experimental results of these studies have shown that appropriate selection of parameters for a learning-based scheme (RBF SVM) leads to a fusion strategy that clearly outperforms the rule-based strategy (Sum rule).

A two-level fusion strategy for audio-visual biometric authentication is proposed in [57]. The fusion is performed sequentially, first at intramodal fusion level, and then at intermodal fusion level. At the intramodal fusion level, the scores of multiple samples

(e.g. utterances or video shots) obtained from the same modality are linearly integrated, and this can be done either by assigning equal weights or different weights to different scores [57]. The process is followed by intermodal fusion. At the intermodal fusion level, the intramodal fused scores obtained from different modalities are combined by using the Sum rule or second-degree polynomial SVM. The experimental results have shown that the two level fusion (intramodal and intermodal) are complementary to each other and that the intermodal fusion is best realised through SVM.

In [58], another comparison of trained rules and fixed rules is presented. In that study, the effectiveness of Arithmetic Mean Rule (AMR) and SVM are compared in a multimodal biometric score fusion based on the integration of the scores for speech and online signature modalities in two different experimental conditions. It is shown that in the first case, clean speech and clean online signature data, AMR with Min-Max gives the best performance, while in the second (noisy) case, with degraded speech data and clean online signatures, SVM gives performance equivalent to those obtained with AMR after score range-normalisation via a Bayes normalisation [58]. Such results are much better than those given by the AMR with Min-Max and Tanh Estimator. For the fusion by AMR, three score range-normalisation methods are used, Min-Max, Tanh Estimator and Bayes normalisation. For the SVM no range-normalisation is used.

The development of a prototype for a multimodal biometric system by using single sensor [59] has shown that Radial Basis Function Support Vector Machine (RBF SVM) outperforms other combined and individual classifiers. Sum rule, Weighted Sum rule and RBF SVM are used in this case to integrate the scores for hand geometry and palm print.

The comparison of 13 different classifier-combination methods based on the fingerprint and voiceprint matching scores for both identification and verification [60] has shown that SVM leads to the best performance out of a wide range of different fusion methods.

A study using Mean rule, Linear Support Vector Machines and Radial Basis Function Support Vector Machines for fusing the scores for fingerprint and iris biometrics has been presented in [61]. The study demonstrates the benefit of integrating the scores for fingerprint and iris modalities. The two SVM approaches outperform the simple Mean rule, and it is shown that a multimodal integration of iris and fingerprints can offer

substantial performance gain that may not be possible with a single biometric indicator alone [61].

The integration of face and speaker voice by using parametric and non-parametric fusion methods is considered in [17]. The experimental results have shown that multimodal verification derived from Logistic Regression works best, although it is indicated that answering the question of which fusion method should be chosen is much more difficult. It has been found that having a large representative database is very important in order to choose a number of potentially powerful fusion paradigms.

The demonstration that multimodal biometric recognition is best realised through trained rules, particularly through the Bayesian method and SVM in [62], involves the combination of four different modalities (fingerprint, face, voice and signature). The fusion methods used in that study include trained rules (represented by Logistic Regression, Multi-Layer Perceptrons, Quadratic Classifiers, Linear Classifiers, SVM, Bayesian classifier) and fixed rules (represented by AND rule, OR rule, majority voting). The results have shown that the Bayesian approach, although considered optimal, requires far more data. However, the data quantity is far less of a concern in the SVM paradigm.

The integration of profile image, frontal image and voice by using parametric and non-parametric fusion methods is considered in [63]. The binary classifier derived from Logistic Regression has produced a more balanced approach and positive indications: a low level of computing time and good results.

In another study, a multimodal identity verification system is proposed using expert fusion to integrate the results obtained from vocal and visual biometric modalities [64]. Paradigms of parametric (Logistic Regression, Quadratic Classifier, Linear classifier and Multi-Layer Perceptrons) and non-parametric (AND rule and OR rule) classes of techniques are used as fusion algorithms. Logistic Regression has seemed to give the overall best performance. Also, results have shown that the parametric methods outperform the non-parametric schemes.

Based on the above investigations, it can be concluded that an accurate design of the best reported trained rules usually outperform the fixed rules even though some examples have been reported in the literatures where the Sum rule has outperformed other trained

rule approaches [4]. However, trained rules still have some disadvantages that may result in degrading the overall accuracy for the combined system. One of the main disadvantages of the trained rules is their rather high degree of complexity [65]. Another disadvantage of trained rules is the scarcity of multimodal data [65] especially because of the trained rules' strong dependence on the quality and size of the development set. Some of the limitations (disadvantages) of trained rules can be overcome by incorporating the quality of the biometric modalities involved in the fusion process (Section 2.3.4).

2.3.4 Quality-Based Fusion

Several studies have shown that the quality of a unimodal biometric sample plays a significant role in the overall system performance [66, 67]. In those studies, poor quality of biometric samples leads to a significant reduction in the accuracy of a unimodal biometric system. As said earlier multimodal biometric systems can overcome this challenge to some extent by integrating the evidence provided by a number of different biometrics. However, one of the important problems associated with score-level fusion for multimodal biometrics is the unpredicted variations in the evidence captured in the scores. Such variation can arise from anomalies such as background noise, communication channel and uncharacteristic disturbances in the modalities. One of the approaches to tackling such anomalies is based on explicit estimation of the quality aspects of the generated scores. This section presents a review of the recent work on the incorporation of quality measures in multimodal biometric systems.

One straightforward way to introduce the quality measures of the input biometric data into the score level fusion approach is through including weights in simple combination approaches (for instances, Weighted Sum rule, Fisher Linear Discriminant). The weights in these approaches can be calculated heuristically, by exhaustive search in order to minimise certain error criterion on a development set (e.g. Brute Force Search), or by using a trained approach based on linear classifiers. After calculating the weights w_m, q_m can be obtained as:

$$q_m = w_m \tag{2.16}$$

and the quality-based score fusion function is achieved as follows:

$$f = \sum_{m=1}^{M} q_m x_m \tag{2.17}$$

where f is the fused score, M is the number of matchers, x_m is the match score from the m^{th} matcher and q_m is the quality measure of the score x_m and w_m is the weight of the score x_m .

In another study [68], it has been proposed to use the margin between impostor and client score distributions as a quality measure. In other words, the quality is derived based on a function of False Acceptance (FA) and False Rejection (FR) Rates, which themselves are estimated at the development stage. A commonly used point to examine the quality of performance is to calculate the point "threshold" of Equal Error Rate (EER), which assumes that the cost of FA and FR are equal. This determines simply how confident a score is. The further the score is from the threshold "decision boundary", the more confident it is. For the fusion purpose, a quality-weighted sum rule is used. The study has shown that fusion using margin information is superior to fusion without the margin information.

The use of confidence measures for multimodal (face and voice) identity verification has also been considered in [69]. A discussion on the influence of using the confidence of unimodal scores on three different fusion techniques: MLP, SVM and Bayesian classifier as density estimators, is carried out in that study. The confidence over a score is estimated based on three different methods: Gaussian hypothesis of the score distribution; Non Parametric Estimation; and the Model Adequacy.

In the Gaussian hypothesis, the measure of confidence for a given score x is the distance between the probability that the score x is from a client and the probability that the score x is from an impostor. This is calculated, under the assumption that all the scores x from clients have been generated by the same Gaussian distribution $N(\mathrm{x}, \mu_c, \sigma_c)$ and all the scores x from impostors have been generated by another Gaussian distribution $N(\mathrm{x}, \mu_i, \sigma_i)$, as follows:

$$m(x) = \left| N(\mathrm{x}, \mu_c, \sigma_c) - N(\mathrm{x}, \mu_i, \sigma_i) \right| \tag{2.18}$$

where $m(x)$ is the measure of confidence for the score x, $N(\mathrm{x}, \mu, \sigma)$ is the hypothesised Gaussian, μ_c and σ_c are the mean and variance for the client scores and μ_i and σ_i are the mean and variance for the impostor scores. All these parameters are computed using the development scores.

In Non-Parametric Estimation, the space of the development scores is partitioned into k distinct subspaces, where each partition contains the same number of development scores. Then, the number of errors that occurred in each partition for the development data (FA and FR), divided by the total number of scores in that subspace is computed as follows[69]:

$$m(\mathrm{x})_k = \frac{(\text{number of FAs})_k + (\text{number of FRs})_k}{(\text{number of accesses})_k} \tag{2.19}$$

where $m(\mathrm{x})_k$ is the confidence measure for the k^{th} subspace. This number gives a simple confidence on the quality of the development scores in corresponding subspace. However, at the test stage, a confidence of a given score x is obtained by finding the subspace corresponding to the given score and returning the associated confidence measure.

In the case of Model Adequacy approach, it is proposed to calculate the gradient of a simple measure of confidence of the decision of the model given an access with respect to each parameter in the model. This is based on the fact that most unimodal verification systems are based on some kind of gradient method optimising a given criterion (e.g. in speaker verification, GMMs are trained to maximise the likelihood while in face verification, MLPs are trained to minimise the mean square error). The average amplitude of such gradient leads to an idea of the adequacy of the parameters to explain the confidence of the model on the access. This is computed as[69]:

$$m(\mathrm{x}) = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{f(\mathrm{x})}{0_i}\right| \tag{2.20}$$

where $m(\mathrm{x})$ is a global confidence measure for the current model, 0_i is one of the N parameters of the model and $f(\mathrm{x})$ is a simple measure of confidence of the model given access x.

Experimental results have shown that all three fusion methods provide better performance than even the best individual modality involved (in this case, voice modality). It has been found that SVM leads to slightly better performance compared to the other two fusion techniques. It has also been found that the performance of both SVM and MLP has been enhanced further through the use of Model Adequacy as a confidence measure. However, the two other confidence methods have not appeared to improve the performance of SVM and MLP significantly. On the other hand, none of the confidence methods has been able to enhance the performance when the Bayesian classifier is used as the fusion technique.

2.4 Summary

Four possible levels of fusion are used for integrating data from two or more biometric systems or sources. These levels are: the sensor level; the feature level; the matching score level; and the decision level. Fusion at the matching score level has been viewed as the most prevalent and useful technique for the integration of biometric data. The score level fusion in multimodal biometrics can be obtained by two different approaches: fixed rules and trained rules. The results of earlier investigations suggest that Multimodal biometric recognition is best realised through trained rules. The investigations have also indicated that the accuracy of multimodal biometric system can be further improved by incorporating quality in the score level fusion. Such incorporation can be achieved by first estimating the quality of the biometric samples and then adaptively weighting the individual biometric scores based on the quality values.

Since the focus of this study is the performance of the fusion methods, the rest of this study concentrates on the trained rules only. Based on the earlier investigations, Support Vector Machines and Logistic regression have shown better performance among the trained rules. The theories of these two techniques plus other various fusion methods are discussed in the next chapter.

Chapter 3

Fusion Techniques for Multimodal Biometrics

3.1 Introduction

The previous chapter indicated that trained rules should lead to better performance than fixed rules. This chapter provides a further description of the trained rules-based fusion methods identified in Section 2.3 as the best performers. Since the description of the fusion methods considered involves range-normalisation techniques as well as the measures used for evaluating identity recognition performance, these are introduced in the first part of the chapter as follows. Section 3.2 discusses the most effective and widely used score range-normalisation techniques. Section 3.3 describes the commonly used measures (those adopted in this study) for evaluating identity recognition performance.

3.2 Range-Normalisation Techniques

Range-normalisation is also known as score normalisation [11, 23, 70-73]. The term range-normalisation is used throughout this thesis to define the task of bringing raw scores from different matchers to the same range. On the other hand, the term score normalisation is used in this thesis to define the process of enhancing the scores from the degraded modalities (see Chapters 6 and 7).

Range-normalisation is a necessary step in any fusion system, as fusing the scores without such normalisation would de-emphasise the contribution of the matcher having a lower range of scores. A number of comparative studies in the literature have discussed the effects of range-normalisation prior to fusion. For example, it is indicated in [70] that range-normalisation is a necessary task because scores from different systems are incomparable. Another study [71] states that in the case of using linear fusion techniques to integrate the scores of the individual modalities, score incomparability affects the system performance. The study concluded that range-normalisation is a necessary task before fusion. The influence of range-normalisation techniques prior to fusion in biometric authentication tasks is also explored in detail in [11, 23, 72, 73]. According to the literature, there are various well-known range-normalisation techniques (i.e. Min-Max, Z-score, Tanh, Median-MAD, Double-sigmoid). Min-Max and Z-score (in most

cases) have shown to be amongst the most effective and widely used methods for this purpose [23, 73].

3.2.1 Min-Max Normalisation (MM)

This linear technique maps the raw scores into the range of [0 1]. Min-Max normalisation conserves the distribution of scores before and after normalisation (Figure 3.1(a and b)). This method uses the following equation

$$x = \frac{n - \min}{\max - \min} \tag{3.1}$$

where, x is the normalised score, n is the raw score, and max and min functions specify the maximum and minimum end-points of the score range respectively and are obtained on some development data.

3.2.2 Z-score Normalisation (ZS)

Z-score normalisation converts the scores to a distribution with the mean of 0 and standard deviation of 1. Like Min-Max normalisation, Z-score normalisation also retains the original distribution of the scores (Figure 3.1(a and c)). However, the numerical range after Z-score normalisation is not fixed. Z-score normalisation is given as

$$x = \frac{n - \mu}{\sigma} \tag{3.2}$$

Where, n is any raw score, and μ and σ are the mean and standard deviation of the stream specific scores and are computed on some development data.

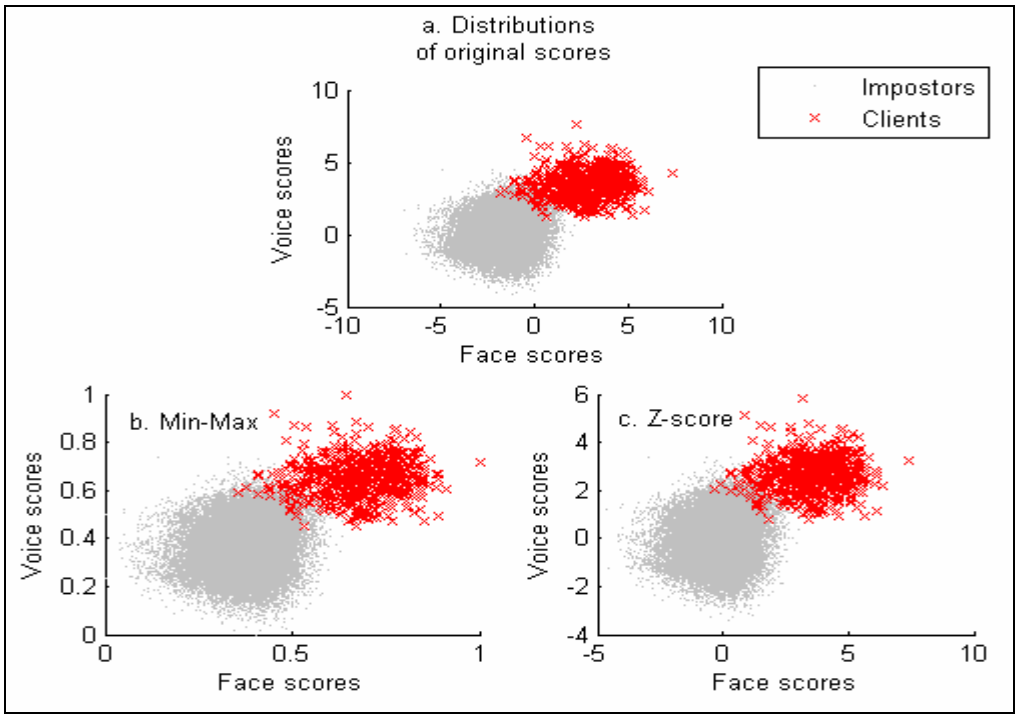


Figure 3.1: Effects of Min-Max and Z-score on the distributions of original face and voice scores.

In Figure 3.1, the subfigure (a) represents the original distributions of face and voice scores whilst the other subfigures (b and c) illustrate the effects of applying Min-Max and Z-score range normalisation techniques respectively. Comparing Figure 3.1(a) with the Figure 3.1(b and c) further illustrates the fact that Min-Max and Z-score normalisation techniques retain the original distribution of the scores. Furthermore, the subfigures show that Min-Max maps the scores in the range of [0 1]. On the other hand, the numerical range after Z-score normalisation is not fixed.

3.3 Evaluation Criteria for identity recognition

As said earlier, a biometric recognition system can operate in one of the two modes of verification and identification. In the verification mode, the user makes an identity claim. In this case the test data is compared only against the reference data (e.g. template, statistical model) associated with the claimed identity. The outcome of this is used to accept or reject the identity claim. In identification, the test data is compared with the data for all the registered individuals to determine the identity of the user. The following

subsections discuss the evaluation criteria for the verification and identification processes.

3.3.1 Evaluation Criterion for identity verification

In the task of verification, there are four different decisions that could be taken in response to a person claiming an identity. These decisions are [1, 39]:

- Accept a client
- Accept an impostor
- Reject a client
- Reject an impostor

Thus, the verification system may make two types of errors:

- False acceptance (FA): when the system accepts an impostor.
- False rejection (FR): when the system rejects a client

The performance of the system can be measured in terms of these two different errors as follows:

$$\text{FAR}=\frac{\text{number of FAs}}{\text{number of impostor accesses}} \tag{3.3}$$

$$\text{FRR}=\frac{\text{number of FRs}}{\text{number of client accesses}} \tag{3.4}$$

In practice, a perfect identity verification (FAR=0 and FRR=0) is unachievable. However, changing the decision threshold can reduce any of the two (FAR, FRR) to an arbitrary small value with the drawback of increasing the other one. The trade-off between FAR and FRR can be graphically represented by a Receiver Operating Characteristics (ROC) plot or a Detection Error Trade-off (DET) plot which is considered in this thesis [74]. This is because the DET plot is on a log scale which can enhance the visual appearance of the curves, whereas the ROC plot is on a linear scale. In a DET plot, the horizontal axis shows the normal deviate of the False Acceptance Rate (in%). The vertical axis of the DET plot represents normal deviate of the False Rejection Rate (in%). In the DET plot, the curves move away from the lower left when performance is low. Each point on a DET curve corresponds with a particular decision threshold.

In order to quantify the system performance into a single number, the verification performance is obtained in terms of Equal Error Rates (EER). EER is obtained when the FAR and FRR are the same at some decision threshold. Another performance measure can be obtained once EER is calculated. This is the so-called Relative Improvement (RI). The Relative Improvement is a measure of performance-enhancement achieved through fusion techniques. In other words, the measure determines the extent to which a fusion approach increases or decreases the biometric verification error rate compared with the best achievable performance without fusion. Mathematically, RI is expressed as [38]:

$$\mathrm{RI} = \frac{\min(\mathrm{EER}_1, \mathrm{EER}_2, ..., \mathrm{EER}_M) - \mathrm{EER}_f}{\min(\mathrm{EER}_1, \mathrm{EER}_2, ..., \mathrm{EER}_M)} \tag{3.5}$$

where $\mathrm{EER}_1, \mathrm{EER}_2, ...\mathrm{EER}_M$ are the equal error rates (EERs) resulting from *M* individual unimodal biometric verification schemes, and EER_f is the EER obtained through the fusion of these. With reference to equation (3.5) it is evident that RI can have a maximum value of one, indicating the fusion scheme adopted has resulted in a zero EER. On the other hand, a zero RI reflects the fact that there has been no improvement through the fusion used over the best individual biometric scheme. Finally, when the fusion adopted leads to the degradation of the verification accuracy, this is reflected by a negative RI. Therefore, for a fusion scheme to be beneficial, RI should be positive, and the closer it is to 1, the better is the effectiveness of fusion.

3.3.2 Evaluation Criterion for identification

Identification can be subdivided into two further categories of Closed-Set and Open-Set Identification problems. The Closed-Set Identification is to identify a person from a group of known (registered) people. On the other hand, in the Open-Set Identification problem, the person to be identified may or may not be one of the known (registered) people. Among these two categories, the Thesis focuses only on the Open-Set Identification problem.

Open-Set Identification consists of two stages of identification and verification. The performance of the verification process is evaluated as discussed in Subsection 3.3.1. In this case, the verification performance is expressed in terms of Open-Set Identification

Equal Error Rates (OSI-EER). On the other hand, the identification performance is expressed in terms of Identification Error Rate (IER) which is evaluated as follows:

$$\text{IER} = \frac{\text{number of incorrectly classified clients}}{\text{number of client accesses}} * 100 \quad \% \tag{3.6}$$

3.4 Effective fusion techniques

This section discusses the most effective trained rules-based fusion methods as identified in Section 2.3. These are Weighted Average, Fisher Linear Discriminant (FLD), Quadratic Discriminant Analysis (QDA), Logistic Regression (LR) and Support Vector Machines (SVM).

3.4.1 Weighted Average Fusion

In weighted average schemes, the fused score for each class (e.g. client or impostor) is computed as a weighted combination of the scores obtained from *M* matching streams as follows:

$$f = \sum_{m=1}^{M} w_m x_m \tag{3.7}$$

where f is the fused score, x_m is the normalised match score from the m^{th} matcher and w_m is the corresponding weight (obtained on some development data) in the interval of 0 to 1, with the condition

$$\sum_{m=1}^{M} w_m = 1 \tag{3.8}$$

As indicated earlier, there are three sub-classes of this scheme, which differ primarily in the method used for the estimation of weight values. These are described below.

3.4.1.1 Brute Force Search (BFS)

This fusion technique can be used in the case of having two matcher types only. The approach is based on using the following equation[75].

$$f = x_1 w + x_2 (1 - w) \tag{3.9}$$

where f is the fused score, x_m is the normalised score of the m^{th} matcher, m=1 or 2 and w is a weighting (combination) factor in the range 0 to 1. The weight (w) is

calculated heuristically, by exhaustive search in order to minimise the Equal Error Rate on the given development data.

3.4.1.2 Matcher Weighting using False Acceptance Rate and False Rejection Rate (MW – FAR/FRR)

This fusion technique can be used again in the case of having two matcher types only. In this technique the performance of the individual matchers determines the weights so that smaller error rates result in larger weights. The performance of the system is measured by False Acceptance Rate (FAR) and False Rejection Rate (FRR). These two types of errors are computed at different thresholds. The threshold that minimises the absolute difference between FAR and FRR on the development set is then taken into consideration. The weights for the respective matchers are computed as follows [76].

$$w_1 = \frac{1-(FAR_1 + FRR_1)}{2-(FAR_2 + FRR_2 + FAR_1 + FRR_1)} \tag{3.10}$$

and

$$w_2 = \frac{1-(FAR_2 + FRR_2)}{2-(FAR_1 + FRR_1 + FAR_2 + FRR_2)} \tag{3.11}$$

where FAR_1, FRR_1 and w_1 are the false acceptance rate, false rejection rate and the weight for one matcher and FAR_2, FRR_2 are the false acceptance rate, false rejection rate for the other matcher with the weight w_2. Note that the weight (obtained on some development data) is in the interval of 0 and 1, with the constraint $w_1 + w_2 = 1$

The fused score using different matchers is given as

$$f = w_1 x_1 + w_2 x_2 \tag{3.12}$$

where, x_m is the normalised score of matcher m and f is the fused score.

3.4.1.3 Matcher Weighting based on Equal Error Rate (MW - EER)

The matcher weights in this case depend on the Equal Error Rates (EER) of the intended matchers for fusion. These EERs are computed using the given development data. EER of matcher m is represented as E_m , $m=1,2,\ldots,M$ and the weight w_m associated with matcher *m* is computed as [12].

$$w_m = \frac{1}{E_m \left(\sum_{m=1}^{M} \frac{1}{E_m} \right)} \tag{3.13}$$

Note that $0 \le w_m \le 1$, with the constraint given in (3.8). It is apparent that the weights are inversely proportional to the corresponding errors in the individual matchers. The weights for less accurate matchers are lower than those of more accurate matchers. The fused score is calculated in the same way as in equation (3.7).

3.4.2 Fisher Linear Discriminant (FLD)

FLD is a simple linear projection of the input vector **x** onto a uni-dimensional space so that a linear boundary between classes can be satisfactorily obtained. The Equation for the linear boundary is given as [38, 39, 77-79]

$$h(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + b \tag{3.14}$$

where, **w** is a transformation vector obtained on the development data using a Fisher criterion (described in the next section), *T* is the transpose operation, and *b* is a threshold determined on the development data to give the minimum error of classification in respective classes. The rule for class allocation of any data vector is given by

$$\mathbf{x} \in \begin{cases} C_1 \\ C_2 \end{cases} \text{ if } \mathbf{w}^T \mathbf{x} + b \begin{cases} > \\ < \end{cases} 0 \tag{3.15}$$

where, C_1, C_2 are the client and impostor classes respectively.

3.4.2.1. Fisher Linear Discriminant for the Data from Two Classes

Given a set of $\mathbf{N_1}$ points for class C_1 and $\mathbf{N_2}$ points for class C_2, with the statistics $[\boldsymbol{\mu}_i, \mathbf{S}_i], i \in 1 and\, 2$, where $\mathbf{S}_i$ and $\boldsymbol{\mu}_i$ are the scatter (covariance) matrix and mean for the particular class *i* obtained on the development data, the scatter matrix is given as [38, 39, 77-79]

$$\mathbf{S}_i = \sum_{k \in C_i} (\mathbf{x}_k - \boldsymbol{\mu}_i)(\mathbf{x}_k - \boldsymbol{\mu}_i)^T \tag{3.16}$$

where, *T* indicates the transpose operation.

