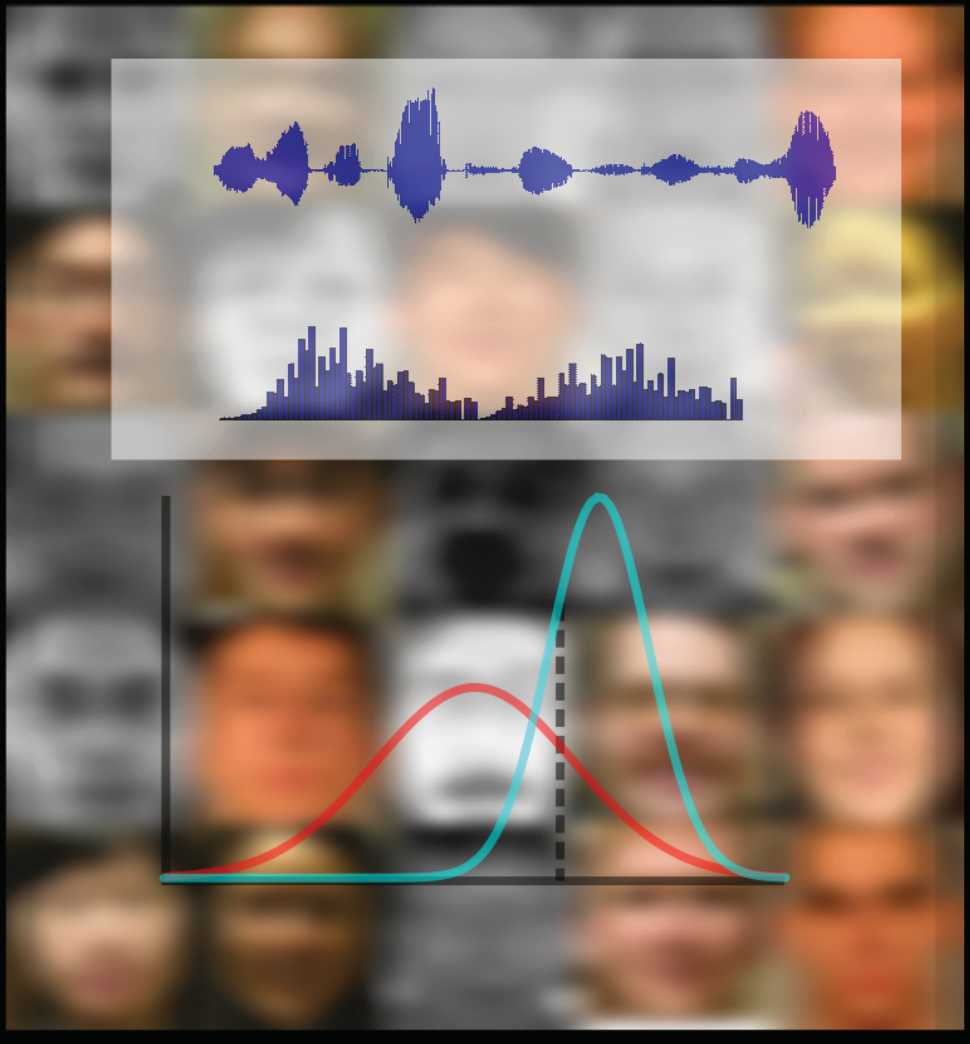
Biometric Score Calibration for Forensic Face Recognition



Tauseef Ali

BIOMETRIC SCORE CALIBRATION FOR FORENSIC FACE RECOGNITION

Tauseef Ali

Graduation committee:

Prof.dr.ir. R.N.J. Veldhuis, University of Twente, The Netherlands
Prof.dr. D. Meuwly, University of Twente and Netherlands Forensic Institute, The Netherlands
Dr. L.J. Spreeuwers, University of Twente, The Netherlands
Prof.dr. R.J. Wieringa, University of Twente, The Netherlands
Prof.dr.ir. C.H. Slump, University of Twente, The Netherlands
Prof.dr. M.J. Sjerps, University of Amsterdam and Netherlands Forensic Institute, The Netherlands
Dr. D. Ramos, Autonomous University of Madrid, Spain

The research is funded by the European commission as Marie-Curie ITN-project (FP7-PEOPLE-ITN-2008) “Bayesian Biometrics for Forensics (BBfor2)". A part of the research is carried out at Autonomous University of Madrid, Netherlands Forensic Institute and IDIAP Research Institute.

CTIT

CTIT Ph.D. Thesis Series No. 14-336, Centre for Telematics and Information Technology P.O. Box 217, 7500 AE, Enschede, The Netherlands.

ISSN: 1381-3617
ISBN: 978-90-365-3689-9
DOI: 10.3990/1.9789036536899

BIOMETRIC SCORE CALIBRATION FOR FORENSIC FACE RECOGNITION

DISSERTATION

to obtain
the degree of doctor at the University of Twente,
on the authority of the rector magnificus,
Prof.dr. H. Brinksma
on account of the decision of the graduation committee,
to be publicly defended
on Thursday 19 June, 2014 at 12.45

by

Tauseef Ali

born on 10 January, 1983
in Charsadda, Pakistan

This dissertation has been approved by:

Prof.dr.ir. R.N.J. Veldhuis and Prof.dr. D. Meuwly (Promoters)
and
Dr. L.J. Spreeuwers (Assistant promoter)

Acknowledgements

I would like to thank my PhD supervisors Prof. Raymond Veldhuis and Dr. Luuk Spreeuwers. Their kindness, dedication and attention to detail have been a great inspiration to me. I would particularly thank them for trusting in me and giving me the freedom to choose and follow any research direction that I wanted to persue while at the same time, they were always there whenever I was stuck and needed guidance and support.

My most special thanks to my parents, family and friends who were the motivation in tough times of my PhD journey.

In particular, I thank:

- Members of my graduation committee for reviewing my work.
- My colleagues, Abhishek dutta, Chris van Dam, Meiru Mu, Jen-Hsuan, C.G. Zeinstra, Chanjuan Liu and Y. Peng.
- A.G.H. Westhoff, G.J. Laanstra, S.E. Engbers and B.F.J. Scholten-Koop.
- Prof.dr. David van Leeuwen for always providing very useful and positive feedbacks during BBfor2 presentations and by emails. I also thank him for providing the speaker recognition scores data.
- Dr. Julian Fierrez and Dr. Daniel Ramos for their supervision and guidance during my research stay at Biometric Recognition Group (ATVS), Autonomous University of Madrid, Spain. I am also thankful to Pedro Tom and all other members of the group for their warm welcome.

- Sebastian Marcel and Manuel Gunther for their supervision and guidance during my research stay at Biometrics group, IDIAP research institute, Switzerland.

- Prof.dr. Didier Meuwly for his supervision and guidance during my research stay at Netherlands Forensic Institute, The Netherlands.

- All members of the BBfor2 project. Besides study and research, we had a lot of fun time during BBfor2 meetings and workshops.

Contents

Summary

When two biometric specimens are compared using an automatic biometric recognition system, a similarity metric called "score" can be computed. In forensics, one of the biometric specimens is from an unknwon source, for example, from a CCTV footage or a fingermark found at a crime scene and the other biometric specimen is obtained from a known source, for example, from a suspect. Automatic biometric recognition systems are gradually replacing the forensic examiners' manual comparison of the two biometric specimens. In forensics, there is a huge interest to use a suitable measure to report the output of the comparison of the two biometric specimens. This has led to the use of the likelihood-ratio, $\frac{P(s|H_p)}{P(s|H_d)}$, where s is the score computed by an automatic biometric recognition system, H_p is the hypothesis of the prosecution (which states that the two biometric specimens are obtained from a same-source) and H_d is the hypothesis of the defense (which states that the two biometric specimens are obtained from different sources). Generally, two sets of training scores, one under H_p and the other under H_d, are needed to compute a likelihood-ratio from a score. In this thesis, we review several methods of likelihood-ratio computation focusing mainly on the issues of the sampling variability in the sets of training scores and the specific conditioning imposed on the pairs of the biometric specimens to compute them. Three different methods are considered in detail: Kernel density estimation, Logistic regression and Pool adjacent Violators.

The effect of the sampling variability is quantified varying : 1) the shapes of the probability density functions which model the distributions of the scores under H_p and under H_d; 2) the sizes of the training sets under H_p and under H_d; 3) the actual value of the score for which the likelihood-ratio is computed. The study proposes a simulation framework which can be used to study several properties of a likelihood-ratio computation method and to quantify the effect of the sampling variability in a likelihood-ratio. This is useful for an appropriate and informed choice of a likelihood-ratio computation method. It is shown that sampling variability is a serious concern when small sets of the training scores are available for likelihood-ratio computation.

Our study of likelihood-ratio computation also focuses on the specific conditioning imposed on the pairs of biometric specimens used for computation of the sets of the training scores. In general, the two sets of training scores are

obtained from a same-source and different-sources comparisons of biometric specimens. However, the same-source and different-sources conditions can be anchored to a specific suspect in a forensic case or it can be generic same-source and different-sources comparisons independent of the suspect involved in the case. This results in two likelihood-ratios which differ in the nature of the training scores they use and therefore consider slightly different interpretations of the two hypotheses. An empirical study is carried out to quantify how much and how frequently the two likelihood-ratios vary considering a speaker, a face and a fingerprint recognition system. Study showed that there is significant variations in the two likelihood-ratios and therefore explicit definition of the training sets and the hypotheses implied by them is very important.

The state-of-the-art towards automated forensic face recognition is reviewed and the concept of likelihood-ratio is applied to several existing biometric face recognition systems. In forensic situations, e.g., when an image from a crime scene is compared with an image from a suspect, forensic face recognition is currently a manual process referred to as "forensic facial comparison" and performed by forensic examiners based on their experience and a limited set of guidelines. A step is taken towards automation of forensic face recognition by studying the discriminating powers of different facial features such as eyes, eye brows, nose, etc. This kind of regional comparison is the essence of forensic facial comparison and prove very useful in situations where a part of the face is available for comparison. Besides the automation, it might also be feasible to use existing automatic face recognition systems for forensic comparison and reporting. To this end, several face recognition systems are calibrated so that they produce likelihood-ratios and their performance is evaluated based on the likelihood-ratios assessment tools.

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Preliminaries

1.1.1 Biometric score

A biometric specimen refers to the acquired biometric data such as a face image, speech segment and fingerprint, which is used in automatic biometric recognition system [1]. Using automatic biometric recognition systems, a pair of biometric specimens can be compared in order to find out whether the two biometric specimens are obtained from same source or different sources. The result of comparison from these systems can generally be represented by a similarity metric called "score". In general, a score quantifies the similarity between the two biometric specimens while taking into account their typicality.

1.1.2 Likelihood-ratio and biometric score calibration

In applications of biometric recognition systems such as access-control to a building and e-passport gates at some airports require the developer of the system to choose a threshold and consequently any score above the chosen threshold implies a positive decision and vice versa. This approach of using a biometric recognition system works well for such applications, however, it has limitations for forensic evidence evaluation [2–4]. When the pair of biometric specimens consists of a biometric specimen from a suspect and a biometric

specimen from a crime scene, the score is not a very useful metric for presentation in court as a result of the comparison. Also a threshold-based decision making is not suitable in forensic casework since there are usually other sources of information about the case at hand which should also be taken into consideration. Furthermore, it has been argued that making a decision is not the province of the forensic practioner [5–7].

The concept of Likelihood-Ratio (LR) can be used to present the result of a biometric comparison in forensic evidence evaluation which is a more informative, useful and objective output than a score. It has been extensively used for DNA evidence [5, 8]. In general, given two biometric specimens, x and y, a LR is defined as follows:

$$LR(x, y) = \frac{P\left(x, y|H_p, I\right)}{P\left(x, y|H_d, I\right)}, \tag{1.1}$$

where H_p is the hypothesis of the prosecution (which states that the two biometric specimens are obtained from a same source) and H_d is the hypothesis of the defense (which states that the two biometric specimens are obtained from different sources). I refers to the background information about the case at hand.

For score-based biometric systems, the score computed by comparing x and y replaces the joint probability of x and y in order to compute a LR [9, 10]. A LR is then the probability of the score given H_p divided by the probability of the score given H_d:

$$LR(s) = \frac{P\left(s|H_p, I\right)}{P\left(s|H_d, I\right)}, \tag{1.2}$$

where s is the score computed by comparing x and y using an automatic biometric recognition system. The process of computation of a LR from a biometric score is referred to as "score calibration" or simply "calibration".

1.1.3 Relation to other fields of study

The LR is a ratio of the two conditional probabilities, $P(s|H_p)$ and $P(s|H_d)$. The background information I is ommited for simplicity. Using two sets of training scores, one under H_p and the other under H_d, these probabilities are computed by estimating the conditional probability densities or from the

posterior probabilities, $P(H_p|s)$ and $P(H_d|s)$, using the Bayes' theorem. Computation of the posterior probabilities is of interest in several other fields of study such as machine learning in general [11]. Specifically, computation of posterior probabilities are carried out in weather forecasting, prediction of the accuracy of a test in medical diagnostics and financial decision-making. In machine learning and data mining, posterior probabilities are more commonly referred to as "class-membership probabilities". In general, pattern classification techniques can be divided into two categories:

- Crispy classification: Given two input feature vectors, a classifier returns the predicted class-label. In biometric recognition, given x and y, this kind of classification will return the output: "x and y are obtained from a same source" (H_p is true) or "x and y are obtained from different sources" (H_d is true). This essentially involves a decision-making process by the classifier.
- Probabilistic classification: Given two input feature vectors, a classifier returns the conditional probability of each class. In biometric recognition, given x and y, the classifier will return probabilities: $P(H_p|x, y)$ and $P(H_d|x, y)$.

Computation of a LR from a score for biometric evidence evaluation is essentially an application of the probabilistic classification. However, there are several issues which are of more serious concern in forensic science and are the focus of this thesis.

1.1.4 Computation of a LR

Generally, the conditional probabilities, $P(s|H_p)$ and $P(s|H_d)$, are unknown in the LR and they are computed empirically using a set of training scores under H_p, $s_p = \{s_j^p\}_{j=1}^{n^p}$ (a set of n^p number of scores given H_p) and a set of training scores under H_d, $s_d = \{s_j^d\}_{j=1}^{n^d}$ (a set of n^d number of scores given H_d) (see Fig.1.1).

1.1.4.1 Sampling variability

Statistically, the training biometric data sets used to compute the s_p and the s_d sets are samples from large populations of biometric data sets. The training biometric data sets, when resampled, will lead to slightly different values of the training scores sets due to the unavoidable sampling variability. This implies that the sets s_p and s_d consist of random draws from large sets of scores.

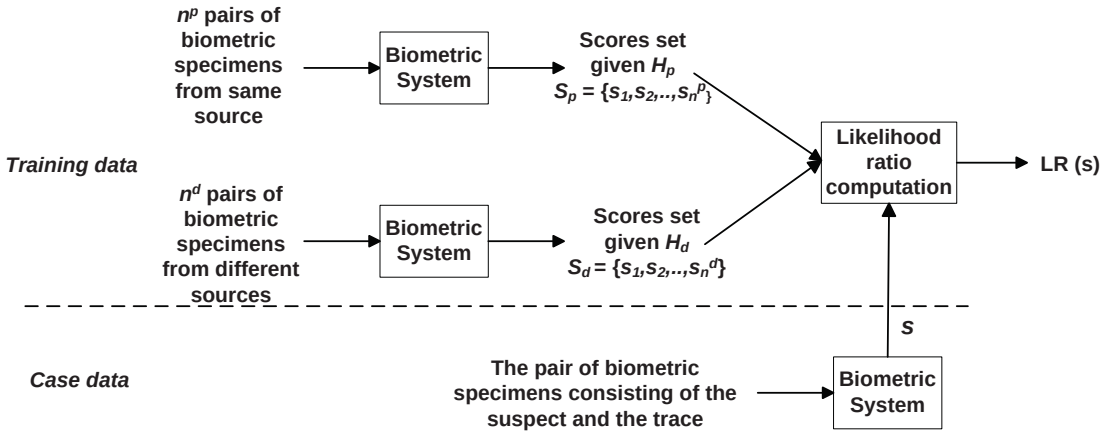


Fig. 1.1: Computation of a LR for a pair of biometric specimens consisting of the suspect's biometric specimen and the trace biometric specimen.

When the resampling is repeated, a slightly different LR is computed for a given score. This is referred to as the "sampling variability" in the computed LR. It is desirable that a given LR computation method is less sensitive to the sampling variability in the training sets. A method which is very sensitive to the sampling variability in the training sets is undesirable as the computed LR is prone to change significantly if it is computed again using another sample of the training data sets.

It should be emphasized that the sampling variability is caused by the training scores sets that is needed to compute the mapping function from scores to LRs. This is because the training biometric data sets (and therefore the training scores) are finite and would vary one to the next in repeated random sampling. In situation where the posterior probabilities or the conditional probabilities in the LR are known in advance and no training scores are needed to compute the mapping function from scores to LRs, there will be no sampling variability in the computed LR.

Sampling variability depends on three factors:

- The sizes of the training scores sets. In general, the sampling variability is expected to be large in LR computation if small training scores sets are used for computation of the mapping functions from scores to LRs.
- The shapes of the distributions of the scores in the two training scores sets.
- The actual value of the score for which the LR is computed. The LR of a score lying at the extremes of the range of the score values is expected to have more sampling variability than a score lying in the middle region of the range of the score values.

1.1.4.2 Training data sets

The biometric data sets used to compute the s_p and the s_d sets depend on the case at hand [4]. In general, the s_d scores are computed by comparing pairs of biometric specimens where the two biometric specimens in each pair are obtained from different sources whereas the s_p scores are computed by comparing pairs of biometric specimens where the two biometric specimens in each pair are obtained from a same source. An important condition in forensic LR computation is that the pairs of biometric specimens used for training should reflect the conditions of the pair of biometric specimens for which the LR is computed. Variability exists in the selection of the different-sources and same-source pairs of biometric specimens for computation of the training scores. For the s_p score, a set of biometric specimens from the suspect can be compared with another set of biometric specimens from the suspect [10,12–14]. An alternate approach is to compare pairs of biometric specimens where each pair is obtained from a same-source [6, 15]. This eliminates the need of the suspect's biometric specimens in computation of the training scores in the s_p set. Similarly, for the s_d scores, a set of pairs of biometric specimens are compared where one biometric specimen in each pair is obtained from the suspect and the other from a person in the relevent potential population. This is suspect-specific approach. The second alternative approach to compute the s_d scores is to compare a set of pairs of biometric specimens where each pair has one biometric specimen from a person in the relevent potential population and the other biometric specimen is the trace. The third alternative approach is to use pairs where the two the biometric specimens are from two different persons in the relevent potential population. [3, 6, 10, 15]. Please refer to [4] for an overview of the biometric data sets collection in forensic casework for a LR computation. The general approaches to compute the s_p and the s_d sets where the only condition on the training pairs of biometric specimens is that they should be obtained from same-source and different-sources respectively ensures a large number of training scores for LR computation. Besides the difference in the sizes of the training sets, it can also be expected that these different approaches result in different values of LR for a given score because of the different nature of the training data sets used for training.

1.1.4.3 Assessment of LRs

Performance assessment techniques such as Area under Receiver Operating Characteristics (AUC) curve and Equal Error Rate (EER) which are tradi-

tionally used for biometric recognition systems producing scores are questioned or even coined as flawed for application to biometric systems producing LRs [16, 17]. The underlying argument is based on the fact that a LR, in contrast to a score, is not used for binary decision-making based on a selected threshold but rather it gives a degree of support for one hypothesis or the other. Therefore, slightly modified techniques such as *Cost of Log LR* (C_{llr}) [17], Tippett plot [18] and Empirical Cross-Entropy (ECE) [16] plot are proposed for assessment of the performance of LR computation systems.

1.2 Forensic face recognition and the likelihood-ratio framework

Forensic face recognition is still mostly a manual process. Forensic examiners compare a face image from an unknown source (e.g., a face image from a CCTV footage of a crime scene) and a face image from a known source (e.g., a face image of a suspect under custody). This comparison is based on experience of the examiner and to a large extent, it is a subjective process [19, 20]. The comparison focus on specific things such as 1) relative distances between different relevant facial features 2) contours of cheek- and chin-lines 3) shape of mouth, eyes, nose, ears, etc. 4) lines, moles, wrinkles, scars, etc. The goal of the comparison is to reach a final conclusion considering details from different individual facial features. The final result is in the form of a LR by combining the conclusions based on different parts. The subjective nature of this process as well as the large amount of human effort needed when a lot of comparisons need to be performed require that the process should be (semi-)automated. Recently there has been an interest in standarization and automation of the forensic examiners way of facial comparison [21, 22].

The requirement that an automatic biometric recognition system used in evidence evaluation should produce a LR instead of a score is also of a concern in forensic face recognition. However, this issue can be addressed by appending a post-processing module with exisitng face recognition systems developed for other applications such as access-control [23, 24]. This only partly address the overall goal of an automatic forensic face recognition system. A desirable system for foreneic examiners is the one which can assist them in their manual process of comparison by, for example, selecting top 10 candidate faces from a large database of facial images with a more descriptive outputs such as how similar a given facial feature is compared to others. After a manual intervention, the system should be able to compute a LR based on a statistical

model. Achieving this goal requires efforts both towards automation of the current practice of forensic facial comparison as well as suitable methods for computation of LRs from the scores [20, 25, 26].

1.3 Research questions

The thesis aims at answering the following specific research questions:

- In computation of a LR, what is the effect of the sampling variability in the training scores sets? How the commonly proposed LR computation methods are affected by the sampling variability varying the sizes of the training scores sets, the shapes of the distributions of the scores in the training scores sets and the actual value of the score for which the LR is computed?
- What is the effect of using the suspect-independent training scores instead of the suspect-specific training scores in computation of a LR? Generally, a larger number of training scores are available if the suspect-independent training sets are used. However, besides the difference in the sizes of the sets of training scores, the nature of these two different ways (suspect-specific and suspect-independent) to compute the training scores sets implies slightly different interpretations of the prosecution hypothesis (H_p) and the defense hypothesis (H_d). It will be investigated that how much and how frequently the two LRs differ. Furthermore, it will also be investigated that, given the two approaches have the same number of scores in the training sets, are there still variations in the two LRs?
- What is the current practice of forensic examiners to perform forensic facial comparison and the current state-of-the-art towards automatic forensic face recognition? Furthermore what is the effective way in which the goal of a (semi-)automatic forensic face recognition can be achieved? What is the discriminating power of different facial features such as eyes, eyebrows, nose, etc?
- What is the performance of commonly proposed LR computation methods for calibration of existing automatic biometric face recognition systems? Is the conclusion drawn from the assessment tools at the score level such as ROC and EER is significantly different than the conclusion drawn from the assessment tools at the LR level such as C_{llr}?

1.4 Contributions

The work carried out has several contributions to the field of statistical biometric evidence evaluation in the form of a LR and forensic face recognition:

- The issue of sampling variability in computation of a LR from biometric scores is addressed in detail. Factors affecting the sampling variability are varied and detail analysis of commonly proposed LR computation methods is provided. These factors are the shapes of the distributions of the scores in the training scores sets, the sizes of the training scores sets and the actual value of the score for which the LR is computed. This analysis is useful for forensic practioners to make an informed and appropriate choice of a LR computation method when a pair of biometric specimens are compared using automatic biometric recognition system.
- The effect of using the suspect-independent (generic) biometric data sets instead of the suspect-specific to learn the mapping function from scores to LRs is investigated. The study is carried out for three different biometric modalities: face, fingerprint and speaker recognition. A state-of-the-art biometric recognition system is used from each biometric modality with same protocol for experiments in order to study the effect that the two different kinds of training biometric data sets have on the resultant LRs and a comparison across the three different biometric modalities is carried out.
- A literature survey on forensic face recognition is carried out which also describes the current practice and guidelines of forensic examiners to perform forensic facial comparison. Inspired by the forensic examiners' way of facial comparison, a study of two state-of-the-art face recognition algorithms is carried out for recognition of individual facial features (such as nose, eye, eye brows). The goal was to investigate how discriminating each facial feature is and rank them based on using the two algorithms for recognition.
- Commonly proposed LR computation methods are applied to several face recognition algorithms and performance is evaluated. The performance of these systems is evaluated before and after applying the score calibration process in order to understand that whether the LR computation stage introduces significant variation or not. Three different public databases are used in the experiments.

1.5 Overview of the thesis

The thesis contains, for the most part, published or submitted papers. Each chapter is preceded by an introduction part which, in case of two papers in the chapter, relates the two papers and links it to the rest of the thesis. This introduction part highlights the main points of the chapter and its overall contribution. The introduction section is followed by the paper(s) which are either submitted for review or published. Each paper is inserted in a separate section. In the papers included, besides small corrections such as typos, no modification is made in the contents.

Chapter 2 reviews the current practice of forensic facial comparison, state-of-the-art towards automated forensic face recognition and reviews the LR framework for evidence evaluation in the context of face recognition.

Chapter 3 reviews different methods of LR computation and the effect of the sampling variability in LRs. In section 3.2, the effect of different locations of the score on the sampling variability of a LR is explored in detail. Only one pair of the probability density functions (PDFs) is considered for generation of the two sets of training scores. In section 3.3, four different pairs of PDFs are considered from which the training scores sets are generated for LR computation. Also the effect of different sizes of the training scores sets is explored. The overall goal of this chapter is to study the effect of the sampling variability in the training scores sets varying the three parameter: sizes of the training sets, shapes of the distributions of the scores in the training sets and the location of the score for which the LR is computed.

Chapter 4 studies the effect of using the suspect-independent (generic) training biometric data sets instead of the suspect-specific (subject-specific) training biometric data sets. Three different biometric modalities (face, speaker and fingerprint) are considered. The slight difference in the two hypotheses that the two different kinds of training data set implies and a quantitative summary of how frequently the two LRs fall in a same range is provided. The effect of the different nature of the training scores sets alone is also studied by using the same sizes of the training scores sets in both the suspect-specific and suspect-independent approaches.

Chapter 5 presents a study of the discriminating power of different facial features using two automatic face recognition algorithms. This chapter also presents study of LR computation from scores computed by different face recognition systems. In section 5.2, a study of the recognition performance of different facial features is presented. In section 5.3, two state-of-the-art

face recognition systems are considered. Assessment of the LRs are performed using Tippett plot. Section 5.4 presents extensive results by considering 10 baseline face recognition systems and three commonly proposed LR computation methods. Assessment is carried out at both the score level and at LR level by using AUC and C_{llr} respectively.

Chapter 6 concludes the work presented in the thesis and gives recommendations for future research. In particular, it is discussed how the research questions posed in this chapter are answered by the work presented in the thesis and points to future research work that can be carried out to further address these and related research questions.

Chapter 2

Forensic face recognition and the LR framework

2.1 Introduction

In this chapter we explore state-of-the-art in forensic face recognition and the current practice of forensic facial comparison. The concept of LR and the use of Bayesian framework is also introduced in the context of evidence evaluation using an automatic face recognition system. The current practice of forensic facial comparison and the existing work towards automated forensic face recognition is described. Issues in the use of the Bayesian framework and court admissibility criteria for scientific evidence are also reviewed.

2.2 Forensic face recognition: A survey [1]

2.2.1 Abstract

The improvements of automatic face recognition during the last 2 decades have disclosed new applications like border control and camera surveillance. A new application field is forensic face recognition. Traditionally, face recognition by

[1] The content of this section are published in [27] "Forensic Face Recognition: A Survey", Book chapter in Face Recognition: Methods, Applications and Technology, Computer Science, Technology and Applications, Nova Publishers, ISBN 978-1-61942-663-4

human experts has been used in forensics, but now there is a quickly developing interest in automatic face recognition as well. At the same time there is a trend towards a more objective and quantitative approach for traditional manual face comparison by human experts. Unlike in most applications of face recognition, in the forensic domain a binary decision or a score does not suffice as a result to be used in court. Rather, in the forensic domain, the outcome of the recognition process should be in the form of evidence or support for or likelihood of a prosecution hypothesis verses a defence hypothesis. In addition, in the forensic domain, trace images are often of poor quality. The available literature on (automatic) forensic face recognition is still very limited. In this survey, an overview is given of the characteristics of forensic face recognition and the main publications. The survey introduces forensic face recognition and reports on attempts to use automatic face recognition in the forensic context. Forensic facial comparison by human experts and the development of guidelines and a more quantitative and objective approach are also addressed. Probably the most important topic of the survey is the development of a framework to use automatic face recognition in the forensic setting. The Bayesian framework is a logical choice and likelihood ratios can in principle be used directly in court. In the statistical evaluation of the trace image, the choice of databases of facial images plays a very important role.

2.2.2 Introduction

Face recognition is one of the most important tasks of forensic examiners during their investigations if there is video or image material available from a crime scene. Forensic examiners perform manual examination of facial images or videos to match a trace with an image of a suspects face or with a large database of mug-shots. The use of automated facial recognition systems will not only improve the efficiency of forensic work performed by various law enforcement agencies but will also standardise the comparison process. However, until now, there is no automatic face recognition system that has been accepted by the judicial system. A face recognition system must be thoroughly evaluated and verified before it can be utilised for forensic applications. Biometric face recognition has of course been used for secure building access, border control, Civil ID and login verification. However, to date no automatic system exists for identification or verification in crime investigation tasks, such as the comparison of images taken by CCTV with available databases of mug-shots. State-of-the-art face recognition systems such as [28, 29] could in principle be used for this purpose, but there are several issues, specific to the forensic

domain, which have to be addressed.

First and foremost, the consequences of a wrong decision made by forensic face recognition are far more severe than for most other biometric face recognition applications. Current face recognition solutions [30] are generally not sufficiently robust [31] to the variability in appearance of faces due to variations in pose, lighting conditions, facial expression and caused by imaging systems such as image quality, resolution and compression.

Secondly, a score or binary decision based biometric recognition system is not suitable to the judicial system where the objective is to give a probability or degree of support for one hypothesis against another incorporating the prior knowledge about the case at hand [32, 33].

Finally it should be mentioned that in the forensic scenario the quality of images available is generally low, e.g. images of a crime scene recorded using CCTV. These images usually have a low resolution and depicted faces are often not frontal and may be partly occluded.

On the other hand, the recognition task in the forensic framework can be carried out "offline" in contrast to other applications where a decision has to be made in real-time, e.g. user access for a building or border control. Forensic face recognition therefore has fewer time constraints and to a certain extent human involvement is allowed and generally does not effect the overall objectivity of the system.

A related field of forensic facial recognition is forensic facial reconstruction which aims to reproduce a lost or unknown face of an individual for the purpose of identification or verification [34]. Well known is the approach to reconstruct a face starting from the skull and using pins to model the thickness of the muscle tissue, then filling in the muscle tissue using clay and thus reconstruct the facial surface [35]

In this survey, we review existing literature on forensic face recognition. There are relatively few papers focusing on the forensic application of face recognition as most effort is put into the improvement of the technology itself. However, as the performance of face recognition systems improves the demand for application in the forensic domain also increases and, hence, there is a great need for integration of the technology with the legal system and a uniform framework for application of face recognition technology in forensics.

The remainder of the chapter is organised as follows: in section 2.2.3, the techniques and methodologies used by forensic examiners for the purpose of facial comparison are discussed. Section 2.2.4 presents a literature review of forensic

face recognition. In section 2.2.5 we discuss the Bayesian framework and how it can be applied to forensic face recognition. Section 2.2.6 discusses reliability and court admissibility issues associated with forensic facial recognition. Section 2.2.7 presents conclusions.

2.2.3 Forensic facial identification

Facial identification refers to manual examination of two face images or a live subject and a facial image to determine whether they are of the same person or not. Facial identification methods generally can be classified into the following four categories:

1. *Holistic Comparison*: In this approach faces are compared by considering the whole face at once.
2. *Morphological Analysis*: In this approach individual features of the face are compared and classified.
3. *Photo-anthropometry*: This approach (sometimes referred to as photogrammetry) is based on the spatial measurements of facial features as well as distances and angles between facial landmarks.
4. *Superimposition*: In this approach, a properly scaled version of one image is overlaid onto another. The two images must be taken from the same angle.

The choice of a specific approach is usually dependent on the face images to be compared and generally combinations of these methods are applied to reach a conclusion. Apart from the above described general categorisation of facial comparison approaches, currently there are no standard procedures and agreed upon guidelines among forensic researchers. Due to the lack of an agreed-upon protocol, the similarities and differences are based on personal probabilities and therefore the opinion of one forensic examiner may vary from those of others.

2.2.3.1 Working groups

There are several working groups active in this area the aim of which is to standardise the procedure of forensic facial comparison as well as the proper training of facial comparison experts. One of the best efforts towards developing standards and guidelines for forensic facial identification is currently carried out by the Facial Identification Scientific Working Group (FISWG) [21]. It works under the Federal Bureau of Investigation (FBI) Biometric Center of

Excellence (BCOE). FISWG is focusing exclusively on facial identification and developing consensus, standards, guidelines, and best practices for facial comparison. Currently they have developed drafts of several useful documents in this regard which include a description of facial comparison, a facial identification practitioner code of ethics and guidelines for training experts to perform facial comparison. These documents are available for public review and comments [21]. Some other working groups active in developing standards and guidelines for forensic facial comparison include the International Association for Identification [22] and the European Network of Forensic Science Institutes (ENFSI) [36]. The standardisation of the process of facial comparison and specific guidelines which are agreed upon by forensic community is, however, still a largely unsolved problem.

2.2.3.2 Manual facial comparison by the forensic expert

In this section we briefly review the forensic experts' way of facial comparison. The discussion is based on the guidelines set forward by the workgroup on face comparison at the Netherlands Forensic Institute (NFI) [20, 37] which is a member of ENSFI [28]. The facial comparison is based on morphological-anthropological features. If possible, for comparison, images with faces depicted at the same size and with the same pose are used. The comparison mainly focuses on:

- Relative distances between different relevant features
- Contours of cheek- and chin-lines
- Shape of mouth, eyes, nose, ears, etc.
- Lines, moles, wrinkles, scars, etc. in the face

When comparing facial images manually, it should be noted that differences may be invisible due to underexposure, overexposure, low resolution, out-of-focus and distortions in the imaging process. On the other hand, due to similar limitations in the image formation process (low resolution, difference in focus and positions of the cameras used to record the images relative to the head and other distortions in the imaging process) may lead to different appearance of similar features in the facial images to compare. Due to the aforementioned effects, which complicate the comparison process, the anthropological facial features are visually compared and classified as: *similar in details*, *similar*, *no observation*, *different* and *different in details*. Apparent similarities and differences are further evaluated by classifying features as: *weakly discriminating*, *moderately discriminating*, and *strongly discriminating*. The conclusion based

on the comparison process is in the form of a measure of support for either of the hypotheses (images show faces of the same person vs. images show faces of different persons) and can be stated as: *no support*, *limited support*, *moderate support*, *strong support* and *very strong support* . The process is subjective and often different experts reach different conclusions. There is a great need to standardise the process. Use of automatic face recognition systems will considerably improve the speed and objectiveness of facial comparison and may also be helpful in standardising the comparison process.

2.2.4 Literature overview

In this section we briefly review existing literature on forensic face recognition. This review focuses on work discussing forensic aspects rather than on work describing techniques for biometric face recognition. Surveys on the latter subject can be found in [30, 38].

2.2.4.1 Forensic biometrics from images and videos at the FBI

Forensic Biometrics from Images and Videos at the Federal Bureau of Investigation (FBI) is described in [39]. The paper gives a description of FBI's Forensic Audio, Video and Image Analysis Unit (FAVIAU) and the forensic recognition activities that they perform. Many of these activities are performed manually. Types of manual tasks include voice comparison, facial comparison, height determination, and other side by side image comparisons. Two types of examinations that involve biometrics are photographic comparisons and photogrammetry [40]. Currently, in both cases, the forensic examinations are performed manually. Photographic comparison means a one-to-one comparison of a trace facial image to facial images from suspects. The characteristics used in photographic comparison can be categorised into class and individual characteristics [41]. Class characteristics such as hair colour, overall facial shape, presence of facial hair, shape of the nose, presence of freckles, etc. place an individual within a class or group. Individual characteristics such as the number of and locations of freckles and scars, tattoos, the number of and positions of wrinkles etc. are unique to an individual and can be used to individualise a person. Photogrammetry [40] determines spatial measurements of objects using photographic images. It is used to determine e.g. the height of a subject or the length of a weapon used in a crime. In [39] several current and past research projects in the field of forensic recognition are discussed and also directions for future research on forensic recognition are proposed.

2.2.4.2 Facial comparison by experts

In [42] the need for facial comparison experts, their role in biometric face recognition development and their training are described. The paper describes the need for facial comparison experts to verify the results of future automatic forensic face recognition systems. It emphasises the systematic training of experts who will be working with these systems. For any future application of an automated face recognition system, the ultimate judgment will be the manual verification of the outcome of the system. Because the implications of an incorrect decision are severe the verification of the outcome of an automated system by an expert is very important. In case of fingerprint technology, there are many experts available working in association with the automated process. Compared to fingerprint technology, forensic application of face recognition is still immature and, therefore, requires even more this manual verification of the results by experts. This means in the near future more experts will have to be trained in order to use automatic face recognition systems. Comparison of images taken under controlled conditions such as passport photos or photos for arrest records requires less expertise compared to images taken under uncontrolled conditions such as snapshots and images from surveillance cameras. The experts also need training in legal issues because they will be working in the judicial system and will present their conclusions in court. The facial image examiners should be trained in three main areas:

1. General background on facial recognition approaches, which includes the history of person identification, current methods in biometrics, underlying principles of photographic comparison [41] and basic knowledge of image formation and processing.
2. Specific knowledge regarding the properties of the face such as the aging process, temporary changes (e.g., makeup and hair change), permanent changes (e.g. formation of scars, loss of hair, cosmetic or plastic surgery), structure of bones and muscles, facial expressions and the involved muscle groups and comparison of ears and iris.
3. Understanding of the judicial system, awareness of the implications of a testimony, admissibility issues of facial comparison in court, presentation of facial comparison results and processes in court and to laymen.

