One-class Classification: An Approach to Handle Class Imbalance in Multimodal Biometric Authentication

Quang Duc Tran

A thesis submitted to City University London for the degree of
Doctor of Philosophy in Information Engineering

City University London
School of Mathematics, Computer Science & Engineering
Department of Electrical and Electronic Engineering

October, 2014

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I am indebted to Professor Panos Liatsis for his excellent supervision, advice, encouragement and patience throughout my PhD, without which this work would have never been accomplished. He has taught me about science, life and friendship and given me the feeling that I can do something useful. With his motivating enthusiasm, he is always available when I have something in my mind. He is definitely someone I found it a pleasure to work with. I would like to thank all the members in the Information Engineering and Medical Imaging Group for their support, discussion and advice.

To my family, words cannot describe how much your unconditional support means to me. And most of all for my loving, encouraging and patient wife, whose support during this PhD is greatly appreciated. Thanks for everything!

DECLARATION

No portion of the work contained in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning. Except where otherwise cited, this thesis is my own work. I hereby grant powers of discretion to the University Librarian to allow this thesis to be copied in whole or in part without further reference to the author. This permission covers only single copies made for study purposes, subject to normal conditions of acknowledgement.

ABSTRACT

Biometric verification is the process of authenticating a person's identity using his/her physiological and behavioural characteristics. It is well-known that multimodal biometric systems can further improve the authentication accuracy by combining information from multiple biometric traits at various levels, namely sensor, feature, match score and decision levels. Fusion at match score level is generally preferred due to the trade-off between information availability and fusion complexity. However, combining match scores poses a number of challenges, when treated as a two-class classification problem due to the highly imbalanced class distributions. Most conventional classifiers assume equally balanced classes. They do not work well when samples of one class vastly outnumber the samples of the other class. These challenges become even more significant, when the fusion is based on user-specific processing due to the limited availability of the genuine samples per user. This thesis aims at exploring the paradigm of one-class classification to advance the classification performance of imbalanced biometric data sets. The contributions of the research can be enumerated as follows.

Firstly, a thorough investigation of the various one-class classifiers, including Gaussian Mixture Model, k-Nearest Neighbour, K-means clustering and Support Vector Data Description, has been provided. These classifiers are applied in learning the user-specific and user-independent descriptions for the biometric decision inference. It is demonstrated that the one-class classifiers are particularly useful in handling the imbalanced learning problem in multimodal biometric authentication. User-specific approach is a better alternative with respect to user-independent counterpart because it is able to overcome the so-called within-class sub-concepts problem, which arises very often in multimodal biometric systems due to the existence of user variation.

Secondly, a novel adapted score fusion scheme that consists of one-class classifiers and is trained using both the genuine user and impostor samples has been proposed. This method also replaces user-independent by user-specific description to learn the characteristics of the impostor class, and thus, reducing the degree of imbalanced proportion of data for different classes. Extensive experiments are conducted on the BioSecure DS2 and XM2VTS databases to illustrate the potential of the proposed adapted score fusion scheme, which provides a relative improvement in terms of Equal Error Rate of 32% and 20% as compared to the standard sum of scores and likelihood ratio based score fusion, respectively.

Thirdly, a hybrid boosting algorithm, called r-ABOC has been developed, which is capable of exploiting the natural capabilities of both the well-known Real AdaBoost and one-class classification to further improve the system performance without causing overfitting. However, unlike the conventional Real AdaBoost, the individual classifiers in the proposed schema are trained on the

same data set, but with different parameter choices. This does not only generate a high diversity, which is vital to the success of r-ABOC, but also reduces the number of user-specified parameters. A comprehensive empirical study using the BioSecure DS2 and XM2VTS databases demonstrates that r-ABOC may achieve a performance gain in terms of Half Total Error Rate of up to 28% with respect to other state-of-the-art biometric score fusion techniques.

Finally, a Robust Imputation based on Group Method of Data Handling (RIBG) has been proposed to handle the missing data problem in the BioSecure DS2 database. RIBG is able to provide accurate predictions of incomplete score vectors. It is observed to achieve a better performance with respect to the state-of-the-art imputation techniques, including mean, median and k-NN imputations. An important feature of RIBG is that it does not require any parameter fine-tuning, and hence, is amendable to immediate applications.

ACRONYMS

A-GMM	**A**dapted score fusion using one-class **GMM** as the base classifier
A-Kmeans	**A**dapted score fusion using one-class **K-means** as the base classifier
A-kNN	**A**dapted score fusion using one-class **k-NN** as the base classifier
A-SVDD	**A**dapted score fusion using one-class **SVDD** as the base classifier
AUC	**A**rea **U**nder the ROC **C**urve
AUDF	**A**dapted **U**ser**-D**ependent **F**usion
DET	**D**etection **E**rror **T**rade-off
EER	**E**qual **E**rror **R**ate
EPC	**E**xpected **P**erformance **C**urve
GMDH	**G**roup **M**ethod of **D**ata **H**andling
GMM	**G**aussian **M**ixture **M**odel
HTER	**H**alf **T**otal **E**rror **R**ate
K-means	**K-means** clustering
k-NN	**k-N**earest **N**eighbour
FAR	**F**alse **A**cceptance **R**ate
FTE	**F**ailure **T**o **E**nrol
FRR	**F**alse **R**ejection **R**ate
LR	**L**ikelihood **R**atio based score fusion
MLP	**M**ulti-**L**ayer **P**erceptron
r-ABOC	Hybrid Boosting Algorithm, based on **R**eal **A**da**B**oost (r-AB) and **O**ne-class **C**lassification (OC)
RBF	**R**adial **B**asis **F**unction
RIBG	**R**obust **I**mputation **B**ased on **G**MDH
ROC	**R**eceiver **O**perating **C**haracteristics
SUM	**Sum** of Scores
SVDD	**S**upport **V**ector **D**ata **D**escription
SVM	**S**upport **V**ector **M**achines
UIFI	**U**ser-**I**ndependent score **F**usion using the **I**mpostor target class

UIFG	**U**ser-**I**ndependent score **F**usion using the **G**enuine user target class
USFI	**U**ser-**S**pecific score **F**usion using the **I**mpostor target class
WER	**W**eighted **E**rror **R**ate

NOTATIONS

$d(\mathbf{s}_T \vert w_k)$	Distance of match score vector $\mathbf{s}_T$ to the target class w_k
$E(x)$	Expected value of x
f_T	Fraction of target objects, which are rejected by a one-class classifiers (i.e., $f_T \in [0, \dots, 1]$)
G	Genuine user
$g_t(\mathbf{s}_T)$	Hypothesis function mapping the test input match score vector $\mathbf{s}_T$ into a real-valued prediction in classifier ensemble learning, where $t \in [1, \dots, \mathrm{T}]$)
$\Phi(\mathbf{s}_i)$	Mapping of vector $\mathbf{s}_i$ to a high dimensional feature space
$h_t(\mathbf{s}_T)$	$h_t(\mathbf{s}_T) = sign[g_t(\mathbf{s}_T)]$, which is able to generate hard class label for the test input match score vector $\mathbf{s}_T$
I	Impostor
$K(\mathbf{s}_i, \mathbf{s}_j)$	Kernel function operating on $\mathbf{s}_i$ and $\mathbf{s}_j$
J	Number of users
j	User index (i.e., $j \in \{1, \dots, J\}$)
M	Number of biometric matchers
m	Matcher index (i.e., $m \in \{1, \dots, M\}$)
μ, $\boldsymbol{\mu}$	Mean and mean vector
N	Number of training samples
n_T^m	Normalized match score for the test match score s_T^m of the m-th matcher
$p(w_k)$	The prior probability for class w_k to occur
$p(w_k \vert \mathbf{s}_T)$	The a posteriori probability that the true class is w_k, given $\mathbf{s}_T$
$p(\mathbf{s}_T \vert w_k)$	Conditional probability density for $\mathbf{s}_T$, given class w_k
s_i^m	Match score, provided by the m-th matcher
$\mathbf{s}_i$	Match score vector
$\mathbf{s}_T$	Test input match score vector
$s_{COM,k}(\mathbf{s}_T)$	Combined match score for match score vector $\mathbf{s}_T$, where $k \in \{G, I\}$

$s_{COM,j,k}(\mathbf{s}_T)$	Combined match score, related to user j for match score vector $\mathbf{s}_T$, where $k \in \{G, I\}$
SV	Set of support vectors (i.e., objects with $\alpha_i > 0$)
σ	Standard deviation
Σ	Covariance matrix
T	Number of individual classifiers in a classifier combination
τ	Decision threshold
$\mathbf{x}_m$	Output feature vector, presented to the m-th classifier
w_k	Class identifier, where $k \in \{G, I\}$
$w_{j,k}$	Class identifier, related to user j, where $k \in \{G, I\}$

PUBLICATIONS

The results of the research described in this thesis have been published in the following papers:

[1] Q. D. Tran, and P. Liatsis, "Optimal weight selection in matching score fusion based face recognition," In: *Int. Conf. Systems, Signals and Image Processing (IWSSIP)*, Sarajevo, Bosnia & Herzegovina, 16-18 June 2011.

[2] Q. D. Tran, P. Liatsis, B. Zhu, and C. He, "An approach for multimodal biometric fusion under the missing data scenario," In: *Int. Conf. Uncertainty Reasoning and Knowledge Engineering (URKE)*, Bali, Indonesia, 4-7 August 2011.

[3] Q. D. Tran, P. Liatsis, B. Zhu, and C. He, "Using Density based Score Fusion for Multimodal Identification Systems under the Missing Data Scenario," In: *4th Int. Conf. Developments in eSystems Engineering (DESE)*, Dubai, 6-8 December 2011.

[4] Q. D. Tran, P. Kantartzis, and P. Liatsis, "Improving Fusion with Optimal Weight Selection in Face Recognition," *Integrated Computer-Aided Engineering*, vol. 19, no. 3, pp. 229-237, 2012.

[5] Q. D. Tran, and P. Liatsis, "A Modified Equal Error Rate based User-Specific Normalization for Multimodal Biometrics," In: *5th Int. Conf. Developments in eSystems Engineering (DESE)*, Bucharest, Romania 5-7 September 2012.

[6] Q. D. Tran, and P. Liatsis, "User-specific Fusion using One-class Classification for Multimodal Biometric System: Boundary Methods," In: *6th Int. Conf. Developments in eSystems Engineering (DESE)*, Abu Dhabi, UAE 16-18 December 2013.

[7] Q. D. Tran, and P. Liatsis, "Improving Fusion with One-Class Classification and Boosting in Multimodal Biometric Authentication,"

Intelligent Computing in Bioinformatics, Springer International Publishing, pp. 438-444, 2014.

[8] Q. D. Tran, and P. Liatsis, “RABOC: An Approach to Handle Class Imbalance in Multimodal Biometric Authentication,” *Neurocomputing*, Accepted, 2014.

CHAPTER 1

INTRODUCTION

This chapter will start with a short introduction of the background associated to this research, including biometric systems, common biometric traits, the sources of biometric evidence; and the type of information to be fused (Sections 1.1-1.3). The problem of imbalanced class distribution that poses serious difficulties to most standard two-class classifiers, when applied in performing multimodal biometric fusion is then discussed in detail in Section 1.4. The contributions originating from this research and the outline of the thesis are given in Sections 1.5 and 1.6.

1.1. Biometric Systems in a Nutshell

Traditional methods of human identification based on credentials (PIN or identification documents) are not able to meet the growing demand for stringent security in applications such as access control, government benefits, border crossings, and national ID cards [1]. As an alternative, biometric recognition, or simply biometrics, has gained much attention over recent years particularly due to rapidly growing demand for person identification applications [1-5]. Biometric recognition refers to the process of verifying an identity claim using a person's behavioural and physiological characteristics, such as hand geometry, iris, face, hand vein, voice, signature, fingerprint, etc. Although biometric recognition has its limitations (e.g., temporal changes in biometric traits, additional cost), it has provided many advantages over existing credential-based mechanisms because it constitutes a strong link between a person and his identity and cannot be forgotten, stolen, lost, shared or forged [1]. In addition, biometric recognition is able to deter users from making false repudiation claims because it requires the users to be present at the time of authentication [5].

The use of biometrics is not new. In the late 19th century, Alphonse Bertillon, a French law enforcement officer, advocated a personal identification system, which entailed the precise measurement of a morphological description of the shape and appearance of the body and a listing of peculiar marks, such as scars, tattoos, moles on the surface of the body [1], [6]. The Bertillon system was short-lived. Soon after its introduction, the distinctiveness of human fingerprints was established. Since the early 1990s, fingerprints have been an accepted method in forensic investigations to identify criminals [1]. With growing concerns about security breaches, financial fraud, other biometric traits, including face, iris, palmprint, and voice have been used for person identification. Now, biometrics is a mature technology, which is widely used in a variety of applications [7-10].

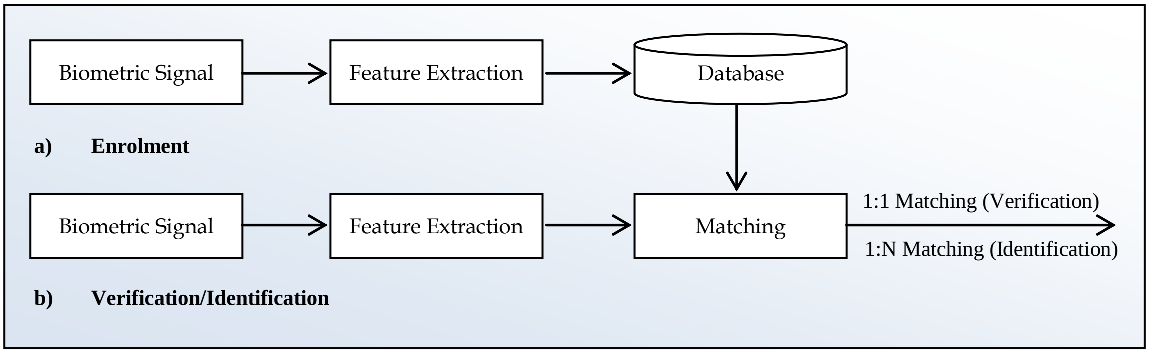


Figure 1.1. The general architecture of a biometric system.

The general architecture of a biometric system, sketched in Figure 1.1 can be divided into two categories: (1) authentication (also referred to as verification), and (2) identification. In authentication applications, the system validates a person's identity by conducting a one-to-one (1:1) matching to compare the captured biometric data with his/her previously enrolled biometric template, pre-stored in the system database. On the contrary, the goal of personal identification applications is to recognize an individual by searching the entire enrolled template database for a match. For these applications, the system conducts a one-to-many (1:N) matching to establish the identity of an individual. This thesis is focused on ***biometric authentication.***

Biometric authentication involves a trade-off between two types of errors: (1) *False Rejection Rate* (FRR), which is an empirical estimate of the probability of falsely rejecting a genuine user, and (2) *False Acceptance Rate* (FAR), which is an empirical estimate of the probability of falsely accepting an impostor as a true user. In general, both FRR and FAR are functions of the decision threshold. The system performance across the various thresholds can be depicted using a *Detection Error Trade-off* (DET) curve. *Equal Error Rate* (EER) is a single valued measure, which refers to an operating point on the DET curve where FRR equals FAR. More details about biometric performance evaluation methodologies will be given in Chapter 3.

1.2. Biometric Modalities

A number of biometric traits are being used in various applications. Behavioural biometric traits include voice, signature, gait, and keystroke. Physiological biometric traits include fingerprint, iris, retina, face, palmprint, and hand geometry (Figure 1.2). Each biometric has its strengths and weaknesses and the applicability of a biometric trait is typically dependent on the following fundamental premises [1], [5], [11-14]:

1) *Universality:* every person should possess the biometric trait

2) *Distinctiveness*: two persons should have different biometric traits.

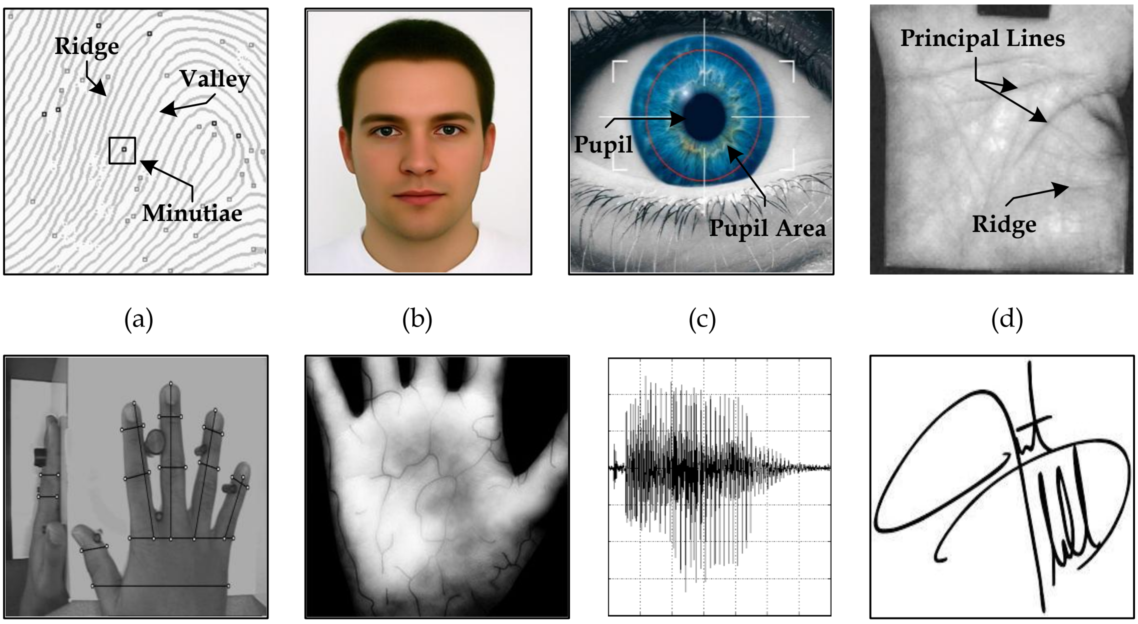
3) *Permanence*: the trait should be invariant over time.

4) *Performance*: this can be measured in terms authentication accuracy and speed (throughput), and resource requirements for biometric system implementation.

5) *Acceptability*: extent to which users are willing to use a particular biometric trait in their daily lives.

6) *Circumvention*: difficulty to fool a biometric system by fraudulent methods.

For the sake of completion, some characteristics of typical biometric traits are summarized in the following paragraphs.



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 1.2. Examples of common biometric traits: (a) Fingerprint, (b) Face, (c): Iris, (d) Palmprint, (e) Hand geometry (f), Hand veins, (g) Voice, and (h) Signature. Adapted from [1] and [21].

Fingerprint: Fingerprint recognition [16] has been used for personal identification for many decades. A fingerprint is the pattern of ridges and valleys on the surface of a fingertip (see Figure 1.2(a)). Ridges are characterised by several landmark points, known as minutiae, whose spatial distribution is claimed to be unique to each finger [5]. Virtually all law enforcement agencies use the Automatic Fingerprint Identification System (AFIS) [1]. The emergence of low cost and compact fingerprint readers has made fingerprint the preferred choice in many commercial applications.

Face: Face recognition [17] has several advantages over other biometric traits. It is natural, non-intrusive and easy to use. Owing to advances in face modelling and analysis techniques, a significant progress has been seen in recent years in face recognition, with many systems capable of achieving recognition rates greater than 90% [17], [18]. However, real-world scenarios remain a challenge, especially for unconstrained tasks, where illumination, viewpoint, expression, occlusion, and accessories vary considerably [18]. The most popular approaches to face recognition are based on either (1) the global analysis of the face image or (2) the location and shape of face attributes, such as the eyes, nose, lips and chin [5]. A survey of existing face recognition challenges and technologies is available in [18].

Iris: The iris is the coloured portion of an individual's eye. An iris image is typically captured under infrared illumination. The complex iris texture, illustrated in Figure 1.2(c), carries very distinctive information, including pits, stripes, and furrows, which allow for highly reliable personal identification [1], [5], [19]. First invented by Daugman [20], iris recognition is extremely accurate and fast on high-resolution well-captured iris images. However, the relatively large failure to enrol (FTE) rate, reported in several studies, and relatively high sensor costs may reduce its applicability to some large-scale government applications [5].

Palmprint: Palmprint recognition [21], just like fingerprint recognition, is based on the aggregate of information, presented in a friction ridge impression, which include ridge flow, ridge characteristics, and ridge structure (see Figure 1.2(d)). As compared to fingerprint, the palm provides a much larger surface area and, as a result, more distinctive information can be extracted [21]. A variety of sensor types can be used for collecting the image of a palm; however, traditional live-scan methodologies have been slow to adapt to the larger capture areas [1]. Palmprint recognition systems have not been deployed for civilian applications (e.g., access control) [5].

Hand geometry: Hand geometry recognition is based on a number of measurements of the human hand, including its size, shape, and the width and length of the fingers [16] (see Figure 1.2(e)). The systems are widely implemented for their ease to use, and integration capabilities. However, the geometry of the hand is not very distinctive, limiting the applications of the hand geometry system to verification tasks [5].

Hand veins: Hand vein recognition works by identifying the pattern of the blood vessels hidden underneath the skin (Figure 1.2(f)), which can be quite distinctive even among identical twins and stable over long periods of time [5],

[16]. Veins could be detected in the finger, palm, and backhand using inexpensive infra-red light emitting diodes (LEDs). Biometric authentication based on hand veins is now available for some commercial applications.

Voice: Voice is a combination of behavioural and physiological characteristics. Voice is not very distinctive since the behavioural characteristics can change over time due to medical conditions (e.g., common cold) and age [1]. Voice recognition is highly suitable for person identification over the telephone, but is quite sensitive to playback spoofing and background noise [5].

Signature: Signature (i.e., the way a person signs their name) is a behavioural characteristic, which has been acceptable in daily business transactions as a method of verification for a long time [16]. Nevertheless, signature is influenced by physical and emotional conditions of the signatories and may change over time. Very few automatic signature recognition systems have been deployed [5].

Table 1.1. Comparison of different biometric traits. H, M and L denote High, Medium and Low, respectively. Adapted from [11].

Biometric Trait	**Universality**	**Distinctiveness**	**Permanence**	**Performance**	**Acceptability**	**Circumvention**
Fingerprint	M	H	H	H	M	M
Face	H	L	M	L	H	H
Iris	H	H	H	H	L	L
Palmprint	M	H	H	H	M	M
Hand geometry	M	M	M	M	M	M
Hand veins	M	M	M	M	M	L
Voice	M	L	L	L	H	H
Signature	L	L	L	L	H	H

The biometric traits, described above are compared in Table 1.1. It has been observed that no single biometric trait is able to meet all criteria. Some of the traits, such as fingerprint, face and iris have sufficient discriminating power to be applicable in large-scale identification applications, while others like hand geometry, are more appropriate for verification applications [1], [5]. It is also important to note that biometric traits are in varying stages of maturity [14].

1.3. Multibiometrics and Multimodal Biometrics

Authentication systems built upon a single biometric trait (also referred as *unimodal biometric systems*) have to contend with a variety of problems [1] such as:

a) *Non-universality*: It may not be possible to extract meaningful biometric data from a subset of users. For example, about 4% of the population have poor quality fingerprints, which are difficult to image with existing fingerprint sensors [11];

b) *Noise in the sensed data*: Noisy data, which results from different ambient conditions or defective sensors, may lead to a genuine user being incorrectly rejected;

c) *Intra-class variations*: Intra-class variations are typically caused by a change in the biometric characteristics of a person over time. Intra-class variations become more prominent in behavioural traits (e.g., voice, signature), as compared to physiological traits (e.g., face, fingerprint);

d) *Inter-class similarities*: Inter-class similarities, which refer to the overlap in feature space of the data describing individual users, will increase the false acceptance rate of a large-scale identification system;

e) *Spoof attacks*: Biometric systems are vulnerable to spoof attacks because some biometric data can be imitated, as highlighted in Section 1.2. The possibility of generating digital artifacts of biometric data in order to circumvent a biometric system has also been demonstrated in [1], [22].

Multibiometric systems seek to alleviate some of the above problems by reconciling multiple, (fairly) independent sources of evidence. They are known to be more robust against individual sensor failure, address the problem of non-universality, and deter spoof attacks as it is difficult for an impostor to spoof multiple biometric traits of a legitimate user [11]. Multibiometric systems are able to meet the stringent performance requirement of various applications [21].

Multiple sources of biometric information can be obtained by using multiple sensors, processing algorithms, instances, samples, and modalities [1], [11-14], [23]. These possibilities are briefly described in the sub-sections below:

a) *Sensors*: A single biometric trait is imaged using multiple sensors, which are capable of extracting diverse information from registered images. As an example, multiple cameras can be deployed to record the two dimensional texture content and the three dimensional surface shape of a person's face.

b) *Processing algorithms*: The same biometric data is processed using different algorithms. For example, a minutiae-based algorithm and a texture-based algorithm can be used simultaneously to obtain complementary feature sets from a fingerprint image.

c) *Instances*: Multiple instances, such as the left and right finger, or the left and right irises can be used to verify a claimed identity. In this context, it is not necessary for the system to use a new sensor or algorithm to process the data and hence, this is a cost effective approach.

d) *Samples*: Multiple samples of the same biometric trait are captured using a single sensor to account for the variations, which may occur in the trait. For example, the frontal face profile together with the left and right profiles are acquired in a face recognition system to account for facial pose variations.

e) *Modalities:* These are also known as *multimodal biometric systems*, and combine the evidence of different body traits for personal recognition. Some of the earliest multimodal biometric systems integrate voice and face features for establishing identity. The deployment of these systems is influenced by several factors, such as cost of deployment, throughput time, expected error rate, enrolment time, etc.

Various fusion scenarios can be performed using the aforementioned sources of evidence. A generic biometric system consists of four modules, i.e., (1) sensor module, which acquires the raw biometric data; (2) feature extraction module, which processes the biometric data to extract a set of compact discriminative features; (3) matching module, which compares the input feature vector with that of the stored template, thus, resulting to match scores (i.e., measures of similarity between the input and template biometric features); and (4) decision module, which outputs the authentication decision with regards to the match scores [1], [13], [14]. The combination of multiple pieces of evidence can be performed at any of these four modules, as illustrated in Figure 1.3, and described below:

(a) *Sensor level fusion* refers to combining the raw data of the same biometric trait, obtained using either a single sensor or various compatible sensors. One example of this type of fusion is the mosaicking of multiple samples of a fingerprint (Figure 1.3(a)).

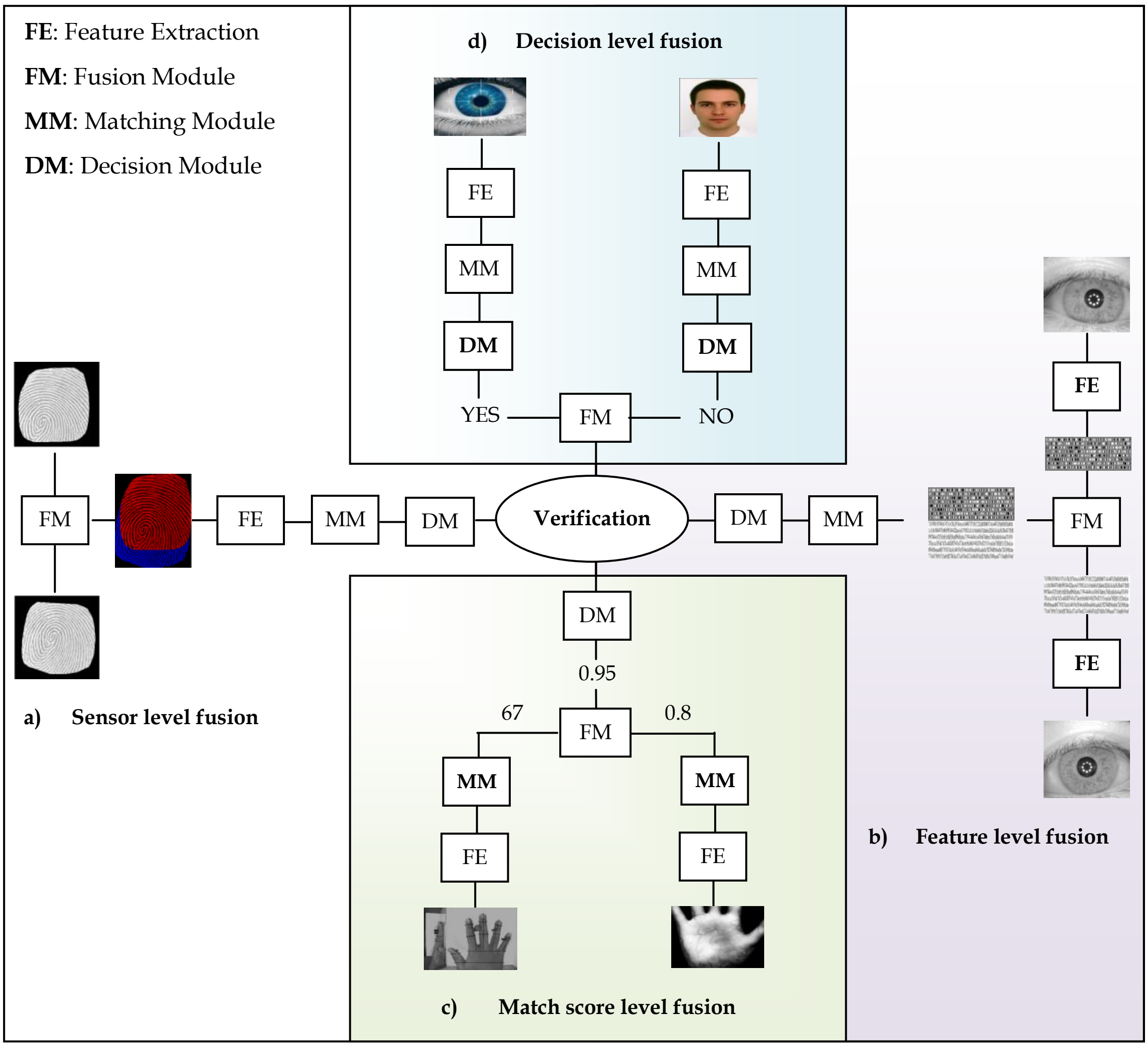


Figure 1.3. Fusion scenarios in multibiometric systems. Adapted from [11], [12] and [14].

(b) *Feature level fusion* refers to the combination of different feature sets, extracted from multiple biometric sources, as shown in Figure 1.3(b) for the example of left and right iris modalities.

(c) *Match score level fusion* refers to the combination of match scores provided by different classifiers pertaining to different biometric modalities (see Figure 1.3(c)). More information regarding this type of fusion will be given in Chapter 2.

(d) *Decision level fusion* refers to combining decisions, made independently by various biometric systems. Methods for decision level fusion include behaviour knowledge space [1], majority voting [59] and Dempster-Shafer theory of evidence based weighted voting [1].

Sensor and feature level fusion are generally expected to be more effective than fusion in the other two levels since they contain richer information about the biometric data. However, fusion in these levels is difficult to achieve in practice due to the following reasons: (1) most commercial biometric systems do not provide access to the raw data (nor the feature sets) for security reasons, (2) the fusion process has to deal with the presence of noise, which becomes prominent in the raw data; and (3) the feature sets of different modalities may be non-homogenous or incompatible [1]. On the other hand, decision level fusion is considered to be the least informative [23]. Thus, match score level fusion is usually preferred, as it is relatively easy to access and combine the match scores generated by the various matchers [11].

1.4. The Class Imbalance Problem

As mentioned above, multimodal biometric fusion at the match score level is a promising research direction. However, combining match scores is a challenging task because the match scores of different matchers may have different probability distributions and can be inhomogeneous [1], [90].

A common practice in many reported works on multimodal biometrics is to view fusion at match score level as a two-class classification problem, where the vector of match scores is treated as a feature vector, and thus, can be classified into one of two classes, i.e., genuine user and impostor. Based on the training set of match scores, the classifier learns the decision boundary between the two classes [1]. The decision boundary can be learned at the training stage, regardless of the claimed identity [24-26] or for each user, enrolled in the system [29-33]. A number of two-class classifiers, such as HyperBF [24], k-Nearest Neighbour using vector quantization [25], C4.5 decision tree, Fisher linear discriminant, Bayesian classifier [26], [33], Multilayer Perceptron [26], Support Vector Machines (SVM) [26], [32], linear classifiers [29], and discriminative classifiers based on reduced polynomial expansion [30] have been used to render the decision in a multimodal biometric verification system. Although these classifiers are capable of learning the decision boundary irrespective of how the feature vectors are generated, their limitation is that it is not always possible to fix one type of error, e.g., False Acceptance Rate (FAR) and then compute the False Rejection Rate at the specified FAR [1].

Table 1.2. Existing multimodal biometric databases and their class imbalance ratios.

Database	# Users	Modalities	# Imp. Samples	# Gen. Samples	Imbalance Ratio
BioSecure DS2 [61], [62]	333	Face. Fingerprint. Iris	78,624	156	524:1
XM2VTS LP1 [63]	295	Face, Speech	40,000	600	66:1
XM2VTS LP2 [63]	295	Face, Speech	40,000	400	100:1
NIST BSSR1 [52]	517	Fingerprint, Face	266,772	517	516:1
MCYT [53]	75	Fingerprint, Signature	750	525	10:7
MSU [107]	10	Face, Fingerprint, Hand-Geometry	12,250	500	49:2

Moreover, recent literature has indicated that the performance of most conventional two-class classifiers deteriorates, when applied to problems characterised by **class imbalance** [34-37]. Class imbalance is a common problem to many application domains, including multimodal biometric authentication. Table 1.2 summarizes several well-known multimodal biometric databases and their class imbalance ratio. It is not uncommon for the class imbalance to be on the order between 25:1 and 500:1, and in each case, the number of genuine user samples (# Gen. Samples) available for training is $O(J)$, but the number of impostor samples (# Imp. Samples) is $O(J^2)$, where J is the number of users (# Users) in the system [90].

Most conventional two-class classifiers assume or expect balanced class distributions, and generally create suboptimal classification models, when complex imbalanced data sets are presented [35]. Particularly, Bayesian networks are learned using certain scoring functions to approximate the dependency patterns, which dominate the data. Since the dependency patterns in the small class are usually not significant, they are hard to be encoded in the networks [115], [158]. In [115], [116], it was reported that the Back propagation (BP) and Radial Basis Function (RBF) neural networks may perform sub-optimally with imbalanced data sets, since the small class is inadequately characterised in the learned solution. In [119], the k-Nearest Neighbour was observed to give higher probabilities to samples from the prevalent class, and hence, the test cases from the small class are prone to being incorrectly classified [115]. Support Vector Machines (SVM) is generally believed to be more robust against the class imbalance than other classification learning algorithms, since the class boundary of SVM is characterised by a few support vectors and the skewed class distribution may not affect this class boundary too much. However, recent works [34], [38], [39] have indicated that SVM may be ineffective in predicting samples of the small class. This is due to the fact that SVM tries to minimize the total error, which is inherently biased towards the prevalent class. Furthermore, if there is lack of samples in the small class, there could be an imbalance of representative support vectors, which can also degrade the overall classification performance.

The most obvious characteristic of an imbalanced data set is the skewed data distribution between the two classes. However, it has been observed that the skewed data distribution is not the only negative impact that class imbalance has on the generalization performance of a classifier's learning algorithm [34], [37]. Other influential factors are as follows: (1) *The existence of within-class sub-concepts* [37], referring to the problem of *small disjuncts*, where a single class is composed of various sub-clusters, which increases the learning concept complexity of the data set, (2) *Separability*, referring to the difficulties in separating the small class from the prevalent class. It is known that if the samples among each class are overlapping, discriminative rules are not easy to induce; and (3) *Small sample size*, where the classifier often fails to generalize the inductive rules over the sample space, leading to overfitting.

Class imbalance has received limited attention in the biometric literature. Under-sampling seems to be the most common approach, which has been used to handle the problem [1], [32], [40]. Obvious shortcomings with the data sampling process are: (1) the optimal class distribution of a training data is usually unknown; (2) additional learning cost for processing data is inevitable in most cases; (3) ineffective data sampling methods may cause the classifier to miss important concepts pertaining to the impostor class [37]. In [30], 3% Gaussian noise with respect to the largest magnitude of the match scores was included into the genuine user class to increase the training sample size and reduce the class imbalance ratio. Nevertheless, this can potentially decrease the matching accuracy of combining non-Gaussian match scores. These factors indicate the need for additional research efforts to advance the classification of imbalanced biometric data.

1.5. Research Contributions of the Thesis

The difficulties originating from the class imbalance problem and its frequent occurrence in practical applications of machine learning have attracted increased research interest [37], [41-43]. Over the years, the machine learning community has addressed this problem in many different ways. Among others, the most perspective technique seems to be one-class classification [34], [44], which is naturally quite robust to the presence of class imbalance by using single class samples rather than differentiating between the samples of the two classes [44], [45]. In [46-48], the authors suggested that one-class classifiers are particularly useful in handling extremely imbalanced data sets with high feature space dimensionality, while two-class classifiers are more suitable for moderately imbalanced data sets [34].

The aim of this research is to investigate one-class classification methods in the context of class imbalanced data sets for systematically enhancing biometric authentication accuracy. The main contributions of the research are as follows:

(1) **A novel user-specific fusion paradigm based on one-class classifiers**: Recent literature has been suggested that biometric systems have varying degrees of accuracy in authenticating users [109]. Due to the existence of user variation, match scores from both the impostor and genuine user classes are scattered into several small regions and the decision boundary may enclose a large sparsely populated area, which could increase the probability of accepting outliers [171]. The proposed score fusion scheme is capable of exploiting user-specific information, and thus, resulting more reliable and compact scatters. Particularly, it trains a different description for each of the users in the biometric system by using his/her corresponding match score patterns. As compared to its conventional user-independent counterpart, this does not only improve verification performance, but also results to significant computational savings. Although a number of works have focused on designing a fusion classifier that differs for each user, no prior work has been conducted in the biometric literature on the use of one-class classifiers in designing user-specific descriptions.

(2) **A novel adapted score fusion scheme based on Bayes theorem**, applied in combining one-class classifiers of the same type to effectively and efficiently exploit the training data from both the genuine user and impostor classes. It also makes use of user-specific instead of user-independent score fusion to learn the characteristics of the impostor class, thus reducing the degree of

imbalance for different classes and counteracting the possible effects of the so-called within-class sub-concepts problem, which frequently occurs in biometric systems due to the existence of user variation.

(3) **A novel hybrid boosting algorithm**, called r-ABOC, which inherits the natural capabilities of both the so-called Real AdaBoost (r-AB) [58] and One-class Classification (OC) to address the issue of extremely imbalanced class distribution in biometric data sets. This algorithm works by developing a weak classifier, which consists of one-class Gaussian Mixture Model and considers the use of training data from the two classes. Real AdaBoost is then applied to combine multiple weak classifiers in order to improve their performance without causing overfitting. However, unlike the conventional Real AdaBoost, the weak classifiers in r-ABOC are learned on the same data set, but with different parameter choices (i.e., fraction rejection rates). It does not only generate the necessary diversity to enable the classifier combination to perform well, but also dramatically reduces the number of parameters to be chosen by the user. An important feature of r-ABOC is that its performance becomes unchanged when the number of rounds of boosting is increased. This suggests that r-ABOC may completely eliminate the possibility of making bad choice of the fraction rejection for the practical biometric authentication problem.

(4) **A comprehensive analysis of the performance of one-class classifiers in the context of multimodal biometric fusion is provided**: A large number of one-class classifiers have been developed in the literature. They differ in their ability to cope with different characteristics of the data [44]. There is no single best classifier for all problems, which confirms the "no panacea" in pattern recognition [59]. Several representative one-class methods have been exhaustively analyzed in this thesis, i.e., Gaussian Mixture Model (GMM), k-Nearest Neighbour (k-NN), K-means clustering (K-means), and Support Vector Data Description (SVDD). The aim of this analysis is to provide guidelines, which can inform the appropriate choice of classifiers to be used in biometric applications. This also highlights the advantages of

one-class classifiers over two-class methods in dealing with imbalanced class distributions of biometric data sets.

(5) **A robust imputation based on Group Method of Data Handling:** Most score level fusion techniques, including also the above proposals, are specifically designed for a complete match score vector and thus, cannot be invoked otherwise [40]. In order to handle incomplete match score vectors and their frequent occurrence in biometric systems, a robust imputation technique, called RIBG was developed. RIBG is based on Group Method of Data Handling, which is able to automatically find interrelations in the data and select the optimal structure of a model and hence, increasing the performance accuracy of existing techniques. RIBG offers many distinct advantages over other state-of-the-art imputation techniques. It is shown to be capable of providing more accurate predictions for the missing elements. It is resistant to noise and does not require any parameter fine-tuning.

The above contributions are supported through systematic empirical evaluation. Extensive experiments, carried out on the BioSecure DS2 [61], [62] and XM2VTS benchmark databases [63], were utilised to illustrate the effectiveness of the proposed approaches. As it can be seen from Table 1.2, XM2VTS and BioSecure DS2 are two publicly available databases, which were developed to benchmark the performance of biometric algorithms at match score level. They contain biometric traits from a large number of users and their class distribution is considered as to be extremely imbalanced. BioSecure DS2 also involves a large number of incomplete match score vectors, which requires to be processed prior to applying any classifier learning algorithms [61].

1.6. Outline of the Thesis

This thesis is organised as follows. Following the introduction, **Chapter 2** provides a literature survey on the different techniques, which have been widely used in multimodal biometric fusion at match score level. These techniques generally can be divided into two broad categories of user-independent and user-specific processing. **Chapter 3** is devoted to explaining the BioSecure DS2 and XM2VTS benchmark databases and evaluation methodologies used to evaluate the authentication performance in multimodal biometric authentication. It discusses the set of experiments, which are conducted in this thesis. The proposed RIBG algorithm (Contribution (5)) to handle the missing values in these databases is also presented in this chapter. **Chapter 4** introduces several representative one-class classifiers, i.e., GMM, K-means, k-NN, and SVDD. Illustrative examples and extensive evaluation are also presented to highlight the advantages of the user-specific fusion scheme over the user-independent counterpart (Contributions (1) and (4)). **Chapter 5** discusses in detail the proposed adapted score fusion scheme (Contribution (2)) and the improved hybrid boosting algorithm, r-ABOC (Contribution (3)). The pros and cons of these techniques are further assessed in this chapter. **Chapter 6** is dedicated to conclusions and future work.