The overall within class scatter matrix $\mathbf{S}_W$ is given by

$$\mathbf{S}_W = \sum_{i=1}^{2} \mathbf{S}_i \tag{3.17}$$

The transformation vector **w** is obtained using the equation

$$\mathbf{w} = \mathbf{S}_W{}^{-1} (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) \tag{3.18}$$

3.4.3 Quadratic Discriminant Analysis (QDA)

This technique is similar to FLD but is based on forming a boundary between two classes using a quadratic equation given as [80]

$$h(\mathbf{x}) = \mathbf{x}^T \mathbf{A} \mathbf{x} + B^T \mathbf{x} + c \tag{3.19}$$

For training data 1 and 2 from two different classes, which are distributed as $M[\boldsymbol{\mu}_i, \mathbf{S}_i], i \in 1 and\, 2$, the transformation parameters A and *B* can be obtained on the development data as:

$$\mathbf{A} = -\frac{1}{2}\left(\mathbf{S}_1^{-1} - \mathbf{S}_2^{-1}\right) \tag{3.20}$$

$$B = \mathbf{S}_1^{-1} \boldsymbol{\mu}_1 - \mathbf{S}_2^{-1} \boldsymbol{\mu}_2 \tag{3.21}$$

c is a constant that depends on the mean vectors and covariance matrices and is computed as follows:

$$c = \boldsymbol{\mu}_1^T \mathbf{S}_1^{-1} \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2^T \mathbf{S}_2^{-1} \boldsymbol{\mu}_2 + \ln \frac{|\mathbf{S}_1|}{|\mathbf{S}_2|} \tag{3.22}$$

3.4.4 Logistic Regression (LR)

Another simple classification method can be used in the case of a two-class problem (Clients / Impostors) is that based on the principles of logistic regression [65, 81-83]. As indicated in the previous chapter the Logistic Regression method classifies the data based on using two functions: logistic regression function (3.23) and logit transformation (3.24) as follows:

$$E(Y \mid \mathbf{x}) = \frac{e^{g(\mathbf{x})}}{1 + e^{g(\mathbf{x})}} \tag{3.23}$$

where, $E(Y \mid \mathbf{x})$ is the conditional probability for the binary output variable Y and where the M-dimensional input vector $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_M)$ exists and $g(\mathbf{x})$ is defined as:

$$g(\mathbf{x}) = w_0 + w_1 \cdot \mathbf{x}_1 + ... + w_M \cdot \mathbf{x}_M \tag{3.24}$$

where w_m is the weight for the m^{th} modality. Due to the fact that each w_m with $i \neq 0$ multiplies one of the M modalities, it is evaluated as the level of the importance of that modality in the fusion process. A high w_m shows an important modality whilst a low w_m shows a modality not contributing a great deal.

Parameters in the above equation $(w_0, w_1, ..., w_M)$ can be calculated with the maximum likelihood approach with an iterative optimisation scheme on some development data as [38]:

$$LH = \prod_{i=1}^{n} p(\mathbf{x}_i; w) \quad (3.25)$$

where *LH* denotes the maximum likelihood function, $p(\mathbf{x}_i; w)$ is a density function with one parameter w for each modality, and a corresponding set of n sample values $\mathbf{x}_i$. In this equation, the maximum likelihood approach associates with each training set a value of w which maximise *LH*.

When the alternate parameters w_m have been worked out on the development data, an unknown test pattern is classified by evaluating $E(Y \mid \mathbf{x})$. The outcome is thus compared with optimal threshold calculated on the development data.

3.4.5 Support Vector Machines (SVM)

SVM is another effective classification technique which can be used in the case of a two-class problem (Clients/ Impostors). It is a new classification technique in the field of Statistical Learning Theory (SLT) [84-89]. SVM is based on the principle of Structural Risk Minimisation (SRM) which aims to find the optimal separating hyper-plane that should classify not only the development data, but also unknown test data. Inversely classical learning approaches are designed to minimise the so-called empirical risk (i.e. error on the development set) based on the Empirical Risk Minimisation (ERM) principle.

3.4.5.1 Linear SVM for linearly separable data

In this case, a linear SVM is trained on linearly separable data. The main aim in this case is to find the optimal hyper-plane which exactly separates the two classes from each other. This optimal separating hyper-plane, as indicated earlier, should classify not only

the development data, but also any unknown data in each class. The said hyper-plane is mathematically presented as:

$$\mathrm{w}^T \cdot \mathbf{x} + b = 0 \tag{3.26}$$

where w is a weight (coefficient) vector, T is the transpose operation, $\mathbf{x}$ is a training vector consisting of the scores for different modalities and b is a bias term estimated on the development set. Using the equation (3.26) leads to a straight line decision boundary (this refers to hyper-plane) that classifies the scores correctly. In this case the error is zero. However, there is actually an infinite number of hyper-planes that could partition the data into two classes (-1 or +1), see Figure 3.2. According to SRM principle, the line that is located half way between the two classes is the intuitive choice for the optimal hyper-plane. This is shown in Figure 3.3. The dashed lines in Figure 3.2 represent some of the possible hyper-planes that can separate the two class data while the solid line in Figure 3.3 is the optimal separating hyper-plane for that data.

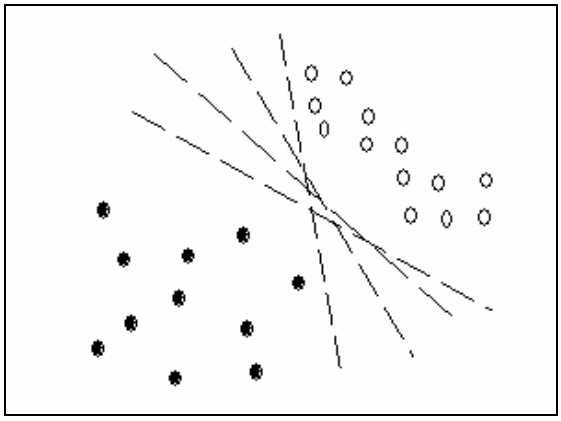


Figure 3.2: Possible separating hyper-planes

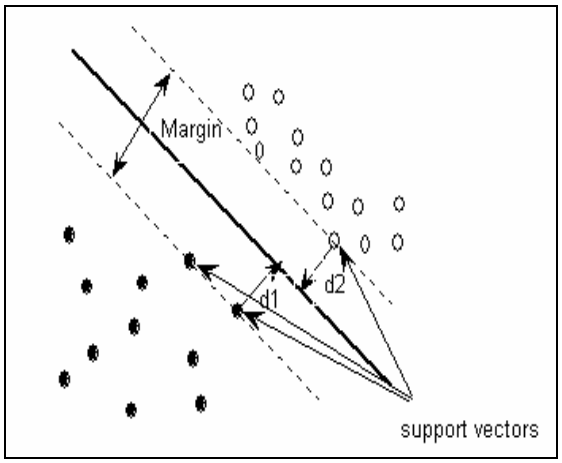


Figure 3.3: Optimal separating hyper-pane.

Assuming that d1 and d2 are the shortest distance from the separating hyper-plane to the closest points of each class; d1+d2 defines the margin of the optimal separating hyper-plane.

This margin is mathematically expressed as follows:

$$\frac{\left|\mathrm{w}^T\mathbf{x}_i + b\right|}{\mathrm{w}} \tag{3.27}$$

For the linearly separable case, the maximal margin can be found by minimising $\mathrm{w}^T\mathrm{w}$ with the constraints [84],

$$y_i\left(\mathbf{x}_i \cdot \mathrm{w}^T + b\right) \geq 1, \forall \mathrm{i} \tag{3.28}$$

where y_i is 1 if $\mathbf{x}_i$ belongs to one set (e.g. Clients) and -1 if $\mathbf{x}_i$ belongs to the other set (Impostors).

This conditional optimisation is accomplished by Lagrange's method as follows:

$$L(\mathrm{w}, b, a_i) = \frac{1}{2}\mathrm{w}^T\mathrm{w} - \sum_i a_i\left(y_i\left(\mathbf{x}_i \cdot \mathrm{w}^T + b\right) - 1\right) \tag{3.29}$$

where a_i are the solutions of the Lagrange's method *L*.

Differentiation with respect to w and *b* leads to [84, 85]:

$$\sum_i a_i y_i = 0 \tag{3.30}$$

and

$$\mathrm{w} = \sum_i a_i y_i \mathbf{x}_i \tag{3.31}$$

Substituting (3.30) and (3.31) into (3.29) leads to:

$$L(\mathrm{w}, b, a_i) = \sum_i a_i - \frac{1}{2}\sum_i \sum_j a_i a_j y_i y_j \mathbf{x}_i^T \mathbf{x}_\mathrm{j} \tag{3.32}$$

This optimisation is reduced to a quadratic programming problem as follows:

$$\max\text{imise} \sum_i a_i - \frac{1}{2}\sum_i \sum_j a_i a_j y_i y_j \mathbf{x}_i^T \mathbf{x}_\mathrm{j} \tag{3.33}$$

subject to

$$\sum_i a_i y_i = 0 \tag{3.34}$$

and

$$a_i \geq 0 \tag{3.35}$$

In the resulting solution, most a_i are equal to zero, which refer to the development data that are not on the margin. The training examples with non-zero a_i are called *support vectors*, which are the input vectors that lie on the edge of the margin (Figure 3.3). Introducing new data outside of the margin will not change the hyper-plane as long as the new data are not on the margin or misclassified. Therefore, the classifier must remember those vectors which define the hyper-plane.

3.4.5.2 Linear SVM for non-linearly separable data

In order to enable linear SVM to classify non-linearly separable data, the formulation in equation (3.28) must be adjusted. A cost for violating the separation constraints (3.28) must be introduced. To achieve this, slack variables are introduced into the inequalities relaxing them so that some points are allowed to lie within the margin or even be misclassified.

$$y_i\left(\mathbf{x}_i \cdot \mathrm{w}^T + b\right) \geq 1 - \xi_i, \forall \mathrm{i} \tag{3.36}$$

For a point to be misclassified, the corresponding ξ_i must exceed unity, so $\sum_i \xi_i$ is an upper bound for the number of classification errors. Hence a logical way to assign an extra cost for errors is to minimise $\mathrm{w}^T\mathrm{w} + C\sum_i \xi_i$

where C is a parameter to be chosen by the user, a larger C corresponding to assigning a higher penalty to errors. Note that the generalised optimal separating hyper-plane is obtained by minimising $\mathrm{w}^T\mathrm{w} + C\sum_i \xi_i$ with the constraints of equation (3.36). This is still a quadratic programming problem.

3.4.5.3 Non-linear SVM

In this case, the data is mapped from the input space into a higher dimensional space by a non-linear transformation. The transformation can be performed through the use of kernel functions. Such functions can have different forms [85-87]. The fundamental concept of kernel functions is to deform the vector space itself to a higher dimensional space. This is as shown in Figure 3.4.

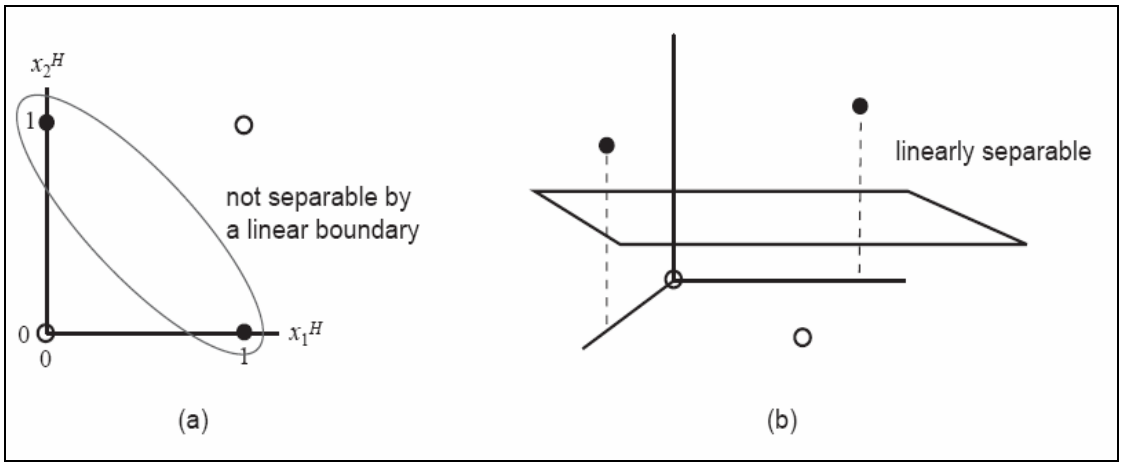


Figure 3.4: Transformation to higher dimension space(Asano(2004)[84]).

Figure 3.4(a) shows an example of the linearly non-separable data. In Figure 3.4, the two-dimensional space (Figure 3.4(a)) is transformed to the three-dimensional one (Figure 3.4(b)). This transformation is applied in order to linearly separate the "black"

vectors from the "white" vectors. Without such an approach, the said vectors can not be linearly separated.

In this case, the kernel function is defined as:

$$k(\mathbf{x},\mathbf{x}') = \Phi(\mathbf{x})^T \Phi(\mathbf{x}') \tag{3.37}$$

where Φ is a transformation to a higher dimensional space.

Equation (3.37) indicates that the kernel function can be represented as the distance between $\mathbf{x}$ and $\mathbf{x}'$ measured in the higher dimensional space transformed by Φ. In this case, the boundary (in the transformed space) is obtained as:

$$\mathrm{w}^T \cdot \Phi(\mathbf{x}) + b = 0 \tag{3.38}$$

and substituting (3.31) into (3.38) leads to:

$$\sum_i a_i y_i \Phi(\mathbf{x}_i^T)\Phi(\mathbf{x}) + b = \sum_i a_i y_i k(\mathbf{x}_i, \mathbf{x}) + b = 0 \tag{3.39}$$

Consequently, the optimisation function of equation (3.33) in the transformed space is obtained by substituting $\mathbf{x}_i^T \mathbf{x}_j$ with $k(\mathbf{x}_i, \mathbf{x}_j)$ in that equation. Thus, the whole calculation can be accomplished based on $k(\mathbf{x}_i, \mathbf{x}_j)$ only. This implies that there is no need to know what Φ or the transformed space actually is.

In this work, linear, radial basis function, and polynomial kernel functions with a degree of 2 (quadratic) are used. These are given by the following equations

$$Linear: k(\mathbf{x},\mathbf{x}') = \mathbf{x}^T \mathbf{x}' \tag{3.40}$$

$$Quadratic: k(\mathbf{x},\mathbf{x}') = (\mathbf{x}^T \mathbf{x}' + 1)^2 \tag{3.41}$$

$$RBF: k(\mathbf{x},\mathbf{x}') = e^{-\frac{\|\mathbf{x}-\mathbf{x}'\|^2}{d^2}} \tag{3.42}$$

where d is a constant that defines the kernel width.

3.5 Summary

Based on the previous studies regarding the range-normalisation techniques, it is concluded that bringing raw scores from different modalities to the same range is a necessary step in any fusion system. The chapter has given a brief description about two

of the most effective and widely used range-normalisation techniques. These are Min-Max and Z-score range-normalisation techniques. Then, the evaluation criteria for identity recognition systems in both cases verification and identification have been discussed. After that, the theories of the currently most effective fusion approaches have been presented. The techniques covered have ranged from weighting schemes that assign weights to the information streams according to their information content, to support vector machines which use the principle of obtaining the best possible boundary for classification according to the development data. The next chapter discusses the results of applying the fusion methods considered in two different cases. The first of these examines the usefulness of fusion in a unimodal biometrics scenario. The second case, on the other hand, involves fusing scores for two different types of biometrics of face and voice.

Chapter 4

Score-level Fusion in Biometric Verification

4.1 Introduction

In the previous chapter, the effective multimodal biometrics fusion techniques were discussed. This chapter details the investigations into the effectiveness of various fusion approaches in both unimodal and multimodal biometrics. In particular, two types of biometrics (i.e. face and voice) are considered in the investigations. The fusion process is performed at the score level. The scores for face and voice biometrics are based on the use of different features extracted from the XM2VTS database[90]. These scores are provided by IDIAP Research Institute [91]. The following section discusses the XM2VTS and gives a brief description of the classifiers used in computing the voice and face verification scores. Section 4.3 details the fusion experiments and provides an analysis of the results.

4.2 Speech and Face data

The XM2VTS database, used for the purpose of this study, is a bimodal database containing synchronised image and speech data from two hundred and ninety five subjects, recorded during four sessions at one month intervals [90]. In each session, two recordings were made, each consisting of a speech shot and a head shot. The speech shot consisted of frontal face and speech recordings of each subject during the recital of a sentence.

The subjects in the database are divided into three sets. These are a set of two hundred clients, a set of twenty five development impostors and a set of seventy test impostors. Two different methods of partitioning the database are in existence. They are called Lausanne Protocols I and II (denoted as LP1 and LP2). As described below, the difference between these two protocols is due to the number of bimodal samples per client used for training and development. In this study, only the scores obtained through LP1 are considered.

In total, there are eight bimodal biometric samples (utterance and face image) per client in the XM2VTS database. The samples are used in the following way. Three are used in

the training phase (i.e. for extracting reference features) in LP1 (and four in LP2). Three samples are for development in LP1 (and only two in LP2). Finally, both LP1 and LP2 involve two samples for testing. Table 4.1 summarises the structures of the two protocols (Norman and Sami (2005) [90]). It should be noted that the number of accesses given are per modality.

Data sets	**Lausanne Protocols**	
	LP1	**LP2**
Training samples (bimodal) per client	3	4
Development client accesses	600 (i.e. 3×200)	400 (i.e. 2×200)
Development impostor accesses	40,000 (i.e. $25 \times 8 \times 200$)	
Test client accesses	400 (i.e. 2×200)	
Test impostor accesses	112,000 (i.e. $70 \times 8 \times 200$)	

Table 4.1: Lausanne Protocols for the XM2VTS database.

4.2.1 Classifiers and features

The unimodal verification scores with this database are based on the use of GMM (Gaussian Mixture Models) for voice and GMM as well as MLPs (Multi-Layer Perceptrons) for face.

Three different types of features are used for face biometrics. These are normalised **F**ace image concatenated with its RBG **H**istogram (FH) and two types of Discrete Cosine Transform (DCT): DCTs and DCTb. **s** in DCTs indicates the use of small images with a size of 40×32 (rows×columns) pixels, whilst **b** in DCTb indicates the use of bigger images with size of 80×64 pixels.

For voice verification, Linear Filter-bank Cepstral Coefficients (LFCC), Phase Auto-Correlation (PAC), and Spectral Subband Centriod (SSC) are used as the three different voice feature types. In total, 5 sets of scores are obtained for face verification and three sets for voice verification [90]. The feature types and classifiers used to extract these are summarised in Table 4.2.

Having different features/classifiers for face and voice modalities leads to critical questions. Such questions are “Are these features complementary to each other at the unimodal level?” and “Would combining the scores for features obtained from the same

sensing modality outperform the best individual feature involved?". For the purpose of addressing such questions, investigations into the effectiveness of various fusion approaches (e.g. Weighted Average, FLD, QDA, LR and SVM) in unimodal biometrics (face or voice) are presented in this chapter.

In the present study, the scores in the development access sets are used to compute the appropriate parameters for various fusion methods. For this purpose, the client and impostor scores from the chosen features are pooled and then range-normalised according to the chosen range-normalisation scheme. The parameters obtained in the development stage are then used in the test phase to fuse the normalised test scores according to the scheme deployed. The verification performance is then obtained on the fused scores in terms of equal error rates (EER).

4.3 Experimental investigations and discussions

This section discusses the results of fusing face and voice scores obtained using the feature and classifier types described above. Nine fusion schemes are used in this study. These are BFS, MW-(FAR/FRR), MW-EER, FLD, QDA, LR, Linear SVM, Poly SVM, and RBF SVM. Details of these fusion schemes can be found in Section 3.4. Each of these is used once with the MM range-normalisation method, and again with ZS range-normalisation. Table 4.2 presents the baseline EERs for the eight combinations of features and classifiers in unimodal verification. The Table shows that FH gives the best EER compared to the other face features, whilst LFCC leads to the lowest EER compared to the other speech features.

Feature		Classifier	EER%
face	FH	MLP	*1.58*
	DCTs	GMM	4.19
	DCTb	GMM	1.69
	DCTs	MLP	3.89
	DCTb	MLP	5.63
speech	LFCC	GMM	*1.08*
	PAC	GMM	6.50
	SSC	GMM	4.58

Table 4.2: Baseline EERs computed using the unimodal verification scores in various cases. The best performance in each of the face and voice modalities is shown in italics.

4.3.1 Score fusion in unimodal biometrics based on multiple matching algorithms

This section presents investigations into the performance of the fusion techniques for combining the score information obtained from the same sensing modality. Since the

scores from two different features are fused at each time, there are thirteen different results for each fusion method with each of the two range-normalisation methods. Ten of them are formed by using the scores for face features and the other three are based on the use of scores for voice features.

4.3.1.1 Score-level fusion results based on MM range-normalisation

The first set of experiments with fusion methods is based on the use of the MM range-normalisation method. Tables 4.3 and 4.4 present the results in terms of EERs for all the possible feature/classifier combinations for face and voice features respectively.

NO.	Fusion candidates	BFS	MW-(FAR/FRR)	MW-EER	FLD	QDA	LR	Linear SVM	Poly SVM	RBF SVM
1	[FH , DCTs-GMM]	1.44	1.29	1.75	1.87	1.67	3.74	3.91	3.92	3.83
2	[FH , DCTb-GMM]	1.25	1.43	1.43	1.94	1.52	1.07	2.12	2.36	2.41
3	[FH , DCTs-MLP]	1.63	1.50	1.72	1.63	1.63	1.58	1.50	1.65	1.72
4	[FH , DCTb-MLP]	1.81	1.77	2.00	1.94	1.83	1.80	1.64	2.15	2.19
5	[DCTs-GMM , DCTb-GMM]	1.76	1.75	1.83	2.58	1.78	2.81	4.96	4.80	4.87
6	[DCTs-GMM , DCTs-MLP]	2.67	2.50	2.59	3.62	3.27	2.98	4.65	4.61	4.71
7	[DCTs-GMM , DCTb-MLP]	4.91	4.25	4.02	7.57	4.35	6.01	4.70	4.68	4.43
8	[DCTb-GMM , DCTs-MLP]	1.21	2.14	1.37	2.38	2.16	2.14	1.84	1.81	1.88
9	[DCTb-GMM , DCTb-MLP]	1.33	4.28	2.41	3.24	2.95	2.25	2.37	2.55	2.53
10	[DCTs-MLP , DCTb-MLP]	2.79	3.25	2.98	2.87	3.23	3.11	3.02	2.78	2.69
Average EER		2.08	2.42	2.21	2.96	2.44	2.75	3.07	3.13	3.13

Table 4.3: Unimodal face verification results in terms of EER (%), based on score-level fusion with MM range-normalisation.

NO.	Fusion candidates	BFS	MW-(FAR/FRR)	MW-EER	FLD	QDA	LR	Linear SVM	Poly SVM	RBF SVM
1	[LFCC , PAC]	1.06	1.78	1.00	3.81	1.19	2.89	2.27	2.30	2.49
2	[LFCC , SSC]	1.00	1.23	0.96	2.96	1.22	2.61	2.74	2.76	2.91
3	[PAC , SSC]	3.78	4.53	4.25	4.32	4.72	4.32	4.48	4.63	4.51
Average EER		1.95	2.51	2.07	3.70	2.38	3.27	3.16	3.23	3.30

Table 4.4: Unimodal voice verification results in terms of EER (%), based on score-level fusion with MM range-normalisation.

The above results clearly show that achieving improvements through the unimodal score-level fusion not only depends on the fusion method adopted but also on the choice of face/voice score combination. Each combination, as stated earlier, differs from the other in terms of feature and/ or classifier for the chosen modality.

Comparing the results in Tables 4.3 and 4.4 with the baseline EERs for face and voice scores in Table 4.2, it is observed that the unimodal score-level fusion based on MM range-normalisation (in most cases) leads to the degradation of the verification accuracy. In some cases though, this type of fusion results in EERs which are just slightly better than the EER offered by the best single feature involved.

Tables 4.5 and 4.6, on the other hand, present the relative effectiveness of various methods (with MM range-normalisation), for fusing the scores obtained with face and voice features at the unimodal level, through the use of RI. As stated earlier (Section 3.3.1) that this type of measure can have a maximum value of one, indicating the fusion scheme adopted has resulted in a zero EER. On the other hand, a zero RI reflects the fact that there has been no improvement through the fusion used over the best individual feature involved. Finally, when the fusion adopted leads to the degradation of the verification accuracy, this is reflected as a negative RI. Therefore, for a fusion scheme to be beneficial, RI should be positive, and the closer it is to 1, the better is the effectiveness of fusion. Unfortunately, an examination of the results (e.g. RI) in tables 4.5 and 4.6 shows that, through the adopted fusion method, most of the RI values are negative. Such behaviour indicates that (in most cases) fusing the scores obtained from the same sensing modality (with MM range-normalisation) is not capable of enhancing the verification accuracy.

It can be seen from the average RI that a positive RI is obtained only by BFS (i.e. RI=0.10) in the case of unimodal fusion for face features (Table 4.5), and BFS (i.e. RI=0.09) and MW-EER (i.e. RI=0.08) in the case of unimodal fusion for voice features (Table 4.6). However, these positive RI values are very close to zero which indicates that there has been inconsiderable improvement.

NO.	Fusion candidates	BFS	MW-(FAR/FRR)	MW-EER	FLD	QDA	LR	Linear SVM	Poly SVM	RBF SVM
1	[FH , DCTs-GMM]	0.09	0.18	-0.11	-0.18	-0.06	-1.37	-1.47	-1.48	-1.42
2	[FH , DCTb-GMM]	0.21	0.09	0.09	-0.23	0.04	0.32	-0.34	-0.49	-0.53
3	[FH , DCTs-MLP]	-0.03	0.05	-0.09	-0.03	-0.03	0	0.05	-0.07	-0.09
4	[FH , DCTb-MLP]	-0.15	-0.12	-0.27	-0.23	-0.16	-0.14	0.04	-0.36	-0.39
5	[DCTs-GMM , DCTb-GMM]	-0.04	-0.04	-0.08	-0.53	-0.05	-0.66	-1.93	-1.84	-1.88
6	[DCTs-GMM , DCTs-MLP]	0.31	0.36	0.33	0.07	0.16	0.23	-0.20	-0.18	-0.21
7	[DCTs-GMM , DCTb-MLP]	-0.17	-0.01	0.04	-0.81	-0.04	0.43	-0.12	-0.12	-0.06
8	[DCTb-GMM , DCTs-MLP]	0.28	-0.27	0.19	-0.40	-0.28	-0.27	-0.09	-0.07	-0.11
9	[DCTb-GMM , DCTb-MLP]	0.21	-1.53	-0.43	-0.92	-0.75	-0.33	-0.40	-0.51	-0.50
10	[DCTs-MLP , DCTb-MLP]	0.28	0.16	0.23	0.26	0.17	0.20	0.22	0.29	0.31
Average RI		0.10	-0.11	-0.01	-0.3	-0.10	-0.16	-0.42	-0.48	-0.49

Table 4.5: Relative improvements (RI) for the unimodal face features using various fusion schemes based on MM range-normalisation.