2.2.4.3 Forensic individualisation from biometric data

In [43] basic concepts of forensic science are reviewed. Also a general forensic face recognition framework is proposed based on the Bayesian likelihood ratio

approach. Although this work is a comprehensive review of forensic concepts and provides a general description of the system, there is no experimental work described to prove the effectiveness of the proposed framework.

In forensic literature there is confusion between the terms identification and individualisation. If the class of individual entities is determined to be the source, it is called identification or classification. If a particular individual is determined to be the source, it is called individualisation. In the former case, the identity is called *qualitative identity* while in the later case the identity is called *numerical identity*.

In forensic science, the individualisation process is usually considered as a process of rigorous deductive reasoning, as a syllogism constituted of a major premise, a minor premise, and a conclusion. The major premise here in forensic face recognition context is the general principle of uniqueness applied to the source face and trace face. However, it is based on inductive reasoning which cannot be considered as a form of rigorous reasoning, because what is true for one instance is not necessarily true for all. While the demarcation criteria of empirical falsifiability reject the uniqueness of properties used for individualisation from face, this does not imply that face recognition cannot be used in forensic individualisation. It rather just puts a limit on the reliability depending on the quality of the images and method used.

To describe the likelihood ratio approach based on the Bayes theorem, two mutually exclusive hypotheses, the prosecution hypothesis (H_p) and defence hypothesis (H_d), can be defined as the set of all possible hypotheses for the inference of the identity of the source of a trace. Let I represent the background information about the case at hand and E the evidence. The likelihood ratio approach requires computation of E, *between-source variability* (BSV) and *within-source variability* (WSV). Fig.2.1 and 2.2 show how to incorporate the likelihood ratio approach into forensic individualisation as described in [43]. A more detailed description of the Bayesian framework and its application to forensic face recognition is presented in section 2.2.5

2.2.4.4 Automatic forensic face recognition from digital images

Automatic forensic face recognition from digital images is addressed in [24]. This paper describes small scale experimental work carried out by the Forensic Science Service in the UK, exploring the performance of an existing automatic face recognition system [44] in the forensic domain. The paper investigates the application of the Bayesian framework for forensic facial comparison and

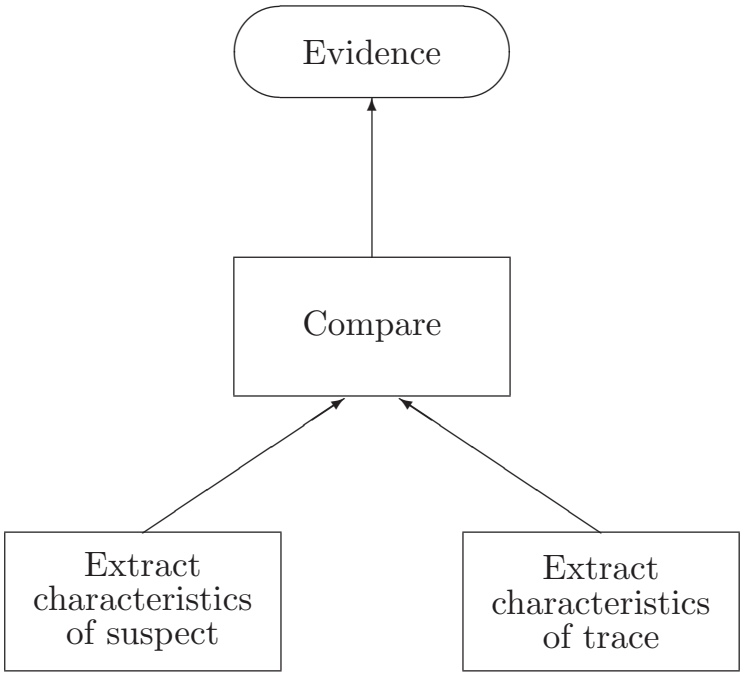


Fig. 2.1: Obtaining evidence in the Bayesian framework

decision making. Experiments are carried out using the Image Metrics OptasiaTM [44] software package for face recognition.

The approach of the Image Metrics OptasiaTM software used for experiments is straightforward. Active shape and appearance models [45], based on a general dataset of faces, are fitted to a new facial image. The fitted model consists of local information around landmark points in the facial image and forms a face template. To compare two faces, the similarity of the two face templates is determined. In [24] the similarity is expressed in a percentage (0-100%) and is called *recognition probability*. Given a database of n facial images, then a query image results in n recognition probabilities. Query images of persons included in the database are presented to the system and for each query image all n similarity scores are computed. The authors carried out three tests for evaluation of the system.

In the first test they used the same images as those in the database for benchmarking to get an idea of the the maximum performance of the technique.

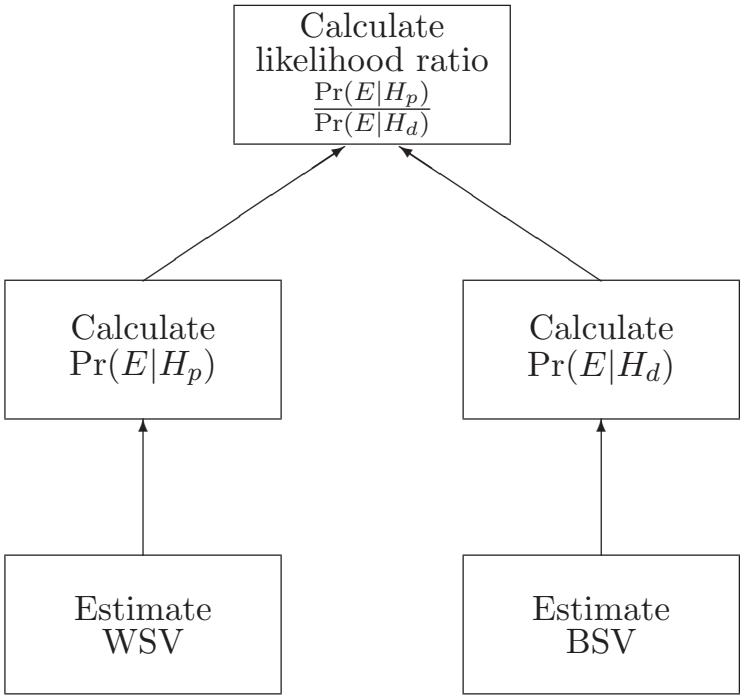


Fig. 2.2: Using the evidence to calculate the likelihood ratio

Twenty pictures chosen at random from the database were used as query images and a similarity score of greater than 95% was obtained for the correct match for each of the query images. The recognition probability sharply drops after the nearest match.

The second test was a feasibility test. For five persons in the database, new images, not present in the database, were obtained and captured exhibiting variation in pose, illumination, age, facial expression, resolution and image quality and used as query images to the system. In this experiment, illumination turned out to have the strongest effect on the recognition probability. The other variations had smaller but significant effects on the recognition probability.

Finally, for evaluation testing, the applicability to the forensic framework was investigated. To be able to calculate likelihoods, the WSV and BSV are needed. Five people of whom images were present in the database were photographed under similar conditions as those used to record images for the

database in order to estimate the WSV and the BSV of the database. Of each person 10 images were recorded, resulting in a set Q of 50 images. From this set Q, the WSV for each person was determined from the matching scores resulting from comparing the templates of the person to the template of the same person in the database. The BSV was obtained by matching all images in the set Q to all images in the database. Using the WSV and BSV, the likelihood ratio for a matching score can be calculated. For the set Q in 58% of the cases the comparison to the correct person in the database resulted in the highest likelihood.

The evaluation test provides a small scale, very limited assessment of the expected value or performance of the system in the forensics domain. There is no discussion on how the population size may influence the results.

2.2.4.5 Face matching and retrieval using soft biometrics

Although it does not directly focus on the forensic aspects of face recognition, the techniques and methodology proposed in [46] seem very attractive for forensic application of face recognition. Soft biometrics (ethnicity, gender and facial marks), if combined with a traditional face recognition system such as [47, 48] can improve the recognition accuracy as well as the ease of use and interpretation of the outcome in the forensic domain.

In [46] first facial landmarks are detected using an Active Appearance Model (AAM) [45]. Using these landmarks primary facial features are extracted and excluded in the subsequent facial marks detection process. First the face image is mapped to a mean facial shape to simplify the subsequent processing. The Laplacian of Gaussian (LoG) operator is utilised to detect facial marks. Each detected facial mark is classified in a hierarchical fashion as linear vs. not linear and circular vs. irregular. Furthermore, each mark is also classified based on its morphology as dark vs. light. In this way, each of the facial marks can be classified as a mole, freckle, scar etc.

Although the demonstrated performance of the proposed approach, using facial marks detection is not robust, facial marks nevertheless give a more descriptive representation of facial recognition accuracy compared to the numerical values obtained from traditional face recognition systems. This representation may be particularly useful in forensic applications. In such an approach semantic based queries can be issued to retrieve a particular image from a database. Furthermore, the facial marks can be used for facial comparison of partly occluded faces, which are quite common for surveillance cameras, and

may even allow differentiation of identical twins. In [46] experimental results are presented, based on the FERET [49] database and a mug-shot database that show that using the soft biometrics in combination with existing face recognition technology can improve the overall performance of the system and is more useful to forensic applications.

2.2.5 A Bayesian framework for forensic face recognition

The aim of a forensic biometric system is to report a meaningful value or expression in court to assess the strength of forensic evidence. The output of a biometric system cannot be used directly in forensic applications as discussed in detail in literature on forensic speaker recognition [2, 32, 33]. Systems using a simple threshold to decide between two classes resulting in a binary decision are not acceptable in the forensic domain [2]. For the purpose of forensic applications, the likelihood ratio framework is agreed upon as a standard way to report evidential value of a biometric system. This framework has been discussed in detail in the speaker recognition domain [32, 33] and the theory presented here benefits from it. However, unlike for forensic speaker recognition, there are very few published works which focus on the forensic aspects of face recognition and there is a serious need for reliable facial comparison and recognition systems which can assist law enforcement agencies in investigation and be used in courts.

The Bayesian framework is a logical approach and can be applied to any biometric system without change in the underlying theory. The likelihood ratio (LR) assessed from a score based biometric system can be used directly in court. While in commercial biometric systems, the objective is to present a score or decisions in a binary form, in forensic applications, the objective is to find the degree of support for one hypothesis against the other. Using the Bayes theorem, given the prior probabilities, the posterior probabilities can be calculated as:

$$\Pr(H_p|E,I) = \Pr(E|H_p,I)\Pr(H_p|I) \tag{2.1}$$

$$\Pr(H_d|E,I) = \Pr(E|H_d,I)\Pr(H_d|I) \tag{2.2}$$

where H_p and H_d are the prosecution and defence hypotheses respectively and E represents forensic information (evidence), while I is background information on the case at hand. The prosecution hypothesis H_p states that the

suspect is the source of the trace (in this case a facial image) while the defence hypothesis H_d states that someone else in a relevant population is the source. Equations 2.1 and 2.2 give the posterior odds required by judicial systems given the prior odds (background knowledge on the case) and likelihood ratio of the evidence E. The likelihood ratio

$$\mathrm{LR} = \frac{\Pr(E|H_p, I))}{\Pr(E|H_d, I))} \tag{2.3}$$

gives a measure of the degree of support for one hypothesis against the other, taking into consideration the circumstances of the case (background information I), and the result of the analysis of the questioned face. It calculates the conditional probability of observing a particular value of the evidence with respect to two competing hypotheses [50]. The numerator of the LR requires the WSV while the denominator requires the BSV to be calculated. This calculation of the LR from the WSV and BSV for a given matching score or evidence E is illustrated in Figure 2.3.

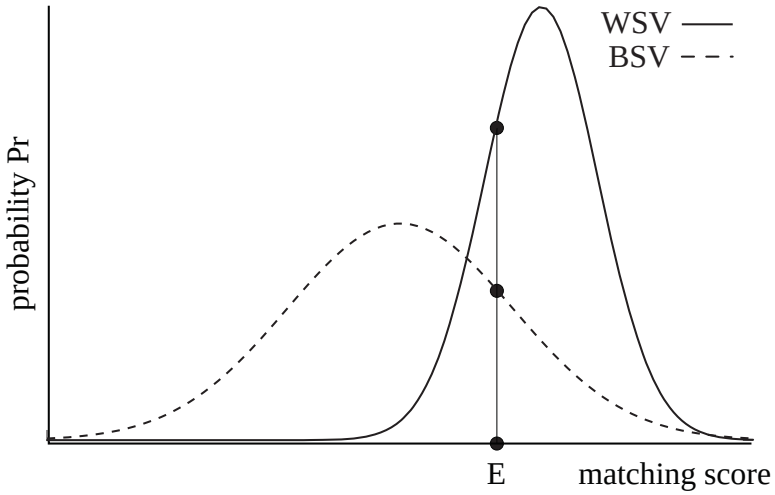


Fig. 2.3: Calculation of the LR from the WSV and the BSV. The solid curve represents the WSV or $\Pr(E|H_p, I)$ and the dashed curve the BSV or $\Pr(E|H_d, I)$. If a trace results in a matching score or evidence E, the LR is obtained by dividing the values of $\Pr(E|H_p, I)$ by $\Pr(E|H_d, I)$. Here the LR would be about 2.

The task of a forensic scientist is to evaluate the LR which is then used by the judicial system to reach a conclusion. Thus in order to use a score based biometric face recognition system, for calculation of the LR we need the following:

- The evidence E, a score obtained by comparing a trace face and a suspect

face.

- A distribution of matching scores obtained by comparing pairs of facial images of the suspect. This gives an estimate of the WSV and used to compute the numerator, $\Pr(E|H_p, I)$ of the likelihood ratio.
- A distribution of matching scores obtained by comparing pairs of facial images where one image is from a person in the relevant population and another is the trace. This gives an estimate of the BSV which is then used to estimate the denominator, $\Pr(E|H_d, I)$ of the likelihood ratio.

Variability exists in the nature of the pairs of images used to estimate the WSV and the BSV as discussed in Chapter 4. Figure 2.4 illustrates the complete procedure.

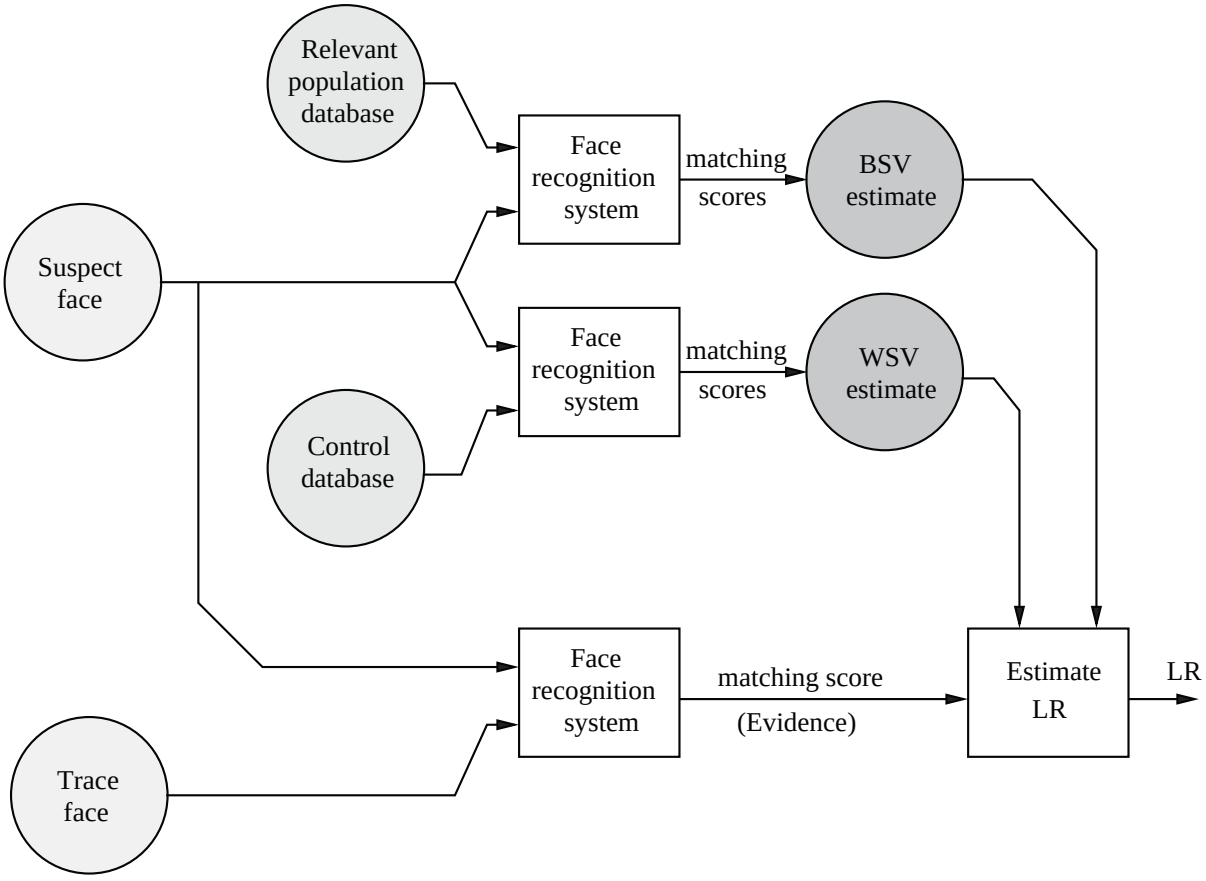


Fig. 2.4: Estimation of the LR. First the WSV and BSV are estimated using a Control database and a Relevant population database with images recorded under the same circumstances as the suspect facial image. Then the LR can be computed by comparing the trace facial image with the suspect facial image and using the WSV and BSV.

2.2.6 Reliability and court admissibility issues

The reliability of forensic face recognition is more critical compared to biometric face recognition where an incorrect decision at most results in e.g. a denial of access for a person to a building or a login restriction, the consequences of which are usually not very serious. In the forensic case, however, the consequences are far more severe. While it is agreed upon that future application of automatic facial recognition systems must be assisted by human experts for the final verification [42], the reliability of these systems themselves is also very important as it will reduce manual efforts and help standardise the process of facial comparison. In order to assess the reliability of forensic face recognition systems, several factors such as lighting conditions, facial expressions and pose etc., which are widely explored in the biometric domain should be considered here as well. If the Bayesian framework is used, other factors such as the number of images used to compute the BSV and WSV must also be taken into consideration. Apart from the sizes of the databases, it should also be ensured that the databases sufficiently cover the variations in imaging conditions (such as lighting conditions, image quality etc.) and facial appearance (such as pose, facial expressions etc). The Bayesian framework is the most logical framework, however it is sensitive to the methods used to determine the BSV and WSV. In particular, the distribution of H_p and H_d scores are probability density functions and their estimation is sensitive to the underlying mathematical model chosen. Therefore, a different modelling method can easily lead to different likelihood ratio values.

As a general rule, in order for the evidence extracted from forensic face recognition to be admissible in a court of law, the employed technology must be thoroughly tested and evaluated. In the United States this was ensured by application of the "Frye rule". It states that the judges should be acting as "gatekeeper" to asses if the technology on which the evidence is based is generally accepted in a relevant scientific community or not. Nowadays, in the United States, mostly a revised version of Frye rule called "Daubert" is in practice. It ensures that, in addition to general acceptance of the technology, the employed technology is tested and can be challenged in some objective way, the technology or theory must be peer-reviewed and a description of the error rate of the technology must be available. Finally, the technology must be maintained and adhere to standards.

In the European judicial system, there is no specific admissibility rule described regarding the scientific evidence. The judges are responsible for the evaluation of the scientific evidence pertaining to the case at hand.

2.2.7 Conclusions

At the moment there are no generally accepted standards for facial comparison by human forensic experts. However, several working groups are working on documents for standardisation and training.

Although much research and effort is put into improving state-of-the-art face recognition systems performance, far less effort is devoted to integrating face recognition technology with the legal system of court and justice. Only few papers have been published on the subject of how automatic face recognition can be adopted to forensic purposes. The output of a biometric face recognition system is not suitable for use in forensic applications and the output of a conventional score based biometric system must be processed in order to make it more useful and acceptable by the court.

The likelihood ratio value depends on the databases used for the estimation of the WSV and BSV and the underlying mathematical model of their distribution, it still provides the most logical framework for the judicial system to incorporate biometric evidence and background information on the case to reach a conclusion. There is an urgent need for "tuning" and integration of face recognition systems or development of new systems which can fulfil the requirements of law enforcement agencies and legal systems of court and justice.

Chapter 3

The effect of sampling variability in LR computation

3.1 Introduction

When a pair of biometric specimens are compared using a biometric recognition system, a score is computed. This score can be converted to a LR using the two sets of training scores s_p and s_d. The use of these two sets for training to learn the mapping function from scores to LRs introduces the problem of sampling variability. Statistically, the training sets are samples from populations and would vary if resampled from population. In this chapter the term "sample" refer to its statistical definition. The focus of this chapter is to introduce the issue of sampling variability in LR computation, review different LR computation methods and quantify the effect of sampling variability on different methods. In section 3.2, we discuss different LR computation methods and using a single pair of distributions of training scores $\{\{s_p\}, \{s_d\}\}$, quantify the effect of sampling variability. The focus is more on how the location (value) of the evidence score affects the sampling variability. Section 3.3 discusses the sampling variability focusing more on the shapes of the distributions of the scores in the training sets and the sizes of the training sets. The contents of section 3.2 are based on a published article while the contents of section 3.3 are based on a submitted article which is an extension of the article contained in section 3.2.

3.2 A review of calibration methods for biometric systems in forensic applications[1]

3.2.1 Abstract

When, in a criminal case there are traces from a crime scene - e.g., finger marks or facial recordings from a surveillance camera - as well as a suspect, the judge has to accept either the hypothesis H_p of the prosecution, stating that the trace originates from the subject, or the hypothesis of the defense H_d, stating the opposite. The current practice is that forensic experts provide a degree of support for either of the two hypotheses, based on their examinations of the trace and reference data - e.g., fingerprints or photos - taken from the suspect. There is a growing interest in a more objective quantitative support for these hypotheses based on the output of biometric systems instead of manual comparison. However, the output of a score-based biometric system is not directly suitable for quantifying the evidential value contained in a trace. A suitable measure that is gradually becoming accepted in the forensic community is the Likelihood Ratio (LR) which is the ratio of the probability of evidence given H_p and the probability of evidence given H_d. In this paper we study and compare different score-to-LR conversion methods (called calibration methods). We include four methods in this comparative study: Kernel Density Estimation (KDE), Logistic Regression (Log Reg), Histogram Binning (HB), and Pool Adjacent Violators (PAV). Useful statistics such as mean and bias of the bootstrap distribution of *LRs* for a single score value are calculated for each method varying population sizes and score location.

3.2.2 Introduction

Use of the Bayesian framework (or the LR framework) is gradually becoming a standard way of evidence evaluation from a biometric system. A general description of this framework can be found in [52]. It is applied to several biometric modalities including forensic-DNA comparison [5] and forensic voice comparison [7,53]. Preliminary results of evidence evaluation using this framework in the context of face recognition systems are presented in [23, 54]. In this framework the responsibility of a forensic scientist is to compute the LR.

[1]The contents of this section are published in [51], "A review of calibration methods for biometric systems in forensic applications", In: 33rd WIC Symposium on Information Theory in the Benelux, 24-25 May, 2012, Boekelo, Netherlands, pp. 126-133, ISBN 978-90-365-3383-6.

The evidence coming from a biometric system can be considered essentially as a realization of some random variable that has a probability distribution and the LR is the ratio of the distribution of this random variable under two hypotheses evaluated at the realized value of evidence:

$$LR(s) = \frac{P(s|H_p)}{P(s|H_d)} \tag{3.1}$$

where s is the evidence which is, in this context, a score value obtained from a biometric system by comparison of data from suspect with data found at the crime scene. This data can be a recording of speech signals or images etc depending on the type of biometric system. P is a Probability Density Function (pdf) if s is continuous or a Probability Mass Function (PMF) if s is discrete. H_p and H_d are two mutually exclusive and exhaustive hypotheses defined as follows:

H_p: The suspect is the source of the data found at the crime scene.
H_d: The suspect is not the source of the data found at crime scene or in other words, someone else is the source of the data found at the crime scene.

The LR calculates a conditional probability of observing a particular value of evidence s with respect to H_p and H_d. It is a concept which provides for evaluation and comparison of the two hypotheses concerning the likely source of the data obtained at the crime scene and which resulted in evidence s after comparison with suspect data using a biometric system. Once a forensic scientist has computed the LR, it can be interpreted as the multiplicative factor which update prior (before observing evidence from a biometric system) belief to posterior (after observing evidence from a biometric system) belief using the Bayesian framework:

$$\frac{P(H_p|s)}{P(H_d|s)} = \frac{P(s|H_p)}{P(s|H_d)} \times \frac{P(H_p)}{P(H_d)} \tag{3.2}$$

In this framework, the judge or jury is responsible for quantification of prior beliefs about H_p and H_d while the forensic scientist is responsible for quantification of evidence in the form of the LR given the evidence. It is clear from the definition of the LR that the distribution of evidence should be considered given the two hypotheses H_p and H_d. The job of forensic scientist is to express the evidence in relation to distibution of evidence given two competing hypotheses while the job of judge or jury is to assess the posterior probabilities of the two competing hypotheses given the evidence. To estimate probability

distribution that suspect is the source of the data found at the crime scene (assuming H_p is true), we need to collect a set of data from suspect under similar conditions to that of data captured at the crime scene. This set of data is compared to the data found at the crime scene using the given biometric system to obtain an estimate of the pdf under the hypothesis H_p. This estimate which is in the form of histogram of score values obtained from the same source comparisons is refered to as Within-Source Variability (WSV). Similarly, estimation of pdf under the defense hypothesis requires a set of data obtained from alternative sources. Comparison of this set of data to the data found at the crime scene results in an estimate of the pdf under the hypothesis H_d. This set of score values obtained from different sources comparison is refered to as Between-Source Variability (BSV). The set of alternative sources are sometimes refered to as relevant population and its choice and size may be affected by the background information about the case.

Obtaining the *LRs* (or calibrated score values) instead of un-normalized score values are desirable in several disciplines beside forensics such as medicine and diagnostics, cost-sensitive decision making and weather forecasting. The focus of this paper is to evaluate and understand different LR computation methods. The remaining of this paper is organized as follows: Section 3.2.3 briefly describes four commonly used LR computation methods. Section 3.2.4 discusses proposed evaluation procedure. Experimental results demonstrating performance of each method are presented in section 3.2.5. Section 3.2.6 finally concludes our work and presents future research work in this direction.

3.2.3 LR computation methods

LR computation methods compared in this study are well-known and therefore we only provide a brief description of each method along with suitable references for detail. MATLAB scripts of specific implementations of these methods used in this comparative study are available from the author.

3.2.3.1 Kernel Density Estimation (KDE)

This approach computes the *LRs* by first modeling pdfs of the WSV and the BSV scores and then finding the ratio of these pdfs at a given score value. A common approach to modeling these densities is using KDE [55]. KDE smooths out the contribution of each observed data point over a local neighborhood of that data point. The contribution of data point s_i to the estimate

at some point s depends on how far apart s_i and s are. The extent of this contribution is dependent upon the shape of the kernel function adopted and the width (bandwidth) accorded to it. If we denote the kernel function as K and its bandwidth by h, the estimated density at any point s is

$$f(s) = \frac{1}{n}\sum_{i=1}^{n} K\left(\frac{(s - s_i)}{h}\right) \tag{3.3}$$

Where n is the total number of data points. In our experiments we use a Gaussian kernel whose size can be optimally computed as [56]:

$$h = \left(\frac{4\hat{\sigma}^5}{3n}\right) \tag{3.4}$$

where $\hat{\sigma}$ is the standard deviation of the samples and n is the number of samples. Once estimated pdfs of the WSV and the BSV are obtained using eq.3, the LR is computed by plugging in values of these pdfs in eq.1. A detailed description of this approach to LR computation is presented by Meuwly [57] for forensic speaker recognition.

3.2.3.2 Logistic Regression (Log Reg)

Log Reg is a commonly used machine learning algorithm which gives probabilistic classification and has been widely used in forensic speaker recognition. Log Reg [58] fits a linear or a quadratic model to the log odds of the $P(H_p|s)$ which can be used in the Bayesian formula to compute the LR. Writing Bayesian formula using logit function

$$logitP(H_p|s) = LogLR(s) + logit(H_p) \tag{3.5}$$

Solving for $LogLR(s)$:

$$LogLR(s) = logitP(H_p|s) - logit(H_p) \tag{3.6}$$

$P(H_p)$ is known and we model $logitP(H_p|s)$ using logistic regression:

$$logitP(H_p|s) = \alpha + f(\beta, s) \tag{3.7}$$

where f is some function of s parameterized by β. Usually $logitP(H_p|s)$ is ordinary linear or quadratic logistic model, i.e., $\alpha + \beta s$. In our experiments

we use a quadratic model:

$$logitP(H_p|s) = \beta_0 + s\beta_1 + s^2\beta_2 \tag{3.8}$$

Parameters β_0, β_1 and β_2 are found from the WSV and the BSV by Maximum Likelihood Estimation (MLE).

3.2.3.3 Histogram Binning (HB)

Histograms of the WSV and the BSV similarity scores can be divided into bins in order to compute posterior probabilities of H_p and H_d given a score value [59]. Using Bayes rule in odds form:

$$\left(\frac{P(H_p|B_i)}{P(H_d|B_i)}\right) = LR(B_i) + \left(\frac{P(H_p)}{P(H_d)}\right) \tag{3.9}$$

where $i = 1, 2, ..n$ represents the number of the bin. $\left(\frac{P(H_p|B_i)}{P(H_d|B_i)}\right)$ is simply the ratio of the number of scores in the set of the WSV to the set of score in the BSV in bin i. For a given s, the LR value of the bin in which s lies is the required LR value of the s.

The choice of bin size is critical for the performance of the method. In [59] author has used fixed bin size by dividing the score axis into 10 bins; however, it results in empty bins when population size is low or when s is very high or very low. We propose an improved implementation by choosing the bin size based on the number of scores required in the sets of the WSV and the BSV for LR computation. For a given score value, the bin is placed symmetrically around the score value and the size is chosen such that it contains a required minimum number of the WSV and the BSV scores. This parameter representing the minimum number of scores of the WSV and the BSV can be varied for different score locations and population sizes to obtain optimal results. However, we do not assume any information about score location and population size and therefore keep this parameter fixed.

3.2.3.4 Pool Adjacent Violators (PAV)

Given data of the WSV and the BSV, PAV [60] sorts and assigns a posterior probability of 1 to scores of the WSV and 0 to scores of the BSV. It then iteratively looks for adjacent group of probabilities which violates monotonicity

and replaces it with average of that group. The process of pooling and replacing violator groups' values with average is continued until the whole sequence is monotonically increasing. The result is a sequence of posterior probabilities where each value corresponds to a score value from either the WSV or the BSV. These posterior probabilities along with the priors are used to obtain the LR values by application of the Bayesian formula. A detailed description of PAV algorithm can be found here [60].
It is interesting to note that computing Receiver Operating Characteristics Convex Hull (ROCCH) is equivalent to computing ROC of PAV transformed score values [61]. This argument leads to another way of implementation; computing ROCCH instead of the PAV procedure described in previous paragraph.

3.2.4 Data simulation and experimental setup

Most of biometric systems output are scores based on comparison of two samples which can be considered as a continuous random variable. Therefore computation of the LR ideally requires two pdfs: one is pdf of s under prosecution hypothesis $P(s|H_p)$ and other is pdf of s under defense hypothesis $P(s|H_d)$. However, in practice, these pdfs are not available and depending upon the LR computation method they are rather estimated from the data of the WSV and the BSV scores from a biometric system or the data is used to estimate the ratio of these pdfs. Given datasets of the WSV and the BSV, it is hard to evaluate performance of different methods of the LR computation for a given score value partially due to the fact that we do not know the ground truth value of the LR for that score value. Suppose we have access to the underlying pdfs which represent the distribution of data in the WSV and the BSV, we can easily evaluate the method by comparing its output LR with the one obtained from ratio of the pdfs. Using simulated data, our evaluation procedure is simple: assume standard pdfs for data of the WSV and the BSV, generate random data from these distribution for calibration of each method. Finally compute the LR for a given similarity score(s). This process of the WSV and the BSV data generation for calibration and the LR computation for a given score is repeated n times so that we have a distribution of n LR values for a given score value. Performance indicators such as mean, bias and standard deviation of the distribution of the LR values for each method can be studied for different parameters such as size of population and location of score value along score axis.

The choice of the types of these distributions and parameters are critical and

it is logical to base it on the background data from a biometric system. We observe the WSV and the BSV data from a speaker verification system [62]. Figure 3.1 shows histograms of the WSV and the BSV obtained from this system. We use MLE to estimate these histograms with different family of distributions. It turns out that to get best possible fit of a standard probability distribution to the data, it should first be flipped along maximum score (both WSV and BSV), then Weibull distributions are fitted to the flipped data of the WSV and the BSV. Once we have the 'best fit' of the flipped data in the form of standard Weibull distributions of certain parameters, we can generate data from these standard distributions and the generated data is flipped back to get a realization of original data of the WSV and the BSV. Figure 3.2 shows the fitted Weibull disstributions to the data shown in figure 3.1.

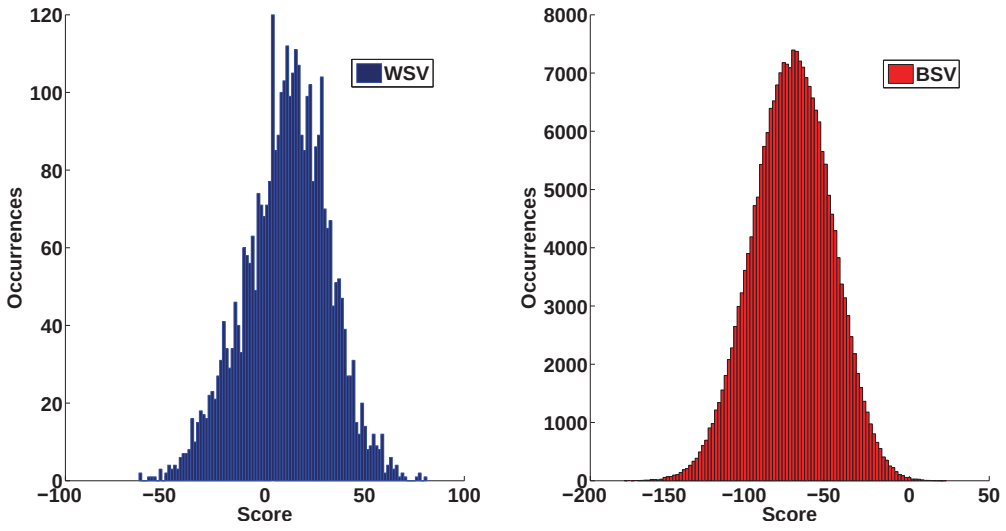


Fig. 3.1: Data of the WSV and the BSV of speaker verification system

3.2.5 Experimental results

We select five score values along the score axis and for each score value we estimate the distribution of the *LRs* using 10000 realizations of the WSV and the BSV data from the distributions shown in figure 3.2. Data is generated in the ratio of 1:63 from the pdf of the WSV and the pdf of the BSV. We study bias and standard deviation of each method using five population sizes. Bias is considered as a measure of accuracy while standard deviation is a measure precision. Figure 3.3-3.7 show the corresponding bias and standard deviation of the distributions of the *LRs* for given score values and population sizes. Population sizes shown are the sizes of the WSV data and the BSV is 63 times

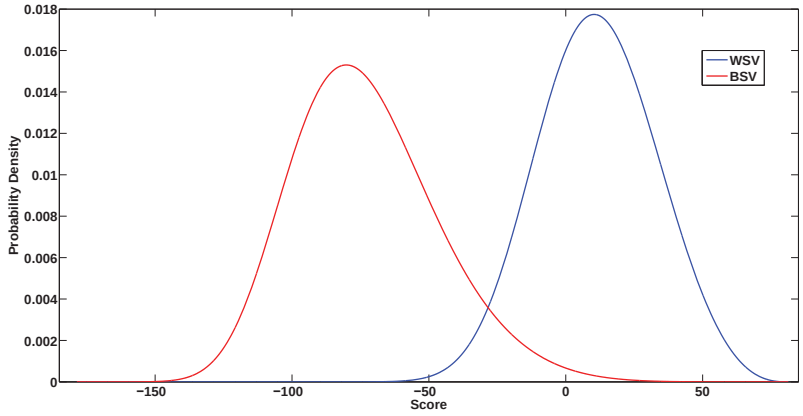


Fig. 3.2: Fitted Weibull distributions using MLE from data of WSV and BSV of speaker verification system

these numbers. Observing the results some comments are in order:

For all score locations, with the increase of population size standard deviation decreases. This is also true for the bias in all cases except for s = -40 where Log Reg gives fluxuating bias when population size is increased from 100.
Standard deviation and bias of HB approach is very high for small population sizes but it decreases faster with increase of population size. The parameter representing the number of scores required in bin to compute the LR in this approach can be adjusted to get improved result for a given score location and population size however we do not assume any knowledge about score location and population size when choosing value of this parameter. There is an interesting trade-off between bias and standard deviation of HB approach and therefore this provide more flexibility whether we can accept more bias or standard deviation. This might be a useful property when using this method for practical cases where we know whether more precise or more accurate value of the LR is desirable.
In most cases Log Reg perform better compared to other methods particularly it can guarantee very low standard deviation compared to other methods. This is due to the fact that the shape of backgorund distributions are closer to family of Gaussian distributions and the corresponding log odds can be estimated with high accuracy by a linear or a quadratic model. It is more biased for score values of -20 and +20 and this bias is not decreasing in usual way with increase of population size which shows that the parametric curve fitted to the log odds of $P(H_p|s)$ is not fitting to the true curve at these score locations. This can be a serious drawback of the parametric approaches that if the model is not appropriate, we cannot compute reliable value of the LR even by increasing population size.