CHAPTER 2

LITERATURE REVIEW ON MATCH SCORE LEVEL FUSION

When the match scores of individual biometric matchers are combined for the decision inference, fusion is said to be done at match score level. Match score level fusion is the most commonly used approach in multimodal biometric systems because the match scores are easily accessible and contain sufficient information to distinguish between a genuine user and impostor [1], [16], [67]. This chapter aims to describe various characteristics of match score level fusion and provide a comprehensive review on the techniques to perform the fusion at this level. This chapter will start with the integration architecture, which refers to the orders of the acquired match scores, being integrated to render the biometric decision (Section 2.1). In Section 2.2, the mathematical framework that describes the classifier combination in the context of statistical pattern recognition is presented. Sections 2.3 and 2.4 discuss in detail the state-of-the-art in match score fusion schemes, and their application to multimodal biometrics. In general, these schemes can be divided into: (1) the traditional user-independent processing, and (2) a new attempt in the literature to exploit the user specificities [15]. Based on this review, several best candidates will be selected to further evaluate and compare with the proposed schemes of this thesis, discussed in Chapters 4 and 5.

2.1. Integration Architecture

The integration architecture refers to the order, where the acquired biometric information is processed in order to render a decision. It is known to significantly impact the throughput time in large-scale identification systems [1]. The various modes of the integration architecture, illustrated in Figure 2.1, are discussed below.

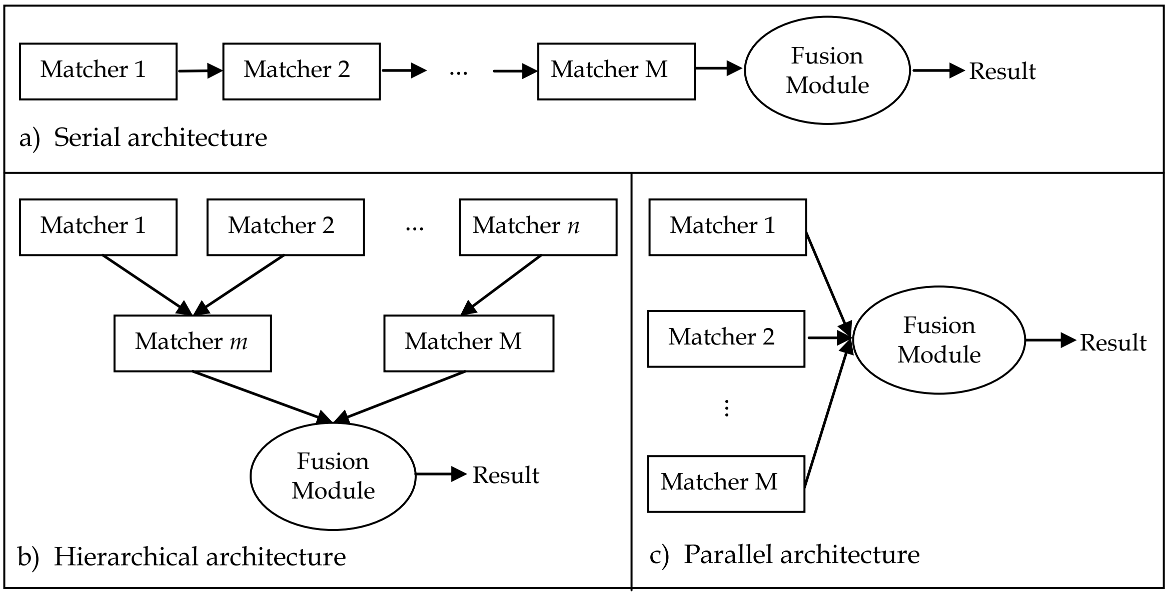


Figure 2.1. General integration architectures of multiple biometric matchers. Adapted from [14].

In the *serial (or cascade) mode*, information processing takes place sequentially. The serial mode can reduce the processing time while improving user convenience since the decision may be made without going through all the biometric matchers [1], [31]. In [93], the author developed a serial architecture where the face matcher is used to retrieve the top *n* matching identities, while the finger matcher is used to determine the final identity. In [31], the serial architecture was employed to select a subset of matchers for each of the users, enrolled in the biometric system. Various aspects were considered in this work, such as the variability of the system performance across the users, and the reliability of a user-ranking criterion based solely on a given data set.

In the *hierarchical architecture*, the different biometric modalities are processed in a tree-like structure. The hierarchical architecture is known to be highly efficient in exploiting the discriminative power of different types of features [14], [16], [147]. In [148], the authors developed a hierarchical palmprint system, where multilevel palmprint features, such as key point distance, tuned mask based texture energy, fuzzy interest lines, and local directional texture energy were extracted and hierarchically matched to facilitate a coarse-to-fine palmprint matching scheme, and thus, reducing the computational time. A similar idea was exploited in [149], where different fingerprint features, namely, the orientation image, minutiae, and pores and ridge contours, were hierarchically consolidated in an identity verification system. It was observed that using the pores and ridge contours in combination with the other features results in a relative reduction of 20% in the EER, and the performance gain is consistently achieved across various quality fingerprint images. In the context of multimodal biometrics, the hierarchical structure, used for combining the various traits can be determined, based on the quality of the individual input samples and the probability of encountering missing data [1]. However, its design is much application-dependent, and hence, it has not received much attention from researchers.

In the *parallel mode*, each biometric matcher independently processes its information, which is then combined to render the decision. The parallel architecture has been the most widely used and investigated in the biometric literature. It generally achieves a lower error rate with respect to that of the serial mode because it uses more evidence about the user for authentication [1], [11], [13]. This PhD thesis is focused on biometric matchers, designed to operate in the parallel mode, since its primary goal has been to reduce the generalization error, and thus, enhancing the matching performance.

2.2. Parallel Classifier Combination Rules

It has been theoretically demonstrated that combining the classifiers has led to improved performance. For instance, in [150], the authors divided the total error into the Bayes error and the added error. The added error is the extra error due to the specific classification model used. They theoretically demonstrated that for independent classifiers, linearly combining their outcomes would result in a smaller added error with respect to the averaged individual added error. A similar observation was found in [151], where the added error was derived for the weighted mean combination of the classifiers.

In this section, several classifier combination rules for the parallel architecture, which can be deployed in the Fusion Module in Figure 2.1, are outlined from a pattern recognition perspective. These combination rules were developed in [95] to consolidate the evidence from multiple classifiers, where each classifier uses a different feature representation of the same input pattern. Consider the problem of classifying an input pattern $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_M]$ into one of the C possible classes $\{w_1, w_2, \dots, w_C\}$, where $\mathbf{x}_m$ is a feature vector, presented to the m-th classifier, and M is the number of classifiers. Each of the classifiers represents the given pattern by a distinct measurement vector. In the feature space, each class w_k is modelled by a probability density function $p(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_M | w_k)$ and its prior probability $p(w_k)$, respectively.

According to the Bayes theorem [96], the input pattern $\mathbf{X}$ should be assigned to class w_r, which maximizes the a posteriori probability, i.e.,

$$\begin{gathered}\text{Assign } \mathbf{X} \to w_r \text{ if} \\ p(w_r | \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_M) = \max_k p(w_k | \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_M)\end{gathered} \tag{2.1}$$

where $k = 1, \dots, C$. The a posteriori probability in Equation (2.1) can be expressed in terms of the joint conditional probability density of the feature vectors as follows:

$$p(w_r | \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_M) = \frac{p(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_M | w_r) p(w_r)}{\sum_{k=1}^{C} p(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_M | w_k) p(w_k)} \tag{2.2}$$

In [95], the authors suggested many approximations to simplify the computation of the a posteriori probability in Equation (2.2), which led to several combination rules, i.e., product, sum, max, min, and median rules [14]. The product rule is obtained by assuming that the M feature representations are statistically independent. This is reasonable in most multimodal biometric systems, where different biometric traits tend to be mutually independent [1]. The sum rule is achieved by further assuming that the a posteriori probabilities of the classifiers do not deviate dramatically from the prior probabilities [95]. The remaining rules are obtained by bounding the a posteriori probabilities [14].

Product Rule: Assign $\mathbf{X} \to w_r$ if

$$\prod_{m=1}^{M} p(w_r|\mathbf{x}_m) = \max_{1\leq k\leq C} \prod_{m=1}^{M} p(w_k|\mathbf{x}_m) \tag{2.3}$$

Sum Rule: Assign $\mathbf{X} \to w_r$ if

$$\sum_{m=1}^{M} p(w_r|\mathbf{x}_m) = \max_{1\leq k\leq C} \sum_{m=1}^{M} p(w_k|\mathbf{x}_m) \tag{2.4}$$

Max Rule: Assign $\mathbf{X} \to w_r$ if

$$\max_{1\leq m\leq M} p(w_r|\mathbf{x}_m) = \max_{1\leq k\leq C} \max_{1\leq m\leq M} p(w_k|\mathbf{x}_m) \tag{2.5}$$

Min Rule: Assign $\mathbf{X} \to w_r$ if

$$\min_{1\leq m\leq M} p(w_r|\mathbf{x}_m) = \max_{1\leq k\leq C} \min_{1\leq m\leq M} p(w_k|\mathbf{x}_m) \tag{2.6}$$

Median Rule: Assign $\mathbf{X} \to w_r$ if

$$\operatorname*{median}_{1\leq m\leq M} p(w_r|\mathbf{x}_m) = \max_{1\leq k\leq C} \operatorname*{median}_{1\leq m\leq M} p(w_k|\mathbf{x}_m) \tag{2.7}$$

Consider a multimodal biometric system operating in the verification mode, a biometric input pattern **X** is classified into one of the two classes, i.e., genuine user (w_G) or impostor (w_I). Equation (2.1) is based on the assumption that all types of errors are equally costly. However, this assumption does not hold in most practical authentication systems, which assign different costs to the False Acceptance Rate (FAR) and False Rejection Rate (FRR) [1]. In general, FAR and FRR are viewed as functions of the decision threshold τ. For a given τ, Equation (2.1) can be rewritten as

$$\text{Assign } \mathbf{X} \to w_G \text{ if}$$

$$\frac{p(w_G|\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_M)}{p(w_I|\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_M)} \geq \tau \tag{2.8}$$

Since the M feature vectors $[\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_M]$ are not available, the a posteriori probabilities have to be estimated based on the vector of match scores. In what follows, a thorough discussion on the state-of-the-art techniques available in the biometric literature to estimate the a posteriori probabilities is provided. I adhere to the taxonomy, described in [15], to divide these techniques into two broad categories: i.e., user-independent processing, and user-specific processing. User-specific processing, as opposed to the user-independent counterpart, takes into account the label of the claimed identity for a given access request. This taxonomy is depicted in Figure 2.2. In the rest of this chapter, a brief review on each category of Figure 2.2 will be given. The contributions originating from this PhD research (drawn in bold) will be detailed in Chapters 4 and 5.

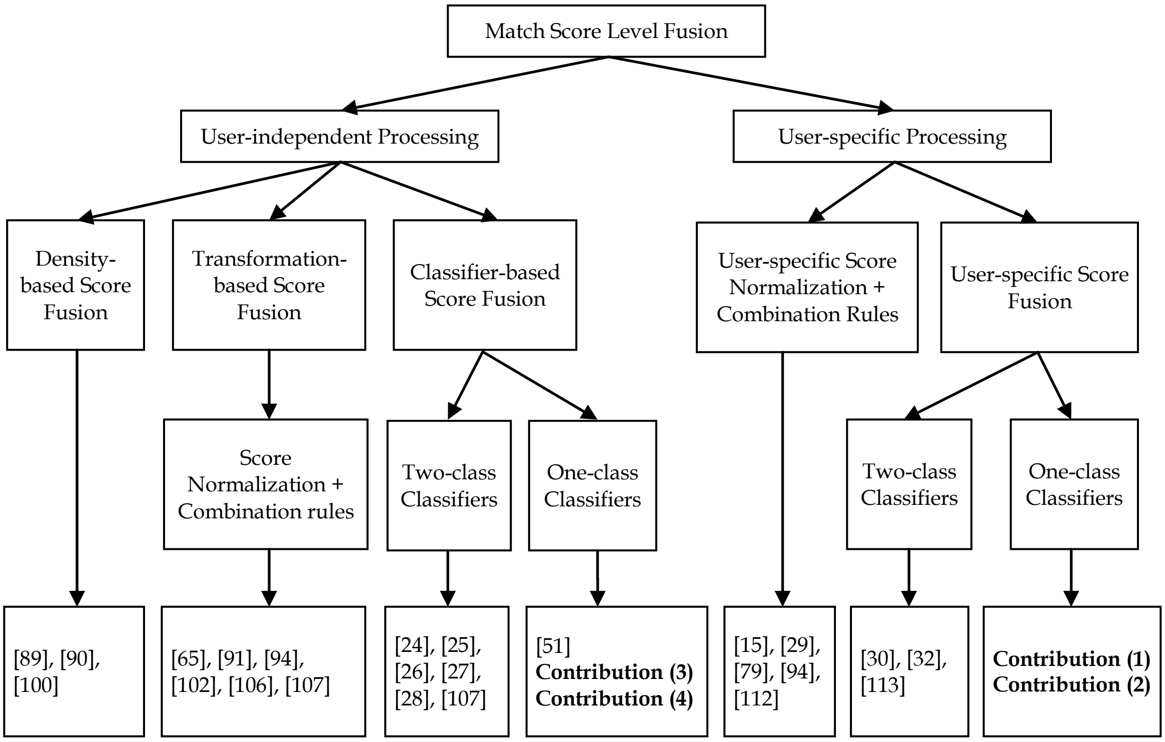

Match Score Level Fusion
User-independent Processing
User-specific Processing
Density-based Score Fusion
Transformation-based Score Fusion
Classifier-based Score Fusion
User-specific Score Normalization + Combination Rules
User-specific Score Fusion
Score Normalization + Combination rules
Two-class Classifiers
One-class Classifiers
Two-class Classifiers
One-class Classifiers
[89], [90], [100]
[65], [91], [94], [102], [106], [107]
[24], [25], [26], [27], [28], [107]
[51] **Contribution (3)** **Contribution (4)**
[15], [29], [79], [94], [112]
[30], [32], [113]
Contribution (1) **Contribution (2)**

Figure 2.2. Approaches to match score level fusion in multimodal biometric authentication. The contributions originating from the research are drawn in bold. Adapted from [15].

2.3. User-independent Processing

Fusion based on user-independent processing is performed independently of the claimed identity. A major research effort in the biometric literature is focused on user-independent processing. A number of techniques have been developed, which can be further grouped into three categories [1], specifically density-based, transformation-based, and classifier-based score fusion. These techniques differ in their ability to estimate the a posteriori probabilities based on the vector of match scores, provided by the individual matchers. They also exhibit different levels of sensitivity to the problems, in which the training data is not sufficient to guarantee the generalization capabilities and may include noisy training samples [1]. Without loss of generality, let $\{\mathbf{s}_i\}_{i=1}^{N}$ be N training samples, where $\mathbf{s}_i = [s_i^1, s_i^2, \dots, s_i^M]$ and M is the number of matchers, and $\mathrm{y}_i \in \{-1, +1\})$ be their corresponding class labels, where -1 denotes an impostor and +1 denotes a genuine user. Suppose that $\mathbf{s}_T$ is the test input match score vector ($\mathbf{s}_T = [s_T^1, s_T^2, \dots, s_T^M]$). These notations will be used throughout this chapter.

2.3.1. Density-based Score Fusion

This approach assumes that the a posteriori probabilities, i.e., $p(w_G|\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_M)$ and $p(w_I|\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_M)$ can be reliably approximated based on the vector of match scores. It requires explicit estimation of the underlying conditional densities $p(\mathbf{s}_T|w_G)$ and $p(\mathbf{s}_T|w_I)$ using the match score vectors. Hence, it is called **density-based score fusion**. By replacing the M feature vectors $[\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_M]$ with the match score vector $\mathbf{s}_T$, Equation (2.8) can be rewritten as

$$\text{Assign } \mathbf{s}_T \rightarrow w_G \text{ if} \\ \frac{p(w_G|\mathbf{s}_T)}{p(w_I|\mathbf{s}_T)} \geq \tau \tag{2.9}$$

According the Bayes theorem [96], the a posteriori probabilities are as follows

$$p(w_G|\mathbf{s}_T) = \frac{p(\mathbf{s}_T|w_G)p(w_G)}{p(\mathbf{s}_T)} \tag{2.10}$$

and

$$p(w_I|\mathbf{s}_T) = \frac{p(\mathbf{s}_T|w_I)p(w_I)}{p(\mathbf{s}_T)} \tag{2.11}$$

By assuming that the prior probabilities are equal, i.e., $p(w_G) = p(w_I)$, Equation (2.9) can be rewritten as

$$\text{Assign } \mathbf{s}_T \rightarrow w_G \text{ if}$$

$$\frac{p(\mathbf{s}_T|w_G)}{p(\mathbf{s}_T|w_I)} \geq \tau \tag{2.12}$$

The terms $p(\mathbf{s}_T|w_G)$ and $p(\mathbf{s}_T|w_I)$ are referred to as the likelihood of the genuine and the impostor classes with respect to $\mathbf{s}_T = [s_T^1, s_T^2, \dots, s_T^M]$. Hence, the ratio in Equation (2.12) is known as the likelihood ratio test. According to Neyman-Pearson [97], when the prior probabilities of the classes are equal (or not known), the likelihood ratio test is considered as the optimal test for classifying an input match score vector. This test is optimal in the sense it will minimize the False Rejection Rate for a fixed False Acceptance Rate and no other decision rule can give a lower FRR [1]. The likelihood ratio test can be alternatively presented in terms of log-likelihood ratio, which is as follow:

$$\log\frac{p(\mathbf{s}_T|w_G)}{p(\mathbf{s}_T|w_I)} = \log\big(p(\mathbf{s}_T|w_G)\big) - \log\big(p(\mathbf{s}_T|w_I)\big) \geq \log(\tau) \tag{2.13}$$

The likelihood ratio test in Equation (2.12) and log-likelihood ratio test in Equation (2.13) are optimal only when the underlying densities are either known or can be estimated accurately [90]. It is well-known that the Gaussian density is not always appropriate for modelling the underlying densities of biometric match scores. This is due to the fact that a biometric baseline matcher may apply thresholds at various stages in the matching process. For example, some fingerprint matchers produce a match score of zero if the number of extracted minutiae is less than a threshold. This leads to discrete components in the match score distribution, which cannot be modelled accurately using the continuous Gaussian density [100].

In [91], [100], the biometric match scores were modelled using the Parzen Window based non-parametric density estimation [96] and kernel density estimators [152]. However, these approaches require a careful choice of the window width and kernel width, which are critical to their performance, respectively. In [90], Gaussian Mixture Model (GMM) was proposed for estimating the densities of match scores. It was reported that GMM is able to achieve a comparable matching accuracy with respect to that of Support Vector Machines (SVM). The theoretical results in [98], [99] also showed that GMM indeed converges to the true density when a sufficient number of training samples is available [90]. Assume that $p(\mathbf{s}_T|w_G)$ $(p(\mathbf{s}_T|w_I))$ represents the densities, which are estimated using the genuine (impostor) score vector. According to [90], these conditional densities are as follows:

$$p(\mathbf{s}_T|w_G) = \frac{1}{N_G}\sum_{i=1}^{N_G} \alpha_{G,i} p(\mathbf{s}_T, \mu_{G,i}, \Sigma_{G,i}) \tag{2.14}$$

and

$$p(\mathbf{s}_T|w_I) = \frac{1}{N_I}\sum_{i=1}^{N_I} \alpha_{I,i} p(\mathbf{s}_T, \mu_{I,i}, \Sigma_{I,i}) \tag{2.15}$$

where $\alpha_{k,i}$ is the mixing coefficient assigned to the i-th mixture component in $p(\mathbf{s}_T|w_k)$, and $p(\mathbf{s}_T, \mu_{k,i}, \Sigma_{k,i})$ is the Gaussian distribution, characterised by the mean $\mu_{k,i}$ and covariance matrix $\Sigma_{k,i}$, where $k \in \{G, I\}$ (genuine user/impostor). When the number of mixture components is known, the means and covariance matrices of the individual mixture components can be estimated using the Expectation Maximization (EM) algorithm [54]. However, selecting the number of mixture components is a challenging problem. As stated in [90], a mixture with too few components may not approximate the true density, while a mixture with too many components may result in overfitting. In this research, these numbers are found by using cross validation and grid search [54].

In [100], the authors combined the match scores from the multiple matchers using the generalized densities, which are derived from the genuine and impostor match scores. The generalized densities were estimated using the multivariate Gaussian copulas [101] to incorporate the correlation (if it exists) among the biometric matchers. From a series of experiments, carried out on the MSU [107] and NIST BSSR1 databases [52], it was observed that the copula model cannot achieve a lower error rate with respect to that of GMM in [90] when the biometric matchers are approximately independent of each other. Even when the degree of correlation among the matchers is more significant (i.e., 0.75 and 0.29 for the genuine and impostor match scores, respectively), the performance gain is not significant. It should be noted that the copula model requires more parameters, which have to be specified by the user. Furthermore, it is also well-known that the different biometric traits tend to be mutually independent. Due to these reasons, the generalized density estimation in [100] is not further considered in this thesis.

2.3.2. Transformation-based Score Fusion

In practice, when the conditional densities $p(\mathbf{s}_T|w_G)$ and $p(\mathbf{s}_T|w_I)$ cannot be accurately estimated due to the limited availability of training data, it would be better to directly combine the match scores without converting them into a posteriori probabilities [1]. However, this brings about to a number of challenges because the match scores of the individual matchers may be inhomogeneous, need not be on the same numerical scale, and may follow different statistical distributions. Due to these reasons, it is essential to transform the match scores into a common reference domain prior to combining them [91]. This transformation is known as score normalization and the corresponding fusion approach is called **transformation-based score fusion** [1].

By definition, score normalization is the change in the location and scale parameters of the match score distributions. It is also referred to as score calibration. For a good score normalization scheme, the estimates of the location and scale parameters have to be efficient and robust [1], [91]. Efficiency refers to the proximity of the obtained estimates with respect to the optimal ones, while robustness refers to the insensitivity to the presence of outliers. A large number of score normalization techniques, such as Min-max, Decimal Scaling, Z-norm, Median and Median Absolute Deviation and Tanh-Estimators have been proposed in the biometric literature. It is therefore essential to determine a technique, which has both the desired characteristics of robustness and efficiency [1]. Without loss of generality, suppose that N is the number of match scores in the training set, and n_T^m is the normalized match score for the match score s_T^m, provided by the m-th matcher.

Min-max

The Min-max normalization is the simplest normalization technique, which is able to retain the original distribution of match scores and transforms all the match scores into a common range [0,1]. It is given by

$$n_T^m = \frac{s_T^m - \min\{s_i^m\}_{i=1}^N}{\max\{s_i^m\}_{i=1}^N - \min\{s_i^m\}_{i=1}^N} \tag{2.16}$$

where $\max\{s_i^m\}_{i=1}^N$ and $\min\{s_i^m\}_{i=1}^N$ are the highest and lowest values of the match scores. In [70], the authors suggested that the Min-max should be chosen, provided that there are no outliers in the training data set. In [94], the authors conducted an empirical evaluation using the state-of-the-art Commercial Off-the-shelf (COTS) fingerprint and face baseline systems on a population approaching 1,000 individuals. They observed that the Min-max achieves lower error rates as compared to those of the Z-norm and Tanh-Estimators. Similar findings were reported in [1] and [51], in which NIST BSSR1 [52] and MCYT databases [53] were used to carry out the experiments.

It has to be noted that the Max-min is not optimal for all kinds of match score data. For instance, it was demonstrated in [90] that the Min-max is not able to achieve a satisfactory performance on the XM2VTS database. The rationale for this is that the score distributions of the face and speech matchers are very different. The Min-max is generally not effective in handling the face match scores because the genuine and impostor face match scores are the outputs of a Multilayer Perceptron (MLP), which are peaked around 1 and -1 [90].

Decimal Scaling

When the match scores from multiple matchers are on a logarithmic scale, it is possible to normalize them using the Decimal Scaling as

$$n_T^m = \frac{s_T^m}{10^{c_m}} \tag{2.17}$$

where $c_m = \log_{10}(\max\{s_i^m\}_{i=1}^N)$. The Decimal Scaling maintains the match scores in the range of $[-1, +1]$. It is however based on an invalid assumption that the match scores vary by a logarithmic factor [16]. The Decimal Scaling is highly sensitive to outliers as it requires a reliable estimation of the maximum values of the match scores. Due to these reasons, it has received a very limited attention in the biometric literature.

Z-norm

The Z-norm is based on the assumption that the match scores of the individual matchers follow a Gaussian distribution. The Z-norm uses the mean μ^m and standard deviation σ^m of the training data. It is therefore expected to perform well if these parameters are reliably estimated. The normalized score is given by

$$n_T^m = \frac{s_T^m - \mu^m}{\sigma^m} \tag{2.18}$$

The Z-norm is the most popular score normalization technique, used in text-independent speaker verification applications, where GMM has become the dominant methodology for building the reference model for each of the users, enrolled in the system [166], [167].

In [91], the authors observed that the Z-norm is among the best evaluated score normalization schemes, which were applied to transform the match scores from the face, fingerprint and hand-geometry modalities of the MSU database [107] into a common numerical domain. In [106], the Z-norm was reported to be better than the Min-max and Tanh-Estimators on the experiments, which were carried out on the XM2VTS benchmark database [63]. However, it should be noted that when the Gaussian assumption is invalid, the Z-norm does not retain the distribution of the data [16]. It is not robust since the estimates of μ^m and σ^m are sensitive to outliers. Also, the Z-norm may not guarantee a common numerical range for the normalized match scores of the different matchers [91].

Median and Median Absolute Deviation (MAD)

The Median and MAD normalized score is given by

$$n_T^m = \frac{s_T^m - median^m}{MAD^m} \tag{2.19}$$

where $median^m = \text{median}\{s_i^m\}_{i=1}^N$ and $MAD^m = \text{median}\{(|s_i^m - median^m|)\}_{i=1}^N$. Both the $median^m$ and MAD^m are insensitive to outliers and points in the extreme tails of the distribution [1]. Nevertheless, the median and MAD estimators have a low efficiency as compared to the mean and standard deviation estimators, i.e., when the match score distribution is not Gaussian, the location and scale parameters can be poorly estimated [91]. As previously discussed, the genuine and impostor match score distributions may contain discrete components and have more than one mode. Hence, the Median and MAD generally cannot perform as well as the Z-norm. The empirical studies, reported in [1], [65] and [91] further support this observation.

Tanh-Estimators

Finally, the Tanh-Estimators [165] is considered, which is fairly robust and efficient. The Tanh-Estimators is given by

$$n_T^m = \frac{1}{2}\left\{tanh\left(0.01\left(\frac{s_T^m - \mu_{GH}^m}{\sigma_{GH}^m}\right)\right) + 1\right\} \tag{2.20}$$

where μ_{GH}^m and σ_{GH}^m are the mean and standard deviation estimates of the genuine score distribution, which are given by Hampel estimators [165]. Hampel estimators are based on the following influence function:

$$\psi(u) = \begin{cases} u & 0 \le |u| < a \\ a \text{ sign}(u) & a \le |u| < b \\ a \text{ sign}(u)\left(\dfrac{c - |u|}{c - b}\right) & b \le |u| < c \\ 0 & c \le |u| \end{cases} \tag{2.21}$$

Since the $\psi(u)$ function is capable of reducing the influence of the match scores at the tails of the genuine score distribution (identified by a, b and c), the Tanh-

Estimators is not sensitive to outliers [1]. In [70], the authors suggested that this technique is most suitable for normalizing non-Gaussian match scores. The Tanh-Estimators is complicated to implement as it requires a careful selection of the parameters a, b and c in the Hampel estimators to obtain good efficiency. In [91], a, b and c were chosen such that 70%, 85%, and 95% of the match scores fall in the range $[median^m - a, median^m + a]$, $[median^m - b, median^m + b]$ and $[median^m - c, median^m + c]$, respectively.

Table 2.1 summarizes the characteristics of the various normalization schemes. The Tanh-Estimators has both desired characteristics, i.e., robustness and efficiency. The Median and MAD is robust but inefficient. On the contrary, the Min-max, Decimal Scaling and Z-norm are efficient but not robust.

Table 2.1. Summary of score normalization techniques. Adapted from [1] and [67].

Score Normalization	Robustness	Efficiency
Min-max	No	High
Decimal Scaling	No	High
Z-norm	No	High
Median and MAD	Yes	Moderate
Tanh-Estimators	Yes	High

On the basis of the previous discussion, it is worth noting that no single score normalization technique has been found to be universally the best. In [1], it was recommended that various normalization schemes would be evaluated to determine the one that gives the best performance on a given data. In this PhD thesis, the match scores of the individual face and speech matchers of the XM2VTS database were normalized using the Z-norm, as recommended in [63].

In the transformed domain, the sum, max, min combination rules in Section 2.2 can be applied to combine the match scores of the different matchers. Since normalized match scores do not have any probabilistic interpretation, the sum, max, and min combination rules can be referred to as sum of scores, max score and min score, respectively [1], [91]. The combined match scores can be computed as a weighted sum of the match scores, which is referred to as the weighted sum method. This is motivated by the fact that biometric matchers are not of identical accuracy. Hence, it is reasonable to give the more competent matcher more weight in making the final decision. In [59], it was observed that the weights only magnify the relevance of each matcher based on its accuracy, and do not take into account the performance of other matchers. In [102], the weights were directly obtained, based on either Equal Error Rate (EER) or Fisher discriminant analysis [54]. Although Fisher discriminant analysis was reported to achieve higher verification performance, it is based on a strong Gaussian assumption with equal covariance matrices, and thus, is particularly sensitive to the underlying joint distribution of the match scores of the individual matchers [106].

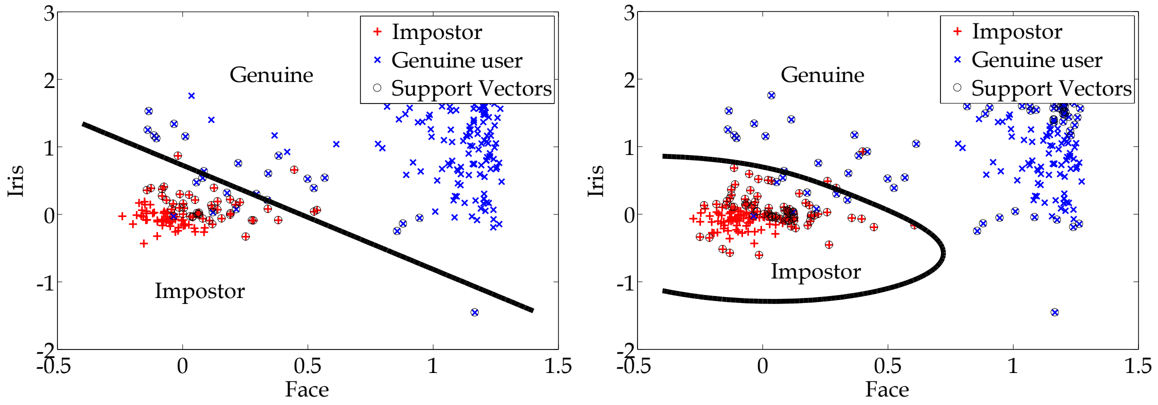
Generally speaking, the weights can be found by adopting the exhaustive search over all possible candidates. Nonetheless, the exhaustive search is very time-consuming and may not be feasible when a large number of biometric matchers are available for rendering the biometric decision. Several authors have examined the evolutionary algorithms to reduce the search space. Among others, one of the most perspective algorithms is known as Bees Algorithm [66], [103]. Intuitively, the Bees Algorithms performs a kind of neighbourhood search, combined with random search in a way that is reminiscent of the food foraging behaviour of swarms of honey bees. It is very efficient at searching optimal solutions and overcoming the problem of local optima [103]. In [65], the Bees Algorithm was employed to search for the weights in order to combine five facial cues (total face, left and right eye, nose and mouth) and optimize the recognition rate in a face identification system. This approach was observed to achieve either better or comparable performance to that of the density-based score fusion (Section 2.3.1) on the experiments, carried out on the CASIA [104] and ORL face databases [105].

It is important to note that there is no single best combination rule for all types of biometric data. There has been a debate in the literature concerning the weighted sum method, such as whether or not the weights should be positive and should be constrained to sum to one [59]. In [106], a systematic study was conducted using the XM2VTS database to assess the possible effect of score normalization on the various combination rules, including the sum of scores and the weighted sum method with the weights, optimised using the exhaustive search, Fisher-ratio [102] and Decision templates [1]. It was confirmed that the Fisher-ratio and Decision templates are particularly sensitive to the underlying joint distributions of the match scores. The sum of scores was reported to achieve the lowest error rates among these schemes. Thus, only the sum of scores will be selected and further studied in the subsequent chapters. The work in [65], related to the application of the Bees Algorithm is not further considered in this thesis since its results are similar or mostly similar to those of the exhaustive search.

2.3.3. Classifier-based Score Fusion

In classifier-based score fusion, the relationship between the match score vector and the a posteriori probabilities is learned using a pattern classifier. In this context, the pattern classifier is viewed as a two-class problem, where the vector of match scores is treated as a feature vector, and thus, can be classified into one of two classes: genuine user/impostor. Based on the training set of match scores, a classifier learns the decision boundary between the two classes. The decision boundary can be quite complex depending on the nature of the classifier. However, it can be learned irrespective of how the feature vectors are generated [1]. Figure 2.3 shows examples of the decision boundary, learned by SVM using the match scores of the face and iris matchers from the BioSecure DS2 database. Since the match scores from these matchers are in the range $[0, \dots, 1]$ and $[0, \dots, 100]$, they are normalized using the F-norm, as suggested in [61] to be better aligned and separated. During authentication, any match score vector that falls in the 'Impostor' region is classified as an impostor, while any match score vector that falls in the 'Genuine' region is classified as a genuine user.

A number of classifiers have been used in the literature to render the biometric decision. In [24], the HyperBF network was used to combine vocal and three facial cues (eyes, nose and mouth) in an identity verification system. It was observed that the use of multiple acoustic and visual cues is beneficial as it achieves a rank-1 recognition rate of 98%, while those of the voice and face matchers were 88% and 91%, respectively. In [25], the authors evaluated three different classifiers, namely, k-Nearest Neighbour classifier using vector quantization, the classifier based on logistic regression model, and decision tree classifier for the fusion of match scores of three biometric matchers, which were based on voice, frontal and profile face images. It was reported that the classier based on logistic regression model achieves the highest verification performance. Such an approach provides a total error rate of 0.1%, while the total error rates of the individual modalities (in the order of profile, frontal face, and voice) are 8.9%, 8.7%, and 3.7%, respectively.



(a)

(b)

Figure 2.3. Examples of the decision boundary, learned using SVM with (a) linear kernel, and (b) RBF kernel in 2-dimensional feature space using match scores of the face and iris matchers from the BioSecure DS2 database. Since the match scores from these matchers are in the range [0, … ,1] *and* [0, … ,100]*, they are normalized using the F-norm [79], as suggested in [61] to be better aligned and separated. Adapted from [1].*

In [27], k-means clustering, fuzzy clustering, and median radial basis function (MRBF) were employed to combine five biometric matchers, which are based on the grey-level and shape information of face image and voice features. All these approaches were reported to achieve lower error rates as compared to those of the individual matchers. No single approach was found to be universally the best. In [107], the face, fingerprint, and hand geometry modalities were consolidated using either decision tree or linear discriminant classifier. In [28], random forest algorithm [138] was used for the classification of 3 dimensional match score vectors, described in [107], including face, fingerprint and hand geometry. The decision tree, linear discriminant classifier and random forest were demonstrated to achieve a much higher performance with respect to that of the individual biometric modalities. However, it is not possible to fix the FAR (and then compute the FRR at the specified FAR) when the multimodal biometric fusion is performed using these techniques [107].

Support Vector Machines (SVM) [120] has largely advanced the situation in terms of decision boundary design. In [26], the authors compared the relative performance of a number of classification schemes, including C4.5 decision tree, Fisher linear discriminant, Bayesian classifier, Multilayer Perceptron (MLP), and SVM with Gaussian and polynomial kernels, as applied in combining the face and speech data from the XM2VTS database. It was reported that SVM achieves the best results among the evaluated schemes. Hence, SVM will be discussed in detail and evaluated in this dissertation.

The aim of SVM is to determine the optimal hyperplane by maximising the margin between the separator hyperplane:

$$\{\mathrm{h} \in \mathbb{H} | \langle \mathrm{w}, \mathrm{h} \rangle_{\mathbb{H}} + w_0 = 0\} \tag{2.22}$$

and the mapped data $\Phi(\mathbf{s}_i)$, where $\langle \cdot,\cdot \rangle$ denotes the inner product in space $\mathbb{H}$, and w are the hyperplane parameter. The optimal hyperplane is often considered as the solution of the following Quadratic Programming problem [120]:

$$\min_{\mathrm{w}, w_0, \xi_i, \ldots, \xi_N} \left(\frac{1}{2} \|\mathrm{w}\|^2 + C \sum_{i=1}^{N} \xi_i \right) \tag{2.23}$$

subject to

$$\begin{aligned} & y_i(\langle \mathrm{w}, \Phi(\mathbf{s}_i) \rangle_{\mathbb{H}} + w_0) - 1 + \xi_i \geq 0 && i = 1, \ldots, N \\ & \xi_i \geq 0 && i = 1, \ldots, N \end{aligned}$$

where N is the number of training samples, ξ_i are the slack variables, which are introduced to take account of the eventual non-separability of $\Phi(\mathbf{s}_i)$ and C is a positive constant to control the trade-off between the slack variable and the size of the margin. The problem in Equation (2.23) is solved by using the dual representation and the kernel trick [54], [120] as

$$\max_{\alpha_i,\ldots,\alpha_N}\left(\sum_{i=1}^{N}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{N}\alpha_i\alpha_j y_i y_j K(\mathbf{s}_i,\mathbf{s}_j)\right) \tag{2.24}$$

subject to

$$0\le\alpha_i\le C \qquad i=1,\ldots,N$$

$$\sum_{i=1}^{N}\alpha_i y_i=0$$

where $K(\mathbf{s}_i,\mathbf{s}_j)=\langle\Phi(\mathbf{s}_i),\Phi(\mathbf{s}_j)\rangle$ is the kernel function, which is introduced to avoid direct manipulation of the samples in $\mathbb{H}$, and α_i are Lagrange multipliers, which can be determined as the solution of a Quadratic Programming problem [120]. In particular, the Radial Basis Function (RBF)

$$K(\mathbf{s}_i,\mathbf{s}_j)=\exp\left(\frac{\left\|\mathbf{s}_i-\mathbf{s}_j\right\|^2}{2\sigma^2}\right) \tag{2.25}$$

is used in this PhD thesis. RBF may result in complex separating surfaces between genuine and impostor score distributions. In [14], the combined match score corresponding to test pattern $\mathbf{s}_T$ is then defined as

$$f(\mathbf{s}_T)=\sum_{i\in\mathrm{SV}}\alpha_i y_i K(\mathbf{s}_i,\mathbf{s}_T)+w_0^* \tag{2.26}$$

where $\mathrm{SV}=\{i|\alpha_i>0\}$ indexes the set of support vectors, and w_0^* is the bias parameter, determined using Equation (2.24). Once the combined match score is obtained, it is compared with a decision threshold τ in order to classify the input match score vector $\mathbf{s}_T$ as belonging to the impostor or genuine user class.

2.4. User-specific Processing

Recent studies have suggested that some users are more difficult to recognize than others, despite the database being captured in similar conditions [15], [108], [109]. Some users do not perform consistently well in term of False Rejection Rate and False Acceptance Rate [109]. For example, it has been empirically estimated that about 4% of the population may have poor quality fingerprints, which are difficult to recognize automatically with existing fingerprint systems [1]. Due to the existence of user-variation, a recent trend in multimodal biometrics is to focus on the fusion techniques, which are capable of making use of the information, related to each of the users, enrolled in the systems. These schemes are called user-specific processing, which will be discussed in detail in the subsequent sections.

2.4.1. The Biometric Menagerie

Naturally, users of a biometric system may have differing degrees of accuracy within the system. This effect is known as the biometric menagerie [108], where the users are characterised by animal labels:

1) **Sheep** refer to the majority of the population, who match poorly against others and well against themselves.

2) **Goats** refer to the users, who are particularly difficult to recognize. Goats contribute significantly to the False Rejection Rate (FRR).

3) **Lambs** are those users, who are easy to imitate. Lambs represent a potential system weakness. They result in relative high match scores when being matched against, accounting for a disproportionate share of false accepts.

4) **Wolves** are those users, who are successful at imitating others. Wolves also increase the False Acceptance Rate (FAR).

The biometric menagerie was first studied in [108], where the population statistics for speaker recognition performance were computed and analysed based on test data from the NIST 1998 speaker recognition evaluation. It was observed that the population of speakers exhibits a continuum of goatish, wolfish and lambish characteristics. Although the true animal-like behaviour is more likely for behavioural biometric systems, this concept is applicable to all areas of biometric verification. In [110], the authors demonstrated the existence of wolves and lambs in fingerprint based data. Evidence for the presence of goats, wolves, and lambs in face recognition was examined in [111].

In [109], the author conducted tests on a variety of biometric modalities, such as fingerprint, iris, 2D face and speech, 3D faces and keystroke. They also suggested a new class of animals, which can be defined in terms of a relationship between genuine and impostor match scores. These animals give further insights into individual performance. It was empirically observed that goats, lambs and wolves appear everywhere, while the presence of the new class of animals varies widely between the systems. This further confirms that users within the biometric systems have their own match score distributions.