NO.	Fusion candidates	BFS	MW-(FAR/FRR)	MW-EER	FLD	QDA	LR	Linear SVM	Poly SVM	RBF SVM
1	[LFCC , PAC]	0.02	-0.65	0.07	-2.53	-0.10	-1.68	-1.10	-1.13	-1.31
2	[LFCC , SSC]	0.07	-0.14	0.11	-1.74	-0.13	-1.42	-1.54	-1.55	-1.69
3	[PAC , SSC]	0.17	0.01	0.07	0.06	-0.03	0.06	0.02	-0.01	0.02
Average EER		0.09	-0.26	0.08	-1.40	-0.09	-1.01	-0.87	-0.90	-0.99

Table 4.6: Relative improvements (RI) for the unimodal voice features using various fusion schemes based on MM range-normalisation.

4.3.1.2 Score-level fusion results based on ZS range-normalisation

In this set of experiments, the considered fusion methods are applied based on the use of the ZS range-normalisation technique. Tables 4.7 and 4.8 present the results in terms of EERs again for all the possible feature/classifier combinations for face and voice features respectively. Their corresponding relative improvements are presented in tables 4.9 and 4.10 respectively.

NO.	Fusion candidates	BFS	MW-(FAR/FRR)	MW-EER	FLD	QDA	LR	Linear SVM	Poly SVM	RBF SVM
1	[FH , DCTs-GMM]	1.26	1.42	1.71	2.25	1.40	1.44	1.57	1.61	1.72
2	[FH , DCTb-GMM]	1.25	1.33	1.39	2.24	1.39	1.15	1.82	1.77	1.85
3	[FH , DCTs-MLP]	1.54	1.22	1.28	2.18	1.08	1.61	1.85	1.91	1.98
4	[FH , DCTb-MLP]	1.79	1.97	1.67	2.34	1.76	1.90	2.03	1.99	2.04
5	[DCTs-GMM , DCTb-GMM]	1.71	1.81	1.68	3.65	1.44	1.77	2.50	2.62	2.53
6	[DCTs-GMM , DCTs-MLP]	2.63	2.67	2.57	4.49	2.52	2.32	2.69	2.73	2.56
7	[DCTs-GMM , DCTb-MLP]	2.61	3.54	3.50	8.22	3.52	2.96	3.27	3.31	3.17
8	[DCTb-GMM , DCTs-MLP]	1.12	1.55	1.24	4.24	1.83	1.98	1.92	1.90	1.88
9	[DCTb-GMM , DCTb-MLP]	1.73	2.37	1.62	7.57	2.45	2.09	2.32	2.26	2.29
10	[DCTs-MLP , DCTb-MLP]	2.75	3.00	2.85	4.47	2.93	2.88	3.08	3.11	3.14
Average EER		1.84	2.09	1.95	4.17	2.03	2.01	2.31	2.32	2.32

Table 4.7: Unimodal face verification results in terms of EER (%), based on score-level fusion with ZS range-normalisation.

NO.	Fusion candidates	BFS	MW-(FAR/FRR)	MW-EER	FLD	QDA	LR	Linear SVM	Poly SVM	RBF SVM
1	[LFCC , PAC]	1.27	1.76	1.83	1.65	1.23	1.28	1.20	1.25	1.16
2	[LFCC , SSC]	1.00	1.17	0.97	1.20	1.01	0.98	1.04	1.02	0.99
3	[PAC , SSC]	3.17	4.26	4.16	3.84	3.56	3.93	3.78	3.83	3.79
Average EER		1.81	2.40	2.32	2.23	1.93	2.06	2.01	2.03	1.98

Table 4.8: Unimodal voice verification results in terms of EER (%), based on score-level fusion with ZS range-normalisation.

NO.	Fusion candidates	BFS	MW-(FAR/FRR)	MW-EER	FLD	QDA	LR	Linear SVM	Poly SVM	RBF SVM
1	[FH , DCTs-GMM]	0.20	0.10	-0.08	-0.42	0.11	0.09	0.01	-0.02	-0.09
2	[FH , DCTb-GMM]	0.21	0.16	0.12	-0.42	0.12	0.27	-0.15	-0.12	-0.17
3	[FH , DCTs-MLP]	0.03	0.23	0.19	-0.38	0.32	-0.02	-0.17	-0.21	-0.25
4	[FH , DCTb-MLP]	-0.13	-0.25	-0.06	-0.48	-0.11	-0.20	-0.28	-0.26	-0.29
5	[DCTs-GMM , DCTb-GMM]	-0.01	-0.07	0.01	-1.16	0.15	-0.05	0.48	-0.55	-0.50
6	[DCTs-GMM , DCTs-MLP]	0.32	0.31	0.34	-0.15	0.35	0.40	0.31	0.30	0.34
7	[DCTs-GMM , DCTb-MLP]	0.37	0.16	0.16	-0.96	0.16	0.29	0.22	0.21	0.24
8	[DCTb-GMM , DCTs-MLP]	0.34	0.08	0.27	-1.51	-0.08	-0.17	-0.14	-0.12	-0.11
9	[DCTb-GMM , DCTb-MLP]	-0.02	-0.40	0.04	-3.48	-0.45	-0.24	-0.37	-0.34	-0.36
10	[DCTs-MLP , DCTb-MLP]	0.29	0.23	0.27	-0.15	0.25	0.26	0.21	0.20	0.19
Average RI		0.16	0.06	0.13	-0.91	0.08	0.06	0.01	-0.09	-0.10

Table 4.9: Relative improvements (RI) for the unimodal face features using various fusion schemes based on ZS range-normalisation.

NO.	Fusion candidates	BFS	MW-(FAR/FRR)	MW-EER	FLD	QDA	LR	Linear SVM	Poly SVM	RBF SVM
1	[LFCC , PAC]	-0.18	-0.63	-0.69	-0.53	-0.14	-0.19	-0.11	-0.16	-0.07
2	[LFCC , SSC]	0.07	-0.08	0.10	-0.11	0.06	0.09	0.04	0.06	0.08
3	[PAC , SSC]	0.31	0.07	0.09	0.16	0.22	0.14	0.17	0.16	0.17
Average EER		0.07	-0.21	0.08	-0.17	0.05	0.01	0.03	0.02	0.06

Table 4.10: Relative improvements (RI) for the unimodal voice features using various fusion schemes based on ZS range-normalisation.

By comparing the results (e.g. average EER and average RI) for the MM range-normalisation method (Tables 4.3-4.6) with the corresponding results for ZS range-normalisation (Tables 4.7-4.10) it is evident that better performance can be obtained with the latter. However, it is observed (Tables 4.9 and 4.10) that, the average RIs obtained (using ZS) are still either negative or very small positive values with all fusion methods. For example, the most significant RI for the unimodal face verification is obtained with BFS (i.e. RI=0.16). This level of performance is closely followed by that of the MW-EER fusion method. For the unimodal voice verification, on the other hand, the best result of RI=0.08 is obtained with MW-EER. It can be seen from these results that fusing scores coming from the same modality cannot lead to considerably lower EERs, even through the use of ZS range-normalisation, compared with the best results without fusion. However, it is believed that with the considered methods of score-level fusion, more benefit can be achieved from the complementary information of the face and voice features at the multimodal level.

4.3.2 Multimodal fusion

This section discusses the results of fusing face and voice scores obtained using the feature and classifier types described in Section 4.2.1. Since the scores from two different modalities are fused each time, there are fifteen different results for each fusion method with each of the two range-normalisation methods.

Tables 4.11 and 4.13 present the results for all the fifteen feature/classifier combinations based on the MM and ZS normalisation techniques respectively. The relative improvements for various fusion methods (with MM and ZS normalisation techniques) are presented in tables 4.12 and 4.14 respectively.

NO.	Fusion candidates			BFS	MW-(FAR/FRR)	MW-EER	FLD	QDA	LR	Linear SVM	Poly SVM	RBF SVM
	Face		Voice									
	Feature	Classifier	Feature									
1	FH	MLP	LFCC	0.68	0.88	0.64	1.78	1.52	1.61	1.04	1.07	1.15
2			PAC	1.00	1.27	1.69	1.96	1.58	0.78	0.95	0.78	0.92
3			SSC	0.93	1.08	1.46	1.94	1.80	1.31	0.96	1.25	1.10
4	DCTs	GMM	LFCC	0.74	0.62	0.75	4.24	2.49	2.11	2.28	2.22	2.27
5			PAC	1.29	1.33	1.42	5.51	3.09	2.94	3.65	3.53	3.68
6			SSC	1.07	1.24	1.12	5.98	6.22	4.49	5.10	4.89	5.07
7	DCTb	GMM	LFCC	0.52	0.51	0.53	1.39	1.89	0.66	2.94	2.88	2.94
8			PAC	1.30	1.25	1.06	1.22	1.25	1.13	1.35	1.34	1.36
9			SSC	0.75	0.76	0.78	1.51	1.44	1.07	2.03	1.89	2.03
10	DCTs	MLP	LFCC	0.69	0.84	0.52	1.92	1.26	0.79	0.58	0.71	0.66
11			PAC	0.89	1.51	2.06	2.96	1.39	0.80	0.75	0.73	0.75
12			SSC	1.20	1.43	1.91	2.41	1.67	1.43	1.08	1.17	1.22
13	DCTb	MLP	LFCC	0.61	2.00	0.66	2.25	1.74	1.75	1.52	1.47	1.55
14			PAC	2.63	3.77	4.11	5.18	3.12	2.43	2.36	2.29	2.34
15			SSC	1.92	3.92	3.65	5.27	2.62	1.85	2.27	2.49	2.40
Average EER				1.08	1.49	1.49	3.04	2.21	1.68	1.92	1.91	1.96

Table 4.11: Bimodal verification results in terms of EER (%), based on score-level fusion with MM normalisation.

NO.	Fusion candidates			BFS	MW-(FAR/FRR)	MW-EER	FLD	QDA	LR	Linear SVM	Poly SVM	RBF SVM
	Face		Voice									
	Feature	Classifier	Feature									
1	FH	MLP	LFCC	0.37	0.19	0.41	-0.65	-0.41	-0.49	0.04	0.01	-0.06
2			PAC	0.38	0.20	-0.07	-0.24	0	0.51	0.40	0.51	0.42
3			SSC	0.41	0.32	0.08	-0.23	-0.14	0.17	0.39	0.20	0.30
4	DCTs	GMM	LFCC	0.31	0.43	0.31	-2.93	-1.31	-0.95	-1.11	-1.06	-1.10
5			PAC	0.69	0.68	0.66	-0.32	0.26	0.30	0.13	0.16	0.12
6			SSC	0.74	0.70	0.73	-0.43	-0.48	-0.07	-0.22	-0.17	-0.21
7	DCTb	GMM	LFCC	0.52	0.53	0.51	-0.29	-0.75	0.39	-1.72	-1.67	-1.72
8			PAC	0.23	0.26	0.37	0.28	0.26	0.33	0.20	0.21	0.20
9			SSC	0.56	0.55	0.54	0.11	0.15	0.37	-0.20	-0.12	-0.20
10	DCTs	MLP	LFCC	0.36	0.22	0.52	-0.78	-0.17	0.27	0.46	0.34	0.39
11			PAC	0.77	0.61	0.47	0.24	0.64	0.79	0.81	0.81	0.81
12			SSC	0.69	0.63	0.51	0.38	0.57	0.63	0.72	0.70	0.69
13	DCTb	MLP	LFCC	0.44	-0.85	0.39	-1.08	-0.61	-0.62	-0.41	-0.36	-0.44
14			PAC	0.53	0.33	0.27	0.08	0.45	0.57	0.58	0.59	0.58
15			SSC	0.58	0.14	0.20	-0.15	0.43	0.60	0.50	0.46	0.48
Average RI				0.51	0.33	0.39	-0.40	-0.07	0.19	0.04	0.04	0.02

Table 4.12: Relative improvements (RI) for various fusion schemes based on MM range-normalisation.

NO.	Fusion candidates			BFS	MW-(FAR/FRR)	MW-EER	FLD	QDA	LR	Linear SVM	Poly SVM	RBF SVM
	Face		Voice									
	Feature	Classifier	Feature									
1	FH	MLP	LFCC	0.56	0.85	0.72	2.23	1.31	0.47	0.47	0.47	0.43
2			PAC	0.76	1.25	1.49	2.25	1.84	0.98	1.08	0.96	0.93
3			SSC	0.81	1.21	1.36	2.24	1.89	0.94	0.73	0.93	1.01
4	DCTs	GMM	LFCC	0.74	0.52	0.59	1.15	0.56	0.53	0.54	0.58	0.66
5			PAC	1.37	1.39	1.23	1.62	1.61	1.59	1.70	1.72	2.03
6			SSC	1.02	1.19	1.18	1.17	1.38	1.13	1.13	1.20	1.33
7	DCTb	GMM	LFCC	0.48	0.44	0.57	2.13	0.66	0.36	0.69	0.67	0.74
8			PAC	1.26	0.79	1.23	3.39	1.09	1.08	0.98	1.13	1.37
9			SSC	0.71	0.71	1.00	2.76	0.87	0.72	0.95	1.02	1.22
10	DCTs	MLP	LFCC	0.58	0.97	0.47	4.39	1.53	0.48	0.47	0.47	0.47
11			PAC	0.73	1.67	2.26	4.35	2.00	0.97	0.96	0.87	0.83
12			SSC	1.05	1.17	1.78	4.22	2.12	1.22	1.12	1.06	1.12
13	DCTb	MLP	LFCC	0.53	1.07	0.58	6.32	1.79	0.60	0.67	0.63	0.69
14			PAC	2.74	3.44	3.89	7.60	3.28	2.75	2.40	2.43	2.53
15			SSC	1.78	2.19	2.56	7.32	2.88	1.73	1.97	2.27	2.66
Average EER				1.01	1.26	1.39	3.54	1.65	1.04	1.08	1.09	1.20

Table 4.13: Bimodal verification results in terms of EER (%), based on score-level fusion with ZS range-normalisation.

NO.	Fusion candidates			BFS	MW-(FAR/FRR)	MW-EER	FLD	QDA	LR	Linear SVM	Poly SVM	RBF SVM
	Face		Voice									
	Feature	Classifier	Feature									
1	FH	MLP	LFCC	0.48	0.21	0.33	-1.06	-0.21	0.56	0.56	0.56	0.60
2			PAC	0.52	0.21	0.06	-0.42	-0.16	0.38	0.32	0.39	0.41
3			SSC	0.49	0.23	0.14	-0.42	-0.20	0.41	0.54	0.41	0.36
4	DCTs	GMM	LFCC	0.31	0.52	0.45	-0.06	0.48	0.51	0.50	0.46	0.39
5			PAC	0.67	0.67	0.71	0.61	0.62	0.62	0.59	0.59	0.52
6			SSC	0.76	0.72	0.72	0.72	0.67	0.73	0.73	0.71	0.68
7	DCTb	GMM	LFCC	0.56	0.59	0.47	-0.97	0.39	0.67	0.36	0.38	0.31
8			PAC	0.25	0.53	0.27	-1.01	0.36	0.36	0.42	0.33	0.19
9			SSC	0.58	0.58	0.41	-0.63	0.49	0.57	0.44	0.40	0.29
10	DCTs	MLP	LFCC	0.46	0.10	0.56	-3.06	-0.42	0.56	0.56	0.56	0.56
11			PAC	0.81	0.57	0.42	-0.12	0.49	0.75	0.75	0.78	0.79
12			SSC	0.73	0.70	0.54	-0.08	0.46	0.69	0.71	0.73	0.71
13	DCTb	MLP	LFCC	0.51	0.01	0.46	-4.85	-0.66	0.44	0.38	0.42	0.36
14			PAC	0.51	0.39	0.31	-0.35	0.42	0.51	0.57	0.57	0.55
15			SSC	0.61	0.52	0.44	-0.60	0.37	0.62	0.57	0.50	0.42
Average RI				0.55	0.44	0.42	-0.82	0.21	0.56	0.53	0.52	0.48

Table 4.14: Relative improvements (RI) for various fusion schemes based on ZS range-normalisation.

The above results are in agreement with the earlier observation that achieving improvements through the score-level fusion not only depends on the fusion method adopted but also on the choice of face-voice score combination. Each combination, as

stated earlier, differs from the other in terms of feature and/ or classifier for one modality or both modalities.

By comparing the results (e.g. average EER and average RI) for the MM range-normalisation method (Tables 4.11 and 4.12) with the corresponding results for ZS range-normalisation (Tables 4.13 and 4.14) it is evident that considerably better performance can be obtained with the latter. Since the focus of this study is the performance of the fusion methods, the discussions presented below, concentrate on the experimental results obtained with the ZS method only.

It can be seen from the average EERs in Table 4.13 that the worst fusion technique in this experiment setup (using ZS) is FLD with an average EER of 3.54%. QDA shows reasonable performance as compared to FLD with an average EER of 1.65%. However, it is observed (Table 4.14) that, with this fusion approach, negative RI's are obtained in a number of cases. The remaining seven fusion methods (i.e. BFS, MW-(FAR/FRR), MW-EER, LR, Linear SVM, Poly SVM, and RBF SVM) appear as the best performers with positive RI's in all cases. In other words, with these fusion methods, the bimodal verification results consistently outperform those for the best single modalities. Based on the average RI's given in Table 4.14 it can be said that, although LR appear as the best method, comparable performance is offered by the other six fusion approaches.

Observing the RI values in Table 4.14 for the top seven fusion methods, it is noted that the best results are obtained when DCTs is used as the face feature. However, Table 4.14 also confirms the earlier suggestion that, in general, the effectiveness of each fusion method varies with the choice of feature and classifier used for each modality.

Another important outcome of the experimental investigations can be observed by considering the results in Section 4.3.2 together with those in Section 4.3.1. Based on these results, it is clearly seen that fusing the scores obtained from the same sensing modality may not necessarily exceed the verification accuracy offered by the best single feature involved. The results in these two sections indicate that higher accuracy is the basic advantage of multimodal biometrics over unimodal biometrics. The reason for such findings is that separate information from different modalities is used to provide complementary evidence about the identity of the users. A direct comparison of the

average RIs, obtained using the fusion methods (with ZS), for the unimodal and multimodal verification is given in Figure 4.1.

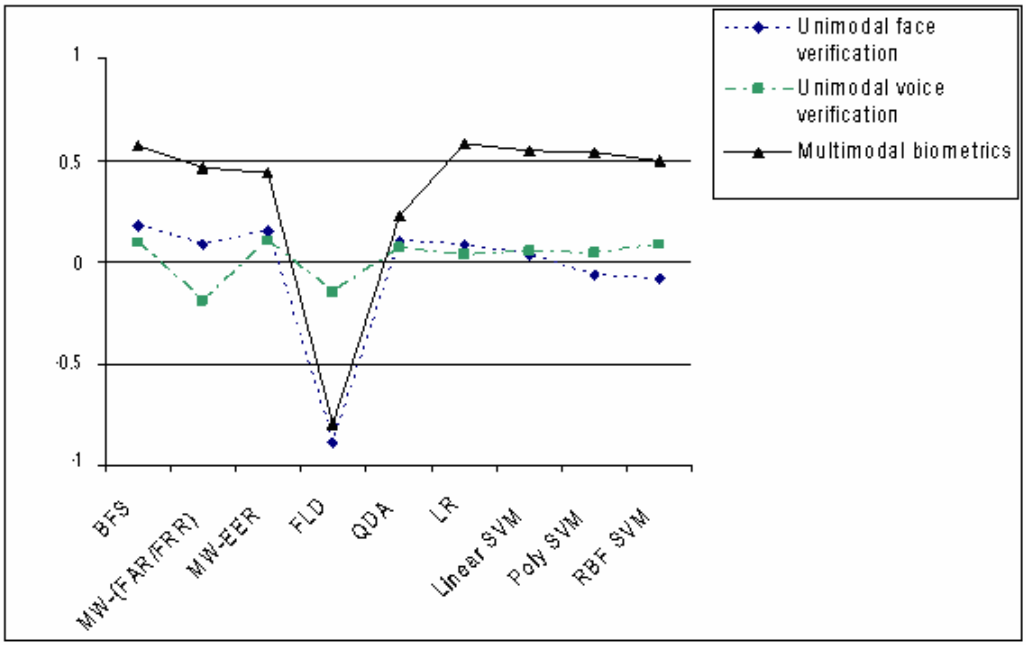


Figure 4.1: Comparison of RIs for the unimodal and multimodal verification experiments based on ZS range-normalisation.

Figure 4.1 clearly shows that, in most cases, combining the score information in the unimodal biometrics (face/ voice) provides comparable performance. However, it is apparent from Figure 4.1 that (in most cases) multimodal biometrics is exhibiting more effectiveness than the unimodal approach. This confirms the earlier suggestion that more benefit can be achieved from the complementary information of the face and voice features for biometric recognition.

Figure 4.2 presents a direct comparison of the effectiveness of multimodal biometrics with those of the individual modalities involved (face and voice) as DET (Detection Error Trade-off) plots. The fusion approach in this case is that of LR, the face and voice features are based on DCTb and LFCC respectively, and the classifier type is GMM. These plots further confirm the advantage in terms of improving the accuracy offered by biometric fusion. It is noted that in this case, the best EER offered by a single modality (voice) is about 1% whereas the EER obtained through the fusion process is around 0.4%.

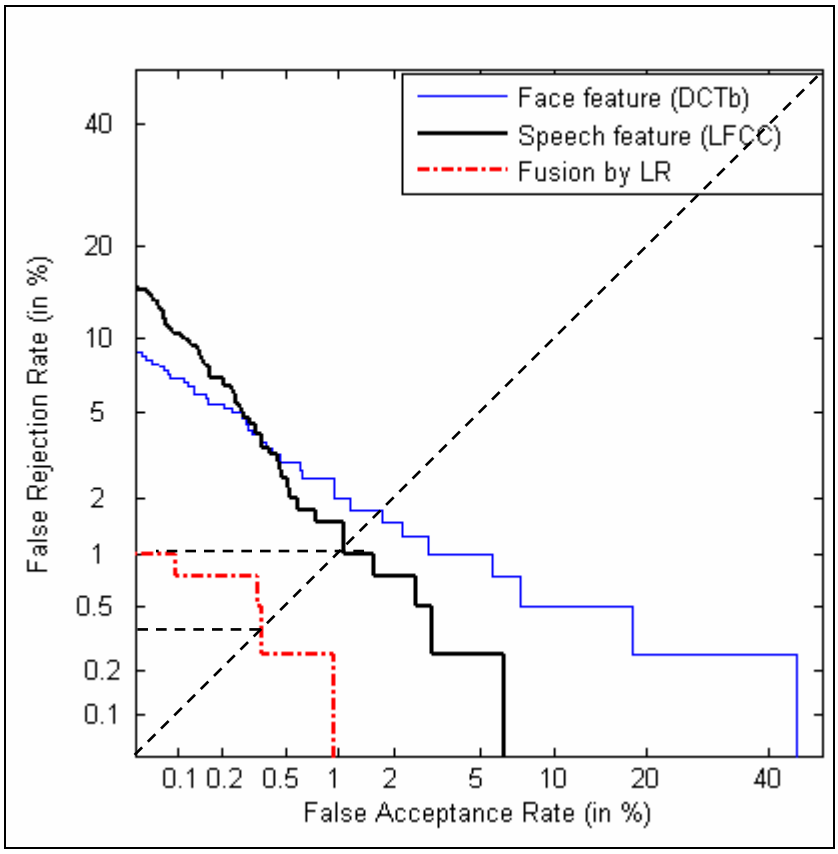


Figure 4.2: Relative performance of fused biometrics (based on LR) and individual modalities (face and voice).

4.4 Summary

The chapter has presented investigations into the performance of the fusion techniques for combining the score information in both unimodal and multimodal biometrics (speaker and face verification). The individual modality scores are obtained using the XM2VTS database. The scores are based on eight baseline systems. Five of the eight baseline systems involve face features and the other three are for speech features. In each experiment, the scores to be fused are subjected to the range-equalisation process prior to fusion. This is based on MM or ZS range-normalisation techniques.

Based on the experimental results presented in this chapter, it has been concluded that higher accuracy is the basic advantage of multimodal biometrics over unimodal biometrics. The reason of such findings is that separate information from different modalities is used to provide complementary evidence about the identity of the users. It is also concluded that, ZS range-normalisation exhibits more effectiveness than MM range-normalisation. With ZS range-normalisation, the fusion process (in most cases) improves the performance beyond that obtainable with the better of the two individual modalities involved. In particular, the seven top fusion methods considered in the multimodal scenario are found to provide consistent improvement regardless of the choice of face-voice score combination. Based on the results it is noted that the usefulness of each fusion method varies with the choice of feature and classifier used for each modality. Next chapter discusses the unpredicted variations problem in the evidence captured in the scores. It proposes a technique to reduce the effects of such variations in multimodal fusion based on estimating the quality aspect of the test scores.

Chapter 5

Multimodal Authentication using Qualitative Support Vector Machines

5.1 Introduction

Fusion techniques can be subdivided into adaptive and non-adaptive ones. Non-adaptive fusion techniques are those where all the fusion parameters are found using the development set, as seen in the previous chapter. With adaptive fusion techniques, all the parameters, or some of them, are found based on the test set. From the definition of the non-adaptive fusion technique, it can be seen that the drawback of this technique is the possible mismatch between the relative variation of the biometric modalities involved in the development and test data respectively. For example, if one modality (e.g. voice) leads to good performance in the development stage, compared to the other modality (e.g. face), but does not retain the same relative performance at the test stage, this can adversely affect the outcome of multimodal biometrics. To tackle this problem, it would be logical to consider the relative levels of contamination in different biometric data not only in the development phase, but also at the test stage.

This chapter presents an adaptive approach to reduce the effects of such relative degradation in multimodal fusion. The proposed approach is based on adjusting the weights for each of the two modalities according to their relative quality. This is performed by estimating the relative quality aspects of the test scores and then passing them on into the Support Vector Machine either as features or weights. The use of SVM is based on earlier investigations (Chapter 4) and other earlier studies which report it as one of the most effective methods for multimodal biometric fusion [1, 87]. Since the fusion process is based on the learning classifier of the Support Vector Machine, the technique is termed Support Vector Machine with Relative Quality Measurement (SVM-RQM). The experimental investigation is conducted using the scores for face and speech modalities. These scores are based on the use of different features extracted from the XM2VTS database. The rest of the Chapter is organised as follows. Section 5.2 provides details of the proposed schemes. Section 5.3 describes the experimental investigation and discusses the results, and Section 5.4 gives the overall conclusions.