KDE perform well in all cases except at high score values where we have fewer score values to estimate the density. The size of the kernel function can be adjusted to improve the results for a given score location however we use the same kernel throughout our experiments.
PAV is also attractive as it shows low bias. It has however the drawback that at very low and very high score values, it can result in zero and infinity values of the LR. In our experiments, values of zeros are considered as valid results while the LR values of infinity output by PAV are replaced by the maximum value in the sequence of the LR mapped from the scores of the WSV and the BSV.

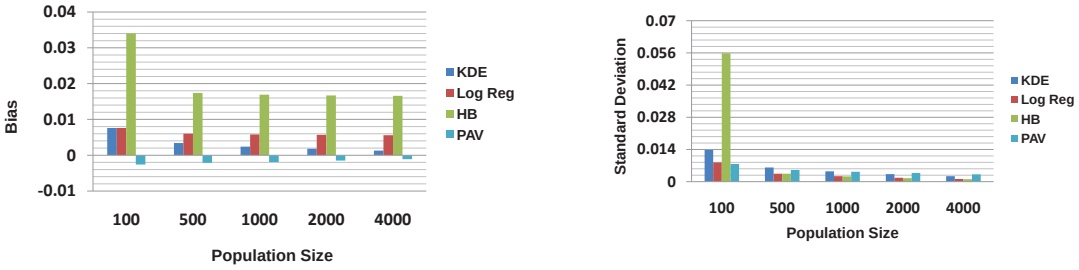


Fig. 3.3: bias and standard deviation of each method for s = -60

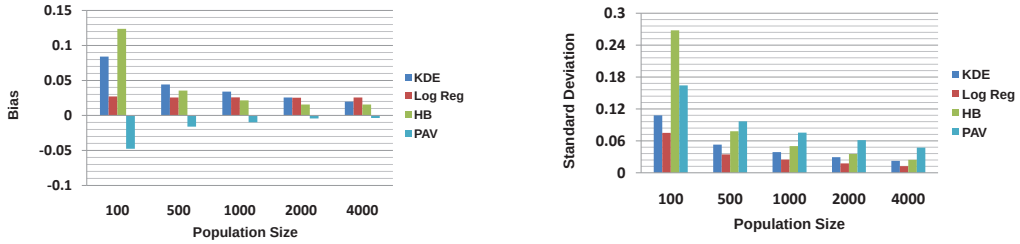


Fig. 3.4: bias and standard deviation of each method for s = -40

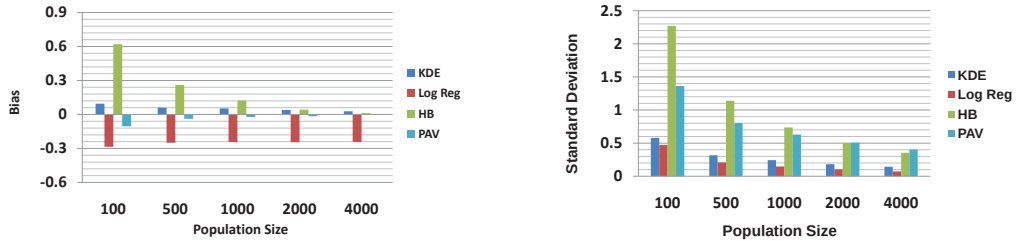


Fig. 3.5: bias and standard deviation of each method for s = -20

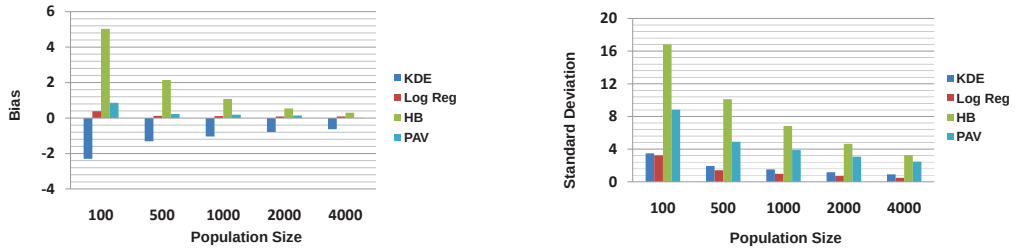


Fig. 3.6: bias and standard deviation of each method for $s = 0$

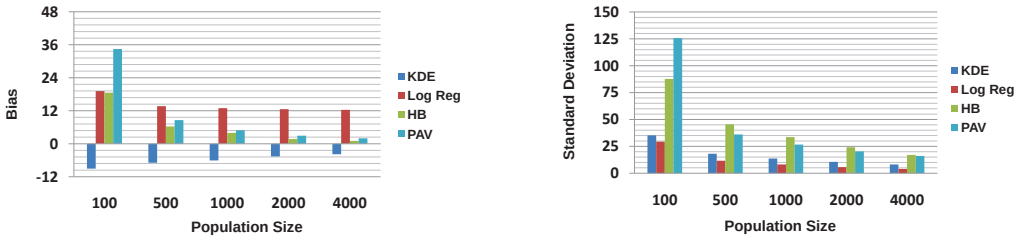


Fig. 3.7: bias and standard deviation of each method for $s = 20$

3.2.6 Conclusions and future work

In this paper we compared different methods of calibration for score-based biomeric systems. A simple methodology is presented for evaluation in terms of bias and standard deviation of the bootstrap distribution of the LRs for a given score value. Bias can be considered as how accurate a method performs while standard deviation is a measure of precision. Performance depends on three dependent parameters: background distributions representing the WSV and the BSV data, population size and location of score value along score axis. Generally it is hard to obtain accurate estimate of the LR when the score is very high or very low. The choice of which method to use depends on all the dependent parameters as well as on the fact whether more accurate or more precise value of the LR is desirable. Future research work includes working with the WSV and the BSV data from real biometric systems. It is expected that when the shape of background distributions deviate from the family of Gaussian distributions, the model fitting procedure will not be giving acceptable results in most cases.

3.3 Quantification of the sampling variability in forensic likelihood-ratio computation from biometric scores [2]

3.3.1 Abstract

Recently in the forensic biometric community there is a growing interest to compute a likelihood-ratio when a pair of biometric specimens are compared using a biometric recognition system. Using an existing biometric recognition system which produces a score, this requires a mapping from score to likelihood-ratio. A likelihood-ratio is the probability of the score given the hypothesis of the prosecution, H_p (the two biometric specimens are originated from a same source), divided by the probability of the score given the hypothesis of the defense, H_d (the two biometric specimens are originated from different sources). Given a set of training scores under H_p and a set of training scores under H_d, several methods exist to map from a score to a likelihood-ratio. In this work, we introduce the issue of sampling variability in the training sets and carry out a detailed empirical study to quantify its effect on commonly proposed likelihood-ratio computation methods. We study the effect of sampling variability varying : 1) the shapes of the probability density functions which model the distributions of scores in the two training sets; 2) the sizes of the training sets; 3) the actual value of the score for which the likelihood-ratio is computed. For this purpose, we introduce a simulation framework which can be used to study several properties of a likelihood-ratio computation method and to quantify the effect of sampling variability in a likelihood-ratio. A large number of training scores from a speaker and a face recognition system are used for devising the framework. It is shown empirically that sampling variability can be significant, particularly when the training sets are small. Furthermore, a given method of likelihood-ratio computation can behave very differently with different shapes of the distributions of the scores in the training sets. Situations are identified where well-known methods have undesirable properties when used for forensic evidence evaluation.

[2] The contents of this section are based on "Quantification of the sampling variability in forensic likelihood-ratio computation from biometric scores", IEEE transactions on information forensics and security (under review)

3.3.2 Introduction

For an automatic comparison of a biometric specimen from a known source and a biometric specimen from an unknown source, a metric called *score* can be computed using a biometric recognition system

$$s = g(x, y), \tag{3.10}$$

where x and y are the two biometric specimens, g is the biometric algorithm (feature extraction and comparison) and s is the computed score. In general, a score quantifies the similarity between the two biometric specimens while taking into account their typicality. In forensics, the known-source biometric specimen could come from a suspect while the unknown-source biometric specimen could come from a crime scene and the score is considered as an evidence. The use of biometric systems in applications such as access-control to a building and e-passport gates at some airports require the developer of the system to choose a threshold and consequently any score above the threshold implies a positive decision and vice versa [1]. This strategy works well in such applications; however, it presents several issues in forensic evaluation and reporting of the evidence from biometric recognition systems [52]. The selection of a threshold and therefore making a decision is not the province of the forensic practitioner. Furthermore, in most criminal cases, scientific analysis of the two biometric specimens provides additional information about the case at hand [63] and a threshold-based hard decision cannot be optimally integrated with other evidences in the case.

3.3.2.1 Likelihood-Ratio (LR)

There is a growing interest among forensic practitioners to use biometric recognition systems to compare a pair of biometric specimens. The concept of LR can be used to present the output of such a comparison. It has been extensively used for DNA evidence evaluation [8]. In general, given two objects, one with a known source and another with an unknown source, it is the joint probability of the occurrence of the two objects given H_p divided by the joint probability of the occurrence of the two objects given H_d [64–66]. When the two objects are the two biometric specimens x and y, compared using a biometric recognition system, the resultant score replaces the joint probability of the occurrence of the two specimens simplifying the computation of the

LR [6, 10]

$$LR(x,y) = \frac{P(x,y|H_p,I)}{P(x,y|H_d,I)} \approx \frac{P(s|H_p,I)}{P(s|H_d,I)}, \tag{3.11}$$

where I refers to background information which may or may not be domain specific.

$$\frac{P(H_p|s)}{P(H_d|s)} = \underbrace{\frac{P(s|H_p)}{P(s|H_d)}}_{\text{LR}} \times \frac{P(H_p)}{P(H_d)}, \tag{3.12}$$

where the background information, I, is omitted for simplicity. This is an appropriate probabilistic framework where the trier of fact is responsible for quantification of the prior beliefs about H_p and H_d while the forensic practitioner is responsible for computation of the LR.

The use of a LR is gradually becoming an accepted manner to report the strength of the evidence computed by biometric recognition systems. This is a more informative, balanced and useful metric than a score for forensic evaluation and reporting [63]. A general description of the LR concept for evidence evaluation can be found in [52, 63]. It is applied to several biometric modalities including speech [3, 7, 53, 67] and fingerprint comparison [68]. Preliminary results of the evidence evaluation using the LR concept in the context of face and handwriting recognition systems are presented in [10, 23, 54, 64].

3.3.2.2 Computation of a LR

In most cases, the conditional probabilities, $P(s|H_p)$ and $P(s|H_d)$, are unknown in the LR and they are computed empirically using a set of training scores under H_p, $s_p = \{s_j^p\}_{j=1}^{n^p}$ (a set of n^p number of scores given H_p) and a set of training scores under H_d, $s_d = \{s_j^d\}_{j=1}^{n^d}$ (a set of n^d number of scores given H_d) (see Fig.3.8). The biometric data sets used to compute s_p and s_d depend on the case at hand [4]. In general, the s_d scores are computed by comparing pairs of biometric specimens where the two biometric specimens in each pair are obtained from different sources whereas the s_p scores are computed by comparing pairs of biometric specimens where the two biometric specimens in each pair are obtained from a same source. An important condition in forensic LR computation is that the pairs of biometric specimens used for training should reflect the conditions of the pair of biometric specimens for which the LR is computed. Variability exists in the selection of the different-sources and

same-source pairs of biometric specimens for training, e.g., either the trace or the trace-like suspect's biometric specimen can be compared to the biometric specimens of the potential population (possible potential sources of the trace biometric specimen) to compute the different-sources scores [3, 6, 10, 15]. Same-source scores can be obtained by comparing trace-like biometric specimens from the suspect to the reference biometric specimens of the suspect or using a general approach, where pairs of biometric specimens from a same-source are compared. The effect of suspect-specific and generic scores in the training sets on a forensic LR is studied in [10, 15] for face and handwriting recognition. Please refer to [4] for an overview of the biometric data sets collection in forensic casework for a LR computation.

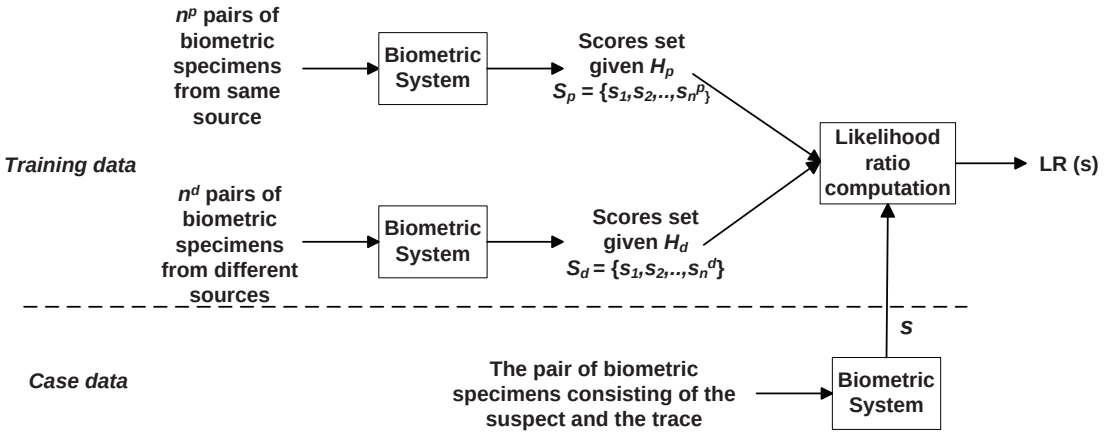


Fig. 3.8: Computation of a LR for a pair of biometric specimens consisting of the suspect's biometric specimen and the trace biometric specimen.

3.3.2.3 Sampling variability

Statistically, the training biometric data sets are samples from large populations of biometric data sets. The training biometric data sets, when resampled, will lead to slightly different values of the training sets due to the unavoidable sampling variability. This implies that the sets s_p and s_d consist of random draws from large sets of scores. When the resampling is repeated, slightly different LRs are computed for a given score. This is referred to as the "sampling variability" in a LR. It is desirable that a given LR computation method is less sensitive to the sampling variability in the training sets. If the PDFs of the scores under H_p and under H_d are known, a LR computed using given sets of training scores can be compared with the LR computed using the ratio of the two PDFs. The closeness of the two values implies the suitability of a given

LR computation method and in this article, we will refer to this performance indicator as "accuracy".

Note that in a given forensic case, the potential population and the suspect are deterministic parameters of the problem. The sampling variability, however, is due to the training scores sets that is needed to compute the mapping function from score to LR. This is because the training biometric data sets (and therefore the training scores) are finite and would vary one to the next in repeated random sampling. In practice, generation of multiple biometric data sets for training by resampling might not be feasible and therefore sampling variability is generally not measured. This is the motivation behind this study which provides an assessment of the sampling variability using a simulation framework. In passing, it should be pointed out that considering sampling variability implies uncertainty in a computed LR and different point of views exist on how to treat uncertainty in LRs [69–73].

3.3.2.4 Goal and organization of the paper

The purpose of this paper is to assess the sampling variability and accuracy of three commonly proposed LR computation methods. They are dependent on three parameters: 1) the shapes of the distributions of the scores in the training sets; 2) the sizes of the training sets; 3) the actual value of the score for which the LR is computed. A detailed empirical study is carried out in order to understand the merits and demirits of each method in evidence evaluation.

The paper is organized as follows. In section 3.3.3, we review existing work on comparison and assessment of different LR computation methods and describe our simulation framework. Section 3.3.4 briefly reviews the LR computation methods compared in this paper. Section 3.3.5 explains the experimental setup by describing the biometric scores sets, selection of the PDFs of the training scores sets and the proposed strategy to avoid infinite values in LR computation. Results are presented in section 3.3.6. Finally conclusions and future research directions are stated in section 3.3.7.

3.3.3 Comparison of LR computation methods

3.3.3.1 Existing work

The training sets s_p and s_d can be divided into two subsets: training subset $\{\{s_{p\_tr}\}, \{s_{d\_tr}\}\}$ and testing subset $\{\{s_{p\_ts}\}, \{s_{d\_ts}\}\}$. The most com-

mon approach to develop and assess LR computation methods is to learn the mapping function from score to LR using $\{\{s_{p\_tr}\},\{s_{d\_tr}\}\}$ while using $\{\{s_{p\_ts}\},\{s_{d\_ts}\}\}$, compute a set of test LRs $\{\{LR_{p\_ts}\},\{LR_{d\_ts}\}\}$ for performance assessment. We appreciate large LRs in $\{LR_{p\_ts}\}$ because they are computed for the pairs of biometric specimens obtained from same source and small LRs in $\{LR_{d\_ts}\}$ because they are computed for pairs of biometric specimens obtained from different sources. Based on this argument, performance of the LRs, $\{\{LR_{p\_ts}\},\{LR_{d\_ts}\}\}$, can be measured in a number of ways including calculation of the rate of misleading evidence in favour of H_p and H_d [10], Tippett plot [18], *Cost of Log LR* (C_{llr}) [17] and ECE plot [9]. This is a 'black-box' approach where performance is assessed using a set of test LRs, $\{\{LR_{p\_ts}\},\{LR_{d\_ts}\}\}$. Such assessments are used extensively and prove useful in practice, however, they do not consider the effect of the sampling variability in the training sets.

Morrison proposed the use of "credible intervals" to assess the sampling variability of a LR [74]. His work is focused on the sampling variability in the evidence score s for which the LR is computed and not the one in the training scores sets. In [71], a strategy based on confidence intervals is used to measure the sampling variability when the frequencies associated with LR computation in DNA evidence are estimated using a sample from the population. For forensic speaker recognition, Alexander [75, section 3.5] used "Leave-one-out" strategy in order to assess the variability in a LR due to small change in the training sets. One score is removed per run, with replacement, from the training scores and a LR is computed. This lead to a distribution of LRs for a given score which can be used to assess the variability. However, this procedure does not address the issue of sampling variability in the training sets where the assumption is that the training sets are samples from populations.

In this work, we carry out a general study, varying the shapes of the distributions of the scores in the training sets as well as the sizes of the training sets in order to understand their effects on the sampling variability and accuracy of LR computation methods. An analysis of the sampling variability constitutes a measure of reliability of a LR computation method and is important in forensic science as recently discussed by Morrison [69, 76].

3.3.3.2 Proposed simulation framework

Since the score output by most biometric systems is a continuous random variable, computation of a LR ideally requires the PDF of scores under H_p

and the PDF of scores under H_d. However, in practice, these PDFs are not known. Only the two sets of training scores, s_p and s_d, are available which can give a rough idea of how the corresponding PDFs look like. Suppose we have access to the underlying PDFs from which the training sets s_p and s_d are generated, a LR of a score s from the PDFs is computed as:

$$LR_{\infty,\infty}(s) = \frac{f_{s_p}(s)}{f_{s_d}(s)}, \tag{3.13}$$

where f_{s_p} is the PDF of scores given H_p and f_{s_d} is the PDF of scores given H_d. For the ease of presentation, we introduce the notation $LR_{n^p,n^d}(s)$ to represent the LR of a given score s computed using n^p number of the s_p scores and n^d number of the s_d scores. Using the LR computed from the PDFs as a benchmark, a form of accuracy, based on the fact that how close the two LRs, $LR_{n^p,n^d}(s)$ and $LR_{\infty,\infty}(s)$ are, can be measured[3]. Our procedure to measure the sampling variability and the accuracy can be summarized as follows:

- Match standard PDFs to large s_p and s_d sets of scores of a biometric recognition system using Maximum Likelihood Estimation (MLE) or assume standard PDFs which are similar to commonly observed distributions of the s_p and the s_d scores.
- Generate n realizations of the s_p and the s_d scores from these standard PDFs by random statistical sampling.
- For a given score, n LRs can be computed using the n random realizations of $\{\{s_p\},\{s_d\}\}$(see Fig.3.9).
- The standard deviation and the difference between the minimum and the maximum values of the set of n LRs of a score can be used to measure the sampling variability while the mean value of the n LRs of a score along with the $LR_{\infty,\infty}$ of the score can be used to measure the accuracy.
- A smaller value of the standard deviation and a smaller value of the difference between the maximum and the minimum values imply the method is less sensitive to the sampling variability in the training sets. Similarly, the closer the mean value of the n LRs and the $LR_{\infty,\infty}$ of the score are, the more accurate a method is. Furthermore, using the mean value, a bias value can be computed as $bias = LR_{\infty,\infty} - mean$.

[3]The term "accuracy" is also used sometimes to refer to the system performance evaluated using the tools mentioned in section 3.3.3.1 (Tippett plot, C_{llr}, etc). That measure of accuracy is based on the ground truth information that whether a pair of biometric specimens has been originated from a same source or from different sources as the benchmark. The measure of accuracy of interest in this paper is based on the knowledge of the true parameters of the probability distributions from which the training sets are generated and uses the LRs obtained from the assumed PDFs as a benchmark.

Note that random statistical sampling of the two PDFs corresponds to the random sampling of the biometric data sets from the suspect and potential population for computation of the training scores.

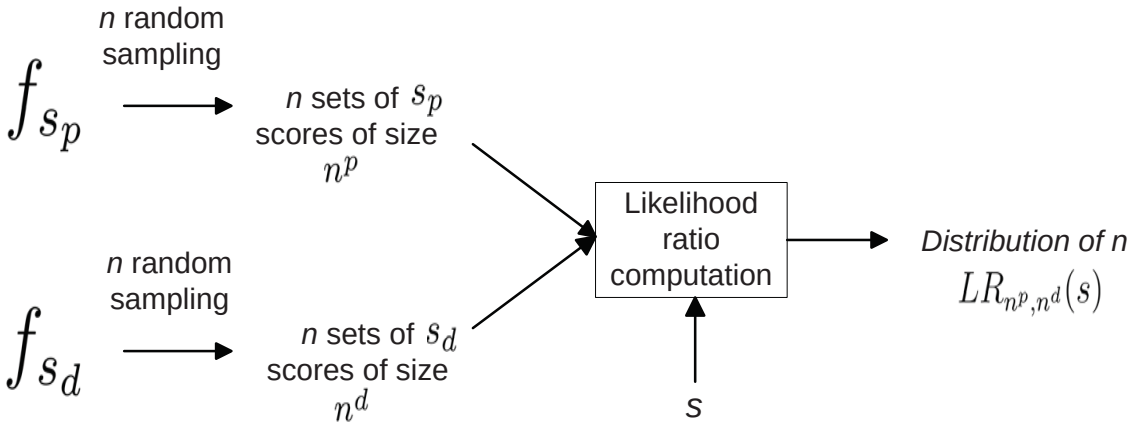


Fig. 3.9: Generation of n realizations of the training sets by random sampling and computation of n LRs of a given score s. The standard deviation, minimum LR, maximum LR and mean LR follow from the set of n LRs of the score s.

Using this procedure, each method of LR computation can be studied based on different shapes of the distributions of the scores in the training sets, different sizes of the training sets and the actual value of the score for which the LR is computed.

It can be argued that the proposed experiment cannot be performed for an exhaustive set of biometric systems or shapes of the distributions of scores. However, based on a few typical PDFs, it can provide a useful insight into the behaviour of different methods of LR computation and in the assessment of the sampling variability in a computed LR. For the purpose of performance assessment, this procedure has several advantages over specific sets of real scores from a biometric system:

- The benchmark value of the LR of a given score s, $(LR_{\infty,\infty}(s))$, is known and can be used to measure the accuracy of different methods of LR computation.
- Multiple realizations of the training sets can be generated by repeated random sampling of the PDFs. These sets simulate the sampling variability when the biometric data sets are repeatedly resampled for computation of the training sets.
- The sizes of the training sets can be increased or decreased easily to quantify its effect on LR computation methods.

- The characteristics of the PDFs can be altered to see how the shapes of the distributions of the scores in the training sets affect the LR computation methods.
- The separation between the assumed PDFs can be increased or decreased which is related to the discriminating power of a biometric system and affects different LR computation methods differently.

The choice of the types and the parameters of these PDFs is critical. We consider four pairs of PDFs; two pairs of PDFs are selected based on their best MLE fitting to the s_p and s_d sets of two biometric recognition systems whereas the other two pairs of PDFs are assumed based on their general proximity to the shapes of the distributions of the s_p and s_d scores sets in the available litrature. With this, we cover a variety of the distributions of scores expected from different biometric recognition systems. The details of the selected PDFs and the fitting procedure will follow in a later section.

3.3.4 LR computation methods

We consider three LR computation methods commonly proposed for the evaluation of evidence. A brief description of these methods is given in the following. MATLAB scripts of all experiments along with the biometric scores sets will be made available online.

3.3.4.1 Kernel Density Estimation (KDE)

This approach estimates the PDFs of the s_p and the s_d scores using KDE [55] and then compute the quotient of these estimated PDFs at the score location s to compute the LR of s. KDE smooths out the contribution of each observed data point over a local neighborhood. The contribution of the score s_i in the training set to the estimate at score location s depends on how far apart s_i and s are. The extent of this contribution is based on the shape and width of the kernel function. If we denote the kernel function as K and its width by h, the estimated density at score s is

$$f(s) = \frac{1}{n}\sum_{i=1}^{n} K\left(\frac{(s - s_i)}{h}\right) \tag{3.14}$$

where for PDF of the s_p scores, n is the total number of scores in the s_p set and for PDF of the s_d scores, n is the total number of scores in the s_d set.

In our experiments, a Gaussian kernel is used where the width is optimally chosen as [56]:

$$h = \left(\frac{4\hat{\sigma}^5}{3n}\right) \tag{3.15}$$

where $\hat{\sigma}$ is the sample standard deviation. A detailed description of this approach to LR computation is presented by Meuwly [57] in the context of forensic speaker identification and subsequently adapted by [23, 54, 77] for other biometric modalities.

3.3.4.2 Logistic Regression (Log Reg)

Log Reg [78] uses a linear or a quadratic function to map a score s to the log odds $\log\left(\frac{P(H_p|s)}{1-P(H_p|s)}\right)$. These log odds along with the sizes of the training sets can be used in the Bayesian formula in eq.3.12 to compute the LR. We choose a linear function since it is more common in forensic likelihood ratio computation [17, 78]:

$$\log\left(\frac{P(H_p|s)}{1-P(H_p|s)}\right) = \beta_0 + s\beta_1 \tag{3.16}$$

parameters β_0 and β_1 are found from the training sets of scores using Iteratively Reweighted Least Squares (IRLS) algorithm [79]. Log Reg is a well-known algorithm in machine learning and statistics and is widely used for LR computation in several biometric modalties including forensic speaker, fingerprint and voice comparison [3, 17, 78].

3.3.4.3 Pool Adjacent Violators (PAV)

Given the sets of the s_p and the s_d scores, PAV algorithm (also called Isotonic Regression) [60] combines them into a single set, sorts and assigns a posterior probability $P(H_p|\,\text{score})$ of 1 to each score of the s_p and 0 to each score of the s_d set. It then iteratively looks for adjacent group of posterior probabilities that violates monotonicity and replaces it with the average of that group. The process of pooling and replacing violator groups' values with the average is continued until the whole sequence is monotonically increasing. The result is a set of posterior probabilities $P(H_p|\,\text{score})$ where each value corresponds to a score from either the s_p or the s_d set. The odds of these posterior probabilities along with the sizes of the training sets are used to obtain the LRs by

application of the Bayesian formula in eq.3. To compute a LR for a score s, linear interpolation is used between the training scores. A detailed description of the PAV algorithm can be found in [60].

It should be mentioned that, another commonly used method to compute LRs from scores is by finding the slope of the Receiver Operating Characteristics Convex Hull (ROCCH). However, it is recently proved that this method is equivalent to PAV [61]. Furthermore, PAV can be considered as an optimized version of the LR computation using histogram binning [80] because PAV chooses optimal bin size depending on the size of the training data in different score locations.

3.3.5 Experimental setup

3.3.5.1 Selection of PDFs

We consider four pairs of PDFs from which the n realizations of the training sets are generated by repeated random sampling. These training sets are used to assess the sampling variability and accuracy of each LR computation method using the simulation framework discussed in section 3.3.3.2. The first pair of PDFs consists of two Normal PDFs shown in Fig.3.10(a). The specific values of the parameters are chosen such that there is some overlap between the two PDFs and the standard deviation of the PDF of the s_d scores is larger than the PDF of the s_p scores (see Table. 3.1). The choice of Normal PDFs

Table 3.1: Parameters of the assumed Normal PDFs

	μ	σ
f_{s_p}	23	2.3
f_{s_d}	13	4

in this comparative study is motivated by its widespread use to model the distribution of scores in the s_p and in the s_d set in various biometric modalities such as handwriting and fingerprint recognition [10, 68].

The second pair of PDFs consists of two reversed Weibull PDFs shown in Fig.3.10(b) and expressed as:

$$f(s;\lambda,k) = \frac{k}{\lambda}\left(\frac{s_{max}-s}{\lambda}\right)e^{\frac{(s_{max}-s)}{\lambda}}, \tag{3.17}$$

where $k > 0$ is the scale parameter and $\lambda > 0$ is the shape parameter in the Weibull distribution. $s_{max} = 4$, found experimentally, is the score value along which the PDFs are reversed. Table 3.2 shows the values of k and λ for the two PDFs. The choice of these specific PDFs is motivated by the shape of the

Table 3.2: Parameters of the assumed Weibull PDFs

	k	λ
f_{s_p}	0.7	0.2
f_{s_d}	5	2.5

distributions of scores obtained by face recognition systems based on Boosted Linear Discriminant Analysis (BLDA) [23, 81].

Next we consider large sets of the s_p and the s_d scores from a speaker recognition system based on probabilistic Linear Discriminant Analysis (PLDA) [62] approach which models the distribution of i-vectors as a multivariate Gaussian. The system is described in [62, section 2.5] and used with the biometric data of the National Institute of Science and Technology (NIST) Speaker Recognition Evaluation (SRE) of 2010 [82]. Fig.3.10(c) shows the s_p set, the s_d set and the matched pair of PDFs using MLE. These are reversed Weibull PDFs, flipped along s_{max} as in equation 3.17, with different parameters as shown in Table 3.3. In this case, s_{max} is the maximum value of the score in the training set $\{\{s_p\}, \{s_d\}\}$.

Table 3.3: Parameters of the Weibull PDFs fitted to the s_p and s_d sets of the speaker recognition system shown in Fig.3.10(c).

	k	λ
f_{s_p}	3.59	77.67
f_{s_d}	6.80	165.29

Lastly we consider a state-of-the-art commercial face recognition system developed by Cognitec [83]. It is used with face images from SCFace database [84] which includes low resolution images captured using surveillance cameras. There are 5 different qualities of surveillance cameras and 130 subjects in the database. Each subject has a frontal mugshot which is compared to the surveillance cameras images. The resultant s_p and s_d scores sets are shown in Fig.3.10(d). Using MLE, the s_p scores set is best fitted by the Uniform PDF and the s_d scores set by the Beta PDF shown in Fig.3.10(d) and expressed as:

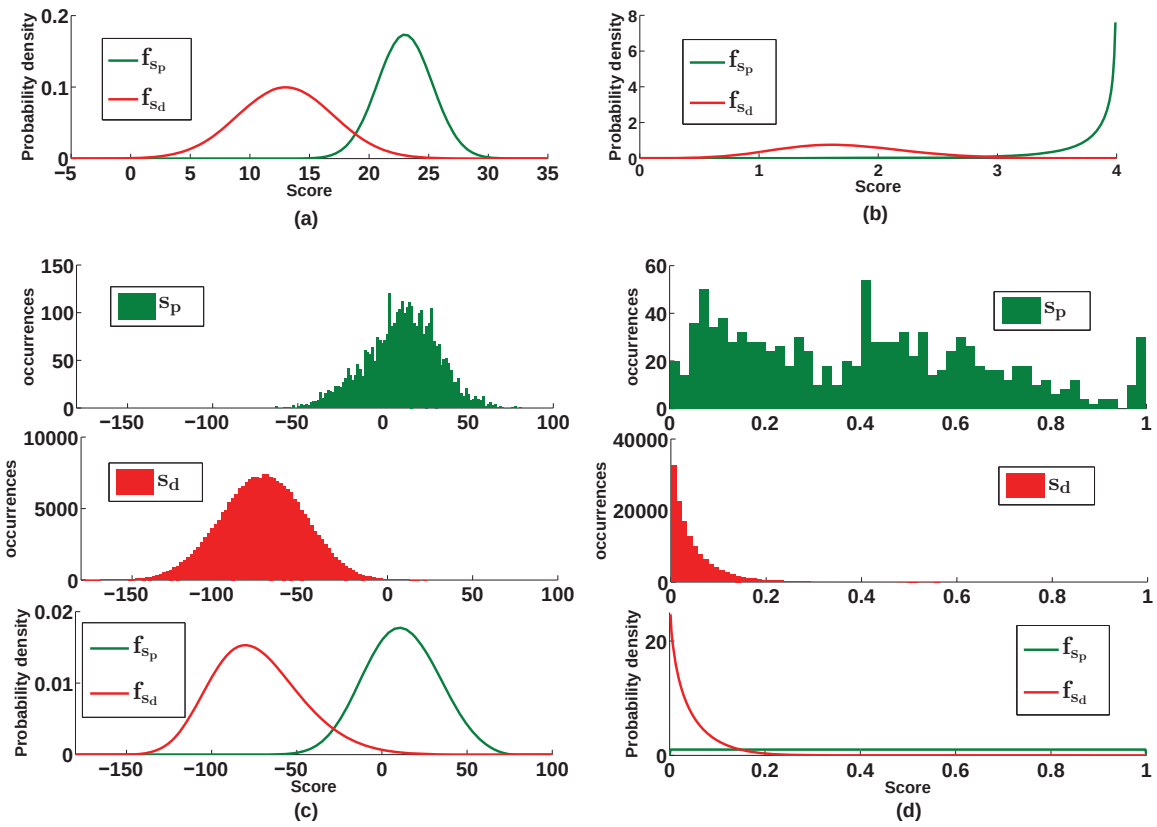


Fig. 3.10: pairs of PDFs from which n realizations of the training sets are generated by random sampling. (a) Assumed Normal PDFs. (b) Assumed reversed Weibull PDFs. (c) Scores sets from the speaker recognition system and the fitted reversed Weibull PDFs. (d) Scores sets from the Cognitec face recognition system and the fitted Uniform and Beta PDFs.

$$f_{s_p}(s;a,b) = \begin{cases} \frac{1}{b-a} & \text{for } a \leq s \leq b \\ 0 & \text{otherwise} \end{cases}$$

$$f_{s_d}(s;\alpha,\beta) = \frac{1}{B(\alpha,\beta)} s^{\alpha-1}(1-s)^{\beta-1}$$

where values of a, b, α and β are 0.0020, 0.9995, 0.8532 and 17.3373 respectively found using MLE.

3.3.5.2 Avoiding infinite Log-likelihood-ratios

In order to avoid computation of infinite log-likelihood-ratios, we insert the score s, for which the LR is computed, both in the s_p and in the s_d set. A slightly different strategy has been proposed in [17] where two additional scores are inserted in these sets: one at the maximum possible score location and another at the minimum possible score location. In general, both of these strategies are motivated by the Laplace's rule of succession. The inserted scores can be considered to represent training scores which were not encountered in the training sets because there is not enough training biometric data but which could have occurred.

3.3.6 Results

PDFs in Fig.3.10 are randomly sampled and 5000 realizations of the training sets $tr_1 = \{\{s_{p_1}\},\{s_{d_1}\}\}, tr_2 = \{\{s_{p_2}\},\{s_{d_2}\}\}, ..., tr_{5000} = \{\{s_{p_{5000}}\},\{s_{d_{5000}}\}\}$ are generated. In practice, the sizes of the s_p and s_d sets available for computation of a LR is case-dependent and can vary significantly. We consider three sizes of the training sets (n^p, n^d): $(2000, 100000)$, $(200, 10000)$ and $(20, 1000)$. The choice of the small size of the s_p set compared to the s_d set is motivated by the fact that it is generally the case in practice.