In [153], the authors proposed explicitly to rank the users according to their performance using criteria such as F-ratio [15], Fisher-ratio and the d-prime [1]. They demonstrated that these criteria are able to partition the users in such a way that the performance of each partition differs by as much as a factor of 2. A similar idea was introduced by the same authors [31], where the users were ranked such that the rank order, obtained based solely on the training data set would maximally correlated with the rank order, which is derived on the test set.

Quantitative method for dealing with the existence of user variation is an active area of research [109]. There are three primary user-specific techniques under investigation: (1) user-specific thresholds, which assign a different decision threshold to each user; (2) user-specific score normalization, which transforms the match scores for each user into a common domain; and (3) user-specific score fusion, which takes into account the label of the claimed identity for a given access. In [15], it was proved that manipulating the threshold or the match score may lead to a similar result. Hence, user-specific thresholds can be considered as a sub-category of user-specific score normalization.

2.4.2. User-specific Score Normalization

The most representative techniques for user-specific score normalization are the Z-norm [112] and the F-norm [15], [79]. In [15], [79], the F-norm and the Z-norm were compared. The majority of the experiments demonstrated that the F-norm is superior to the Z-norm. Similar observation and findings can be found in [114]. The rationale for this is that the F-norm considers the genuine and impostor class distributions simultaneously, while the Z-norm exclusively relies on the impostor class distribution.

Let $\mu_{j,G}^{m}$ ($\mu_{j,I}^{m}$) be the mean of the genuine (impostor) match scores, provided by the m-th matcher for the user j. The F-norm is given as

$$n_{T,j}^{m} = \frac{s_{T,j}^{m} - \mu_{j,I}^{m}}{\alpha\mu_{j,G}^{m} + (1-\alpha)\mu_{G}^{m} - \mu_{j,I}^{m}} \tag{2.27}$$

where $n_{T,j}^{m}$ is the normalized match score for the test score $s_{T,j}^{m}$. Since $\mu_{j,G}^{m}$ cannot be reliably estimated due to the small sample-size of the user-specific genuine match scores, it is compensated by the system-wide (user-independent) genuine mean score μ_{G}^{m} via a tuneable parameter $\alpha \in [0,\ldots,1]$. In [15], where α is associated with the so-called "relevance factor". The fundamental idea is to further parameterize α as a function of the number of user-specific genuine samples [31]. In this thesis, α is found through directly optimization using the training data. As suggested in [61], the F-norm is used to normalize the match scores of the BioSecure DS2 database.

2.4.3. User-specific Score Fusion

User-specific score fusion was first studied in [29], where different set of weights, indicating the importance of individual biometric matchers, were determined on a user-by-user basis. For instance, users with persistently dry fingerprint may not provide a good quality fingerprint. It is therefore reasonable to decrease the weight of the fingerprint matcher of such users, and increase the weights associated with the other matchers [1]. This process involves searching the weight space for a user, such that the error rate on the training set is minimized.

However, using a brute-force search to find these weights can be extremely computationally expensive and hence, significantly reducing its applicability to large-scale biometric systems.

In [94], the user-specific weight associated to each of the biometric matchers was determined using on the biometric menagerie concept (see Section 2.4.1). The so-called d-prime metric [1] was used as a measure of the degree of the separability between the genuine and impostor match scores in formulating the lambish metric for every pair of user and matcher (j, m). When the d-prime is small, user j is a lamb in the space of matcher m. The weight associated with this matcher is then decreased for user j. The main aim is to reduce the lambish characteristics of user j in the space of combined matchers [94], which results in the improved verification performance.

In [113], a MLP was used with $M + 1$ inputs to combine M matchers and a user-identity index. The user-specific weights and bias on individual matchers are automatically computed using the training data. It was demonstrated that exploiting user identity as an additional feature can improve the performance accuracy, albeit insignificantly [15].

In [30], four global and local learning and decision paradigms were investigated, making use of both user-specific and user-independent parameters. For the training data, 3% Gaussian noise with respect to the largest magnitude of the match scores was added to the ten genuine user samples in order to address the problem of imbalanced class distributions of data for the two classes. It was observed that making use of both the user-specific score fusion and user-specific threshold can achieve about 50% improvement in Equal Error Rate (EER) over the user-independent counterpart.

A potential weakness of user-specific score fusion is the need for a substantial amount of training match scores, especially those from the genuine user class, before it is able to provide better performance as compared to the user-independent counterpart [15], [31]. For instance, ten genuine samples were required in [30]. The d-prime in [94], which is based on the first-order and second-order moments, is not able to be reliably estimated with one or two genuine training samples.

2.4.4. Adapted User-Dependent Fusion

The training data scarcity problem, related to the user-specific genuine samples can be partially overcome by trading-off the general class knowledge and the local characteristics of the user at hand. This approach is called "Adapted User-Independent Fusion" (AUDF) [32], [33] to distinguish it from the conventional user-independent and user-specific techniques.

In [32], the SVM classifiers were trained using either the user-specific training set or the training set from a pool of the users. The resulting user-specific $f_{j,US}(\mathbf{s}_T)$ and user-independent score fusion schemes $f_{UI}(\mathbf{s}_T)$ were then consolidated to form the final decision for a given input vector of match scores $\mathbf{s}_T$, and the claimed identity j as

$$f_{j,Adapted}\,(\mathbf{s}_T) = \alpha f_{j,US}(\mathbf{s}_T) + (1-\alpha) f_{UI}(\mathbf{s}_T) \qquad (2.28)$$

where $\alpha \in [0,1]$ is a trade-off parameter. This approach was reported to achieve a relative improvement of 42% and 35% in terms of EER, compared to the user-independent and user-specific counterparts. Although the idea can be extended easily to train the fusion schemes with other two-class classifiers, it has to contend with the extremely imbalanced class distributions in the biometric training data as previously mentioned in Section 1.4

A similar idea was presented by the same authors in [33], which is based on Bayesian adaptation (instead of SVM) to exploit the user-independent data. Although a relative improvement of 80% was obtained as compared to user-specific score fusion, this approach, which assumes a single Gaussian component with a diagonal covariance matrix, may not be applicable to combine non-Gaussian match scores [31]. Hence, it will not be considered further in the context of this thesis.

2.5. Chapter Summary

Since the multimodal biometric matchers tend to be mutually independent of each other, combining their outcomes is believed to result in a significant improvement in performance. In this chapter, a comprehensive review on the techniques for the parallel combination of match scores of the individual biometric matchers has been provided. Although a large number of techniques have been developed in the literature, these are divided into two main categories, i.e., user-independent and user-specific processing. In user-independent processing, three main groups of techniques can be identified: density-based, transformation-based, and classifier-based schemes. They differ in their ability to estimate the a posteriori probabilities based on the vector of match scores. User-specific processing, as opposed to the user-independent counterparts, takes into account the label of the claimed identity for a given access request. It is motivated by the concept of the biometric menagerie, which suggested that users of a biometric system may have differing degrees of accuracy within the system. It has been also empirically reported to be better alternative with respect to the user-independent counterpart, when sufficient training data is available for individual users.

Based on this comprehensive review, it is worth noting that a common practice in multimodal biometrics is to view match score fusion as a pattern classification problem, where the two-class classifiers are employed to train the decision boundary between the genuine user and impostor classes. This decision boundary can be learned regardless of the claimed identity (user-independent processing) or for each user, enrolled in the system (user-specific processing). However, as mentioned in Section 1.4, most conventional two-class classifiers are inadequate when applied to the extremely imbalanced biometric data. It should be noted that such a problem becomes even more prominent when the classifiers are trained to learn the user-specific decision boundary due the limited availability of the genuine user samples. In Chapters 4 and 5, several techniques, based on the paradigm of one-class classifiers in order to tackle the class imbalance problem will be developed. These techniques will be evaluated and compared with the most representative state-of-the-art solutions, which were selected and discussed in this chapter, namely the likelihood ratio based score [90], sum of scores [95], two-class SVM [120] and the SVM based Adapted User-Dependent Fusion (AUDF) [32].

CHAPTER 3

PERFORMANCE MEASURES AND DATABASES

This chapter summarizes the guidelines on biometric performance evaluation methodologies. It also provides information on the two databases, namely XM2VTS [63] and BioSecure DS2 [61], [62] databases, used in the experiments, reported in this thesis. Although research in multimodal biometrics has entailed an enormous effort on data collection, most of the works in the literature have concentrated on treatment of the individual matchers, and downplayed the subject of multimodal fusion [63]. The XM2VTS and BioSecure DS2 are two publicly available databases, which contain biometric traits from a large number of users. They also come with well-defined experimental protocols such that different fusion algorithms can be benchmarked [61-63]. BioSecure DS2 contains a number of missing values due to the failure of the segmentation process or other stages of biometric authentication [61]. Since most score level fusion techniques cannot be invoked when score vectors are incomplete [80], a novel Robust Imputation based on Group Method of Data Handling (RIBG) is also developed in this section to handle to missing data problem.

This chapter is organized as follows: Section 3.1 summarizes the performance measures to assess accuracy and usability of biometric authentication systems. Sections 3.2 and 3.3 describe the XM2VTS and BioSecure DS2 databases. Section 3.4 discusses in detail the proposed RIBG. RIBG is also compared with other state-of-the-art imputation techniques, as applied in dealing with the missing values in the BioSecure DS2 database. Section 3.5 is dedicated to provide an overview of the sets of experiments, which will be conducted to assess the various biometric algorithms, developed in this PhD thesis.

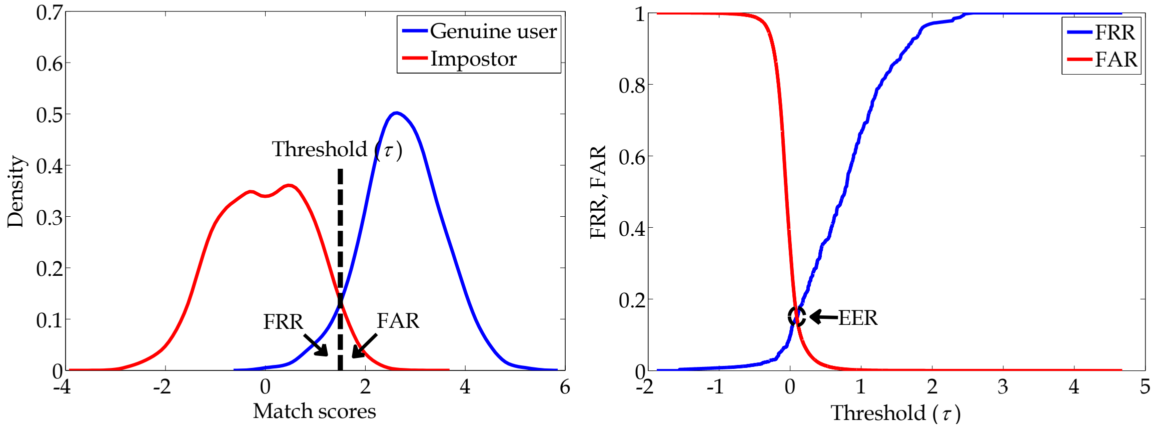
3.1. Performance Measures

Biometric systems as a pattern recognition system are inevitable to make errors [16]. This is due to the fact that the biometric systems, which rely on simplistic models of biometric data, generally fail to capture the richness of information in a biometric sample [1], [16], [157]. In addition, these systems are not able to correctly model the invariance relationship among different patterns from the same user, even when these patterns are captured under different representation conditions [67], [157]. Performance measures are used to determine the range of errors with the aim of predicting the real-world performance of the biometric systems. The typical metrics for verification accuracy are *False Rejection Rate* (FRR), *False Acceptance Rate* (FAR) and *Genuine Acceptance Rate* (GAR). Besides these fundamental performance measures, a few "compact" indicators, namely *Equal Error Rate* (EER), *Half Total Error Rate* (HTER), *relative change of EER* and *relative change of HTER* are defined for more accurate evaluation of the system performance. These metrics were employed in many biometric algorithm contests such as Fingerprint Verification Contests (FVCs) [155] and BioSecure DS2 Evaluation Campaign [156]. On the other hand, the authentication accuracy can also be graphically represented using *Detection Error Trade-off* (DET), *Receiver Operating Characteristics* (ROC) curves, or *Expected Performance Curve* (EPC). In this section, a brief overview on these performance measures is presented. Their formal definitions and discussion can be found in ISO/IEC IS 19795-1 [68], and 19785-2 [69]. In practice, it has been suggested that different tasks should explore distinct performance measures [67]. For instance, the compact indicators are best suited to summarize the results obtained on a large number of experiments, whereas DET and ROC should be considered when a few experiments are presented.

3.1.1. Quantitative Measures

It is well-known that the features, extracted from the same biometric trait of an individual can be significantly different from each other, leading to an *imperfect* match. This is caused by changes in sensing conditions, the manner a user interacts with a biometric device, and natural alternation of biometric trait due to sickness [1], [71].

The match score is a measure of similarity between the input and template biometric features. A match score is known as ***genuine*** if it is a result of matching two features of the same biometric trait of a user. It is known as ***impostor*** if it is obtained by comparing the features originating from different users. By definition, the *False Rejection Rate* (FRR) is the fraction of genuine match scores falling below a threshold τ. Similarly, the *False Acceptance Rate* (FAR) is the fraction of impostor match scores exceeding a threshold τ.



(a)

(b)

Figure 3.1. (a) FAR and FRR for a given threshold (τ), (b) FAR and FRR versus threshold (τ). Adapted from [70].

Since FRR and FAR are functions of the threshold τ, it is possible to express these two types of errors in terms of class conditional distribution of the match scores [1]. Assume that $p(s|genuine)$ and $p(s|impostor)$ are the probability density functions of the genuine and impostor match scores, respectively. For a given threshold τ, FRR and FAR are as follows

$$\mathrm{FRR}(\tau) = \int_{-\infty}^{\tau} p(s|genuine)ds \tag{3.1}$$

$$\mathrm{FAR}(\tau) = \int_{\tau}^{+\infty} p(s|impostor)ds \tag{3.2}$$

The FRR and FAR, defined in Equations (3.1) and (3.2) are shown in Figure 3.1(a). Apart from these error rates, another performance measure, known as *Genuine Acceptance Rate* (GAR) can be used to denote the correct decision in a biometric system

$$\mathrm{GAR}(\tau) = 1 - \mathrm{FRR}(\tau) \tag{3.3}$$

Figure 3.1(b) demonstrates the values of FRR and FAR versus the threshold τ. It is observed that if the value of τ is increased, FRR shows an increasing trend, while FAR shows a decreasing trend and vice versa. When the threshold τ is varied, there is a point where the two curves (FAR and FRR) cross each other, called the Equal Error Rate (EER) (see Figure 3.1(b)) [1]. A lower EER value indicates better verification accuracy. In practice, because of the discrete nature of FRR and FAR plots, the exact calculation of EER may be ambiguous, and hence, an operational procedure for computing EER must be followed. In this dissertation, the procedure for computing EER, described in [63], has been applied.

The optimal threshold can be selected using a threshold criterion, which has to be optimized on the training data set. A common threshold criterion is known as Weighted Error Rate (WER) [15], [63]:

$$\mathrm{WER}(\alpha, \tau) = \alpha\mathrm{FAR}(\tau) + (1 - \alpha)\mathrm{FRR}(\tau) \tag{3.4}$$

where $\alpha \in [0,1]$ provides a balance between FAR and FRR. Having chosen the optimal threshold τ^* using the WER threshold criterion, another performance measure, known as Half Total Error Rate (HTER), is defined as

$$\mathrm{HTER}(\tau^*) = \frac{\mathrm{FAR}(\tau^*) + \mathrm{FRR}(\tau^*)}{2} \tag{3.5}$$

The HTER is called *a priori* HTER if the threshold is selected prior to the measurement of the verification performance. Otherwise, it is called *a posteriori*. The a priori HTER is more realistic for use than the a posteriori, since the distribution of the test data set is usually different from the distribution of the training data set in most practical biometric systems [15], [63], [70]. Hence, in this thesis, only the a priori HTER is used.

Finally, the relative change of EER and the relative change of HTER are considered. These measures quantify the performance gain obtained from the specific fusion approach with respect to the baseline system

$$\text{rel. change of EER} = \frac{\text{EER}_{\text{new}} - \text{EER}_{\text{baseline}}}{\text{EER}_{\text{baseline}}} \tag{3.6}$$

$$\text{rel. change of HTER} = \frac{\text{HTER}_{\text{new}} - \text{HTER}_{\text{baseline}}}{\text{HTER}_{\text{baseline}}} \tag{3.7}$$

It is clear that a negative (positive) change of EER and HTER implies a performance improvement (decrease), whereas zero change implies no change in performance. The relative change of EER and the relative change of HTER are useful because they take into account the fact that when an error rate is already low, making more progress can be rather difficult [15], [63].

In general, the requirements related to the authentication accuracy are very much application-dependent. For instance, in forensic applications, it is the FRR that is of more concern than the FAR in minimising the risk of manually examining a large number of potential FAs. At the other extreme, a very low FAR is the most important factor in the highly secure access control applications, where the primary aim is to not let in any impostor [16]. As a consequence, the choice of quantitative measures for a particular biometric system is also application-dependent. EER has been the most commonly used quantitative measure in biometric literature. Nevertheless, in some cases, when the exact EER cannot be reliably estimated due to the limited availability of the genuine and impostor match scores, the a priori HTER should be considered as an alternative.

3.1.2. Graphical Measures

When presenting test results, the verification performance can be graphically represented using the Receiver Operating Characteristics (ROC), Detection Error Trade-off (DET), or Expected Performance Curve (EPC) measures.

Receiver Operating Characteristics (ROC)

The ROC curve [72], [73] is a traditional method for summarizing the performance of pattern-matching systems. It is threshold independent, allowing performance comparison of different biometric systems under similar conditions, or a single system under different conditions. The ROC visualizes the trade-off relationship between the GAR and the FAR. Figure 3.2(a) is a sample of the ROC curve. Varying the system's threshold moves the operating point along its ROC curve.

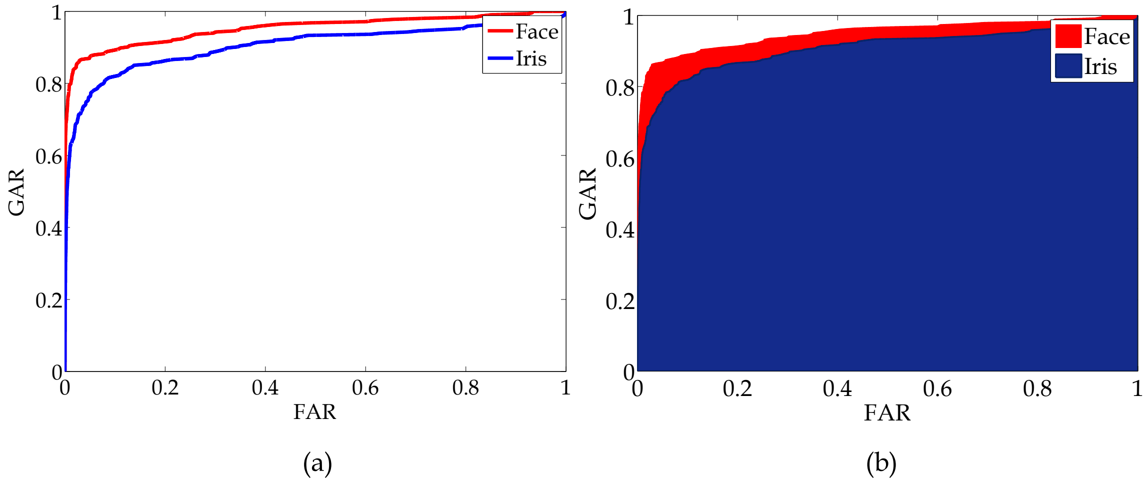


Figure 3.2. (a) ROC curves and (b) Area under the ROC curve (AUC) from the face and iris matchers of the BioSecure DS2 database.

The Area under the ROC curve (AUC) (see Figure 3.2(b)) corresponds to the probability of correctly identifying the positive case when presented with a randomly chosen pair of cases, where one case is positive and the other is negative. It can also be interpreted as the average GAR over the entire range of possible FAR. Hence, the AUC provides a single measure of a classifier's performance for the evaluation of which model is better on average. The AUC seems to be the most commonly used performance measure in imbalanced domains [34], [36], which is concerned with the performance of learning algorithms in the presence of underrepresented data and severe class distribution skews [34]. The AUC is shown in Figure 3.2(b). In [36], the AUC measure is computed as

$$\mathrm{AUC} = \frac{1 + \mathrm{GAR} - \mathrm{FAR}}{2} \tag{3.8}$$

Since FRR + GAR = 1, Equation (3.8) is rewritten as

$$\mathrm{AUC} = \frac{1 + (1 - \mathrm{FRR}) - \mathrm{FAR}}{2} \tag{3.9}$$

Hence

$$\mathrm{AUC} = \ 1 - \mathrm{HTER} \tag{3.10}$$

Generally, a lower HTER means a higher AUC and a better performance.

Detection Error Trade-off (DET)

The FAR and FRR at various values of τ can be summarized using the Detection Error Trade-off (DET) curve [74]. The DET, which is illustrated in Figure 3.3, plots the FRR against the FAR at various thresholds. It has a distinct advantage over the ROC curve for presenting system performance, where the trade-off of the two error types is involved. It also enables the user to select the threshold according to the system requirements.

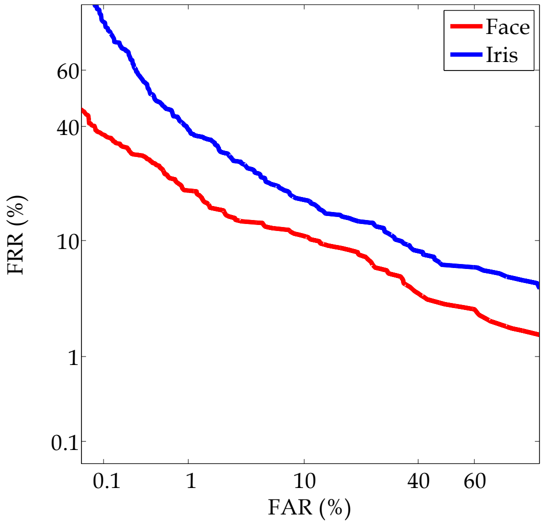


Figure 3.3. The DET curves from the face and iris matchers of the BioSecure DS2 database.

Expected Performance Curve (EPC)

The Expected Performance Curve (EPC) [15], [63], [70], depicted in Figure 3.4, plots the HTER versus the value of α, as defined in Equation (3.4). The HTER is measured using the a priori threshold τ^*, which is optimized for a given value of α as in Equation (3.4). The EPC curve can be interpreted in a similar manner to the DET curve, i.e., the lower the curve, the better performance. It is used for the comparison of two systems for a given cost, controlled by α.

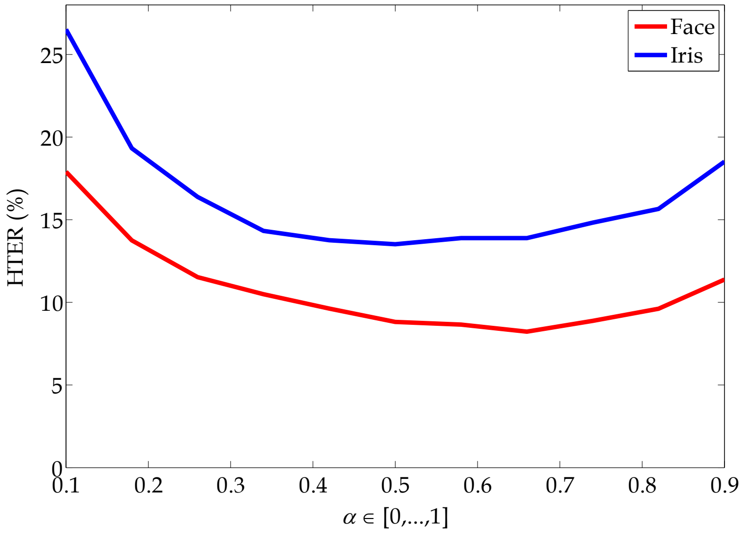


Figure 3.4. The EPC curves from the face and iris matchers of the BioSecure DS2 database.

The ROC and DET have an distinct advantage over the EPC because they can provide a direct view of the system performance at all operating points (decision threshold τ). A system designer may not know in advance the particular applications, where the biometric systems would be deployed. In such circumstance, it is advisable to report the system performance using the ROC and DET [67]. However, in this thesis, I decided not to use graphical measures due to their inherent difficulties to summarize results obtained on a large number of experiments. Instead, quantitative measures, such as EER, a priori HTER and relative change of EER and HTER will be employed for performance reporting.

3.2. XM2VTS Database

The XM2VTS database [63] contains synchronized face video and speech data of 295 persons. There are 200 genuine users in the database. The remaining 95 users serve as zero-effort impostors. The zero-effort impostors refer to impostors, who submit their biometric traits only to claim to be another user. There are 25 impostors in the development set and 70 in the evaluation set. The data was processed independently by 5 face matchers and 3 speech matchers. The match scores of these matchers were obtained by two experimental protocols, known as Lausanne Protocol I (LP1) and II (LP2) [63]. All 8 channels of data are used in LP1, while 5 out of 8 channels are used in LP2. Note that there are 8 biometric samples in the XM2VTS databases on a per user basis. As demonstrated in Table 3.1, they are used in the following decomposition: three samples are used to train the template in LP1 Train (and 4 in LP2 Train). Three samples are remaining in the LP1 Eva (and two in LP2 Eva) for learning the user-specific descriptions. Finally, in both protocols, two genuine accesses are dedicated to the testing in the LP Test. The class imbalance ratios in LP1 and LP2 are in the order of 66:1, and 100:1, respectively.

Table 3.1. The Lausanne Protocols of the XM2VTS database. LP Eva (LP Test) denotes the fusion protocols' development (evaluation) set. Adapted from [63].

Data sets	Lausanne Protocols		Fusion Protocols
	LP1	LP2	
LP Train genuine accesses	3	4	NIL
LP Eva genuine accesses	600 (3x200)	400 (2x200)	Development set
LP Eva impostor accesses	40,000 (25x8x200)		Development set
LP Test genuine accesses	400 (2x20)		Evaluation set
LP Test impostor accesses	112,000 (70x8x200)		Evaluation set

The detailed descriptions of the speech and face matchers are shown in Table 3.2. Each matcher is characterised by a feature representation and a classifier. All speech matchers are based on Gaussian Mixture Model (GMM) [54]. They differ by the nature of the employed feature representations, i.e., Spectral Sub-band Centroids (SSC) [75], Linear Frequency Cepstral Coefficients (LFCC) [76], and Phase-Auto-Correlation (PAC) [77]. On the other hand, the face matchers are based on the normalized face image concatenated with its RGB Histogram (abbreviated as FH) [63], and Discrete Cosine Transform (DCT) coefficients [78]. The DCT operates with two image block dimensions: i.e., small (denoted as

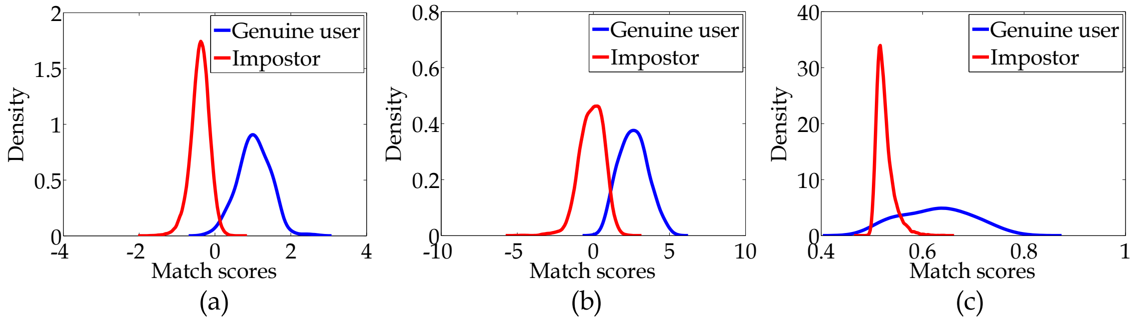
DCTs) and big (denoted as DCTb). Two types of classifiers are used, i.e., GMM and Multilayer Perceptron (MLP).

The inhomogeneity of the match scores of the different matchers raises a number of challenges [1], [23]. Figures 3.5(a) and 3.5(b) show the match score distributions of the speech and face matchers from the XM2VTS LP1 database. It is observed that these match scores are not on the same numerical scale, and follow different statistical distributions. Due to these reasons, it is essential to normalize the match scores of individual matchers into a common reference framework prior to combining them. The match scores of the XM2VTS database were normalized using the Z-norm (see Section 2.3.2), as recommended in [63].

Table 3.2. The 13 biometric matchers of the XM2VTS database, and their performance in terms of EER (%). Note that P1 (P2) indicates LP1 (LP2).

Labels	Modalities	Features	Classifiers	EER (%)
XM2VTS Lausanne Protocol I (LP1)				
P1:1	Face	FH	MLP	1.94
P1:2	Face	DCTs	GMM	4.22
P1:3	Face	DCTb	GMM	1.82
P1:4	Face	DCTs	MLP	3.53
P1:5	Face	DCTb	MLP	6.61
P1:6	Speech	LFCC	GMM	1.15
P1:7	Speech	PAC	GMM	6.62
P1:8	Speech	SSC	GMM	4.53
XM2VTS Lausanne Protocol II (LP2)				
P2:1	Face	FH	MLP	1.73
P2:2	Face	DCTb	GMM	0.55
P2:3	Speech	LFCC	GMM	1.37
P2:4	Speech	PAC	GMM	5.39
P2:5	Speech	SSC	GMM	3.33

As can be seen from Table 3.2, P1:6 achieves the lowest error rate with an EER of 1.15% among the biometric matchers in LP1. This is followed by P1:3 (1.82%), P1:1 (1.94%), P1:4 (3.53%), P1:2 (4.22%), and P1:8 (4.53%). The worst performance is produced by P1:5 and P1:7 with an EER of 6.61% and 6.62%, respectively. The biometric matchers in LP2 have lower error rates with respect to those in LP1, since more training samples can be used to extract the biometric features in LP2. For instance, as shown in Figures 3.5(b) and 3.5(c), the verification performance significantly depends on the overlapping zone between the impostor and genuine user match score distributions. P2:2 has a smaller overlapping region, and thus, it has better verification accuracy, as compared to P1:3, even when they are based on similar feature representation (i.e., DCTb) and classifier (GMM). Among the biometric matchers in LP2, P2:2 (EER of 0.55%) performs best, while the highest error rate is produced by P2:4 (EER of 5.39%).



(a) (b) (c)

Figure 3.5. Examples of the match score distributions: (a) speech matcher (P1:8), (b) face matcher (P1:3) and (c) face matcher (P2:2) of the XM2VTS database. Note that P1:3 and P2:2 are based on the same feature representation (DCTb), and classifier (GMM).

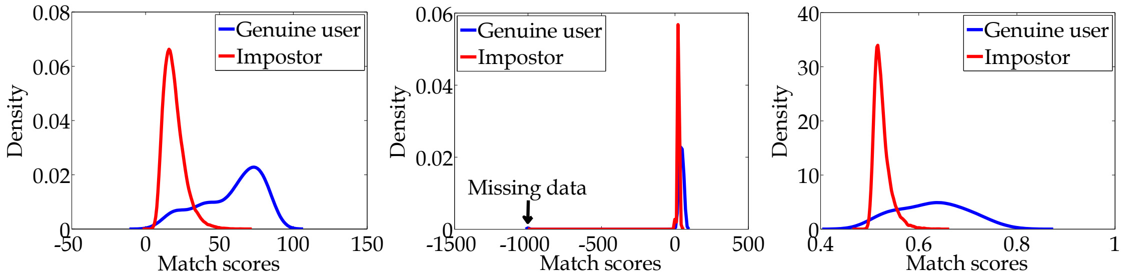
3.3. BioSecure DS2 Database

The BioSecure DS2 [61], [62] is the desktop scenario subset of the BioSecure database, which contains still face, 6 fingerprint (i.e., thumb, middle, and index fingers of both hands) and iris matchers from 333 persons. These 8 matchers are divided into two score data sets: the development (training) and the evaluation (test) sets. The development set is used for algorithm development, e.g., finding the optimal parameters of a fusion algorithm, while the evaluation set is used specifically for performance reporting. There are 51 genuine users in the development set and 156 in the evaluation set. The remaining 126 subjects are considered as an external population of users, who serve as zero-effort impostors. The zero-effort impostors refer to the impostors, who submit their biometric traits only to claim to be another user.

Each user contributes 4 samples. The first sample serves as a reference/template, while the second sample is used to obtain a genuine match score for training the user-specific descriptions. The remaining two samples are used to evaluate the performance of different fusion approaches. The BioSecure DS2 database has a class imbalance in the order of 524:1. Table 3.3 provides a summary of the BioSecure DS2 data. To overcome the challenges, related to the inhomogeneity of the match scores of individual matchers as shown in Figure 3.6, the F-norm [79] was applied, based on the recommendation of [61] to transform the match scores into a common domain.

Table 3.3. The experimental protocol for the BioSecure DS2 database. Adapted from [61].

Data sets		No. of match scores	
		Development set (51 persons)	Evaluation set (156 persons)
S1	Genuine accesses	1x51	1x156
	Impostor accesses	103x4x51	126x4x156
S2	Genuine accesses	2x51	2x156
	Impostor accesses	103x4x51	126x4x51



(a) (b) (c)

Figure 3.6. Examples of the match score distributions: (a) face matcher, (b) right thumb fingerprint matcher and (c) iris matcher of the BioSecure DS2 database. The value of the missing elements is denoted as "-999", as shown in (b).

It has to be noted that the BioSecure DS2 database contains a number of missing elements as a result of the failure of the segmentation process or other stages of biometric authentication [61]. In the event of any failure, the missing element is denoted as "**-999**" (see Figure 3.6(b)). The missing elements occur only on the evaluation data set, and not the training data set. The rationale for this is that the training data set is often better controlled. It is well-known that most techniques for match score level fusion have been implicitly designed for complete score vectors, assuming that all the match scores to be fused are available. They are not well equipped to handle the problem of incomplete score vectors [40].

3.4. Handling the Missing Information

It is important to distinguish between different patterns of missing data, since it can determine the method used to handle the problem [80-82]. In *Missing Not At Random* (MNAR), the probability that an entry is missing depends on both the observed data and the value of the missing data. In *Missing Completely At Random* (MCAR), the missing entry is neither dependent on the observed values nor the unobserved values in the data set. In *Missing At Random* (MAR), the probability that an entry is missing is a function of the observed data. It should be noted that it is not always possible to determine the reason behind missing information, and hence, it cannot be guaranteed that the occurrence of a missing observation is truly random [80].

3.4.1. Missing Data in Biometrics

In multimodal biometric systems, there are three common reasons for missing information [40], [80], [83], as follows:

1) *Temporary/permanent alternation of biometric traits*: a cough may temporarily change the voice of a person, while some drugs are known to permanently change fingerprint features. In such circumstances, the biometric traits should not be used because they may lead to false rejection;

2) *Desire to increase the authentication throughput*: for some applications, where access requests is much higher than expected, it is sensible to reduce the number of biometric traits needed in order to increase the throughput of the authentication process;

3) *Device malfunction*: biometric devices may not be operational as they have been worn over time.

The missing data problem in biometric fusion has received some attention. In [84], separate Support Vector Machine (SVM) classifiers were implemented for each possible combination of input modalities, and the appropriate SVM was selected based on the available data of the query sample. In [85], a Bayesian approach was proposed, which handles missing elements by assigning a fixed rank value to the marginal likelihood ratio, corresponding to the missing sample. While the former is obviously not scalable, and hence, inefficient, the latter is quite sensitive to the number of match scores available during training [80].

In [83], the authors developed a discriminative classifier capable of dealing with missing values using a kernel-based SVM with Neutral Point Substitution (NPS). This approach replaces the missing element by one, which is unbiased with regards to the classification, called a neutral point. Particularly, the NPS adopts a decision-agnostic approach with respect to the substitution of the missing elements by assuming that they do not contribute to any bias in the discrimination of one class from another. The estimation of the missing elements (i.e., neutral points) was implicitly incorporation within the SVM training framework. Although this technique was demonstrated to achieve a better generalization performance than the sum of scores, it is strongly related to the SVM, and may not be applicable to other fusion techniques [40].

Data Imputation, which substitutes the missing elements with predicted values, has the following distinct advantages: (1) it does not delete any score vectors, which may contain useful information for authentication, and (2) the treatment of missing data is independent of the succeeding learning algorithm [40], [64], [80], [92]. Among others, one of the most perspective imputation techniques seems to be RIBG (Robust Imputation Based on Group Method of Data Handling) [64]. The aim of RIBG is to provide an accurate prediction for the missing elements. It is known to be resistant to noise [64], which frequently occurs in biometric applications. In [64], RIBG demonstrated a lower error rate, as compared to other state-of-the-art techniques, such as mean imputation [82], median imputation [82], regression imputation [82], EM imputation [87], grey-based k-NN imputation [86], and Multiple imputation [81]. In [89], the capability of RIBG was exploited to fill the missing match scores of individual biometric matchers. Since the complete match score vectors were obtained, the likelihood ratio based score fusion [90] was utilized to deduce the evidence for making the final decision in a personal recognition system. The experiments were conducted on the three partitions of the NIST BSSR1 database [52], which demonstrate that the proposed approach is capable of achieving 95% rank-1 recognition rate even when the missing rate is set to 25%. This led to the investigation of RIBG in this research and adapting it to be more suitable in the context of biometric authentication.

3.4.2. RIBG Algorithm based Data Imputation

RIBG is based on the Group Method of Data Handling (GMDH) [88], which is applied in a great variety of areas for data mining and knowledge discovery, forecasting and systems modelling, optimization, and pattern recognition. GMDH is a self-organizing approach, which is substantially different from deductive methods used commonly for modelling. It has inductive nature and provides a mean of automatically finding interrelations in the data, selecting the optimal structure of a model and increasing the performance accuracy of existing techniques. GMDH guarantees that most unbiased models will be found even for real problems with noised and short data.

Without loss of generality, assume that $\mathbf{s}_T$ is the input vector of match scores, which contains missing elements, D_k is the initial complete set of match scores, $k \in \{G, I\}$ indicates the genuine user or impostor class, A_m is a variable, which corresponds to the scores of the m-th matcher, and M is the number of biometric matchers. The RIBG algorithm works as follows:

Input: incomplete vector of match scores $\mathbf{s}_T$

Output: complete vector of match scores $\hat{\mathbf{s}}_T$

1) Generate the data set, which includes $\mathbf{s}_T$:

$$D = \begin{cases} D_G \cup \mathbf{s}_T & if\ d(\mathbf{s}_T, \mu_G) \geq d(\mathbf{s}_T, \mu_I) \\ D_I \cup \mathbf{s}_T & otherwise \end{cases} \tag{3.11}$$

where $d(\mathbf{s}_T, \mu_k) = \sum_{h \in O_h} (s_{T,h} - \mu_{k,h})^2$, μ_k is the mean vector of the initial complete set D_k ($k \in \{G, I\}$), and O_h ={ h | the h-th element of the observation $\mathbf{s}_T$ is observed}.

2) Divide D into two disjoint data subsets $D = B \cup C$.

3) Replace each missing element $s_{T,i}$ in $\mathbf{s}_T$ by the mean of the observed match scores in D.

4) Select the variable A_i, which corresponds to the missing element $s_{T,i}$, as output variable ($\hat{y}_i = A_i$). All the remaining variables are treated as input variables $x_m = A_m$, where $m = 1, \dots, M$ and $m \neq i$, to enter the first layer of the GMDH network.

5) Exhaustively combine the input variables in pairs (x_m, x_n), where $1 \leq m, n \leq M$ and $m, n \neq i$ and generate model candidates from each combination using the following quadratic polynomial:

$$\hat{y}_i = c_0 + c_1 x_m + c_2 x_n + c_3 x_m x_n + c_4 x_m^2 + c_5 x_n^2 \tag{3.12}$$

where $c_0, c_1, \dots, c_5$ are parameters estimated by the Ordinary Least Squares (OLS). The OLS is the method for estimating unknown parameter using a linear regression model [54].

6) Evaluate the external criterion of each model using the RM criterion:

$$RM = \left\{ \left(\sum_{i \in B} (y_i - \hat{y}_i^C)^2 + \sum_{i \in C} (y_i - \hat{y}_i^B)^2 \right) \right\} + \sum_{i \in B \cup C} (\hat{y}_i^B - \hat{y}_i^C)^2 \tag{3.13}$$

where y_i is of the actual value and $\hat{y}_i^B$ and $\hat{y}_i^C$ are the predicted values, corresponding to the model, constructed on dataset B and C, respectively.

7) Record the minimum of the external criterion from the current layer, and the best model with lower criterion values. Their outputs are implied as new input variables for the second layer of the GMDH network.

8) Repeat steps 5-7 to produce model candidates for the second layer, the third layer, etc. until the lowest value of the external criterion at the current layer

is greater than that in the previous layer, in which the model with the minimum external criterion is selected as the final optimal complex model.

9) Use the estimate $\hat{A}_i$ of the optimal complex model to update the missing variable A_i.

10) Repeat steps 4-9 until the change of missing element estimate A_i becomes smaller than a predefined threshold or the maximum number of iterations is reached. The current value of A_i is assigned to the corresponding missing element in $\hat{\mathbf{s}}_T$.

11) Repeat steps 4-10 to predict and update the remaining missing elements.

3.4.3. Comparison of Data Imputation Techniques

In Table 3.4, the robustness of RIBG is illustrated. RIBG is also compared with other state-of-the-art techniques, including mean, median, and k-NN imputations, when applied in handling the missing data in the BioSecure DS2 database. It is observed that the face matcher does not contain any missing elements (see Figure 3.6(a)). Hence, in such case, no performance gains are observed. RIBG consistently produces the best verification accuracy, while the highest error rates are achieved when no imputation technique was used.