5.2 Proposed approach

Improving fusion with a quality learning process has already been examined by several studies [53, 68, 69]. The experimental results for these studies have all verified that quality-based fusion schemes outperform those raw fusion strategies with no consideration of quality of input biometric data. However, these techniques still have some limitations which might adversely influence the overall effectiveness of biometric recognition. For example, quality estimation in [53, 68, 69] is derived from labeled (development) data. In other words, the parameters of the fusion technique are adopted based on the quality of the development data with no consideration of the possible mismatch between the relative degradation of the biometric modalities involved in the development and test data respectively. Therefore, this technique might be less effective in real-applications.

As indicated earlier, such limitations might adversely affect the overall accuracy of a multimodal biometrics system. Therefore, it is suggested that the relative quality aspect of the development as well as test data be incorporated in the fusion process. Such an approach should ideally tackle the effect of the possible mismatch between the relative quality in the biometric data. This, as indicated earlier, is because the approach is believed to provide a useful means of adjusting appropriately the weights for each of the two modalities according to their relative quality.

In this technique the quality aspect of the test samples is quantified and then passed on into a SVM. This process involves estimating the quality of the development data by measuring some parameters for the development score data and then incorporating these parameters in the quality estimation of the test scores. This quantification is similar to that described in [9] and is described as follows:

in the case of a two-class problem (Clients / Impostors), let $M(f/s)$ be the development scores for face or speech, (where (f/s) is used to denote that a measure is applied to either face or speech modality) and let the client and impostor scores from each modality be given as

$$C_{M(f/s)} \cong \left\{\mu^{C}_{M(f/s)}, \sigma^{C}_{M(f/s)}\right\} \tag{5.1}$$

$$I_{M(f/s)} \cong \left\{\mu^{I}_{M(f/s)}, \sigma^{I}_{M(f/s)}\right\} \tag{5.2}$$

where $\mu^{C}_{M(f/s)}$ and $\sigma^{C}_{M(f/s)}$ are the mean and variance for the client scores from each modality - face or speech, $\mu^{I}_{M(f/s)}$ and $\sigma^{I}_{M(f/s)}$ are the mean and variance for the impostor scores from each modality - face or speech.

The quality of samples of a modality (face or speech in this chapter) is determined by the characteristics of the scores obtained with the development and test samples of that modality. The quality of the face scores (Q_f) and speech scores (Q_s) are calculated as follows:

$$Q_{(f/s)} = D_{M(f/s)} \times T_{(E(f/s)/E1(f/s))} \tag{5.3}$$

where $Q_{(f/s)}$ is the quality for face or speech, D is the quality of the development data, T is the quality of the test (sample) data, $E(f/s)$ is the subset of scores from the test data which is used to determine the quality of the test samples and $E1(f/s)$ is the rest of the scores from the test data which is used to investigate the performance for the proposed scheme.

Based on the equation (5.3), the computation for the quality of samples is divided into two steps. These are described in the following sections.

5.2.1. Estimation of the quality aspects for the development data samples

$D_{M(f/s)}$ in equation (5.3) denotes the quality of the development data for face or speech scores. It is computed based on the scores obtained in the development phase as follows.

$$D_{M(f)} = \frac{l_{M(s)}}{l_{M(s)} + l_{M(f)}}\ , \tag{5.4}$$

$$D_{M(s)} = \frac{l_{M(f)}}{l_{M(s)} + l_{M(f)}}, \tag{5.5}$$

where $l_{M(f)}, l_{M(s)}$ are computed during the development phase using equation (5.6).

$$l_{M(f/s)} = \sqrt{\frac{\left(\sigma^{C}_{M(f/s)}\right)^2}{N^{C}_{M(f/s)}} + \frac{\left(\sigma^{I}_{M(f/s)}\right)^2}{N^{I}_{M(f/s)}}}\ , \tag{5.6}$$

where $N^{C}_{M(f/s)}$ is the total number of clients in the development data for each modality - face or speech, and $N^{I}_{M(f/s)}$ is the total number of impostors in the development data for each modality - face or speech.

5.2.2. Estimation of the quality aspects for the test data samples

$T_{E(f/s)}$ in equation (5.3) represents the quality of the test data for face or speech scores. These quality aspects are calculated using a subset of the test data as follows

$$T_{E(f)} = \frac{k_{E(s)}}{k_{E(s)} + k_{E(f)}}, \tag{5.7}$$

$$T_{E(s)} = \frac{k_{E(f)}}{k_{E(s)} + k_{E(f)}}, \tag{5.8}$$

where $k_{E(f/s)}$ is computed during the test phase as follows

$$k_{E(f/s)} = \frac{\left| \frac{(E(f/s) - \mu^{C}_{M(f/s)})^2}{\sigma^{C}_{M(f/s)}} - \frac{(E(f/s) - \mu^{I}_{M(f/s)})^2}{\sigma^{I}_{M(f/s)}} \right|}{\mu^{C}_{M(f/s)}}. \tag{5.9}$$

In the test phase, $T_{E1(f/s)}$ is computed same as $T_{E(f/s)}$ but using the test data *E1*.

The quality measurements for face scores Q_f and speech scores Q_s are passed to a SVM using two different approaches. These two approaches and the motivation behind them are discussed in the next section.

5.2.3. Methods of passing the quality aspects to SVM

In this chapter, two approaches for passing the relative quality of the test scores to SVM have been studied. The first approach is based on passing relative quality aspects in the individual modality as a separate feature for SVM. In the second approach, relative quality aspects in each of the modalities are fused with the respective scores and then the combined scores are passed as a feature to support vector machine. These approaches are described in the following subsections.

5.2.3.1. Relative quality aspects as independent features (RQ-IF)

In this approach, SVM is fed with four input vectors/ data values, two of these (vectors/ data values) present the actual individual biometric scores (face/ speech) based on the current stage (development/ test) whilst the other two present the relative quality of both the development and test data, as shown in Figures 5.1 and 5.2.

During the development stage, as indicated above, the estimation of the quality of the face and speech scores (Q_f, Q_s) is passed on into the SVM as new features alongside the actual development scores (M_f, M_s). The SVM uses these four input vectors (particularly the former two input vectors) to generate prior knowledge of the expected level of degradation of each biometric data type involved (in the test phase). This helps SVM to tune its parameters to fit the incoming test data.

In the test stage, four input data values are passed on into the classifier (fusion stage), with two of them presenting the quality of the test data (Q_{f_i}, Q_{s_i}). These are computed based on the parameters obtained from the development data (Equation 5.3). The other two data values present the test data itself $(E1_{f_i}, E1_{s_i})$. In the fusion stage, the four input data values are combined and then classified based on the tuned SVM parameters obtained from the development stage.

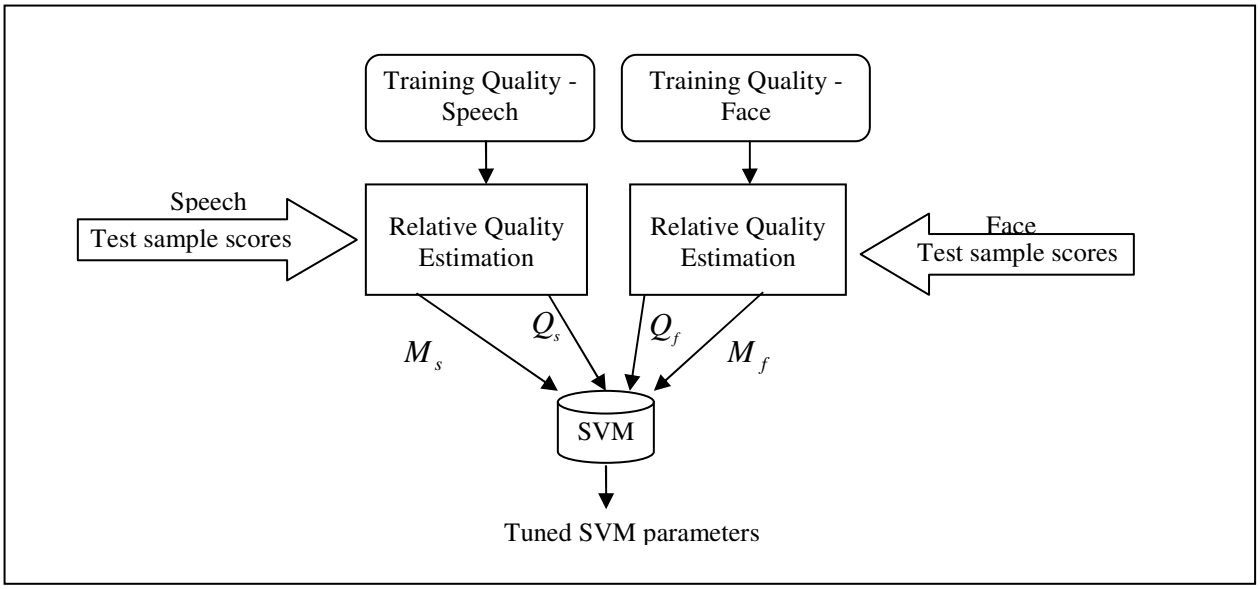


Figure 5.1: Proposed Scheme of SVM-RQM using quality aspect as separate features in the development stage.

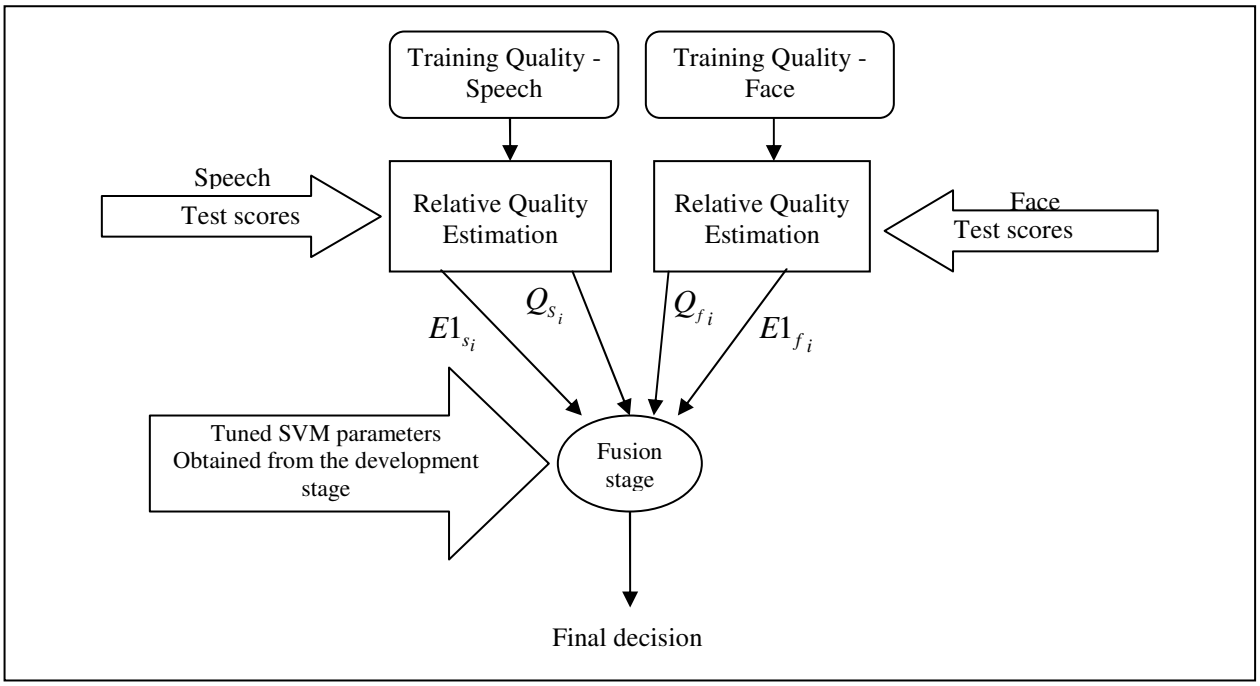


Figure 5.2: Proposed Scheme of SVM-RQM using quality aspect as separate features in the test stage.

5.2.3.2. Modality specific fusion of relative quality aspects (RQ-MSF)

In this approach, the quality of face and speech scores is considered as weights. These weights must be in the interval of 0 and 1 with the condition of $\sum Q_{(f/s)} = 1$. To achieve

this, the quality obtained from equation (5.3) is normalised using the following two steps.

$$S_i = Q_{f_i} + Q_{s_i} \tag{5.10}$$

where S_i is the summation of the i^{th} face quality Q_{f_i} and its corresponding i^{th} speech quality Q_{s_i}. The weight for the i^{th} face or speech scores $W_{(f/s)_i}$ is obtained as,

$$W_{(f/s)_i} = \frac{Q_{(f/s)_i}}{S_i} \tag{5.11}$$

These weights for face or speech scores, which are computed based on their respective test (sample) scores, are then multiplied by their corresponding face or speech scores, respectively.

In the development phase (Figure 5.3), the results of the above multiplications, two weighted input vectors, are used in order to optimise (tune) the parameters of SVM. This is because these parameters are believed to provide useful information about the relative degradation in the different types of biometric data in the test phase since they are partly based on the test sample scores.

In the test phase, weights for face or speech scores, which are computed based on their respective test scores, are multiplied by their corresponding face or speech scores, respectively (Figure 5.4). The resulted two weighted input data values are fused and then classified (in the fusion stage) based on the tuned SVM parameters obtained in the previous phase.

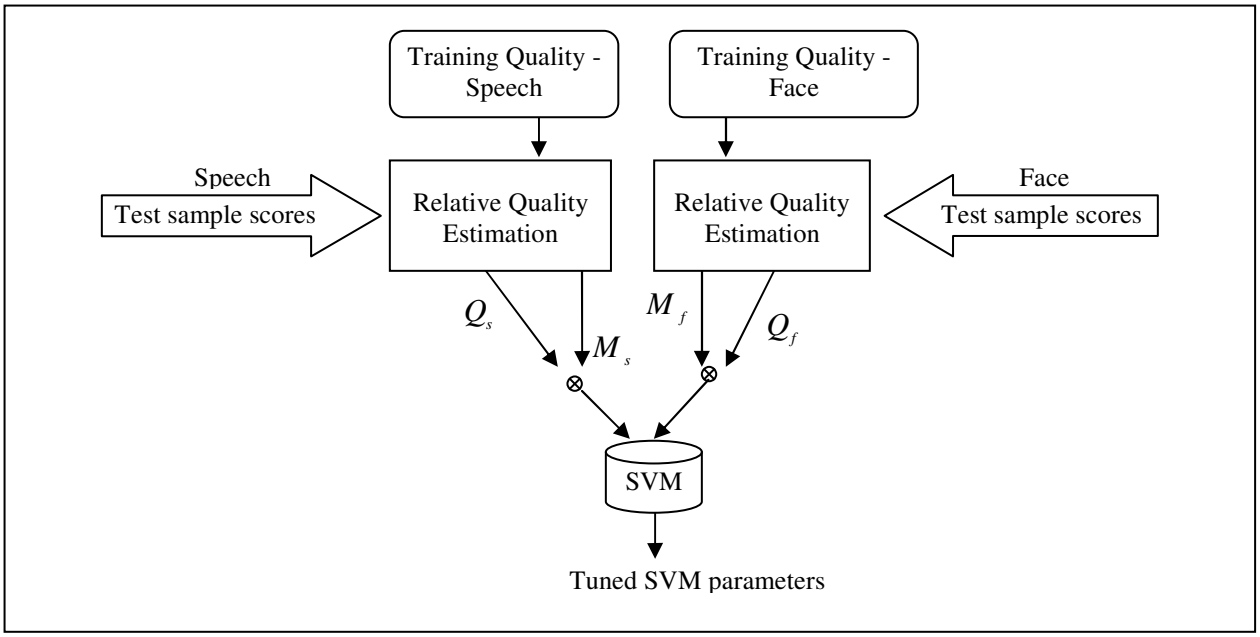


Figure 5.3: Proposed Scheme of SVM-RQM using quality aspect as weights at the development stage.

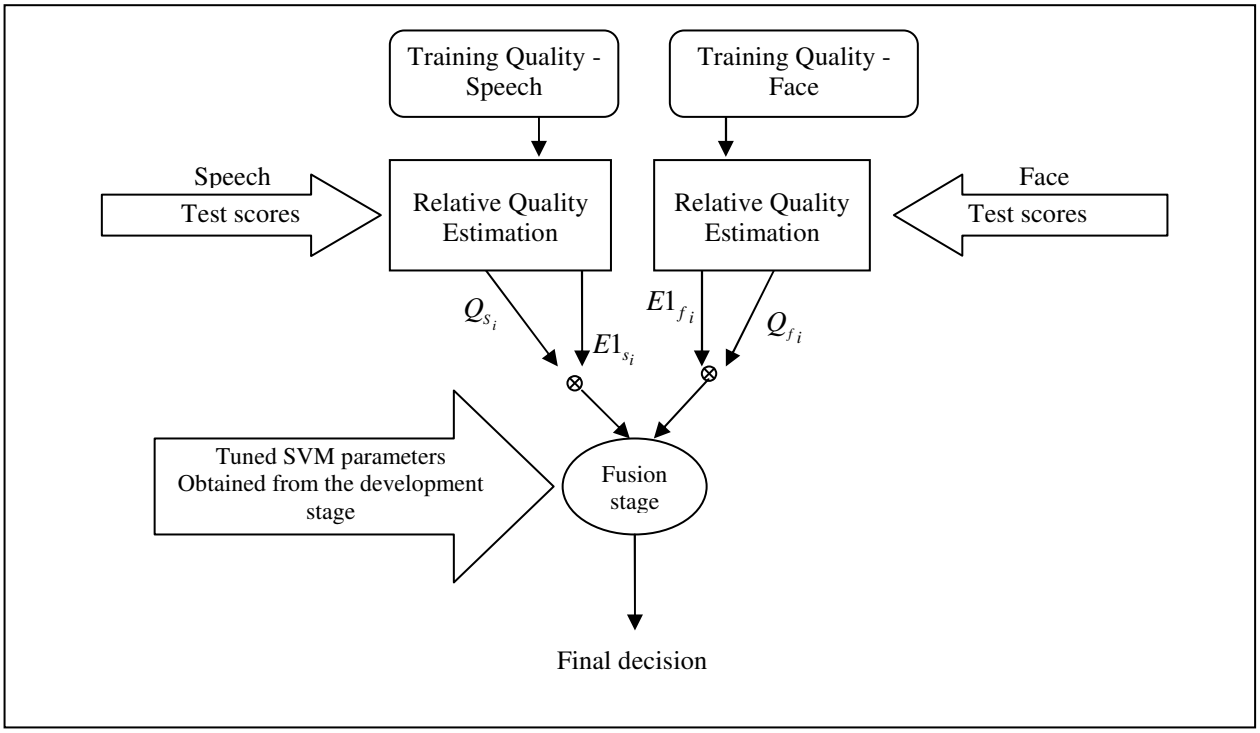


Figure 5.4: Proposed Scheme of SVM-RQM using quality aspect as weights at the test stage.

5.3. Experimental Investigation

5.3.1. Speech and Face data

Experiments are conducted using a subset of the XM2VTS database [90]. This database, as indicated earlier, is a multi-session database containing synchronised image and speech data obtained from 295 subjects, recorded during four sessions which are taken at one month intervals [90]. For the purpose of this study, the subjects in the database are divided into the following sets: the training set is used to train client models; the development set as well as a subset of the test set which is denoted as *E* in equation 5.3 are used to obtain various parameters in the proposed schemes and the test set *E1* is used to investigate the performance. The training set consists of 200 client subjects, the development set consists of 25 non-client subjects and the test set consists of 70 non-client subjects. The total number of 200 client tests and 40000 non-client tests is used from the development data while the total number of client and non-client tests used in finding the relative quality of the test data is 200 and 40000 respectively. The rest of the test set, 200 client tests and 72000 non-client tests, is used to investigate the performance for the proposed scheme. This division is based on the framework of Lausanne protocol which is further described in [90].

The experiments in this chapter are conducted using the same features and classifiers as in the previous chapter (Section 4.2.1). More details about the XM2VTS database are given in Section 4.2.

5.3.2. Testing with Fusion

In the XM2VTS database, the complementary verification scores are based on eight baseline systems which are all included in the configuration 1 of the Lausanne Protocol. Five of the eight baseline systems involve face features and the other three are for speech features. The testing procedure involves combining the scores obtained from two different modalities at each time. In other words, the scores for a face feature are fused with the scores for a speech feature at each time. Therefore, there are fifteen different combinations of features for the fusion purpose. In each experiment, the individual biometric score types involved are subjected to the range equalisation process using the

ZS normalisation [12]. In this work, the fusion process is based on using SVM classifiers. The use of SVM in this chapter is based on earlier investigations (Chapter 4) as well as earlier studies reporting it as one of the most effective methods for multimodal biometrics fusion [1, 87]. The tests are conducted with and without learning the relative quality aspect of the test data.

5.3.3. Results and Discussions

In this study, the results obtained for the authentication tests are given in terms of Equal Error Rate (EER%). Table 5.1 shows the baseline results obtained using the individual features. The Table shows that FH gives the best EER compared to the other face features, whilst LFCC leads to the lowest EER compared to the other speech features.

Feature		Classifier	EER%
face	FH	MLP	*1.78*
	DCTs	GMM	4.15
	DCTb	GMM	1.87
	DCTs	MLP	3.50
	DCTb	MLP	6.49
speech	LFCC	GMM	*1.06*
	PAC	GMM	6.56
	SSC	GMM	4.53

Table 5.1: Baseline EERs computed using the unimodal verification scores in various cases. The best performance in each of the face and voice modalities is shown in italics.

As the experimental results show in Table 5.1, the accuracy rates for the individual modalities in this chapter are observed to be different from the corresponding ones in the previous chapter (Table 4.2). The reason for such behaviour is due to the use of different size of the XM2VTS database for each chapter. This also leads to different results in the fusion stage compared to the corresponding results (using linear SVM) in the previous chapter (Table 4.13). The results for the fusion exercise with and without learning the relative quality of face and speech scores are presented in Table 5.2 in terms of Equal Error Rate (EER%). On the other hand, Table 5.3 presents the Relative Improvements (RI) for the fusion process again with and without learning the relative quality.

No.	Fusion candidates		SVM (without RQM)	SVM-RQM	
	Face	Voice		RQ-IF	RQ-MSF
1	FH	LFCC	0.46	0.34	0.41
2		PAC	0.99	0.86	0.83
3		SSC	0.93	0.65	0.75
4	DCTs - GMM	LFCC	0.92	0.53	0.48
5		PAC	1.76	1.45	1.42
6		SSC	1.17	0.79	1.00
7	DCTb-GMM	LFCC	0.65	0.38	0.22
8		PAC	1.22	0.35	0.43
9		SSC	1.05	0.36	0.35
10	DCTs-MLP	LFCC	0.57	0.41	0.29
11		PAC	1.02	0.98	0.77
12		SSC	1.34	1.16	0.86
13	DCTb-MLP	LFCC	0.64	0.44	0.49
14		PAC	2.24	1.79	1.37
15		SSC	1.84	1.41	1.50
Average EER			1.12	0.79	0.74

Table 5.2: Bi-modal authentication results in terms of EER (%), with and without relative quality learning.

No.	Fusion candidates		SVM (without RQM)	SVM-RQM	
	Face	Voice		RQ-IF	RQ-MSF
1	FH	LFCC	0.57	0.68	0.61
2		PAC	0.44	0.52	0.53
3		SSC	0.48	0.64	0.58
4	DCTs - GMM	LFCC	0.13	0.50	0.55
5		PAC	0.58	0.65	0.66
6		SSC	0.72	0.81	0.76
7	DCTb-GMM	LFCC	0.39	0.64	0.79
8		PAC	0.35	0.81	0.77
9		SSC	0.44	0.81	0.81
10	DCTs-MLP	LFCC	0.46	0.61	0.73
11		PAC	0.71	0.72	0.78
12		SSC	0.62	0.67	0.75
13	DCTb-MLP	LFCC	0.40	0.58	0.54
14		PAC	0.65	0.72	0.79
15		SSC	0.59	0.69	0.67
Average RI			0.50	0.67	0.69

Table 5.3: Relative improvements for the bi-modal authentication with and without relative quality learning.

By comparing the results (e.g. EERs and RIs) for the linear SVM in Tables 4.13 and 4.14 with the corresponding results in Tables 5.2 and 5.3, it can be noticed that such results are not the same although the classifier (linear SVM) is the same in these experiments. This, as indicated earlier, is due to the use of a different size XM2VTS database for each chapter. It is observed from the results in Tables 5.2 and 5.3 that the fusion processes (with and without relative quality learning process) which have been considered

consistently improve the performance beyond that obtainable with the better of the two individual modalities involved. It is also apparent from the results that, in all cases, incorporating the relative quality learning process into the fusion scheme exhibits greater effectiveness than using the fusion process without the relative quality having been learnt. The results also clearly show that the choice of face-voice score combination can have significant impact on the final result. Each combination differs, as stated earlier, from the other in terms of feature and/ or classifier for one modality or both modalities. Based on the EER and RI values in Tables 5.2 and 5.3, it is worth noting that the capabilities of the relative quality-based fusion process in decreasing the verification error rates is considerably higher when DCTb-GMM is used as the face feature. Thus, the discussion presented hereafter concentrates on the experimental results obtained when DCTb-GMM is used as the face feature.

It can be observed that the best results where no relative quality learning process has taken place are obtained by combining the scores obtained from DCTb and LFCC feature. It can also be observed that the reduction in EER obtained by learning the relative quality of the data is quite significant. The lowest EER (0.22%) is observed in the case of DCTb-LFCC combination with SVM-RQM. Such a result is observed when the relative quality is passed to the SVM as weights. The EER reduction in this case is 66% compared with the best result obtained without relative quality learning.

These results clearly show that learning the relative quality information of a score is useful for improving the performance of the multimodal authentication systems. A direct comparison of the results obtained using fusion with and without relative quality learning, together with the baseline results for each of the two cases of DCTb and LFCC is given in terms of DET (Detection Error Trade-off) plots in Figure 5.5.