In order to understand the effect of the shape of the distribution of the scores in the s_p and in the s_d set and the sizes of these two sets used for training, a fixed score s is considered. This score is shown by a vertical line in each pair of PDFs in the leftmost column in Fig.3.11. The value of s is chosen such that its Log_{10} Likelihood Ratio (LLR) computed from the PDFs is 0; i.e., $LLR_{\infty,\infty}(s) = 0$ or $LR_{\infty,\infty}(s) = 1$. The right three columns of Fig.3.11 show, for each method, the histograms of 5000 $LLR_{2000,100000}(s)$ values computed for the score s using the training sets $tr_1, tr_2, .., tr_{5000}$. The size of tr_i is $(n^p, n^d) = (2000, 100000)$. The solid vertical line in the histograms shows the $LLR_{\infty,\infty}(s)$

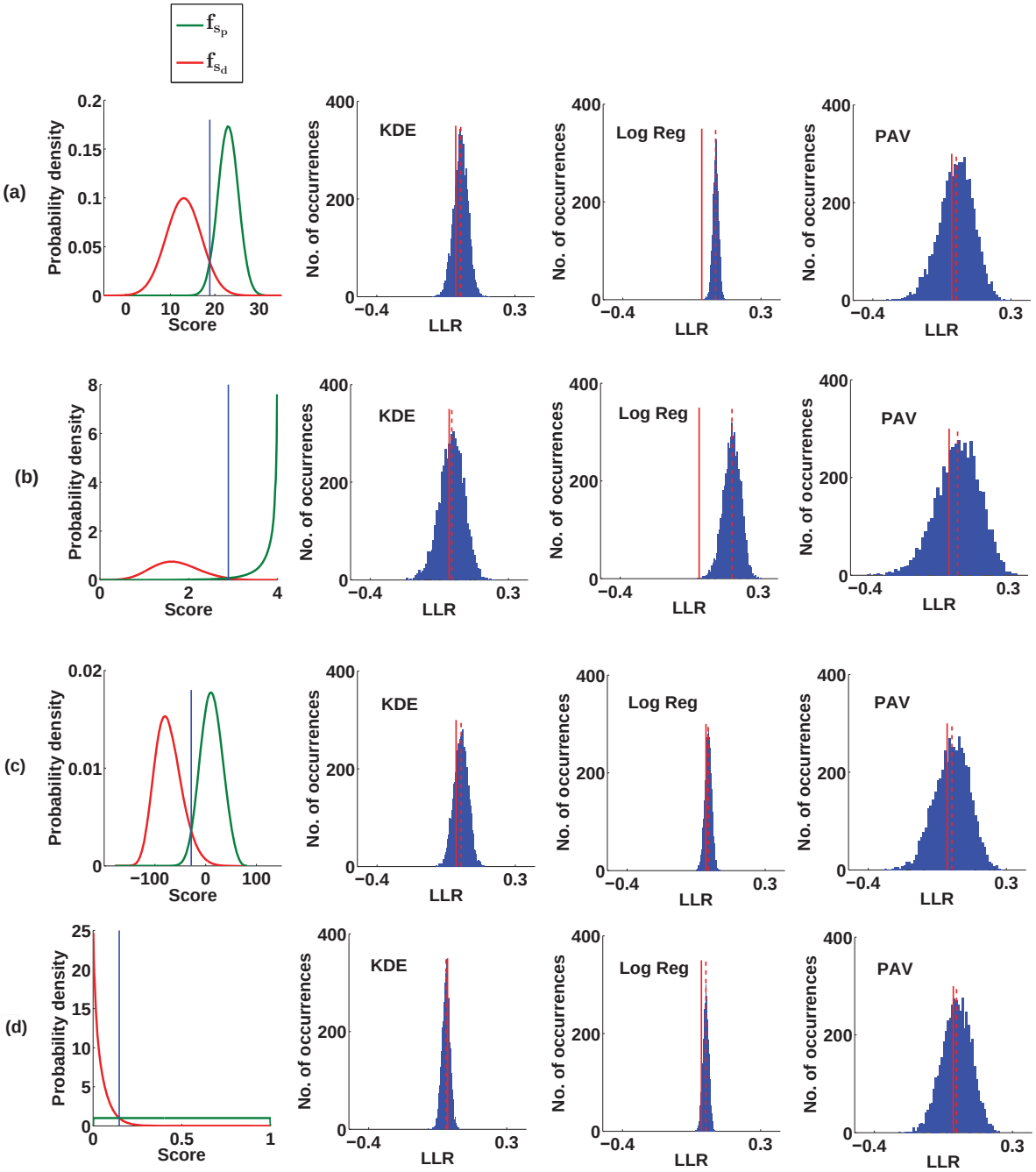


Fig. 3.11: The leftmost column shows the PDFs with the considered score s shown as a vertical line. The next three colums show histograms of the 5000 $LLR_{2000,100000}(s)$ values computed by each method using the training sets $tr_1, tr_2, .., tr_{5000}$ where size of $tr_i = (n^p, n^d) = (2000, 100000)$ and generated using random sampling from the corresponding pairs of PDFs.

whereas the dotted vertical line shows the mean LLR of s computed from the set of 5000 LLRs. The distance between the solid line and the dotted line gives an estimate of the accuracy of the method whereas the standard deviation of the histograms of LLRs gives an estimate of the sensitivity of a method to the sampling variability in the training sets. The use of a logarithmic scale is preferred for plotting purposes as well as it has intuitive apeal for forensic practitioners. As an example, a LLR of 1 ($LR = 10$) can be interpreted as "The probability of the evidence is 10 times more given the two biometric specimens are originated from a same source than if they were originated from different sources" whereas a LLR of -1 ($LR = 0.1$) can be interpreted as "The probability of the evidence is 10 times more given the two biometric specimens are originated from different sources than if they were originated from a same source". From Fig.3.11, it can be concluded that, based on these large sizes of the training sets, Log Reg is least sensitive to the sampling variability in the training sets followed by KDE and then PAV. However, the mean LLR in case of PAV and KDE is closer to $LLR_{\infty,\infty}$ for distribution pairs in Fig.3.11(a,b,d) demonstrating the better generalization ability of these non-parametric approaches across different shapes of the scores distributions. The mean values computed by Log Reg in Fig.3.11(a) and Fig.3.11(c) also show the fact that a small difference in the shapes of the distributions of the scores in the training sets can significantly affect the accuracy of the Log Reg approach.

To see the effect of reduction in the sizes of training sets, Fig.3.12 repeats the results when the sizes of the s_p and the s_d sets are 20 and 1000 respectively. Note the larger range of the x-axis in the histograms, showing, in general, a large standard deviation of the LLRs due to reduction in the sizes of the training sets. For these small training sets, Log Reg outperforms KDE and PAV in terms of the accuracy of the computed LRs, however, it is more sensitive to the sampling variability in the training sets compared to KDE.

Sampling variability can be significant in case of small training sets. As an example, when the size is (20,1000), the minimum and maximum LRs computed by KDE in Fig.3.12(d) are $\frac{1}{6.56}$ and 2.04 respectively, resulting in the closed interval, $[\frac{1}{6.56}, 2.04]$, in which the LR value could lie. This implies that, given the two input biometric specimens, the exact value of the LR lies in this interval and therefore it suggests that a range of LRs should be presented in court. Sampling variability may not be a serious concern in cases where a very large or a very small value of LR is computed. However, it will still be a good practice to perform and include an assessment of the sampling variability when reporting the strength of a biometric evidence in the form of a LR. The corresponding intervals of LRs in Fig.3.12(d) in case of Log Reg and PAV are

$[\frac{1}{6.83}, 2.30]$ and $[\frac{1}{1.43}, 9.53]$ respectively. Note that these intervals of LRs are for a score lying in a location which is expected to be less sensitive to the sampling variability in the training sets because of the large number of data points in this score location. We will discuss the effect of varying the score location later.

Fig.3.13 summarizes the effect of the sizes of the training sets on the standard deviation of the 5000 LLRs and on the bias ($mean - LR_{\infty,\infty}$) of each method. Some notes are in order:

- The sizes of the training sets has little effect on the bias of the Log Reg method. The bias, however, is dependent on the shapes of the distributions of the scores. This is one of the main drawback of most parametric approaches, in general; Once the model is not appropriate, the sizes of the training sets cannot compensate for it. In contrast to the bias, the sampling variability in case of the Log Reg method is dependent on both the sizes of the training sets as well as the shapes of the distributions of the scores.
- Instead of the shapes of the distributions of the scores in the training sets, the bias and sampling variability in case of the PAV method is strongly dependent on the sizes of the training sets. This can be concluded from Fig.3.13 by noticing similar behaviour of the PAV method across the four different pairs of the PDFs of scores.
- The KDE method follows similar behaviour as the PAV method but it is less biasd and less sensitive to the sampling variability when the sizes of the training sets are reduced significantly, e.g., the case of $(20, 1000)$, as shown in Fig.3.13.

Table 3.4 shows the mean, maximum and minimum values of the 5000 LLRs computed by each method considering different sizes of the training sets.

In order to study the effect of the value of the score s for which the LR is computed, the sizes of the training sets are kept fixed. Since sampling variability is significant when the sizes of the training sets are small, we consider the sizes $(n^p, n^d) = (20, 1000)$. A set of 50 equidistant scores are generated in the range $[min\,(s_{d_i})\, max(s_{p_i})]$ for $[i = 1, 2, .., 5000]$. Then, as previously, for each score in this set, 5000 LLRs are computed using the 5000 training sets $t_{r_1}, t_{r_2}, .., t_{r_{5000}}$ generated by random sampling of the PDFs. The mean, maximum and minimum values of the 5000 LLRs are plotted for each score in the set of 50 equidistant scores (see Fig.3.14). Only results for the pairs of PDFs in Fig.3.11-3.13(c,d) are shown.

The sampling variability in a LR is more significant when the score location is

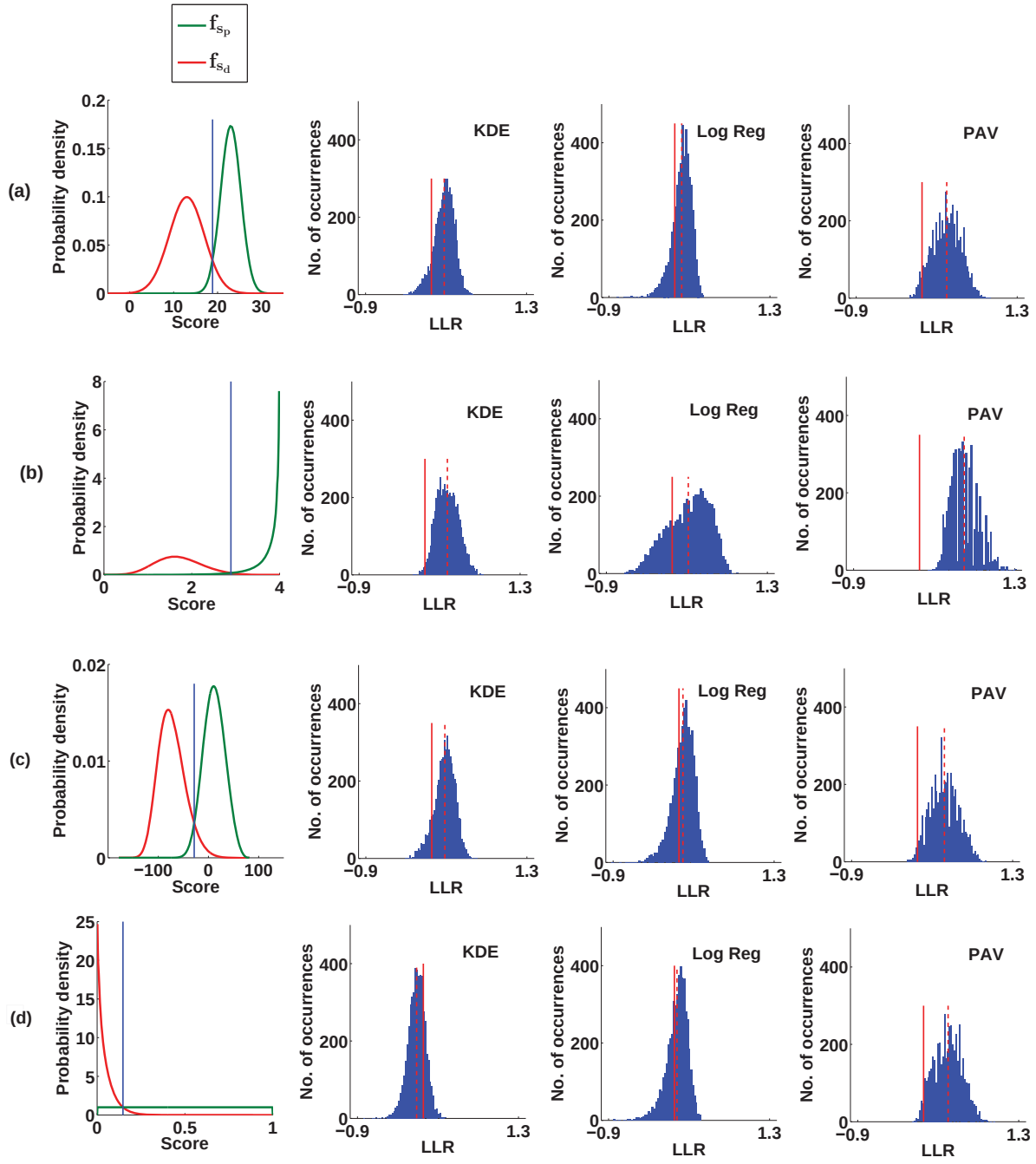


Fig. 3.12: The leftmost column shows the PDFs with the considered score s shown as a vertical line. The next three colums show histograms of the 5000 $LLR_{20,1000}(s)$ values computed by each method using the training sets $tr_1, tr_2, .., tr_{5000}$ where size of $tr_i = (n^p, n^d) = (20, 1000)$ and generated using random sampling from the corresponding pairs of PDFs.

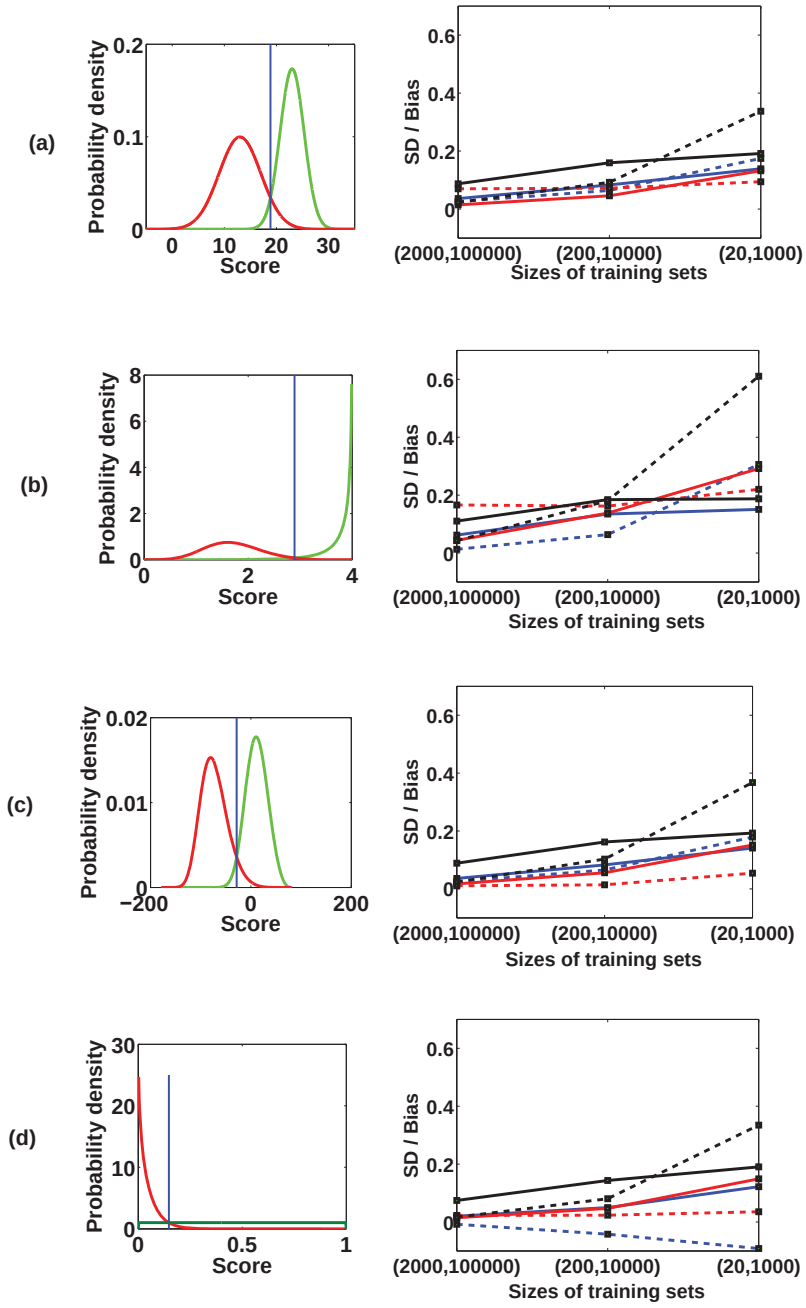


Fig. 3.13: The leftmost column shows the PDFs with the considered score value shown as a vertical line. The next colum shows the Standard Deviation (SD) and bias of each method for the three different sizes of the training sets.

Table 3.4: The mean and the interval between the maximum (Max) and the minimum (Min) LLRs for the three different sizes of the training sets. For each size, the mean LLR closest to the $LLR_{\infty,\infty}$ and the smallest interval is highlighted.

(a) For PDFs in Fig.3.11-3.13(a)

	KDE		Log Reg		PAV	
	Mean	[Min,Max]	Mean	[Min,Max]	Mean	[Min,Max]
(2000,100000)	0.0258	[-0.1172,0.1699]	0.0700	**[0.0101,0.1204]**	**0.0229**	[-0.3306,0.2976]
(200,10000)	**0.0643**	[-0.2744,0.3041]	0.0718	**[-0.1471,0.2299]**	0.0922	[-0.5293,0.4899]
(20,1000)	0.1747	**[-0.3841,0.5735]**	**0.0942**	[-0.7927,0.4034]	0.3374	[-0.1730,0.9379]

(b) For PDFs in Fig.3.11-3.13(b)

	KDE		Log Reg		PAV	
	Mean	[Min,Max]	Mean	[Min,Max]	Mean	[Min,Max]
(2000,100000)	**0.0130**	[-0.2197,0.2328]	0.1661	**[-0.0110,0.3139]**	0.0426	[-0.4139,0.3461]
(200,10000)	**0.0636**	**[-0.5754,0.4315]**	0.1622	[-0.6047,0.4937]	0.1806	[-0.4399,0.7116]
(20,1000)	0.3062	**[-0.1645,0.7999]**	**0.2200**	[-0.6569,0.9012]	0.6109	[0.1219,1.3260]

(c) For PDFs in Fig.3.11-3.13(c)

	KDE		Log Reg		PAV	
	Mean	[Min,Max]	Mean	[Min,Max]	Mean	[Min,Max]
(2000,100000)	0.0254	[-0.1017,0.1442]	**0.0109**	**[-0.0515,0.0715]**	0.0240	[-0.3162,0.2819]
(200,10000)	0.0653	[-0.3135,0.3048]	**0.0140**	**[-0.2543,0.1829]**	0.1030	[-0.6170,0.5829]
(20,1000)	0.1790	**[-0.3659,0.6237]**	**0.0542**	[-0.9233,0.4088]	0.3674	[-0.1413,0.9478]

(d) For PDFs in Fig.3.11-3.13(d)

	KDE		Log Reg		PAV	
	Mean	[Min,Max]	Mean	[Min,Max]	Mean	[Min,Max]
(2000,100000)	**-0.0078**	[-0.1014,0.0599]	0.0228	**[-0.0312,0.0710]**	0.0166	[-0.2779,0.2368]
(200,10000)	-0.0423	[-0.2485,0.1296]	**0.0234**	**[-0.1688,0.1769]**	0.0803	[-0.5180,0.4731]
(20,1000)	-0.0915	**[-0.8170,0.3097]**	**0.0353**	[-0.8341,0.3626]	0.3347	[-0.1543,0.9792]

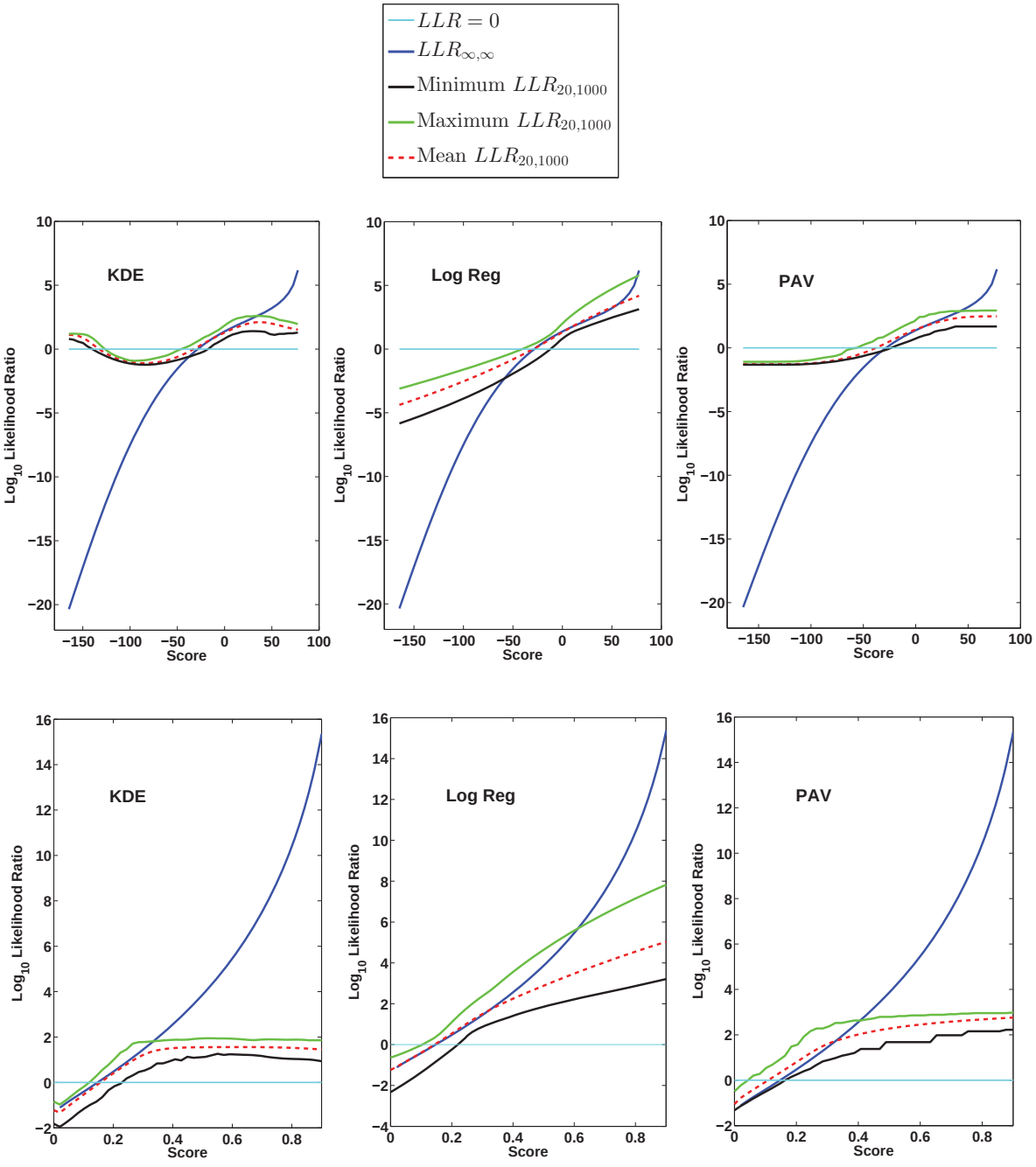


Fig. 3.14: The mean, maximum and minimum LLRs computed from the set of 5000 LLRs of each of the score in the set of 50 equidistant scores.

far from the intersecting point of the two PDFs. In case of the KDE method, the function from scores to LLRs may not be monotonically increasing. This is a very undesirable property in evidence evaluation where the score increases with the degree of similarity between the two biometric specimens. A specific regularization procedure such as imposing a condition on s as "$min(s_p) \leq s \leq max(s_d)$ and mapping scores outside this range to the end-points of this range" is needed in the KDE approach. There is a large sampling variability in Log Reg approach which strictly warns its use when small sizes of the training sets are available. Assuming in a forensic case when a LR is computed and sampling variability is not assessed, this leaves the PAV method as the most attractive choice since it LRs are monotonically increasing over the whole range of scores (unlike the KDE method) and has smaller sampling variability than the Log Reg method.

3.3.7 Conclusions and future work

In this paper we studied three LR computation methods commonly proposed for forensic evidence evaluation. The focus was to understand the effect of the sampling variability in the training sets of scores and the accuracy. A simulation framework is proposed for this purpose and a comparative study is carried out based on different shapes of the PDFs from which the training sets are generated, the sizes of the training sets and the value of the score for which the LR is computed. It is observed that these three parameters are important and should be considered in order to appropriately select a method of LR computation since all of them affect the computed LR. It is shown that sampling variability is a serious concern when the sizes of the training data sets are small. Furthermore, for small sizes of the training sets, the whole range of scores is mapped to LLRs and it is found that the two commonly used methods, KDE and Log Reg, give undesirable results. The study suggests that a range of LRs should be reported which incorporates the sampling variability.

For future work, research can be carried out in order to derive the functional relationship between the sizes of the training sets and the amount of sampling variability in the computed LR. Furthermore, specific strategies can be incorporated in the LR computation methods to make them more robust for cases where small training sets are available for LR computation.

Chapter 4

Suspect-specific and generic training scores for computation of LRs

4.1 Introduction

The two sets of training scores s_p and s_d can be suspect-specific (also referred as subject-specific or suspect-anchored) or they can be generic (also referred to as suspect-independent or subject-independent). This chapter studies the effect on resultant LRs when suspect-independent training scores sets are used instead of suspect-dependent. This is useful since in cases where enough biometric data sets cannot be collected from a suspect and the suspect-independent approach is followed, it can give an estimation of the differece in the resultant LRs. In section 4.2, we consider only a face recognition system. The Pool Adjacent Violators method is used for LR computation. In section 4.3, the work is extended to two other biometric modalities as well: speaker and fingerprint recognition. Same protocol is used for collection of training scores in each biometric modality. This enables us to study the effect of different training biometric data sets and the robustness of these biometric modalities to this difference in the training scores sets. A quantitative analysis of how frequently the two LRs fall in a same range is carried out. The ranges are based on the verbal equivalents represented by the LRs.

4.2 Effect of calibration data on forensic likelihood ratio from a face recognition system [1]

4.2.1 abstract

A biometric system used for forensic evaluation requires a conversion of the score to a likelihood ratio. A likelihood ratio can be computed as the ratio of the probability of a score given the prosecution hypothesis is true and the probability of a score given the defense hypothesis is true. In this paper we study two different approaches of a forensic likelihood ratio computation in the context of forensic face recognition. These approaches differ in the databases they use to obtain the score distribution under the prosecution and the defense hypothesis and therefore consider slightly different interpretation of these hypotheses. The goal of this study is to quantify the effect of these approaches on the resultant likelihood ratio in the context of evidence evaluation from a face recognition system. A state-of-the art commercial face recognition system is employed for facial images comparison and computation of scores. A simple forensic case is simulated by randomly selecting a small subset from the FRGC database. Images in this subset are used to estimate the score distribution under the prosecution and the defense hypothesis and the effect of different approaches of a likelihood ratio computation is demonstrated and explained. It is observed that there is a significant variation in the resultant likelihood ratios given the databases which are used to model the prosecution and defense hypothesis are varied.

4.2.2 Introduction

A score obtained from a face recognition system quantifies the similarity between the pair of input images while taking into account their typicality. In biometric applications such as access control to a building and e-passport gates at some airports, we choose a threshold from the range of the score and consequently any score above the threshold implies a positive decision and vice versa [1]. However, in a criminal case, there are facial images from a crime scene, e.g., facial recordings from a surveillance camera as well as images from the suspect. The responsibility of a forensic scientist is to give a Likelihood

[1] The contents of this section are published in [15] "Effect of calibration data on forensic likelihood ratio from a face recognition system", In: IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems, BTAS 2013, Washington, DC, U.S.A.. pp. 1-8, IEEE explore digital library, ISBN 978-1-4799-0527-0

Ratio (LR) instead of a decision to the court [85]. Then, it is the responsibility of the judge or the jury to make a decision which involves other sources of information about the case at hand. The use of a LR to report the output of a biometric comparison is gradually becoming a standard way of evidence evaluation from score-based biometric systems. A LR is a more objective and useful output in forensic evaluation than simply a score [52]. A general description of the LR framework for evidence evaluation from biometric systems can be found in [52,63]. It is applied to several biometric modalities including forensic voice [53], speaker [3,32,67] and fingerprint comparison [68]. Preliminary results of evidence evaluation using this framework in the context of face and handwriting recognition systems are presented in [10,23,24]. A LR is the probability of the score given the prosecution hypothesis is true divided by the probability of the score given the defense hypothesis is true:

$$LR(s) = \frac{P(s|H_p)}{P(s|H_d)} \tag{4.1}$$

where s, considered as the evidence, is the score obtained by comparison of the image from the suspect with the image found at the crime scene. H_p and H_d are two mutually exclusive and exhaustive source-level hypotheses defined as follows:

H_p: The pair of input images that produced score s originated from the same source.
H_d: The pair of input images that produced score s originated from different sources.

The LR computes a conditional probability of observing a particular value of the evidence s with respect to H_p and H_d. It is a concept which provides for evaluation and comparison of the two hypotheses concerning the likely source of the trace image found at the crime scene. Once a forensic scientist has computed the LR, it can be interpreted as a multiplicative factor which updates the prior odds (before observing the evidence from a biometric system) to the posterior odds (after observing the evidence from a biometric system) using the Bayesian framework:

$$\frac{P(H_p|s)}{P(H_d|s)} = \frac{P(s|H_p)}{P(s|H_d)} \times \frac{P(H_p)}{P(H_d)} \tag{4.2}$$

In this framework, the judge or the jury is responsible for quantification of the

prior beliefs about H_p and H_d while the forensic scientist is responsible for the quantitative evaluation of the evidence s in the form of a LR.

The hypotheses H_p and H_d can be specifically interpreted in a slightly different way so that the same-source and different-sources condition is linked with the specific suspect in a forensic case. These different interpretations correspond to differences in the pairs of images used to obtain the score distribution under H_p and H_d. In forensic evaluation, these pairs of images are called calibration data. The purpose of this paper is to quantify the evidential value from facial images using the likelihood ratio framework and to quantify the effect of different calibration data used to obtain the score distribution under the prosecution and defense hypothesis [86]. The study is carried out in the specific context of face recognition, however, the concepts and procedure described apply to any biometric system which computes a score for an input pair of samples.

The paper is organised as follows. In section 4.2.3 we briefly review the LR computation process from biometric scores and discuss the employed score-to-LR conversion method. Section 4.2.4 discusses the two different approaches and the differences in the interpretation of the hypotheses it implies. Section 4.2.5 reviews some existing work which studies the effects of different calibration data on LRs in the context of forensic speaker and handwriting recognition and presents the comparison procedures. Section 4.2.6 explains experimental setup and the degradation process to obtain trace-like images. Section 4.2.7 presents results by mapping score-axis to Log_{10} LR (LLR) using the two different approaches of LR computation. Finally section 4.2.8 draws conclusions and points toward future research directions.

4.2.3 Computation of a LR

4.2.3.1 Computation of calibration scores

Score-based biometric systems output two classes of scores. The first one is the result of the comparison of two samples produced by the same source. When comparing a set of samples produced by the same source, there is some variation in the score values output by a biometric system. Each modality has different nature of variations in the samples produced by the same source, for example, in case of face recognition systems it is caused by lighting condition, facial expressions and partial occlusion of the face, etc. A set of scores obtained by comparing samples from the same source represent the within-source

variability of the score and is referred to as the within-source scores. Similarly, comparing a set of samples produced by different sources results in a set of scores that represent the between-source variability of the score and is referred to as the between-source scores (see Fig.4.1). Scores in the within-source and in the between-source sets are collectively called calibration scores where the pair of samples to obtain these calibration scores are referred to as calibration data.

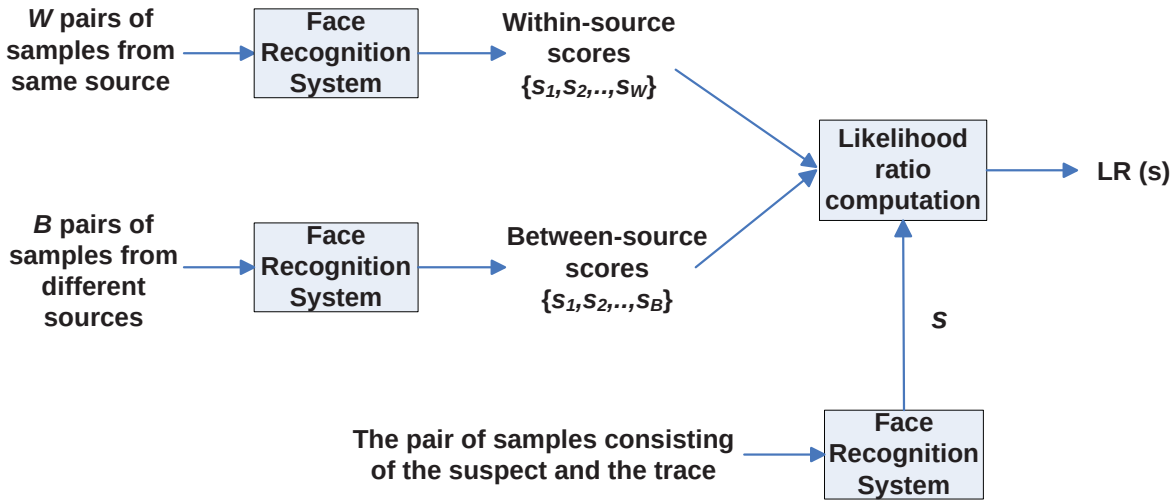


Fig. 4.1: Computation of a score-based LR

In order to compare a pair of input images and obtain a score, we use a state-of-the-art commercial face recognition system [83]. This system computes a score value in the range of 0 and 1.

4.2.3.2 Mapping score-axis to likelihood ratios using calibration scores

Score-based LR computation can be considered as a mapping function from score to LR. Given a set of calibration scores, there are several methods to map the score-axis to Likelihood Ratios (LRs). These methods as described in [51] can be classified as parametric or non-parametric. When the distribution of the within-source and the between-source scores sets are similar to a standard probability density function (pdf), the pdf of score under H_p and H_d can be estimated by fitting standard pdfs with certain parameters using maximum likelihood estimation to these sets of scores [68]. Another possible parametric approach is to estimate the ratio of the pdf under H_p and pdf under H_d using logistic regression [78]. In nonparametric category, there are

histogram binning, Kernel Density Estimation (KDE) and finding slope of Receiver Operating Characteristic Convex Hull (ROCCH) [61]. For the scores obtained from the face recognition system in this study, we propose the use of ROCCH procedure due to the variability in the distribution of scores obtained and the mismatch of the distribution of scores to any standard family of pdf. This approach is preferred because it can ensure that the resultant variation in LRs are due to the differences in the within-source and between-source sets and not due to the poor fitting of the densities to the calibration score sets.

Once ROCCH is computed for a given set of calibration scores, LR for a given score is the slope of the corresponding segment of the ROCCH and can be computed as follows:

$$LR(s) = \frac{W}{B} \times \frac{P(H_d)}{P(H_p)} \tag{4.3}$$

where W and B are the number of the within-source and the between-source scores respectively in the corresponding segment of the ROCCH on which score s lies. The value $\frac{P(H_d)}{P(H_p)}$ is computed from the size of the within-source and the between-source sets. It is interesting to note that computing ROCCH is equivalent to computing ROC of the posterior probabilities obtained by Pool Adjacent Violators (PAV) algorithm [61]. This argument leads to an alternative way of implementation; computing posterior probabilities using PAV and then plugging it into Bayesian formula along with $\frac{P(H_d)}{P(H_p)}$ to compute LRs.

Once the ROCCH procedure is applied to compute LRs, there are a group of scores for which the posterior probabilities are either 0 or 1. The log odds of these posterior probabities results in minus infinities and plus infinities respectively. To avoid this problem, a procedure similar to [17] is followed. We replace a score in the between-source set by the maximum score in the within-source set and a score in the within-source set by the minimum score of the between-source set. These replaced scores can be considered to represent scores which were not encountered in the calibration scores because there is not enough calibration data, but which could have occurred.

4.2.4 Suspect-anchored and suspect-independent calibration data

Based on the quantity of the available data from the suspect, the within-source and the between-source scores can be either anchored to the suspect or it can be general within-source and between-source matches using all suspects from the potential population database.

Suspect-anchored approach

To compute the within-source scores, a set of images from the suspect can be compared with another set of images from the suspect. This is referred to as suspect-anchored approach [14]. The two sets of images are referred to as suspect-reference and suspect-control data sets. For better calibration, the images in the suspect-control data set should be as close as possible to the trace and the images in the suspect-reference data set should be as close as possible to the potential population database. Cross comparison of all images in the suspect-reference and suspect-control data set results in a set of scores that can be used to model the distribution of scores under the prosecution hypothesis. Similarly, for modelling the distribution of scores under the defense hypothesis, images in the suspect-control data set are compared with images in the suspect-reference data set of all the suspects in the potential population [87]. The suspect-anchored approach implies considering the following interpretations of the prosecution and defense hypotheses:

- H_p: The score s arises from the distribution of scores obtained by pairing suspect images in the suspect-control data set with the suspect images in the suspect-reference data set.
- H_d: The score s arises from the distribution of scores obtained by pairing suspect images in the suspect-control data set with the reference images in the potential population database.