Table 3.4. EER (%) of different imputation techniques, as applied to handle the missing data in BioSecure DS2 database.

Modalities	No Imp.	Mean	Median	k-NN	RIBG
Face	**10.67**	**10.67**	**10.67**	**10.67**	**10.67**
Right thumb finger	16.40	16.18	16.18	15.99	**15.80**
Right index finger	7.99	7.75	7.75	7.42	**7.36**
Right middle finger	12.50	12.27	12.27	12.22	**11.91**
Left thumb finger	16.48	16.37	16.37	16.00	**15.89**
Left index finger	10.34	10.18	10.18	9.81	**9.73**
Left middle finger	14.51	14.45	14.45	14.01	**13.89**
Iris	**14.89**	15.49	15.39	14.93	14.90

k-NN imputation can achieve a better performance, with respect to the median and mean imputation techniques since it is capable of selecting the most suitable candidates for the estimation process. However, k-NN imputation requires a careful choice of the number of nearest neighbours k. As it was observed in [40], a large value of k may include samples, which are significant different from the sample containing missing elements, while a small value of k may lead to deterioration in the classifier's performance after imputation due to the overemphasis on a few dominant samples. On the contrary, RIBG does not require such parameter, making it amendable to immediate applications.

As it can be seen from Table 3.4, among the biometric modalities, the right index finger has the highest authentication accuracy with an EER of 7.36%. This is followed by the left index finger (9.73%), face (10.67%), right middle (11.91%), left middle finger (13.89%), iris (14.90%), and right thumb finger (15.80%). The worst performance (i.e., 15.89%) is achieved by the left thumb finger.

3.5. Experimental Protocol

As previously mentioned, this PhD thesis aims to exploit the paradigm of one-class classifiers to address the highly imbalanced class distribution of biometric data sets, and thus, advancing the classification performance of multimodal biometric fusion (i.e., using different biometric traits). Using the data from the XM2VTS and BioSecure DS2 databases, I conduct four sets of experiments in order to assess the various biometric algorithms that are investigated and developed in this PhD dissertation. Table 3.5 summarizes the properties of these sets, i.e., the class imbalance ratio, number of modalities (# Modalities), which were combined for the decision inference, number of combination possibilities (# Combinations) and number of genuine samples per user (# Gen. Samples/user).

Specifically, in the first three sets, two biometric matchers of the BioSecure DS2, XM2VTS LP1 and LP2 were exhaustively paired in the context of multimodal fusion. Hence, Experiment (1) involves **13** combination possibilities (*1 face matcher* × *6 fingerprint matchers* + *1 iris matcher* × *6 fingerprint matchers* + *1 face matcher* × *1 iris matcher*), while those numbers of Experiments (2) and (3) are **15** (*5 face matchers* × *3 speech matchers*) and **6** (*2 face matchers* × *3 speech matchers*), respectively. It should be noted that each of the databases has a different class imbalance ratio (i.e., 524:1 in BioSecure DS2, 66:1 in XM2VTS LP1 and 100:1 in LP2), and contains different number of genuine match scores per user (i.e., 1 in BioSecure DS2, 3 in XM2VTS LP1 and 2 in LP2). The objective of conducting these first three sets is to determine whether one-class classifiers are able to achieve a better performance compared to that of the two-class methods, such as two-class SVM when dealing with extremely imbalanced biometric data. The possible effect of the number of genuine match scores per user on the authentication accuracy is also evaluated.

In the remaining set, three out of the eight biometric matchers of the BioSecure DS2 database were combined at the same time. Hence, **6** combination possibilities (*1 face matcher* × *6 fingerprint matchers* × *1 iris matcher*) were obtained to analyse the impact of having additional pieces of evidence from another biometric on the performance of the various biometric algorithms.

Table 3.5. Summary of the characteristics of different sets of experiments in this thesis. These characteristics include the class imbalance ratio, number of modalities (# Modalities), number of combination possibilities (# Combinations) and number of genuine samples per user (# Gen. Samples/user).

Characteristics	Experiment (1)	Experiment (2)	Experiment (3)	Experiment (4)
Database	BioSecure DS2	XM2VTS LP1	XM2VTS LP2	BioSecure DS2
Class Imbalance Ratio	524:1	66:1	100:1	524:1
# Modalities	2	2	2	3
# Combinations	13	15	6	6
# Gen. Samples/user	1	2	3	1

These sets of experiments were conducted on an Intel(R) 2.00 GHz, 3.25 GB of RAM and MATLAB platform using the functions provided by the Data Description Toolbox 1.9.1 [163], and the MATLAB SVM and Kernel Methods Toolbox [164]. The performance in terms of EER, HTER, relative change of EER and HTER was evaluated using the tools, developed in [63].

3.6. Chapter Summary

In this chapter, several popular quantitative and graphical performance measures for determining system performance have been presented. Some of these measures, such as EER, HTER, the relative change of EER and the relative change of HTER will be used in the subsequent chapters to allow a comparison of the fusion techniques, investigated and developed. I also provided an overview of the main multimodal biometric databases, namely BioSecure DS2 and XM2VTS databases, and their corresponding sets of experiments, which will be conducted in this thesis for the performance evaluation. Some data imputation techniques were also introduced to deal with the missing information in the BioSecure DS2 database. Among others, RIBG was observed to achieve the best performance. Moreover, it does not require any parameter fine-tuning, thus making it amenable to immediate applications.

CHAPTER 4

ONE-CLASS LEARNING

One-class classifiers aim to recognize instances of a concept by using mainly or only a single class of samples [34]. They are known to be naturally quite robust to the class imbalance problem [41], [45-50]. This chapter aims to exploit the one-class classifiers to advance the classification performance of extremely imbalanced biometric data sets. The main contributions of this work consist of the followings: (1) Designing user-specific score fusion scheme based on the paradigm of one-class methods. An extensive empirical evaluation is also presented to highlight advantages of this scheme over the user-independent counterparts; (2) Exhaustively analysing several representative one-class classifiers, namely Gaussian Mixture Model, k-Nearest Neighbour, K-means clustering, and Support Vector Data Description. The target is to determine whether their performance outranks the performance of the standard two-class methods, such as two-class SVM when dealing with extremely imbalanced biometric data. It also provides an indication, which can drive a proper choice of the classifier to be used in different biometric applications.

This chapter is organized as follows: Section 4.1 describes the various one-class classifiers considered in this work, pointing out their characteristics, advantages and disadvantages. Section 4.2 explains how these classifiers can be used for rendering the multimodal biometric decision. The user-specific score fusion scheme is also presented in this section. Sections 4.3 and 4.4 present extensive experiments using data from the BioSecure DS2 and XM2VTS databases, and discuss the results.

4.1. One-class Classification Methods

One-class classification has been successfully applied in a wide variety of application domains, including Bioinformatics [121], Information Retrieval [122], and Face Recognition applications [123]. Its main objective is to learn a description around the target set of samples, and to detect which objects resemble this training set [44]. Unlike conventional two-class classification, where the decision boundary is supported from samples of both classes, in one-class classification only samples of one-class are available. The objects from this class are called the target objects, while all other objects are known as the outliers.

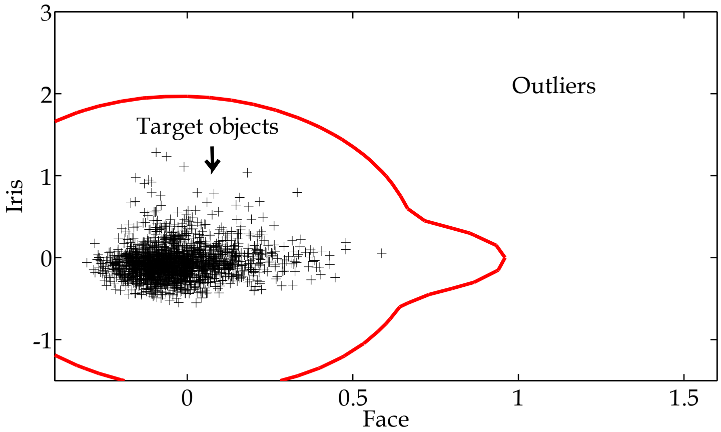


Figure 4.1. Example of the description, learned around the impostor class using match scores from the face and iris matchers of the BioSecure DS2 database. Because the match scores from these matchers are in the range $[0, \dots, 1]$ *and* $[0, \dots, 100]$, *they are normalized using the F-norm [79], as suggested in [61] to be better aligned and separated. A one-class classifier is shown, which distinguishes impostor samples from all other possible outliers. Adapted from [44].*

Table 4.1. Summary of one-class classification methods. N *is the number of training samples. Adapted from [127].*

Methods	Robustness to outliers	Number of free parameters	Computation
GMM	No	1	$O(1)$
k-NN	No	1	$O(N)$
K-means	No	0	$O(1)$
SVDD	Yes	1	$O(N)$

In the context of multimodal biometrics, ***a match score vector can be treated as the target object (sample)***, which is then fed into a trained classifier, whereas the target class can be either the genuine user or impostor class. As illustrated in

Figure 4.1, a one classifier is applied to learn the description around the impostor class by using samples from the face and iris matchers of the BioSecure DS2 database. No samples of the genuine user class are needed for the training of the description. The red solid line shows that the one-class classifier is able to distinguish between impostor class samples and the outliers (i.e., genuine user samples).

There are number of methods for one-class classification, such as Gaussian Mixture Model [54], Support Vector Data Description [44], [56], k-Nearest Neighbour [124], etc. They differ in their ability to exploit or cope with different characteristics of the data, which include the scaling of features in the data, the convexity of the data distribution, etc. In [44], these one-class classifiers were compared using the following characteristics:

1) **Robustness to outliers** refers to the ability of rejecting outliers, when the training data is already contaminated by some of those.
2) **Free parameters and ease of configuration** refer to the number of parameters to be defined by the user. It is obvious that when a large number of free parameters is required, finding a good set of parameters may be hard and time consuming.
3) **Computation** is defined as the computational time of a classifier, when evaluating a single test object.

In the rest of this chapter, I present only some representative one-class classifiers, such that Gaussian Mixture Model, k-Nearest Neighbour (k-NN), K-means clustering (K-means), and Support Vector Data Description (SVDD), which can be utilized to render the biometric decision. The characteristics of these classifiers are summarized in Table 4.1.

It must be noted that all the one-class classifiers require the fraction rejection $f_T \in [0, \dots, 1]$, which has to be specified by the user. By definition, this parameter controls the amount of target samples, which the classifier can reject during training. It allows the classifier to be more robust against outliers in the training data set [44]. For instance, when $f_T = 0.1$, a decision threshold on the classifier's outcomes is optimized such that the 10% most dissimilar target data (possibly containing outliers/noisy samples) are rejected. In practice, it is challenging to determine whether a given sample represents the general class characteristics or is merely attributed to noise. It is therefore essential for the different values of f_T to be tested. The fraction rejection is not a free parameter for a particular one-class method, thus, it is omitted in Table 4.1. Without loss of generality, assume that w_T is the target class, $\mathbf{s}_T$ is a test object, $\|\mathbf{s}\|$ is the Euclidean length of vector $\mathbf{s}$, and $p(\mathbf{s}_T, \mu_j, \Sigma_j)$ is the Gaussian distribution characterised by mean μ_j and covariance matrix Σ_j. These notations will be used throughout this chapter.

4.1.1. Gaussian Mixture Model

The Gaussian Mixture Model (GMM) is often employed in multimodal biometric systems due to its capability of forming smooth approximations to arbitrary shape densities of the match scores. The Gaussian Mixture Model (GMM) is a linear combination of Gaussian distributions as given by:

$$p_{GMM}(\mathbf{s}_T | w_T) = \frac{1}{N_{GMM}} \sum_{j=1}^{N_{GMM}} \alpha_j p(\mathbf{s}_T, \mu_j, \Sigma_j) \tag{4.1}$$

where α_j are the mixing coefficients, N_{GMM} is the number of mixture components, which is used to model the match score distributions of either the genuine user class or the impostor class. When the number of mixture components N_{GMM} is known, μ_j and Σ_j of the individual mixture components can be estimated using the Expectation Maximization (EM) algorithm [54]. As already mentioned, selecting the appropriate number of components is a vital requirement for the success of the mixture. With too few components, a mixture may not flexible enough to approximate the true underlying densities, while a mixture with too many components may overfit the data [90]. In general, this number can be selected in many different ways. Penalized likelihood approaches, such as Bayesian inference criterion (BIC) [159], Minimum description length (MDL)

[160] are typical derived from approximations based on asymptotic arguments as the training data size approaches ∞ [168]. The fully Bayesian approach treats the number of component as a parameter and obtains a posterior distribution on N_{GMM}, given the data. The models and the posterior can be then estimated analytically or via Markov Chain Monte Carlo sampling [161]. Nevertheless, none of the above approaches is able to achieve the results, which is independent of the quality of the underlying approximation or simulations [162], [168]. Cross validation [54] is another tool for automatically determining the true value of N_{GMM}. It is simple and straightforward in the sense that the model is directly judged on their out-of-sample predictive ability [168]. Due to this reason, it has been applicable to a wide variety of practical problems. In this PhD thesis, the value of N_{GMM} is also determined through cross validation and grid search.

4.1.2. k-Nearest Neighbour

When limited training data is available, k-Nearest Neighbour (k-NN) [44], [124] is preferred, compared to the GMM. In general, k-NN can be derived from a local density estimation, which can be defined as:

$$p_{k-NN}(\mathbf{s}_T|w_T) = \frac{k/N}{V_k(\|\mathbf{s}_T - NN_k(\mathbf{s}_T)\|)} \tag{4.2}$$

where N is the number of training objects, $NN_k(\mathbf{s}_T)$ is the k^{th} Nearest Neighbour of the test pattern $\mathbf{s}_T$ in the training set, and V_k is the volume surrounding $\mathbf{s}_T$. This algorithm requires the user to define the number of nearest neighbours k, which heavily relies on the distance between objects, as it is sensitive to the scaling of the match scores, provided by multiple matchers [44].

There is almost no computation cost for training k-NN, but testing is expensive. Such classifiers require all the training samples to be stored and, during testing, distances to all training samples have to be calculated and sorted [44], [54], [125]. This effect can be offset by constructing tree-based search structures, which allow for the nearest neighbours to be found efficiently without carrying out an exhaustive search on the training data [54].

4.1.3. K-means Clustering

K-means clustering (K-means) [54] is one of the simplest methods for one-class classification, which assumes that the data $\mathbf{s}_i$ is clustered and described by a set of prototype objects μ_k. The position of these prototype objects is determined by minimizing the following error:

$$\varepsilon_{K-means} = \sum_i \min_k \|\mathbf{s}_i - \mu_k\|^2 \tag{4.3}$$

The classification of a test object $\mathbf{s}_T$ is based on its distance to the nearest prototype object, defined as:

$$d_{K-means}\left(\mathbf{s}_T | w_T\right) = \min_k \|\mathbf{s}_T - \mu_k\|^2 \tag{4.4}$$

K-means uses the squared Euclidean distance as the measure of similarity between the test and prototype objects. Not only does this limit the scaling of data, which can be considered, however it may impact on the robustness of the cluster means estimation, with regards to outliers [54]. Note that the error in Equation (4.3) can be then minimized using a so-called batch algorithm, which is comparable to the EM algorithm of GMM [44]. A distinct advantage of K-means over other one-class classifiers is that it has no free parameters, which have to be specified by the user apart from the fraction rejection, and thus, making it more amenable to immediate biometric applications.

4.1.4. Support Vector Data Description

Support Vector Data Description (SVDD) [44] aims to directly fit a closed boundary around the target data set, without estimating a complete probability density. Thus, such a method has the ability to obtain the data boundary from a limited data set. The boundary can then be described by a few training objects, known as, the support vectors. In order to obtain more flexible data descriptions, SVDD replaces the normal inner products by kernel functions $K(\mathbf{s}_i, \mathbf{s}_j)$. Polynomial and Radial Basis Functions (RBF) are the most common kernel functions used in practice. Since SVDD is a kernel-based method, it is particularly suitable to deal with hyper-dimensional feature spaces. The classification of a test object $\mathbf{s}_T$ is then based on its distance from the centre of the hypersphere, which is calculated as follows:

$$\begin{aligned} d_{SVDD}(\mathbf{s}_T|w_T) &= \|\mathbf{s}_T - \mathbf{a}\|^2 \\ &= K(\mathbf{s}_T, \mathbf{s}_T) - 2\sum_i \alpha_i K(\mathbf{s}_T, \mathbf{s}_i) + \sum_{i,j} \alpha_i \alpha_j K(\mathbf{s}_i, \mathbf{s}_j) \end{aligned} \tag{4.5}$$

where $\mathbf{a} = \sum_i \alpha_i \mathbf{s}_i$ is the centre of the sphere, $\alpha_i \geq 0$ are the Lagrange multipliers, which can be determined as the solution of a Quadratic Programming problem [44], [56]. Obviously, the centre of the sphere is a linear combination of the samples $\mathbf{s}_i$ with weights α_i. Since a large fraction of the weights becomes 0, the description can be characterised by a few samples with positive weights. These samples are called the support vectors of the description. It was observed that when an insufficient number of samples is available, the number of support vectors remains high, indicating that more data is necessary. Hence, for very small sample sizes (less than ten), the SVDD breaks down due to its requirement for support vectors [44], [56].

In general, SVDD is different from the ν-SVM, proposed in [126], for one-class classification. While the former computes a closed hypersphere around the data, the latter estimates the largest margin hyperplane, used to separate the data and the origin of the space, where the data resides [45]. However, it has been shown that when the RBF is used as the kernel function, these methods are equivalent and produce similar performances [44], [126]. The width of the RBF kernel is the single free parameter, which is used in SVDD and can be also found by using cross validation [54].

4.2. Multimodal Fusion via One-class Classifiers

It has been demonstrated that most conventional two-class classifiers are inadequate when applied to authenticate the identity claim in multimodal biometric systems, where very often the training data is extremely imbalanced. One-class classifiers on the other hand are capable of learning the description using samples exclusively from one-class. They are therefore known to be naturally quite robust to the class imbalance problem [44], [45]. In [46], [47], the authors suggested that one-class classifiers are particularly useful in handling extremely imbalanced data sets with high feature space dimensionality, while two-class classifiers are more suitable for moderately imbalanced data sets. In [47], it was reported that with 5% or lower small class data, the performance of the ν-SVM surpassed that of the two-class SVM.

In [48], a variety of one-class classifiers, including GMM, k-NN, K-means and SVDD, and two-class classifiers, including Logistic Regression, Naïve Bayes, MLP and two-class SVM, was evaluated on a selection of credit score datasets as the class imbalance is manipulated. It is important to note that one-class classifiers offer a viable solution to the low-default portfolio problem, when the minority class constitutes approximately 4% or less of the data (i.e., the imbalance rate is severe). In [49], it was observed that the auto-encoder may be superior to the MLP neural network under certain conditions, such as multimodal domains. The pros and cons of the auto-encoder, as compared to other data sampling techniques were demonstrated in [50].

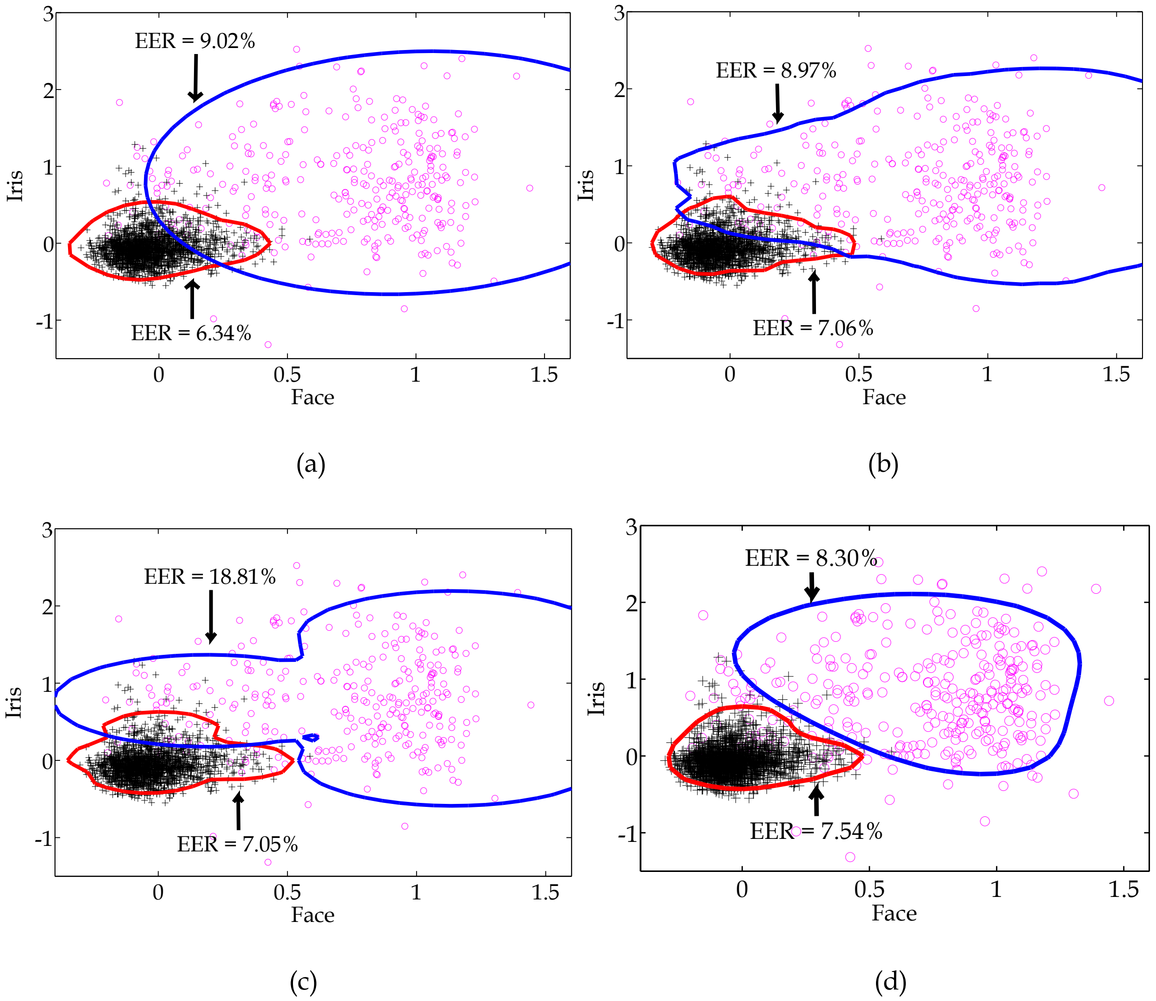
In [51], a user-independent approach based on the ν-SVM was developed, which aims to classify a given test match score vector based on the description learned around the data, containing all impostor match score vectors from a number of different users. It was observed to be comparable to the two-class SVM, and surpass other conventional classifier combination rules, including the sum of scores in the experiments, carried out on the NIST BSSR1 [52] and MCYT databases [53].

One-class classifiers have not been sufficiently exploited in multimodal biometric authentication. The remainder of this chapter aims to provide a thorough investigation on the various one-class classifiers and directly extend the method in [51]. To this end, I propose user-independent and user-specific score fusion schemes based on the paradigm of these classifiers. Both of the two methods are capable of using match score vectors from either the genuine user or impostor class. User-specific score fusion scheme, as opposed to the user-independent counterpart takes into account the claimed identity and learn a different description for each of the users, enrolled in the systems. It should be noted that all the reported works in biometric literature have been focused on designing user-specific descriptions using the two-class methods [30], [32]. No attempt has been made to exploit the one-class classifiers to learn these descriptions.

As already mentioned, essentially, one-class classifiers can provide two types of outcomes, i.e., (1) a distance or (2) a probability estimate of the test object to the target class. K-means and SVDD are examples of classifiers belonging to the first group, while GMM and k-NN are examples of classifiers of the second group. In the context of multimodal biometrics, the combined match score can be directly defined using the outcomes of the one-class classifiers with the assumption that the test object $\mathbf{s}_T$ is the test match score vector. This section discusses in detail the user-independent and user-specific score fusion schemes. Illustrative examples are also presented to highlight the advantages of the user-specific approach over the user-independent counterpart.

4.2.1. User-independent Score Fusion

In the user-independent score fusion, a single description is learned around the target class w_k using the match score vectors from a number of different users. Since the target class can be either impostor or genuine user, one-class classifiers applied separately to each of these classes can produce two different descriptions. Figure 4.2 shows graphical representations of the decision boundaries, learned by GMM, k-NN, K-means and SVDD. Specifically, the continuous red lines correspond to the descriptions of the impostor class, while the continuous blue lines correspond to those of the genuine user class. Obviously, each of the one-class classifiers has a different ability to learn the characteristics of the training data, leading to a difference in their description, and the associated error rates.



(a)

(b)

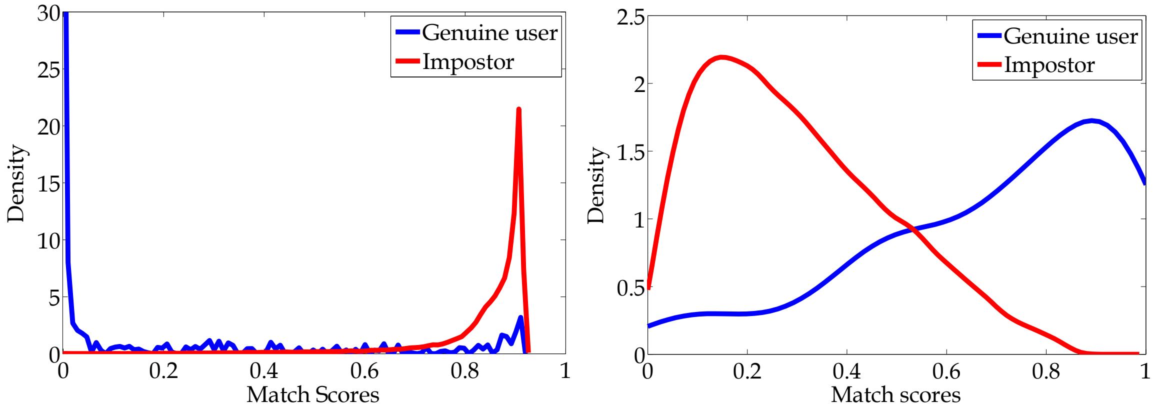
(c)

(d)

Figure 4.2. Examples of the user-independent descriptions around the impostor class (continuous red lines) and genuine user class (continuous blue lines), learned by (a) GMM, (b) k-NN, (c) K-means, and (d) SVDD using the match scores from face and iris matchers of the BioSecure DS2 database. The match scores of these matchers were normalized using the F-norm [79].

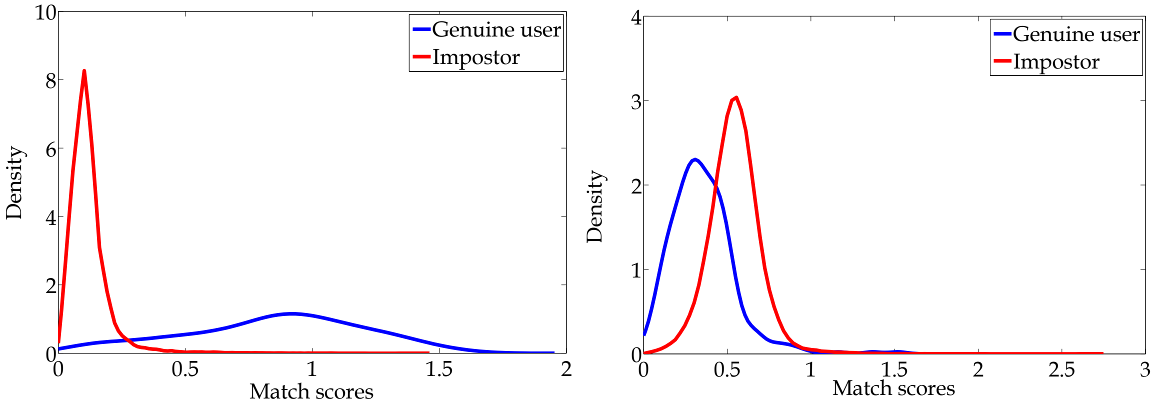
Since samples of the two classes are scattered into several small regions, the decision boundary encloses a large empty area, which could augment the probability of accepting outliers [51]. This situation has been referred to as the problem of *within-class sub-concepts* in the literature [34], [37], [115], which is observed to occur in any biometric system. The rationale for this is that some users may particularly be vulnerable to impersonation while others may have trouble authenticating. The manner that a user interacts with devices also leads to a difference in the biometric signals, which are captured at different time instances. It is therefore expected that there will be some impostor matches with high scores, and there will be some genuine matches with low scores [109].

Because of the vast representation of the impostor class, the occurrence of such problem is more notable in the genuine user class [34]. As illustrated in Figure 4.2, a large number of impostor match score vectors are accepted by the description, learned around the genuine user class, leading to significantly higher error rates. This will be further elaborated in Section 4.3.



(a) (b)

Figure 4.3. The distributions of the combined match scores, generated by GMM when the descriptions are learned around (a) impostor, and (b) genuine user classes using the match scores from face and iris matchers of the BioSecure DS2 database.



(a) (b)

Figure 4.4. The distributions of the combined match scores, generated by K-means when the descriptions are learned around (a) impostor, and (b) genuine user classes using the match scores from face and iris matchers of the BioSecure DS2 database.

In Figures 4.3 and 4.4, examples of the distributions of the outcomes of GMM and K-means are provided in both cases, i.e., when the descriptions are learned around the impostor and genuine user classes. The solid red line corresponds to the distribution of the outcomes, obtained when classifying the test samples from the impostor class. On the other hand, the solid blue line corresponds to the distribution of the outcomes for those samples from the genuine user class. The verification performance depends much on the overlapping zone between the distributions of these outcomes. It is clear that the impostor class is more reliable source of information. Its corresponding distributions have a much smaller overlapping zone (see Figures 4.3(a) and 4.4(a)) as compared to that of the distributions, which are generated when the description is learned around the genuine user class (see Figures 4.3(b) and 4.4(b)). The distributions of the outcomes, related to the SVDD and k-NN are not demonstrated here, since they show a similar trend to those of K-means and GMM.

In general, the combined match scores can be directly obtained based on the outcomes of the one-class classifiers. It should be noted the target class can be either genuine user or impostor class. Due to the choice of the target class, the combined match scores, even when generated by the same one-class method, are inhomogeneous. Obviously, when the target class is the impostor w_I, user-independent score fusion, as illustrated in Figures 4.3(a) and 4.4(a), produces a low probability or high distance for a test object $\mathbf{s}_T$ of a genuine user, while a higher probability or lower distance is obtained when the test object belongs to an impostor. Thus, the distance in this case is a similarity measure, while the probability is a dissimilarity measure and has to be transformed into a similarity one. To do this, the combined match scores $s_{COM,I}(\mathbf{s}_T)$ can be defined as:

$$s_{COM,I}(\mathbf{s}_T) = \begin{cases} -p(\mathbf{s}_T|w_I), & if\ the\ measure\ is\ a\ probability \\ d(\mathbf{s}_T|w_I), & if\ the\ measure\ is\ a\ distance \end{cases} \tag{4.6}$$

In contrast, when the target class is the genuine user class w_G (see Figures 4.3(b) and 4.4(b)), the probability becomes a similarity measure, while the distance turns out to be a dissimilarity measure. The combined match scores $s_{COM,G}(\mathbf{s}_T)$ are as follows:

$$s_{COM,G}(\mathbf{s}_T) = \begin{cases} p(\mathbf{s}_T|w_G), & if\ the\ measure\ is\ a\ probability \\ -d(\mathbf{s}_T|w_G), & if\ the\ measure\ is\ a\ distance \end{cases} \tag{4.7}$$

It should be noted that the former case (Equation (4.6)) was considered in [51] using the ν-SVM [126]. I directly extend the former to obtain the latter case (Equation (4.7)). Since the ν-SVM is equivalent to the SVDD when RBF is used as the kernel function (see Section 4.1.4), such a classifier is not further considered in this PhD thesis. In addition to SVDD, a number of one-class classifiers, such as GMM, k-NN, and K-means are evaluated in both user-independent scenarios.

4.2.2. User-specific Score Fusion

In user-specific score fusion, different descriptions around the target class are determined for each user enrolled in the system using exclusively their corresponding match score vectors. Its basic motivation is to address biometric menagerie effect, and hence, enhancing the verification performance. It also differs from the conventional user-specific score fusion since its description is learned using the one-class classifiers (see Section 2.4.3 for more details). Similarly to user-independent score fusion, the classifier's outcomes can be either similarity or dissimilarity measure depending on the choice of the target class, which can be the genuine user ($w_{j,G}$) or impostor ($w_{j,I}$). Subsequently, the combined match score, related to the user j is defined as:

$$s_{COM,j,I}(\mathbf{s}_T) = \begin{cases} -p(\mathbf{s}_T|w_{j,I}), & if\ the\ measure\ is\ a\ probability \\ d(\mathbf{s}_T|w_{j,I}), & if\ the\ measure\ is\ a\ distance \end{cases} \tag{4.8}$$

if the target class is the impostor, and:

$$s_{COM,j,G}(\mathbf{s}_T) = \begin{cases} p(\mathbf{s}_T|w_{j,G}), & if\ the\ measure\ is\ a\ probability \\ -d(\mathbf{s}_T|w_{j,G}), & if\ the\ measure\ is\ a\ distance \end{cases} \tag{4.9}$$

if the target class is the genuine user. The latter case (Equation (4.9)) is impractical to achieve due to the limited availability of genuine match scores per user. GMM, as mentioned in Section 4.1.1 requires a large number of target samples to converge to the true density, while SVDD suffers from the lower bound on the number of support vectors required for its description. In [44], the authors observed that GMM and SVDD do not work at all with a sample size of less than ten. Due to this reason, I only consider user-specific score fusion in

Equation (4.8), when the descriptions are learned using the impostor class samples.

As the impostor is the target class, user-specific score fusion can be generally considered as a better alternative, relative to the user-independent counterpart for the following two reasons.

- Firstly, it is observed that user-specific score fusion is much faster in testing time, since fewer samples are used to construct the classifier. This is particularly true in the case of k-NN, where, during testing, the distances of the test object from all training samples have to be calculated and sorted. A similar observation can also be made when SVDD is used as a classifier, where the reduction in the number of training samples leads to a smaller number of support vectors, and associated computational savings. More results, supporting for this observation can be found in [127].

- Next, it was observed that one-class classifiers fail to achieve a good verification performance when the impostor class is composed of various sub-clusters. As already mentioned, this problem is known as the problem of *within-class sub-concepts*, which becomes even more prominent when the biometric decision is based on user-independent score fusion. A straightforward explanation for this is that such method trains the classifiers using the match score patterns from different users, while the literature on biometric menagerie, mentioned in Section 2.4.1, suggests that users of a biometric system may have differing degrees of accuracy within the system [108]. Particularly, the sheepish users can be easily recognized, matching poorly against others, and well against themselves, while other users (i.e., lambish and wolfish users) are particularly successful at impersonation, receiving high match scores for all verifications even when matching against others. Clearly, these users, when enrolled in the system, amplify the within-class sub-concepts problem in the impostor class distribution. User-specific score fusion partially alleviates this problem by defining a different reference model for each user, thus forming a more reliable and compact scatter, which reduces both types of errors, i.e., False Acceptance Rate (FAR) and False Rejection Rate (FRR).

Figure 4.5 provides an overview of the effect of user-specific vs. user-independent descriptions, learned using impostor class samples of the first three users in the BioSecure DS2 database using samples of the face and iris matchers. Specifically, the three continuous lines (drawn in red, green and blue) correspond to distinct user-specific descriptions. The dashed line (drawn in black) corresponds to the user-independent description and is obtained from the three user-specific training data by learning the decision boundaries, using GMM and k-NN. Impostor samples are indicated by '+', while genuine samples are given by 'o'. In this context, the false rejects correspond to those genuine samples, which are accepted by the descriptions. It can be seen that user-specific score fusion, as compared to user-independent score fusion, has resulted in a reduction in the number of false rejects by two and three samples, when GMM, and k-NN are used to train the descriptions, respectively. The descriptions, related to the SVDD and K-means are not demonstrated here, since they show a similar trend to those of GMM and k-NN.

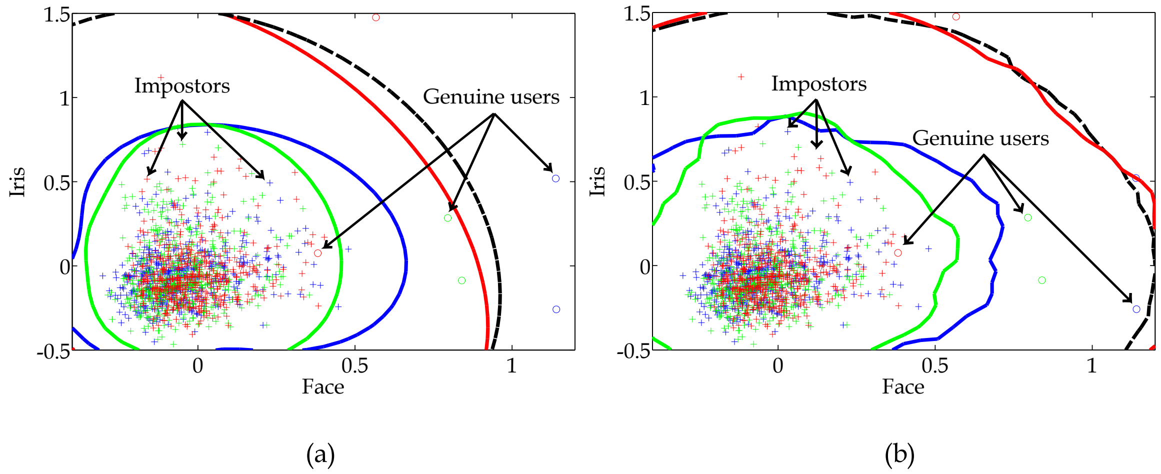


Figure 4.5. User-specific descriptions (continuous lines) and user-independent description (dashed line) around the impostor class, learned by (a) GMM and (b) k-NN for the first 3 users in the BioSecure DS2 database using match scores of the face and iris matchers. The match scores of these matchers were normalized using the F-norm [79].

4.3. Empirical Evaluation of One-class Learning

In this section, extensive experiments are conducted to evaluate the user-specific and user-independent score fusion schemes with various one-class classifiers, used to learn the descriptions: i.e., Gaussian Mixture Model (GMM) k-Nearest Neighbour (k-NN), K-means clustering (K-means) and Support Vector Data Description (SVDD). These classifiers have a common parameter to adjust, i.e., the fraction rejection $f_T \in [0, \dots, 1]$. Figure 4.6 shows the user-independent descriptions learned by GMM with two values of f_T (e.g., $f_T = 0$ and $f_T = 0.1$). It can be seen that by increasing f_T, a smaller decision boundary is obtained, leading to a reduction in the probability of accepting outliers, while enhancing the probability of rejecting target samples. The matching accuracy of the one-class classifiers is therefore dependent on the choice of the fraction rejection. A similar observation can also be made when k-NN, K-means and SVDD are selected to train the description. In general, f_T can take values between 0 and 1. Because the data collection is assumed to be accurate and there is a low density of noisy samples in the training set, the rejection rate of more than 50% (i.e., $f_T = 0.5$) of the target data is considered excessive. In this thesis, the choice of $f_T \in [0, \dots, 0.5]$ will be evaluated in terms of the performance of the user-independent and user-specific score fusion schemes. Other parameters specific to one-class classifiers were determined using grid search and cross validation [54].

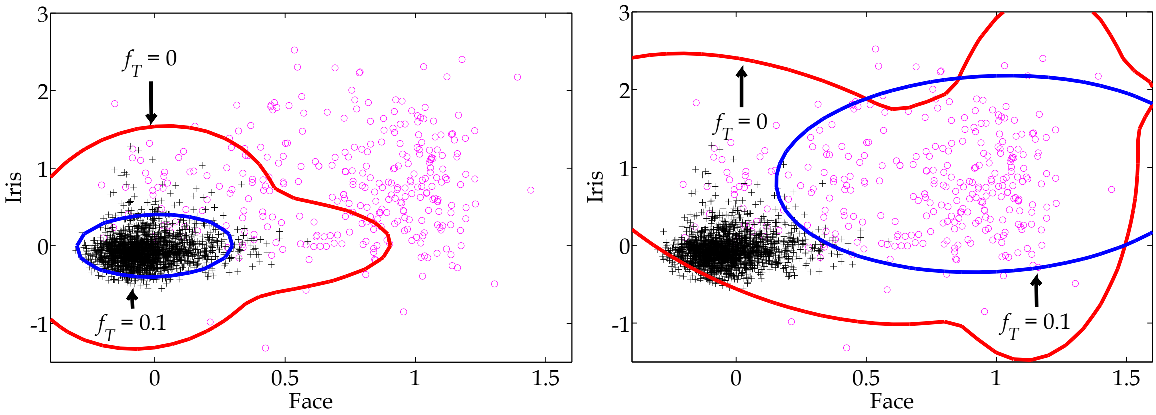


Figure 4.6. User-independent descriptions around (a) the impostor class, and (b) the genuine user class, learned by GMM with various values of f_T. The continuous red (blue) lines correspond to the descriptions learned with $f_T = 0$ ($f_T = 0.1$).

User-specific and user-independent approaches are also compared with other state-of-the-art techniques, discussed in Chapter 2, namely the sum of scores (abbreviated as SUM), likelihood ratio based score fusion (abbreviated as LR), and conventional two-class SVM (abbreviated as SVM). In this section, the following abbreviations will be often referred: UIFG (for User-independent score fusion using the genuine user target class), UIFI (for User-independent score fusion using the impostor target class), and USFI (for User-Specific score fusion using the impostor target class).