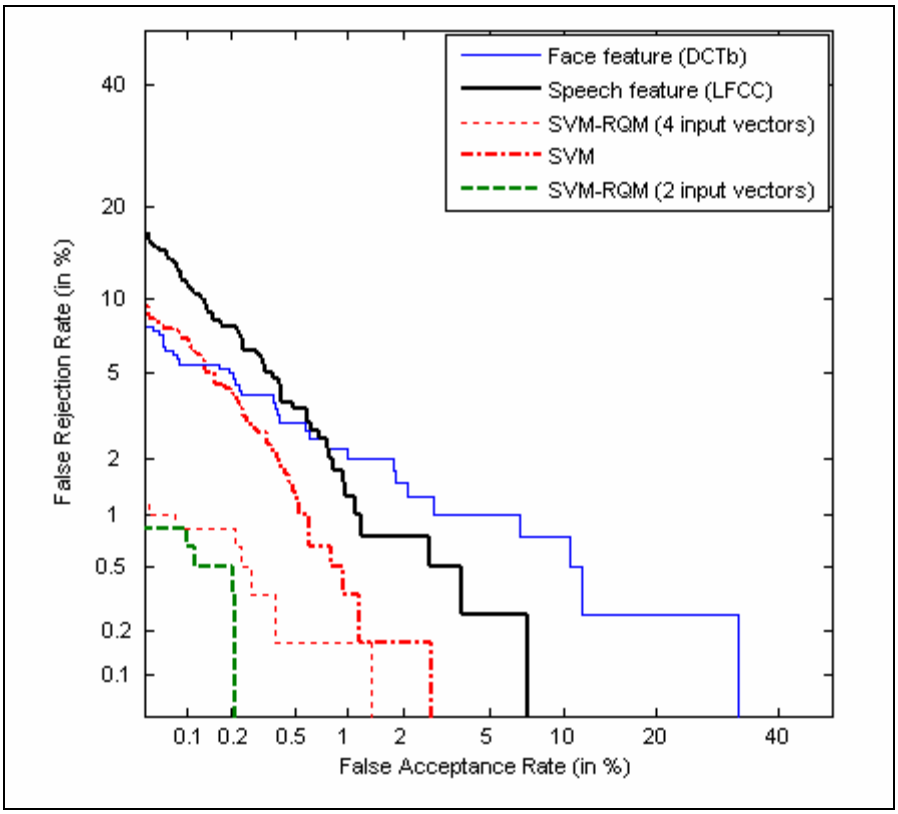


Figure 5.5: DET plots for SVM fusion with and without relative quality learning in bi-modal fusion together with the baseline performers.

Figure 5.6 gives a direct comparison of the RIs obtained using the fusion method (SVM) with and without learning the relative quality of face and speech scores.

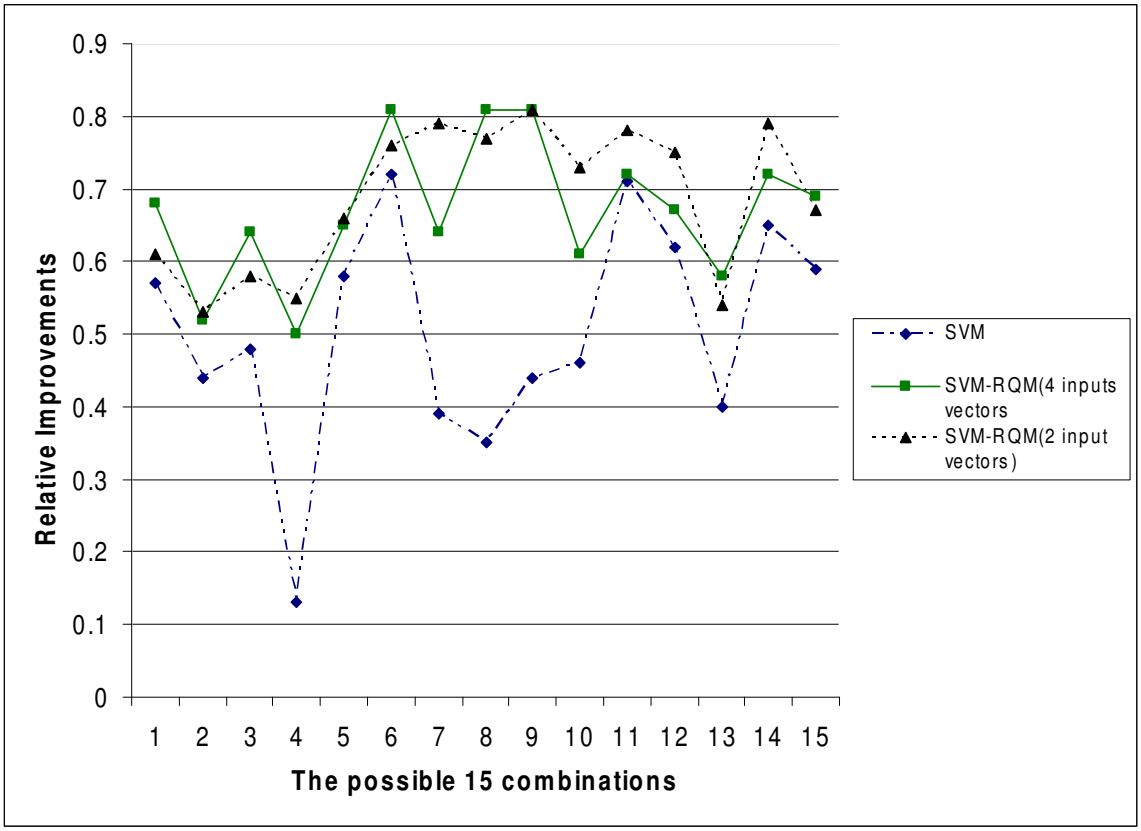


Figure 5.6: Relative improvements for the bi-modal authentication with and without relative quality learning.

It was indicated in Section 3.3.1 that for a fusion scheme to be beneficial, RI should be positive, and the closer it is to 1, the better is the effectiveness of fusion. Based on the above statement it can be observed from Figure 5.6 that amongst the two fusion schemes considered (SVM and SVM-RQM), SVM-RQM scheme has appeared to provide better performance in terms of reducing error rates. Such results prove that Linear SVM can benefit from the relative quality of the testing data in order to decrease the system error rates. However, the choice of face-voice score and quality combination can have significant impact on the final result, as shown in Figure 5.6.

5.4 Summary

The chapter has proposed an approach to enhancing the accuracy of multimodal biometrics (speaker and face verification). The proposed approach is based on adjusting the weights for each of the two modalities according to their relative quality. This is performed by passing the relative quality aspects of the test scores into the Support Vector Machine either as features or weights. Such features and weights provide prior information about the relative degradation in the different types of test biometric data. Such information helps SVM to optimise its parameters to fit the test data. This approach is termed Support Vector Machine with Relative Quality Measurement (SVM-RQM).

The chapter has compared such an approach to the linear SVM. Experimental comparisons of fusion schemes as well as quality measures have been carried out using the XM2VTS database. It is concluded from this chapter that the combination of complementary information from the face and speech can improve the performance over single-modality. Amongst the two fusion schemes considered (SVM and SVM-RQM), SVM-RQM scheme has appeared to provide better performance in terms of reducing error rates. Such results prove that Linear SVM can benefit from the relative quality of the testing data in order to decrease the system error rates. The next chapter presents discussions about the unpredicted variations problem in the evidence captured in the scores. The discussion includes an investigation into the effects, on the accuracy of multimodal biometrics, of introducing score normalisation into the score level fusion process.

Chapter 6

Enhancement of Multimodal Biometric Accuracy

6.1 Introduction

The previous chapter was concerned with the effects, on the accuracy of multimodal biometrics, of relative degradation of the individual biometric modalities involved. An adaptive approach, based on adjusting the weights for each of the two modalities according to their relative quality, was introduced and investigated. This was by estimating the quality of the development data by measuring some parameters for the development score data and then incorporating these parameters in the quality estimation of the test scores.

This chapter proposes an approach to enhancing the accuracy of multimodal biometrics in uncontrolled environments. In general, one of the important problems associated with any multimodal or unimodal technique is the undesired variations in the biometric data. Such variations are reflected in the corresponding biometric scores, and thereby can adversely influence the overall effectiveness of biometric recognition. The said variations can arise due to the effects of data capturing apparatus and various non-ideal operating conditions such as background noise and ambient lighting effects.

The chapter presents investigations into the enhancement of the accuracy of multimodal biometrics, through the introduction of unconstrained cohort normalisation (UCN) into the field. Whilst score normalisation has been widely used in voice biometrics [92, 93], its effectiveness in other biometrics has not been previously investigated. The chapter aims to explore the potential usefulness of the said score normalisation technique in face as well as voice biometrics and to investigate its effectiveness for enhancing the accuracy of multimodal biometrics. The fusion process is performed by SVM (support vector machine). The use of SVM is based on earlier investigations (Chapters 4 and 5) and other earlier studies which report it as one of the most effective methods for multimodal biometric fusion [1, 87]. However, because of the generality of the approach proposed in this chapter, the outcomes should be applicable to other fusion methods as well. The experimental investigations involve the two recognition modes of verification and open-set identification in clean, degraded and mixed-quality data conditions.

The rest of the chapter is organised as follows. Section 6.2 introduces the proposed approach and discusses the motivation behind its use. The experimental investigations and an analysis of the results are presented in Section 6.3, and the overall summary is given in Section 6.4.

6.2. Motivation and proposed approach

An important requirement for the effective operation of a multimodal biometric system in practice is the existence of capability for minimising the effects of variations in the data from the individual modalities deployed. This then leads to maximisation of recognition accuracy in the presence of variation (e.g. due to contamination) in some or all of the types of biometric data involved. In reality, however, this is a challenging requirement as data variation can be due to a variety of reasons, and can have different characteristics. Another aspect of difficulty in multimodal biometrics is the lack of information about the relative variation in the different types of biometric data.

In recent years, there has been considerable research into methods for dealing with data quality in fusion based biometrics [68, 94-98]. However, the work carried out to date has, in general, been concerned with adjusting the balance of weighting in fusion in favour of modalities of better quality. In other words, emphasising or deemphasising the scores for the individual biometric modalities in the fusion process, based on an estimate of their relative degradation. The results of these studies have all verified that the introduction of an appropriate weighting scheme can be beneficial in multimodal fusion. However, it is believed that the effectiveness of multimodal biometrics can be further improved if, through some means, the scores from the degraded modalities can be corrected appropriately. According to the literature, an approach with the potential for offering the above desired capability is that of score normalisation. To date, this method has been used only in the context of speaker recognition [92, 93]. The approach is based on the concept that if anomalous events in the test utterance cause a speaker's score against his (her) own model to degrade, then the scores obtained for the same speaker against certain other background models are also affected in the same way. As a result, the ratio of the score for the target model to a statistic of scores for the considered background models remains relatively unchanged. The use of this ratio instead of the

absolute score for the target model has been shown to improve the verification performance.

The development of the concept of score normalisation in speaker recognition has been based on the fact that the statistical speaker classifiers provide the verification score as the probability of the observed test utterance *x*, given the target model λ. In other words, they compute the probability for the target model producing the observed utterance. However, since the observed test material is in fact the test utterance, what is required to be computed is the probability of the target model, given the test utterance. These two properties are related through the Bayes' theorem as [93, 99]

$$p(\lambda \mid x) = \frac{p(x \mid \lambda)\, p(\lambda)}{p(x)}, \tag{6.1}$$

where $p(.)$ is the probability function. In this equation, the speaker model probability, $p(\lambda)$, can be assumed equal for all speakers, and therefore ignored. $p(x)$, on the other hand, will need to be approximated. To date, diverse approximation approaches have been introduced for this purpose, leading to different score normalisation methods [92, 93, 100]. A slightly different approach to score normalisation in speaker recognition is that based on the standardisation of score distributions, which aims to facilitate the use of a single threshold for all registered speakers [92]. A major difficulty in setting a global threshold in speaker verification (SV) is that both impostor score distribution and true speaker score distribution have different characteristics for different registered speakers. An approach to tackling this issue is that of fixing the characteristics of one of the score distribution types for all registered speakers. Currently, the common practice is to focus on standardising the impostor score distributions. The main reason for operating on the impostor score distributions, rather than on the true speaker score distributions, is the unavailability of sufficient data (in the existing databases) for reliable estimation of the standardisation parameters in the latter approach. The different methods in these two categories of score normalisation (i.e. Bayesian and standardisation) have already been subjected to thorough comparative evaluations in the context of speaker recognition [99, 101]. The normalisation methods considered for this purpose are Cohort Normalisation (CN), Unconstrained Cohort Normalisation (UCN), Universal Background Model (UBM) Normalisation, T-norm and Z-norm. The outcomes, which have been based on

the use of decoupled reference modelling, have indicated UCN as the best performing normalisation technique. The study has also shown that whilst T-norm is amongst the best performers in speaker verification, it provides one of the worst results in the verification stage of open-set identification, even when combined with Z-norm.

The current state-of-the-art in speaker recognition, involves the use of GMM-UBM [74]. The advantage of this approach is twofold. First, it helps alleviate the adverse effects of unseen data. Second, it provides a useful means for score normalisation. However, the method requires the use of UBM-based adapted modelling which is developed specifically for speaker recognition, and is not applicable to other biometric modalities. According to the study in [101], T-norm is extremely effective for open-set speaker identification as well as speaker verification, only when speaker models are obtained by appropriately adapting a universal background model (UBM). Since such adapted modelling is only feasible in the context of speaker recognition, for the purpose of consistency, both biometric modalities considered in this study are based on decoupled reference material. In this case, UCN appears as the best choice for the purpose of score normalisation, and is therefore deployed in this study. It should be pointed out that, in general, such consistency across different modalities involved is not essential. In other words, in multimodal biometrics involving voice, the speaker representation can be based on adapted models, whilst the decoupled representation approach is used for other modalities. In such a scenario, certain other established methods may also be considered for the normalisation of speaker recognition scores, but UCN is still the most appropriate choice for modalities involving decoupled reference material.

In UCN, $p(x)$ in equation (6.1) is approximated as [99, 101]:

$$p(x) \approx \left[\prod_{k=1}^{K} p\left(x|\lambda_k\right) \right]^{\frac{1}{K}}, \tag{6.2}$$

where $p\left(x|\lambda_k\right), k = 1,...K$, are the top K probabilities obtained for the observation, using a set of M background speaker models ($M > K$). These top scoring models are called competing models and their selection is carried out dynamically based on their closeness to the observed utterance in the test phase.

Based on the above, the normalised score can be expressed in the log domain as:

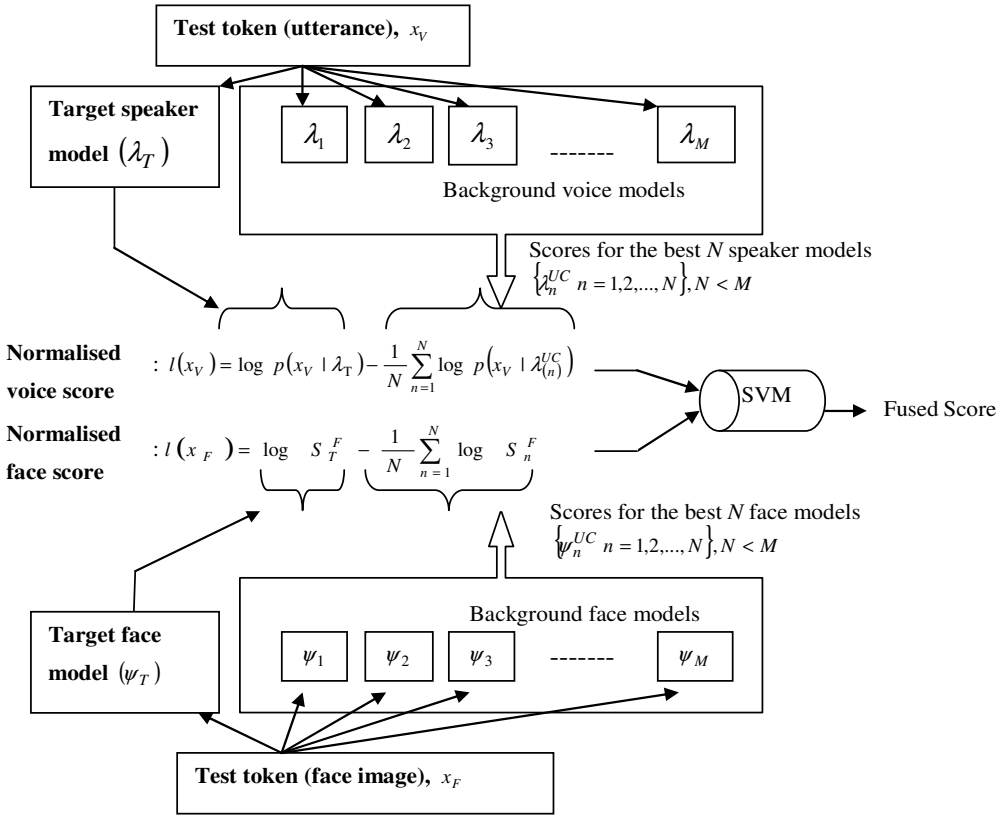
$$S_{UCN} = \log p(x \mid \lambda) - L(x), \quad (6.3)$$

where $L(.) = log\ p(.)$.

This equation suggests that the effects of data degradation can be significantly reduced if these are reflected similarly in $L(x)$ and the target model score. As already shown in [102], this approach works effectively regardless of whether the operating framework is probabilistic or non-probabilistic. Therefore, provided UCN exhibits similar characteristics with other types of biometrics, its application to multimodal biometric fusion can be of considerable value for enhancing the reliability of the process in uncontrolled/varied operational conditions. This is because the approach provides a useful means for appropriately adjusting the individual biometric scores for a client, without any prior knowledge of the level of degradation of each biometric data type involved. However, to date, there have been no reported investigations into the use of UCN with any biometrics other than voice. The aim of this chapter is therefore to explore the potential usefulness of score normalisation in an additional modality (i.e. face biometrics) and to investigate its effectiveness for enhancing the accuracy of multimodal biometrics. Figure 6.1 illustrates the concept of deploying UCN in a multimodal biometric recognition scenario. As observed in this figure, the given test tokens for the individual modalities (e.g. voice, face) are compared against the corresponding reference models for the target identity to produce unimodal scores. For each modality, the test token is also compared against a set of (M) corresponding background models. The top N (<M) background model scores obtained (in the case of each modality) are transferred into the log domain and then averaged together to produce the normalisation term for the considered modality. The normalised score for each modality ($l(x_V)$ or $l(x_F)$in Figure 6.1) is then obtained by subtracting the relevant normalisation term from the logarithm of the score for the target model. The resultant normalised scores for the individual modalities are subsequently fused together through SVM to produce the final multimodal score.

Another interesting and beneficial aspect of using UCN in multimodal biometrics is that it can potentially facilitate the separation of the scores for a given client from those for impostors targeting that client. This is based on the suppression of all the individual

biometric scores for the latter in relation to those for the former. The reason is that, for a given type of biometrics and an adequately large set of background models, an impostor targeting a particular client model is likely to match one or few of the background models more closely. As a result, the application of UCN can result in reducing the impostor biometric scores relative to those of the client. The combination of the above two characteristics of UCN suggests that the technique can help enhance the biometrics' reliability in both clean and adverse conditions. It is also thought that these capabilities should significantly increase the multimodal biometric accuracy. This is because the technique operates on the individual biometric scores involved independently, and the accuracy of the final fused score in multimodal recognition can benefit from the enhancement achieved in all these individual scores.



Note: s_T^F and s_n^F are the scores obtained for the target model (ψ_T) and background models respectively, using the test face image

Figure 6.1: Unconstrained cohort normalisation of scores in multimodal biometric fusion.

6.3. Experimental investigations and results

The experimental studies are concerned with the fusion of face and voice biometrics in the two recognition modes of verification and open-set identification. The modelling and pattern matching approaches used with each modality are not discussed here, as these are outside the scope of this study. The investigations in each mode involve four different data conditions. Two of them are based on the use of scores for clean face images together with scores for either clean or degraded utterances. The other two are based on the use of scores for degraded face images together with scores for either clean or degraded utterances.

In each experiment, the individual biometric score types involved are subjected to the range equalisation process using the ZS normalisation [12]. The process of score-level fusion is based on the use of linear support vector machine (SVM) [85]. The fusion process is applied to the biometric scores with and without subjecting them to the UCN process. This is to determine the level of effectiveness enhancement offered by unconstrained cohort normalisation. The competing models required for UCN are selected from within the set of registered users during the test phase. The cohort size of the competing models is set to 1 and 3 in the cases of clean and degraded data respectively. This is in agreement with the findings in some earlier studies [93, 99]. The procedures for speech feature extraction and speaker classification are as detailed in [99, 101]. The face recognition scores are based on the approaches detailed in [103, 104].

6.3.1. Fusion under Clean Data Conditions

The aim of the experiments in this part of the chapter is to investigate the effectiveness of UCN in enhancing the reliability of multimodal fusion when the biometric datasets are free from degradation. The datasets considered for the face and voice modalities in this investigation are extracted from the XM2VTS and TIMIT databases respectively [103, 105]. Using these biometric datasets, a total of 235 chimerical identities are formed. These consist of 140 clients, 25 development impostors and 70 test impostors. The

development data comprises 140 and 22960 (i.e. 140×{25+[140-1]}) score tokens from the same-users and impostors (including cross-users) respectively. The corresponding score tokens used in the testing phase are 140 and 29260 (i.e. 140×{70+[140-1]}) respectively.

The results for the verification experiments in this part of the chapter are presented as equal error rates (EERs) in Table 6.1. As observed, the use of UCN has resulted in reduction of the verification EERs for the individual modalities and for the fused biometrics. These outcomes confirm the earlier suggestion (Section 6.2) that the use of UCN in clean data conditions is still beneficial. The effectiveness of UCN under such an operating condition is due to its ability to suppress the scores for impostors in relation to those for true users. It is noted that the usefulness of UCN in fused biometrics is mostly due to its performance with the voice modality. However, the corrective effect that UCN has on the face modality is also seen to be considerable. This in turn has helped further enhance the accuracy of classification based on the fused data. It should be emphasised that this is the first time that the use of UCN with face biometrics has been investigated and its effectiveness demonstrated.

Modality	EER% (Without UCN)	EER% (With UCN)
Voice (TIMIT)	2.61	0.05
Face (XM2VTS)	3.57	2.86
Fused: voice and face	0.11	≈0.00

Table 6.1: Effectiveness of UCN in Multimodal verification based on clean biometric data.

Table 6.2 presents the results of open-set identification (OSI) experiments with clean data. These are expressed in terms of IER (identification error rate) and OSI-EER that occur in the first and second stages of the process respectively. An interesting aspect of these results is that the use of UCN does not change the IER for any of the single modalities, whilst it successfully reduces IER (to zero in this case) for the fused biometrics. The reason for this phenomenon can be described as follows. Firstly, like any

other score normalisation method, UCN cannot be expected to correct any misidentification occurring in the first stage of unimodal OSI [101]. However, what is achieved through UCN is the suppression of the scores which lead to the misidentification in the individual modalities in relation to the scores for the correct identities. Although this does not lead to the re-ranking of the unimodal identity scores, it facilitates the reduction of misidentification in the fusion stage. It is also interesting to note that, in this case, the use of UCN appears to ensure that the lowest error rates are obtained through the fused biometrics.

Modality	Without UCN		With UCN	
	IER%	OSI-EER%	IER%	OSI-EER%
Voice (TIMIT)	≈0.00	17.14	≈0.00	2.86
Face (XM2VTS)	9.29	12.86	9.29	8.57
Fused: voice and face	0.71	2.86	≈0.00	≈0.00

Table 6.2: Experimental results for open-set identification based on clean biometric data.

The results in Tables 6.1 and 6.2 indicate that the OSI-EERs in unimodal biometrics are considerably larger than the EERs for the verification experiments. This is due to the fact that the verification stage in open-set identification is more challenging than the standard biometric verification [99]. The reason is that in the former process, each unknown (unregistered) user will need to be discriminated from his/her best matched registered user. In other words, the verification stage in open-set identification can be considered as a specific (but unlikely) scenario in the standard verification process in which each impostor targets only his or her closest model in the registered set.

In multimodal biometrics, however, it is very unlikely that different biometric modalities of an impostor are best matched to the corresponding modalities of an individual registered user. Consequently (as the experimental results show), fusing the biometric scores leads to a significant improvement in the verification accuracy. The use of UCN in this case is observed to maximise the fused biometrics accuracy as well as considerably to reduce the OSI-EER for each of the modalities involved. As indicated

earlier, this is achieved through UCN suppressing the scores for unknown users in relation to those of registered users.

6.3.2. Fusion under Varied Data Quality Conditions

The purpose of the experiments presented in this section is to investigate the usefulness of UCN in multimodal fusion when the qualities of the biometric data types differ considerably.

6.3.2.1. Fusion under clean face data and degraded voice data

The datasets considered for the face and voice modalities in this case are extracted from the XM2VTS (clean images) [103] and from the 1-speaker detection task of the NIST Speaker Recognition Evaluation 2003 (degraded speech) databases respectively [101]. Using these datasets, again a total of 235 chimerical identities are formed. These consist of the same number of clients, development impostors and test impostors as in the previous experiments (Section 6.3.1). The development and test datasets also consist of the same number of score tokens from the same-users and impostors as those considered in the previous section.

The results of verification and open-set identification in this case are presented in Tables 6.3 and 6.4 respectively. It is noted that whilst the error rates for the face modality are exactly the same as those in the previous investigation, due to the use of a degraded speech database, the accuracy rates for the voice modality are in this case lower than the corresponding ones in Section 6.3.1.

Modality	EER% (Without UCN)	EER% (With UCN)
Voice (NIST)	26.24	10.00
Face (XM2VTS)	3.57	2.86
Fused: voice and face	2.86	0.78

Table 6.3: Performance of UCN in biometric verification based on mixed-quality data (clean face data and degraded voice data).

The results in Table 6.3 demonstrate the capability of UCN in reducing the verification error rate, particularly that in fused biometrics. UCN achieves this by a combination of enhancing the client scores when these are affected by data degradation, and suppressing the impostor scores in relation to the client ones. It is noted that without UCN, the fusion process results in improving the EER associated with the better modality by about 20%. According to the results, this reduced EER (2.86%) is further decreased by about 73% through the use of UCN.

Modality	Without UCN		With UCN	
	IER%	OSI-EER%	IER%	OSI-EER%
Voice (NIST)	40	45.71	40	15.71
Face (XM2VTS)	9.29	12.86	9.29	8.57
Fused: voice and face	6.43	12.86	4.29	5.71

Table 6.4: Experimental results for open-set identification based on mixed-quality data (clean face data and degraded voice data).