The difficulty in following the suspect-anchored approach is that in most cases it might not be possible to obtain a large set of data from the suspect in similar conditions to the trace. The lack of enough calibration data increases the uncertainty in the resultant estimate of a LR.

Suspect-independent approach

Certain specific solutions have appeared as how to increase the number of the within-source and the between-source scores to estimate the distribution of scores under the prosecution and defense hypothesis [14, 77]. A general solution to this problem is to compute the within-source scores by considering pairs of images from multiple potential suspects in the potential population database. Excluding images of the original suspect, images in the suspect-control and suspect-reference database of all the other suspects are paired with each other where each pair originate from the same source. This is referred to as suspect-independent approach. Similarly to obtain the suspect-independent between-source scores, reference images in the potential population are paired with the suspect-control data set of each suspect where each pair originate from different sources [3,14]. Using suspect-independent approach to LR computation implies the following interpretations of the prosecution and defense hypotheses:

- H_p: Score s arises from the distribution of scores obtained by pairing images in the suspect-control and suspect-reference data set of all the suspects in potential population where the paired images are obtained from the same source.
- H_d: Score s arises from the distribution of scores obtained by pairing images in the suspect-control and suspect-reference data set of all the suspects in potential population where the paired images are obtained from different sources.

For the between-source scores, besides the suspect-anchored and suspect-independent approaches, another commonly used approach is to compute trace-anchored scores. In this approach, trace image is compared with all the reference images of the potential population to compute the between-source scores [4].

4.2.5 Comparing the resultant LRs

Since it is preferable to compute a suspect-anchored LR, there is some research on how to compute a LR for a biometric comparison when there is a limited calibration data available from the suspect. Ramos [14] proposed a strategy which is based on the adaptation of the suspect-independent within-source score distribution to the suspect-anchored scores via Maximum A Posteriori (MAP) estimation. Similarly, in forensic handwriting recognition, Davis [77] gener-

ated simulated writing samples from a small set of suspect samples to form a database for computation of the suspect-anchored within-source scores. These specific approaches do not generalize in most cases and usually a suspect-independent approach is considered as a last resort to compute a reliable LR for the evidence s [88]. In [88] suspect-independent approach is proposed as a feasible alternative when a single sample is available from the suspect.
Given the common use of the suspect-independent approach as an alternative to the suspect-anchored approach, it is important to study and analyse the differences in the score-to-LR functions produced by these approaches. Quantifying the variability between the suspect-anchored and suspect-independent approach is still under investigation in most of biometric modalities including face recognition. Ramos [14] studied the effect of using suspect-independent within-source scores instead of suspect-anchored approach on the resultant LRs in the context of speaker recognition. [77] describes the effect(s) of different calibrations data used to construct the denominator distribution in the context of handwriting recognition.
We compute two functions from score to LLR using the suspect-anchored and suspect-independent sets of calibration scores. The behavior of these two score-to-LLR functions is studied in different regions of the score-axis which correponds to different evidence values. For a more quantitative evaluation, we generate 100 evidence values by uniformly sampling the score-axis and compute the number of cases in which these two approaches agree and disagree on a given range of LRs. A disagreement is reported when one approach produces a LR that falls into a different range. These ranges correspond to verbal equivalents which can be used in certain situations to report the forensic evaluation of the evidence. These ranges along with their corresponding verbal equivalents are shown in table 2 [85].

4.2.6 Experimental setup

To simulate a forensic case, we randomly select a small subset of five subjects from the FRGC [89] database. Each subject has 36 frontal images taken in different illumination condition. For each subject, half of the images are used to create the suspect-control database while the remaining half are used as a suspect-reference database. Images in the suspect-control database are degraded by adding motion blur of 15 pixels with zero angle and downsampling them by half of the original resolution. Face regions are manually cropped where eye detection is performed automatically by the face recognition system [83]. Figure 4.2 shows an example of the degradation applied to an image for

creation of the suspect-control data set. The goal of the degradation process is to make the images in the suspect-control data set similar to a trace image.

Fig. 4.2: An example of the degradation process applied to obtain suspect-control data set.

Figure 4.3 illustrates computation of the within-source and the between-source sets in each approach for a single image per subject assuming subject 1 is the suspect.

Table 4.1 shows the number of unique comparisons (and hence the number of scores) in each appraoch of the within-source and the between-source scores sets computation given 5 subjects and 18 images per subject in the probe and gallery set.

within-source scores	
Suspect-anchored	18 X 18 = 324
Suspect-independent	4 X 324 = 1296
between-source scores	
Suspect-anchored	18 X (18 X 4) = 1296
Suspect-independent	4 X 1296 = 5184

Table 4.1: Number of scores in the set of the within-source and the between-source scores.

4.2.7 Experimental results

The score-axis is mapped to LLR using suspect-specific as well as suspect-independent approach in order to compare their score-to-LLR mapping functions. Figure 4.4 shows the frequency histograms of the scores in the within-

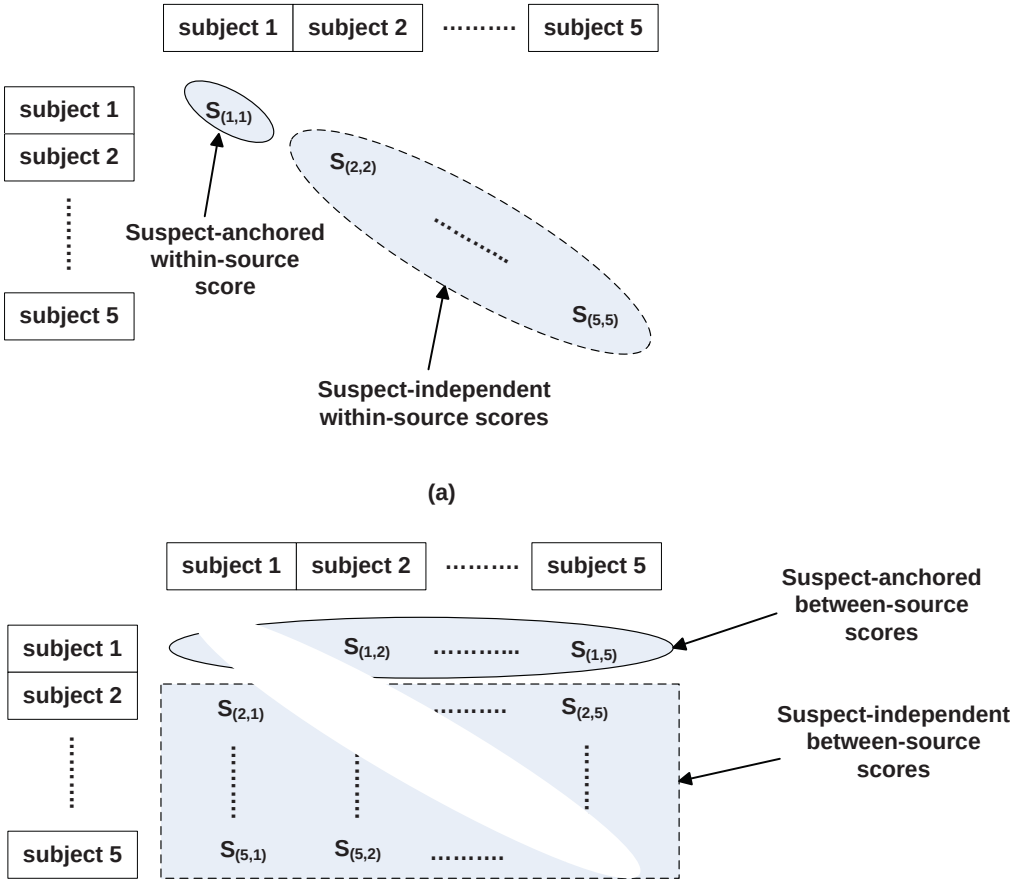


Fig. 4.3: The within-source and the between-source scores sets assuming the first subject as the suspect and 1 image per subject. a) Computation of the within-source scores sets b) Computation of the between-source scores sets.

source and the between-source sets, the ROCs of the calibration scores in the suspect-anchored and suspect-independent approach and the score-to-LLR functions obtained by the ROCCH procedure as described in section 4.2.3.2. Note the greater variation in the within-source scores for the suspect-anchored approach. Images of each subject are selected in such a way so that the variation in illumination and facial expressions are as close as possible across different subjects. Other conditions such as resolution and face pose is the same across all the subjects. However, still we observe considerable variation in the shape of the suspect-anchored frequency histograms of the within-source scores. Variation in the histograms of the suspect-anchored within-source scores are caused by either the slight difference in the illumination conditions and facial expressions or due to the difference in identity. Illumination conditions and facial expressions are very similar across different subjects, however, they are not the same for all 36 images of each subject. Besides the slight difference in illumination condition and facial expression, identity itself has effect on the suspect-anchored within-source scores distribution. A face recognition algorithm may perform differently for different subjects when it is used to match images of the same subject. Generally, it is expected that the suspect-anchored approach produces scores and subsequent LRs which are more discriminative as in the case of the first three subjects. However, this is not true in case of subject 4 and 5 where the suspect-anchored within-source scores have more standard deviation than the suspect-independent within-source scores.

There is a significant variation in the values of the LRs computed using the suspect-anchored and suspect-independent approach. For example in the case of subject 1, at the score location of 0.44, the suspect-anchored and the suspect-independent LR is 1052 and 78 respectively. The horizontal lines in the mapping functions are due to the proposed strategy to avoid infinite LLRs. Values of LRs along this horizontal line are referred to as saturated LRs. These LRs, as illustrated later, depend on the size of the set of calibration scores used to map score-axis to LRs. In this region of LRs, the suspect-independent approach results in higher values of LRs than the suspect-anchored approach. Given the fact that LRs in the higher ranges are more important and useful in practice than in the lower and in the middle ranges, it can be argued that, in practice, the anchoring plays a crucial role. Note that the suspect-independent mapping functions across different subjects can be considered to reflect the possible variation in likelihood ratios when a different suspect is considered keeping the calibration data constant.

In most cases the exact numerical value of a likelihood ratio is of less importance than the range in which it lies. These ranges can be taken into

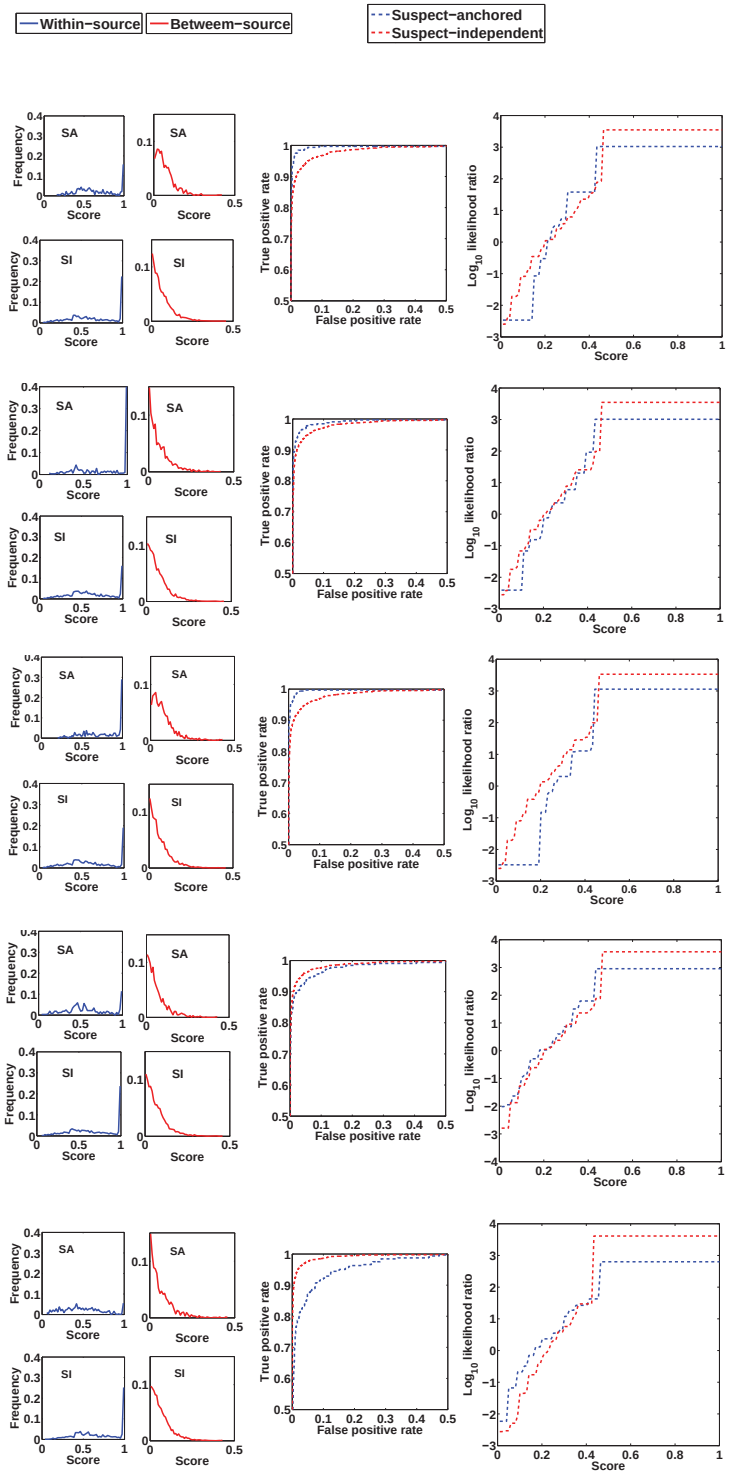


Fig. 4.4: The first two columns show the frequency histograms of the suspect-anchored (SA) and suspect-independent (SI) within-source and between-source scores sets. The third columns plots the ROCs from the corresponding sets of the within-source and between-source scores. Last column shows the mapping function from score to LLR using the ROCCH procedure. Row 1 through 5 repeat the same experiment considering each of the 5 subjects in the selected subset as the suspect.

consideration when performing such a comparative study. The score-axis is uniformly sampled to simulate 100 values of evidence s. These scores are converted to LLRs using each of the LR computation approach. Table 4.2 shows the number of cases in which the two approaches compute LLRs which fall into the same range. As seen from table 4.2, in 296 cases out of 500, the two LRs agree on same verbal equivalents resulting in 59.2% agreement rate. A considerable difference results from the fact that for subject 4 and subject 5, the saturated LRs are in different ranges.

Ranges	Verbal equivalents	Number of agreements					
		P1	P2	P3	P4	P5	Total
$4 <$ LLR	Very strong evidence to support H_p	0	0	0	0	0	0
$3 <$ LLR≤ 4	Strong evidence to support H_p	54	54	54	0	0	162
$2 <$ LLR≤ 3	Moderately strong evidence to support H_p	0	0	0	0	0	0
$1 <$ LLR≤ 2	Moderate evidence to support H_p	9	8	10	8	9	44
$0 <$ LLR≤ 1	Limited evidence to support H_p	9	10	6	13	7	45
$-1 <$ LLR≤ 0	Limited evidence to support H_d	2	6	0	4	3	15
$-2 <$ LLR≤ -1	Moderate evidence to support H_d	0	2	0	5	0	7
$-3 <$ LLR≤ -2	Moderately strong evidence to support H_d	4	4	4	2	4	18
$-4 <$ LLR≤ -3	Strong evidence to support H_d	0	0	0	0	0	0
LLR < -4	Very strong evidence to support H_d	1	1	1	1	1	5
	Total	79	85	75	33	24	296

Table 4.2: Number of times in which the LRs computed by the two approaches falls into same ranges. For each subject considered as the suspect, there are 100 values of s generated by uniformly sampling the score-axis. Out of a total of 500 LRs computed by the two approaches, 296 times the LRs agree on one range of LLRs.

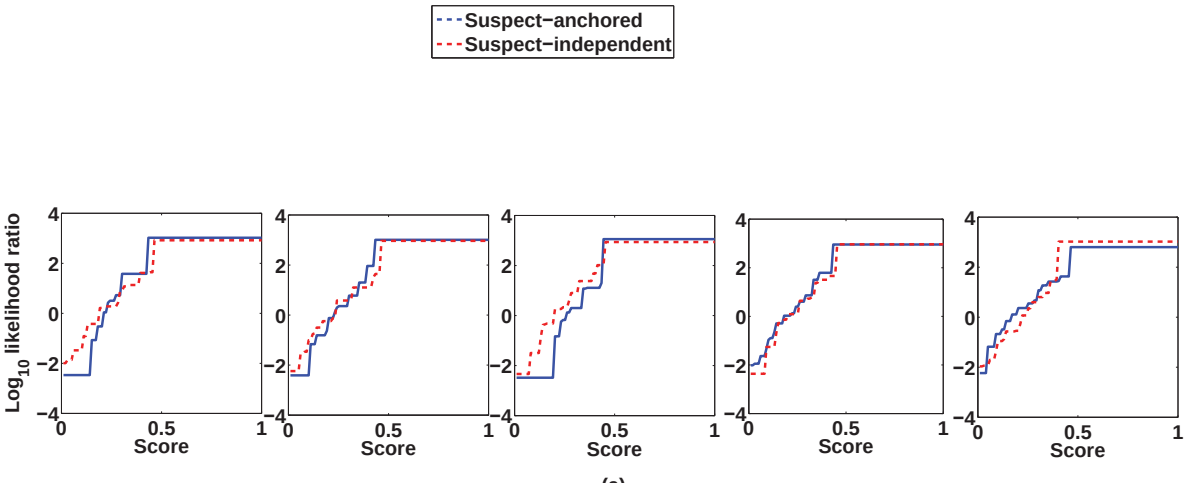


Fig. 4.5: Score-axis is mapped to LLRs using the same sizes of the within-source and the between-source sets in the suspect-anchored and suspect-independent approach.

Note that using a different face recognition system to compute scores and using

a different method to map score-axis to LRs might lead to completely different results. Similarly a different database of images might also influence the variations between the suspect-anchored and suspect-independent approach of LR computation.

The effect of the difference in the size of the calibration sets between the suspect-anchored and suspect-independent approach can be investigated by randomly sampling without replacement a number of scores equal to the size of the suspect-anchored sets from the suspect-independent sets. Given the size of the within-source and between-source sets in the suspect-anchored and suspect-independent approach is the same, the variation in LRs is only caused by the nature of the distributions of the scores. Figure 4.5 shows the mapping function obtained by the two approaches when the within-source and the between-source sets are equally sized by random subsampling the suspect-independent within-source and between-source sets so that the sizes of these sets in the suspect-independent approach is equal to those in the suspect-anchored approach. Note that reduction in the size of the calibration scores reduces the range of LRs that can be computed. This can be seen by comparing the mapping functions of the suspect-independent approach in figure 4 and figure 5. Besides the saturated region of LRs, the difference in the size of the calibration sets has less effect on the resultant mapping function from score to LLRs.

4.2.8 Conclusions and future work

We discussed the effect of different calibration data on the resultant forensic LR in the context of face recognition. The process of conversion of a score, obtained from the comparison of the crime scene image with the suspect image, to a forensic LR is described. It is observed that there is a significant variation between the LRs obtained using suspect-anchored and suspect-independent approach. The differences are more prominent in the higher ranges of LRs and therefore more caution should be taken if one approach is used as an altenative to the other. Future work will include quantifying the influence of images from other databases, different face recognition systems and other score-to-LR conversion methods. Furthermore, it is also of interest to study and compare these results from other modalities such as speech.

4.3 Biometric evidence evaluation: an empirical assessment of the effect of different training data [2]

4.3.1 Abstract

For an automatic comparison of a pair of biometric specimens, a similarity metric called *score* is computed by the employed biometric recognition system. In forensic evaluation, it is desirable to convert this score into a likelihood ratio. This process is referred to as calibration. A likelihood ratio is the probability of the score given the prosecution hypothesis (which states that the pair of biometric specimens are obtained from the same source) is true divided by the probability of the score given the defense hypothesis (which states that the pair of biometric specimens are obtained from different sources) is true. In practice, a set of scores (called training scores) obtained from the same-source and different-sources comparisons is needed to compute a likelihood ratio value for a score. In likelihood ratio computation, the same-source and different-sources conditions can be anchored to a specific suspect in a forensic case or it can be generic same-source and different-sources comparisons independent of the suspect involved in the case. This results in two likelihood ratio values which differ in the nature of training scores they use and therefore consider slightly different interpretations of the two hypotheses. The goal of this study is to quantify the differences in these two likelihood ratio values in the context of evidence evaluation from a face, a fingerprint and a speaker recognition system. For each biometric modality, a simple forensic case is simulated by randomly selecting a small subset of biometric specimens from a large database. In order to be able to carry out a comparison across the three biometric modalities, the same protocol is followed for training scores set generation. It is observed that there is a significant variation in the two likelihood ratio values.

4.3.2 Introduction

For a given pair of biometric specimens, a score computed by a biometric recognition system quantifies the similarity between the input pair of biometric specimens while taking into account their typicality. In biometric applications

[2] The contents of this section are based on "Biometric evidence evaluation: an empirical assessment of the effect of different training data", IET Biometrics (under review).

such as access control to a building and e-passport gates at some airports, the developer of the system choose a threshold from the range of the score and consequently any score above the threshold implies a positive decision and vice versa [1, 90]. However, in a criminal case, it is desirable to report a Likelihood Ratio (LR) instead of a score or a decision based on a selected threshold [85]. This distinction between biometric and forensic applications is addressed in detail recently by Meuwly [90]. Once a forensic scientist has computed the LR, it is the responsibility of the judge or the jury to make a decision which involves other sources of information about the case at hand such as other types of evidences. Use of a LR value to report the output of a biometric comparison is gradually becoming a standard way of evidence evaluation from score-based biometric systems. A LR is a more informative, balanced and useful output in forensic evaluation than simply a score [52]. A general description of the LR concept for evidence evaluation from biometric systems can be found in [52, 63]. It is applied to several biometric modalities including forensic voice [53], speech [3, 7, 67] and fingerprint comparison [68]. Preliminary results of evidence evaluation using a LR value in the context of face and handwriting recognition systems are presented in [10, 23, 54]. A LR is the probability of the score given the prosecution hypothesis is true divided by the probability of the score given the defense hypothesis is true:

$$LR(s) = \frac{P(s|H_p, I)}{P(s|H_d, I)}, \tag{4.4}$$

where s, considered as the evidence, is the score obtained by comparison of the biometric specimen from the suspect with that found at the crime scene. I refers to background information which may or may not be domain specific. H_p and H_d are two mutually exclusive and exhaustive source-level hypotheses defined as follows:

H_p: The pair of biometric specimens are originated from a same source.
H_d: The pair of biometric specimens are originated from different sources.

Once a forensic scientist has computed the LR value, one way to interpret it is as a multiplicative factor which updates the prior odds (before observing the evidence from a biometric system) to the posterior odds (after observing the evidence from a biometric system) using the Bayesian probabilistic framework:

$$\frac{P(H_p|s, I)}{P(H_d|s, I)} = \frac{P(s|H_p, I)}{P(s|H_d, I)} \times \frac{P(H_p|I)}{P(H_d|I)}. \tag{4.5}$$

In this framework, the judge or the jury is responsible for quantification of the

prior beliefs about H_p and H_d while the forensic scientist is responsible for the scientific analysis of the pair of biometric specimens and quantification of its evidential value in the form of a LR.

The hypotheses H_p and H_d can be specifically interpreted in a slightly different way so that the same-source and different-sources conditions are linked with the specific suspect in a forensic case. These different interpretations correspond to difference in the pairs of biometric specimens used to obtain the distribution of scores under H_p and H_d. In forensic evaluation, these pairs of biometric specimens are called training data (or calibration data). The purpose of this paper is to quantify the evidential value from a face, a fingerprint and a speaker recognition system using the likelihood ratio concept and to study the effect of different training data on resultant likelihood ratio values using a simple simulated forensic LR evaluation scenario. The study is carried out for a single biometric recognition system from each of the three biometric modalities; however, the concepts and proecedure described to obtain the training data and compute LR values apply to any biometric system which computes a score for an input pair of biometric specimens.

The paper is organised as follows. In section 4.3.3 we briefly review the procedure of LR computation from scores and discuss the employed score-to-LR computation method. Section 4.3.4 discusses the two different approaches for selection of the training data and the differences in the interpretation of the hypotheses they imply. Section 4.3.5 reviews existing work which studies the effects of different training data on LR values and presents the comparison procedures followed in this paper. Section 4.3.6 explains experimental setup by introducing the three biometric recognition systems, databases of biometric specimens and the way the training scores sets are constructed. Section 4.3.7 presents results by mapping score values to Log_{10} LR (LLR) values using the two different sets of the training scores. Finally section 4.3.8 draws conclusions and points toward future research directions.

4.3.3 Computation of a LR from a score

4.3.3.1 Computation of training scores

Score-based biometric systems output two classes of scores. The first one is the result of the comparison of two biometric specimens produced by the same source. When comparing a set of biometric specimens produced by the same source, there is some variation in the score values output by a biometric

system. Each biometric modality has different nature of variations in the biometric specimens produced by the same source, for example, in case of face recognition systems it is caused by lighting condition, facial expressions, partial occlusion of the face, etc. A set of scores obtained by comparing biometric specimens from the same source represent the within-source variability of the score and is referred to as the within-source scores. Similarly, comparing a set of biometric specimens produced by different sources results in a set of scores that represent the between-source variability of the score and is referred to as the between-source scores [3] (see Fig.4.6). Scores in the within-source and in the between-source sets are collectively called training scores where the pair of biometric specimens to obtain these training scores are referred to as training data.

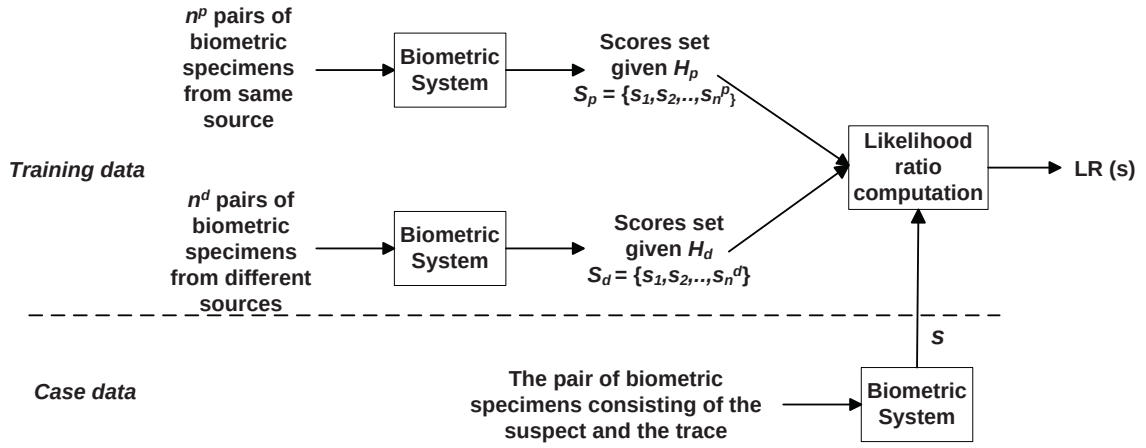


Fig. 4.6: Computation of a score-based LR for a given pair of biometric specimens consisting of the trace biometric specimen and the suspect biometric specimen. The same biometric system must be used to compute the within-source scores, the between-source scores and the evidence score s.

4.3.3.2 Mapping a score to a LR using training scores

Score-based LR computation can be considered as a mapping function from score to LR. Given a set of training scores, there are several methods to map the score-axis to LLR-axis. LR values in logarithmic scale are preferred for plotting purposes as well as it has intuitive apeal for forensic practitioners. As an example, a LLR value of 1 can be interpreted as "It is 10 times more likely that the two biometric specimens are originated from a same source than if they were originated from different sources". Similarly, a LLR value of -1 can

[3] Researchers in different biometric modalities use different terminologies for the within-source and between-source scores such as 'genuine and impostor scores' and 'same-source and different-sources scores'.

be interpreted as "It is 10 times more likely that the two biometric specimens are originated from different sources than if they were originated from a same source". Several methods of conversion of biometric scores to LR values are described in [3, 51] and can be classified as parametric or non-parametric. When the distribution of the scores in the within-source and in the between-source sets are similar to a standard probability density function (PDF), the PDF of the scores under H_p and H_d can be estimated by fitting standard PDFs with certain parameters using maximum likelihood estimation to these sets of scores [68]. Another possible parametric approach is to estimate the ratio of the PDF under H_p and the PDF under H_d using logistic regression [78]. In non-parametric category, there are histogram binning, Kernel Density Estimation (KDE) and finding slope of the Receiver Operating Characteristic Convex Hull (ROCCH) [51, 61]. Logistic regression and ROCCH approaches have a desirable property: both of them produce a monotonically increasing function from score to LR values. For the purpose of this study, we propose the use of ROCCH procedure because it can ensure to a greater extent that the resultant variation in LR values are due to the difference in the training scores set and not due to the poor fitting of the PDFs or the logistic regression model to the training score sets.

Readers are referred to [61] for the algorithm to construct the ROCCH from a given set of training scores. Once the ROCCH is constructed, LR value for a given score s is the slope of the corresponding segment of the ROCCH on which score s lies and can be computed as follows:

$$LR(s) = \frac{w_s}{b_s} \times \frac{W}{B} \tag{4.6}$$

where w_s and b_s are the number of the within-source and the between-source scores respectively in the corresponding segment of the ROCCH on which score s lies. The value W and B are number of scores in the complete sets of the within-source and the between-source in the training scores set. It is interesting to note that computing ROCCH is equivalent to computing Receiver Operating Characteristic (ROC) curve of the posterior probabilities obtained by Pool Adjacent Violators (PAV) algorithm [61]. This argument leads to an alternative way of implementation; computing posterior probabilities using PAV and then plugging it into the Bayesian formula along with W and B to compute LR values. PAV algorithm (or equivalently, the ROCCH approach) is extensively used in forensic speech recognition for computation of LR values [17].

Once the ROCCH procedure is applied to compute LR values, there is a group of scores for which the LR value is either zero or infinity. The logarithm of these LR values results in minus infinities and plus infinities respectively. To avoid this problem, a procedure similar to [17] is followed. We insert a score in the between-source set which is equal to the maximum score in the within-source set and a score in the within-source which is equal to the minimum score of the between-source set. These replaced scores can be considered to represent scores which were not encountered in the training scores set because there is not enough training data, but which could have occurred.

4.3.4 Choice of the training data

Based on the available number of biometric specimens from the suspect, the within-source and the between-source conditions can either be anchored to the suspect or it can be general within-source and between-source comparisons using all persons from the potential population defined in a given forensic case.

4.3.4.1 Suspect-specific training data

To compute the suspect-specific within-source scores, a set of biometric specimens from the suspect can be compared with another set of biometric specimens from the suspect [3]. The two sets of biometric specimens are referred to as *reference* and *test* data sets. For better calibration, the biometric specimens in the test data set should be as close as possible to the trace and the biometric specimens in the reference data set should be as close as possible to the database of biometric specimens from the potential population. Cross-comparison of all the biometric specimens in the reference and the test data set results in a set of scores that can be used to model the distribution of scores under the prosecution hypothesis. Similarly, for modelling the distribution of scores under the defense hypothesis, biometric specimens in the test data set are compared with the reference biometric specimens of the potential population database [87]. The suspect-specific approach implies considering the following interpretations of the prosecution and defense hypotheses:

- H_p: The pair of biometric specimens is originated from the suspect.
- H_d: The pair of biometric specimens is not originated from the suspect (or alternatively, the pair of biometric specimens is originated from someone else in the potential population).

The difficulty in following the suspect-specific approach is that in most cases it may not be possible to obtain a set of biometric specimens from the suspect. Availability of fewer specimens from the suspect leads to fewer scores in the training scores set, particularly the within-source scores set.

4.3.4.2 Suspect-independent (generic) training data

Certain specific solutions have been proposed as how to increase the number of the within-source scores when following the suspect-specific approach [14, 77]. A general solution is to construct the within-source scores set by combining the suspect-specific within-source scores sets of multiple persons from the potential population database. Similarly, to obtain the suspect-independent between-source scores, suspect-specific between-source scores sets of multiple persons are combined. Using the suspect-independent approach to LR computation implies the following interpretations of the prosecution and defense hypotheses:

- H_p: The pair of biometric specimens is obtained from a same source.
- H_d: The pair of biometric specimens is obtained from different sources.

For computation of the between-source scores, besides the suspect-specific and suspect-independent approaches, another commonly used approach is to compute trace-anchored scores. In this approach, the trace biometric specimen is compared with all the reference biometric specimens of the potential population to compute the between-source scores [4].

4.3.5 Comparing the resultant LR values

4.3.5.1 Motivation

It is preferred to compute a suspect-specific LR because it takes into account more relevant information about the case at hand. Therefore, there is some research on how to compute a suspect-specific LR for a biometric comparison when there is limited training data available from the suspect. Ramos [14] proposed a strategy which is based on the adaptation of the suspect-independent within-source scores distribution to the suspect-specific scores via Maximum A Posteriori (MAP) estimation. Similarly, in forensic handwriting recognition, Davis [77] generated a large set of simulated writing specimens from a small set of suspect specimens to form a data set for computation of the suspect-specific within-source scores. These specific approaches do not generalize in most cases and usually a suspect-independent approach is considered as a last

resort to compute a reliable LR for the given pair of biometric specimens [88]. In [88], suspect-independent approach is proposed as a feasible alternative when a single specimen is available from the suspect. Given the common use of the suspect-independent approach as an alternative to the suspect-specific approach to compute a LR value, it is important to study and analyse the differences in the LR values produced by these two approaches.

4.3.5.2 Existing work

Quantifying the difference between the LR values using the suspect-specific and suspect-independent training data is still under investigation in most biometric modalities. In [77], authors describe the effect(s) of different training data used to construct the between-source scores set in the context of handwriting recognition. For fingermark evidence, Alberink et al. [13] recently discussed different theoritical possiblities of conditioning such as conditioning on specified fingers, fingerprints and fingermarks in order to compute the training scores set. They also studied the asymmetric conditioning in LR computation which is, however, subject to further debate. Similarly, Ramos et al. [14] studied the effect of using suspect-independent within-source scores instead of suspect-specific on the resultant LR values in the context of forensic speaker recognition.

4.3.5.3 Comparison approach

A common approach to compare systems producing forensic LR values is to compute a set of test LR values for a set of pairs of biometric specimens whose origin is known. Then the criterion is that a better system should result in a larger value of LR for a same-source pair of biometric specimens and a smaller value of LR for a different-sources pair of biometric specimens. Two common tools that compare systems (more precisely, sets of test LR values produced by systems) based on this criterion are Tippett plot [18] and *Cost of Log LR* (C_{llr}) [17]. Such a comparison approach is very useful in practice, however, the focus of this work is to study how close the two LR values are instead which LR value is preferred. Therefore, in this work, instead of following the traditional approach to compare forensic LR computation systems, we propose to study the whole mapping function from score to LLR values and for a given random score, observe how much the LR values differ in magnitude. We compute the functions from score to LLR using the suspect-specific and suspect-independent sets of training scores and sample them uniformly for a

quantitative analysis of the differences in the two LR values. The behavior of the two score-to-LLR functions is studied in different regions of LLR values. Furthermore, using random subsampling, the effect of the different sizes of the training scores set used in each approach is also investigated.

4.3.6 Experimental setup

Figure 4.7 illustrates computation of the within-source and the between-source scores in the suspect-specific and suspect-independent approach for a single specimen per person case assuming person 1 is the suspect.

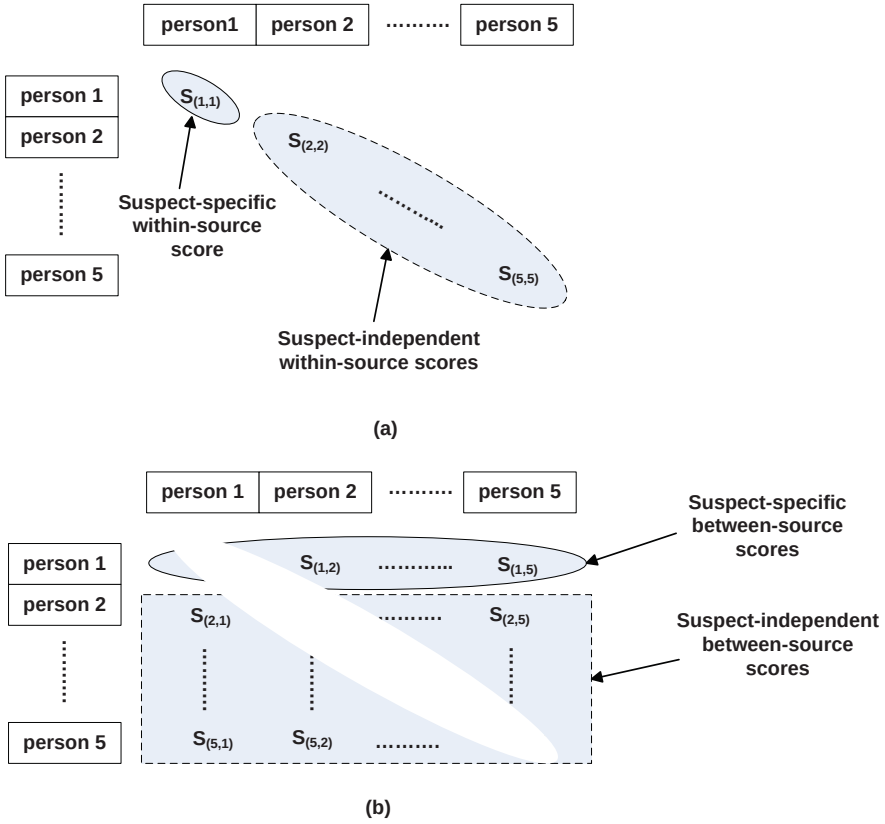


Fig. 4.7: The within-source and the between-source scores sets assuming the first person as the suspect and 1 biometric specimen per person. a) Computation of the within-source scores sets b) Computation of the between-source scores sets.

