A FRAMEWORK FOR COMBINING ANCILLARY INFORMATION WITH PRIMARY BIOMETRIC TRAITS

By

Yaohui Ding

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Computer Science – Doctor of Philosophy

2018

ABSTRACT

A FRAMEWORK FOR COMBINING ANCILLARY INFORMATION WITH PRIMARY BIOMETRIC TRAITS

By

Yaohui Ding

Biometric systems recognize individuals based on their biological attributes such as faces, fingerprints and iris. However, in several scenarios, additional ancillary information such as the biographic and demographic information of a user (e.g., name, gender, age, ethnicity), or the image quality of the biometric sample, anti-spoofing measurements, etc. may be available. While previous literature has studied the impact of such ancillary information on biometric system performance, there is limited work on systematically incorporating them into the biometric matching framework. In this dissertation, we develop a principled framework to combine ancillary information with biometric match scores.

The incorporation of ancillary information raises several challenges. Firstly, ancillary information such as gender, ethnicity and other demographic attributes lack distinctiveness and can be used to distinguish population groups rather than individuals. Secondly, ancillary information such as image quality and anti-spoof measurements may have different numerical ranges and interpretations. Further, most of the ancillary information cannot be automatically extracted without errors. Even the direct collection of ancillary information from subjects may be susceptible to transcription errors (e.g., errors in entering the data). Thirdly, the relationships between ancillary attributes and biometric traits may not be evident.

In this regard, this dissertation makes three contributions. The first contribution entails the design of a Bayesian Belief Network (BBN) to model the relationship between biometric scores and ancillary factors, and exploiting the ensuing structure in a fusion framework. The ancillary information considered by the network includes image quality and anti-spoof measures. Experiments convey the importance of explicitly incorporating such information

in a biometric system. The second contribution is the design of a Generalized Additive Model (GAM) that uses spline functions to model the correlation between match scores and ancillary attributes, and then learns a transformation function to normalize the match scores prior to fusion. The resulting framework can also be used to *predict in advance* if fusing match scores with certain demographic attributes is beneficial in the context of a specific biometric matcher. Experiments indicate that the proposed method can be used to significantly improve the recognition accuracy of state-of-the-art face matchers. The third contribution is the design of an ensemble of One Class Support Vector Machines (OC-SVMs) to combine multiple anti-spoofing measurements in order to mitigate the concerns associated with the issue of "imbalanced training sets" and "insufficient spoof samples" encountered by conventional anti-spoofing algorithms. In the proposed method, the spoof detection problem is formulated as a one-class problem, where the focus is on modeling a real fingerprint using multiple feature sets. The one-class classifiers corresponding to these multiple feature sets are then combined to generate a single classifier for spoof detection. Experimental results convey the importance of this technique in detecting spoofs made of materials that were not included in the training data.

In summary, this dissertation seeks to advance our understanding of systematically exploiting ancillary information in designing effective biometric recognition systems by developing and evaluating multiple statistical models.

ACKNOWLEDGEMENTS

To be a Ph.D. is the dream of my life, although I did not expect how much I need to pay to chase it. Looking back, it was all worth it.

I am most grateful to my advisor, Prof. Arun Ross. He is not only a great advisor but also a great mentor to me. He has always been supportive of my research and my life. He opened up many opportunities for me and respected my thoughts with great patience. Several times, he was even sitting with me at the front of my desktop, polishing the research paper word-by-word. That is something I can never forget. I appreciate his kindness, compassion and immense knowledge as a person. My sincere thanks also goes to Prof. A.K. Jain, Dr. Xiaoming Liu, and Dr. Yuehua Cui, for their insightful comments and encouragement, but also for the hard questions which incent me to widen my research from various perspectives.

I was very lucky to join the iPRoBe lab and work along with so many self-disciplined but also warm-hearted labmates. We worked together in close proximity every day, which is so convenient to share the joys and frustrations we had. I should put a list here with all the names, and all the stimulating discussions we have had in lab seminars, and all those sleepless nights we worked together, and all the fun we have had in the past few years ... and this list must be as long as the thesis itself. To be necessarily simplified, this thesis would not be possible without the overwhelming kindness and support that they have all given me throughout this journey.