It is observed from the results in Table 6.4 that the use of fusion process, in this case, leads to reducing the lowest IER offered by unimodal biometrics. However, it is also seen that this capability of fused biometrics is considerably improved through UCN. On the other hand, it is observed that, in this case, the fusion process can only reduce the OSI-EER% when used together with UCN. The reduction in OSI-EER achieved with such a combination is in excess of 55%.

Another important outcome of the experimental investigations can be observed by considering the results in Table 6.4 together with those in Table 6.2. Based on these results, it is clear that the fusion process on its own may not necessarily lead to the reduction of IER or OSI-EER offered by the best single biometric modality involved. The results in these two tables indicate that it is by the deployment of UCN that the fused biometrics consistently outperforms unimodal biometrics.

6.3.2.2. Fusion under degraded face data and clean voice data

The datasets considered for the face and voice modalities in this investigation are extracted from the BANCA (degraded images) and TIMIT (clean speech) databases respectively [104, 105]. Using these biometric datasets, a total of 52 chimerical identities consisting of 26 clients and 26 impostors is formed. The face recognition scores are

obtained based on images captured in a single session, and affected by two different forms of distortion [104]. Based on these and the corresponding score data for TIMIT, a development score dataset is formed for the experiments. This consists of 26 and 1326 (i.e. 26×{26+[26-1]}) score tokens from the same-users and impostors (including cross-users) respectively. The corresponding score tokens used in the testing phase are also 26 and 1326 (i.e. 26×{26+[26-1]}) respectively.

The results of verification and open-set identification in this case are presented in tables 6.5 and 6.6 respectively.

Modality	EER% (Without UCN)	EER% (With UCN)
Voice (TIMIT)	3.99	0.15
Face (BANCA)	15.38	11.54
Fused: voice and face	3.85	≈0.00

Table 6.5: Performance of UCN in biometric verification based on mixed-quality data (degraded face data and clean voice data).

It can be seen in Table 6.5 that fusion without UCN leads to an EER which is just slightly better than the one offered by the best modality involved. However, the use of UCN appears to ensure that the lowest error rate is obtained through the fused biometrics. This outcome again demonstrates the capability of UCN in reducing the error rates for the fused biometrics through enhancing the separation of the scores for each one of the involved modalities.

Modality	Without UCN		With UCN	
	IER%	OSI-EER%	IER%	OSI-EER%
Voice (TIMIT)	≈0.00	19.23	≈0.00	3.85
Face (BNACA)	30.77	26.92	30.77	23.08
Fused: voice and face	≈0.00	11.54	≈0.00	≈0.00

Table 6.6: Experimental results for open-set identification based on mixed-quality data (degraded face data and clean voice data).

It is observed from the results in Table 6.6 that the use of UCN has resulted in reducing the verification OSI-EERs for the individual modalities as well as for the fused biometric. The results also show that, although the usefulness of UCN in fused biometrics is mostly due to its performance with the voice modality, its corrective effect

on the face modality is also beneficial. This as shown has helped reduced OSI-EER (to zero) for the fused biometrics. For IER, the use of fusion process (with and without UCN) successfully reduces IER to zero.

6.3.3. Fusion under degraded Data Conditions

The experiments in this section investigate the effectiveness of UCN in enhancing the reliability of multimodal fusion when the two biometric data types adopted are both degraded. The dataset for the face modality in this investigation is extracted from the BANCA (degraded images) database [104] whilst the data for the speech modality is extracted from the 1-speaker detection task of the NIST Speaker Recognition Evaluation 2003 (degraded speech) database [101]. Using these biometric datasets, a total of 52 chimerical identities consisting of 26 clients and 26 impostors is formed. The face recognition scores are obtained based on images captured in four sessions, and affected by two different forms of distortion [104]. Based on these and the corresponding score data for NIST, a development score dataset is formed for the experiments. This consists of 104 (i.e. 4×26) and 5304 (i.e. 4×{26×[26+(26-1)]}) score tokens from the same-users and impostors (including cross-users) respectively. The corresponding score tokens used in the testing phase are also 104 (i.e. 4×26) and 5304 (i.e. 4×{26×[26+(26-1)]}) respectively. Tables 6.7 and 6.8 present the results obtained in this case for verification and open-set identification respectively.

It can be seen from the experimental results in Table 6.7 that the use of UCN has again resulted in the reduction of the verification EERs for the individual modalities as well as for the fused biometrics. It can also be observed that the fusion process on its own outperforms the best individual modality involved. On the other hand, it is seen that the verification accuracy offered by fused biometrics increases significantly (by about 61%) through the use of UCN prior to fusion. It is worth noting that the accuracy of fused biometrics without UCN (Table 6.7) is below the accuracy obtained by using UCN with any of the two single modalities involved. These results are in agreement with the earlier suggestions (Section 6.2) that the use of UCN in degraded data conditions is beneficial. The effectiveness of UCN under such operating conditions is due to the twofold characteristic of UCN. Firstly it provides a means of enhancing the scores when the test

data is degraded, and secondly, it suppresses the scores from impostors in relation to those for clients.

Modality	EER% (Without UCN)	EER% (With UCN)
Voice (NIST)	35.69	15.38
Face (BANCA)	18.27	13.46
Fused: voice and face	16.35	6.35

Table 6.7: Effectiveness of UCN in multimodal verification based on degraded data.

Modality	Without UCN		With UCN	
	IER%	OSI-EER%	IER%	OSI-EER%
Voice (NIST)	26.92	48.08	26.92	15.38
Face (BANCA)	38.46	30.77	38.46	25.00
Fused: voice and face	25.00	31.73	18.27	9.62

Table 6.8: Experimental results for open-set identification based on degraded biometric data.

In Table 6.8, it is observed that the fusion process results in an IER which is slightly better than the IER offered by the best unimodal biometrics. However, using UCN together with the fusion process leads to a considerably lower IER. It is also observed that, in this scenario, the fusion process reduces the OSI-EER only when used in conjunction with UCN. In fact, without UCN, the OSI-EER obtained with fused biometrics is worse than that for the better of the two modalities. The use of UCN is seen to reduce the OSI-EER for the fused biometrics by about 70%. Again it is noted that, in terms of OSI-EER, the performance of fused biometrics without UCN is well below that of either of the modalities with UCN. In brief, the results in this chapter indicate that it is only through the deployment of an appropriate score normalisation technique, in this case UCN, that the fused biometrics can consistently outperform the unimodal biometrics involved.

Figures 6.2 and 6.3 further illustrate the results obtained for the verification and the second stage of open-set identification experiments in this part of the chapter

respectively. Figure 6.2 clearly shows the significant increase in the reliability of fused biometrics obtained through the use of UCN. The plots in this figure also illustrate the considerable performance improvements achieved through the use of UCN with the individual modalities, which is the cause of the above mentioned enhancement in the accuracy of fused biometrics.

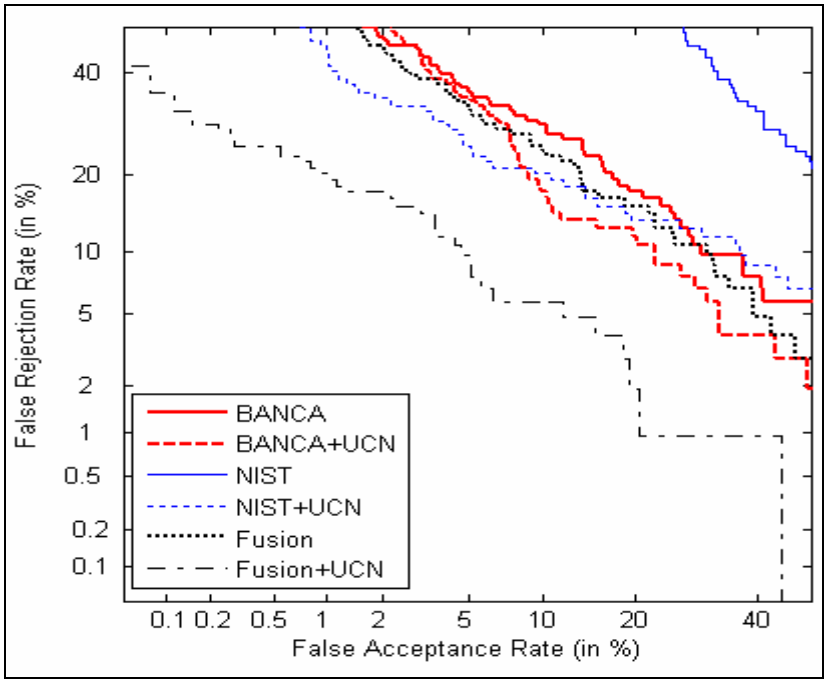


Figure 6.2: DET plots for the verification experiments with degraded data.

The DET plots in Figure 6.3 further emphasise the role of UCN in enhancing the reliability of fused biometrics. In fact, it is observed that, without UCN, the fused biometrics accuracy is highly influenced by the worse of the two modalities involved and does not even match the performance of the better of the two individual modalities. On the other hand, by applying UCN to the individual modalities, the fusion process is observed to provide the highest reliability in the experiments.

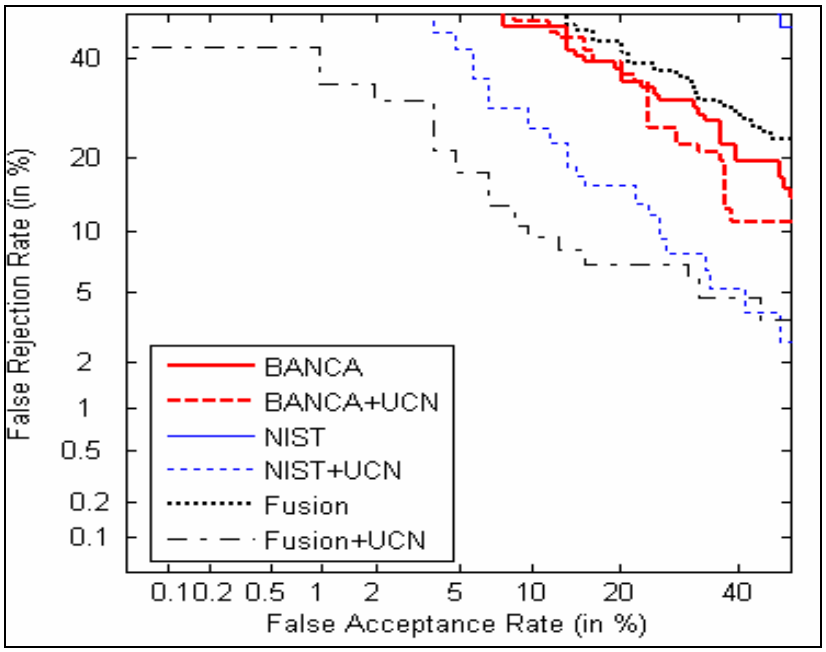


Figure 6.3: DET plots for the verification process in the second stage of open-set identification experiments with degraded data.
Note: the plot for NIST dataset (without UCN) is mostly outside the scale due to the excessively high error rate in this case.

6.4 Summary

The chapter has proposed and investigated the usefulness of unconstrained cohort normalisation (UCN) in face biometrics and also in a multimodal biometric scenario. The experimental investigations have been concerned with the fusion of face and voice biometrics in the two recognition modes of verification and open-set identification. The investigations in each mode have involved four different data conditions.

Based on the experimental investigations, it has been shown that UCN offers considerable improvements to the accuracy of multimodal biometrics in both degraded and clean data conditions. This is shown to be due to the twofold characteristic of this score normalisation method. Firstly it provides a means of enhancing the scores when the test data are degraded, and secondly, it aims to suppress the scores from impostors in relation to those from clients. The investigations have also confirmed the usefulness of UCN in face recognition as well as in speaker recognition for which the technique had originally been developed. Additionally, through a set of open-set identification

experiments, it has been shown that multimodal fusion can consistently outperform the accuracy offered by the best single modality performer when it is combined with UCN.

Chapter 7

Combined Approach to Enhancing Multimodal Biometric Accuracy

7.1 Introduction

The previous two chapters presented two different techniques for tackling the effects of data variations on the fusion process (i.e. variation in relative degradation of data and variation arising from uncontrolled operating conditions). This chapter aims to explore the usefulness of combining UCN with relative quality learning mechanism for the purpose of enhancing accuracy in multimodal biometrics. A two-stage process is adapted. Firstly, the matching scores obtained for face and voice biometrics are normalised. Then, the quality of the normalised scores for each modality is measured. With this knowledge, score-level fusion using SVM (support vector machine) is carried out. The experimental investigations involve the two recognition modes of verification and open-set identification in clean, degraded and mixed-quality data conditions.

The rest of the chapter is organised as follows. Section 7.2 introduces the proposed approach and discusses the motivation behind its use. The experimental investigations together with an analysis of them are presented in Section 7.3, and the overall summary is given in Section 7.4.

7.2. Proposed approach

As indicated earlier, data variations are considered one of the main problems in multimodal fusion. Such variations are reflected in the corresponding biometric scores, and can for this reason adversely influence the overall effectiveness of biometric recognition. As a result, there has been considerable research recently into ways of tackling the problem of data variations, through quality learning schemes [68, 94-98] or score normalisation [106] in fusion-based biometrics. As described in Chapter 5 the quality learning schemes are, in general, concerned with adjusting the balance of weighting in fusion in favour of better quality modalities. In other words, emphasising or deemphasising the scores for individual biometric modalities in the fusion process,

depending on an estimate of their relative degradation [68, 94-98]. On the other hand, it has been shown that the use of unconstrained cohort normalisation in fusion based biometrics helps improve the robustness of multimodal biometrics [106]. This, as indicated earlier, is because the approach provides a useful means for appropriately adjusting the individual biometric scores for a client, without any prior knowledge of the level of degradation of each biometric data type involved. Another motivation for using UCN in multimodal biometrics is that it facilitates the suppression of all the individual biometric scores for impostors in relation to those for the clients. However, it is believed that the accuracy of multimodal biometrics can be further enhanced if the scores from the individual modalities involved are first subjected to UCN [106] and then passed on to the relative quality learning mechanism [98]. This process is expected to enhance the overall accuracy of score level fusion in multimodal biometrics due to the individual capabilities of each technique. The combined method should help enhance the multimodal biometrics reliability in clean, degraded and mixed-quality data conditions. Figure 7.1 illustrates the concept of deploying the proposed method in a multimodal biometric recognition scenario.

7.3. Experimental investigations and results

The fusion of face and voice biometrics in the two recognition modes of verification and open-set identification is again the subject of further experimental studies. The investigations in each mode involve three[1] different data conditions. Two of them use scores for clean face images together with scores for either clean or degraded utterances. The third uses scores for degraded face images together with scores for degraded utterances.

The individual biometric score types involved (in each experiment) are subjected to the range equalisation process using the ZS normalisation [12]. The fusion process is applied to the biometric scores with and without subjecting them to the UCN process. The fusion process, with UCN, is achieved via three different fusing configurations, as shown in Figure 7.1. In the first configuration, the normalised scores for face and voice are combined using the simple linear SVM. This approach is termed Support Vector

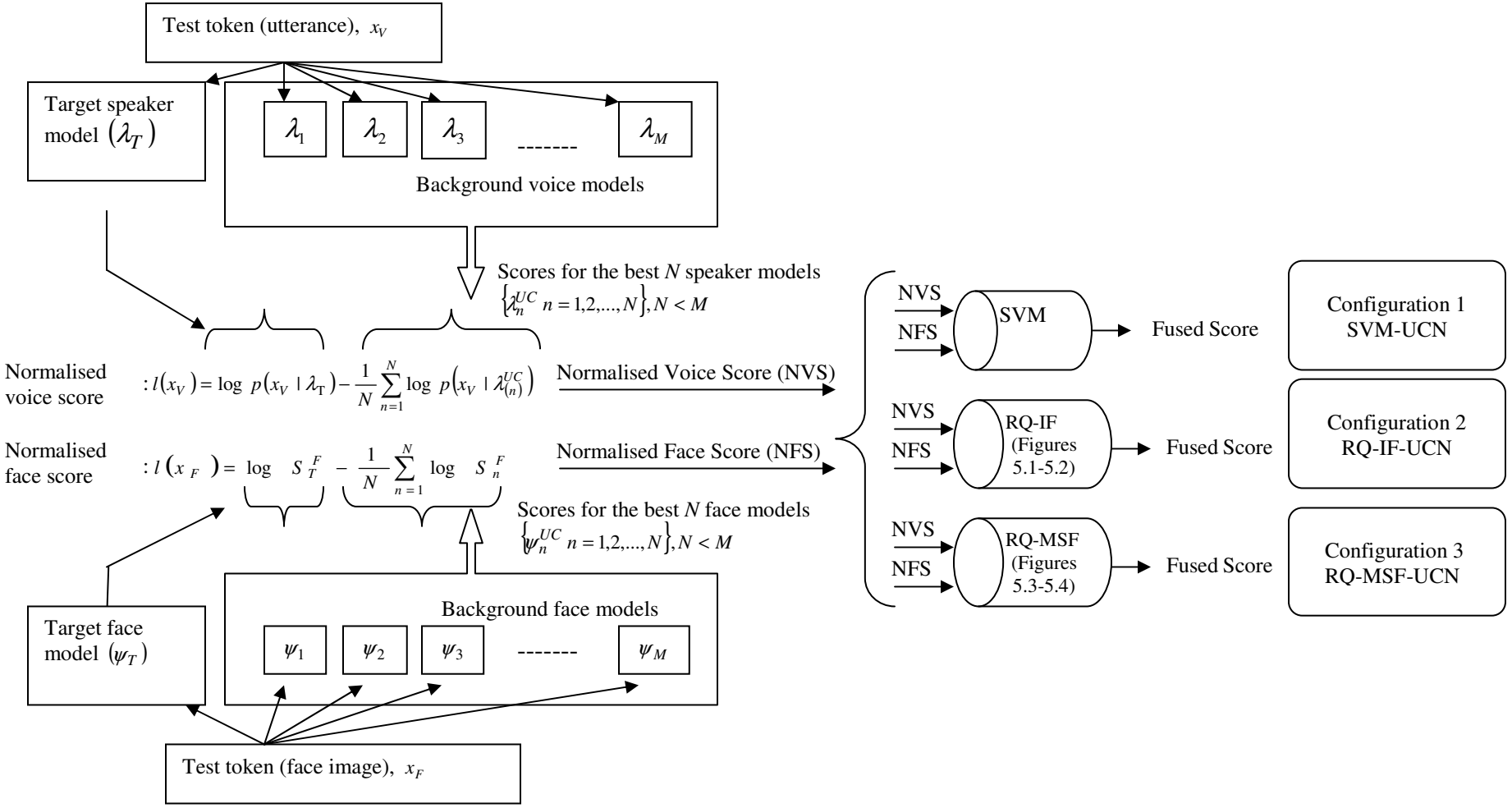
[1] This chapter does not include the fusion of the scores for degraded face images with the scores for clean utterances because of the lack of sufficient amount of data.

Machine with Unconstrained Cohort Normalisation (SVM-UCN). However, in the other two configurations, the normalised scores for face and voice are passed on to SVM-RQM (i.e. RQ-MSF or RQ-IF) in order to measure the relative quality aspects for the individual modalities. The structures and motivation behind RQ-MSF and RQ-IF are discussed in Chapter 5. Such fusing configurations are denoted as RQ-MSF-UCN and RQ-IF-UCN respectively. The competing models required for UCN are selected from within the set of registered users during the test phase. The cohort size of the competing models is set to 1 and 3 for clean and degraded data respectively (see Section 6.3). The experimental results (i.e. verification in both modes) are accompanied by a 95% confidence interval.

7.3.1. Fusion under Clean Data Conditions

The purpose of the experiments in this part of the study is to investigate the effectiveness of the proposed method in enhancing the reliability of multimodal fusion when the biometric datasets are free from degradation. The datasets considered for the face and voice modalities in this investigation are extracted from the XM2VTS and TIMIT databases respectively [103, 105]. The experimental investigations in this part of the study utilise a subset of the database described in Section 6.3.1. A total of 165 chimerical identities are formed in this section. These consist of 70 clients, 25 development impostors and 70 test impostors.

The development data comprises 70 and 6580 (i.e. 70×{25+[70-1]}) tokens from the same-users and impostors (including cross-users) respectively whilst the total number of client and impostor tests used in finding the quality of the test data is 70 and 6580 (i.e. 70×{25+[70-1]}) respectively. In order to investigate the performance of the proposed scheme, 70 client tests and 7980 (i.e. 70×{45+[70-1]}) impostor tests are used.



Note: S_T^F and S_n^F are the scores obtained for the target model (ψ_T) and background models respectively, using the test face image;
RQ is the relative quality; IF and MSF are the two methods of passing on the relative quality to SVM (i.e. Independent Features, Modality Specific Fusion).

Figure 7.1: Unconstrained cohort normalisation with relative quality learning of scores in multimodal biometric fusion.

The experimental results for this part of the study are presented as equal error rates (EERs) with a 95% confidence interval in Table 7.1. The second column in Table 7.1 shows that the use of RQ-MSF and RQ-IF resulted in better performance than the best individual modalities and the fused biometrics with linear SVM. Moreover, it is observed that the use of UCN has resulted in a further reduction of EERs for the individual modalities and for the fused biometrics. It is also noted that the use of qualitative SVM with UCN successfully reduces the EER to zero for the fused biometrics which in this case is comparable to the results obtained using linear SVM with UCN. The advantages of performing quality measurements on the normalised data prior to fusion are not clearly visible in this case because of the use of clean datasets for both modalities.

Modality	EER ± CI 95(%) (Without UCN)	EER ± CI 95(%) (With UCN)
Voice (TIMIT)	2.86 ± 0.36	0.03 ± 0.04
Face (XM2VTS)	3.44 ± 0.39	1.56 ± 0.27
Fusing by SVM	0.16 ± 0.09	$\approx 0.00 \pm 0$
RQ-MSF(2 inputs)	0.08 ± 0.06	$\approx 0.00 \pm 0$
RQ-IF(4 inputs)	0.09 ± 0.07	$\approx 0.00 \pm 0$

Table 7.1: Effectiveness of combining qualitative linear SVM with UCN based on clean biometric data.

Table 7.2 presents the results of open-set identification (OSI) experiments with clean data. These are expressed in terms of IER (identification error rate) and OSI-EER that occur in the first and second stages of the process respectively. As before, the advantages of performing the proposed method on the scores for the biometric data involved are not clearly visible in the case of IER since the databases contain clean data. It is noted that, as in the verification scenario, the use of qualitative SVM results in better OSI-EER compared to linear SVM or the individual modalities involved. It is also observed that subjecting the individual biometric scores to UCN prior to fusion in each of the three different configurations effectively reduces the error rates of the fused scores to zero.

Modality	Without UCN		With UCN	
	IER%	OSI-EER ± CI 95(%)	IER%	OSI-EER ± CI 95(%)
Voice (TIMIT)	≈0.00	18.57 ± 0.85	≈0.00	1.43 ± 0.26
Face (XM2VTS)	12.86	11.11 ± 0.69	12.86	3.57 ± 0.41
Fusing by SVM	≈0.00	4.28 ± 0.44	≈0.00	≈0.00 ± 0
RQ-MSF(2 inputs)	≈0.00	2.86 ± 0.36	≈0.00	≈0.00 ± 0
RQ-IF(4 inputs)	≈0.00	2.94 ± 0.37	≈0.00	≈0.00 ± 0

Table 7.2: Experimental results for open-set identification based on clean biometric data.

7.3.2 Fusion under Varied Data Conditions

The purpose of the experiments presented in this section is to investigate the effectiveness of combining qualitative SVM with UCN when the biometric data types have different levels of quality. The datasets considered for the face and voice modalities in this case are extracted from the XM2VTS (clean images) and from the 1-speaker detection task of the NIST Speaker Recognition Evaluation 2003 (degraded speech) databases respectively [103, 101]. Using these datasets, again a total of 165 chimerical identities are formed, which consist of the same number of clients, development impostors and test impostors, as in the previous experiments (Section 7.3.1). The development and test datasets also consist of the same number of tokens from the same-users and impostors as those considered in the previous section.

The results of verification and open-set identification for this part of the study are presented in Tables 7.3 and 7.4 respectively. There are several observations to be made from these results. Firstly, it is noted that whilst the error rates for the face modality are exactly the same as those in the previous investigation, due to the use of degraded speech database, the accuracy rates for the voice modality in this case are lower than the corresponding ones in Section 7.3.1. It is observed from the results in Table 7.3, that the fusion process (SVM) on its own may not necessarily lead to the reduction of EER offered by the best single biometric modality involved. However, it is noted that the use of SVM-RQM, particularly, using the quality aspects as independent features (RQ-IF), results in improvement of the EER associated with the better modality by about 17%. On the other hand, using linear SVM together with UCN reduces this EER by about 58%. It is interesting to note that the use of relative quality learning mechanisms (i.e. RQ-MSF and RQ-IF) together with UCN results in considerable improvement in the accuracy. A

reduction in EER of 75 % is obtained with such a combination, when the best qualitative SVM performer is RQ-IF.

Modality	EER ± CI 95(%) (Without UCN)	EER ± CI 95(%) (With UCN)
Voice (NIST)	30 ± 1.00	11.43 ± 0.70
Face (XM2VTS)	3.44 ± 0.39	1.56 ± 0.27
Fusing by SVM	3.69 ± 0.41	1.43 ± 0.26
RQ-MSF(2 inputs)	3.32 ± 0.39	0.97 ± 0.21
RQ-IF(4 inputs)	2.86 ± 0.36	0.86 ± 0.20

Table 7.3: Performance of UCN and quality learning in biometric verification based on mixed-quality data.

Figure 7.2 presents a direct comparison of the EERs obtained using fusion based on SVM, RQ-MSF, RQ-IF, SVM-UCN, RQ-MSF-UCN and RQ-IF-UCN together with the EER for the best individual modality involved (face).