To simulate a forensic case, we randomly select five persons from a large database of biometric specimens in each of the biometric modality. For this purpose, Face Recognition Grand Challenge (FRGC) [89] database is used for face data, Dutch National Police Services Agency (KLPD) database is

used for fingerprint data and National Institute of Standards and Technology (NIST) 2010 Speaker Recognition Evaluation (SRE) extended database is used for speech data [82]. The important condition for setting up a rational experimental protocol for such a comparative study across the three different biometric modalities is to use equal number of scores per person in the within-source and between-source sets.

For face data, a set of 36 biometric specimens are chosen randomly for each of the 5 persons in the selected subset. Then half of the biometric specimens are used to generate the test data set while the remaining half are used as a reference data set. Cross-comparison of the 18 biometric specimens in the test data set and reference data set results in 324 scores in the suspect-specific within-source scores set. For face data, images used as test data set are degraded by adding motion blur of 15 pixels with zero angle and downsampling them by half of the original resolution. The goal of the degradation process is to make the images in the test data set similar to a trace image because the original face images in FRGC database are of very high resolution (231 X 251). The comparison is performed using a commercial face recognition system developed by Cognitec [83].

For fingerprint data, there is only one fingermark available which is used as the test data set. However, the number of reference fingerprints, obtained from the KLPD database, are very large. In order to have equal number of scores in the training scores set for each biometric modality, the size of the suspect-reference data set is increased to 324. Comparison of the one fingermark in the test data set with the 324 fingerprints in the reference data set results in 324 scores in the suspect-specific within-source scores. Motorola Biometric Identification System (BIS) software (version 9.1) is used for comparison of the fingermark to the reference fingerprints.

For speech data, similar to the face data, 18 specimens are used as the test data set and 18 specimens are used as the reference data set. The recognition algorithm is based on Probabilistic Linear Discriminant Analysis (PLDA) approach [62] which models the distribution of i-vectors as a multivariate Gaussian. The system is described in [62, section 2.5] in detail.

Table 4.3 shows the number of unique comparisons (and hence the number of scores) in each approach of the within-source and the between-source scores sets computation given there are 5 persons in the selected subset.

Beside studying the overall score-to-LLR functions for comparison, for a more quantitative analysis of the differences in LR values, we define score-axis as starting from the minimum value of the score in the suspect-independent

within-source scores	
Suspect-specific	324
Suspect-independent	4 X 324 = 1296
between-source scores	
Suspect-specific	4 X 324 = 1296
Suspect-independent	4 X 1296 = 5184

Table 4.3: Number of scores in the set of the within-source and the between-source scores.

between-source scores set and ending at the maximum value of the scores in the suspect-independent within-source scores set. Then we generate 100 evidence scores by uniformly sampling the score-axis and compute the number of cases in which the two LR values agree and disagree on a given range of LR. A disagreement is reported when one approach produces a LR that falls into a different range. These ranges correspond to different verbal equivalents of the numerical LR values which can be used in certain situations to report the forensic evaluation of the evidence. These ranges along with their corresponding verbal equivalnets are shown in the left two columns of table 4.4 [85].

4.3.7 Results

The score-to-LLR functions are computed using the suspect-specific as well as the suspect-independent training scores set in order to compare the general behavior of these functions. Figures 4.8-4.10 show the frequency histograms of the scores in the within-source and in the between-source sets, the ROC curves of the training scores in the suspect-specific and suspect-independent approach and the score-to-LLR functions computed by the ROCCH procedure as described in section 4.3.3.2.

Note the large variations in the histograms of the within-source scores for the suspect-specific approach. Within-source biometric specimens of each person are selected in such a way so that the variations are as close as possible across the five persons. However, still we observe considerable variation in the suspect-specific frequency histograms of scores. These variations are caused by either the slight variation in the specimen acquisition process or due to the fact that some people are easy to be recognised or differentiated from others [86]. As can be observed from the histograms of scores, besides the slight difference in the within-source specimens from person to person, identity itself has

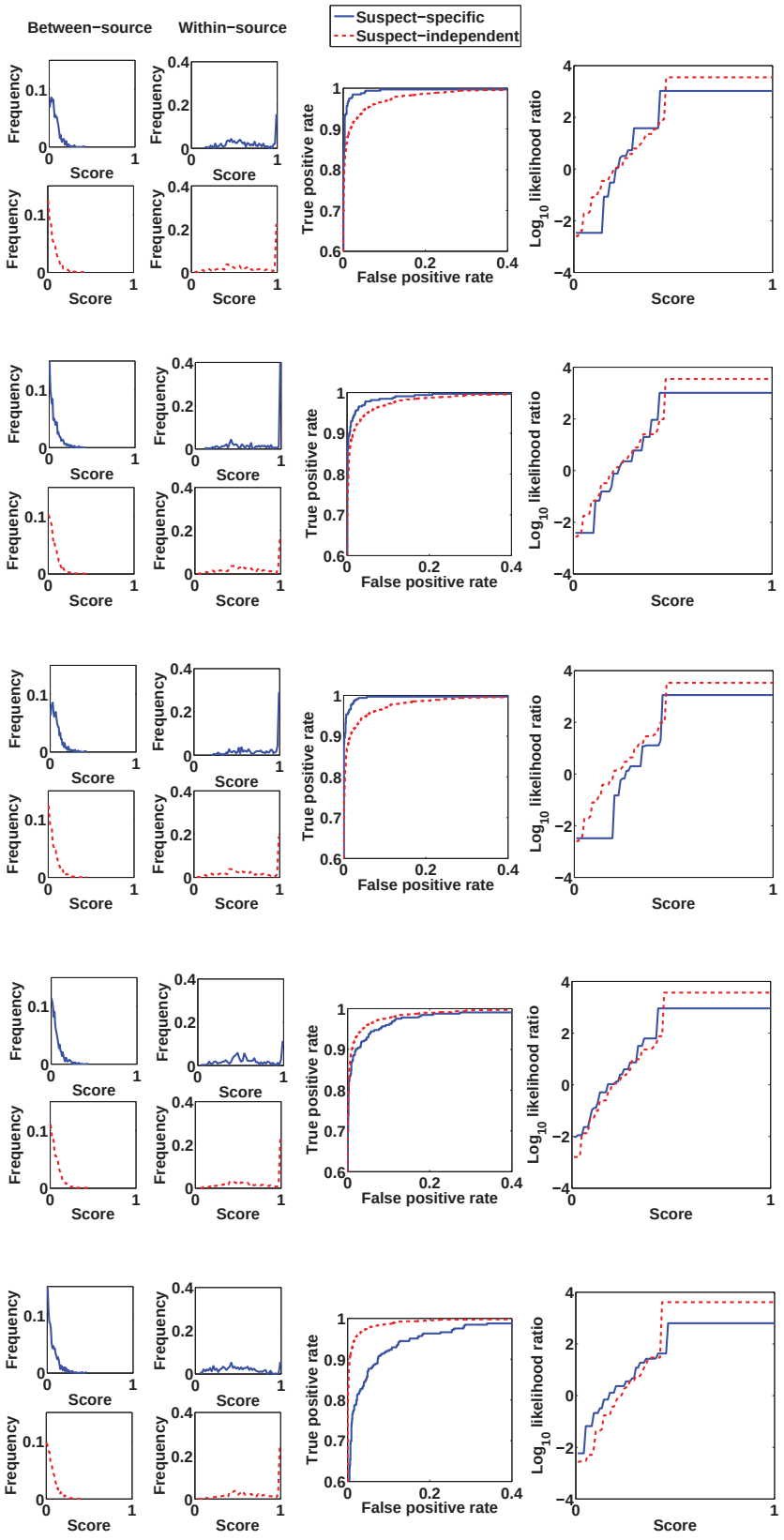


Fig. 4.8: Frequency histograms of scores, ROC curves and score-to-LLR functions for the five persons in the selected subset of FRGC face images database.

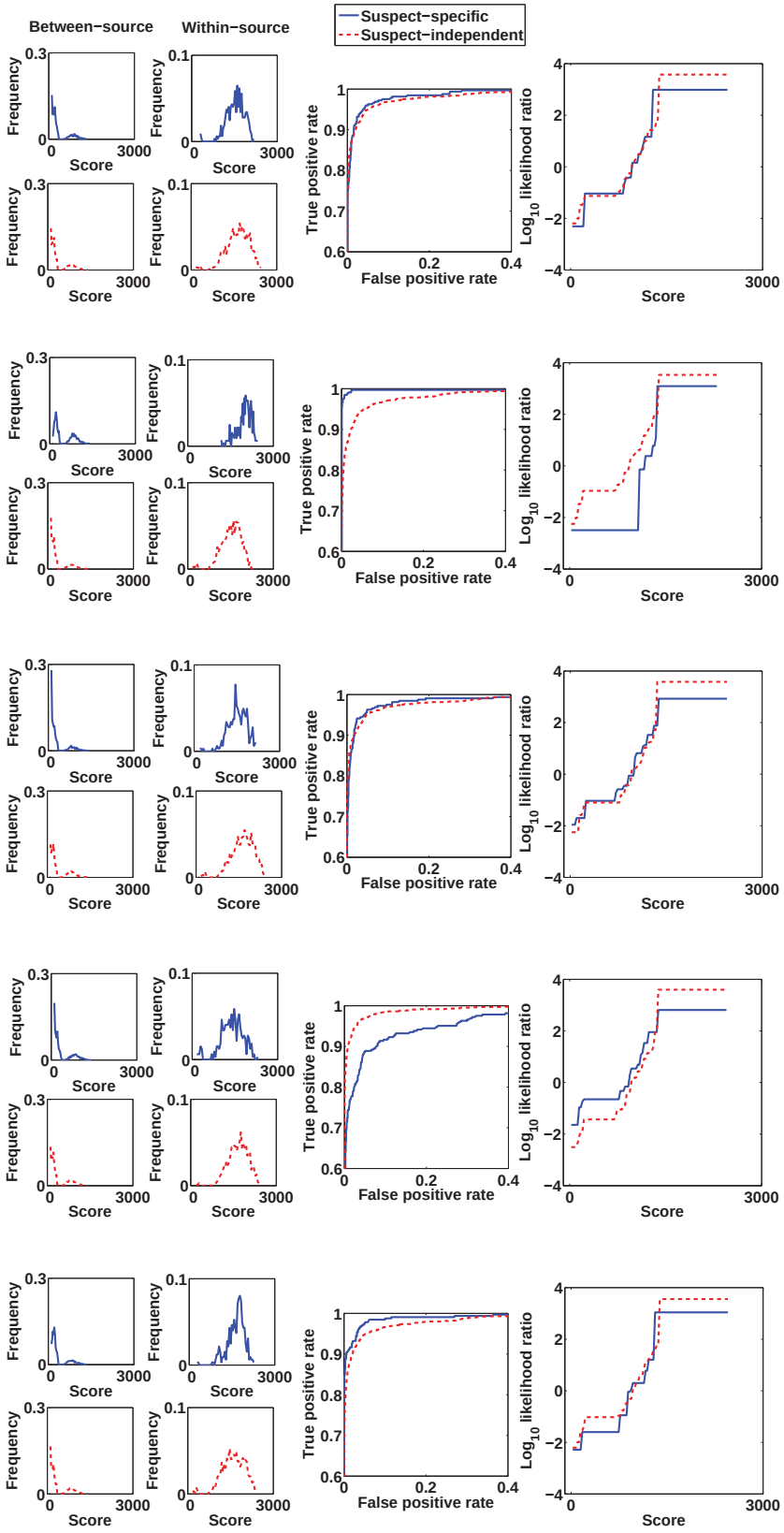


Fig. 4.9: Frequency histograms of scores, ROC curves and score-to-LLR functions for the five persons in the selected subset of KLPD fingerprints database.

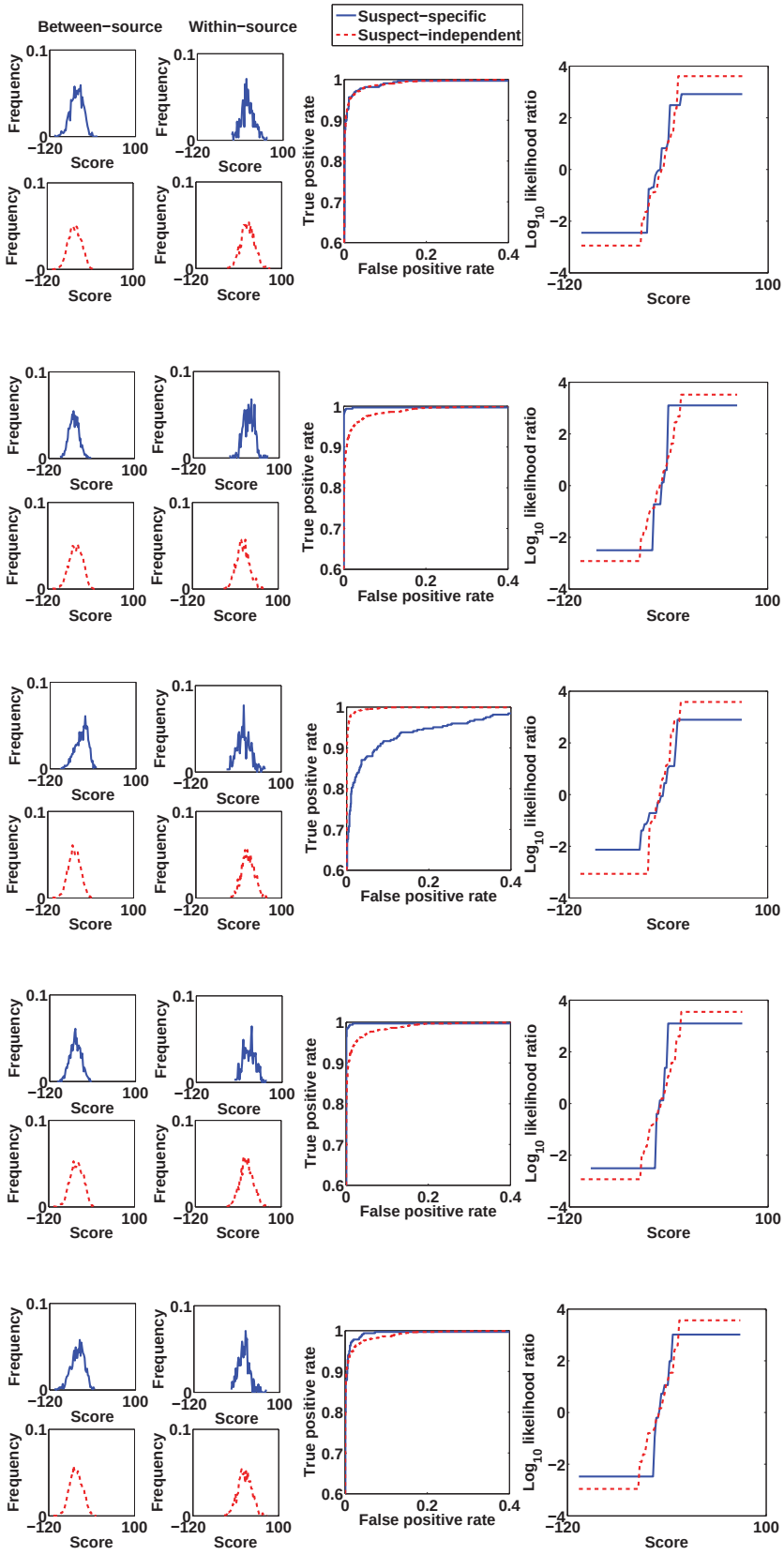


Fig. 4.10: Frequency histograms of scores, ROC curves and score-to-LLR functions for the five persons in the selected subset of NIST SRE speech recordings database.

a considerable effect on the suspect-specific within-source scores distribution. A biometric recognition system may perform differently for different persons when it is used to match a set of pairs of within-source and between-sources specimens.

The area under the ROC (AUROC) curve is used as a summary metric to assess the discrimination power of a set of within-source and between-sources scores. The motivation behind plotting ROC curves in this context is to demonstrate that there is no general conclusion about the discrimination power when comparing the suspect-specific and suspect-independent training scores.

There is a significant difference in the LR values computed using the suspect-specific and suspect-independent approach. For example, for person 1 in face recognition case, at the score location of 0.44, the suspect-specific and the suspect-independent LR values are 1052 and 78 respectively. The uppermost and the lowermost horizontal lines in the mapping functions are due to the proposed strategy to avoid infinite LLR values. We will refer to LR values along these horizontal lines as "saturated LR values". The magnitude of these LR values is directly proportional to the size of the training scores set and therefore, the suspect-independent approach results in saturated LR values of larger magnitude than the suspect-specific approach. In general, for all of the three biometric systems, it can be stated that there are significant differences in the suspect-specific and suspect-independent LR values. Therefore, it can be argued that anchoring plays a crucial role in computation of a LR for a given pair of biometric specimens.

In most cases, the exact numerical value of a likelihood ratio is of less importance than the range in which it lies. This fact should be taken into consideration when performing such a comparative study. To this end, the score-axis is uniformly sampled to simulate 100 values of evidence score s. These scores are converted to LLR values using both suspect-specific and suspect-independent training data. For 5 persons, this implies computation of 500 LLR values using suspect-specific as well as suspect-independent approach in each biometric modality. Table 4.4 shows the number of cases in which the two approaches compute LLR values which fall into a same range. As seen from table 4.4, in 296 cases out of 500 for face recognition, in 241 cases out of 500 for fingerprint recognition and in 294 cases out of 500 for speaker recognition system, the two LR values agree on a same verbal equivalents resulting in 59.2%, 48.2% and 58.8% agreement rates for face, fingerprint and speaker recognition systems respectively.

Note that using a different method to map from score values to LR values may

Ranges	Verbal equivalents		Number of agreements					
			P1	P2	P3	P4	P5	Total
		Face	0	0	0	0	0	0
$4 <$ LLR	Very strong evidence to support H_p	Fingerprint	0	0	0	0	0	0
		Speech	0	0	0	0	0	0
		Face	54	54	54	0	0	162
$3 <$ LLR≤ 4	Strong evidence to support H_p	Fingerprint	0	40	0	0	44	84
		Speech	0	36	0	38	38	112
		Face	0	0	0	0	0	0
$2 <$ LLR≤ 3	Moderately strong evidence to support H_p	Fingerprint	0	0	0	0	0	0
		Speech	2	0	2	0	0	4
		Face	9	8	10	8	9	44
$1 <$ LLR≤ 2	Moderate evidence to support H_p	Fingerprint	5	0	8	5	4	22
		Speech	1	0	2	0	4	7
		Face	9	10	6	13	7	45
$0 <$ LLR≤ 1	Limited evidence to support H_p	Fingerprint	8	0	5	6	8	27
		Speech	4	4	1	2	3	14
		Face	2	6	0	4	3	15
$-1 <$ LLR≤ 0	Limited evidence to support H_d	Fingerprint	5	0	7	4	7	23
		Speech	7	4	3	2	3	19
		Face	0	2	0	5	0	7
$-2 <$ LLR≤ -1	Moderate evidence to support H_d	Fingerprint	21	0	23	0	23	67
		Speech	0	0	0	0	0	0
		Face	4	4	4	2	4	18
$-3 <$ LLR≤ -2	Moderately strong evidence to support H_d	Fingerprint	5	5	0	0	3	13
		Speech	37	28	0	31	37	133
		Face	0	0	0	0	0	0
$-4 <$ LLR≤ -3	Strong evidence to support H_d	Fingerprint	0	0	0	0	0	0
		Speech	0	0	0	0	0	0
		Face	1	1	1	1	1	5
LLR < -4	Very strong evidence to support H_d	Fingerprint	1	1	1	1	1	5
		Speech	1	1	1	1	1	5
		Face	79	85	75	33	24	296
	Total number of agreements	Fingerprint	45	46	44	16	90	241
		Speech	52	73	9	74	86	294

Table 4.4: Number of times in which the LR values computed by the two approaches fall into a same range considering all of the five persons (P1, P2, P3, P4 and P5) in the selected subset. For each person considered as a suspect, there are 100 values of s generated by uniformly sampling the score-axis. Out of a total of 500 LR values computed by the two approaches, 296, 241 and 294 times the two LR values agree on one range for face, fingerprint and speaker recognition systems respectively.

lead to completely different results. Similarly, a different database of biometric specimens and a different biometric recognition system to compute scores can also slightly influence the difference between the suspect-specific and suspect-independent approach of LR computation. Authors are currently investigating the effect on the results when other methods of score-to-LR conversion such as KDE and logistic regression are used.

Effect of the size of training scores set

An obvious difference between the two approaches is the use of different sizes of the training scores sets. One way to study the effect of the difference in the sizes of the training sets between the suspect-specific and suspect-independent approach is to randomly sample a number of scores equal to the size of the suspect-specific sets from the suspect-independent sets. Given the size of the within-source and between-source sets in the two approach is the same, the variation in the LR values is only caused by the nature of the distributions of the scores. Figure 4.11 shows the mapping functions computed by the two approaches when the within-source and the between-source sets are equally sized by random subsampling the suspect-independent within-source and between-source sets so that the sizes of these sets in the suspect-independent approach is equal to those in the suspect-specific approach. Note that reduction in the size of the training scores reduces the range of LR values that can be computed. Besides the saturated region of LR values, the difference in the size of the training sets has very small effect on the resultant mapping function from score to LLR values.

4.3.8 Conclusions and future work

We discussed the effect of the different training data on the resultant LR values in the context of face, fingerprint and speaker recognition systems. The process of conversion of a score, computed from the comparison of the crime scene biometric specimen with the suspect biometric specimen, to a forensic LR is described. It is observed that there is a significant variation between the LR values computed using the suspect-specific and the suspect-independent approach. The differences are more prominent in the higher ranges of LR values and therefore more caution should be taken if one approach is used as an altenative to the other. Future work will include quantifying the influence of biometric specimens from other databases, different biometric recognition systems and other score-to-LR computation methods.

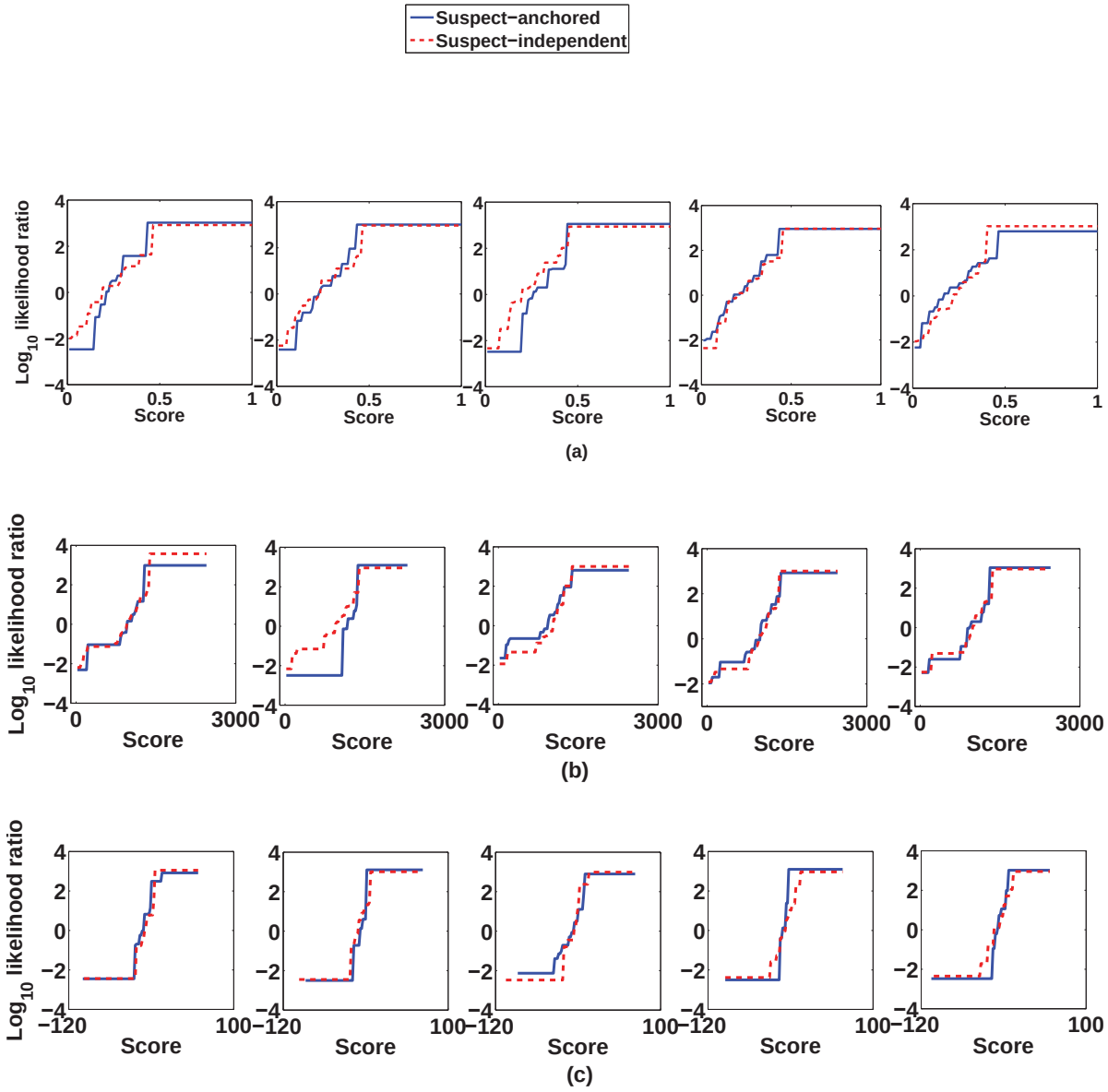


Fig. 4.11: Score-to-LLR functions using equal number of specimens in the within-source and between-source sets of the suspect-specific and suspect-independent approach. The suspect-independent within-source and between-source sets are randomly subsampled so that there are equal number of scores in these sets for the two approaches. (a) Face recognition system (b) Fingerprint recognition system (c) Speaker recognition system.

Chapter 5

Towards automated forensic face recognition and the LR framework

5.1 Introduction

In chapter 2 we explored the current practice of forensic facial comparison. Based on that, in this chapter we take a step towards automation of the process of forensic facial comparison performed by forensic examiners. The concept of LRs is also used for calibration of existing biometric face recognition systems. In section 5.2, inspired by the current practice of forensic facial comparison, we carry out a study to compare the recognition performance of two automatic face recognition algorithms when they are applied to perform the recognition task based on different facial features (such as eyes, nose, mouth, etc) separately. Studying the discriminating performance of facial features (also referred to as "facial regions" in this chapter) is important for future automated forensic face recognition systems. For example, it can be used to weight similarity scores computed for different facial features. Also a face recognition system based on the result of several independent facial features recognitions is more robust in cases when a small part of the face is available which is usually the case in forensic scenarios.

This chapter also presents a detailed study of the computation of LRs from face recognition systems. This is important since, in forensic evaluation, there is a growing interest to compute LRs for a comparison of a pair of biometric specimens instead of a score. To this end, in section 5.3, two state-of-the-art

face recognition systems are considered. Using kernel density estimation, LRs are computed and evaluation is performed using Tippett plot. Section 5.4 provide a more detailed study by considering 10 baseline face recognition algorithms and 3 different LR computation methods. The goal is to understand how well a system performs when instead of the scores, LRs are considered in assessment. It is also investigated that whether the choice of a LR computation method makes a significant effect on the performance of the overall LR computation system or not.

5.2 A study of identification performance of facial regions from CCTV images [1]

5.2.1 Abstract

This paper focuses on automatic face identification for forensic applications. Forensic examiners compare different parts of the face image obtained from a closed-circuit television (CCTV) image with a database of mug shots or good quality image(s) taken from the suspect. In this work we study and compare the discriminative capabilities of different facial regions (also referred to as facial features) such as eye, eyebrow, mouth, etc. It is useful because it can statistically support the current practice of forensic facial comparison. It is also of interest to biometrics as a more robust general-purpose face recognition system can be built by fusing the similarity scores obtained from the comparison of different individual parts of the face. For experiments with automatic systems, we simulate a very challenging recognition scenario by using a database of 130 subjects each having only one gallery image. Gallery images are frontal mug shots while probe set consist of low quality CCTV camera images. Face images in gallery and probe sets are first segmented using eye locations and recognition experiments are performed for the different face regions considered. We also study and evaluate an improved recognition approach based on AdaBoost algorithm with Linear Discriminant Analysis (LDA) as a week learner and compare its performance with the baseline Eigenface method for automatic facial feature recognition.

[1] The contents of this section are published in [25] "A study of identification performance of facial regions from CCTV images", In: 5th International Workshop on Computational Forensics (IWCF), 2012, Tsukuba, Japan.

5.2.2 Introduction

The difficulty of automatic face recognition mainly depends on the type of facial images we want to compare. A lot of research has been carried out to perform automatic face recognition and as a result several systems are available [47, 83, 91–94]. Problems such as different facial expressions, illumination conditions and poses have been studied and to certain extent some solutions have been proposed [27, 92, 95]. A relatively less investigated problem is the automatic face recognition from low quality images taken using CCTV camera. To date, there is no automatic system available which can reliably compare CCTV images with high quality images in mug shot database or image(s) taken from the suspect. This task is manually performed by forensic examiners where instead of following a holistic approach they use a "feature-based" approach. Each part such as nose, eyes, mouth, etc. is compared separately and a conclusion is reached by observing similarities and differences. Finally conclusions based on the different facial features along with the relative importance of each is used to state an opinion in the form of a ratio of how likely is that the two images being compared are obtained from the same person to how likely is that the two images being compared are obtained from different persons [23, 27].

The task of facial feature comparison is very challenging when one or both images under consideration are taken using CCTV camera because of the low quality. An automatic system comparing individual facial features is highly desirable as it will not only make the manual comparison of forensic examiners faster but will also help standardize this process. It is not possible with current state-of-the art recognition technologies to replace the manual comparison process in forensic face recognition; however, an automatic system can reduce, to a great extent, the manual effort. This can be, for instance, displaying top 10 candidate matches from a database of thousands of images based on a facial feature extracted from a criminal face image taken at a crime scene from a CCTV camera. Individual facial feature recognition is also important in cases such as having partial occlusion of the face and when only one facial feature is visible. In such cases even state-of-the art commercial face recognition systems such as [83] fail to work. Studies like the one presented in this paper are also necessary to scientifically support and help to establish procedures to assign relative weights to the opinions that can be inferred from different parts of the face.

In this paper we study the recognition performance of different facial features using two automatic recognition systems. The first system is the baseline

Eigenface approach [47] while the second system is based on AdaBoost algorithm where we use LDA as a weak learner. The remaining of this paper is organized as follows: section 5.2.3 reviews the protocol followed by forensic examiners to carry out the facial comparison which is the main motivation for this work. Section 5.2.4 describes the database, evaluation protocol and the segmentation of face images. Section 5.2.5 briefly describes the improved boosting-based LDA approach. Experimental results based on the Eigenface method and the boosting approach are presented in section 5.2.6. Finally, in section 5.2.7 we draw conclusions and mention future research directions.

5.2.3 Forensic examiners' facial comparison

In this section we briefly review the forensic experts' way of facial comparison which is the main motivation behind our work. The discussion is based on the guidelines set forward by the workgroup on face comparison at Netherlands Forensic Institute (NFI) [19, 20] which is a member of the European Network of Forensic Science Institutes [96]. The facial comparison is based on morphological-anthropological facial features. In most cases the pictures are obtained or processed to be in the same posture. The comparison mainly focuses on:

- Shape of mouth, eyes, nose, ears, eyebrows, etc.
- Relative distance among different relevant facial features
- Contour of cheek- and chin-lines
- Lines, moles, wrinkles, scars, etc. in the face

When comparing faces manually, it should be noted that differences can be invisible due to underexposure, overexposure, resolution too low, out-of-focus and distortions in imaging process, specifically when considering information from surveillance camera. Furthermore, similar facial features can result in different depictions due to the camera position regarding the head, insufficient resolution, difference in focusing of two images, and distortion in imaging process.

Due to the aforementioned effects, which usually make the comparison process difficult, the anthropological facial features are visually compared and classified as: similar in details, similar, no observation, different, different in details. Apparent similarities and differences are further evaluated by classifying facial features as: *weakly discriminative*, *moderately discriminative*, and *strongly discriminative*. A conclusion based on this comparison process is a form of support for either the prosecutions or defense hypothesis and can be stated

as no support, limited support, moderate support, strong support, and very strong support. The process is subjective to great extent and the conclusion of one expert can be different than other. The final result is based on the combination of the comparison results of different individual features. This is in contrast to automatic biometric face recognition systems where the whole face image is usually considered as a single entity [47, 92].

5.2.4 Database description and face segmentation

We use SCFace database [84] in our experiments which consists of 130 subjects each having one frontal mug shot image and 5 surveillance camera images. This database presents novel and challenging tests for automatic face recognition systems due to the very low quality images taken by surveillance cameras. A few examples of mug shot and surveillance camera images used in our experiments are shown in figure 5.1. There are five different surveillance cameras used each with three different distances from the subjects. For simplicity in our experiments we consider only one surveillance camera with the closest distance to the subjects.

All of the frontal mug shot and surveillance camera images are segmented using the ground truth locations of the eyes. Segmentation of the face image into different parts is based on standard facial proportions [26]. An example of the set of segments into which a face image is divided is shown in figure 5.2. As shown in figure 5.2, pixels outside the region of interest are masked by setting them to zero. Given a probe patch of a facial feature extracted from a surveillance camera image, it is matched with each of the 130 patches extracted from the frontal mug shot images.

5.2.5 Facial feature recognition

To handle the complex nature of individual facial feature recognition from low quality CCTV images we use LDA [97] as a weak learner in Adaboost.M2 [98] for feature [2] extraction while classification is performed using simple Euclidean distance. The performance of traditional LDA-based approach [93] is improved by incorporating it in the boosting framework. Since both LDA and AdaBoost

[2]Here the term "feature" refers to a vector of values describing the characteristics of an image patch. This is the common use of the term "feature" in pattern recognition. In order to avoid ambiguity we always use the term "facial features" for referring to the parts of the face such as eye, eyebrow, nose, etc.

Fig. 5.1: A few samples of gallery (first row) and probe images (second row) used in our experiments.

are well known algorithms we only provide a brief description of our employed recognition system highlighting the way LDA is integrated in AdaBoost.M2. Each round of the boosting generates a new LDA subspace particularly focusing on examples which are misclassified in previous LDA subspace. The final feature extractor module is an ensemble of several specific LDA solutions. In order to incorporate LDA in boosting framework, slight modifications are introduced in the way the within-class and the between-class scatter matrices are constructed at the end of each boosting iteration by incorporating the weight associated with each sample. Please refer to [94] for a detailed description of using LDA as a weak leaner in AdaBoost algorithm.

This kind of ensemble based approach takes advantage of both LDA and boosting and outperforms simple LDA based systems in complex face recognition tasks. This is particularly important where a small number of training samples for each subject are available (1 image patch per facial feature in this case) compared to the number of dimensions of the samples i.e., the small-sample-size problem [99] and when non-linear variations are present in facial images. Our employed face recognition system is more robust when performing recognition of low resolution face images. This result is also verified by the authors

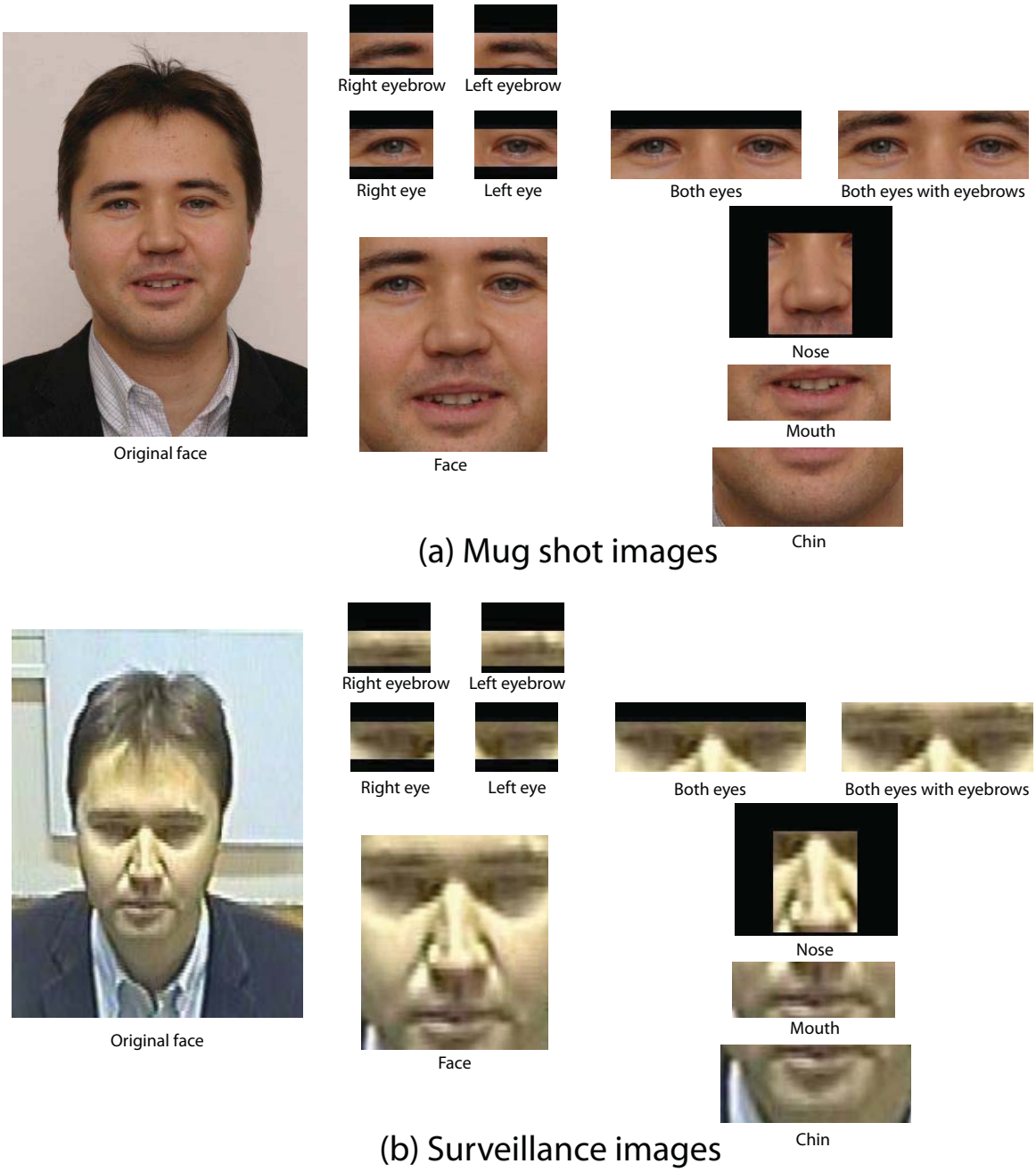


Fig. 5.2: (a) Mug shot images (b) Surveillance camera images.

in [94] where they use similar approach for face recognition.