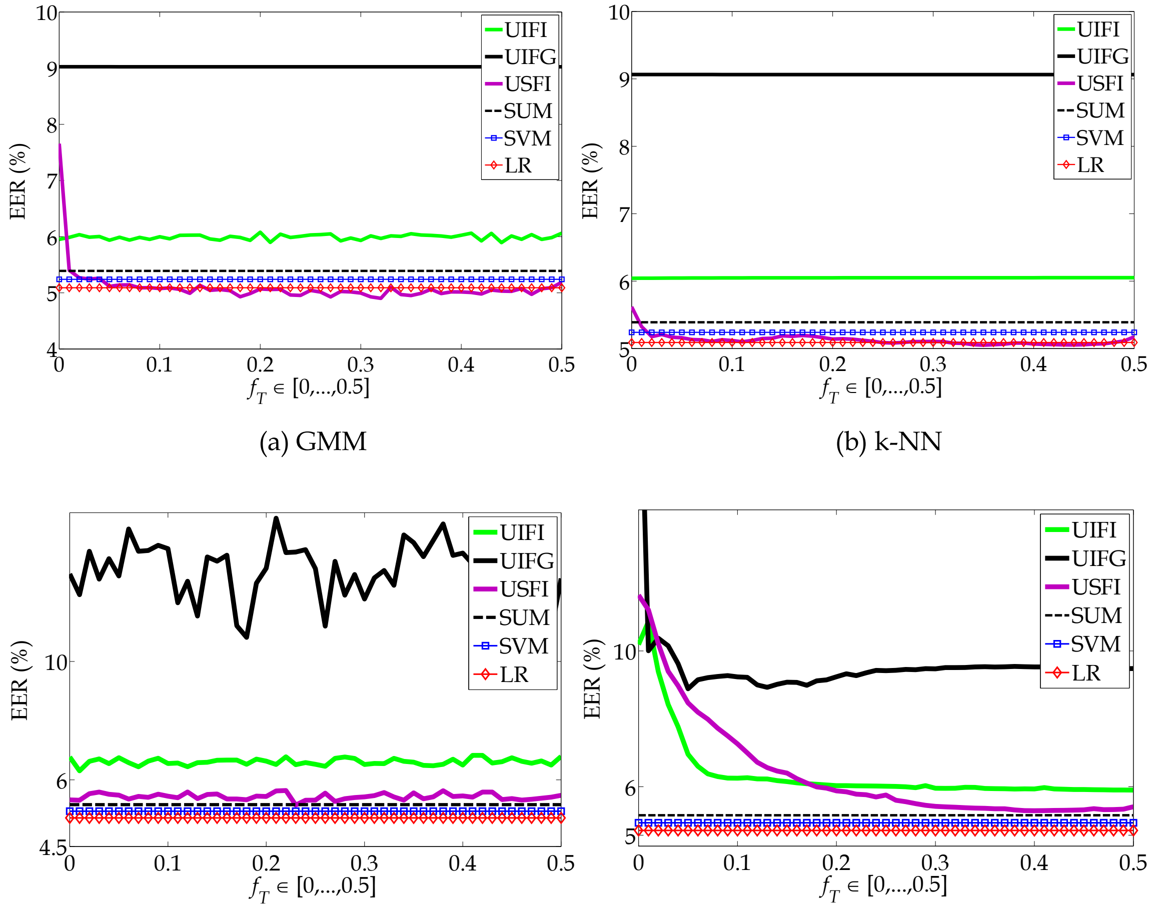
As already mentioned in Section 3.5, four sets of experiments will be presented. In the first three sets, two biometric matchers were combined simultaneously in the context of multimodal fusion, leading to 13, 15 and 6 combination possibilities for the BioSecure DS2, XM2VTS LP1, and LP2 databases. In the remaining set, three out of the eight biometric matchers of BioSecure DS2 database were consolidated at the same time. Hence, 6 multimodal combination possibilities were performed to assess the impact of having additional pieces of evidence from another biometric trait on the combination techniques. The reported results for a given set of experiments will correspond to the average EER over all the combination possibilities.

Note that there is a large number of impostor match score vectors in the XM2VTS and BioSecure DS2 databases. Using all these samples to train the SVM and the one-class classifiers in UIFI can be computationally expensive and increase memory requirements. To avoid this, a subset of 2,000 impostor match score vectors is randomly selected to carry out training. The selection process is repeated 20 times, and the mean authentication rates over 20 trials will be reported.

4.3.1. Experiment (1): Combining Two Biometric Matchers of the BioSecure DS2 Database

The performance of GMM, k-NN, K-means and SVDD as applied in learning user-specific and user-independent descriptions is illustrated in Figure 4.7.

Among the user-independent approaches, UIFI consistently outperforms UIFG, when any of the one-class classifiers are selected. For example, UIFI with K-means classifier achieves an average of 6.24% in terms of EER at $f_T = 0.01$, while that of UIFG is only 9.06% in EER at $f_T = 0.18$. Since USFI learns a different description for each of the users, it is shown to be better than both its user-independent counterparts.



(a) GMM

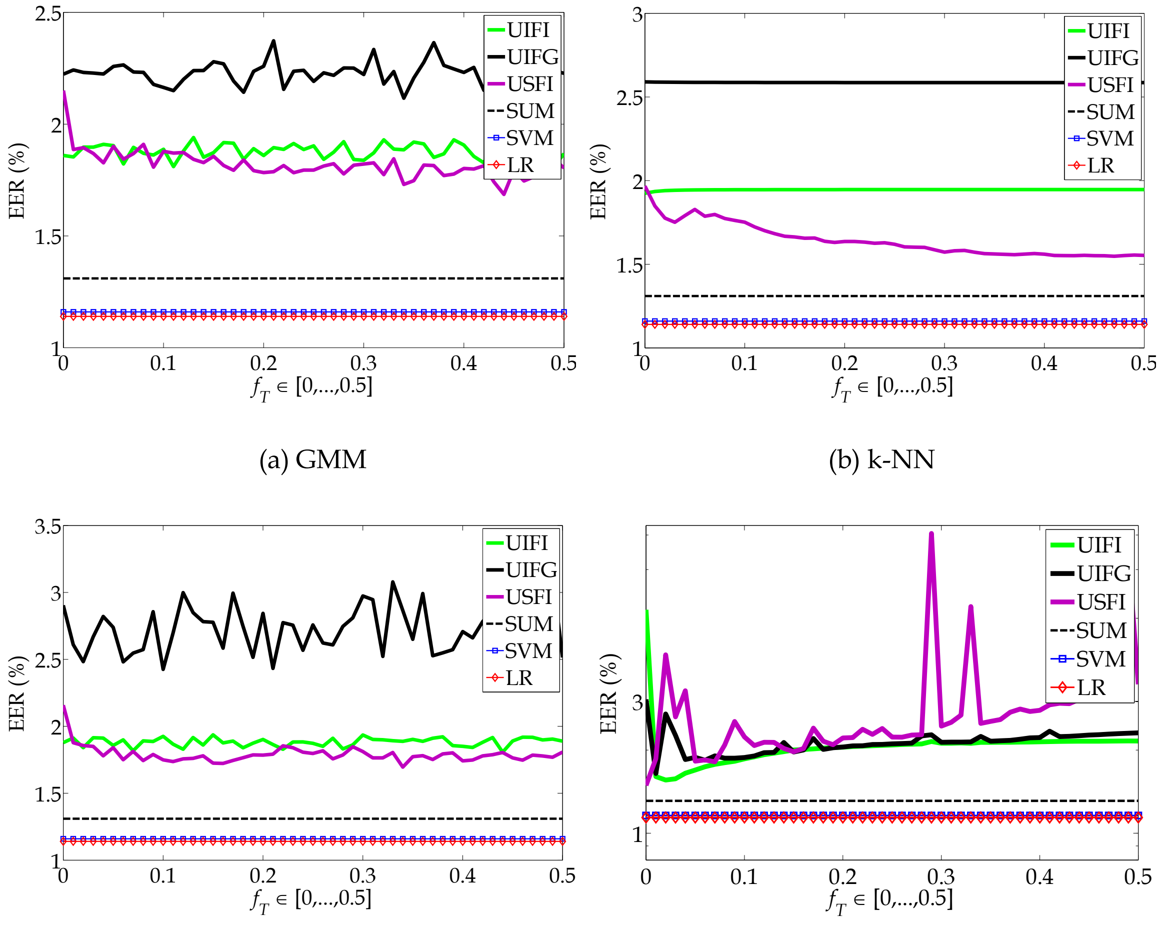
(b) k-NN

(c) K-means

(d) SVDD

Figure 4.7. Combining two biometric matchers of the BioSecure DS2 database: The average EER (%) of UIFI, UIFG, and USFI when applied in conjunction with various one-class classifiers, is plotted as a function of fraction rejection $f_T \in [0, \ldots, 0.5]$. *The schemes are also compared with state-of-the-art techniques, such as SUM, SVM and LR.*

Among the various one-class classifiers used in USFI, it is observed that GMM performs best as it achieves an average EER of 4.90% at $f_T = 0.32$. This is followed by k-NN (5.05% EER at $f_T = 0.35$). Both of these one-class classifiers are shown to be better than SUM (5.39% EER), SVM (5.24% EER) and LR (5.09% EER), even when they use samples exclusively from the impostor class. SVDD is shown to be highly sensitive to the choice of f_T, while the performances of GMM, k-NN, and K-means do not change much for $f_T \in [0.02, \dots, 0.5]$.



(a) GMM

(b) k-NN

(c) K-means

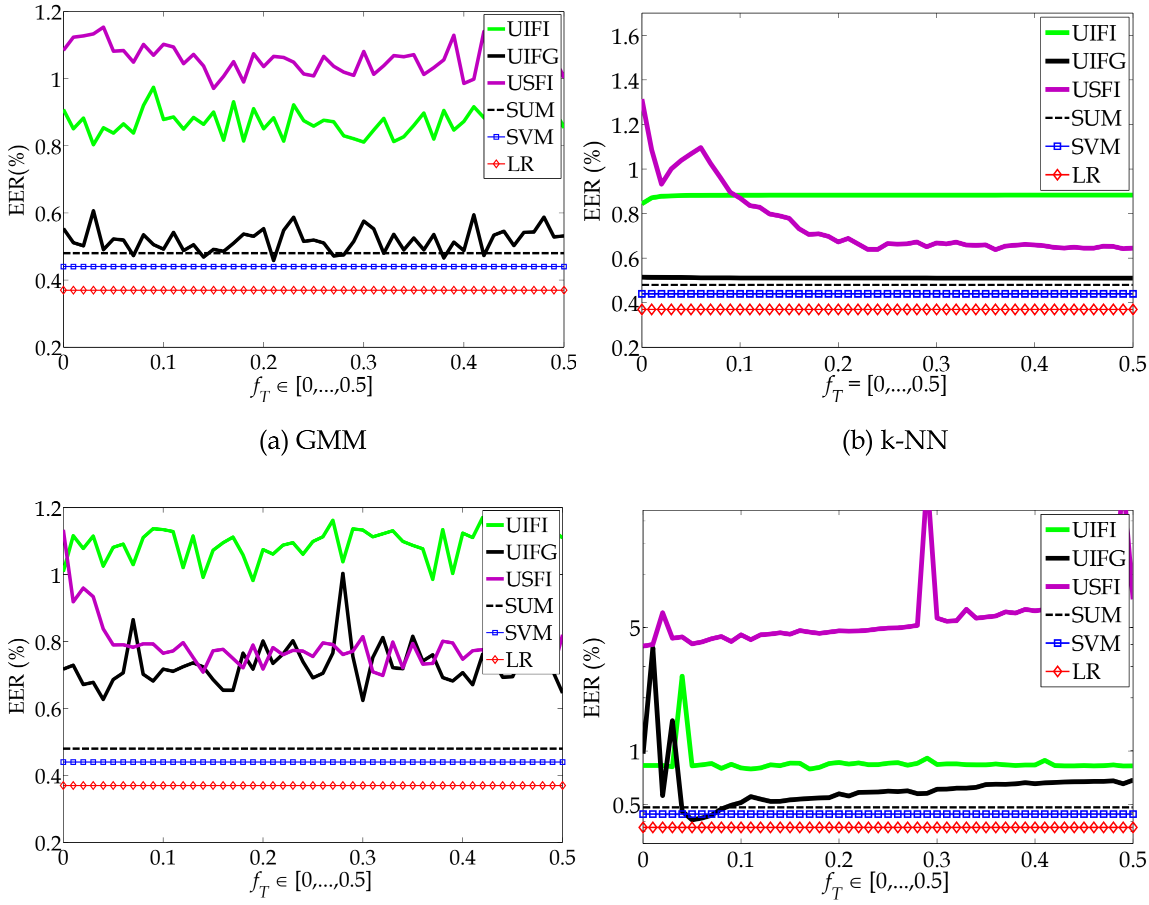
(d) SVDD

Figure 4.8. Combining two biometric matchers of the XM2VTS LP1 database: The average EER (%) of UIFI, UIFG, and USFI when applied in conjunction with various one-class classifiers, is plotted as a function of the fraction rejection $f_T \in [0, \dots, 0.5]$. The schemes are also compared with state-of-the-art techniques, such as SUM, SVM and LR.

4.3.2. Experiment (2): Combining Two Biometric Matchers of the XM2VTS LP1 database

In Figure 4.8, the performances of UIFI, UIFG and USFI show a similar trend to those of Experiment (1), where UIFI is shown to be better than UIFG, while the lowest error rates are produced by the USFI. However, none of these schemes is able to achieve better performance, with respect to those of SUM (1.31% EER), SVM (1.16% EER) and LR (1.14% EER), even when various one-class classifiers are used to learn the decision boundaries.

Among the one-class classifiers used in USFI, SVDD performs best with an average EER of 1.49% at $f_T = 0$. This is followed by k-NN (1.55% EER for $f_T \in [0.41, \dots, 0.5]$), GMM (1.69% EER at $f_T = 0.44$), and K-means (1.70% EER at $f_T = 0.38$). However, while the performances of GMM and K-means do not vary significantly for $f_T \in [0.01, \dots, 0.5]$), that of k-NN decreases by increasing the values of f_T. SVDD is shown to be more dependent on the choice of such parameter, as its performance becomes even lower with respect to that of the user-independent schemes when $f_T > 0$.



(a) GMM

(b) k-NN

(c) K-means

(d) SVDD

Figure 4.9. Combining two biometric matchers of the XM2VTS LP2 database: The average EER (%) of UIFI, UIFG, and USFI when applied in conjunction with various one-class classifiers, is plotted as a function of the fraction rejection $f_T \in [0, \dots, 0.5]$. The schemes are also compared with state-of-the-art techniques, such as SUM, SVM and LR.

4.3.3. Experiment (3): Combining Two Biometric Matchers of the XM2VTS LP2 database

Figure 4.9 shows the performance of various match score fusion schemes, as applied in combining two biometric matchers of the XM2VTS LP2 database. As opposed to Experiments (1) and (2), UIFG is observed to consistently achieve lower error rates, with respect to UIFI and USFI. USFI is shown to be better than UIFI when k-NN and K-means are employed to learn the decision boundaries. However, such method, when using GMM and SVDD to perform the fusion demonstrates the poorest verification performance.

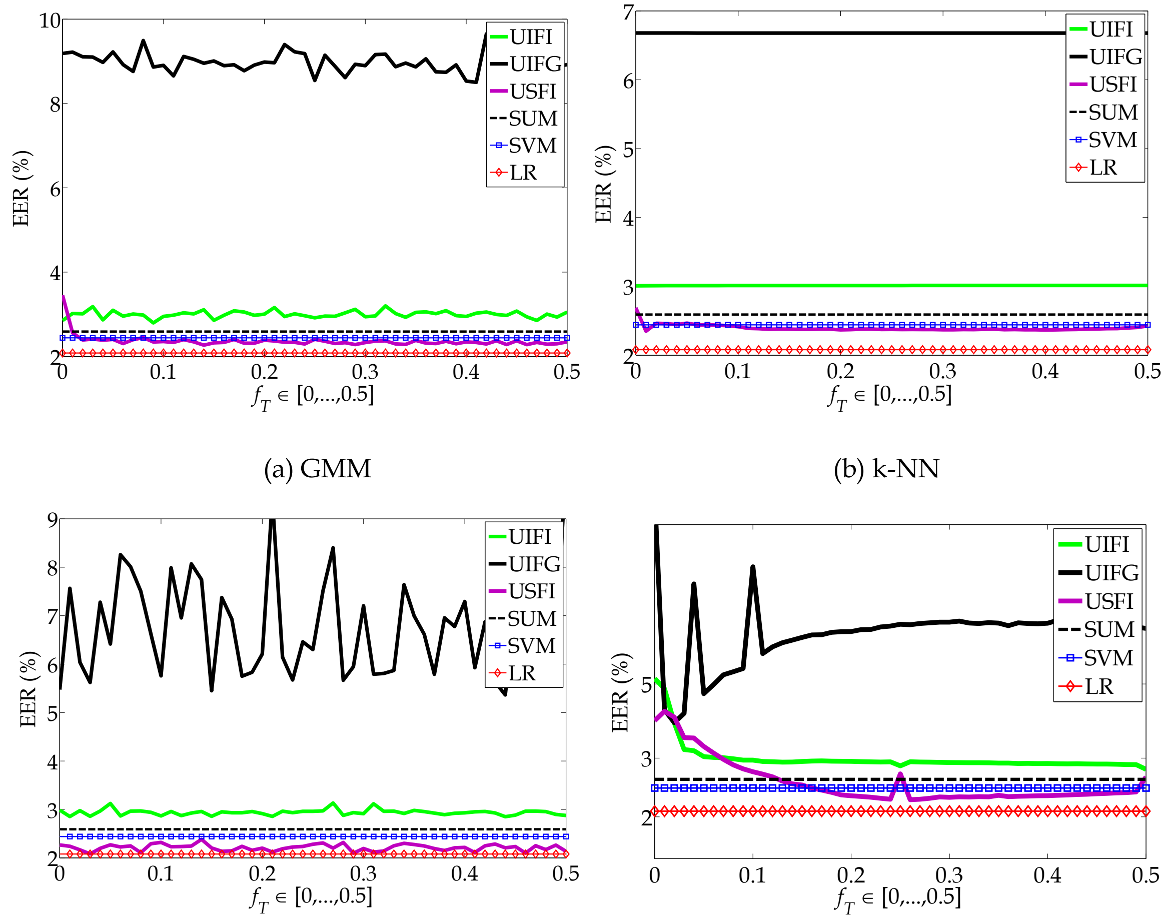
Among the various one-class classifiers used in UIFG, SVDD achieves the lowest error rate (0.41% EER at $f_T = 0.05$), which is better than the SUM (0.48% EER) and SVM (0.44% EER). On the contrary, GMM (0.46% EER at $f_T = 0.21$), k-NN (0.51% EER), and K-means (0.62% EER at $f_T = 0.32$) cannot perform as well as the state-of-the-art solutions. It is also noted that SVDD cannot provide a good result for $f_T \neq 0.05$. In contrast to SVDD, the verification accuracy of k-NN appears to be unchanged for $f_T \in [0, \dots, 0.5]$. Overall, LR with average EER of 0.37% outperforms other evaluated schemes in this experiment.

4.3.4. Experiment (4): Combining Three Biometric Matchers of the BioSecure DS2 database

For the fusion of three biometric matchers from the BioSecure DS2 database, Figure 4.10 shows the verification performance for the various one-class classification methods, as applied to learn the user-specific and user-independent descriptions. It is observed that all of these schemes demonstrate significantly improved EER performance, as compared to the two biometrics case in Experiment (1). It is therefore beneficial to have additional traits to render the biometric decision. Similarly to Experiments (1) and (2), the highest error rates are produced by UIFG, while USFI consistently outperforms UIFI.

Among the various one-class classifiers used in USFI, K-means achieves the lowest error rates (2.08% EER at $f_T = 0.03$), which is similar to that of the LR.

This is followed by SVDD (2.25% EER at $f_T = 0.26$), GMM (2.27% EER at $f_T = 0.14$), and k-NN (2.35% EER at $f_T = 0.01$). Obviously, USFI even when applied in conjunction with the one-class classifiers can provide a better performance with respect to the SVM (2.44% EER) and SUM (2.59% EER). Apart from SVDD, GMM, k-NN and K-means are able to achieve a consistently high performance for $f_T \in [0.01, \dots, 0.5]$.



(a) GMM

(b) k-NN

(c) K-means

(d) SVDD

Figure 4.10. Combining three biometric matchers of the BioSecure DS2 database: The average EER (%) of UIFI, UIFG, and USFI when applied in conjunction with various one-class classifiers, is plotted as a function of the fraction rejection $f_T \in [0, \dots, 0.5]$. The schemes are also compared with state-of-the-art techniques, such as SUM, SVM and LR.

4.3.5. Computational Complexity

In the previous experiments, I focused on the error performance of the various one-class classifiers as applied in learning the user-independent and user-specific decision boundaries. It is also important to evaluate the computational complexity of these methods. In most cases, it is possible to train a classifier off-line and the training time is not of interest. In practical applications however, the evaluation time might be critical to improve the degree of user convenience.

From Table 4.2, it can be observed that when SVDD and k-NN are applied in rendering the biometric decision, UIFI is the most computationally expensive. UIFG has the lowest evaluation time in Experiments (1) and (4), while in Experiments (2) and (3) the least computational complexity is produced by USFI. This naturally meets the expectation, shown in Table 4.1, which demonstrates that SVDD and k-NN have time complexity of order N, where N is the sample size. It has to be noted that in all the experiments, the descriptions of UIFI are trained on a subset of 2000 impostor match score vectors. On the other hand, UIFG learns its descriptions using 156, 600, 400 and 156 samples in Experiments (1), (2), (3) and (4), whereas those numbers, which were used by USFI are 524, 200, 200 and 524, respectively. The evaluation time of SVDD is much higher than that of k-NN since it has to compute the kernel between the input match score vector and the support vectors.

Table 4.2. Evaluation time (ms) of the one-class classification methods, as applied in learning the user-specific and user-independent descriptions

Methods		Experiment (1)	Experiment (2)	Experiment (3)	Experiment (4)
GMM	USFI	0.40	0.54	0.54	0.45
	UIFG	0.30	0.55	0.59	0.31
	UIFI	0.34	0.57	0.57	0.35
k-NN	USFI	2.79	1.08	1.08	3.11
	UIFG	0.75	2.25	1.67	0.76
	UIFI	6.26	6.22	6.22	6.47
K-means	USFI	0.18	0.31	0.32	0.21
	UIFG	0.17	0.27	0.28	0.17
	UIFI	0.17	0.28	0.28	0.19
SVDD	USFI	7.78	6.59	6.62	8.00
	UIFG	7.30	7.07	7.40	7.44
	UIFI	8.27	7.33	7.60	8.87

As opposed to SVDD and k-NN, the evaluation time of GMM and K-means does not change much in all the experiments, even when they are applied in learning the user-specific or user-independent descriptions. This implies that these classifiers are not influenced by the sample size. In general, their computational complexity is shown to be dependent only on the number of the mixture components and the prototype objects, which can be automatically found by adopting cross validation [54] or the batch algorithm (see Section 4.1.3). The obtained results in Table 4.2 provide concrete evidence to support the argument, made in Section 4.2.2, which suggested that when the decision boundaries are learned using the samples exclusively from impostor class, user-specific approach is much faster in testing time with respect to user-independent counterpart.

It can be also observed that the evaluation time of all the one-class methods is increased in Experiment (4) as compared to that in Experiment (1). The straightforward reason for this is that in Experiment (4), additional pieces of evidence (biometric traits) have been used for rendering the biometric decision and hence, the dimensionality of the feature vectors (i.e., match score vectors) is increased. Among the various one-class classifiers, SVDD demonstrates to be the most expensive. This is followed by k-NN, while K-means requires the least evaluation time. In terms of computational complexity, GMM is in the middle between k-NN and K-means.

4.3.6. Summary and Discussion

Based on the previous experiments, the following observations can be made:

- Impostor match score vectors are a more reliable source of information, which can be used to learn the descriptions, since UIFI and USFI consistently demonstrate better performances with respect to those of UIFG in most of the experiments. USFI is also observed to be a better alternative as compared to its user-independent counterpart in terms of both verification accuracy and computational time. In practice, USFI can be directly applied in those biometric systems, where only data coming from the impostor class is available for training the decision boundaries.

- It is observed that the user-specific score fusion is able to achieve a lower error rate as compared to the standard two-class SVM in Experiments (1) and (4), which are carried out on the BioSecure DS2 database. However, this observation does not hold in Experiments (2) and (3), which were conducted on the XM2VTS database. A possible reason for this is that XM2VTS database contains only 200 impostor samples for learning the user-specific decision boundaries, while the number of impostor samples per user in the BioSecure DS2 database is 524. The significant reduction in sample size may have a detrimental effect on the overall generalization error. This observation will be further analysed in Section 4.4.1.
- Various one-class classifiers have been applied in performing fusion. There is no one-class classifier that performs best in all of the experiments. GMM achieves the lowest error rate in Experiment (1), while SVDD and K-means produce the highest performance in Experiments (2), (3), and (4). The performance of SVDD is highly sensitive to the choice of the fraction rejection f_T, while those of GMM, k-NN and K-means do not vary significantly for $f_T \in [0.02, \dots, 0.5]$. In term of time complexity, SVDD and k-NN are computationally expensive, while GMM and K-means demonstrate the least evaluation time.

4.4. The Influence of Data Characteristics on the Performance of One-class Classifiers

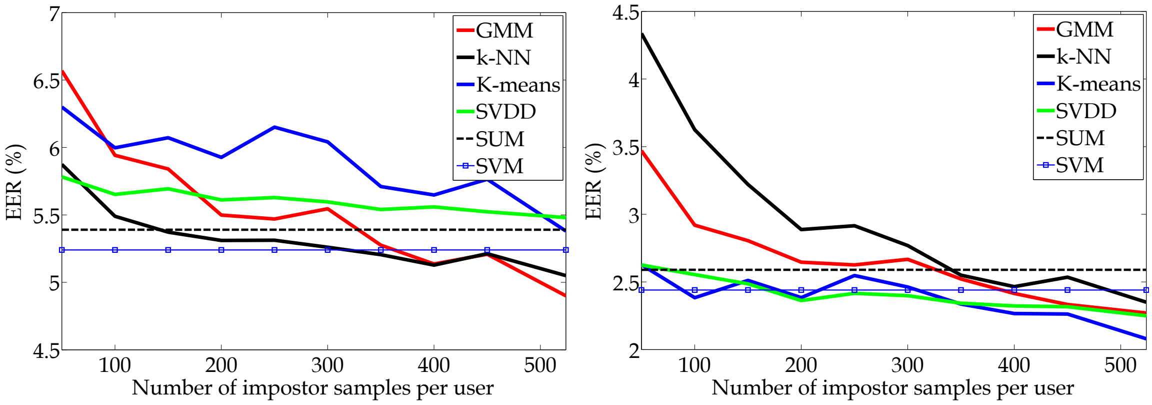
In [44], the influence of sample size, multimodality, non-convexity, subspaces, robustness, and score scaling on various one-class classification methods were investigated from a pattern recognition perspective. This section is therefore dedicated on the factors, which have a significant impact on the practical applications of USFI in multimodal biometric systems. These factors include sample size and score scaling. Here I do not consider the remaining factors, such as multimodality, non-convexity and subspaces because they are much dependent on the nature of biometric data, and cannot be controlled through the data collection and post-processing processes.

4.4.1. Sample Size

Since USFI uses only samples from the impostor class, the problem of training data scarcity due to the limited availability of genuine class samples per user can be completely eliminated. However, the number of user-specific impostor match scores available for training the description can be relatively small due to the time, effort and cost involved in collecting multimodal biometric data [1]. Due to these reasons, in this section, the following two essential questions will be addressed:

(1) *How does the sample size influence the verification accuracy of various one-class classifiers?*

(2) *How many of the user-specific impostor samples are needed to guarantee the good performance of USFI as compared to other state-of-the-art techniques, such as SUM, and SVM?*



(a) Combing two biometric modalities (b) Combining three biometric modalities

Figure 4.11. EER (%) of USFI when using various one-class classification methods as a function of the number of training samples.

To this end, the average EER of the one-class classification methods is evaluated with varying numbers of impostor samples, used for learning the user-specific descriptions. Experiments were carried out on the BioSecure DS2 database, which involves a reasonably large number of impostor samples per user. It should be noted that addressing the second question will provide obvious evidence to support the observations, made in Section 4.3.6.

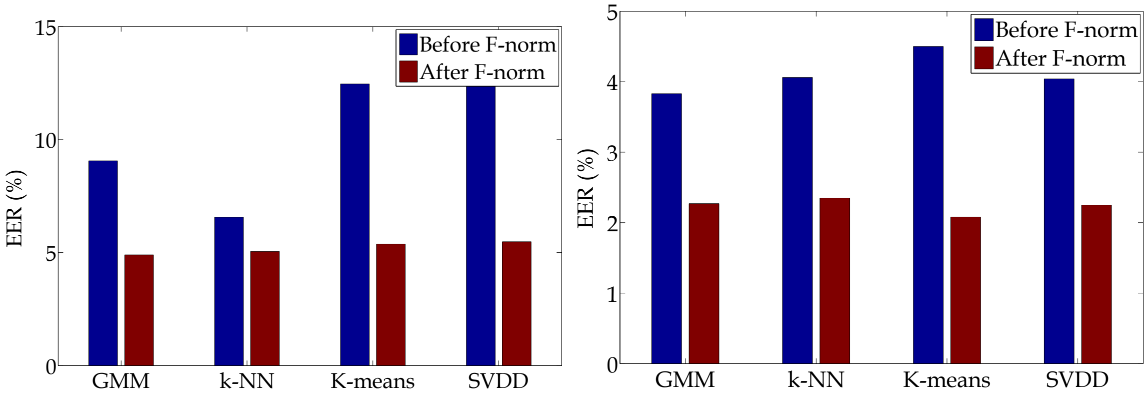
Note that SUM is a non-trainable combination rule, while SVM is a user-independent approach. Both of them are not affected by the sample size. As illustrated in Figure 4.11, SVDD seems to be least sensitive to the sample size, with respect to other one-class classifiers. In general, the description, learned by SVDD can be described by few training samples, known as support vectors. Hence, removing all other samples would leave the description unchanged.

For GMM, k-NN and K-means, increasing the sample size leads to an error decrease. GMM and k-NN seem to be mostly influenced by this factor since their EERs significantly deteriorate by 30% when the number of user-specific impostor match score vectors is decreased from 524 to 50. It is observed that these classifiers require at least 400 training impostor samples in order to provide a higher matching accuracy with respect to SVM.

SVDD and K-means cannot achieve a good performance when combining two biometric modalities (see Figure 4.11(a)). However, their performances are significantly improved in the three biometrics case (see Figure 4.11(b)). These classifiers are shown to perform better than SVM when at least 200 training samples are used to learn the user-specific decision boundaries. Recall that USFI when using any of the one-class classifiers was not able to produce lower error rates as compared to SVM in Experiments (2) and (3) in Section 4.3. A straightforward rationale for this may be due to the small sample size as the Experiments (2) and (3) were carried out on the XM2VTS database, where a limited number (i.e., 200) of impostor samples is available for learning the user-specific descriptions.

4.4.2. Data Scaling

It is well-known that the match scores, generated by individual matchers are not in the same numerical scale (range). It is generally believed that one-class classifiers are capable of learning the decision boundary irrespective of how the feature vectors are generated. However, recent studies have indicated that some methods are heavily dependent on the appropriate definition of a well-scaled feature [44]. This holds most explicitly for the SVDD, but also for the K-means, k-NN, and GMM. It has to be noted that the inhomogeneity of the data is more notable in the BioSecure DS2 database, where the iris matcher has match scores in the range of [0, ... ,1], while the range of other matchers is [0, ... ,100]. Since this characteristic is not really shown in the XM2VTS database, the influence of data scaling on the various one-class classifiers will be demonstrated using the BioSecure DS2 database.

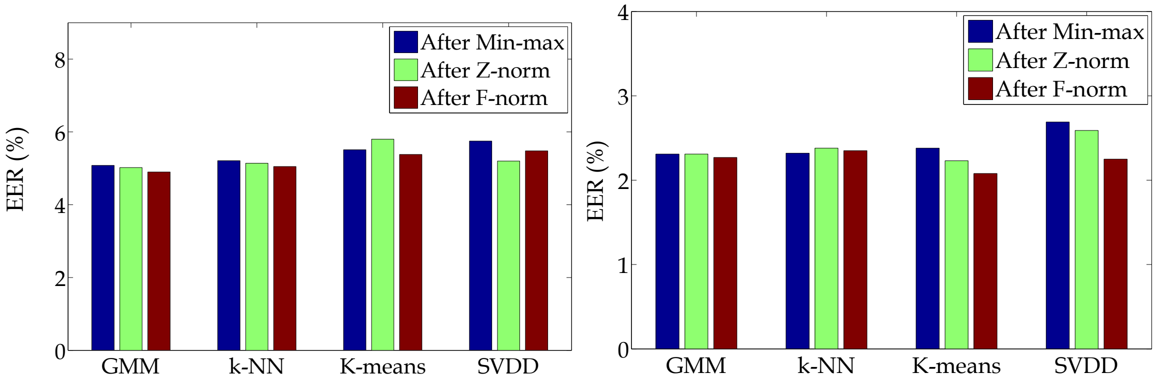


(a) Combing two biometric modalities

(b) Combining three biometric modalities

Figure 4.12. EER (%) of various one-class classification methods, as applied to train the user-specific score fusion when no normalization is performed and when the F-norm is used prior to fusion.

In Figure 4.12, the EER of various one-class classification methods is shown in both cases, when the F-norm is applied to transform the match scores into a common scale, and when no normalization is carried out. It is observed that SVDD and K-means are most sensitive to the scaling of the data as their performances heavily rely on the distance of the test samples from the centre of the sphere or the prototype objects. Applying the F-norm prior to fusion can lower the error rate of SVDD and K-means by as much as 57%. GMM and k-NN are less influenced by the scaling of the data. Nevertheless, a decreasing trend is still observed. Hence, score normalization is beneficial in all these cases.



(a) Combing two biometric modalities (b) Combining three biometric modalities

Figure 4.13. EER (%) of various one-class classification methods, as applied to train the user-specific score fusion when the Min-max, Z-norm and F-norm are used prior to fusion.

Figure 4.13 illustrates the error performance of the one-class classifiers when various score normalization techniques, such as Min-max, Z-norm and F-norm (see Sections 2.3.2 and 2.4.2) were applied in rescaling the match scores of the individual matchers. Clearly, since GMM and k-NN are probability based methods, they consistently achieve their best generalization performance irrespective of which score normalization techniques are chosen to pre-process the data. This somewhat confirms the finding in [106]. On the contrary, the EER of SVDD and K-means is much dependent on the choice of the score normalization techniques. Applying the F-norm can reduce their error rates by as much as 16% with respect to those of the Min-max and Z-norm. Based on these empirical results, it can be concluded that selecting the appropriate scaling solutions is crucial to the success of SVDD and K-means. In this section, the influence of the score scaling on GMM, k-NN, K-means and SVDD was assessed in the context of the user-specific score fusion. However, the above observations can be applicable to the user-independent counterparts, where the various one-class classifiers are equally influenced by the scaling factor.

4.5. Chapter Summary

In this chapter, I have provided a thorough investigation of various one-class classifiers, such as GMM, k-NN, K-means and SVDD, as applied in training the user-independent and user-specific score fusion schemes using data from either genuine user or impostor class. It was observed that the user-specific approach is a better alternative with respect to the user-independent counterpart. Impostor match scores are a more reliable source of information, which can be used to learn the descriptions.

As expected, one-class classifiers are particularly useful in handling the extremely imbalanced biometric data sets. They have been demonstrated to be either better or comparable to the two-class SVM when the experiments were conducted on the BioSecure DS2 database, which has the class imbalance ratio to be on the order of 524:1.

Among the various one-class classifiers, SVDD is able to achieve a good performance but is computationally expensive. It is also highly sensitive to the selection of the fraction rejection. On the contrary, K-means has the least evaluation time, but cannot produce a satisfactory performance in most cases. Both SVDD and K-means are heavily influenced by the scaling of the match scores, provided by the individual matchers. GMM is the best trade-off between verification accuracy and time complexity. However, it is highly sensitive to the training sample size.

CHAPTER 5

COMBINING AND BOOSTING ONE-CLASS CLASSIFIERS

In Chapter 4, the use of the various one-class classifiers has been investigated in order to advance the classification performance of the extremely imbalanced biometric data sets. This chapter aims at systematically improving the performance of these methods. Towards this end, the following two contributions have been made. In the first contribution, described in Section 5.1, a novel adapted score fusion is proposed, which is based on Bayes Decision Theory, as applied in combining one-class classifiers to effectively exploit the training data from both classes (genuine user/impostor). It also makes use of user-specific instead of user-independent score fusion to learn the characteristics of the impostor class, and thus, decreasing the degree of class imbalance and counteracting the effects of the within-class sub-concepts problem. In the second contribution (see Section 5.2), a hybrid boosting algorithm, called r-ABOC is developed, which inherits the naturally capabilities of Real AdaBoost in order to enhance the system performance without causing overfitting. However, unlike the conventional Real AdaBoost, the individual classifiers in r-ABOC are trained on the same data set, but with different parameter choices. This does not only generate the necessary diversity to make r-ABOC perform well, but also reduces the number of user-specified parameters.

5.1. Combining Descriptions

This section will start with the related topic of multiple classifier combination. Next, the proposed adapted score fusion scheme is presented. Finally, I provide an extensive empirical evaluation and discussion of the results.

5.1.1. Combining Pattern Classifiers

By combining descriptions, we aim at a more accurate classification decision at the expense of increased complexity [59]. The idea is not to rely on a single decision making classifier. Instead, all the designs, or their subsets, are used for decision making by combining their individual outputs in order to render a consensus decision [95]. This may not only increase the performance, but also the robustness of classification [44].

A large number of combining methodologies exist, which can be divided into two categories, depending on the nature of the outcomes of the individual classifiers [44], [129]. In the first category, each classifier outputs hard class labels, and these labels can be combined using majority voting [59], [130] and label ranking [59]. The second category involves the combination of continuous outputs (i.e., the degrees of support for a given input pattern) for each of the classes. The continuous outputs can be either posterior probabilities [131] or evidences [1], [132].

In [133], the authors investigated the effect of regularization on averaging the estimated posterior probabilities of individual classifiers. They compared four different averaged regularized combiners, namely simple averaging, bagging, variance-based weighting and variance-based bagging. It was empirically reported that bagging and variance-based bagging achieve the lowest overall error rates over a wide range of degrees of regularization. Simple averaging also improves the performance of the individual classifiers. However, the degree of improvement may be application-specific.

In [134], the Bayes theorem was adopted to combine probability estimates of individual classifiers. Under the assumption of statistical independence, the outputs of these classifiers can be multiplied and normalized. This is called logarithmic opinion pool, which was reported to achieve a significant improvement in overall classification accuracy as compared to the individual classifiers in the experiments, carried out on the multisource remote sensing/geographic data and very high dimensional data.

In [95], the authors developed various types of classifier combination rules, namely the sum, product, max, min, and median decision rules, which were introduced in Section 2.2. It was shown that when the data representations are independent, classifiers' outcomes should be multiplied to gain maximally from the independent representations. Otherwise, they should be summed to reduce the errors in the posterior probability estimation. In [44], [135], the sum decision rule was observed to be particularly useful in combining classifiers with highly correlated feature spaces.

For the combination of one-class classifiers, another approach is introduced. The posterior probabilities have to be estimated using information exclusively from the target class. In [136], by assuming a uniform distribution over the feature space, several combining schemes, such as mean vote, mean weighted vote, product of weighted votes, mean of the estimated probabilities and product combination of the probabilities, were evaluated to combine the different types of one-class classifiers on a handwritten digit dataset. It was observed that in most cases, the product combination of the estimated probabilities achieves the lowest error rates, whereas the mean of these probabilities suffers from the fact that the area covered by the target class tends to be overestimated, and hence, more outlier samples are accepted. Combining various one-class classifiers does not always result in better performance when one single classifier is already highly accurate or when the posterior probability outcomes of the individual classifiers are poorly estimated.

Unlike [136], this research is not dealing with the problem of combining various families of one-class classifiers. Instead, the focus is to combine the outputs of the one-class classifiers of the same type to efficiently use both the genuine user and impostor class samples, which are always available during training in most practical biometric systems. It is clear that training data from two classes definitely provide more information to define the description than sampling only on one side/data from one class. In general, the following criterion can be adopted [45]:

$$\arg\left\{\max_{k\in\{G,I\}} \delta(\mathbf{s}_T|w_k)\right\} \tag{5.1}$$

where $\delta(\mathbf{s}_T|w_k)$ can either be $p(\mathbf{s}_T|w_k)$ or $d(\mathbf{s}_T|w_k)$ depending on the nature of the one-class classifier. Although this criterion is very simple and straightforward, it is not based on any solid (Bayesian) foundation and may not be suitable for those biometric applications, where the reliability of each of the predictions has to be estimated.