Finally, I thank my entire family who truly understood me, respected my decisions, and supported me spiritually throughout writing this thesis and my life in general.

TABLE OF CONTENTS

LIST OF TABLES viii

LIST OF FIGURES xi

CHAPTER 1 INTRODUCTION 1
1.1 Biometrics and Biometric Fusion 1
1.2 Ancillary Information 3
1.2.1 Demographic Attributes 3
1.2.2 Anti-Spoofing Measurements 6
1.2.3 Sample Quality Assessment 8
1.3 Challenges and Possible Solution 10
1.4 Dissertation Contributions 11
1.5 Notation 14

CHAPTER 2 COMBINING DEMOGRAPHIC ATTRIBUTES WITH BIOMETRIC TRAITS 15
2.1 Background 15
2.2 Related Work 18
2.3 Analytical Investigation on Fusion Schemes 21
2.3.1 Partitioned Score Matrix 21
2.3.2 Formulation of Stratified Matching Scheme 23
2.3.3 Formulation of Decision-Level Fusion Schemes 26
2.3.4 Generalization and Optimization 29
2.4 Additive Model and Extension 31
2.4.1 Additive Model with Interaction 31
2.4.2 Fitting AM via Penalized B-Splines 32
2.4.3 Generalized Additive Model 37
2.5 Experimental Results 40
2.5.1 Databases and Tools 40
2.5.2 Experimental Design 44
2.5.3 Experiment 1. Matching Accuracy 45
2.5.4 Experiment 2. Scalability to Multiple Attributes 47
2.5.5 Experiment 3. Predicting Gain 50
2.5.6 Experiment 4. Robustness to Mislabeling Problem 52
2.6 Summary and Future Work 55

CHAPTER 3 COMBINING ANTI-SPOOFING MEASUREMENTS WITH BIOMETRIC MATCH SCORES 57
3.1 Background 57
3.2 Related Work 59
3.2.1 Feature Extraction for Anti-Spoofing 59

3.2.2 Compromised Templates . . . 61
3.2.3 Performance Evaluation Metrics . . . 63
3.3 Fusion Schemes: Sequential vs. Parallel . . . 67
3.4 Bayesian Belief Networks in Biometrics . . . 70
3.4.1 Existing Bayesian Belief Networks . . . 72
3.4.2 Proposed Bayesian Belief Networks . . . 74
3.5 Databases and Protocol . . . 81
3.6 Experimental Results on LivDet 2009 Database . . . 85
3.7 Experimental Results on LivDet 2011 Database . . . 87
3.7.1 EXP1. Baseline . . . 87
3.7.2 EXP2. Performance Under Zero-Effort Impostors . . . 88
3.7.3 EXP3. Spoof Detection Accuracy . . . 93
3.7.4 EXP4. Overall Recognition Accuracy . . . 95
3.7.5 EXP5. Performance Across Fabrication Materials . . . 97
3.7.6 EXP6. BBN-Based Validation . . . 99
3.8 Summary and Future Work . . . 103

CHAPTER 4 COMBINING ONE-CLASS SVMS FOR ANTI-SPOOFING . . . 106
4.1 Background . . . 106
4.2 An Overview of Image-Based Spoof Detection . . . 109
4.2.1 Feature Extraction for Spoof Detection . . . 110
4.2.2 Availability of Training Data . . . 112
4.2.3 Learning Classifiers . . . 113
4.3 Proposed Ensemble of OC-SVMs Approach . . . 117
4.3.1 Conventional OC-SVM . . . 117
4.3.2 Proposed Ensemble of OC-SVMs . . . 119
4.4 Experimental Results . . . 121
4.4.1 Database and Protocol . . . 121
4.4.2 Conventional B-SVM and OC-SVM . . . 126
4.4.3 Analysis of Proposed Ensemble Strategy . . . 129
4.4.4 Proposed Ensemble of OC-SVMs . . . 130
4.4.5 Validation Using Spoof Samples . . . 132
4.4.6 Performance on Cross-Sensor Training . . . 134
4.5 Summary and Future Work . . . 135