5.2.6 Experimental results

There are 130 subjects each having only one image both in gallery and probe sets. Each face image is segmented and as a result we have 130 patches for each facial feature both in gallery and probe set. Figure 5.3(a) shows the Cu-

mulative Match Characteristics (CMC) curves of different facial feature when the Eigenface [47] method is applied to this close-set identification task. Only components whose eigenvalues are equal to or greater than 1 are retained. Simple Euclidean distance is used for classification. Very low identification results are observed mainly due to the very low quality probe patches obtained from surveillance camera images, only one training sample, and relatively high size of gallery. For the same identification scenario, we see improved identification rates for all facial features using the AdaBoost approach discussed in section 5.2.5 (figure 5.3 (b)).

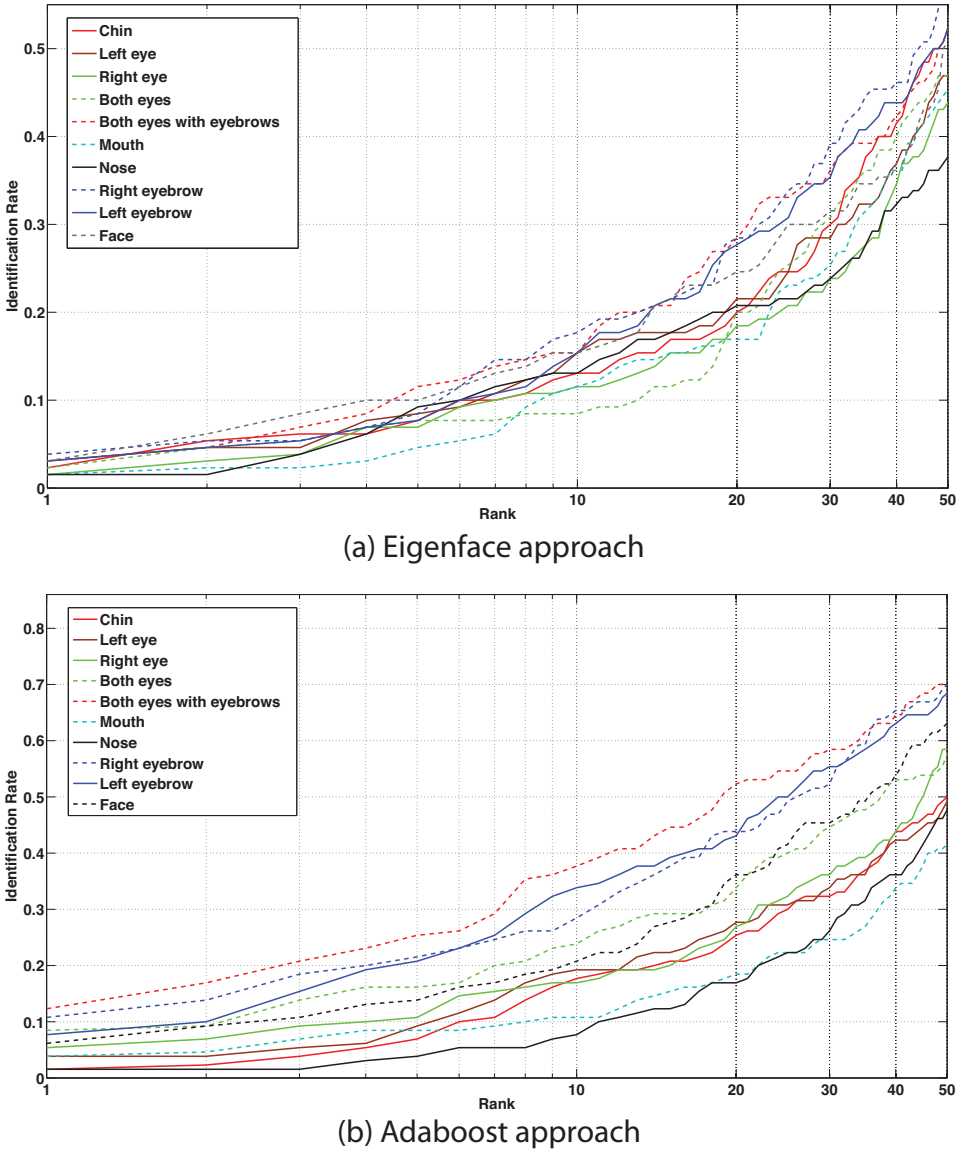


Fig. 5.3: Identification performance of different facial features.

Table 5.1-5.2 list the rank 1 and rank 10 identification rate of each facial feature. It can be concluded that different automatic systems might rank different

	Left eye	Right eye	Left EB	Right EB	Mouth	Nose	Eyes with EB	Eyes	Chin	Face
Eigenface approach	2.31	1.54	3.08	3.85	1.54	1.54	3.08	2.31	2.31	3.84
AdaBoost approach	3.85	5.39	7.69	10.77	3.85	1.54	12.31	8.46	2.31	6.15

Table 5.1: Rank 1 identification rate (%) (EB stands for eyebrow).

	Left eye	Right eye	Left EB	Right EB	Mouth	Nose	Eyes with EB	Eyes	Chin	Face
Eigenface approach	11.54	11.54	15.38	17.69	11.54	13.08	15.38	8.46	13.08	15.38
AdaBoost approach	19.23	16.92	33.85	28.46	10.77	7.69	37.69	23.85	17.69	20.77

Table 5.2: Rank 10 identification rate (%) (EB stands for eyebrow).

facial feature differently with respect to their discriminative capabilities. It is important to note that since the segmentation process is based on eyes location, the eye regions are expected to be better aligned than other regions. However, it is a standard practice in automatic face recognition to locate eyes positions and normalize face based on eyes positions.

Besides identification performance, it is also important to consider performance in verification scenario. In forensic facial comparison, a simple verification situation happens when an image from a suspect is compared with an image obtained from a crime scene. This is a one-to-one comparison for which different evaluation metrics such as Area under Receiver Operating Characteristics (ROC) curve (verification rate vs. false acceptance rate) and Equal Error Rate (EER) are used. In our experiments we use area under the ROC to summarize the verification performance of both systems for different facial features. Higher value of area under the ROC implies better verification performance of a system and vice versa. Table 5.3 summarizes results of verification experiments using the area under ROC curve metric. In table 5.4 we rank facial feature according to their verification performance using each method.

	Left eye	Right eye	Left EB	Right EB	Mouth	Nose	Eyes with EB	Eyes	Chin	Head
Eigenfaces	57	56	61	63	55	53	60	56	60	59
AdaBoost approach	63	66	75	77	55	57	79	69	61	72

Table 5.3: Verification performance using percentage of area under ROC (EB stands for eyebrow).

Eigenface method	**AdaBoost approach**
Right eyebrow	Both eyes with eyebrow
Left eyebrow	Right eyebrow
Both eyes with eyebrows	Left eyebrow
Chin	Face
Face	Eyes
Left eye	Right eye
Eyes	Left eye
Right eye	Chin
Mouth	Nose
Nose	Mouth

Table 5.4: Ranking facial feature based on verification performance.

5.2.7 Conclusions and future work

Comparing individual facial features of two or more faces is a common practice that forensic examiners carry out during their investigation of a crime when there are facial images from a crime scene. In this paper we presented preliminary experiments to compare and evaluate the discriminative capabilities of different facial features. We studied a boosting based LDA approach and compared its performance with the standard Eigenface method for individual facial feature recognition. The studied method has shown improved performance, however, still it is far from the point where it can be used in real applications. It is however important to study and understand the recognition performance of different facial features by recognition algorithms. This can lead to future research such as building more robust face recognition systems by the weighted sum of all facial features recognition results. Also it is more important in cases where crime scene images are partially occluded or only a few facial features are visible. Our future research will include improving the recognition performance as well as combining evidence from different facial feature comparison to single evidence for forensic face recognition.

5.3 Towards automatic forensic face recognition [3]

5.3.1 Abstract

In this paper we present a methodology and experimental results for evidence evaluation in the context of forensic face recognition. In forensic applications, the matching score (hereafter referred to as similarity score) from a biometric system must be represented as a Likelihood Ratio (LR). In our experiments we consider the face recognition system as a 'black box' and compute LR from similarity scores. The proposed approach is in accordance with the Bayesian framework where the duty of a forensic scientist is to compute a LR from biometric evidence which is then incorporated with prior knowledge of the case by the judge or jury. In our experiments we use a total of 2878 images of 100 subjects from two different databases. Our experimental results prove the feasibility of our approach to reach a LR value given an image of a suspect face and a questioned face. In addition, we compare the performance of two biometric face recognition systems for forensic LR computation.

5.3.2 Introduction

Output of a score-based biometric system is not suitable for forensic applications where the objective is to obtain a degree of support for one hypothesis (prosecution) against the other (defense). This issue is discussed in detail in previous literature on forensic speaker recognition [32, 33, 52], forensic voice comparison [53] and some other fields of forensic science such as DNA [5]. Systems using a threshold to decide between two classes are not acceptable in forensic domain [52]. For forensic applications, the Bayesian interpretation framework is an agreed upon standard way to report evidence value from a biometric system. However, less effort has been done to utilize this framework in forensic face recognition in contrast to forensic speaker, voice, fingerprints, DNA, etc. There are very few published works [54] which focus on the forensic aspects of face recognition and there is an utmost need for a reliable facial comparison and recognition systems which can assist law enforcement agencies in investigation and whose output can be readily used in judicial system. In [54] the author performs preliminary experiments to reach to LR values for

[3] The contents of this section are published in [23] "Towards automatic forensic face recognition", In: International Conference on Informatics Engineering and Information Science (ICIEIS), 2011, Malaysia, Communications in Computer and Information Science 252, pp. 47-55, Springer Verlag, ISSN 1865-0929.

forensic face recognition. However, the approach lacks suitable modeling of Within-Source Variability (WSV) and Between-Source Variability (BSV) before LR computation. In a typical forensic face recognition scenario, a forensic expert is provided with two face images; one of a suspect (usually obtained from a mugshot database) and other face image is of a person whose identity is in question (the perpetrator). The duty of forensic expert is to reach to a LR value which is interpreted as a degree of support for one hypothesis against the other. The first hypothesis, called prosecution hypothesis, states that the suspect is the source of unknown face. Second hypothesis, called defense hypothesis, states that someone else in potential population (not the suspect) is the source of unknown face. Usually as a part of forensic face recognition, forensic experts also have to compare a questioned face to a database of mugshots. It is highly desirable to automate the process of forensic facial comparison which will not only speed up comparison but will also standardize the process. The remaining of the paper is organized as follows: In section 5.3.3 we discuss general idea of Bayesian framework. Section 5.3.4 presents the computation process of LR from a similarity score. Section 5.3.5 reviews briefly our employed face recognition systems. Section 5.3.6 shows experimental results and finally in section 5.3.7 we conclude our work and show future research directions.

5.3.3 Bayesian interpretation framework

Bayesian framework (or the likelihood ratio framework) is a logical approach to evaluation of evidence from a biometric system and can be applied to any biometric system without change in the underlying theory. A general description of this framework can be found in [63]. A description of this framework in the context of forensic speaker recognition, voice comparison, and DNA analysis can be found in [5, 32, 33, 52, 53]. In this framework the task of a forensic scientist is to compute a LR based on the evidence from a biometric system. This LR assessed from a score based biometric system is then provided to judge or jury where they combine it with the prior knowledge about the case (I) to reach to a conclusion. The basic idea of this framework is that evidence does not consist uniquely of scientific data [100] and the forensic scientist while evaluating evidence from a biometric system should report a LR. While in commercial biometric systems, the objective is to make a decision in binary form, in forensic applications, the objective is to find the degree of support for one hypothesis against the other. Using the Bayes theorem, given the prior odds (prior knowledge of the case) and the LR, the posterior odds

can be calculated as:

$$\frac{Pr(H_p|E,I)}{Pr(H_d|E,I)} = \frac{Pr(E|H_p,I)}{Pr(E|H_d,I)} \times \frac{Pr(H_p|I)}{Pr(H_d|I)}. \tag{5.1}$$

where H_p and H_d are the prosecution and defense hypothesis respectively and E represents forensic information (evidence) while I is background information on the case at hand. The prosecution hypothesis H_p states that the suspect is the source of the questioned face while the defense hypothesis H_p states that someone else in the relevant population is the source of the questioned face. In this framework, the likelihood ratio $\frac{Pr(E|H_p,I)}{Pr(E|H_d,I)}$ gives a measure of the degree of support for one hypothesis H_p against the other H_d based on the scientific analysis of the questioned face. It calculates the conditional probability of observing a particular value of the evidence with respect to two competing hypotheses. The task of a forensic scientist in this framework is to compute a LR value from the evidence (similarity score) which is then used by the judicial system to reach to a decision. A LR greater than 1 supports prosecution hypothesis, for instance, if LR is 10, it will be interpreted as 10 times stronger belief that the suspect is the perpetrator regardless of the prior information about the case. Similarly a LR less than 1 supports the defense hypothesis. A LR of 1 supports both hypothesis equally or in other words no additional information can be derived from the biometric evidence.

5.3.4 Computation of a LR

The numerator of the LR is the probability of observing the evidence (score value) given the prosecution hypothesis is true. It requires WSV of the suspect to be computed. The denominator of the LR is the probability of observing the same evidence (score value) given the defense hypothesis is true. It requires the BSV to be estimated from the relevant population. Fig.5.4-5.5 illustrate general approach of LR computation given a suspect and a questioned face.

Estimation of the WSV. The WSV is estimated by obtaining similarity scores using a database of a number of images of the suspect. This database is called 'control database' and contain images of suspect taken under similar conditions as that of the image obtained at the crime scene (questioned face image). Variability exists in the image pairs used to estimate the WSV. For example, the suspect-dependant approach requires that in all comparisons for the WSV estimation, there must be images of the suspect which is under investigation. The second approach called suspect-independent allows

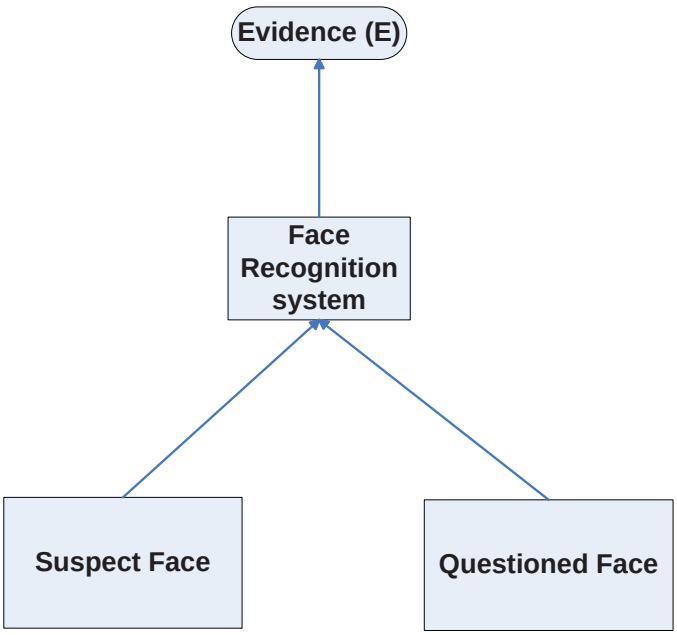


Fig. 5.4: Evidence from a face recognition system.

to accumulate a number of different subjects each having several images taken under similar conditions. Face images of each subject are then compared with other images of that subject to estimate the WSV. Due to the difficulty of obtaining enough number of images from one subject to estimate the WSV, the suspect-independent approach is generally followed.

Estimation of the BSV. To estimate the BSV, we need a large scale database of relevant population. This database referred to as 'potential population database' should be made of a large number of facial images taken under similar conditions to that of the questioned image. Generally it is hard to decide as to the exact number of images to be used in this database as it should ideally be dependent on the case at hand. However once a system is employed and tested on sample database it can be easily adjusted to the particular case at hand by changing the size and nature of the relevant population in the potential population database.

Modeling the WSV and the BSV. Once the WSV and the BSV of similarity scores are obtained, the next step is to model the distributions of score using a probability density function. In our work we use Kernel Density Estimation (KDE) [56] for WSV and BSV modeling. The use of KDE to model

WSV and BSV is also demonstrated by Meuwly [57] for forensic speaker recognition. KDE smooths out the contribution of each observed data point over a local neighborhood of that data point. The contribution of data point s_i to the estimate at some point s depends on how apart s_i and s are. The extent of this contribution is dependent upon the shape of the kernel function adopted and the width (bandwidth) accorded to it. If we denote the kernel function as K and its bandwidth by h, the estimated density at any point s is

$$f(s) = \frac{1}{n} \sum_{i=1}^{n} K \left(\frac{(s - s_i)}{h} \right) \tag{5.2}$$

The size of the kernel function can be optimally computed as:

$$h = \left(\frac{4\hat{\sigma}^5}{3n} \right) \tag{5.3}$$

where $\hat{\sigma}$ is the sample standard deviation of the samples and n is the size of samples. Once the background probability density function (pdf) of the WSV and the BSV are in hands, LR is simply computed by dividing the pdf of the WSV by the pdf of the BSV at value of the evidence (similarity score).

5.3.5 Face recognition systems

We use two face recognition systems in our experiments. Both systems are used as a 'black box' however a general description of each system is presented in this section.

System A. This system is based on AdaBoost [98] algorithm with LDA [93] as a week learner is used for feature selection in LDA subspace while classification is performed using a classic nearest center classifier. This approach is based on the work of [81] and is partially inspired by Viola's and Jone's [101] work as boosting is used for feature extraction and not for classification. The performance of traditional LDA-based approach is improved by incorporating it in the boosting framework. Each round of boosting generalizes a new LDA subspace particularly focusing on examples which are misclassified in the previous LDA subspace. The final feature extractor module is an ensemble of several specific LDA solutions. This kind of ensemble based approach take advantage of both the LDA and the boosting and outperforms only LDA based systems in complex face recognition tasks such as the case where less number of training samples for each subject are available compared to number of dimensions of the samples (small-sample-size problem) [99] and when non-linear

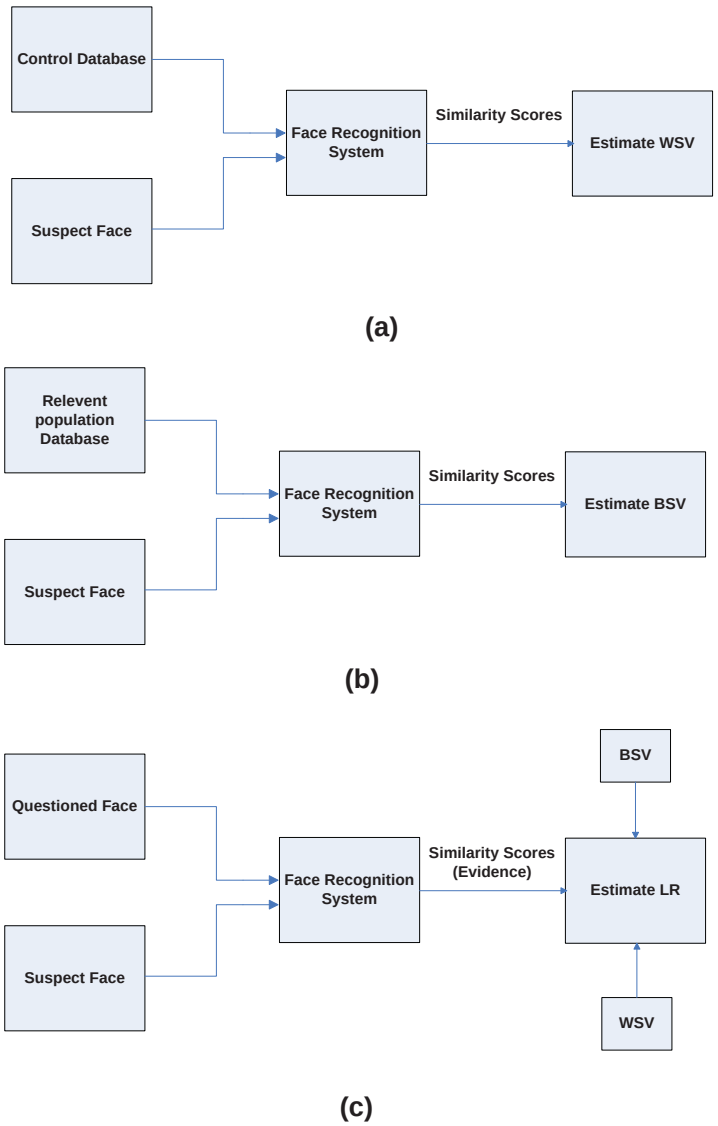


Fig. 5.5: (a) Estimation of WSV (b) Estimation of BSV (c) Computation of LR.

variations are present in facial images. Refer to [81] for a detailed description of this employed face recognition system.

System B. This system used in our experiments is Cognitec [83] commercial face recognition system. This system has better discriminating ability and handles the pose and illumination problem better than system A. Detail about the Cognitec system can be found here [83]. Both systems are essentially used as a 'black box' and the main objective is to explore and evaluate their use in forensic evidence evaluation.

5.3.6 Experimental results

In our experiments we use a total of 2878 images of 100 different subjects. Images are collected from two public databases: BioID [102] and FRGC [89]. Variation in terms of facial expression and lighting conditions exist in both databases. Mostly images are frontal and only small variation exist in pose of facial images. Fig.5.6 shows some example images used in our experiments. 60% images of each subject are used for training and remaining 40% for testing the face recognition system. Depending on the number of test images available for each subject, 1 to 3 images in the testing set are used for suspect trails. This corresponds to a total of 1736 training images, 954 testing images and 188 suspect images. Each test results in 100 similarity scores corresponding to each of the subject in the database. For each test image, the target similarity score is used for the WSV estimation and the remaining 99 similarity scores are used for the BSV estimation. Following this procedure for 954 test images we get a total of 954 similarity scores for the WSV and 94446 (954 X 100 - 954) similarity scores for estimation of the BSV. Fig.5.7 shows histograms of similarity scores obtained for target matches or the WSV and non-target matches or the BSV using system A. Fig.5.8 shows the pdfs of the WSV and the BSV estimated from the histograms of the similarity scores in Fig.5.7 using KDE. LR is then computed by dividing the pdf of the WSV by the pdf of the BSV at value of similarity score for which we want to find LR. Fig.5.8 also illustrates calculation of LR for an evidence value (similarity score) of 20. LR of 18.79 implies that a suspect image whose similarity score with the questioned face is 20 is 18.79 times more likely to be the source of the questioned face compared to the hypothesis that someone else is the source of the questioned face. The LR value of 18.79 can be multiplied to prior odds (prior knowledge of the case) by the judicial system to reach to posterior odds (conclusion).

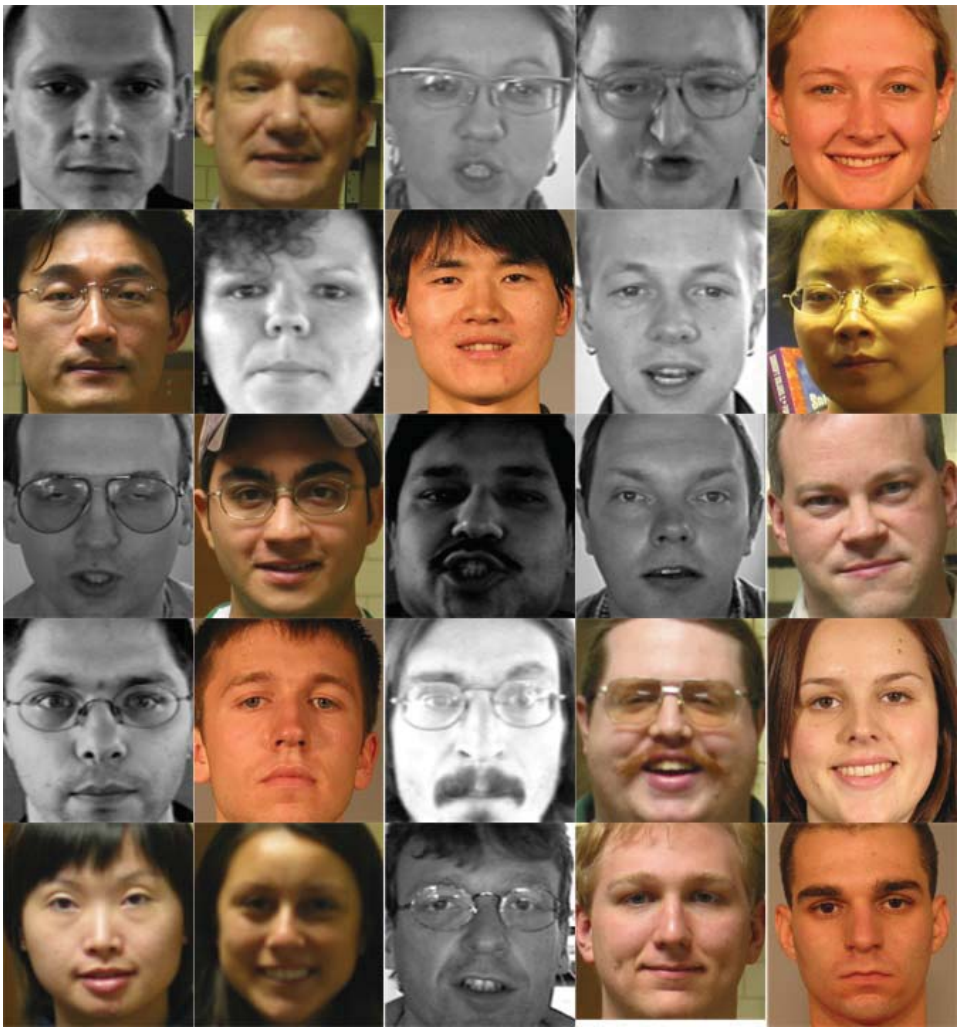


Fig. 5.6: Example images from BioID and FRGC database used in experiments

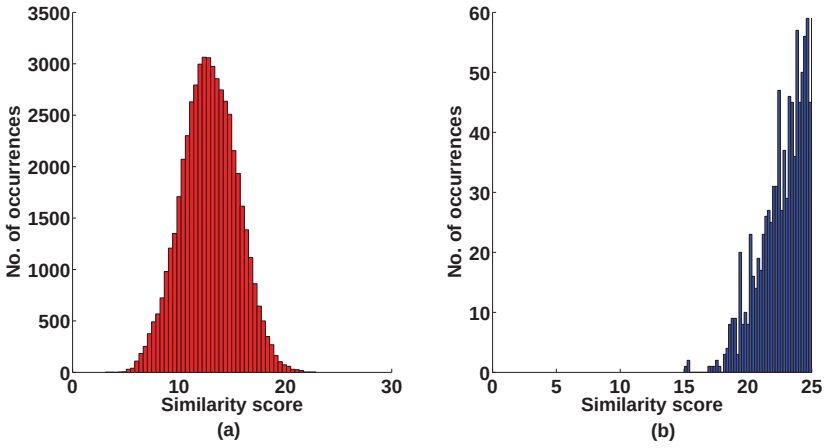


Fig. 5.7: (a) Histogram of similarity scores obtained for non-target matches (BSV); (b) Histogram of similarity scores obtained for target matches (WSV)

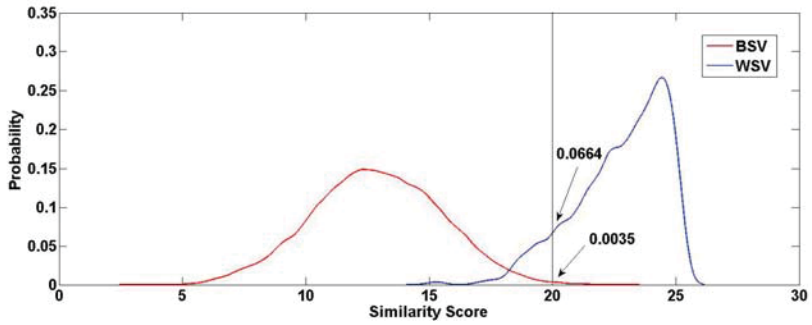


Fig. 5.8: Probability density functions of the WSV and the BSV estimated using KDE. To compute LR value for similarity score of 20, the pdf of the WSV is divided by the pdf of the BSV at value 20, 0.0664 / 0.0035 = 18.79

Fig.5.9 shows result of modeling the WSV and the BSV of the similarity scores using same dataset for system B. Score range of System B is from 0 to 1, however, for the ease of comparison, it is scaled to match score range of System A. As can be seen from fig.5.9, system B better separates the WSV and the BSV scores values which shows better discriminating ability of the system.

The performance of a forensic system is better evaluated using the Tippett plot. Fig.5.10 shows the tippet plot when 50000 non-target and 1000 target LR values are computed using both systems. Greater separation between the curves of the target and the non-target LR values is desirable and shows the system reliability for computing LR values.

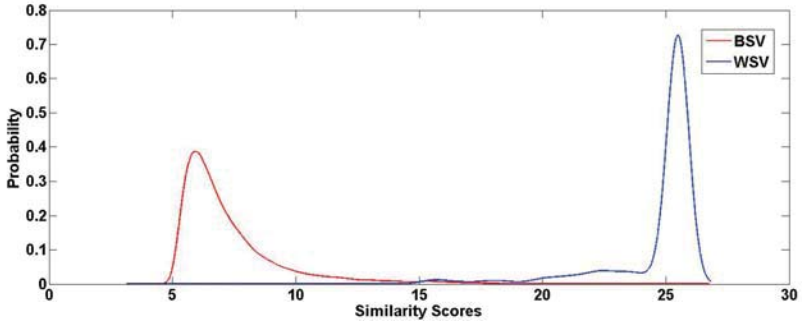


Fig. 5.9: Probability density functions of the WSV and the BSV estimated for similarity scores obtained from System B.

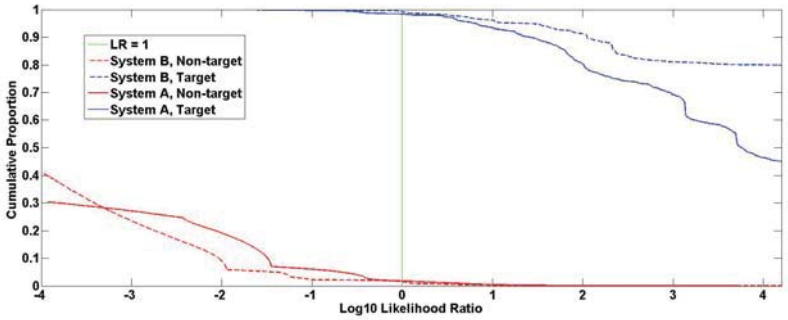


Fig. 5.10: Tippett plot computed for the 1000 target and the 50000 non-target LR values

5.3.7 Conclusions and future work

This work focused on the use of existing biometric face recognition systems in forensics. The process of LR computation is explained in face recognition context. Large numbers of face images from two public databases are used to estimate the WSV and the BSV. The target and non-tart LRs for two different face recognition systems are evaluated for same set of face data and compared using Tippett plot. Future research will include better methods of LR computation and working with facial data taken under forensic conditions.

5.4 Calibration and comparison of baseline face recognition algorithms

5.4.1 Abstract

In this section, we measure performance of 10 baseline face recognition algorithms after they have been calibrated using three commonly used calibration (LR computation) methods. The effect of ZT score normalization technique [103] is also reported. Experiments are carried out using SCFace database which, for each subject, consists of a mugshot image and images from surveillance cameras at three different distances.

5.4.2 Face recognition algorithms

We use 10 baseline face recognition algorithms in this comparative study. These algorithms are Probabilistic Linear Discriminant Analysis (PLDA) [104], Local Gabor Binary Patterns Histogram Sequences (LGBPHS) [105], Linear Discriminant Analysis (LDA) [106], Inter-Session Variability (ISV) [107], Gaussian Mixture Model (GMM) [103], Gabor graphs [108], Bayesian Intrapersonal/extrapersonal Classifier (BIC) [109], Eigenfaces [47], Local Region Principal Component Analysis (LRPCA) [110] and Cohort Linear Discriminant Analysis (CLDA) [111]. The choice of these algorithms is based on their wide use and availability of standard implementations. Refer to [112] for details of the specific implementations of these algorithms. These stanadard implementations are available online here [113].

5.4.3 Performance evaluation

We use two performance evaluation tools. The first is Area under ROC (AUC) which gives a summary metric of the discriminating power of a given set of scores. The larger the value of AUC, the better the discriminating power of a given set of scores. AUC is not a suitable measure for performance evaluation of a set of LR values because LR values are not used in threshold-based decision making systems. To this end, another metric of performance evaluation called Cost of Log-Likelihood Ratios (C_{llr}) [17] is used to measure performance of a

given set of LR values:

$$C_{\text{llr}} = \frac{1}{2}\left(\frac{1}{N_{ss}}\sum_{i=1}^{N_{ss}}\log_2(1+\frac{1}{LR_{ss_i}}) + \frac{1}{N_{ds}}\sum_{j=1}^{N_{ds}}\log_2(1+LR_{ds_j})\right) \tag{5.4}$$

where N_{ss} and N_{ds} are the number of the same-source and different-source LRs respectively. LR_{ss} and LR_{ds} are the sets of the same-source and different-source LRs whose performance need to be assessed. The smaller the value of the C_{llr} for a given set of LR values, the better the corresponding biometric recognition system and the calibration method which produces those LR values. C_{llr} measures both the discrimination as well as the calibration.

5.4.4 Experimental results

Three LR computation methods are used in combination with the 10 baseline FR algorithms. The LR computation methods are discussed in chapter 3. Another procedure called "score normalization" is also considered in performance evaluation. Performance is evaluated at both the score level and LR level. There are 4 different protocols of the SCFace database: "close", "medium", "far" and "combined". Table 5.5-5.8 show performance evaluation of scores using AUC and of LRs using C_{llr} for each of the four protocols. Fig.5.11 shows the score to LLR functions using the three different LR computation methods and two different FR algorithms to compute scores. Note that the range of LRs that can be computed is dependent on the score computation algorithm. Fig.5.12-5.13 show Tippett plots of the evaluation sets of LLRs computed using the three methods of LR computation. For Tippett plot in Fig.5.12, the scores are computed using LGBPHS approach while for Tippett plot in Fig.5.13, the scores are computed using Eigenfaces approach. Comparing the Tippett plots in Fig.5.12-5.13, it can be concluded that the LR computation method has slight effect on the performance of resultant LRs whereas the score computation algorithm has significant effect on resultant LRs.

5.4.5 Conclusions and future work

A comparative study of 10 baseline face recognition systems is carried out using standard implementations of these methods. Three LR computation methods are then used to calibrate the output of these systems. Performance

Table 5.5: Results using the "close" protocol of the SCFace database.