5.1.2. Adapted Score Fusion Scheme

Assume that $\mathbf{s}_T$ is the test match score vector and w_k is the target class, where $k \in \{G, I\}$ indicates the genuine user or impostor class. In Chapter 4, it was shown that an one-class classifier of the same type can produce four different outputs for a given $\mathbf{s}_T$ of the claimed identity j. Particularly, GMM and k-NN outputs are $p(\mathbf{s}_T|w_k)$ and $p(\mathbf{s}_T|w_{j,k})$, while the SVDD and K-means outputs are $d(\mathbf{s}_T|w_k)$ and $d(\mathbf{s}_T|w_{j,k})$. It should be, however, noted that $p(\mathbf{s}_T|w_{j,G})$ and $d(\mathbf{s}_T|w_{j,G})$ are impractical to achieve due to the limited availability of genuine match score vectors per user. Here, a novel adapted score fusion scheme is proposed to combine the remaining outputs so as to improve system performance. According to the Bayes Decision Theory,

$$\begin{gathered}\text{Assign } \mathbf{s}_T \rightarrow w_G \text{ if} \\ p(w_G|\mathbf{s}_T) \geq p(w_I|\mathbf{s}_T)\end{gathered} \tag{5.2}$$

In general, the a posteriori probabilities of $\mathbf{s}_T$ belonging to the genuine user or impostor class are computed as

$$p(w_k|\mathbf{s}_T) = \frac{p(\mathbf{s}_T|w_k)p(w_k)}{\sum_{k\in\{G,I\}} p(\mathbf{s}_T|w_k)p(w_k)} \tag{5.3}$$

By assuming that the prior probabilities are equal ($p(w_G) = p(w_I)$), Equation (5.2) can be rewritten as

$$\text{Assign } \mathbf{s}_T \rightarrow w_G \text{ if} \tag{5.4}$$

$$p(\mathbf{s}_T|w_G) \geq p(\mathbf{s}_T|w_I)$$

Or

$$\text{Assign } \mathbf{s}_T \rightarrow w_G \text{ if}$$
$$p(\mathbf{s}_T|w_G) - p(\mathbf{s}_T|w_I) \geq 0 \tag{5.5}$$

Equation (5.5) is only achieved when all types of errors are equally costly. Most biometric verification systems assign different costs to the two types of error rates, i.e., False Acceptance Rate (FAR) and False Rejection Rate (FRR) [1]. Thus, FAR and FRR can be considered as functions of the decision threshold τ. For a given τ, Equation (5.5) is as follows

$$\text{Assign } \mathbf{s}_T \rightarrow w_G \text{ if}$$
$$p(\mathbf{s}_T|w_G) - p(\mathbf{s}_T|w_I) \geq \tau \tag{5.6}$$

Equation (5.6) holds when the prior probabilities are assumed to be equal. This assumption is, however, not valid in biometric systems due to the highly imbalanced class distributions. A better alternative is to replace $p(\mathbf{s}_T|w_I)$ of the user-independent score fusion with $p(\mathbf{s}_T|w_{j,I})$ of the user-specific counterpart. The rationale is that both the outcomes are based on the descriptions, which exploit the same source of information, i.e., impostor match score vectors. This not only alleviates the within-class sub-concepts problem in the impostor score distribution (see Section 4.2), but also reduces the degree of imbalance for different classes. For example, the BioSecure DS2 database has the imbalance ratio to be in the order of 524:1. Replacing $p(\mathbf{s}_T|w_I)$ with $p(\mathbf{s}_T|w_{j,I})$ reduces this ratio to 5:1. The adapted score fusion scheme can be finally written as:

$$\text{Assign } \mathbf{s}_T \rightarrow w_G \text{ if}$$
$$p(\mathbf{s}_T|w_G) - p(\mathbf{s}_T|w_{j,I}) \geq \tau \tag{5.7}$$

Previously, it is assumed that one-class classifiers directly approximate the posterior probabilities of $\mathbf{s}_T$ for the target class. When some classifiers, such as K-means, SVDD, are not based on some type of density estimation, the posterior probabilities can be estimated by a heuristic mapping as in [44], [136]:

$$p(\mathbf{s}_T|w_k) = \frac{1}{c_1}\exp\left(-\frac{d(\mathbf{s}_T|w_k)}{c_2}\right) \tag{5.8}$$

where c_1 is a normalization constant and c_2 is a scale parameter. Both can be fitted to the distribution of $d(\mathbf{s}_T|w_k)$ of the training (target) class. It is clear that the probability estimate decreases to zero when the distance is very large. On the contrary, the probability becomes maximal when the distance drops to zero. Since the probability is always bounded between 0 and 1, the proposed adapted score fusion scheme is bounded between -1 and 1.

In general, the proposed adapted score fusion scheme can be considered as a unified framework for combining the outcomes of one-class classifiers regardless of whether they are distance-based or probability-based. It has the following three significant advantages. Firstly, it uses one-class methods instead of two-class ones to counteract the class imbalance problem, which is encountered very often in biometric systems. Secondly, it is believed to be a better alternative as compared to the ν-SVM in [51], which is a simple application of ν-SVM for multimodal fusion, because it combines user-specific and user-independent fusions to effectively exploit the entire training data set. Finally, since user-specific score fusion is trained using only samples from the impostor class, the problem of training data scarcity of the genuine class samples per user can be completely eliminated.

5.1.3. Experimental Setup

In the reminder of Section 5.1, extensive experiments are carried out using the data from the BioSecure DS2 and XM2VTS databases. In order to assess the effectiveness of the proposed adapted score fusion scheme, various one-class classifiers, described in the Chapter 4, have been tested, i.e., Gaussian Mixture Model (GMM), k-Nearest Neighbour (k-NN), K-means clustering (K-means) and SVDD (with RBF kernel). For notation purposes, this scheme when used with

GMM, k-NN, K-means and SVDD is abbreviated by A-GMM, A-kNN, A-Kmeans and A-SVDD, respectively.

Similarly to Section 4.3, A-GMM, A-kNN, A-Kmeans and A-SVDD will be evaluated with different fraction rejections f_T, which can take values between 0% and 50% (i.e., $f_T \in [0, \dots, 0.5]$). Next, their performance will be compared with that of the state-of-the art solutions. In addition to the user-independent approaches, namely the sum of scores (abbreviated as SUM), likelihood ratio based score fusion (abbreviated as LR), and two-class SVM (abbreviated as SVM), the Adapted User-Dependent Fusion (abbreviated as AUDF) (see Section 2.4.4) is also evaluated for comparison purposes. Cross validation [54] was adopted to find the optimal parameters for each of these methods.

The AUDF approach (see Section 2.4.4), presented in this research is based on SVM. For this method, there is a trade-off parameter $\alpha \in [0,1]$, which reflects the reliability of user-specific information and thus influences verification performance. Particularly, $\alpha = 0$ implies that no user-specific score fusion is needed and AUDF is equivalent to SVM, while $\alpha = 1$ indicates that user-specific score fusion is beneficial. In order to obtain the best performance from AUDF, different values for $\alpha \in [0,1]$ will be evaluated. The performance of the user-specific fusion technique, based on SVM is equivalent to that of AUDF for $\alpha = 1$.

Similarly to Section 4.3, four sets of experiments will be presented. In the first three sets, two biometric matchers are combined with respect to the multimodal biometric fusion. It should be noted that these sets of experiments were based on the BioSecure DS2 and XM2VTS databases, which differ not only in term of the class imbalance ratio, but also in the number of genuine user samples, which can be used to learn the user-specific descriptions (i.e., 1 in BioSecure DS2, 3 in XM2VTS LP1 and 2 in LP2). Hence, the possible effect of the number of genuine match scores per user on the verification performance is also evaluated. In the remaining set, three out of the eight biometric matchers of BioSecure DS2 database are combined at the same time. The performance in terms of Equal Error Rate (EER) and relative change of EER will be evaluated. In

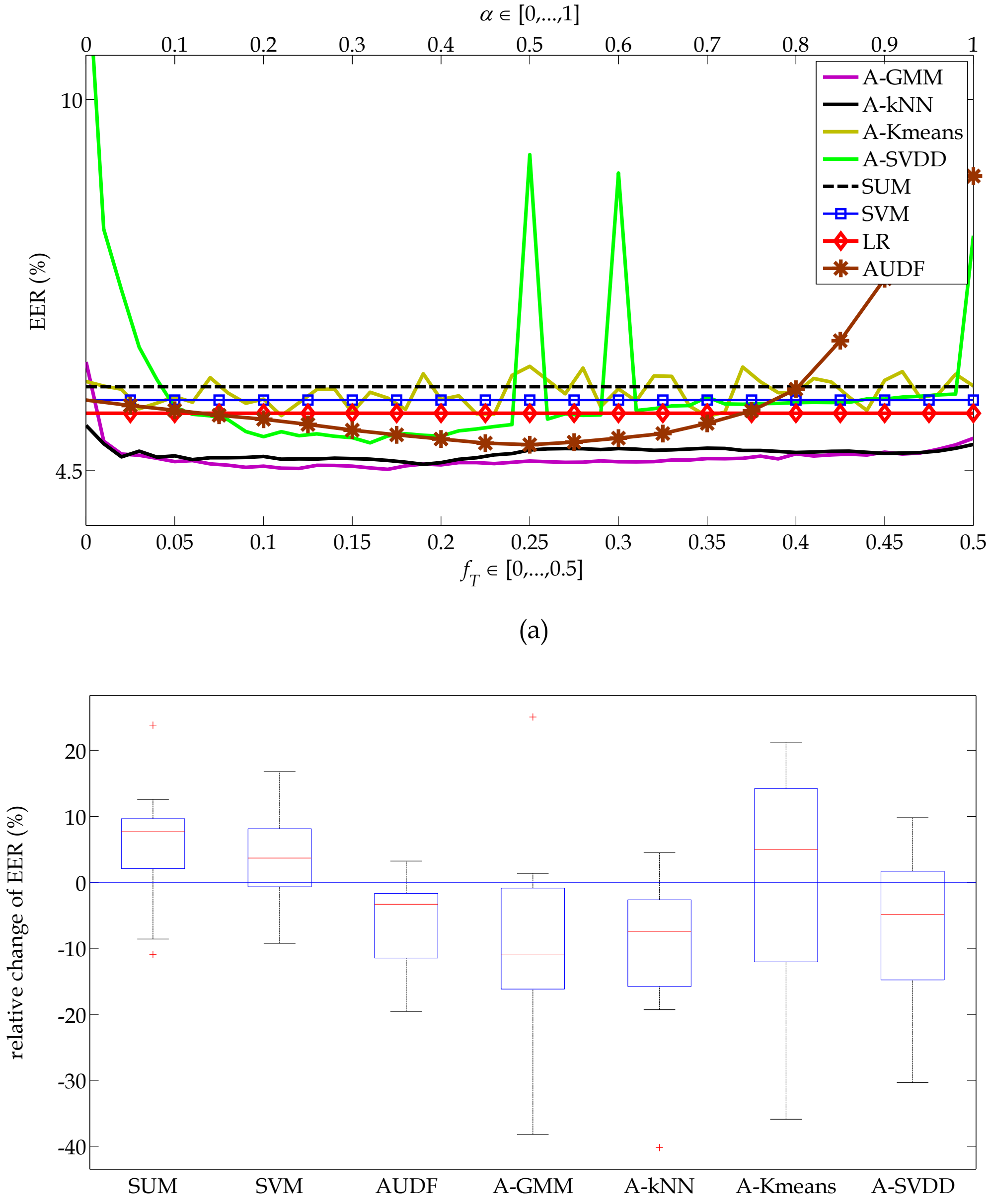
Section 4.3, it was observed that no one-class classifiers, even when applied to learn the user-specific descriptions in order to render a biometric decision, were able to achieve a higher matching accuracy than LR. Due to this reason, LR is selected as the baseline system to estimate the relative change of EER. The relative change of EER will be illustrated using the box plot (i.e., a non-parametric approach showing the median, the 25-th and 75-th, as well as the 5-th and 95-th percentiles of the data). Obviously, a negative (positive) change of EER implies a performance improvement (decrease), whereas zero change implies no change in performance.

5.1.4. Experiment (1): Combining Two Biometric Matchers of the BioSecure DS2 Database

The performance of SUM, SVM, LR, AUDF, A-GMM, A-kNN, A-Kmeans and A-SVDD schemes is illustrated in Figure 5.1(a). Note that the effect of two tunable parameters is considered in the Figure. The range of fraction rejection f_T is considered for A-GMM, A-kNN, A-Kmeans and A-SVDD and displayed at the bottom of the figure, while α is considered for the case of AUDF and displayed at the top of the figure.

As it can be seen, LR performs best among all user-independent schemes, as it provides an average EER of 5.09%. This is followed by SVM (5.24%) and SUM (5.39%). AUDF is observed to outperform user-independent approaches for $\alpha \in [0.15,..,0.75]$. When $\alpha = 1$, its performance reduces to that of the user-specific score fusion. It is noted that user-specific score fusion cannot provide a good performance (EER of 8.48%) since there is only one genuine match score per user available to train the decision boundary for each user. AUDF achieves its lowest average EER (4.76%) at $\alpha = 0.5$, which is similar to that of A-SVDD at $f_T = 0.16$.

A-Kmeans achieves its lowest error rate, i.e., an average EER of 5.06% at $f_T = 0.35$. The performances of A-kNN and A-GMM do not change much for $f_T \in [0.01,..,0.5]$. The lowest EER (4.56%) for A-kNN is found at $f_T = 0.19$, while that of A-GMM (EER of 4.51%) is found at $f_T = 0.17$. The authentication performance of A-GMM demonstrates relative EER improvements of approximately 5%, 11%, 14% and 16% with respect to AUDF, LR, SVM and SUM, respectively.



(a)

(b)

Figure 5.1. Combining two biometric matchers of the BioSecure DS2 database: (a) Average EER (%) for different learning settings (i.e., fraction rejection f_T, and trade-off parameter α), (b) Relative change of EER (%) across 13 combination possibilities of SUM, SVM, AUDF ($\alpha = 0.5$), A-GMM ($f_T = 0.17$), A-kNN ($f_T = 0.19$), A-Kmeans ($f_T = 0.35$) and A-SVDD ($f_T = 0.16$) with respect to LR.

In terms of relative change of EER, shown in Figure 5.1(b), it is observed that AUDF, A-GMM, A-kNN and A-SVDD demonstrate an improvement over the baseline system, i.e., LR, since their median values are less than zero. AUDF has a similar median value to that of A-SVDD. The lowest median value, which indicates the best performance, is attributed to A-GMM.

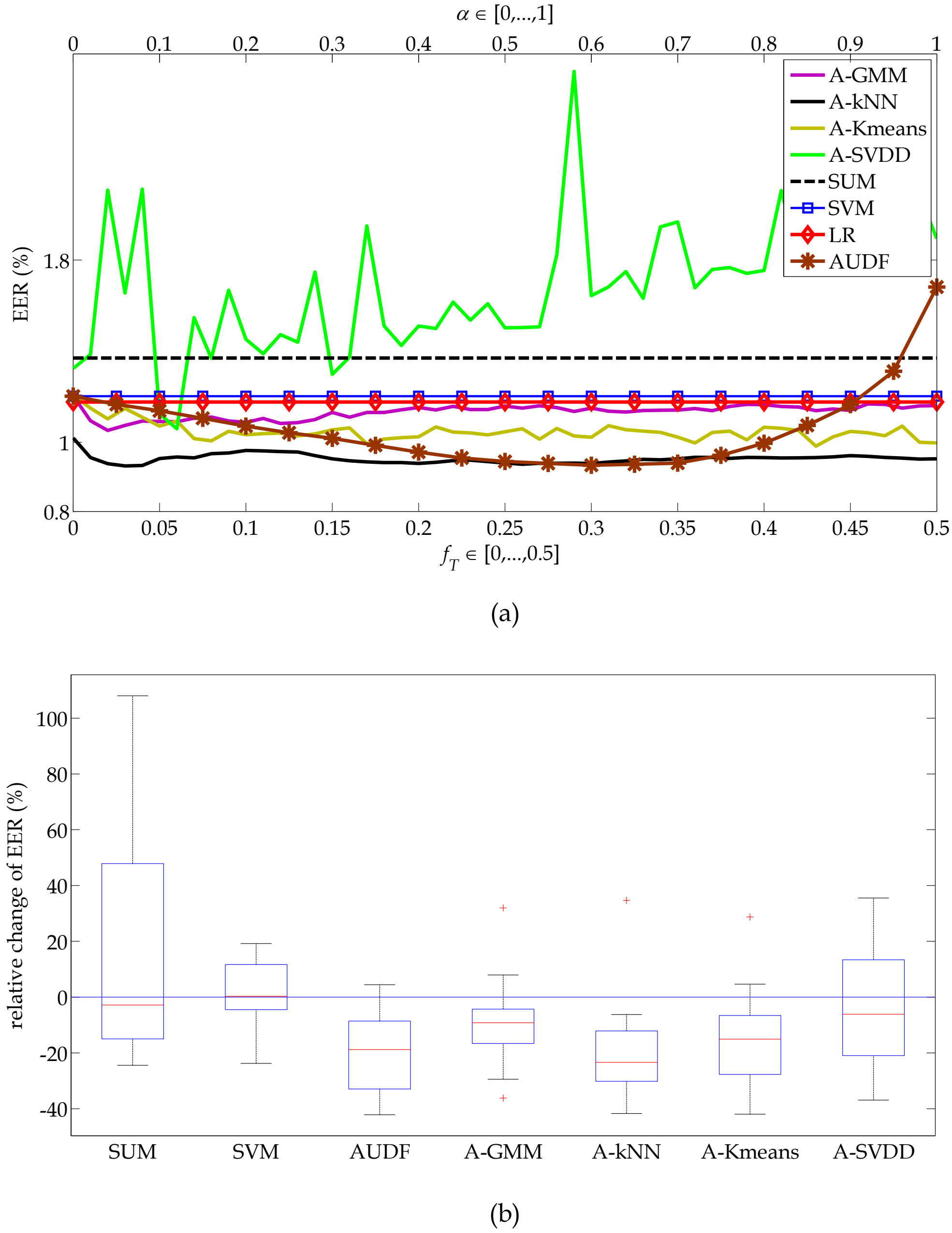


Figure 5.2. Combining two biometric matchers of the XM2VTS LP1 database: (a) Average EER (%) for different learning settings (i.e., fraction rejection f_T, and trade-off parameter α), (b) Relative change of EER (%) across 15 combination possibilities of SUM, SVM, AUDF ($\alpha = 0.6$), A-GMM ($f_T = 0.02$), A-kNN ($f_T = 0.03$), A-Kmeans ($f_T = 0.43$) and A-SVDD ($f_T = 0.06$) with respect to LR.

5.1.5. Experiment (2): Combining Two Biometric Matchers of the XM2VTS LP1 Database

In Figure 5.2(a), the performance of LR (1.14% EER) is observed to be either superior or comparable to that of SVM (1.16% EER), and SUM (1.31% EER). The evaluation of the influence of parameters was carried out as in Figure 5.1(a). Compared to these techniques, A-GMM has a lower EER (1.04%) at $f_T = 0.02$. A similar EER is obtained by A-SVDD when $f_T = 0.06$. However, while the good performance of A-GMM is consistently achieved for $f_T \in [0.01, \dots, 0.5]$, that of the A-SVDD significantly decreases when $f_T \neq \{0.05, 0.06\}$.

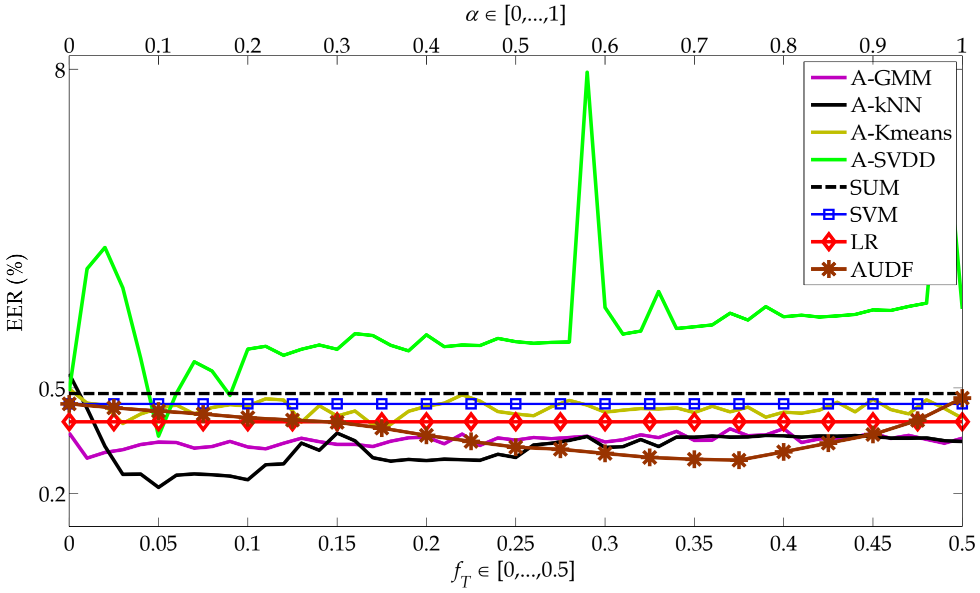
A-Kmeans has its best authentication accuracy (0.99% EER) at $f_T = 0.43$. The lowest EER (0.93%) is achieved by A-kNN at $f_T = 0.03$. In addition, for $f_T \in [0.01, \dots, 0.5]$, A-kNN is better than LR, SVM and SUM with a relative improvement in terms of average EER of 18%, 18% and 29%, respectively. An EER of 0.93% is also achieved by AUDF at $\alpha = 0.6$ since more (three) genuine match score vectors per user were used for training the user-specific fusion. However, the error rate of the user-specific score fusion (see Figure 5.2(a), performance of the AUDF at $\alpha = 1$) is still higher than that of the user-independent counterpart, i.e., SVM.

As illustrated in Figure 5.2(b), the highest performance in terms of relative change of EER is attributed to AUDF and A-kNN. These are followed by A-Kmeans, A-GMM, A-SVDD, and SUM. All these approaches have median values less than zero, and thus, demonstrate a performance improvement over LR.

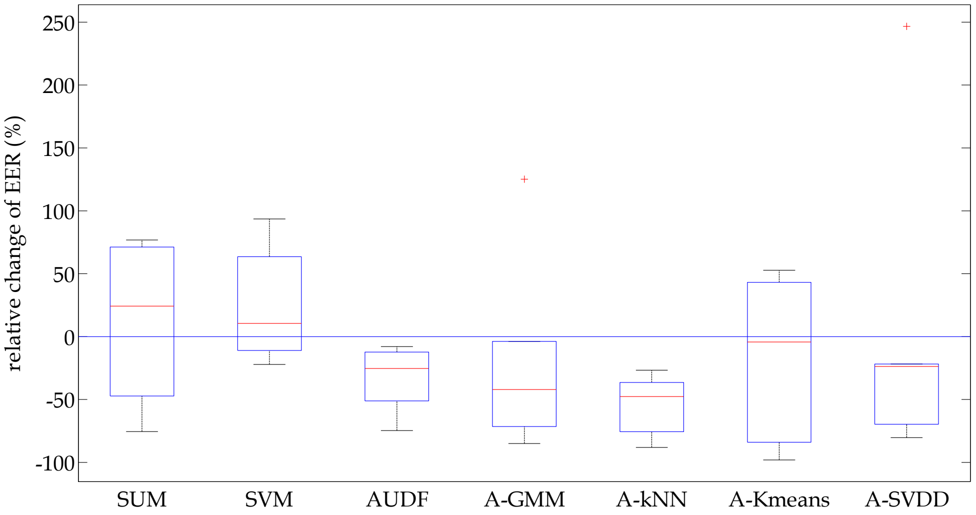
5.1.6. Experiment (3): Combining Two Biometric Matchers of the XM2VTS LP2 Database

Similarly to Experiments (1), and (2), for A-GMM, A-kNN, A-Kmeans and A-SVDD, various fraction rejection values were tested ($f_T \in [0, \dots, 0.5]$), and for AUDF, the trade-off parameter α was chosen in the interval of $[0, \dots, 1]$.

In Figure 5.3(a), it is observed that LR with an average EER of 0.37% outperforms other user-independent approaches, including SVM (0.44% EER) and SUM (0.48% EER). The best performance for AUDF (0.27% EER) is obtained at $\alpha = 0.75$, which is similar to that of A-GMM at $f_T = 0.01$. User-specific score fusion, which is equivalent to AUDF at $\alpha = 1$, cannot produce a better performance, as compared to user-independent counterpart, i.e., SVM.



(a)



(b)

Figure 5.3. Combining two biometric matchers of the XM2VTS LP2 database: (a) Average EER (%) for different learning settings (i.e., fraction rejection f_T, and trade-off parameter α), (b) Relative change of EER (%) across 6 combination possibilities of SUM, SVM, AUDF ($\alpha = 0.75$), A-GMM ($f_T = 0.01$), A-kNN ($f_T = 0.05$), A-Kmeans ($f_T = 0.18$) and A-SVDD ($f_T = 0.05$) with respect to LR.

A-SVDD has its lowest EER (0.33%) at $f_T = 0.05$, while that of A-Kmeans is 0.36% at $f_T = 0.18$. A-SVDD seems to be highly sensitive to the choice of fraction rejection as it cannot provide a satisfactory results for $f_T \neq 0.05$. It is observed that A-kNN, A-Kmeans and A-GMM are able to achieve a low EER for different f_T values. The highest authentication accuracy (0.21% EER) is obtained by A-kNN with $f_T = 0.05$, providing a relative EER improvement of 22%, 43%, 52% and 56% with respect to AUDF, LR, SVM and SUM, respectively.

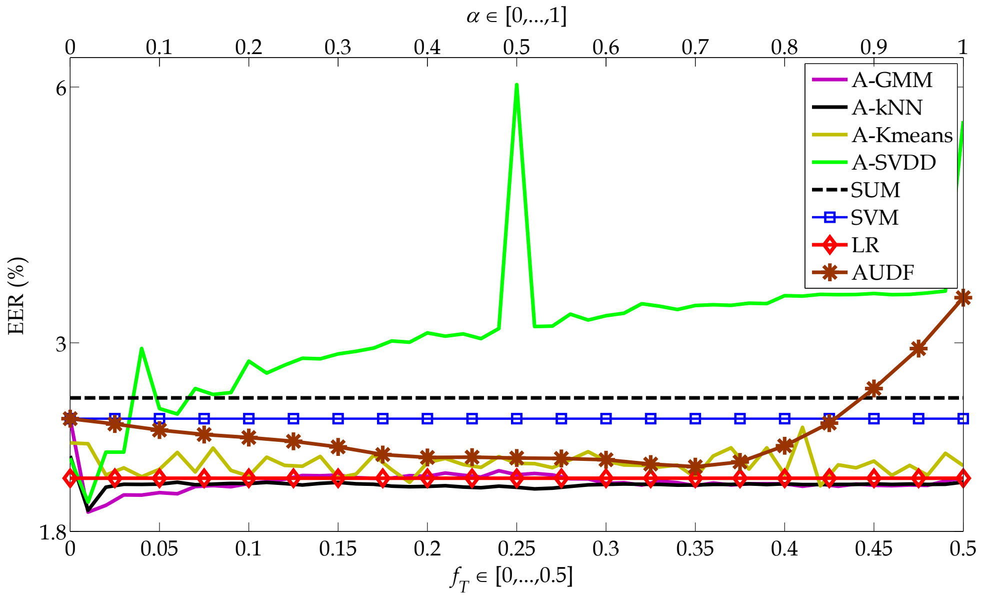
Concerning the relative change of EER in Figure 5.3(b), A-Kmeans has a lower median value (i.e., better performance) compared to that of SUM, SVM and LR. A-SVDD, AUDF and A-GMM are shown to be better relative to A-Kmeans, while the lowest median is achieved by A-kNN.

5.1.7. Experiment (4): Combining Three Biometric Matchers of the BioSecure DS2 Database

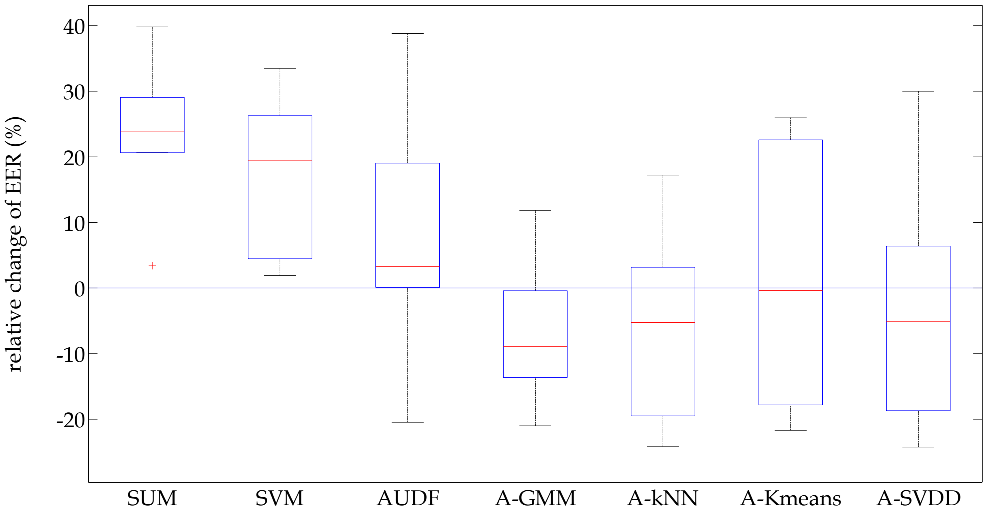
For the fusion of three biometric matchers from the BioSecure DS2 database, Figure 5.4(a) shows the verification performance for the various approaches. All these methods demonstrate significantly improved EER performance, compared to the two biometrics cases in Experiment (1). Particularly, a relative improvement of 60% is produced by A-Kmeans, while those of A-SVDD, A-GMM and A-kNN are 59%, 58%, and 58%, respectively.

Among the user-independent techniques, LR has the lowest EER (2.08%), while those of SUM, and SVM are 2.59%, and 2.44%, respectively. The lowest EER (2.15%) for AUDF is at $\alpha = 0.7$. Surprisingly, this error rate is higher than that of LR. A straightforward reason for this is that AUDF makes use of the SVM based user-specific and user-independent score fusions, which do not produce a good performance for this dataset.

A-Kmeans, A-SVDD, A-GMM and A-kNN are better than other techniques, providing a relative improvement of 9%, 12%, 22% and 27% with respect to LR, AUDF, SVM and SUM, respectively. The performance of these techniques for various values of f_T demonstrates similar trends to those in Experiment (1).



(a)



(b)

Figure 5.4. Combining three biometric matchers of the BioSecure DS2 database: (a) Average EER (%) for different learning settings (i.e., fraction rejection f_T, and trade-off parameter α), (b) Relative change of EER (%) across 6 combination possibilities of SUM, SVM, AUDF ($\alpha = 0.7$), A-GMM ($f_T = 0.01$), A-kNN ($f_T = 0.01$), A-Kmeans ($f_T = 0.42$) and A-SVDD ($f_T = 0.01$) with respect to LR.

With regards to the relative change of EER (see Figure 5.4(b) for more details), A-Kmeans is demonstrated to be better than SVM, SUM, LR and AUDF. However, its performance is still lower relatively to that of A-SVDD and A-kNN. The lowest median value is provided by A-GMM.

5.1.8. Summary and Discussion

Tables 5.1 and 5.2 summarize the findings of the previous experiments, by ranking the techniques based on their performance in terms of the average EER, and relative change of EER. The following observations can be made:

Table 5.1. Ranking of different approaches with respect to the EER

	Experiment (1)	Experiment (2)	Experiment (3)	Experiment (4)	Mean
A-GMM	**1**	4	2	**1**	2.00
A-kNN	2	**1**	**1**	**1**	**1.25**
A-Kmeans	5	3	5	4	4.25
A-SVDD	3	4	4	3	3.50
SUM	8	8	8	8	8.00
SVM	7	7	7	7	7.00
LR	6	6	6	5	5.75
AUDF	3	**1**	2	6	3.00

Table 5.2. Ranking of different approaches with respect to the relative change of EER

	Experiment (1)	Experiment (2)	Experiment (3)	Experiment (4)	Mean
A-GMM	**1**	4	2	**1**	2.00
A-kNN	2	**1**	**1**	2	**1.50**
A-Kmeans	7	3	5	4	4.75
A-SVDD	3	4	3	2	3.00
SUM	8	6	8	8	7.50
SVM	6	7	7	7	6.75
LR	5	7	6	5	5.75
AUDF	4	**1**	3	6	3.50

- The best authentication accuracy is achieved by A-kNN. This is followed by A-GMM, AUDF, A-SVDD, A-Kmeans, LR, and SVM. As compared to these techniques, SUM typically provides higher error rates due to the simplicity of the fusion rule [1].
- In Chapter 4, no one-class classifiers, even when applied in learning the descriptions for each of the users, were shown to be better as compared to LR, since they use samples exclusively from one class (either genuine user

or impostor class) to train the descriptions. On the contrary, the performance in terms of EER of A-GMM, A-kNN, A-Kmeans and A-SVDD is significantly better than LR, although the base classifier is a one-class method; however, these use training data from both classes.

- The state-of-the-art user-specific score fusion approach, considered in this research, is equivalent to AUDF at $\alpha = 1$. In brief, user-specific score fusion cannot provide a lower error rate relative to the user-independent counterpart, i.e., SVM for all the experiments. The performance of AUDF is improved by increasing the number of samples in the training of the decision boundary specific to a user. For this reason, it is observed that AUDF has a comparable performance to that of the A-kNN in Experiment (2), where three genuine match score vectors are available for each user, enrolled in the system. However, such method requires careful selection of the trade-off parameter $\alpha \in [0,1]$, which is shown to be influenced by the C parameter of SVM [15], [31], and the classification problem at hand. For example, the optimal choice of α is different for Experiments (1) to (4), where SVMs are trained using the same training data set, parameters, and kernel width. In addition, the lowest EER is achieved using a smaller α value in Experiment (2), as compared to those in Experiments (3), and (4), although the user-specific information in Experiment (2) is more reliable due to the higher number of genuine match score vectors per user, which can be used during training. A-GMM, A-kNN, A-Kmeans and A-SVDD combine both user-specific and user-independent information, similarly to AUDF, but do not require the choice of a trade-off parameter.
- A-GMM, A-kNN, A-Kmeans and A-SVDD have a single parameter to adjust, i.e., fraction rejection value $f_T \in [0, \dots, 0.5]$. It was demonstrated that A-SVDD is highly sensitive to the selection of f_T. As opposed to A-SVDD, the performance of A-GMM, A-kNN and A-Kmeans does not change significantly for different $f_T \in [0.01, \dots, 0.5]$, implying that these approaches are more robust to the selection of this parameter.

5.2. Boosting for Learning Imbalanced Biometric Data

Imbalance learning problems have drawn growing research interest due to their classification difficulty caused by the imbalanced class distributions. It should be noted that research solutions to handle the class imbalance are not solely in the form of one-class learning. For instance, Real AdaBoost [58] is the technique, which can be used to improve the classification performance of any classifier regardless of whether the data is imbalanced or not [34-38]. The aim of Real AdaBoost is to combine multiple (***weak***) classifiers in order to develop a highly accurate (***strong***) classifier system [58]. It is known to reduce bias and variance errors as it focuses on the samples, which are harder to classify. Particularly, Real AdaBoost weighs each sample to reflect its importance, and places the most weights on those samples, which are most often misclassified by the preceding classifiers [115]. Real AdaBoost is very effective at handling the class imbalance problem because the small class samples are most likely to be misclassified.

In this section, a novel hybrid boosting algorithm, called r-ABOC, is developed, which is capable of exploiting the natural capabilities of both Real AdaBoost (r-AB) and One-class Classification (OC) to address the problem of highly imbalanced biometric data sets. The proposed r-ABOC works by first considering the classifier, given in Equation (5.6), as the weak classifier. The paradigm of Real AdaBoost is then applied to further improve the performance of this classifier without causing overfitting. It has been recognized that diversity is a key requirement for the success of Real AdaBoost. Conventional Real AdaBoost generates diversity by training weak classifiers on different data subsets, constructed from the original training data. In the proposed paradigm, a new training procedure is introduced to train these classifiers on the same data set, but with different parameter choices. The target is to reduce the number of user-specified parameters, while still generating the diversity necessary to enable the classifier ensemble to perform well. Extensive experiments are carried out on the BioSecureDS2 and XM2VTS benchmark databases, which demonstrate that the proposed r-ABOC algorithm achieves significantly improved results in terms of Half Total Error Rate (HTER) as compared to state-of-the-art solutions.

The remainder of this section will start with a thorough discussion on the choice of the weak classifier algorithm. The proposed r-ABOC and other related classifier ensemble learning algorithms are then presented in Sections 5.2.2-5.2.4. Finally, sections 5.2.6-5.2.11 report the extensive experiments using the XM2VTS and BioSecure DS2 databases and discuss the results.

5.2.1. Weak Classifier Algorithm

The weak classifier algorithm, developed in this research, consists of one-class classifiers, trained using data from both the genuine user and impostor classes. A large number of one-class methods have been developed in the literature. Among others, Gaussian Mixture Model (GMM) has been demonstrated to successfully estimate the biometric match score distributions, and converge indeed to the true density with a sufficient number of training samples [90]. In sections 4.3.5, it was observed that the use of GMM also results in significant saving in testing time with respect to other classifiers, such as SVDD and k-NN. Hence, GMM is selected as the initial one-class classifier in this research.

In general, it is possible to combine the one-class GMMs using Equations (5.6), (5.7) or the log-likelihood ratio (see Section 2.3.1). Although the log-likelihood ratio is the optimal test for deciding that the test match score vector $\mathbf{s}_T$ corresponds to a genuine user or an impostor, it is not for use as weak classifiers in boosting algorithms, including r-ABOC. This is due to fact that the log-likelihood ratio is numerically unstable, leading to very large updates in the margins of some of the training samples. This, according to margin-based theory, can have a detrimental effect on the overall generalization error [58], [170]. On the contrary, Equations (5.6) and (5.7) have their continuous outputs to be in the range $[-1, +1]$ and hence, offer a distinct advantage especially when stability is an issue. Equation (5.7) is better alternative because it is capable of learning a different description for each of the users in the system, and thus, counteracting the within-class sub-concepts problem in the impostor class distribution. Its main drawback is however the need for substantial amount of match scores to train and validate the user-specific descriptions. Practical biometric systems only contain a limited number of genuine user samples per user in the training data sets. If all data is used for training and the same data is used for validation, the classification model might be over-trained so that it perfectly learns the available data and fails on unseen data [59]. It has to be also noted that a reliable model validation is a key requirement for the success of classifier ensemble learning algorithms. Due to these reasons, Equation (5.6) will be used as the weak classifier in the present work.

As previously demonstrated, such a classifier is naturally quite sensitive to the choice of the fraction rejection value. It is well-known that biometric data suffers from various forms of degradation, caused by being sampled in different circumstances, such as the manner a user interacts with a biometric device, the changes in the acquisition environment, and even the natural alteration of biometric traits due to sickness [71]. As a result, the fraction rejection value, obtained through an optimization process on the training set [44] does not necessarily provide optimal performance on the testing set. In order to eliminate the risk of making a bad choice for the fraction rejection for the problem at hand, a possible solution is to generate a number of classifiers with different values for this parameter and combine their outcomes to form the final hypothesis.

5.2.2. The Proposed r-ABOC Algorithm

Classifier ensemble learning has received much attention in recent years to solve complex recognition problems [59], [115]. This increased interest has been reflected in the introduction of a series of annual International Workshops on Multiple Classifier Systems, which have been held since 2000. A well researched survey in this field can be found in [59]. As already mentioned, the main motivation of classifier ensembles is to create a highly accurate classification model by combining multiple weak classifiers, which are only moderately accurate. This idea follows the human natural behaviour, which tends to seek several opinions before making any important decision.

In the literature, the effect of combining classifiers is studied in terms of statistical concepts of bias-variance decomposition. There is a trade-off between bias and variance as decreasing the bias will likely result in a higher variance, and vice versa. For example, increasing k in the k-NN classifier is believed to reduce the variance, and increase the bias [59]. For the decision tree, heavily pruned trees will have smaller variance and larger bias than trees, which are fully grown to classify correctly all training samples. In general, bias is associated with underfitting, while variance is associated with overfitting [59], [115], [137]. The improved performance of a classifier ensemble is therefore often a result of a reduction in variance, rather than a reduction in bias.

Boosting is known as one of the most successful classifier ensemble algorithms, because it is able to reduce bias (in addition to variance), and similarly to SVM boost the distance margin [34], [35]. Boosting can be applicable to most classification system and is capable of combining multiple classifiers with little risk of model overfitting [37]. Discrete AdaBoost [143] and Real AdaBoost [58] are the most representative boosting algorithms. Real AdaBoost is the generalization of Discrete AdaBoost. Real AdaBoost improves over Discrete AdaBoost because it generates not only hard class labels, but also real valued "confidence-rated" predictions, and thus, is more tolerant to classification noise, which naturally appears in biometric applications [169].

As previously mentioned, r-ABOC algorithm is based on Real AdaBoost The key to the success of Real AdaBoost is to build a set of diverse classifiers using different subsets of the training data [34-38]. The proposed r-ABOC algorithm offers the same advantages as Real AdaBoost. It is, however, different from the conventional Real AdaBoost, since the diverse classifiers in r-ABOC are trained with different fraction rejections.

Let $\mathbf{X} = \langle(\mathbf{s}_1, y_1), \dots, (\mathbf{s}_N, y_N)\rangle$ be a sequence of N training samples, where $\mathbf{s}_i$ is a match score vector and y_i is its associated class label, i.e., $y_i \in \{-1, +1\}$ (-1 denotes an impostor and +1 denotes a genuine user). At the t^{th} iteration, $t \in [1, \dots, \mathrm{T}]$, the weak classifier's continuous outcomes for the match score vector

$\mathbf{s}_T$ can be generated as $g_t(\mathbf{s}_T) = p(\mathbf{s}_T|w_G) - p(\mathbf{s}_T|w_I)$. Assume that $D_t(i)$ is the weight assigned to $\mathbf{s}_i$. The value of the fraction rejection f_T will be selected in the range $[0.01, \dots, 0.01 \times \mathrm{T}]$, where 0.01 is equivalent to 1% of the most dissimilar target samples that the classifier would reject during training. This level of detail is to ensure that the individual classifiers are not over-trained, while the single best classifier is still obtained.

```
Given $\mathbf{X} = \langle(\mathbf{s}_1, y_1), \dots, (\mathbf{s}_N, y_N)\rangle; y_i \in \{-1, +1\}$.
1)  Initialize $D_1(i) = 1/N$
2)  For $t = 1, \dots, \mathrm{T}$
    (a)  Train the classifier $g_t(\mathbf{s}_T) = p(\mathbf{s}_T|w_G) - p(\mathbf{s}_T|w_I)$ using $f_T = 0.01 \times t$
    (b)  Determine the weight updating parameter $\alpha_t$
             $\alpha_t = \frac{1}{2}\log\frac{\sum_{i=1}^{N} D_t(i)\big(1 + y_i g_t(\mathbf{s}_i)\big)}{\sum_{i=1}^{N} D_t(i)\big(1 - y_i g_t(\mathbf{s}_i)\big)}$
    (c)  Update and normalize $D_{t+1}(i)$ such that
             $D_{t+1}(i) = \frac{D_t(i)\exp\big(-\alpha_t y_i g_t(\mathbf{s}_i)\big)}{Z_t}$
         where $Z_t = \sum_{i=1}^{N} D_t(i)\exp(-\alpha_t y_i g_t(\mathbf{s}_i))$
3)  Output the final hypothesis
             $\mathrm{G}(\mathbf{s}_T) = sign\left[\sum_{t=1}^{\mathrm{T}} \alpha_t g_t(\mathbf{s}_T)\right]$
```

Figure 5.5. The process of r-ABOC algorithm.