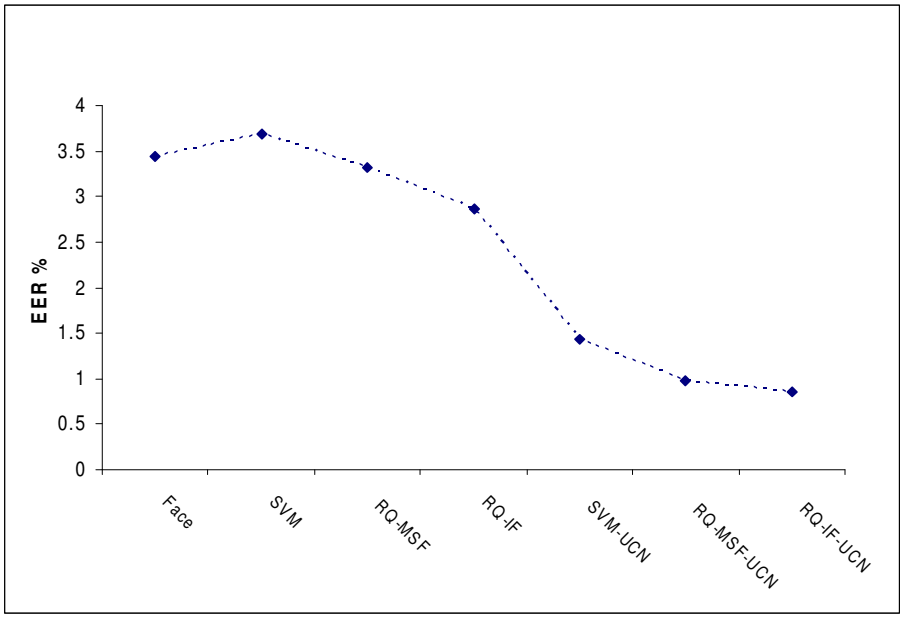


Figure 7.2: Comparison of EERs for various fusion methods with the baseline EER for face modality based on varied quality data.
Recognition mode: Verification.

It is observed from Figure 7.2 that the integration of scores for face and voice by SVM leads to a higher EER compared with the best result without fusion, i.e. face. The two methods of incorporating the quality of the biometric scores (RQ-MSF and RQ-IF) in the

fusion process result in EERs which are just slightly better than the best baseline EER (face). The results obtained with SVM-UCN, RQ-MSF-UCN and RQ-IF-UCN are found to be very encouraging.

It is observed from the results in Table 7.4 that the use of SVM and SVM-RQM leads to lower IERs than that offered by the unimodal biometrics. It is also seen that the capabilities of these fusion processes in decreasing the identification error rates, is considerably higher when combined with UCN. In the second stage of open-set identification, it is shown from the results in Table 7.4 that multimodal biometrics is exhibiting more effectiveness than the unimodal approach. However, the most significant improvement is obtained with RQ-MSF-UCN with an OSI-EER of 2.11%. This level of performance is closely followed by that of the RQ-IF-UCN fusion method. This confirms the earlier suggestion that the combination of normalised biometric scores together with learning the relative quality of the data yields the best OSI-EER.

Modality	Without UCN		With UCN	
	IER%	OSI-EER ± CI 95(%)	IER%	OSI-EER ± CI 95(%)
Voice (NIST)	45.71	41.43 ± 1.08	45.71	15.56 ± 0.79
Face (XM2VTS)	12.86	11.11 ± 0.69	12.86	3.57 ± 0.41
Fusing by SVM	11.43	7.14 ± 0.56	5.71	2.66 ± 0.35
RQ-MSF(2 inputs)	8.89	6.43 ± 0.54	4.33	2.11 ± 0.31
RQ-IF(4 inputs)	8.57	6.67 ± 0.55	4.29	2.16 ± 0.32

Table 7.4: Experimental results for open-set identification based on mixed-quality data.

Figure 7.3 gives a visual representation of the OSI-EERs obtained using fusion based on SVM, RQ-MSF, RQ-IF, SVM-UCN, RQ-MSF-UCN and RQ-IF-UCN together with the error rate for the best individual modality (face).

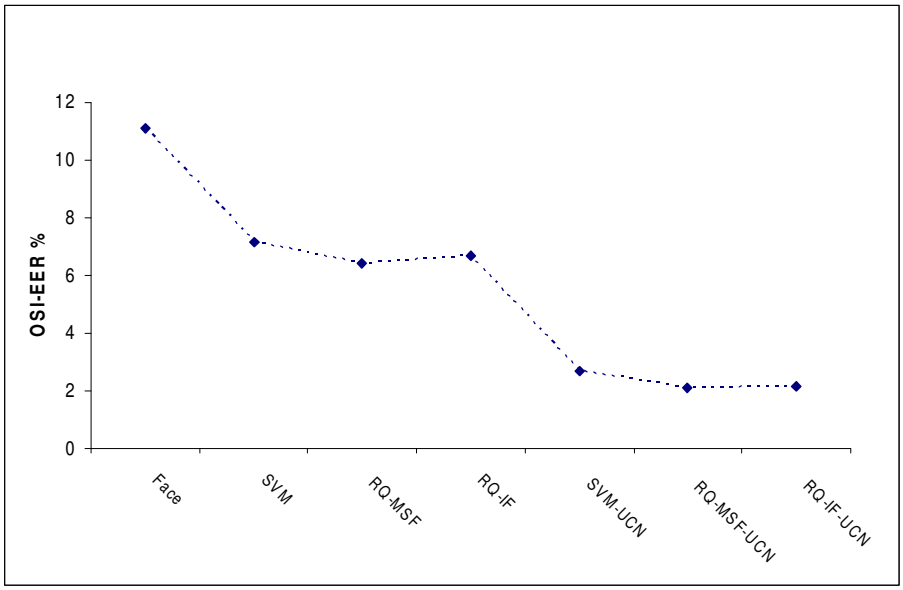


Figure 7.3: Comparison of OSI-EERs for various fusion methods with the baseline OSI-EER for face modality based on varied quality data.
Recognition mode: Verification process in the second stage of open-set identification.

7.3.3 Fusion under Degraded Data Conditions

The aim of the experiments in this part of the chapter is to investigate the effectiveness of combining qualitative SVM with UCN in enhancing the reliability of multimodal fusion when the biometric datasets are contaminated. The datasets considered for the face and voice modalities in this investigation are extracted from the BANCA [104] and NIST Speaker Recognition Evaluation 2003 [101] databases respectively. Using these biometric datasets, a total of 52 chimerical identities consisting of 26 clients and 26 impostors are formed. The face recognition scores are obtained based on images captured in six sessions [104]. These sessions are separated as follows. Two sessions are used for the development data, while four sessions are used for the test data. Two of these four sessions are used to measure the relative quality of the test data whilst the other two are used to investigate the performance of the proposed scheme. Based on these and the corresponding score data for NIST, a development score dataset is formed for the experiments. This consists of 52 (i.e. 2×26) and 2652 (i.e. 2×{26×[26+(26-1)]}) score tokens from the same-users and impostors (including cross-users) respectively whilst the total number of score tokens from the same-users and impostors (including cross-users) used for finding the relative quality of the test data is 52 (i.e. 2×26) and 2652 (i.e.

2×{26×[26+(26-1)]}) respectively. In order to investigate the performance of the proposed scheme 52 (i.e. 2×26) client tests and 2652 (i.e. 2×{26×[26+(26-1)]}) impostor tests are used.

The experimental results for the verification and open-set identification scenarios are presented in tables 7.5 and 7.6 respectively. The performances of the two recognition modes are measured in terms of EERs for the former and IERs and OSI-EERs for the latter.

As the experimental results show in Table 7.5, the accuracy rates for the face modality are lower than the corresponding ones in the previous sections. This is due to the use of a degraded face database. On the other hand, although the speech database is degraded as in the previous section, the accuracy rates for the speech modality in this section are observed to be lower. The reason for such behaviour is the use of a different size subset of the NIST data in this case. It is noted that the use of SVM on its own does not lead to performance better than the best individual modality involved. This is also shown by the results in Section 7.3.2. The results in Table 7.5 demonstrate the capability of reducing the verification error rates by combining UCN with the qualitative SVM. This is thought to result from the three-fold characteristics of this combination. The first is that UCN provides a means for enhancing the scores when the test data is degraded; the second that it aims to suppress the scores for impostors in relation to those for clients; finally, that the use of relative quality measurements further facilitates the reduction in error rates. This is achieved by either assigning higher weights (RQ-MSF) to the best biometric scores or by feeding the SVM with new features (RQ-IF). A direct comparison of the performance (EERs) obtained using the various fusion techniques described above, together with baseline EER for face modality is given in Figure 7.4.

Modality	EER ± CI 95(%) (Without UCN)	EER ± CI 95(%) (With UCN)
Voice (NIST)	40.09 ± 1.85	11.98 ± 1.22
Face (BANCA)	17.68 ± 1.44	13.46 ± 1.29
Fusing by SVM	20.93 ± 1.53	5.42 ± 0.85
RQ-MSF(2 inputs)	12.65 ± 1.25	4.15 ± 0.75
RQ-IF(4 inputs)	14.78 ± 1.34	4.94 ± 0.82

Table 7.5: Effectiveness of UCN and quality learning in verification based on degraded data.

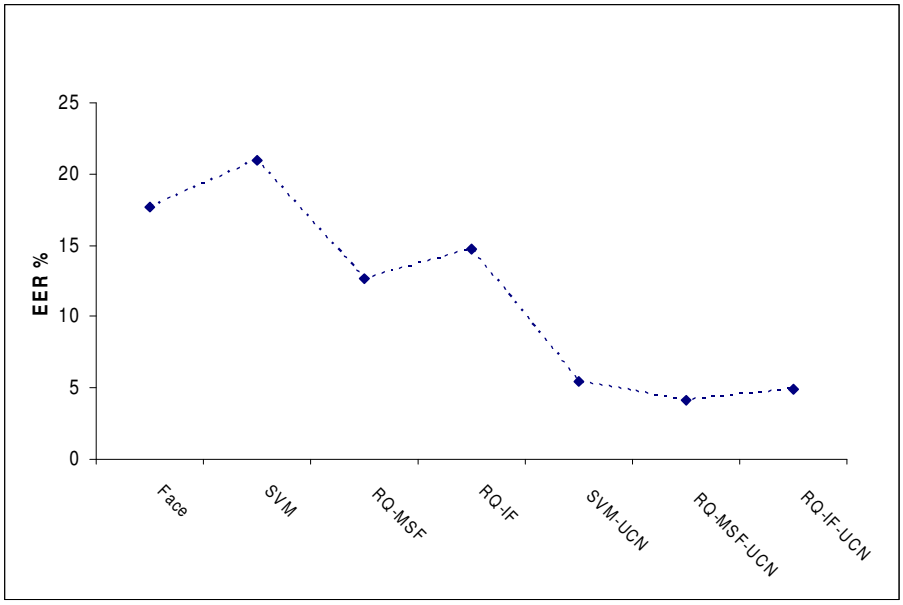


Figure 7.4: Comparison of EERs for various fusing configurations with the baseline EER for face modality based on degraded data.
Recognition mode: Verification.

This figure clearly shows that the proposed technique significantly increases the reliability of fused biometrics.

It can be seen from the results in Table 7.6 that the use of SVM alone results in an increase in both of the IER and OSI-EER obtained with the best single modality. The use of SVM-RQM (i.e. RQ-MSF and RQ-IF) on the other hand can not reduce both types of error together. This is in conflict with the results in Sections 7.3.1 and 7.3.2. The reason for such a phenomenon is that the two databases involved in this part of the study are much more degraded than in the previous sections. However, it should be emphasised

that the combination of UCN with qualitative SVM, successfully reduces both the IER and OSI-EER.

A performance assessment of the results in terms of OSI-EERs is presented in Figure 7.5. The results for each of the two individual modalities used in this investigation are given as baselines.

Modality	Without UCN		With UCN	
	IER%	OSI-EER ± CI 95(%)	IER%	OSI-EER ± CI 95(%)
Voice (NIST)	28.85	59.62 ± 1.85	28.85	15.38 ± 1.36
Face (BANCA)	32.69	32.69 ± 1.77	32.69	25 ± 1.63
Fusing by SVM	46.15	34.62 ± 1.79	19.23	7.69 ± 1.00
RQ-MSF(2 inputs)	21.15	32.69 ± 1.77	19.23	5.77 ± 0.88
RQ-IF(4 inputs)	32.69	30.77 ± 1.74	23.08	3.85 ± 0.72

Table 7.6: Experimental results for open-set identification based on degraded data.

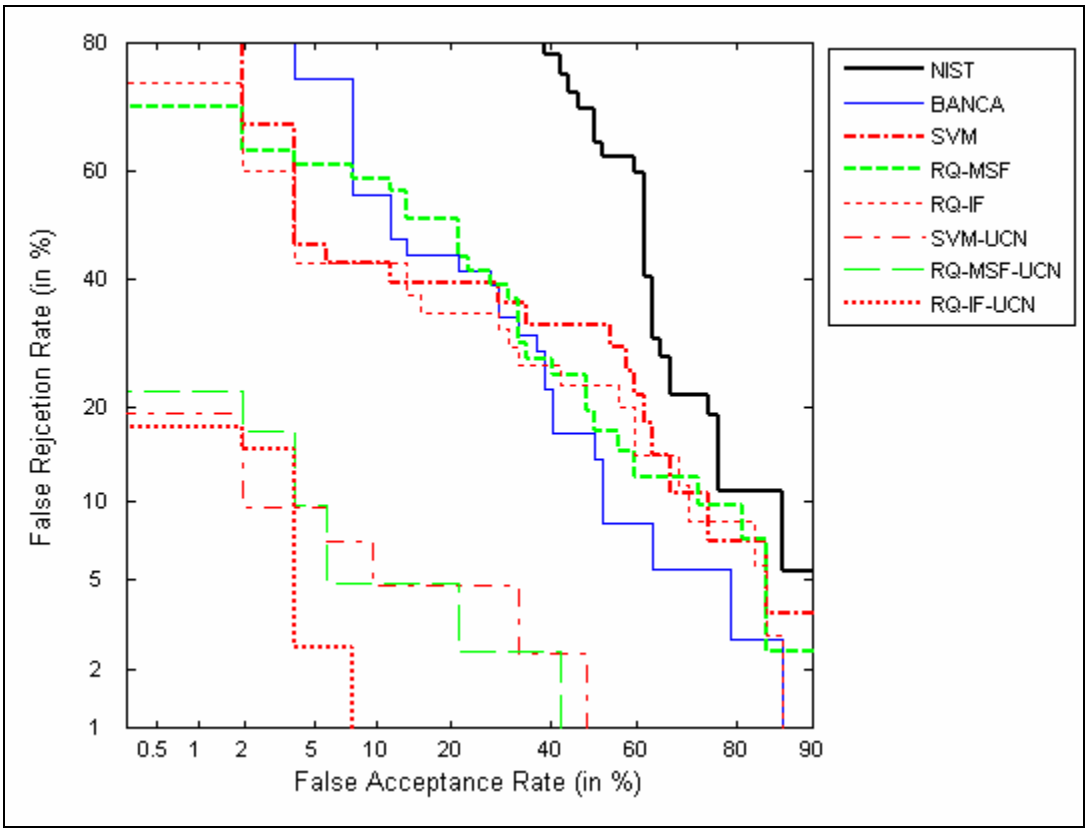


Figure 7.5: DET plots showing the effects of qualitative SVM and UCN on the verification process in the second stage of open-set identification experiments with degraded data.

It can be observed from Figure 7.5 that combining UCN with the qualitative SVM appears to provide better performance in terms of reducing error rates (OSI-EERs). These outcomes confirm the earlier suggestion (Section 7.2) that the reliability of multimodal biometrics can be further increased if the scores from the individual modalities involved are first subjected to UCN and then passed on to the relative quality learning mechanism.

Figure 7.6 gives a visual representation of the OSI-EERs obtained using fusion based on SVM, RQ-MSF, RQ-IF, SVM-UCN, RQ-MSF-UCN and RQ-IF-UCN together with the error rate for the best individual modality (face).

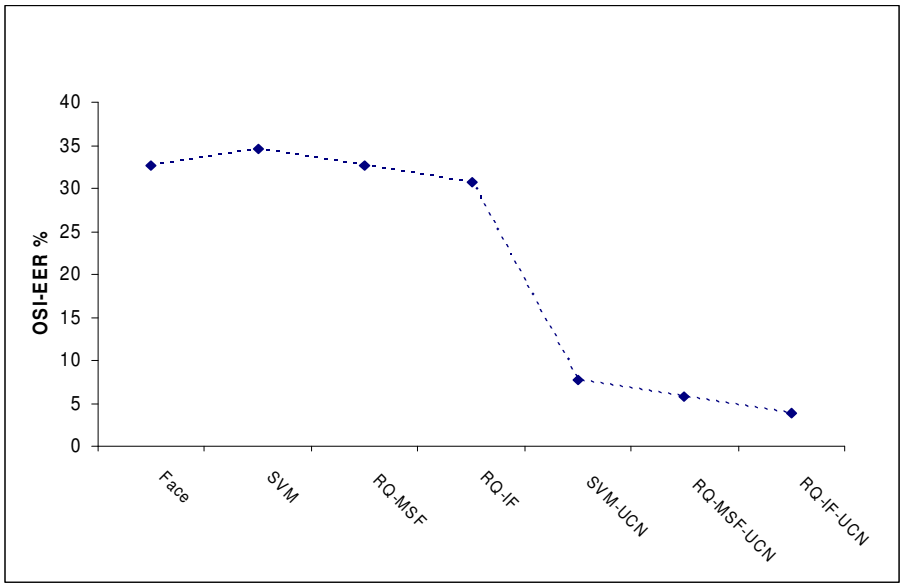


Figure 7.6: Comparison of OSI-EERs for various fusing configurations with the baseline OSI-EER for face modality based on degraded data.
Recognition mode: Verification process in the second stage of open-set identification.

As shown in Figure 7.6, a sharp drop in OSI-EERs is obtained with the fusion technique SVM-UCN. However, passing on the normalised face and voice scores to the relative quality learning mechanism further improves the accuracy of multimodal biometrics.

Some important outcomes of the experimental investigations can be observed by considering the results in all the tables shown above. From these results, it is clearly seen that in all three data conditions, combining UCN with relative quality learning mechanism consistently lead to the best performance whether it is in verification or

open-set identification mode. It can also be seen that neither of RQ-IF and RQ-MSF appears to perform consistently better than the other. This is thought to be due to the different processes involved in passing on the quality of the scores to the SVM. Another possible reason for such behaviour is the biometric data conditions involved. It should also be pointed out that the error rates obtained in this chapter, for the individual modalities and the fused biometrics scores using non-qualitative SVM with or without UCN, differ from those obtained in the previous chapter. This is due to the use of different sizes of databases.

7.4 Summary

An investigation into the use of unconstrained cohort normalisation (UCN) combined with qualitative score-level fusion for multimodal biometrics has been presented. The experimental investigations have been carried out under three different data conditions. The experimental results have shown that, in the two cases of verification and open-set identification, the combination of UCN with relative quality learning measurements is more effective than either the best single modality performer or the use of only one of these techniques with SVM. The reason for this seems to relate to the individual characteristics of the two techniques: UCN aims to compensate for degraded scores and to suppress the impostor scores with respect to the client scores; whilst SVM-RQM makes use of the knowledge of the relative level of degradation of biometric data types involved (in the test phase).

Chapter 8

Conclusions and Future work

8.1 Summary and conclusions

Combining multimodal data has shown to be a very promising trend, both in experiments and in real-life biometric authentication applications. Multimodal biometric systems can overcome some of the limitations of unimodal systems. For example, the problem of non-universality is addressed since multiple traits can ensure sufficient population coverage. Also, multimodal biometric systems make it difficult for an intruder to simultaneously spoof the multiple biometric traits of a registered user. The key to multimodal biometrics is the fusion of various biometric data. Fusion can occur at various levels, the most popular one is the score level where the scores output by the individual modalities are integrated.

A critical question is how to integrate these scores. As part of this study, a review of well-established fusion methods has been carried out. The experimental investigations have included the use of fusion methods in both unimodal and multimodal biometrics. In particular, two types of biometrics (i.e. face and voice) have been considered in the investigations. The individual modality scores are obtained using the XM2VTS database. The scores are based on eight baseline systems. Five of the eight baseline systems involve face features and the other three are for speech features. In each experiment, the scores to be fused are subjected to the range-equalisation process prior to fusion (to achieve values in a common range). This is based on MM or ZS range-normalisation techniques. Nine fusion schemes are used in the fusion stage. These are BFS, MW-(FAR/FRR), MW-EER, FLD, QDA, LR, Linear SVM, Poly SVM, and RBF SVM. These techniques are known as non-adaptive as they involve determining all the fusion parameters using the development set.

Based on the baseline EERs computed using the unimodal verification scores, it is noted that, FH gives the best EER compared to the other face features, whilst LFCC leads to the lowest EER compared to the other speech features. By comparing the results obtained

for the two cases (i.e. unimodal and multimodal biometrics), it is evident that higher accuracy is a main advantage of multimodal biometrics over unimodal biometrics. The reason of such findings is that separate information from different modalities is used to provide complementary evidence about the identity of the users. On the other hand, comparing the results for the MM range-normalisation method with the corresponding results for ZS range-normalisation shows that better performance can be obtained with the latter normalisation method. With ZS range-normalisation, the fusion process (in most cases) improves the performance beyond that obtainable with the better of the two individual modalities involved. In particular, the seven top fusion methods (i.e. BFS, MW-(FAR/FRR), MW-EER, LR, Linear SVM, Poly SVM, and RBF SVM) considered in the multimodal scenario are found to provide consistent improvement regardless of the choice of face-voice features considered. For these seven fusion methods, it is noted that the best results are obtained when DCTs is used as the face feature. Based on the results it is noted that the usefulness of each fusion method varies with the choice of feature and classifier used for each modality.

Another issue of concerns in this thesis is the effect of the data variation on the recognition performance of biometric systems. Such variations are reflected in the corresponding biometric scores, and thereby can adversely influence the overall effectiveness of biometric recognition. The term data variation, as used in this thesis, is subdivided into two types. These are, variation in each data type arising from uncontrolled operating conditions, and variation in the relative degradation of data. The former variation can be due to operating in uncontrolled conditions (e.g. poor illumination of a user's face in face recognition), or user generated (e.g. carelessness in using the sensor for providing fingerprint samples). The variation in the relative degradation of data is due to the fact that in multimodal biometrics different data types are normally obtained through independent sensors and data capturing apparatus. Therefore, any data variation of the former type (discussed above) may in fact result in variation in the relative degradation (or goodness) of different biometric data deployed. The thesis has made a number of contributions aimed at tackling the above-mentioned variations. For the relative degradation problem, it was found from the definition of the non-adaptive fusion technique, that the drawback of this technique is the possible

mismatch between the relative variation of the biometric modalities involved in the development and test data respectively. For example, if one modality (e.g. voice) leads to good performance in the development stage, compared to the other modality (e.g. face), but does not retain the same relative performance at the test stage, this can adversely affect the outcome of multimodal biometrics. To tackle this problem, it would be logical to consider the relative levels of contamination in different biometric data not only in the development phase, but also at the test stage. Therefore, the thesis presents an adaptive approach to reduce the effects of such relative degradation in multimodal fusion. The proposed approach is based on adjusting the weights for each of the two modalities according to their relative quality. This is performed by estimating the relative quality aspects of the test scores and then passing them on into the Support Vector Machine either as features or weights. The former approach is based on passing the relative quality aspects in the individual modality as a separate feature for SVM. This technique is termed Relative Quality aspects as Independent Features (RQ-IF). In the latter approach, the relative quality aspects in each of the modalities are fused with the respective scores and then the combined scores are passed on as a feature to SVM. This is referred to as Modality Specific Fusion of Relative Quality aspects (RQ-MSF). Since the fusion process is based on the learning classifier of the Support Vector Machine, the technique is termed Support Vector Machine with Relative Quality Measurement (SVM-RQM). Such an approach is compared with the linear SVM. Experimental comparisons of fusion schemes as well as quality measures have been carried out using the XM2VTS database. Amongst the two fusion schemes considered (SVM and SVM-RQM), SVM-RQM scheme has appeared to provide better performance in terms of reducing error rates. Such results prove that Linear SVM can benefit from the relative quality of the testing data in order to decrease the system error rates. This is because SVM-RQM provides prior information about the relative degradation in the different types of test biometric data. Such information helps SVM to optimise its parameters to fit the test data.

Although SVM-RQM has helped decrease the system error rates, it is believed that the effectiveness of multimodal biometrics can be further improved if, through some means, the scores from the degraded modality can be corrected appropriately. An approach with

the potential for offering the above desired capability is that of score normalisation. To date, this method has been used only in the context of speaker recognition. According to the literature, there have been different approximation approaches introduced for this purpose, leading to different score normalisation methods (i.e Cohort Normalisation (CN), Unconstrained Cohort Normalisation (UCN), Universal Background Model (UBM) Normalisation, T-norm and Z-norm). UCN appears as the best choice for the purpose of score normalisation, and therefore deployed in this thesis. This is because the approach provides a useful means for appropriately adjusting the individual biometric scores for a client, without any prior knowledge of the level of degradation of each biometric data type involved. However, to date, there have been no reported investigations into the use of UCN with any biometrics other than voice. Therefore, the thesis has explored the potential usefulness of UCN in face biometrics and investigated its effectiveness in enhancing the accuracy in a multimodal biometric scenario. The experimental investigations have been concerned with the fusion of face and voice biometrics in the two recognition modes of verification and open-set identification. The investigations in each mode have involved four different data conditions. Two of them are based on the use of scores for clean face images (XM2VTS) together with scores for either clean (TIMIT) or degraded utterances (NIST Speaker Recognition Evaluation 2003). The other two are based on the use of scores for degraded face images (BANCA) together with scores for either clean (TIMIT) or degraded utterances (NIST Speaker Recognition Evaluation 2003).

In each experiment, the individual biometric score types involved are subjected to the range equalisation process using the ZS normalisation. The linear support vector machine (SVM) is used for the purpose of fusion. The fusion process is applied to the biometric scores with and without subjecting them to the UCN process. This is to determine the level of effectiveness enhancement offered by unconstrained cohort normalisation. The fusion process, with UCN, is denoted as SVM-UCN.

Based on the experimental investigations, it has been shown that UCN offers considerable improvements to the accuracy of multimodal biometrics in both degraded and clean data conditions. This is shown to be due to the twofold characteristic of this score normalisation method. Firstly it provides a means for enhancing the scores when

the test data is degraded, and secondly, it aims to suppress the impostor scores in relation to those for clients. The investigations have also confirmed the usefulness of UCN in face recognition as well as in speaker recognition for which the technique had originally been developed. Additionally, through a set of open-set identification experiments, it has been shown that multimodal fusion can consistently outperform the accuracy offered by the best single modality performer, when it is combined with UCN.