Performance evaluation of scores using AUC

Score Normalization	Face recognition algorithm									
	PLDA	LGBPHS	LDA	ISV	GMM	Gaborgraph	BIC	Eigenfaces	LRPCA	LDAIR
None	0.6756	0.7803	0.6991	0.8654	0.7963	0.7397	0.5225	0.6315	0.6819	0.8358
ZT	0.6584	0.8764	0.8257	0.8575	0.8360	0.8448	0.5458	0.6551	0.6788	0.8136

Performance evaluation of LR values using C_{llr}

Normalization technique	Calibration Method	Face recognition algorithm									
		PLDA	LGBPHS	LDA	ISV	GMM	Gaborgraph	BIC	Eigenfaces	LRPCA	LDAIR
None	KDE	0.9326	0.8141	0.9152	0.6569	0.7955	0.8529	1.006	0.9583	0.9132	0.7100
	Log Reg	0.9293	0.8037	0.9121	0.6451	0.7759	0.8576	1.0005	0.9577	0.9305	0.7090
	PAV	0.9432	0.8054	0.9081	0.6484	0.7890	0.8543	0.9987	0.9675	0.9193	0.7197
ZT	KDE	0.9391	0.6698	0.7634	0.6809	0.7428	0.7085	1.0002	0.9638	0.9250	0.7586
	Log Reg	0.9361	0.6540	0.7501	0.6703	0.7196	0.7033	0.9998	0.9454	0.9355	0.7560
	PAV	0.9574	0.6518	0.7497	0.6761	0.7364	0.7181	0.9968	0.9476	0.9339	0.7724

Table 5.6: Results using the "medium" protocol of the SCFace database.

Performance evaluation of scores using AUC

Score Normalization	Face recognition algorithm									
	PLDA	LGBPHS	LDA	ISV	GMM	Gaborgraph	BIC	Eigenfaces	LRPCA	LDAIR
None	0.6850	0.8245	0.7260	0.8218	0.7054	0.8014	0.5252	0.6666	0.6258	0.8638
ZT	0.6795	0.8981	0.8407	0.8187	0.7671	0.9060	0.5323	0.6772	0.6267	0.8477

Performance evaluation of LR values using C_{llr}

Normalization technique	Calibration Method	Face recognition algorithm									
		PLDA	LGBPHS	LDA	ISV	GMM	Gaborgraph	BIC	Eigenfaces	LRPCA	LDAIR
None	KDE	0.9160	0.7528	0.8839	0.7487	0.8911	0.7741	1.0212	0.9421	0.9470	0.6533
	Log Reg	0.9198	0.7627	0.9227	0.7402	0.8968	0.7865	0.9993	0.9416	0.9869	0.6532
	PAV	0.9317	0.7626	0.8923	0.7453	0.8977	0.7770	1.0023	0.9464	0.9671	0.6573
ZT	KDE	0.9258	0.5688	0.7171	0.7602	0.8373	0.5904	1.004	0.9357	0.9417	0.6910
	Log Reg	0.9217	0.5678	0.7192	0.7501	0.8297	0.5689	1.0004	0.9258	0.9977	0.6910
	PAV	0.9247	0.5781	0.7165	0.7527	0.8262	0.5917	0.9977	0.9198	0.9675	0.6924

Table 5.7: Results using the "far" protocol of the SCFace database.

Performance evaluation of scores using AUC

Score Normalization	Face recognition algorithm									
	PLDA	LGBPHS	LDA	ISV	GMM	Gaborgraph	BIC	Eigenface	LRPCA	LDAIR
None	0.6527	0.6958	0.6851	0.6000	0.5472	0.6386	0.4515	0.6328	0.6304	0.8151
ZT	0.6470	0.7868	0.8251	0.5987	0.5693	0.7308	0.4405	0.6244	0.6253	0.7879

Performance evaluation of LR values using C_{llr}

Normalization technique	Calibration Method	Face recognition algorithm									
		PLDA	LGBPHS	LDA	ISV	GMM	Gaborgraph	BIC	Eigenfaces	LRPCA	LDAIR
None	KDE	0.9435	0.9050	0.9315	0.9750	1.0052	0.9517	1.0102	0.9647	0.9656	0.7631
	Log Reg	0.9460	0.9094	0.9395	0.9748	0.9966	0.9500	1.0005	0.9600	0.9584	0.7606
	PAV	0.9474	0.9109	0.9427	0.9778	1.0228	0.9523	1.007	0.9730	0.9589	0.7672
ZT	KDE	0.9642	0.8094	0.7487	0.9772	1.0001	0.8764	1.0068	0.9497	0.9577	0.8244
	Log Reg	0.9509	0.7961	0.7405	0.9767	0.9926	0.8641	1.0042	0.9632	0.9588	0.8199
	PAV	0.9582	0.8200	0.7444	0.9787	1.0024	0.8706	1.0124	0.9451	0.9623	0.8405

Table 5.8: Results using the "combined" protocol of the SCFace database.

Performance evaluation of scores using AUC

Score Normalization	Face recognition algorithm									
	PLDA	LGBPHS	LDA	ISV	GMM	Gaborgraph	BIC	Eigenfaces	LRPCA	LDAIR
None	0.6745	0.7435	0.6955	0.7597	0.6362	0.6961	0.5004	0.6410	0.6468	0.8406
ZT	0.6646	0.8540	0.8334	0.7564	0.7214	0.8283	0.5076	0.6517	0.6449	0.8185

Performance evaluation of LR values using C_{llr}

Normalization technique	Calibration Method	Face recognition algorithm									
		PLDA	LGBPHS	LDA	ISV	GMM	Gaborgraph	BIC	Eigenfaces	LRPCA	LDAIR
None	KDE	0.9296	0.8504	0.9192	0.8243	0.9190	0.8871	1.0050	0.9556	0.9416	0.7087
	Log Reg	0.9316	0.8597	0.9271	0.8218	0.9519	0.9050	1.0000	0.9550	0.9559	0.7087
	PAV	0.9320	0.8552	0.9226	0.8208	0.9251	0.8838	1.0003	0.9577	0.9447	0.7162
ZT	KDE	0.9421	0.6859	0.7379	0.8342	0.8829	0.7349	1.0000	0.9453	0.9413	0.7594
	Log Reg	0.9363	0.6849	0.7354	0.8321	0.8853	0.7279	0.9999	0.9475	0.9600	0.7556
	PAV	0.9430	0.6895	0.7347	0.8277	0.8828	0.7356	0.9997	0.9385	0.9505	0.7636

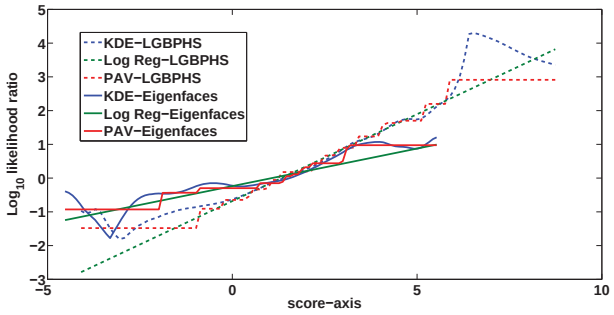


Fig. 5.11: Mapping functions from Score-axis to Log10-likelihood-ratio-axis for the "close" protocol using ZT-normalized scores. The score-axis ranges from the minimum and maximum value in the calibration scores set. A set of 100 scores are sampled uniformly from the score-axis to generate the functions.

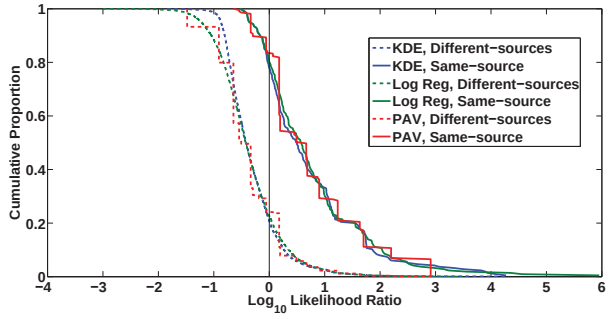


Fig. 5.12: Tippett plots of the likelihood ratio values for LGBPHS face recognition algorithm using ZT-normalized scores and the "close" protocol.

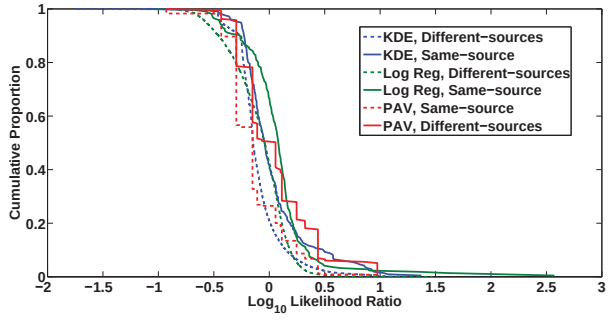


Fig. 5.13: Tippett plots of the likelihood ratio values for Eigenfaces face recognition algorithm using ZT-normalized scores and the "close" protocol.

is evaluated both at the score as well as at the LR level. It is observed that the three LR computation methods, KDE, Log Reg and PAV, do not significantly affect the results in terms of the value of the C_{llr} or the Tippett plot evaluation. Future work includes fusion of these baseline face recognition systems and a study of the distributions of the scores of each of the face recognition system in order to choose an appropriate method of LR computation.

Chapter 6

Conclusion

This chapter concludes the work and presents recommendations for future research. The thesis focused on different aspects of LR computation for biometric evidence evaluation. The concept of LR is applied to forensic face recognition. In the following, we highlight the main contributions of this work by reviewing the research questions posed in chapter 1 and explain how the work carried out answers these questions.

6.1 Answers to the research questions

- In computation of a LR, what is the effect of the sampling variability in the training scores sets? How the commonly proposed LR computation methods are affected by the sampling variability varying the sizes of the training scores sets, the shapes of the distributions of the scores in the training scores sets and the actual value of the score for which the LR is computed?

It is observed that the sampling variability can be significant if small sets of the training scores are used. The three parameters are important and should be considered in order to appropriately select a method of LR computation since all of them affect the sampling variability in the computed LR. Different methods of LR computation are affected differently by the sampling variability. For example, the KDE and the PAV methods are more sensitive to the sizes of the training sets whereas the Log Reg method is more sensitive to the shapes of the distributions of the scores in the training sets. When the

whole range of scores is mapped to LLRs, it has been observed that Log Reg has large sampling variability whereas KDE has the undesirable property that its mapping function from scores to LLRs may not be monotonically increasing. It can also be concluded that a range of LRs should be reported which incorporates the sampling variability.

- What is the effect of using suspect-independent training scores instead of suspect-specific training scores in computation of a LR? Generally, a larger number of training scores are available if suspect-independent training sets are used. However, besides the difference in the sizes of the sets of training scores, the nature of these two different ways (suspect-specific and suspect-independent) to compute the training scores sets implies slightly different interpretations of the prosecution hypothesis (H_p) and the defense hypothesis (H_d). It will be investigated that how much and how frequently the two LRs differ. Furthermore, it will also be investigated that, given the two approaches have the same number of scores in the training sets, are there still variations in the two LRs?

It is observed that there is a significant variation between the two LRs computed using the suspect-specific and the suspect-independent approach. The differences are more prominent in the higher ranges of the LRs and therefore more caution should be taken if one approach is used as an altenative to the other. The two LRs agree on a same verbal equivalents 59.2%, 48.2% and 58.8% of the time for face, fingerprint and speaker recognition systems respectively. This at least shows that it is extremely important to explain how the LR is calculated and what it represents. When equal sizes of the training scores sets in each approach are considered, the variations still exist, however, the large variations in the higher regions are not present.

- What is the current practice of forensic examiners to perform forensic facial comparison and the current state-of-the-art towards automatic forensic face recognition? Furthermore, what is the effective way in which the goal of a (semi-)automatic forensic face recognition can be achieved? What is the discriminating power of different facial features such as eyes, eyebrows, nose, etc?

The way forensic examiners perform forensic facial comparison is different than the mainstream automatic biometric face recognition systems. In forensic comparison, the examiners pay attention to small and regional details such as the number and location of moles and scars in the face, shapes of different facial features and the relative distances among different facial features. Currently, there is no automatic face recognition system which can be used for forensic

facial comparison without human intervention. One of the best strategy towards automation of forensic facial comparison is to build a system which is the combination of different individual classifiers each performing comparison of a facial feature. For this purpose, two existing recognition algorithms are used to perform the recognition task of different individual facial features. It has been observed that eyes and eye brows regions are more discriminating than other facial regions.

- What is the performance of commonly proposed LR computation methods for calibration of existing automatic biometric face recognition systems? Is the conclusion drawn from the assessment tools at the score level such as ROC and EER is significantly different than the conclusion drawn from the assessment tools at the LR level such as C_{llr}?

Based on experiments with 10 baseline face recognition algorithms and three commonly proposed LR computation methods (KDE, Log Reg and PAV), it is observed that the choice of the face recognition algorithm has a significant effect while the choice of the LR computation method has negligible effect on the value of the C_{llr}. The same is true if the assessment is performed using the Tippett plot. When Tippett plots are observed, the two classes of LLRs overlap for the most part, reflecting similar performance of LRs by different LR computation methods for a given face recognition algorithm. This shows the fact that when only the LR computation stage differs, a comparison using Tippett plot and C_{llr} might not be useful in order to choose a method of LR computation. Instead a detailed analysis is required considering factors such as the sizes of the training scores sets and the shapes of the distributions of the scores. In order to study and compare different methods of LR computation considering these factors, a simulation framework is used in chapter 3.

6.2 Final remarks

With the rapid improvement in biometric recognition technology, completely automated forensic-comparison systems seems feasible in the future for several biometric modalities such as face, fingerprint and speech. For a system to be more useful in forensic evidence evaluation, in addition to the robust recognition algorithm, a reliable calibration method is also needed to convert a score to a LR. This was the focus of chapter 3 and chapter 4 considering general aspects of biometric recognition systems while in chapter 5, the discussion was specifically about face recognition systems. The field of forensic face recognition is less mature compared to fingerprint and speaker recognition and

therefore, in chapter 2, we presented a survey on forensic face recognition and a step is taken towards automation of the process of forensic facial comparison in chapter 5.

6.3 Recommendations for future work

Sampling variability in LRs has received less attention in forensic biometric community. In future work more formal study can be carried out in order to derive the functional relationship between the sizes of the training sets and the sampling variability in a computed LR. Based on the analysis presented in chapter 3, modifications can be proposed to make a LR computation method robust. Particularly, the KDE method has the undesirable property that the mapping function from scores to LRs is not monotonically increasing. At least specific modifications are needed in the KDE method to control its behaviour at the outer tails of the two distributions.

The effect of using the suspect-independent training biometric data sets instead of the suspect-specific biometric data sets can be further investigated. The results in chapter 4 are based on using the PAV method to convert scores to LRs. It is interesting to investigate by how much the results and conclusions vary if another method of LR computation is used. Furthermore, a rigorous study of the distributions of the scores of the suspect-specific and suspect-independent approaches might help formulate statistical methods to estimate the suspect-specific distributions of scores from the suspect-independent distributions of scores. An example of a similar study is described in [114]. The proposed method estimates the PDF of the s_p scores from the estimates of the mean and variance of the s_p scores and the PDF of the s_d scores by minimizing the Kullback-Leibler distance.

In chapter 5, a study of the recognition performance of different facial features is presented. A useful extension of this work is to fuse the recognition results of different facial features, either at the score level or at the LR level. The recognition result of each facial features can be weighted based on the recognition performance of a facial feature on a training data set. Apart from this, a more useful study towards automation of forensic facial comparison is to develop automatic detectors for moles, scars, tattoos, etc. To make this process easier, a semi-automatic segmentation can be performed such as based on the eye locations. The study in chapter 5 performs segmentation based on the eye coordinates. Alternatively, for improved recognition results, more manual labeling can be used for segmentation. For example, providing the coordinates

of all of the facial features. This kind of manual labeling will not compromise the usefulness of a forensic face recognition system. This is because the goal is not necessarily a complete automation but also to assist the forensic examiners in their practice.

Bibliography

[1] A. K. Jain, P. Flynn, and A. Ross, *Handbook of biometrics*. Springer, 2007.

[2] C. Aitken, *Statistics and evaluation of evidence for forensic scientists*. John Wiley & Sons, 1997.

[3] D. R. Castro, *Forensic evaluation of the evidence using automatic speaker recognition systems*. PhD thesis, Universidad autónoma de Madrid, 2007.

[4] D. Meuwly, "Forensic individualisation from biometric data," *Science & Justice*, vol. 46, no. 4, pp. 205–213, 2006.

[5] J. Buckleton, "A framework for interpreting evidence," in *Forensic DNA Evidence Interpretation* (S. J. W. J. Buckleton, C.M. Triggs, ed.), pp. 27–63, CRC, Boca Raton, FL.

[6] J. Gonzalez-Rodriguez, J. Fierrez-Aguilar, D. Ramos-Castro, and J. Ortega-Garcia, "Bayesian analysis of fingerprint, face and signature evidences with automatic biometric systems," *Forensic science international*, vol. 155, no. 2, pp. 126–140, 2005.

[7] C. Champod and D. Meuwly, "The inference of identity in forensic speaker recognition," *Speech Communication*, vol. 31, no. 2, pp. 193–203, 2000.

[8] D. J. Balding, *Weight-of-evidence for Forensic DNA Profiles*. John Wiley & Sons, 2005.

[9] D. Ramos and J. Gonzalez-Rodriguez, "Reliable support: Measuring calibration of likelihood ratios," *Forensic science international*, vol. 230, no. 1, pp. 156–169, 2013.

[10] A. B. Hepler, C. P. Saunders, L. J. Davis, and J. Buscaglia, "Score-based likelihood ratios for handwriting evidence," *Forensic science international*, vol. 219, no. 1, pp. 129–140, 2012.

[11] C. M. Bishop *et al.*, *Pattern recognition and machine learning*, vol. 1. springer New York, 2006.

[12] G. S. Morrison, F. Ochoa, and T. Thiruvaran, "Database selection for forensic voice comparison," in *Proceedings of Odyssey*, pp. 62–77, 2012.

[13] I. Alberink, A. Jongh, and C. Rodriguez, "Fingermark evidence evaluation based on automated fingerprint identification system matching scores: The effect of different types of conditioning on likelihood ratios," *Journal of forensic sciences*, vol. 59, no. 1, pp. 70–81, 2014.

[14] D. Ramos-Castro, J. Gonzalez-Rodriguez, A. Montero-Asenjo, and J. Ortega-Garcia, "Suspect-adapted MAP estimation of within-source distributions in generative likelihood ratio estimation," in *Speaker and Language Recognition Workshop, 2006. IEEE Odyssey 2006: The*, pp. 1–5, 2006.

[15] T. Ali, L. Spreeuwers, R. Veldhuis, and D. Meuwly, "Effect of calibration data on forensic likelihood ratio from a face recognition system," in *proceedings of IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems, BTAS*, pp. 1–8, 2013.

[16] D. Ramos, J. Gonzalez-Rodriguez, G. Zadora, and C. Aitken, "Information-theoretical assessment of the performance of likelihood ratio computation methods," *Journal of forensic sciences*, vol. 58, no. 6, pp. 1503–1518, 2013.

[17] N. Brümmer and J. du Preez, "Application-independent evaluation of speaker detection," *Computer Speech & Language*, vol. 20, no. 2, pp. 230–275, 2006.

[18] C. Tippett, V. Emerson, M. Fereday, F. Lawton, A. Richardson, L. Jones, and M. S. Lampert, "The evidential value of the comparison of paint flakes from sources other than vehicles," *Journal of the Forensic Science Society*, vol. 8, no. 2, pp. 61–65, 1968.

[19] http://www.forensicinstitute.nl, "Netherlands forensic institute."

[20] N. F. I. Addendum 1, internal document, "General method of facial comparison."

[21] "Facial identification scientific working group, http://www.fiswg.org/."

[22] "International association for identification, http://www.theiai.org/."

[23] T. Ali, L. Spreeuwers, and R. Veldhuis, "Towards automatic forensic face recognition," in *Informatics Engineering and Information Science*, pp. 47–55, Springer, 2011.

[24] C. Peacock, A. Goode, and A. Brett, "Automatic forensic face recognition from digital images," *Science & Justice*, pp. 29–34, 2004.

[25] T. Ali, P. Tom, J. Fierrez, R. Vera-Rodriguez, L. Spreeuwers, and R. Veldhuis, "A study of identification performance of facial regions from CCTV images," in *5th International Workshop on Computational Forensics, Tsukuba, Japan*, pp. 1–9, 2012.

[26] P. Tome, J. Fierrez, R. Vera-Rodriguez, and D. Ramos, "Identification using face regions: Application and assessment in forensic scenarios," *Forensic science international*, vol. 233, no. 1, pp. 75–83, 2013.

[27] T. Ali, L. Spreeuwers, and R. Veldhuis, "Forensic face recognition: A survey," in *Face Recognition: Methods, Applications and Technology*, Nova Publishers, 2012.

[28] L. Wiskott, J.-M. Fellous, N. Kruger, , and C. von der Malsburg, "Face recognition by elastic bunch graph matching," *IEEE Transactions on PAMI*, pp. 775–779, 1997.

[29] F. Wang, J. Wang, C. Zhang, and J. T. Kwok, "Face recognition using spectral features," *Pattern Recognition*, pp. 2786–2797, 2007.

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

[31] P. J. Phillips, W. T. Scruggs, A. J. O' Toole, P. J. Flynn, K. W. Bowyer, C. L. Schott, and M. Sharpe, "FRVT 2006 and ICE 2006 large-scale results," *National Institute of Standards and Technology*, vol. 7408, 2007.

[32] C. Champod and D. Meuwly, "The inference of identity in forensic speaker recognition," *Speech Communication*, vol. 31, pp. 193–203, 2000.

[33] J. Gonzalez-Rodriguez, J. Fierrez-Aguilar, and J. Ortega-Garcia, "Forensic identification reporting using automatic speaker recognition systems," in *Proc. ICASSP*, 2003.

[34] M. Y. Aulsebrook, W. A. Iscan, and J. H. Slabbert, "Superimposition and reconstruction in forensic facial identification: a survey," *Forensic Science International*, vol. 75, pp. 101–102, 1995.

[35] C. Wilkinson and R. Neave, "Skull re-assembly and the implications for forensic facial reconstruction," *Science & Justice*, vol. 41, no. 3, pp. 233–234, 2001.

[36] "European network of forensic science institutes, http://www.enfsi.eu/."

[37] "Netherlands forensic institute (NFI), http://www.forensicinstitute.nl/."

[38] A. Samal and P. A. Iyengar, "Automatic recognition and analysis of human faces and facial expressions: A survey," *Pattern Recognition*, pp. 65–77, 1992.

[39] N. A. Spaun, "Forensic biometrics from images and video at the federal bureau of investigation," in *Proc. BTAS*, pp. 1–3, 2007.

[40] D. Alspaugh, *A brief history of photogrammetry*, pp. 1–14. 2004.

[41] H. Tuthill and G. George, *Individualization: Principles and Procedures in Criminalistics, 2nd Ed.* Jacksonville, FL: Lightning Powder Company, Inc., 2002.

[42] N. A. Spaun, "Facial comparisons by subject matter experts: Their role in biometrics and their training," in *Proc. ICB*, pp. 161–168, 2009.

[43] D. Meuwly, "Forensic individualization from biometric data," *Science & Justice*, pp. 205–213, 2006.

[44] www.imagemetrics.com, "Image Metrics OptasiaTM operating instructions," 2002.

[45] T. Cootes, G. Edwards, and C. Taylor, "Active appearance models," *IEEE Transactions on PAMI*, pp. 681–685, 2001.

[46] U. Park and A. K. Jain, "Face matching and retrieval using soft biometrics," *IEEE Transactions on Information Forensic and Security (TIFS)*, pp. 406–415, 2010.

[47] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, pp. 71–86, 1991.

[48] h.-s. Cognitec Systems GmbH, "FaceVACS software developer kit."

[49] J. Phillips, H. Wechsler, J. S. Huang, and P. Rauss, "The Feret database and evaluation procedure for face recognition algorithms," *Image and Vision Computing*, pp. 295–306, 1998.

[50] F. Botti, A. Alexander, and A. Drygajlo, "An interpretation framework for the evaluation of evidence in forensic automatic speaker recognition with limited suspect data," in *ODYSSEY04-The Speaker and Language Recognition Workshop*, 2004.

[51] T. Ali, L. Spreeuwers, and R. Veldhuis, "A review of calibration methods for biometric systems in forensic applications," in *33rd WIC Symposium on Information Theory in the Benelux, Netherlands*, pp. 126–133, 2012.

[52] C. G. Aitken, F. Taroni, and J. Wiley, *Statistics and the evaluation of evidence for forensic scientists*, vol. 10. Wiley Online Library, 2004.

[53] G. Morrison, "Forensic voice comparison," in *Expert Evidence* (I. Freckelton and H. Selby, eds.), pp. Ch–99, Thomson Reuters, Sydney, Australia, 2010.

[54] C. Peacock, A. Goode, and A. Brett, "Automatic forensic face recognition from digital images," *Science & Justice*, vol. 44, no. 1, pp. 29–34, 2004.

[55] E. Parzen *et al.*, "On estimation of a probability density function and mode," *Annals of mathematical statistics*, vol. 33, no. 3, pp. 1065–1076, 1962.

[56] B. W. Silverman, *Density estimation for statistics and data analysis*, vol. 26. CRC press, 1986.

[57] D. Meuwly and A. Drygajlo, "Forensic speaker recognition based on a bayesian framework and gaussian mixture modelling (gmm)," in *2001: A Speaker Odyssey-The Speaker Recognition Workshop*, 2001.

[58] A. Agresti, "Building and applying logistic regression models," *Categorical Data Analysis, Second Edition*, pp. 211–266, 2002.

[59] B. Zadrozny and C. Elkan, "Learning and making decisions when costs and probabilities are both unknown," in *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 204–213, ACM, 2001.

[60] B. Zadrozny and C. Elkan, "Transforming classifier scores into accurate multiclass probability estimates," in *Proceedings of the eighth ACM*

SIGKDD international conference on Knowledge discovery and data mining, pp. 694–699, ACM, 2002.

[61] T. Fawcett and A. Niculescu-Mizil, "PAV and the ROC convex hull," *Machine Learning*, vol. 68, no. 1, pp. 97–106, 2007.

[62] M. I. Mandasari, M. McLaren, and D. A. van Leeuwen, "The effect of noise on modern automatic speaker recognition systems," in *Acoustics, Speech and Signal Processing (ICASSP), International Conference on*, pp. 4249–4252, IEEE, 2012.

[63] B. Robertson and G. A. Vignaux, "Interpreting evidence: evaluating forensic science in the courtroom," 1995.

[64] Y. Tang and S. N. Srihari, "Likelihood ratio estimation in forensic identification using similarity and rarity," *Pattern Recognition*, vol. 47, no. 3, pp. 945–958, 2014.

[65] G. Zadora, "Evaluation of evidence value of glass fragments by likelihood ratio and bayesian network approaches," *Analytica chimica acta*, vol. 642, no. 1, pp. 279–290, 2009.

[66] T. Grant, "Quantifying evidence in forensic authorship analysis.," *International Journal of Speech, Language & the Law*, vol. 14, no. 1, 2007.

[67] P. Rose, "Technical forensic speaker recognition: Evaluation, types and testing of evidence," *Computer Speech & Language*, vol. 20, no. 2, pp. 159–191, 2006.

[68] C. Neumann, I. Evett, and J. Skerrett, "Quantifying the weight of evidence from a forensic fingerprint comparison: a new paradigm," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 175, no. 2, pp. 371–415, 2012.

[69] G. S. Morrison, "Measuring the validity and reliability of forensic likelihood-ratio systems," *Science & Justice*, vol. 51, no. 3, pp. 91–98, 2011.

[70] J. Curran, J. Buckleton, C. Triggs, and B. Weir, "Assessing uncertainty in DNA evidence caused by sampling effects," *Science & justice*, vol. 42, no. 1, pp. 29–37, 2002.

[71] G. W. Beecham and B. S. Weir, "Confidence interval of the likelihood ratio associated with mixed stain DNA evidence," *Journal of forensic sciences*, vol. 56, no. 1, pp. 166–171, 2011.

[72] C. Agostinelli and L. Greco, "A weighted strategy to handle likelihood uncertainty in bayesian inference," *Computational Statistics*, vol. 28, no. 1, pp. 319–339, 2013.

[73] D. V. Lindley, *Understanding uncertainty*. John Wiley & Sons, 2013.

[74] G. S. Morrison, C. Zhang, and P. Rose, "An empirical estimate of the precision of likelihood ratios from a forensic-voice-comparison system," *Forensic science international*, vol. 208, no. 1, pp. 59–65, 2011.

[75] A. Alexander, "Forensic automatic speaker recognition using bayesian interpretation and statistical compensation for mismatched conditions," *Swiss Federal Institute of Technology, Lausanne, Switzerland*, 2005.

[76] G. S. Morrison, "A comparison of procedures for the calculation of forensic likelihood ratios from acoustic–phonetic data: Multivariate kernel density (MVKD) versus gaussian mixture model–universal background model (GMM–UBM)," *Speech Communication*, vol. 53, no. 2, pp. 242–256, 2011.

[77] L. J. Davis, C. P. Saunders, A. Hepler, and J. Buscaglia, "Using subsampling to estimate the strength of handwriting evidence via score-based likelihood ratios," *Forensic science international*, vol. 216, no. 1, pp. 146–157, 2012.

[78] G. S. Morrison, "Tutorial on logistic-regression calibration and fusion: Converting a score to a likelihood ratio," *Australian Journal of Forensic Sciences*, vol. 45, no. 2, pp. 173–197, 2013.

[79] R. Chartrand and W. Yin, "Iteratively reweighted algorithms for compressive sensing," in *Acoustics, speech and signal processing, IEEE international conference on*, pp. 3869–3872, IEEE, 2008.

[80] J. Platt *et al.*, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," *Advances in large margin classifiers*, vol. 10, no. 3, pp. 61–74, 1999.

[81] J. Lu, K. N. Plataniotis, A. N. Venetsanopoulos, and S. Z. Li, "Ensemble-based discriminant learning with boosting for face recognition," *Neural Networks, IEEE Transactions on*, vol. 17, no. 1, pp. 166–178, 2006.

[82] "NIST speaker recognition evaluation 2010."

[83] "Cognitec systems gmbh. faceVACS c++ sdk version 8.4.0., 2010."

[84] M. Grgic, K. Delac, and S. Grgic, "SCface–surveillance cameras face database," *Multimedia tools and applications*, vol. 51, no. 3, pp. 863–879, 2011.

[85] D. Lucy, *Introduction to statistics for forensic scientists*. John Wiley & Sons, 2006.

[86] 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," tech. rep., DTIC Document, 1998.

[87] A. Nordgaard and T. Höglund, "Assessment of approximate likelihood ratios from continuous distributions: A case study of digital camera identification," *Journal of forensic sciences*, vol. 56, no. 2, pp. 390–402, 2011.

[88] F. Botti, A. Alexander, and A. Drygajlo, "An interpretation framework for the evaluation of evidence in forensic automatic speaker recognition with limited suspect data," in *ODYSSEY04-The Speaker and Language Recognition Workshop*, 2004.

[89] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek, "Overview of the face recognition grand challenge," in *Computer vision and pattern recognition, 2005. CVPR 2005. IEEE computer society conference on*, vol. 1, pp. 947–954, IEEE, 2005.

[90] D. Meuwly and R. Veldhuis, "Forensic biometrics: From two communities to one discipline," in *In proceedings: Biometrics Special Interest Group (BIOSIG)*, pp. 1–12, IEEE, 2012.

[91] 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.

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

[93] P. N. Belhumeur, J. P. Hespanha, and D. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 19, no. 7, pp. 711–720, 1997.

[94] J. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos, "Boosting linear discriminant analysis for face recognition," in *Proceedings of IEEE ICIP*, vol. 1, pp. I–657, IEEE, 2003.

[95] S. G. Kong, J. Heo, B. R. Abidi, J. Paik, and M. A. Abidi, "Recent advances in visual and infrared face recognition: a review," *Computer Vision and Image Understanding*, vol. 97, no. 1, pp. 103–135, 2005.

[96] http://www.enfsi.eu/, "European network of forensic science institutes."

[97] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*. John Wiley & Sons, 2012.

[98] 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.

[99] S. J. Raudys and A. K. Jain, "Small sample size effects in statistical pattern recognition: Recommendations for practitioners," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 13, no. 3, pp. 252–264, 1991.

[100] S. Lewis, "Philosophy of speaker identification. police applications of speech and tape recording analysis," *Proceeding of the Institute of Acoustics*, vol. 6, no. 1, pp. 69–77, 1984.

[101] P. Viola and M. J. Jones, "Robust real-time face detection," *International journal of computer vision*, vol. 57, no. 2, pp. 137–154, 2004.

[102] K. J. Kirchberg, O. Jesorsky, and R. W. Frischholz, "Genetic model optimization for hausdorff distance-based face localization," in *Biometric Authentication*, pp. 103–111, Springer, 2002.

[103] R. Wallace, M. McLaren, C. McCool, and S. Marcel, "Cross-pollination of normalization techniques from speaker to face authentication using gaussian mixture models," *Information Forensics and Security, IEEE Transactions on*, vol. 7, no. 2, pp. 553–562, 2012.

[104] S. J. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Proceedings of 11th ICCV*, pp. 1–8, IEEE, 2007.

[105] W. Zhang, S. Shan, W. Gao, X. Chen, and H. Zhang, "Local gabor binary pattern histogram sequence (LGBPHS): A novel non-statistical model for face representation and recognition," in *Proceedings of ICCV*, vol. 1, pp. 786–791, IEEE, 2005.

[106] W. Zhao, A. Krishnaswamy, R. Chellappa, D. L. Swets, and J. Weng, "Discriminant analysis of principal components for face recognition," in *Face Recognition*, pp. 73–85, Springer, 1998.

[107] R. Wallace, M. McLaren, C. McCool, and S. Marcel, "Inter-session variability modelling and joint factor analysis for face authentication," in *Biometrics (IJCB), 2011 International Joint Conference on*, pp. 1–8, IEEE, 2011.

[108] L. Wiskott, J.-M. Fellous, N. Kuiger, and C. Von Der Malsburg, "Face recognition by elastic bunch graph matching," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 19, no. 7, pp. 775–779, 1997.

[109] B. Moghaddam, "Principal manifolds and probabilistic subspaces for visual recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 6, pp. 780–788, 2002.

[110] P. J. Phillips, J. R. Beveridge, B. A. Draper, G. Givens, A. J. O'Toole, D. S. Bolme, J. Dunlop, Y. M. Lui, H. Sahibzada, and S. Weimer, "An introduction to the good, the bad, & the ugly face recognition challenge problem," in *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pp. 346–353, IEEE, 2011.

[111] Y. M. Lui, D. Bolme, P. J. Phillips, J. R. Beveridge, and B. A. Draper, "Preliminary studies on the good, the bad, and the ugly face recognition challenge problem," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*, pp. 9–16, IEEE, 2012.

[112] M. Günther, R. Wallace, and S. Marcel, "An open source framework for standardized comparisons of face recognition algorithms," in *Computer Vision–ECCV 2012. Workshops and Demonstrations*, pp. 547–556, Springer, 2012.

[113] A. Anjos, L. E. Shafey, R. Wallace, M. Günther, C. McCool, and S. Marcel, "Bob: a free signal processing and machine learning toolbox for researchers," in *20th ACM Conference on Multimedia Systems (ACMMM), Nara, Japan*, ACM Press, Oct. 2012.

[114] T. Ali, L. Spreeuwers, and R. Veldhuis, "Computation of likelihood ratio from small sample set of within-source variability," in *European Academy*

of Forensic Science conference, European Academy of Forensic Science, 2012.

List of Publications

- Ali, T., Spreeuwers, L.J., Veldhuis, R.N.J., and Meuwly, D., "Quantification of the sampling variability in forensic likelihood-ratio computation from biometric scores", submitted to *IEEE transactions on information forensic and security*.

- Ali, T., Spreeuwers, L.J., Veldhuis, R.N.J., and Meuwly, D., "Biometric evidence evaluation: an empirical assessment of the effect of different training data", submitted to *IET Biometrics*.

- Ali, T., Spreeuwers, L.J., Veldhuis, R.N.J., and Meuwly, D., "Effect of calibration data on forensic likelihood ratio from a face recognition system", In: *IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, 2013, Washington DC, U.S.A., IEEE explore digital library, ISBN 978-1-4799-0527-0.

- Ali, T., Spreeuwers, L.J., and Veldhuis, R.N.J., "A review of calibration methods for biometric systems in forensic applications", In: *33rd WIC Symposium on Information Theory in the Benelux*, 2012, Netherlands, pp. 126-133, ISBN 978-90-365-3383-6.

- Ali, T., Spreeuwers, L.J., and Veldhuis, R.N.J., "Forensic Face Recognition: A Survey", Book chapter in: *Face Recognition: Methods, Applications and Technology*, Nova Publishers, ISBN 978-1-61942-663-4.

- Ali, T., Tom, P., Fierrez, J., Vera-Rodriguez, R., Spreeuwers, L.J., and Veldhuis, R.N.J., "A study of identification performance of facial regions from CCTV images", In: *5th International Workshop on Computational*

Forensics (IWCF), 2012, Tsukuba, Japan.

- Ali, T., Veldhuis, R.N.J., and Spreeuwers, L.J., "Computation of likelihood ratio from small sample set of within-source variability", In: *6th European Academy of Forensic Science Conference (EAFS)*, 2012, The Hague, The Netherlands.

- Ali, T., Spreeuwers, L.J., and Veldhuis, R.N.J., "Towards automatic forensic face recognition", In: *International Conference on Informatics Engineering and Information Science (ICIEIS)*, 2011, Kuala Lumpur, Malaysia, pp. 47-55, Communications in Computer and Information Science 252, Springer Verlag, ISSN 1865-0929.

- Ali, T., Veldhuis, R.N.J., and Spreeuwers, L.J., "Forensic Face Recognition: A Survey", *Technical Report*, TR-CTIT-10-40, Centre for Telematics and Information Technology University of Twente, Enschede, ISSN 1381-3625.