Figure 5.5 illustrates the process of r-ABOC algorithm. In step (1), the weights of each sample are selected to be uniformly distributed for the entire training data set. Thus, the weights of the samples are initialized to $1/N$. In step (2), T classifiers are trained, as shown in steps (2a)-(2c). In step (2a), the classifier $g_t(\mathbf{s}_T)$ is trained with a different value of $f_T = 0.01 \times t$. In step (2b), the weight updating parameter α_t is selected. Next, the weight distributions for the next iteration (i.e., $D_{t+1}(i)$) are updated and normalized (step (2c)). After T iterations of step (2), the final hypothesis $\mathrm{G}(\mathbf{s}_T)$ is obtained as a linear combination of the T classifiers $g_t(\mathbf{s}_T)$ (step (3)), i.e., $\mathrm{G}(\mathbf{s}_T) = sign[\sum_{t=1}^{\mathrm{T}} \alpha_t g_t(\mathbf{s}_T)]$.

As it can be seen, it is very essential to choose the appropriate value of α_t at each round of r-ABOC. By unravelling the weight updating rule in step (2c), $D_{\mathrm{T}+1}(i)$ can be written as

$$D_{\mathrm{T}+1}(i) = \frac{\exp(-\sum_{t=1}^{\mathrm{T}} \alpha_t y_i g_t(\mathbf{s}_i))}{m \prod_{t=1}^{\mathrm{T}} Z_t} \tag{5.9}$$

where

$$\mathrm{Z}_t = \sum_{i=1}^{N} D_t(i) \exp(-\alpha_t y_i g_t(\mathbf{s}_i)) \tag{5.10}$$

Moreover, if $\mathrm{G}(\mathbf{s}_i) \neq y_i$, then $\sum_{t=1}^{\mathrm{T}} \alpha_t y_i g_t(\mathbf{s}_i) \leq 0$, implying that $\exp(-\sum_{t=1}^{\mathrm{T}} \alpha_t y_i g_t(\mathbf{s}_i)) \geq 1$. Thus,

$$[\![\mathrm{G}(\mathbf{s}_i) \neq y_i]\!] \leq \exp(-\sum_{t=1}^{\mathrm{T}} \alpha_t y_i g_t(\mathbf{s}_i)) \tag{5.11}$$

where $[\![\mathrm{G}(\mathbf{s}_i) \neq y_i]\!] = 1$ if $\mathrm{G}(\mathbf{s}_i) \neq y_i$, and $[\![\mathrm{G}(\mathbf{s}_i) \neq y_i]\!] = 0$ otherwise. Combining Equations (5.9) and (5.11) gives the bound on training error since

$$\frac{1}{N}\sum_{i=1}^{N} [\![\mathrm{G}(\mathbf{s}_i) \neq y_i]\!] \leq \frac{1}{N}\sum_{i=1}^{N} \exp(-\sum_{t=1}^{\mathrm{T}} \alpha_t y_i g_t(\mathbf{s}_i)) \tag{5.12}$$

Hence,

$$\frac{1}{N}\sum_{i=1}^{N} [\![\mathrm{G}(\mathbf{s}_i) \neq y_i]\!] \leq \sum_{i=1}^{N} \left(\prod_t Z_t\right) D_{\mathrm{T}+1}(i) \tag{5.13}$$

Since $\sum_{i=1}^{N} D_{\mathrm{T}+1}(i) = 1$, Equation (5.13) can be rewritten as

$$\frac{1}{N}\sum_{i=1}^{N} [\![\mathrm{G}(\mathbf{s}_i) \neq y_i]\!] \leq \prod_t Z_t \tag{5.14}$$

Equation (5.14) suggests that in order to minimize the training error, a reasonable approach might be to minimize Z_t at each round of r-ABOC. Since $g_t(\mathbf{s}_i) \in [-1,+1]$, implying that $y_i g_t(\mathbf{s}_i) \in [-1,+1]$, Z_t in Equation (5.10) is bounded as

$$\mathrm{Z}_t \leq \sum_{i=1}^{N} D_t(i)\left(\frac{1+y_i g_t(\mathbf{s}_i)}{2}\exp(-\alpha_t) + \frac{1-y_i g_t(\mathbf{s}_i)}{2}\exp(\alpha_t)\right) \tag{5.15}$$

The value of α_t is then obtained by zeroing the first derivative of the right hand side of Equation (5.15), i.e.,

$$\alpha_t = \frac{1}{2}\log\frac{\sum_{i=1}^{N} D_t(i)\big(1+y_i g_t(\mathbf{s}_i)\big)}{\sum_{i=1}^{N} D_t(i)\big(1-y_i g_t(\mathbf{s}_i)\big)} \tag{5.16}$$

The aim of the weight updating scheme in step (2c) is to decrease the weight of training samples, which are correctly classified and increase the weight of the opposite part [36], [37]. Hence, α_t should be a positive number. To ensure $\alpha_t > 0$, the following condition should hold:

$$\sum\nolimits_{i=1}^{N} D_t(i)(1+y_i g_t(\mathbf{s}_i)) > \sum\nolimits_{i=1}^{N} D_t(i)(1-y_i g_t(\mathbf{s}_i)) \tag{5.17}$$

5.2.3. Diversity among weak classifiers

The lack of performance improvement could be attributed to the lack of diversity in the ensemble [59], [60], [150], [151]. In general, an ensemble is not needed if there is a perfect classifier, which makes no errors. However, since classifiers do make errors, then a potential classifier should be complemented with others, which make errors on different samples. In [59], it was also observed that if all classifiers in the ensemble make errors on the same samples, combining them would result in overfitting. Hence, diversity of the classifier outputs is a key factor for the success of the classifier ensemble learning algorithms [60].

As previously mentioned, r-ABOC is different from the conventional Real AdaBoost because it builds a set of diverse classifiers using the entire training data set, but with different fraction rejections. This section is dedicated to demonstrate that the new training procedure is capable of generating the necessary diversity to make r-ABOC work efficiently.

Table 5.3. Average HTER (%) of r-ABOC and the conventional Real AdaBoost, as applied to combine $T \in \{5,10,15,20,25,50\}$ *classifiers over the four sets of experiments*

	$T = 5$	$T = 10$	$T = 15$	$T = 20$	$T = 25$	$T = 50$
Experiment (1)						
r-ABOC	**4.57**	**4.38**	**4.36**	**4.36**	**4.36**	**4.36**
Real AdaBoost	6.13	6.25	6.14	6.11	6.11	6.11
Experiment (2)						
r-ABOC	**1.01**	**0.98**	**0.98**	**0.98**	**0.98**	**0.98**
Real AdaBoost	1.08	1.05	1.05	1.04	1.04	1.04
Experiment (3)						
r-ABOC	**0.29**	**0.29**	**0.29**	**0.29**	**0.29**	**0.29**
Real AdaBoost	0.36	0.32	0.32	0.32	0.32	0.32
Experiment (4)						
r-ABOC	**1.90**	**1.89**	**1.85**	**1.85**	**1.85**	**1.85**
Real AdaBoost	2.47	2.38	2.38	2.38	2.38	2.38

It should be noted that measuring diversity is not straightforward since there is no formal definition of what is perceived as diversity [59]. Hence, in this research, a MATLAB program was implemented to directly gauge the conventional Real AdaBoost and r-ABOC as applied to combine $T \in \{5,10,15,20,25,50\}$ classifiers. The classifiers in Real AdaBoost were learned on different data subsets, which were generated by randomly removing the impostor samples until the number of impostor and genuine user samples are equal [35]. The fraction rejection value in r-ABOC was selected in the range $[0.01, \dots, 0.01 \times T]$. From Table 5.3, the following observations can be made:

- The conventional Real AdaBoost is not able to provide a satisfactory performance in all the experiments. Increasing T does not lead to a significant change in its error rates. In some cases (see Experiment (1)), doing so may result in overfitting. A possible reason for this is due to the lack of variability/diversity among the individual classifiers in Real AdaBoost. Moreover, the fraction rejection value, used in training these classifiers is optimised on the training set. It cannot guarantee the optimal performance on the testing set.

- The proposed r-ABOC consistently produces higher verification accuracy with respect to that of Real AdaBoost. This implies that the new training procedure is a better alternative in order to generate the diversity in the classifier combination. The issue here is whether the combination is justified. *Would it be able to achieve lower error rates with respect to the single best classifier?* These questions will be addressed in Sections 5.2.6-5.2.11

5.2.4. Other Related Classifier Ensemble Approaches

Many alternative algorithms can be applied in classifier ensemble learning. Bagging [142] has been also reported to be successful at the bias and variance reduction. Bagging is known as bootstrap aggregating. Its basic aim is to fit and combine T weak classifiers using T new training sets by uniformly sampling samples from the original training data with replacement. Hence, bagging is not suitable for combining the classifiers, which are generated using different values of the fraction rejection.

Random Forest [138] is a variant of Bagging. The difference lies in the construction of the decision tree. The best feature at each node is selected among M randomly chosen features, where M is the parameter of the algorithm. Random Forest specially designed for decision tree classifiers. It also requires a sufficient number of data features in order to perform well, and hence, reducing its applicability to multimodal biometric score data with only two or three features.

In [37], [145], several cost-sensitive boosting techniques for imbalance learning, such as AdaCost, AdaC1, AdaC2, and AdaC3 have been proposed, which are motivated by the work of Discrete AdaBoost. Their focus is on how to introduce the cost into the weight updating rule of the Discrete AdaBoost and how to find an appropriate cost matrix. In general, the costs denote the uneven classification importance among classes, such that the boosting algorithms can intentionally bias the learning towards the small class, and thus, improving the classification performance on it [37]. While various heuristic manipulations are available, these manipulations do not guarantee the asymptotic convergence to a good cost-sensitive decision rule [146]. Moreover, given a data set, the cost matrix is often unavailable. Genetic Algorithm (GA) has been widely employed to find the cost matrix; however, it can be very time-consuming due to its very nature. Because of these reasons, cost-sensitive boosting techniques will not be further considered.

The "AND" and "OR" decision rules are the simplest means of combining the class label of the weak classifiers, which is implemented as $h_t(\mathbf{s}_T) = sign[g_t(\mathbf{s}_T)]$. The decision of the "AND" rule is +1 only when all the classifiers produce +1 outputs. On the contrary, the "OR" rule outputs +1 as long as the decision of at least one classifier is +1. The "AND" and "OR" rules are duals of each others. The "OR" results in lowering the FRR but increases the FAR, while the "AND" results in lowering FAR but increases FRR. In [139], the authors observed that these decision rules could yield worse performance that the best of the individual classifiers. Due to this reason, the "AND" and "OR" decision rules are rarely used in practice [1]. The Majority Vote is perhaps the most commonly used method for decision making. Given a set of $h_t(\mathbf{s}_T), t = 1, \dots, \mathrm{T}$, the Majority Vote gives an accurate class label if at least $\mathrm{T}/2 + 1$ classifiers give correct decision. In [59], the accuracy of the Majority Vote is given by

$$P_{maj} = \sum_{i=\frac{\mathrm{T}}{2}+1}^{\mathrm{T}} \binom{\mathrm{T}}{i} p^i (1-p)^{\mathrm{T}-i} \tag{5.18}$$

where p is the probability for each classifier to give the correct class label. This supports the intuition that one can expect performance improvement over the individual accuracy p only when p is higher than random guessing, i.e., $p > 0.5$. This observation is valuable even for unequal p [140]. In [141], a theoretical analysis was done to establish the limits on the accuracy of the Majority Vote based on the number of classifiers, the pair-wise dependence between these classifiers, and their individual accuracy.

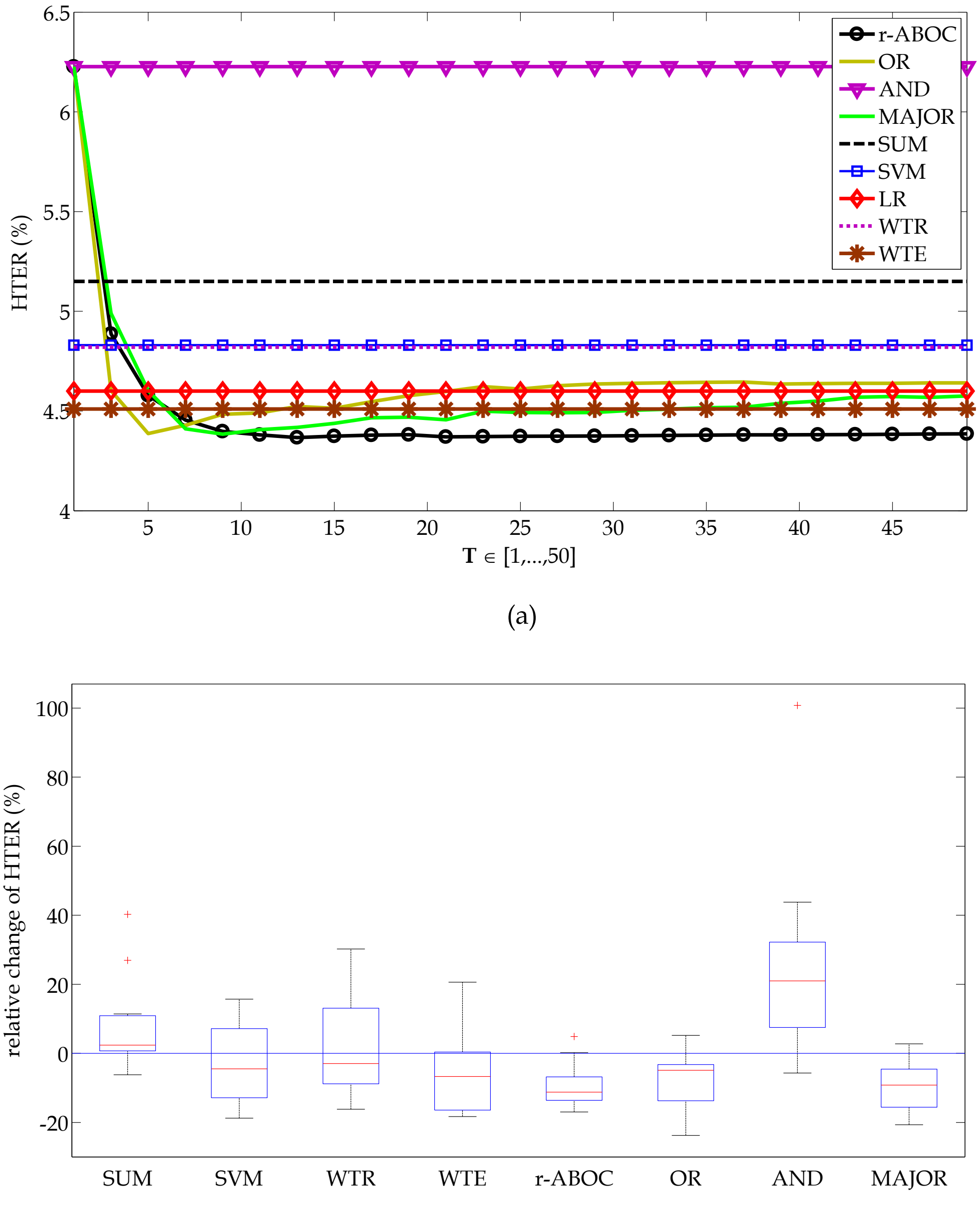
In general, the "AND", "OR", and Majority Vote decision rules have the following distinct advantages: (1) no a priori knowledge about the classifiers is needed; (2) no training is required to come up with the final decision [1]. Apart from r-ABOC, these decision rules will be evaluated in this research.

5.2.5. Experimental Setup

In the reminder of Section 5.2, an extensive empirical evaluation is carried out using data from the BioSecure DS2 and XM2VTS databases. The performance of various classifier combination algorithms, such as "AND" (abbreviated as AND), "OR" (abbreviated as OR), Majority Vote (abbreviated as MAJOR) decision rules and r-ABOC will be evaluated with different numbers of weak classifiers T. In these experiments, the maximum T that is selected is 50. The rationale for this is that T is much dependent on the choice of the fraction rejection f_T. For $\mathrm{T} = 50$, the range of f_T should be from 0.01 and 0.5 (i.e., $0.01 \times \mathrm{T}$). Because the data collection is assumed to be accurate and there is a low density of noisy samples in the training data, a rejection rate of more than 50% (i.e., $f_T = 0.5$) of the training samples is considered excessive [44].

It should be noted that the classifier combination algorithms, investigated and developed in this research, work independently of the claimed identity. Hence, experiments were only conducted to compare these algorithms with other state-of-the-art user-independent solutions, namely the sum of scores (abbreviated as SUM), two-class SVM (abbreviated as SVM), and likelihood ratio based score fusion (abbreviated as LR). The results, corresponding to the weak classifier, given in Equation (5.6) are also presented in this section for comparison purposes. The weak classifier will be evaluated in both scenarios: (1) the fraction rejection is optimized on the training set (WTR) and (2) the fraction rejection is directly on the testing set (WTE).

Similarly to Section 5.1, four sets of experiments will be presented using the BioSecure DS2 and XM2VTS databases. Because AND, OR and MAJOR are only able to produce the hard class label outcomes, their error rates cannot be tuned with different decision thresholds. In other words, it is not possible to obtain their EER. Due to this reason, the a priori HTER and relative change of HTER will be used for performance reporting. In order to determine the relative change of HTER, LR is selected as the baseline system. The obtained results will be demonstrated using the box plot representation.



(a)

(b)

Figure 5.6. Combining two biometric matchers of the BioSecure DS2 database: (a) Average HTER (%) as a function of the number of weak classifiers, used in classifier ensembles, (b) Relative change of HTER (%) across 13 combination possibilities of SUM, SVM, WTR, WTE, r-ABOC, OR (T = 5), *AND, and MAJOR* (T = 7) *with respect to LR.*

5.2.6. Experiment (1): Combining Two Biometric Matchers of the BioSecure DS2 Database

Figure 5.6(a) shows the performance of SUM, SVM, LR, WTR, WTE, r-ABOC, OR, AND, and MAJOR schemes as the functions of the number of weak classifiers in the classifier combination. As it can be seen, WTE is better than other single classifier models as it provides an average HTER of 4.51%. This is followed by LR (4.60% HTER), WTR (4.82% HTER), SVM (4.83% HTER) and SUM (5.15% HTER).

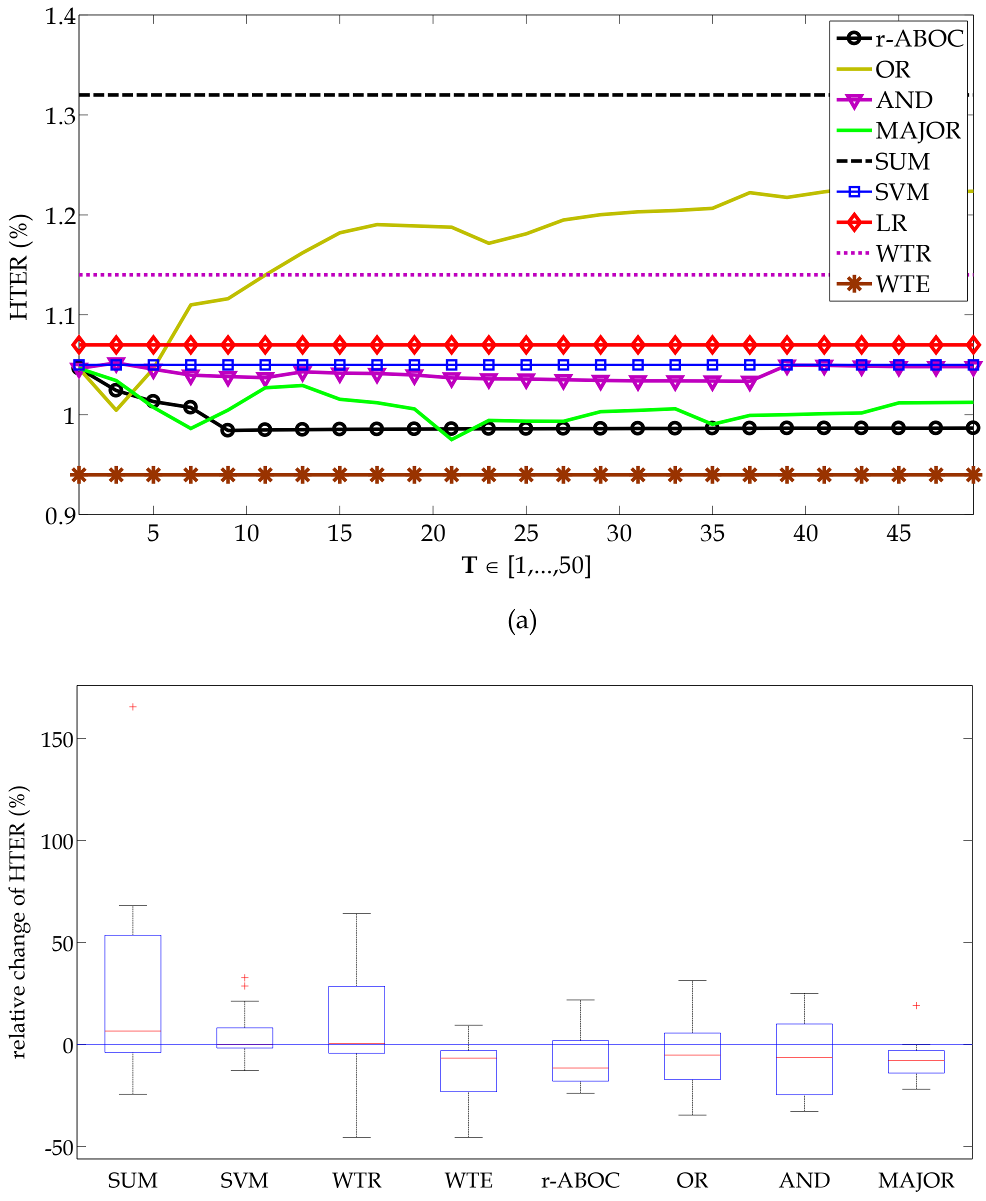
AND is the worst performer. It is observed that its error rates (6.24% HTER) are not even decreased when the value of T is increased. OR has its lowest HTER (4.42%) at $T = 5$, while that of MAJOR is found at $T = 7$. r-ABOC achieves the highest authentication accuracy (4.36% HTER), providing an HTER improvement of 3%, 5%, 10%, 10%, and 15% as compared to WTE, LR, WTR, SVM and SUM, respectively.

It should be noted that while the performance of OR and MAJOR seems to be susceptible to the selection of T, that of r-ABOC becomes stable when $T > 10$. This in turn implies that r-ABOC is more robust to the choice of this parameter.

In terms of relative change of HTER, shown in Figure 5.6(b), SVM, WTR, WTE, r-ABOC, OR and MAJOR demonstrate an improvement over the baseline system, i.e., LR, since their median values are less than zero. MAJOR has mostly similar median value to that of WTE, while the lowest median value, indicating the best performance, is attributed to r-ABOC.

5.2.7. Experiment (2): Combining Two Biometric Matchers of the XM2VTS LP1 Database

Similarly to Experiment (1), for r-ABOC, OR, AND, and MAJOR various values of $T \in [1, \dots, 50]$ are tested. The performance is shown in Figure 5.7(a). The lowest HTER of SVM is 1.05%, which is either superior or comparable to that of LR (1.07% HTER), WTR (1.14% HTER) and SUM (1.32% HTER). WTE (0.94% HTER) demonstrates the best performance when the value of f_T is optimized directly on the testing set.



(a)

(b)

Figure 5.7. Combining two biometric matchers of the XM2VTS LP1 database: (a) Average HTER (%) as a function of the number of weak classifiers, used in classifier ensembles, (b) Relative change of HTER (%) across 15 combination possibilities of SUM, SVM, WTR, WTE, r-ABOC, OR (T = 3)*, AND* (T = 38)*, and MAJOR* (T = 21) *with respect to LR.*

AND has its lowest error rate (1.03% HTER) at T = 38, while that of OR (1.00% HTER) is found at T = 3. OR seems to be highly dependent on the selection of T as its performance significantly decreases for T ≠ 3. The lowest HTER, achieved by r-ABOC and MAJOR is 0.98%, which demonstrates a relative improvement of 7%, 8%, 14% and 26% as compared to SVM, LR, WTR and SUM, respectively. It should be noted that r-ABOC is able to consistently produce a good performance for T > 10.

As illustrated in Figure 5.7(b), the highest performance in terms of relative change of HTER is attributed to r-ABOC and MAJOR. WTE, AND, and OR have median values less than zero, indicating a performance improvement over LR.

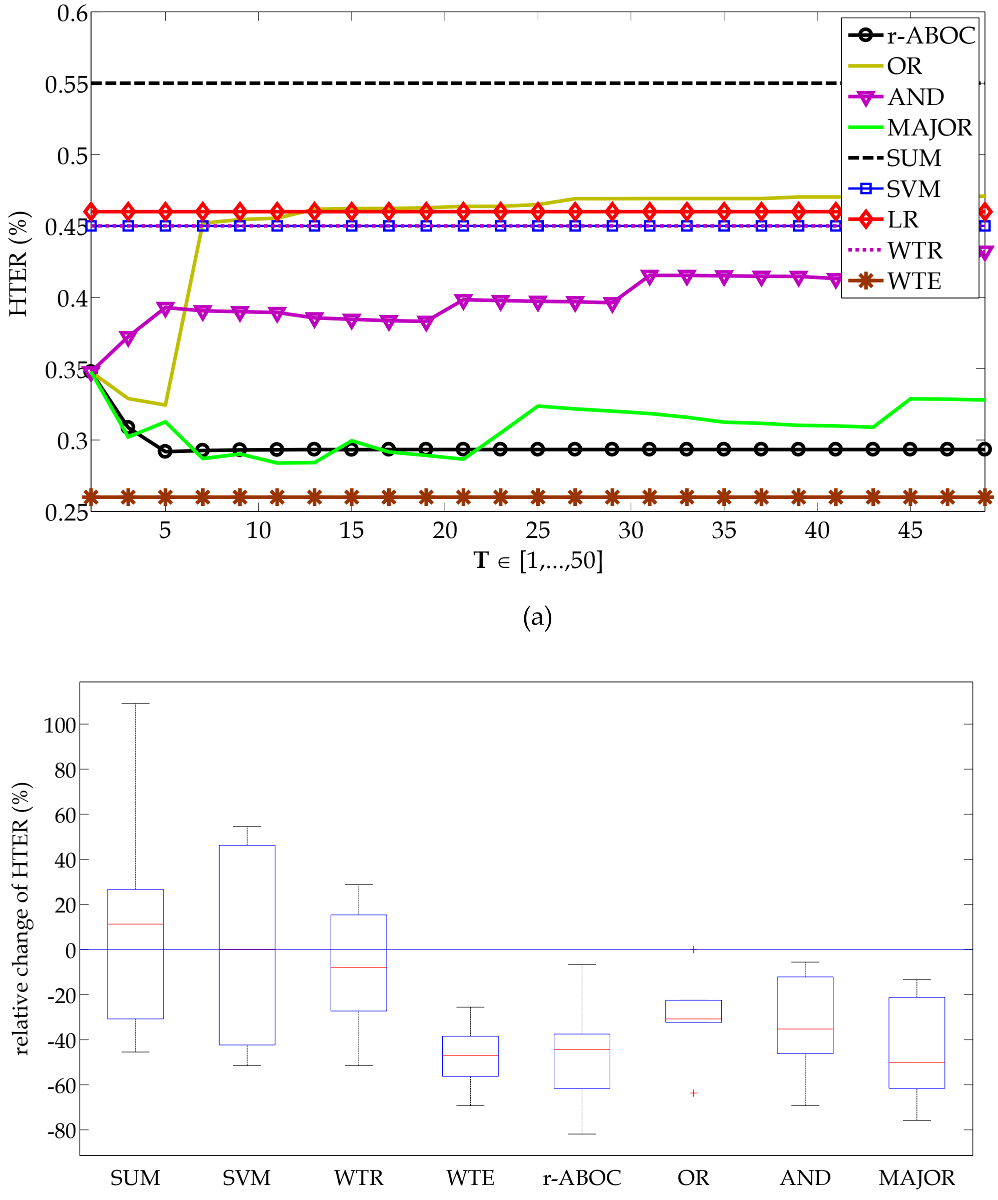
5.2.8. Experiment (3): Combining Two Biometric Matchers of the XM2VTS LP2 Database

In Figure 5.8(a), it is demonstrated that SUM with an average HTER of 0.55% is the worst performer. LR has a mostly similar error rate to that of WTR and SVM. It is observed that when the value of fraction rejection f_T is optimized directly on the testing set, WTE offers significant advantages as it is shown to produce the highest authentication accuracy (HTER of 0.26%) in this experiment.

OR, AND, MAJOR and r-ABOC are able to provide a better performance as compared to the single classifier models. MAJOR has its lowest HTER (0.28%) at T = 13, while those of OR (0.32% HTER) and AND (0.35% HTER) are achieved at T = 5, and T = 1, respectively.

The performance of OR, AND, and MAJOR is shown to be highly sensitive to the choice of T. Similarly to Experiment (1), an average HTER of 0.29% is consistently achieved by r-ABOC for T > 5, which provides a relative improvement of 29%, 29% and 47% as compared to LR, SVM and SUM, respectively.

Concerning the relative change of HTER in Figure 5.8(b), apart from SUM and SVM, all other techniques demonstrate an improvement over the baseline system, i.e., LR since their median values are less than zero. AND has slightly lower median value (i.e., better performance) than OR. The proposed r-ABOC and WTE are shown to be better alternative relative to these techniques, while the lowest median value is achieved by MAJOR.



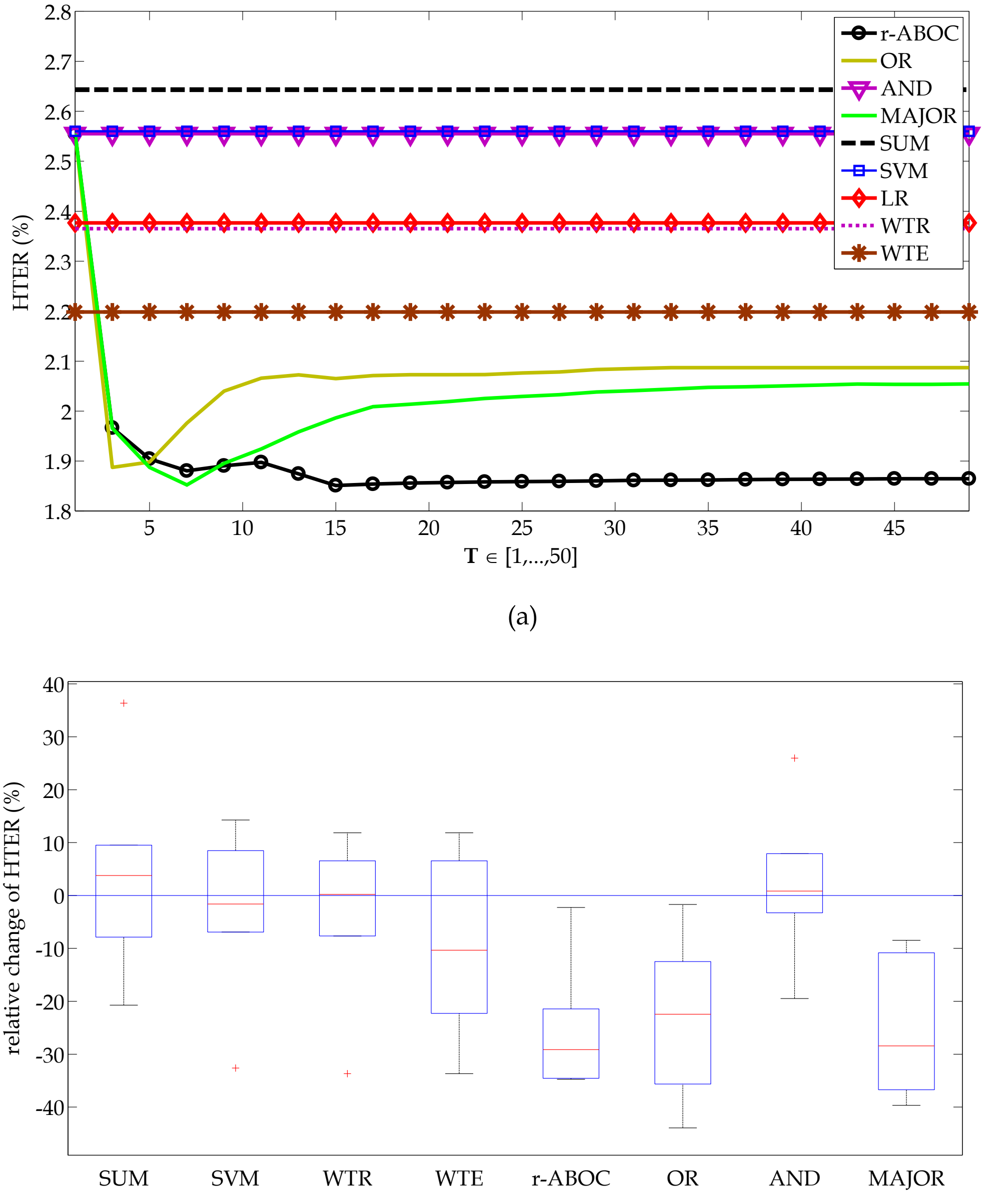
(a)

(b)

Figure 5.8. Combining two biometric matchers of the XM2VTS LP2 database: (a) Average HTER (%) as a function of the number of weak classifiers, used in classifier ensembles, (b) Relative change of HTER (%) across 6 combination possibilities of SUM, SVM, WTE, WTE, r-ABOC, OR (T = 5)*, AND* (T = 1)*, and MAJOR* (T = 13) *with respect to LR.*

5.2.9. Experiment (4): Combining Three Biometric Matchers of the BioSecure DS2 Database

Figure 5.9(a) shows the performance in terms of HTER for SUM, SVM, LR, WTR, WTE, r-ABOC, OR, AND, and MAJOR as applied in combining three biometric matchers from the BioSecure DS2 database. The authentication accuracy of all these schemes is significantly improved, as compared to the two biometrics case in Experiment (1) (see Section 5.2.6).



(a)

(b)

Figure 5.9. Combining three biometric matchers of the BioSecure DS2 database: (a) Average HTER (%) as a function of the number of weak classifiers, used in classifier ensembles, (b) Relative change of HTER (%) across 6 combination possibilities of SUM, SVM, WTR, WTE, r-ABOC, OR (T = 3)*, AND, and MAJOR* (T = 7) *with respect to LR.*

WTE demonstrates the lowest error rate (2.19% HTER) among the single classifier models, including WTR (2.36% HTER), LR (2.38% HTER), and SUM (2.64% HTER). The performance of SUM is similar to that of AND. Similarly to

Experiment (1), the error rates of AND are not decreased when the value of T is increased.

OR has its lowest HTER (1.88%) at T = 3. The best authentication accuracy is achieved by r-ABOC and MAJOR (1.85% HTER), providing a relative improvement of 22%, 22%, 27% and 30% with respect to WTR, LR, SVM and SUM, respectively. The error rates achieved by OR, and MAJOR are shown to be highly sensitive to the choice of T, while that of r-ABOC becomes unchanged for T > 15.

With regards to the relative change of HTER, illustrated in Figure 5.9(b), WTE, OR, MAJOR and r-ABOC are shown to be better than the baseline system, i.e., LR. Overall, the lowest median value is provided by r-ABOC.

5.2.10. Computational Complexity

Figure 5.10 shows the evaluation time of r-ABOC. In most cases, it is possible to carry out the training off-line and the training time is not a major concern. For practical uses, evaluation time might be critical. The evaluation time was recorded on an Intel(R) 2.00 GHz.

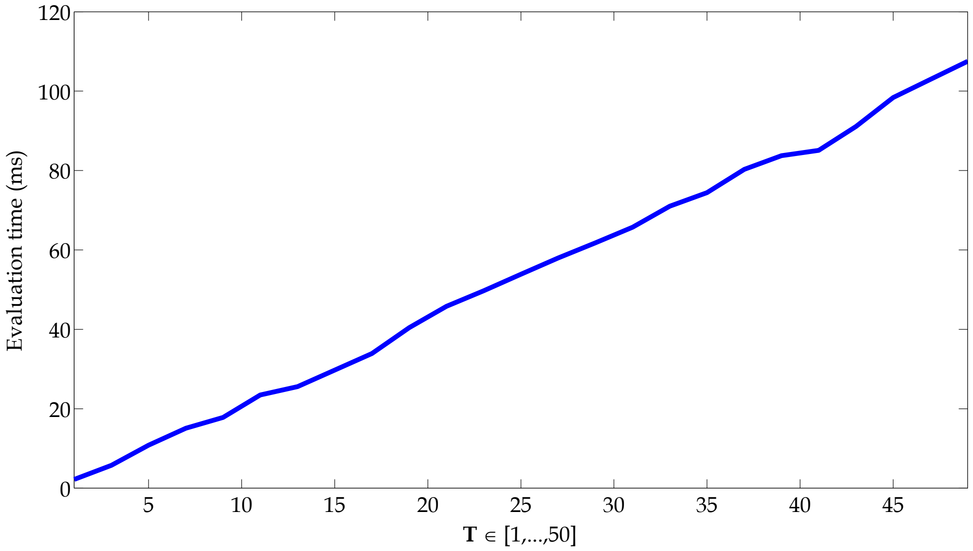


Figure 5.10. The evaluation time of r-ABOC as a function of the number of weak classifiers

In general, the evaluation time of r-ABOC is not comparable to that of SUM, LR and SVM since it has to render the biometric decision using the outcomes of multiple weak classifiers. It is observed that its computational complexity is almost linear and increases with the number of classifiers in the ensemble. However, it should be noted that at $T = 50$, the evaluation time of this algorithm is approximately 110 ms, which is still suitable for most practical multimodal biometric systems.

5.2.11. Summary and Discussion

Tables 5.4 and 5.5 summarize the findings of the previous experiments (Sections 5.2.6-5.2.9), by ranking the techniques based on their performance in terms of the average HTER, and relative change of HTER. The following observations can be made:

Table 5.4. Ranking of different approaches with respect to the HTER

	Experiment (1)	Experiment (2)	Experiment (3)	Experiment (4)	Mean
r-ABOC	**1**	3	3	**1**	2.00
OR	3	4	4	3	3.50
AND	9	5	5	7	6.50
MAJOR	2	2	2	**1**	**1.75**
SUM	8	9	9	9	8.75
SVM	6	6	6	7	6.25
LR	5	7	6	6	6.00
WTR	6	8	6	5	6.25
WTE	4	**1**	**1**	4	2.50

Table 5.5. Ranking of different approaches with respect to the relative change of HTER

	Experiment (1)	Experiment (2)	Experiment (3)	Experiment (4)	Mean
r-ABOC	**1**	**1**	3	**1**	**1.50**
OR	4	5	5	3	4.25
AND	9	3	4	8	6.00
MAJOR	2	2	**1**	2	1.75
SUM	8	9	9	9	8.75
SVM	4	6	7	5	5.50
LR	7	6	7	6	6.50
WTR	6	6	8	6	6.50
WTE	3	3	2	4	3.00

- When the training data is sufficiently representative to guarantee the parameter estimation and the optimal value of the fraction rejection f_T is determined, WTE is demonstrated to be better with respect to other state-of-the-art solutions, namely SUM, SVM and LR in all the experiments. In practice, training data may be significantly different from the testing data because they are collected in different sessions. It is therefore difficult to determine the optimal value of f_T. The results of WTR, which optimizes f_T on the training data, further confirm this observation. Despite that, it should be noted that WTR is able to achieve either better or comparable performance to that of the LR and SUM. Its error rates are only slightly higher than those of the SVM.
- MAJOR, r-ABOC and OR are among the best performers. Indeed, MAJOR and r-ABOC are demonstrated to achieve lower error rates, as compared to WTE. Although WTE is based on an unrealistic assumption and is impractical to achieve, its corresponding results are still presented in this section in order to give the answer to the question, which arose in Section 5.2.3. The results also highlight the distinct advantages of r-ABOC, which is able to provide lower error rates relative to the single best classifier.
- It is observed that the authentication accuracy of MAJOR and OR is highly susceptible to the selection of the number of the weak classifiers T, while that of r-ABOC becomes unchanged for $\mathrm{T} \in [15, \dots ,50]$. This, in turn, implies that the proposed r-ABOC is naturally robust to the selection of this parameter. As already mentioned, T is directly related to the fraction rejection. Hence, it can be concluded that r-ABOC is able to eliminate the risk of making a bad choice of the fraction rejection for the multimodal biometric authentication problem. In practice, one may select a reasonably large value of T (e.g., $\mathrm{T} = 50$) for this algorithm to provide its lowest error rates. The lowest error rates are achieved at the expense of increased complexity. However, even when $\mathrm{T} = 50$, the increased complexity is still suitable for the practical biometric authentication systems.

5.3. Chapter Summary

In this chapter, an adapted score fusion scheme and a novel hybrid boosting algorithm, called r-ABOC have been developed to advance the classification performance of extremely imbalanced class distribution in multimodal biometric systems. Both these approaches are based on the Bayes Decision Theory, as applied in combining one-class classifiers to effectively use the training data from both the genuine user and impostor classes and efficiently overcome the class imbalance problem.