The encouraging results of the previous techniques (i.e. SVM-RQM and SVM-UCN) motivate further research in order to introduce a new approach to enhancing the accuracy of multimodal fusion. Such an approach is based on a two-stage process. Firstly, the matching scores obtained for face and voice biometrics are normalised. Secondly, the quality of the normalised scores for each modality is then measured. Using this knowledge, score-level fusion is carried out using SVM. The experimental investigations have been carried out under three different data conditions. The experimental results have shown that, in the two cases of verification and open-set identification, the combination of UCN with relative quality learning measurements is more effective than either the best single modality performer or the approaches based on using only one of these techniques with SVM. This has been attributed to the individual characteristics of the two techniques: UCN aims to compensate for degraded scores and to suppress the impostor scores with respect to the client scores; whilst SVM-RQM makes use of the knowledge of the relative level of degradation of biometric data types involved (in the test phase).

8.2 Suggestions for future work

The experimental investigations, in this thesis, involve the two recognition modes of verification and open-set identification, in clean, mixed-quality and degraded data conditions. The results show that the performance of biometric systems can benefit from score level fusion, but that this depends highly on the types of fusion technique as well as the range-normalisation method used. Hence, future work will focus on these two factors which play important roles in the effectiveness of multimodal biometric systems. An explanation of the nature of the problem of multimodal biometric systems (how these two factors can affect the performance) and suggested solutions to obtain an optimal multimodal biometric system is discussed in the rest of this section.

8.2.1 Range-Normalisation

Range-normalisation (Section 3.2) refers to the transformation of single modality scores into a common domain prior to combining them. Several studies have shown the significant influence of the range-normalisation techniques prior to fusion in biometric recognition task. For example, Srihari et al [70] claimed that range-normalisation is a necessary task because scores from different systems are incomparable. In [71] Altinay et al mentioned that in the case of using linear fusion techniques to integrate the scores of the individual modalities, score incomparability affects the system performance. Indovina et al. [12] evaluated the effects of range-normalisation techniques (Min-Max, Z-score, Tanh, Quadric-Line-Quadric) and fusion methods (Simple Sum, Min score, Max Score, Matcher Weighting, User Weighting) on the performance of a multimodal biometric system using face and fingerprint modalities. Their experiments showed that Min-Max and Quadric-Line-Quadric normalisation methods lead to the best performance except for Min score fusion technique. However, they do not offer any reasons for such a behavior. Although there exists a number of studies regarding range-normalisation, there still exist some questions to be addressed. These are, “Does range-normalisation affect the original score distribution for clients and impostors? What are the effects of linear

and non-linear range normalisation techniques on the performance of linear and non-linear fusion methods?".

8.2.2 Fusion techniques

The experimental results have shown that although non-adaptive fusion techniques (i.e. linear SVM) might lead to good performance in clean conditions, they fail in noisy conditions. This, as expected, is in agreement with the results presented in [9]. Sanderson et al [9] has indicated that in clean conditions, the integration of the scores for face and voice by SVM obtains performance better than either face or voice features. However, in high noise levels (SNR=-8dB) [9], the SVM performance has been found to be worse than the face feature. This is expected since SVM is a non-adaptive fusion technique. Therefore, the thesis has presented several approaches to help the fusion process (i.e. linear SVM) to enhance its performance regardless of the biometric data conditions. In keeping with this line of research, this section presents possible ways for future work.

8.2.2.1 Quality estimation

An important aspect of the future work is to investigate further methods for evaluating the quality of the testing data. The distance between a reference model and the model associated with a claimant can be useful for evaluating the quality of testing data. However, this area should be subjected to thorough investigations to identify the most appropriate approach.

8.2.2.2 Unconstrained cohort normalisation at feature level

Chapters 6 and 7 show that introducing UCN into the score level fusion process can improve the system performance. Since, for each modality, the biometric score is obtained by accumulating the feature scores, it would be interesting to investigate UCN usefulness at such a level.

8.2.2.3 Unconstrained cohort normalisation for other types of biometrics

Whilst this thesis has confirmed the effectiveness of UCN for face modality as well as fused voice and face biometrics, further investigations are required to determine what other types of biometrics can benefit from such forms of score normalisation.

8.2.2.4 Unconstrained fusion techniques

In the opinion of the author, another area worth investigating is that of unconstrained fusion methods. These methods are defined here as the fusion approaches requiring no development data. In fact, unconstrained fusion methods can be defined as a subset of adaptive fusion methods. These should only require information about the quality of the test data in order to provide multimodal-based discrimination between the clients and impostors. The investigations into unconstrained fusion techniques will be carried out over clean and noisy databases and the results will be compared with these of other approaches.

References

[1] A. K. Jain, A. Ross, and S. Prabhakar, "An Introduction to Biometric Recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, pp. 4-19, 2004.

[2] S. Prabhakar, S. Pankanti, and A. K. Jain, "Biometric recognition: security and privacy concerns," *IEEE Security & Privacy*, pp. 33-42, 2003.

[3] K. Delac and M. Grgic, " A Survey of Biometric Recognition Methods " *46th International Symposium Electronics*, pp. 184-193, 2004.

[4] A. Ross and A. Jain, "Information Fusion in Biometrics," *Pattern Recognition Letters, Special Issue on Multimodal Biometrics*, vol. 24, pp. 2115-2125, 2003.

[5] U. M. Bubeck, "Multibiometric authentication: An overview of recent developments," in *Term Project CS574 Spring*. San Diego State University, 2003.

[6] C. Sanderson and K. K. Paliwal, "Identity Verification using Speech and Face Information," *Digital Signal Processing*, vol. 14, pp. 449-480, 2004.

[7] M. Faundez-Zanuy, "Data fusion in biometrics," *IEEE Aerospace and Electronic Systems Magazine*, vol. 20, pp. 34-38, 2005.

[8] S. Pigeon and L. Vandendorpe, "Image-based multimodal face authentication," *Signal Processing*, pp. 59-79, 1997.

[9] C. Sanderson and K. K. Paliwal, "Information Fusion and Person Verification Using Speech and Face Information," *IDIAP-RR 02-33*, 2003.

[10] A. K. Jain and A. Ross, "Multibiometric Systems," *Interagency Information Exchange on Biometrics*, 2003.

[11] A. K. Jain, K. Nandakumar, and A. Ross, "Score normalisation in multimodal biometric systems" *Pattern Recognition*, vol. 38, pp. 2270–2285, 2005.

[12] M. Indovina, U. Uludag, R. Snelick, A. Mink, and A. Jain, "Multimodal Biometric Authentication Methods: A COTS Approach," *Proceedings of Multi-Modal User Authentication (MMUA)*, pp. 99-106, 2003.

[13] J. Fierrez-Aguilar, "Adapted Fusion Schemes for Multimodal Biometric Authentication ", PhD Thesis: University of Madrid 2006.

[14] A. Ross and A. Jain, "Multimodal biometrics: an overview," *Proceedings of the 12th European Signal Processing Conference (EUSIPCO), (Vienna, Austria)*, pp. 1221-1224, 2004.

[15] J. Fierrez-Aguilar, J. Ortega-Garcia, and J. Gonzalez-Rodriguez, "Fusion Strategies in Multimodal Biometric Verification," *Proceedings of the IEEE International Conference on Multimedia and Expo, ICME '03*, pp. 5 - 8, 2003.

[16] F. Roli, J. Kittler, G. Fumera, and D. Muntoni, "An Experimental Comparison of Classifier Fusion Rules for Multimodal Personal Identity Verification Systems," *Proceedings of Multiple Classifier Systems, Sringer-Verlag, LNCS 2364*, pp. 325-336, 2002.

[17] P. Verlinde and M. Acheroy, "A contribution to multi-modal identity verification using decision fusion," *Royal Military Academy, Signal and Image Centre*, 2000.

[18] R. C. Luo and M. G. Kay, "Introduction," *Multisensor Integration and Fusion for Intelligent Machines and systems*, pp. 1-26, 1995.

[19] P. K. Varshney, "Distributed Detection and Data Fusion," *Springer-Verlag, New York*, 1997.

[20] D. Genoud, F. Bimbot, G. Gravier, and G. Chollet, "Combining methods to improve the phone based speaker verification " *Proceedings of the 4th International Conference on Spoken Language Processing, philadelphia*, vol. 3, pp. 1756-1759, 1996.

[21] S. S. Lyengar, L. Prasad, and H. Min, "Advances in Distributed Sensor Technology," *Prentice Hall PTR, New Jersey*, 1995.

[22] V. Radova and J. Psutka, "An approach to speaker identification using multiple classifiers," *Proceedings of the IEEE Conference on Acoustics, Speech and Signal Processing*, vol. 2, pp. 1135-1138, 1997.

[23] R. Snelick, U. Uludag, A. Mink, M. Indovina, and A. K. Jain, "Large Scale Evaluation of Multimodal Biometric Authentication Using State-of-the-Art Systems," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, pp. 450-455, 2005.

[24] R. Snelick, M. Indovina, J. Yen, and A. Mink, " Multimodal Biometrics: Issues in Design and Testing " *Proceedings of the Fifth International Conference on Multimodal Interfaces*, pp. 68-72, 2003.

[25] R. Brunelli and D. Falavigna, "Person identification using Multiple Cues," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, pp. 955-966, 1995.

[26] J. Kittler, M. Hatef, R. P. W. Duin, and J. Mates, "On Combining Classifiers," *IEEE*, vol. 20, pp. 226-239,1998.

[27] N. Poh and J. J. Korczak, "Hybrid Biometric Person Authentication Using Face and Voice Features," *Proceedings of the Third International Conference on Audio- and Video-Based Biometric Person Authentication*, pp. 348-353, 2001

[28] C. Vielhauer, S. Schimke, V. Thanassis, and Y. Stylianou, " Fusion Strategies for Speech and Handwriting Modalities in HCI," *SPIEEI 2005 – Conference on Multimedia on Mobile Devices*, pp. 63-71, 2005.

[29] R. Frischholz and U. Dieckmann, "BioID: A Multimodal Biometric Identification System," *IEEE*, vol. 33, pp. 64-68, 2000.

[30] G. Shakhnarovich and T. Darrell, "On probabilistic combination of face and gait cues for identification," *Proceedings of the 5th IEEE International Conference on Automatic Face and Gesture Recognition*, 2002.

[31] L. Hong and A. Jain, "Integrating faces and fingerprints for personal identification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, pp. 1295-1307, 1998.

[32] A. Kale, A. Roy-Chowdhury, and R. Chellappa, "Fusion of gait and face for human identification," *International Conference on Acoustics, Speech and Signal Processing*, 2004.

[33] A. Kale, A. K. R. Chowdhury, and R. Chellappa, "Towards a view invariant gait recognition algorithm," *Proceedings of IEEE Conference on Advanced Video and Signal Based Surveillance*, pp. 143–150, 2003.

[34] S. Zhou and R. Chellappa, "Probabilistic human recognition from video," *Proceedings of ECCV*, 2002.

[35] F. Roli and G. Fumera, "Analysis of linear and order statistics combiners for fusion of imbalanced classifiers. ," *3rd International Workshop on Multiple Classifier Systems (MCS 2002)*, pp. 252-261, 2002.

[36] K. Tumer and J. Ghosh, "Linear and Order Statistics Combiners for Pattern Classification," *Combining Artificial Neural Nets. Springer (1999)*, pp. 127-161, 1999.

[37] L. Alexandre, A. Campihlo, and M. Kamel, "On combining classifiers using sum and product rules," *Pattern Recognition Letters*, vol. 22, pp. 1283-1289, 2001.

[38] C. Bishop, Neural Networks for Pattern Recognition. New York, 1996.

[39] S. Bengio, J. Mariethoz, and S. Marcel, " Evaluation of biometric technology on XM2VTS " *Technical Report IDIAP-RR 01-21, Dalle Molle Institute for Perceptual Artificial Intelligence*, 2001.

[40] T. M. Mitchell, Machine learning: Mc Graw-Hill, 1997.

[41] P. Verlinde, P. Druyts, G. Chollet, and M. Acheroy, "Applying Bayes based classifiers for decision fusion in a multi-modal identity verification system," *International Symposium*, 1999.

[42] S. Ribaric, D. Ribaric, and N. Pavesic, "A Multimodal Biometric User-identification System for Network-based Applications," *IEE Proceedings on Vision, Image and Signal Processing*, vol. 150, pp. 409-416, 2003.

[43] A. K. Jain and A. Ross, "Learning User-specific Parameters in a Multibiometric System," *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, vol. 1, pp. 57-60, 2002.

[44] T. Hazen, E. Weinstein, and A. Park, "Towards robust person recognition on handheld devices using face and speaker identification technologies," *Proceedings of the International Conference on Multimodal Interfaces*, pp. 289-292, 2003.

[45] C. Sanderson and K. K. Paliwal, "Multi-Modal Person Verification System Based on Face Profiles and Speech," *Fifth International Symposium on Signal Processing and its Applications*, vol. 2, pp. 947-950, 1999.

[46] S. Marcel, J. Mariéthoz, Y. Rodriguez, and F. Cardinaux, "Bi-modal face and speech authentication: a biologin demonstration system," *Workshop on Multi-Modal User Authentication (MMUA)*, IDIAP-RR 06-18, 2006.

[47] J. Czyz, S. Bengio, C. Marcel, and L. Vandendorpe, "Scalability analysis of audio-visual person identity verification," *Audio- and Video-based Biometric Person Authentication*, pp. 752–760, 2003.

[48] J. Luettin and S. Ben-Yacoub, "Robust Person Verification based on Speech and Facial Images," *Proceedings of the European Conference on Speech Communication and Technology*, pp. 991-994, 1999.

[49] J. Bigun, B. Duc, F. Smeraldi, S. Fischer, and A. Makarov, "Multimodal Person Authentication," *Advanced Study on Face Recognition*, pp. 26-50, 1997.

[50] J. Bigun and J. M. H. du Buf, "N-folded symmetries by complex moments in Gabor space," *Proceedings of the IEEE on Pattern Analysis and Machine Intelligence (IEEE-PAMI)*, vol. 16, pp. 80-87, 1994.

[51] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 29, pp. 254-272, 1981.

[52] S. Ben-Yacoub, J. Luttin, K. Jonson, J. Matas, and J. Kittler, "Audio-visual person verification," *Proceedings of the IEEE on Computer Vision and Pattern Recognition (CVPR)*, pp.580-585,1999.

[53] C. Sanderson and K. K. Paliwal, "Adaptive Multi-Modal Person Verification System," *Proceedings of the First IEEE Pacific-Rim Conference on Multimedia*, 2000.

[54] J. Kittler, Y. Li, J. Matas, and M. U. Sanchez, "Combining Evidence in Multimodal Personal Identity Recognition Systems," *Proceedings of the First International Conference on AVBPA*, pp. 327-334, 1997.

[55] F. Roli, S. Raudys, and G. L. Marcialis, "An experimental comparison of fixed and trained fusion rules for crisp classifier outputs," *3rd International Workshop on Multiple Classifier Systems (MCS 2002)*, pp. 232-241, 2002.

[56] J. Fierrez-Aguilar, J. Ortega-Garcia, and J. Gonzalez-Rodriguez, "A Comparative Evaluation of Fusion Strategies for Multimodal Biometric Verification,"

Proceedings of the International Conference on Audio- and Video-Based Person Authentication, (AVBPA'03), pp. 830-837, 2003.

[57] M. C. Cheung, M. W. Mak, and S. Y. Kung, "A Two-level Fusion Approach to Multimodal Biometric Verification," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP'05),*, vol. 5, pp. 486-489, 2005.

[58] S. Garcia-Salicetti, M. A. Mellakh, L. Allano, and B. Dorizzi, "Multimodal Biometric Score Fusion: the Mean Rule vs. Support Vector Classifiers," *Proceedings of the European Signal Processing Conference (EUSIPCO'05)*, 2005.

[59] M. G. K. Ong, T. Connie, A. T. B. Jin, and D. N. C. Ling, "A Single-sensor Hand Geometry and Palmprint Verification System," *Proceedings of the 2003 ACM SIGMM Workshop on Biometrics: Methods and Applications*, pp. 100-106, 2003.

[60] Y. Wang, Y. Wang, and T. Tan, "Combining Fingerprint and Voiceprint Biometrics for Identity Verification: an Experimental Comparison," *Proceedings of the ICBA*, pp. 663-670, 2004.

[61] A. Lumini and L. Nanni, "When Fingerprints Are Combined with Iris - A Case Study: FVC2004 and CASIA," *International Journal of Network Security*, vol. 3, pp. 317-324, 2006.

[62] J. Ortega-Garcia, J. Gonzalez-Rodriguez, D. Simon-Zorita, and S. Cruz-Llanas, "From biometrics technology to applications regarding face, voice, signature and fingerprint recognition systems," *Biometric solutions for authentication in a e-world*, pp. 289-337, 2002.

[63] P. Verlinde and G. Chollet, "Comparing decision fusion paradigms using k-NN based classifiers, decision trees and logistic regression in a multi-modal identity verification application," *Proceedings of the International Conference on Audio and Video-Based Biometric Person Authentication*, pp. 188-193, 1999.

[64] P. Verlinde, G. Chollet, and M. Acheroy, "Multi-modal identity verification using expert fusion," *Information Fusion 1*, pp. 17-33, 2000.

[65] P. Verlinde, P. Druyts, G. Chollet, and M. Acheroy, "A multi-level data fusion approach for gradually upgrading the performances of identity verification

systems," *Sensor Fusion: Architectures, Algorithms and Application III*, vol. 3719, pp. 14-25, 1999.

[66] Y. Chen, S. C. Dass, and A. K. Jain, "Fingerprint Quality Indices for Predicting Authentication Performance," *Proceedings of the Fifth International Conference on Audio and Video-Based Person Authentication,(AVBPA'05)*, pp. 160–170, 2005.

[67] Y. Chen, S. C. Dass, and A. K. Jain, " Localized Iris Image Quality Using 2-D Wavelets," *Proceedings of the International Conference on Biometrics (ICB)*, pp. 373-381, 2006.

[68] N. Poh and S. Bengio, "Improving Fusion with Margin-Derived Confidence in Biometric Authentication Tasks," *Proceedings of the fifth International Conference on Audio and Video-Based Biometric Person Authentication, (AVBPA'05)*, pp. 474–483, 2005.

[69] S. Bengio, C. Marcel, S. Marcel, and J. Mariethoz, " Confidence Measures for Multimodal Identity Verification," *IDIAP Research Report No. IDIAP-PR 01-38*, 2002.

[70] T. K. Ho, J. Hull, and S. N. Srihari, "Decision Combination in Multiple Classifier Systems," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 16, pp. 66–75, 1994.

[71] H. Altinay and M. Demirekler, "Why Does Output Normalisation Create Problems in Multiple Classifier Systems?," *Proceedings of the 16-th International Conference on Pattern Recognition (ICPR)*, pp. 20775–20778, 2002.

[72] N. Poh and S. Bengio, "A Study of the Effects of Score Normalisation Prior to Fusion in Biometric Authentication Tasks," *IDIAP Research Report No. IDIAP-RR 04-69*, 2004.

[73] k. Nandakumar, "Integration of Multiple Cues in Biometric Systems," M.S. Thesis: Michigan State University, 2005.

[74] G. R. Doddington, M. A. Przybycki, A. F. Matin, and D. A. Reynolds, "The NIST speaker recognition evaluation - Overview, methodology, systems, results, perspective," *Speech Communication*, vol. 31, pp. 225-254, 2000.

[75] Y. Ma, B. Cukic, and H. Singh, "A Classification Approach to Multi-biometric Score Fusion," *Proceedings of the International Conference on Audio and Video-Based Biometric Person Authentication, (AVBPA)*, pp. 484-493, 2005.

[76] Y. Wang, T. Tan, and A. K. Jain, "Combining Face and Iris Biometrics for Identity Verification," *Proceedings of the Fourth International Conference on Audio and Video-Based Biometric Person Authentication, (AVBPA)*, pp. 805-813, 2003.

[77] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, vol. 7, pp. 179-188, 1936.

[78] S. Ben-Yacoub, Y. Abdeljaoued, and E. Mayoraz, "Fusion of face and speech data for person identity verification," *IEEE Transactions on Neural Networks*, vol. 10, pp. 1065-1074, 1999.

[79] S. Balakrishnama and A. Ganapathiraju, "Linear Discriminant Analysis- A Brief Tutorial," *Mississippi Sate University*, 1998.

[80] B. Flury, Common Principle Components and Related Multivariate Models. USA: John Wiley and Sons, 1988.

[81] B. D. Ripley, Pattern Recognition and Neural Networks. U.K: Cambridge University 1996.

[82] D. W. Hosner and S. Lemeshow, Applied logistic regression: John Wiley & Sons, 1989.

[83] Y. So, "A Tutorial on Logistic Regression," *SAS Institute Inc*, 1995.

[84] A. ASANO, "Support vector machine and kernel method," *Pattern information processing*, 2004.

[85] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, pp. 955-974, 1998.

[86] V. N. Vapnik, Statistical Learning Theory: Wiley Interscience, 1997.

[87] B. Gutschoven and P. Verlinde, " Multi-modal identity verification using support vector machines (SVM)," *Proceedings of the International Conference on Information Fusion*, pp. 3-8, 2000.

[88] V. N. Vapnik, The nature of statistical learning theory: Springer, 1995.

[89] S. Gunn, "Support vector machines for classification and regression," University of Southampton ISIS, 1998.

[90] N. Poh and S. Bengio, "Database, Protocol and Tools for Evaluating Score-Level Fusion Algorithms in Biometric Authentication," *Proceedings of the Fifth International Conference on Audio- and Video-Based Biometric Person Authentication, (AVBPA)*, 2005.

[91] N. Poh and S. Bengio, "Biometric Authentication Fusion Benchmark Database," http://www.idiap.ch/~norman/fusion.

[92] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score Normalisation for Text-independent Speaker Verification Systems," *Digital Signal Processing*, vol. 10, pp. 42-54, 2000.

[93] A. Ariyaeeinia and P. Sivakumaran, "Analysis and comparison of score normalisation methods for text-dependent speaker verification," *Proceedings of the Eurospeech'97*, vol. 3, pp. 1379-1382, 1997.

[94] J. Fierrez-Aguilar, J. Ortega-Garcia, J. Gonzalez-Rodriguez, and J. Bigun, "Discriminative Multimodal Biometric Authentication based on Quality Measures," *Pattern Recognition,* vol. 38, pp. 777–779, 2005.

[95] J. Bigun, J. Fierrez-Aguilar, J. Ortega-Garcia, and J. Gonzalez-Rodriguez, "Multimodal Biometric Authentication using Quality Signals in Mobile Communications," *Proceedings of the International Conference on Image Analysis and Processing, ICIAP 2003*, pp. 2-11, 2003.

[96] J. Fierrez-Aguilar, J. Ortega-Garcia, J. Gonzalez-Rodriguez, and J. Bigun, "Kernel-based multimodal biometric verification using quality signals," *Proceedings of The International Society for Optical Engineering ,SPIE 5404 (2004)*, pp. 544-554, 2004.

[97] K. Nandakumar, Y. Chen, A. K. Jain, and S. C. Dass, "Quality-based Score Level Fusion in Multibiometric Systems," *Proceedings of the 18th International Conference on Pattern Recognition(ICPR'06)*, pp.473-476, 2006.

[98] F. Alsaade, A. Ariyaeeinia, L. Meng, and A. Malegaonkar, "Multimodal Authentication using Qualitative Support Vector Machines," *Interspeech 2006*, pp. 2454-2457, 2006.

[99] A. Ariyaeeinia, J. Fortuna, P. Sivakumaran, and A. Malegaonkar, "Verification Effectiveness in Open-Set Speaker Identification," *Proceedings of the IEE on Vision, Image and Signal Processing,* vol. 153, pp. 618-624, 2006.

[100] D. Reynolds, "Comparison of background normalisation methods for text-independent speaker verification," *Proceedings of the Eurospeech'97*, pp. 963-966, 1997.

[101] J. Fortuna, P. Sivakumaran, A. Ariyaeeinia, and A. Malegaonkar, "Relative effectiveness of score normalization methods in open-set speaker identification," *Proceedings of the Speaker and language Recognition Workshop (Odyssey 2004)*, pp. 369-376, 2004.

[102] A. M. Ariyaeeinia, P. Sivakumaran, M. Pawlewski, and M. J. Loomes, "Dynamic weighting of the distortion sequence in text-dependent speaker verification," *Proceedings of the Eurospeech' 99*, pp. 967-970, 1999.

[103] S. Zafeiriou, A. Tefas, I. Buciu, and I. Pitas, " "Exploiting Discriminant Information in Non-negative Matrix Factorization with Application to Frontal Face Verification"," *IEEE Transactions on Neural Networks*, vol. 17, pp. 683-695, 2006.

[104] S. Bengio, F. Bimbot, J. Mariethoz, V. Popovici, F. Poree, E. Bailly-Bailliere, G. Mate, and B. Ruiz, "Experimental protocol on the BANCA database," *Technical Report IDIAP-RR 02-05, IDIAP*, 2002.

[105] F. Alsaade, A. Malegaonkar, and A. Ariyaeeinia, "Fusion of Cross Stream Information in Speaker Verification," *Proceedings of the COST 275 Workshop on Biometrics on the Internet*. pp. 63-66, 2005.

[106] F. Alsaade, A. M. Ariyaeeinia, A. S. Malegaonkar, M. Pawlewski, and S. G. Pillay, "Enhancement of Multimodal Biometric Segregation Using Unconstrained Cohort Normalisation," *Pattern Recognition, special issue on multimodal biometrics,* vol. 41, pp. 814-820, 2007.

APPENDIX A. Publications

[1] F. Alsaade, A. Malegaonkar, and A. Ariyaeeinia, "Fusion of Cross Stream Information in Speaker Verification," *Proceedings of the COST 275 Workshop on Biometrics on the Internet*. pp. 63-66, 2005.

[2] F. Alsaade, A. Ariyaeeinia, L. Meng, and A. Malegaonkar, "Multimodal Authentication using Qualitative Support Vector Machines, *Proceedings of Interspeech'06*, pp. 2454-2457, 2006.

[3] F. Alsaade, A. M. Ariyaeeinia, A. S. Malegaonkar, M. Pawlewski, and S. G. Pillay, "Enhancement of Multimodal Biometric Segregation Using Unconstrained Cohort Normalisation," *Pattern Recognition, special issue on multimodal biometrics,* vol. 41, pp. 814-820, 2007.