As it can be seen, the adapted score fusion is capable of exploiting both the user-specific information and general class knowledge, provided by a number of different users. It offers many advantages over the stage-of-the-art solutions as it is able to completely or partially overcome the within-class sub-concepts in the impostor score distribution and training data scarcity of the genuine user class samples. The adapted score fusion scheme has been tested with different one-class classifiers. It was observed to consistently outperform other techniques evaluated, providing a relative improvement of 20% and 10% as compared to the likelihood ratio based score fusion and adapted user-dependent fusion. Among the one-class classifiers, k-NN and GMM provides the lowest error rates. In addition, these classifiers are insensitive to the choice of the fraction rejection, and thus, are more amendable to immediate applications.

One the other hand, r-ABOC inherits the natural capabilities of Real AdaBoost to combine multiple classifiers in order to find a highly accurate classifier system. However, unlike the conventional Real AdaBoost, the individual classifiers in r-ABOC were trained on the same data set, but with different fraction rejection values. It has been shown that this training procedure not only generates the necessary diversity to make r-ABOC work efficiently, but also eliminates the risk of making a bad choice of the fraction rejection for the problem at hand. Extensive experiments, carried out on the BioSecure DS2 and XM2VTS databases, demonstrated the potential of the proposed r-ABOC, which provides a significant relative improvement of 28%, 24% and 22% with respect to the sum of score, likelihood ratio based score fusion and SVM, respectively. It is also considered as a better alternative as compared to other classifier ensemble learning algorithms, namely the "AND", "OR" and Majority Vote decision rules. Another important feature of r-ABOC is that it does not require any parameter fine-tuning, making it easily handled in practical multimodal biometric systems.

CHAPTER 6

CONCLUSIONS AND FUTURE WORK

This chapter will start by summarizing the main contributions of the research reported in this thesis (Section 6.1). Next, the possible extensions and recommendations for future research will be given in Section 6.2.

6.1. Conclusions

Biometric authentication is the process of verifying a human identity using his behavioural and physiological characteristics. It is well-known that multimodal biometric systems can further improve the verification performance of a biometric system by combining the evidence, presented by multiple biometric traits at various levels, namely, sensor, feature, match score and decision levels. Fusion at match score level is generally preferred due to the trade-off between information availability and fusion complexity and is normally treated as a two-class classification problem, i.e., genuine user and impostor. However, as previously mentioned, two-class methods suffer when applied to imbalanced data sets.

In this research, the paradigm of one-class classification has been exploited to advance the classification performance of extremely imbalanced biometric data sets. A robust imputation technique based on Group Method of Data Handling (RIBG) was also developed to handle incomplete match score vectors in multimodal modal biometric system. During a series of experiments, carried out on the BioSecure DS2 and XM2VTS databases, four main scientific achievements can be identified in this research:

(1) **RIBG, a robust imputation technique** was proposed in Section 3.4 to handle incomplete match score vectors in BioSecure DS2 database, which is primary problem of any match score level fusion techniques. RIBG is based on Group Method of Data Handling. It is able to find an optimal structure of model to provide accurate predictions for the missing elements in multimodal biometric systems. From a comprehensive empirical analysis, RIBG was observed to outperform other state-of-the-art imputation techniques, namely mean and median imputations. It is also comparable to k-NN imputation. However, while k-NN requires a careful selection of the number of nearest neighbours to achieve a satisfactory performance, RIBG does not require any parameter fine-tuning.

(2) **Design of user-specific score fusion**: In Chapter 4, the one-class classifiers were employed to learn the user-specific and user-independent descriptions around either the impostor or genuine user class. It was illustrated that user-specific approach is a better alternative and consistently demonstrates a better authentication performance when compared to user-independent approaches. It is also shown to be able to partially overcome the problem of within-class sub-concepts, which arises when the target class is scattered into several small regions due to the existence of user variations. Various one-class classification methods, such as Gaussian Mixture Model (GMM), k-Nearest Neighbour (k-NN), K-means clustering (K-means), and Support Vector Data Description (SVDD) have been applied in designing user-specific score fusion. There is no single best classifier for all problems. However, they are all demonstrated to be better than the standard two-class SVM on the experiments, carried out on the BioSecure DS2 database, where the class imbalance is on the order of 524:1. This in turn implies that one-class methods are particularly useful in handling the extremely imbalanced class distributions of biometric data sets. As it is expected, SVDD, K-means are most sensitive to the scaling of the data since they heavily reply on the distance of the test samples to the center of the sphere or the prototype objects. On the other hand, GMM and k-NN seem to be most influenced by

the sample size. In term of computational complexity, SVDD and k-NN are the most expensive, while K-means shows the least evaluation time. GMM demonstrates the best trade-off between authentication accuracy and computational complexity.

(3) **Adapted Score Fusion Scheme**: In Section 5.1, a novel adapted score fusion scheme based on both user-specific information and general class characteristics, has been developed for multimodal biometric authentication. It is different from the well-known Adapted User-Dependent Fusion (AUDF), which is based on two-class SVM, because it consists of one-class classifiers, trained using training data from both the genuine user and impostor classes. It is also shown to offer many distinct advantages over the state-of-the-art solutions as it can completely or partially alleviate the problems of training data scarcity and imbalanced class distribution. Extensive experiments with various one-class classifiers were carried out in order to gauge the adapted score fusion scheme. They demonstrated that the proposed scheme is able to provide an improvement in Equal Error Rate (EER) of 32%, 27% and 20% with respect to the user-independent approaches, such as the sum of scores, two-class SVM, and likelihood ratio based score fusion, respectively. It is also either comparable or better to the AUDF. The comparable result was achieved only when experiments were conducted on XM2VTS LP1 database and three genuine samples per user were available for training the user-specific descriptions. The proposed adapted score fusion scheme was tested with various one-class classification methods. Among others, it was noted that k-NN and GMM provide the lowest error rate. Moreover, these classifiers were shown to be insensitive to the choice of fraction rejection, hence, making them amenable to implementation without requiring fine-tuning.

(4) **r-ABOC, a hybrid Boosting Algorithm** was developed, which was capable of exploiting the natural capabilities of both Real AdaBoost and one-class classification. This algorithm works by developing a weak classifier, which

is also based on the Bayes Decision Theory as applied in combining the one-class methods in order to make use of training data from both classes. Real AdaBoost is then applied to further improve the performance of the weak classifier without causing overfitting. However, unlike the conventional Real AdaBoost, the weak classifiers in r-ABOC are learned on the same data set but with different values of the fraction rejection. It has been shown that this training procedure is able to generate a high diversity, which is the key requirement for the success of this algorithm. An extensive empirical evaluation was carried out to illustrate the effectiveness of r-ABOC. Overall, the proposed algorithm was shown to consistently achieve better performance, relative to the "AND", "OR" decision rules and other state-of-the-art solutions, such as the sum of scores, likelihood ratio based score fusion and SVM. It is also comparable to the Majority Vote decision rule. However, while the Majority Vote decision rule is highly sensitive to the choice of the number of weak classifiers in the ensemble T, the performance of r-ABOC does not change much for $T > 15$. It should be noted that T is directly related to the choice of the fraction rejection. This in turn implies that r-ABOC is able to completely eliminate the possibility of making bad choice of this parameter for the practical biometric authentication problem.

6.2. Future Work

Several research directions arise from the dissertation work are enumerated as follows:

(1) It was observed in Chapter 4 that none of the one-class classification methods shows a clear dominance above the remaining ones, which confirms the "no panacea" principle in pattern recognition. It means that defining the best fitting data description for a given biometric data as well as building a user-specific classification model for each of the users, enrolled in the system, is still an open field for researchers. It is also possible to combine the various classifiers to further improve the

authentication accuracy, while increasing the robustness of the classification [163].

(2) There is growing evidence that making use of cohort templates can improve biometric verification performance [70], [71]. The match scores, obtained using the cohort temples, are called cohort match scores, which are captured online by comparing the query with a set of competing templates, including the template of the claimed identity. These match scores are subject to the same degradation, and are expected to reduce the effect of varying score distributions due to the degradation factor, caused by the change in environmental conditions, and the user interaction. Future research can be focused on the incorporation of cohort match scores into the proposed schemes of this thesis order to improve the learning capabilities of different characteristics of impostor class distributions, and thus, enhance the verification performance of the user-specific and adapted score fusion schemes.

(3) It is also possible to capture the various degradation factors by using a set of quality measures. Very often, high quality values are associated with good verification performance, and vice versa for low quality values. Thus, it is reasonable to consider quality measures of the input biometric signals and weight the contributions of various biometric traits based on this information. A common practice in many reported works is to treat quality measures as another set of features, which is then fed into the classifiers in order to render the biometric decision [15]. A similar idea can be exploited to further enhance the performance of the match score fusion schemes, developed in this thesis.

(4) The user-specific and adapted score fusion schemes were employed to perform the classification task on the BioSecure DS2 and XM2VTS databases, where a limited (up to three) genuine samples can be used for learning the user-specific descriptions. It is therefore important to investigate their behaviours on a different database with a large availability

of training data from the genuine user class, and hence, to determine whether they are consistently better or at what level they are better than state-of-the-art solutions. Nevertheless, there is no publicly available database, which can serve this purpose. Hence, a new database should be also developed in order to address these questions.

(5) The proposed r-ABOC, developed in this work, provides the hard class label outcomes by assuming that the decision threshold is firmly selected as 0. Although it was demonstrated to be better than the state-of-the-art solutions in terms of a priori HTER, its limitation is that it is not easy to fix one type of errors (say FAR) and compute the FRR at the specified FAR. As already mentioned, the system requirements, concerning the authentication accuracy are very much dependent on the application. Some applications tend to have a low FRR to make them more tolerant to the input variations, while others require an extremely low FAR to not let in the impostors. Hence, future research is also focused on the classifier ensemble algorithm, where the trade-off between the FAR and FRR can be adjusted to better meet the system requirements.

REFERENCES

[1] A. Ross, K. Nandakumar, and A. K. Jain, *Handbook of Multibiometrics*, Springer, 2006.

[2] R. Heyer, "Biometrics Technology Review 2008", 2008. Available at http://dspace.dsto.defence.gov.au/dspace/handle/1947/9704.

[3] J. Elliott, "Biometrics roadmap for police applications," *BT technology journal*, vol. 23, no. 4, pp. 37-44, 2005.

[4] M. C. Fairhurst, S. T. France, and J. M. Matthias, "Promoting Biometrics in the UK: Bridging the gap between research and exploitation," *Cyber Security Knowledge Transfer Network,* University of Kent, 2008.

[5] A. K. Jain, and A. Kumar, "Biometrics of next generation: An overview," *Second Generation Biometrics,* 2010.

[6] A. A. Moenssens, *Fingerprint Techniques*, Chilton Book Company, 1971.

[7] J. Daugman, "Iris recognition border-crossing system in the UAE," *International Airport Review,* vol. 8, no. 2, 2004.

[8] Y. B. Kwon, "Biometrics in Asia", 2014. Available at http://biometrics.org.

[9] Planning Commission, "Ensuring Uniqueness: Collecting iris biometrics for the Unique ID Mission". Available at http://eprints.cscsarchive.org.

[10] UKBA, "Using IRIS to enter the UK", 27 Feb. 2013. Available at http://www.ukba.homeoffice.gov.uk.

[11] A. Ross, K. Nandakumar, and A. K. Jain, "Introduction to multibiometrics," *Handbook of biometrics*, Springer US, pp. 271-292, 2008.

[12] A. Ross, "Information fusion in fingerprint authentication," Ph.D. dissertation, Michigan State University, 2003.

[13] K. Nandakumar, "Multibiometric Systems: Fusion Strategies and Template Security," Ph.D. dissertation, Michigan State University, 2008.

[14] J. Fierrez-Aguilar, "Adapted Fusion Schemes for Multimodal Biometric Authentication," Ph.D. dissertation, Technical University of Madrid (UPM), 2006.

[15] N. Poh, “Multi-system Biometric Authentication: Optimal Fusion and User-Specific Information,” Ph.D. dissertation, Dept. STI, Univ. EPFL, Lausanne, Switzerland, 2006.

[16] D. Maltoni, D. Maio, A. K. Jain, and S. Prabhakar, *Handbook of fingerprint recognition*, Springer, 2009.

[17] A. K. Jain, and S. Z. Li, *Handbook of face recognition*, Springer, 2005.

[18] A. F. Abate, M. Nappi, D. Riccio, and G. Sabatino, “2D and 3D face recognition: A survey,” *Pattern Recognition Letters*, vol. 28, no. 14, pp. 1885-1906, 2007.

[19] J. Daugman, “How iris recognition works,” *IEEE Trans. Circuits and Systems for Video Technology,* vol. 14, no. 1, pp. 21-30, 2004.

[20] J. Daugman, “High confidence visual recognition of persons by a test of statistical independence,” *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 15, no. 11, pp. 1148-1161, 1993.

[21] D. D. Zhang, *Palmprint authentication*, vol. 3, Springer, 2004.

[22] A. Ross, J. Shah, and A. K. Jain, “Towards Reconstructing Fingerprints from Minutiae Points,” In: *Proc. SPIE Conference on Biometric Technology for Human Identification II*, vol. 5779, pp. 68-80, Orlando, USA, 2005.

[23] A. Ross, and A. K. Jain, “Multimodal biometrics: An overview,” In: *Proc. of 12th European Signal Processing Conference*, pp. 1221-1224, Sep. 2004.

[24] R. Brunelli, and D. Falavigna, “Person Identification Using Multiple Cues,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 17, no. 10, pp. 995-966, Oct. 1995.

[25] P. Verlinde, and G. Cholet, "Comparing decision fusion paradigms using k-NN based classifiers, decision trees and logistic regression in a multi-modal identity verification application," In: *Proc. Int. Conf. Audio and Video-Based Biometric Person Authentication (AVBPA)*, pp. 188-193, USA, 1999.

[26] S. Ben-Yacoub, Y. Abdeljaoued, and E. Mayoraz, “Fusion of Face and Speech data for Person Identity Verification,” *IEEE Trans. Neural Network,* vol. 10, no. 6, pp. 1065-1075, Sep. 1999.

[27] V. Chatzis, A. G. Bors, and I. Pitas, “Multimodal Decision-level Fusion for Person Authentication,” *IEEE Trans. Systems, Man, and Cybernetics, Part A: Systems and Humans*, vol. 29, no. 6, pp. 674-681, 1999.

[28] Y. Ma, B. Cukic, and H. Singh, “A Classification Approach to Multi-biometric Score Fusion,” In: *Proc. 5th Int. Conf. Audio- and Video-based Biometric Person Authentication (AVBPA)*, pp. 484-493, Rye Brook, USA, 2005.

[29] A. K. Jain, and A. Ross, “Learning user-specific parameters in a multibiometric system,” In: *Proc. Int. Conf. Image Processing (ICIP),* pp. 57-60. 2002.

[30] K.-A. Toh, X. Jiang, and W.-Y. Yau, “Exploiting Global and Local Decision for Multimodal Biometrics Verification,” *IEEE Trans. Signal Processing,* vol. 52, no. 10, pp. 3059-3072, Oct. 2004.

[31] N. Poh, A. Ross, W. Lee, J. Kittler, "A User-specific and Selective Multimodal Biometric Fusion Strategy by Ranking Subjects," *Pattern Recognition*, vol. 46, no. 12, pp. 3341 - 3357, Dec. 2013.

[32] J. Fierrez-Aguilar, D. Garcia-Romero, J. Ortega-Garcia, and J. Gonzalez-Rodriguez, “Adapted User-dependent Mutlimodal Biometric Authentication Exploiting General Information,” *Pattern Recognition Letters,* vol. 26, no. 16, pp. 2628-2639, Dec. 2005.

[33] J. Fierrez-Aguilar, J. Ortega-Garcia, D. Garcia-Romero, and J. Gonzalez-Rodriguez, “Bayesian Adaptation for User-Dependent Multimodal Biometric Authentication,” *Pattern Recognition,* vol. 38, no. 8, pp. 1317-1319, Aug. 2005.

[34] H. He, and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263-1284, 2009.

[35] C. Seiffert, T. M. Khoshgoftaar, J. Van Hulse, and A. Napolitano, “RUSBoost: A hybrid approach to alleviating class imbalance,” *IEEE Trans. Systems, Man and Cybernetics, Part A: Systems and Humans*, vol. 40, no. 1, pp. 185-197, 2010.

[36] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, "A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches," *IEEE Trans. Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 42, no. 4, pp. 463-484, 2012.

[37] Y. Sun, M. S. Kamel, A. K. Wong, and Y. Wang, "Cost-sensitive boosting for classification of imbalanced data," *Pattern Recognition*, vol. 40, no. 12, pp. 3358-3378, 2007.

[38] B. X. Wang, and N. Japkowicz, "Boosting support vector machines for imbalanced data sets," *Knowledge and Information Systems*, vol. 25, no. 1, pp. 1-20, 2010.

[39] G. Wu, and E. Y. Chang, "Adaptive feature-space conformal transformation for imbalanced data learning," In: *Proc. the 20th Int. Conf. Machine Learning*, 2003.

[40] Y. Ding, and A. Ross, "A comparison of imputation methods for handling missing scores in biometric fusion," *Pattern Recognition*, vol. 45, no. 3, pp. 919-933, 2012.

[41] "Learning from Imbalanced Data Sets," In: *Proc. Am. Assoc. For Artificial Intelligence (AAAI) Workshop*, N. Japkowicz, ed., 2000, (Technical Report WS-00-05).

[42] "Workshop Learning from Imbalanced Data Sets II," In: *Proc. Int. Conf. Machine Learning*, N. V. Chawla, N. Japkowicz, and A. Kolcz, eds., 2003.

[43] N. V. Chawla, N. Japkowicz, and A. Kolcz, "Editorial: Special Issue on Learning from Imbalanced Data Sets," ACM SIGKDD Explorations Newsletter, vol. 6, no. 1, pp. 1-6, 2004.

[44] D. M. J. Tax, "One-class Classification: Concept-learning in the absence of counter-examples," Ph.D. dissertation, Dept. Intelligent Systems, Univ. TU Delft, 2001.

[45] J. Munoz-Marf, Lorenzo Bruzzone, and G. Camps Vails, "A support vector domain description approach to supervised classification of remote sensing images." *IEEE Trans. Geoscience and Remote Sensing*, vol. 45, no. 8, pp. 2683-2692, 2007.

[46] B. Raskutti, and K. Adam, "Extreme re-balancing for SVMs: a case study," *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 1, pp. 60-69, 2004.

[47] H. J. Lee, and S. Cho, “The Novelty Detection Approach for Different Degrees of Class Imbalance,” *Lecture Notes in Computer Science*, vol. 4233, pp. 21-30, 2006.

[48] K. Kennedy, B. Mac Namee, and S. J. Delany, “Learning without default: A study of one-class classification and the low-default portfolio problem,” *Artificial Intelligence and Cognitive Science,* pp. 174-187, Springer Berlin Heidelberg, 2009.

[49] N. Japkowicz, “Supervised versus unsupervised binary-learning by feedforward neural networks,” *Machine Learning*, vol. 42, no. 1-2, pp. 97-122, 2001.

[50] N. Japkowicz, C. Myers, and M. Gluck, “A Novelty Detection Approach to Classification,” In: *Proc. Joint Conf. Artificial Intelligence*, pp. 518-523, 1995.

[51] C. Bergamini, L.S. Oliveira, A.L. Koerich, and R. Sabourin, “Combining Different Biometric Traits with One-class Classification”, *Signal Processing,* vol. 89, no. 11, pp. 2117-2127, Nov. 2009.

[52] National Institute of Standards and Technology (NIST) multimodal database. Available at http://www.nist.gov/index.html.

[53] J. Ortega-Garcia, J. Fierrez-Aguilar, D. Simon, M. Faundez, J. Gonzalez, V. Espinosa, A. Satue, I. Hernaez, J. J. Igarza, C. Vivaracho, D. Escudero, and Q. I. Moro, "MCYT baseline corpus: A bimodal biometric database", *IEE Proceedings Vision, Image and Signal Processing*, vol. 150, no. 6, pp. 395-401, Dec. 2003.

[54] C. M. Bishop, *Pattern Recognition and Machine Learning,* Springer, 2006.

[55] M. Breunig, H.-P. Kriegel, R. Ng, and J. Sander, “LOF: Indentifying density-based local outliers,” In: *Proc. ACM SIGMOD 2000 Int. Conf. Management of Data*, pp. 93-104, USA, 2000.

[56] D. M. J. Tax, and R. P. Duin, “Support vector data description,” *Machine Learning*, vol. 54, no. 1, pp. 45-66, 2004.

[57] R. P. Duin, "The combining classifier: to train or not to train?" In: *Proc. Int. Conf. Pattern Recognition,* vol. 2, 2002.

[58] R. E. Schapire, and Y. Singer, "Improved boosting algorithms using confidence-rated predictions," *Machine learning*, 37(3), 297-336, 1999.

[59] L. I. Kuncheva, *Combining pattern classifiers: methods and algorithms*, John Wiley & Sons, 2004.

[60] K. Venkataramani, "Optimal classifier ensembles for improved Biometric Verification," Ph.D dissertation, Carnegie Mellon University, 2007.

[61] N. Poh, T. Bourlai, and J. Kittler, "A Multimodal Biometric Test Bed for Quality-dependent, Cost-sensitive and Client-specific Score-level Fusion Algorithms", *Pattern Recognition Journal,* vol. 43, no. 3, pp. 1094-1105, Mar. 2010.

[62] J. Ortega-Garcia, et al., "The Multiscenario Multienvironment BioSecure Multimodal Database (BMDB)," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 32, no. 6, pp. 1097-1111, June, 2010.

[63] N. Poh, and S. Bengio, "Database, protocols and tools for evaluating score-level fusion algorithms in biometric authentication," *Pattern Recognition*, vol. 39, no. 2, pp. 223-233, 2006.

[64] B. Zhu, C. He, and P. Liatsis, "A robust missing value imputation method for noisy data", *Springer Science + Business Media*, LLC 2010.

[65] Q. D. Tran, P. Kantartzis, and P. Liatsis, "Improving fusion with optimal weight selection in Face Recognition," *Integrated Computer-Aided Engineering*, vol. 19, no. 3, pp. 229-237, 2012.

[66] D. T. Pham, A. Ghanbarzadeh, E. Koc, S. Otri, S. Rahim, and M. Zaidi, "The bees algorithm-a novel tool for complex optimization problems," In: *Proc. 2nd Virtual Int. Conf. Intelligent Production Machines and Systems* (IPROMS 2006), pp. 454-459, 2006.

[67] S. Z. Li, and A. K. Jain, *Encyclopedia of Biometrics: I-Z,* vol. 2, Springer, A. K. (2009).

[68] ISO/IEC JTC1/SC37 IS19795-1: Biometric Performance Tasting and Reporting- Part 1: Principles and Framework, 2006.

[69] ISO/IEC JTC1/SC37 FDIS 19795-2: Biometric Performance Tasting and Reporting- Part 2: Testing methodologies for technology and scenario evaluation, 2006.

[70] A. Merati, "Multi-Modal Biometric Authentication with Cohort-Based Normalization," Ph.D. dissertation, Centre for Vision, Speech and Signal Processing, University of Surrey, Sep. 2011.

[71] N. Poh, A. Merati, and J. Kittler, "Making Better Biometric Decisions with Quality and Cohort Information: A Case Study in Fingerprint," In: *IEEE Conf. Biometrics: Theory, Applications and Systems (BTAS)*, 2009.

[72] T. Fawcett, "ROC Graphs: Notes and Practical Considerations for Data Mining Researchers," *Technical Report HPL-2003-4*, HP Labs, 2003.

[73] T. Fawcett, "An Introduction to ROC Analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861-874, 2006.

[74] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of decision task performance," In: *Proc. ESCA Eur. Conf. Speech Communication and Technology (EuroSpeech)*, pp. 1895-1898, 1997.

[75] K. K. Paliwal, "Spectral Subband Centroids Features for Speech Recognition," In: *Proc. Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2, pp. 617–620, Seattle, 1998.

[76] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition,* Oxford University Press, 1993.

[77] S. Ikbal, H. Misra, and H. Bourlard, "Phase Auto-Correlation (PAC) derived Robust Speech Features," In: *Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 133–136, Hong Kong, 2003.

[78] C. Sanderson, and K. Paliwal, "Fast Features for Face Authentication under Illumination Direction Changes," *Pattern Recognition Letters*, vol. 24, no. 14, pp. 2409–2419, 2003.

[79] N. Poh, and S. Bengio, "An Investigation of F-ratio Client-Dependent Normalisation on Biometric Authentication Tasks," *Research Report 04-46,* IDIAP, Martigny, Switzerland, 2004.

[80] Y. Ding, and A. Ross, "When data goes missing: methods for missing score imputation in biometric fusion," In: *Proc. SPIE Conf. Biometric Technology and Human Identification VII*, Orlando, USA, 2010.

[81] D. B. Rubin, "Inference and missing data," *Biometrika,* vol. 63, no. 3, pp. 581–592, 1976.

[82] R. Little, and D. Rubin, *Statistical analysis with missing data*, Wiley, New York, 2002.

[83] N. Poh, D. Windridge, V. Mottl, A. Tatarchuk, and A. Eliseyev, "Addressing Missing Values in Kernel-based Multimodal Biometric Fusion Using Neutral Point Substitution," *IEEE Trans. Information Forensics and Security*, vol. 5, no. 3, pp. 461–469, 2010.

[84] S. Dinerstein, J. Dinerstein, and D Ventura, "Robust Multi-Modal Biometric Fusion via Multiple SVMs," In: *IEEE Int. Conf. Systems, Man and Cybernetics*, pp. 1530–1535, Oct. 2007.

[85] K. Nandakumar, A.K. Jain, and A. Ross, "Fusion in multibiometric identification systems: what about the missing data?" In: *Proc. IEEE/IAPR Int. Conf. Biometrics*, Springer, pp. 743–752, 2009.

[86] G. E. Batista, and M. C. Monard, "An analysis of four missing data treatment methods for supervised learning," *Applied Artificial Intelligence*, vol. 17, no. 5-6, pp. 519-533, 2003.

[87] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm (with discussion)," *Journal of the Royal statistical Society*, vol. 39, no. 1, pp. 1-38, 1977.

[88] A. G. Ivakhnenko, "The group method of data handling-a rival of the method of stochastic approximation," *Soviet Automatic Control*, vol. 13, no. 3, pp. 43-55, 1968.

[89] Q. D. Tran, P. Liatsis, B. Zhu, and C. He, "Using Density based Score Fusion for Multimodal Identification Systems under the Missing Data Scenario," In: *Developments in E-systems Engineering (DeSE)*, pp. 238-242, Dec. 2001.

[90] K. Nandakumar, Y. Chen, S. C. Dass, and A. K. Jain, "Likelihood Ratio Based Biometric Score Fusion", *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 30, no. 2, Feb. 2008.

[91] A. K. Jain, K. Nandakumar, K., and A. Ross, "Score normalization in multimodal biometric systems," *Pattern Recognition*, vol. 38, no. 12, pp. 2270-2285, 2005.

[92] O. Fatukasi, J. Kittler, and N. Poh, "Estimation of missing values in multimodal biometric fusion," In: *Proc. IEEE Int. Conf. Biometrics: Theory, Applications and Systems (BTAS)*, 2008.

[93] L. Hong, and A. K. Jain, "Integrating Faces and Fingerprints for Personal Identification," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 20 no. 12, pp. 1295-1307, 1998.

[94] 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 Trans. Pattern Analysis and Machine Intelligence*, vol. 27, no. 3, pp. 450-455, 2005.

[95] J. Kittler, M. Hatef, R. P. Duin, and J. Matas, "On combining classifiers," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, pp. 226-239, 1998.

[96] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification.* John Wiley & Sons, 2001.

[97] E. L. Lehmann, and J. P. Romano, *Testing Statistical Hypotheses*, Springer, 2005.

[98] J. Q. Li, and A. Barron, "Mixture Density Estimation," *Advances in Neural Information Processings Systems 12*, S. A. Solla, T. K. Leen, and K.-R. Muller, eds., 1999.

[99] A. Rakhlin, D. Panchenko, and S. Mukherjee, "Risk Bounds for Mixture Density Estimation," *ESAIM: Probability and Statistics*, vol. 9, pp. 220-229, June 2005.

[100] S. C. Dass, K. Nandakumar, and A. K. Jain, "A Principled Approach to Score Level Fusion in Multimodal Biometric Systems," In: *Proc. Fifth Int.*

Conf. on Audio- and Video-based Biometric Person Authentication (AVBPA), pp. 1049-1058, Rye Brook, USA, 2005.

[101] U. Cherubini, E. Luciano, and W. Vecchiato, *Copula Methods in Finance,* Wiley, 2004.

[102] Y. Wang, T. Tan, and A.K. Jain, "Combining Face and Iris Biometrics for Identity Verification," In: *Proc. 4th Int. Conference on Audio- and Video-based Biometric Person Authentication (AVBPA)*, pp. 805-813, Guildford, UK, 2003

[103] D. T. Pham, and A. Ghanbarzadeh, "Multi-Objective Optimization using Bees Algorithm," In: *Proc. Innovative Production Machines and Systems Virtual Conference*, 2007.

[104] CASIA 3D face database, Center for Biometrics and Security Research, http://www.cbsr.ia.ac.cn/.

[105] ORL face database, AT&T Laboratories Cambridge, http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html.

[106] N. Poh, and S. Bengio, "A Study of the Effects of Score Normalisation Prior to Fusion in Biometric Authentication Tasks," *IDIAP Research Report 69*, IDIAP, 2004.

[107] A. Ross, and A. K. Jain, "Information Fusion in Biometrics," *Pattern Recognition Letters*, vol. 24, no. 13, pp. 2115-2125, 2003.

[108] G. Doddington, W. Liggett, A. Martin, M. Przybocki, and D. Reynolds, "Sheep, goats, lambs, and wolves: a statistical analysis of speaker performance in the NIST 1998 Speaker Recognition Evaluation," In: *Int. Conf. Spoken Language Processing (ICSLP)*, Sydney, 1998.

[109] N. Yager, and T. Dunstone, "The biometric menagerie," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 32, no. 2, pp. 220-230, 2010.

[110] J. L. Wayman, "Multi-Finger Penetration Rate and ROC Variability for Automatic Fingerprint Identification Systems," *Technical report*, National Biometric Test Center, 1999.

[111] M. Wittman, P. Davis, and P. Flynn, "Empirical Studies of the Existence of the Biometric Menagerie in the FRGC 2.0 Color Image Corpus," In: *Proc. Computer Vision and Pattern Recognition Workshop*, 2006.

[112] J. Fierrez-Aguilar, J. Ortega-Garcia, and J. Gonzalez-Rodriguez, "Target dependent score normalization techniques and their application to signature verification," *IEEE Trans. Systems, Man, and Cybernetics, Part C: Applications and Reviews,* vol. 35, no. 3, pp. 418-425, 2005.

[113] A. Kumar, and D. Zhang, "Integrating Palmprint with Face for User Authentication," In: *Workshop on Multimodal User Authentication (MMUA 2003)*, pp. 107–112, Santa Barbara, 2003.

[114] Q. D. Tran, and P. Liatsis, "A Modified Equal Error Rate Based User-specific Normalization for Multimodal Biometrics," In: *Proc. Fifth Int. Conf. Developments in e-Systems Engineering (DESE)*, Bucharest, Romania, Sep. 2012.

[115] Y. Sun, "Cost-Sensitive Boosting for Classification of Imbalanced Data," Ph.D. dissertation, Univ. Waterloo, Waterloo, Ontario, Canada, 2007.

[116] K. Carvajal, M. Chacon, D. Mery, and G. Acuna, "Neural network method for failure detection with skewed class distribution," *INSIGHT, Journal of the British Institute of Non-Destructive Testing*, vol. 46, no. 7, pp. 399-402, 2004.

[117] N. Japkowicz, and S. Stephen, "The class imbalance problem: A systematic study," *Intelligent Data Analysis Journal*, vol. 6, no. 5, pp. 429-450, 2002.

[118] M. Zhu, H. Chipman, and W. Su, "An adaptive method for statistical detection with applications to drug discovery," In: *2003 Proc. Am Stat Assoc-Biopharm*, pp. 4784-4789, 2003.

[119] G. E. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *SIGKDD Explorations Special Issue on Learning from Imbalanced Datasets*, vol. 6, no. 1, pp. 20-29, 2004.

[120] V. N. Vapnik, *Statistical Learning Theory*, Wiley-Interscience, New York, 1998.

[121] H. T Alashwal, S. Deris, and R. M. Othman, "One-class Support Vector Machines for Protein-Protein Interactions Prediction", *International Journal Biomedical Sciences,* vol. 1, no. 2, pp. 120-127, 2006.

[122] L.M. Manevitz, and M. Yousef, "One-class SVMs for Document Classification," *Journal of Machine Learning Research*, vol. 2, pp. 139-154, Jan. 2002.

[123] Q. Wang, L. S. Lopes, and D. M. J. Tax, "Visual Object Recognition through One-class Learning," In: *Proc. Int. Conf. Image Analysis and Recognition*, pp. 463-470, 2004.

[124] M. Breunig, H.-P. Kriegel, R. Ng, and J. Sander, "LOF: Indetifying density-based local outliers," In: *Proc. ACM SIGMOD 2000 Int. Conf. Management of Data*, pp. 93-104, USA, 2000.

[125] B. W. Silverman, *Density Estimation for Statistical and Data Analysis,* Chapman and Hall, 1986.

[126] B. Scholkopf, R. Willianmson, A. Smola, and J. Shawe-Taylor, "SV Estimation of a Distribution's Support" In: *Proc. Neural Information Processing Systems (NIPS)*, pp. 582-588, 1999.

[127] A. Tavakkoli, M. Nicolescu, M. Nicolescu, and G. Bebis, "Incremental svdd training: Improving efficiency of background modeling in videos," In: *Proc. the 10th IASTED International Conference*, vol. 623, 2008.

[128] N. Poh, A. Merati, and J. Kittler, "Adaptive client-impostor centric score normalization: A case study in fingerprint verification," In: *IEEE 3rd Int. Conf. Biometrics: Theory, Applications, and Systems*, pp. 1-7, 2009.

[129] L. Xu, A. Krzyzak, and C. Suen, "Methods of combining multiple classifiers and their applications to handwriting recognition," *IEEE Trans. Systems, Man and Cybernetics*, vol. 22, no. 3, pp. 418–435, 1992.

[130] J. Franke, and E. Mandler, "A Comparison of Two Approaches for Combining the Votes of Cooperating Classifiers," In: *Proc. 11th IAPR Int. Conf. Pattern Recognition, Conf. B: Pattern Recognition Methodology and Systems*, vol. 2, pp. 611-614, 1992.

[131] R. Jacobs, "Method for combining experts' probability assessments," *Neural Computation*, vol. 7, no. 5, pp. 867-888, 1995.

[132] G. Rogova, "Combining the results of several neural network classifiers," *Neural Networks*, vol. 7, no. 5, pp. 777–781, 1994.

[133] M. Tanigushi, and V. Tresp, "Averaging regularized estimators," *Neural Computation*, vol. 9, no. 5, pp. 1163–1178, 1997.

[134] J. Benediktsson, and P. Swain, "Consensus theoretic classification methods," *IEEE Trans. Systems, Man and Cybernetics*, vol. 22, no. 4, pp. 688–704, 1992.

[135] D. M. J. Tax, M. Breukelen, R. P. Duin, and J. Kittler, "Combining multiple classifiers by averaging or multiplying?," *Pattern Recognition*, vol. 33, no. 9, pp. 1475–1485, 2000.

[136] D. M. J. Tax, and R. P. Duin, "Combining one-class classifiers," *Multiple Classifier Systems*, pp. 299-308, Springer Berlin Heidelberg, 2001.

[137] R. Kohavi, and D. H. Wolpert, "Bias plus variance decomposition for zero–one loss functions," In: *Proc. 13th Int. Conf. Machine Learning*, pp. 275–283, Morgan Kaufmann, 1996.

[138] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[139] J. Daugman, "Combining Multiple Biometrics". Available at http://www.cl.cam.ac.uk/~jgd1000/combine/combine.html.

[140] L. Shapley, and B. Grofman, "Optimizing group judgemental accuracy in the presence of interdependencies," *Public Choice*, vol. 43, no. 3, pp. 329–343, 1984.

[141] O. Matan, "On voting ensembles of classifiers (extended abstract)," In: *Proc. AAAI-96 workshop on Integrating Multiple Learned Models*, pp. 84–88, 1996.

[142] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123-140, 1996.

[143] Y. Freund, R. E. Schapire, and N. Abe, "A short introduction to boosting," *Journal of Japanese Society For Artificial Intelligence*, vol. 14, no. 5, pp. 771-780, 1999.

[144] Y. Freund, and R. E. Schapire, "A decision–theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119–139, 1997.

[145] W. Fan, S. J. Stolfo, J. Zhang, and P. K. Chan, "AdaCost: misclassification cost-sensitive boosting," In: *Int. Conf. Machine Learning*, pp. 97-105, 1999.

[146] H. Masnadi-Shirazi, and N. Vasconcelos, "Cost-sensitive boosting," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 33, no. 2, pp. 294-309, 2011.

[147] A. K. Jain, R. P. Duin, and J. Mao, "Statistical pattern recognition: A review," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 22, no. 1, pp. 4-37, 2000.

[148] J. You, W. K. Kong, D. Zhang, and K. H. Cheung, "On hierarchical palmprint coding with multiple features for personal identification in large databases," *IEEE Trans. Circuits and Systems for Video Technology,* vol. 14, no. 2, pp. 234-243, 2004.

[149] A. K. Jain, Y. Chen, and M. Demirkus, "Pores and ridges: high-resolution fingerprint matching using level 3 features," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 29, no. 1, pp. 15-27, 2007.

[150] K. Tumer, and J. Ghosh, "Error correlation and error reduction in ensemble classifiers," *Connection Science*, vol. 8, no. 3-4, pp. 385-404, 1996.

[151] G. Fumera, and F. Roli, "Linear combiners for classifier fusion: some theoretical and experimental results," In: *T. Windeatt and F. Roli, editors, Proc. 4th Int. Workshop on Multiple Classifier Systems (MCS 2003)*, vol. 2709 of Lecture Notes in Computer Science LNCS, Guildford, UK, 2003, Springer-Verlag, pp. 74–83, 2003.

[152] A. W. Bowman, and A. Azzalini, *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations*, Oxford University Press, 1997.

[153] N. Poh, and J. Kittler, "A Methodology for Separating Sheep from Goats for Controlled Enrolment and Multimodal Fusion," In: *Proc. 6th Biometrics Symposium*, pp. 17–22, Tampa, 2008.

[154] N. Poh, A. Rattani, M. Tistarelli, and J. Kittler, "Group-specific Score Normalization for Biometric Systems", In: *IEEE Computer Society Workshop on Biometrics (CVPR)*, 2010.

[155] D. Maio, D. Maltoni, R. Cappelli, J. L. Wayman, and A. K. Jain, "FVC2000: fingerprint verification competition," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 24, no. 3, pp. 402–412, 2000.

[156] N. Poh, T. Bourlai, J. Kittler, L. Allano, F. Alonso-Fernandez, O. Ambekar, J. Baker, B. Dorizzi, O. Fatukasi, J. Fierrez, H. Ganster, J. Ortega-Garcia, D. Maurer, A. A. Salah, T. Scheidat, and C. Vielhauer, "Benchmarking Quality-dependent and Cost-sensitive Multimodal Biometric Fusion Algorithms," *IEEE Trans. on Information Forensics and Security*, vol. 4, no. 4, pp. 849–866, 2009.

[157] A. K. Jain, S. Pankanti, S. Prabhakar, L. Hong, and A. Ross, "Biometrics: a grand challenge," In: *Proc. 17th IEEE International Conference on Pattern Recognition (ICPR),* vol. 2, pp. 935-942, Aug. 2004.

[158] H. Kuck, "Bayesian formulations of multiple instance learning with applications to general object recognition," Master's thesis, University of British Columbia, Vancouver, BC, Canada, 2004.

[159] J. Campbell, C. Fraley, F. Murtagh, and A. Raftery, "Linear Flaw Detection in Woven Textiles Using Model-Based Clustering," *Pattern Recognition Letters*, vol. 18, no. 14, pp. 1539-1548, 1997.

[160] J. Rissanen, *Stochastic Complexity in Statistical Inquiry*, Singapore: World Scientific, 1989.

[161] S. Richardson, and P. Green, "On Bayesian Analysis of Mixtures with Unknown Number of Components," *Journal of the Royal Statistical Society: series B (statistical methodology)*, vol. 59, no. 4, pp. 731-792, 1997.

[162] M. A. Figueiredo, and A. K. Jain, "Unsupervised learning of finite mixture models," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 24, no. 3, pp. 381-396, 2002.

[163] D.M.J Tax, "DDtools, the Data Description Toolbox for Matlab", May 2012.

[164] S. Canu, Y. Grandvalet, and A. Rakotonam, "SVM and Kernel Methods Matlab Toolbox," Perception Systems et Information, INSA de Rouen, France, 2003.

[165] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel, *Robust statistics: the approach based on influence functions*, John Wiley & Sons, 2011.

[166] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, vol. 10, no. 1, pp. 42-54, 2000.

[167] D. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1, pp. 19-41, 2000.

[168] P. Smyth, "Model selection for probabilistic clustering using cross-validated likelihood," *Statistics and Computing*, vol. 10, no. 1, pp. 63-72, 2000.

[169] P. M. Long, and R. A. Servedio, "Random classification noise defeats all convex potential boosters," *Machine Learning*, vol. 78, no. 3, pp. 287-304, 2010.

[170] J. Friedman, T. Hastie, and R. Tibshirani, "Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors)," *The Annals of Statistics*, vol. 28, no. 2, pp. 337-407, 2000.

[171] D. Wang, D. S. Yeung, and E. C. Tsang, "Structured one-class classification," *IEEE. Trans. Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 36, no. 6, pp. 1283-1295, 2006.