CHAPTER 5 PROPOSED FRAMEWORK FOR COMBINING ANCILLARY INFORMATION WITH BIOMETRIC TRAITS . . . 136
5.1 Background . . . 136
5.2 Related Literature . . . 138
5.2.1 Introduction of Fingerprint Sample Quality . . . 138
5.2.2 Taxonomy on Fusion Frameworks against Spoof Attacks . . . 140
5.3 Experimental Results . . . 143
5.4 Summary and Future Work . . . 146

CHAPTER 6 SUMMARY . . . 149

BIBLIOGRAPHY 152

LIST OF TABLES

Table 1.1: Table of recent work on the automated extraction of demographic information from biometric data. For an elaborate treatment of the subject, see [20]. 5

Table 1.2: Examples of schemes incorporating quality information in the biometric recognition process. This table is not intended to be exhaustive. It merely highlights a few examples of existing studies that use quality measures as ancillary information. 9

Table 2.1: Overview of recent face-based and fingerprint-based gender estimation algorithms using biometric data. Abbreviations used: a Deep Multi-Task Learning (DMTL), Principal Component Analysis (PCA), Support Vector Machines (SVM), Discrete Wavelet Transform (DWT), Convolutional Neural Networks (CNN). 16

Table 2.2: A demonstration of how the demographic labels are distributed in one fold of the 5-fold cross-validation protocol that was executed on the MOR face database. Subjects are organized according to their gender information to retain class balance. 41

Table 2.3: A demonstration of how the demographic labels are distributed in one fold of the 5-fold cross-validation which was performed using the WVU multimodal dataset. In this example, the distribution of race labels is intentionally kept balanced for the two categories (i.e., Caucasian and Non-Caucasian), while the gender distribution may not be balanced at the same time. 42

Table 2.4: A summary of the experimental design in this work. An extensive experiments were carried on three biometric databases with three biometric modalities. The match scores are generated using three commercial biometric matchers. The demographic attributes are labelled by: i) a direct collection (marked as "D") from subjects, ii) a manual annotation (marked as "M"), or iii) a machine learning based gender estimation module from COTS-A (marked as "L"). 43

Table 2.5: The matching accuracy of the proposed GAM fusion scheme on the LFW face database. The true match rates (TMRs) and standard errors are reported under the category of "Image-Restricted, No Outside Data" on the LFW face database. The performance is compared with multiple existing algorithms reported under the same protocol. 45

Table 2.6: The true match rates (TMRs) on Morph face and WVU face databases before and after integrating the gender attribute via the proposed GAM scheme. The match scores are generated using the COTS-B face matcher. The gender label is from: i) a direct collection (marked as "D") from subjects, ii) a manual annotation (marked as "M"), or iii) a built-in gender estimation module in COTS-A with binary outputs (marked as "2L") or 3-level outputs (marked as "3L"). 49

Table 2.7: The P-Values generated from a statistical analysis on interaction effects in the GAM scheme. The highlighted P-Values denote the interaction effects between demographic factors and match scores are significant at the significance level 0.001. 52

Table 2.8: Matching accuracy of the proposed GAM when the demographic labels are incorrect. The proportion of mislabeled data in indicated in the left-most column. 54

Table 3.1: Examples of features that have been proposed for fingerprint spoof detection. A more detailed review can be found in [70]. 60

Table 3.2: Eight possible events during the biometric system operation for a pair of enrolled and input fingerprint image. These events are distinguished by the input state of the pair of fingerprint images, which can be live or spoof, and whether they are from the claimed identity or not. The desirable classification decisions are provided as well. 62

Table 3.3: Number of match scores, liveness scores and quality values corresponding to different states based on 5-fold cross validation. These scores are used for used for training and testing the fusion frameworks against spoof attacks. 85

Table 3.4: Comparison of all the methods from verification, spoof detection and global error perspective (silicone samples) 86

Table 3.5: Comparison of all the methods from verification, spoof detection and global error perspective (gelatin samples) 87

Table 3.6: Spoof detection performance of the various BBN frameworks on the LivDet 2011 database. 102

Table 3.7: Performance of the various frameworks when all eight events are considered for the Biometrika and Italdata sensors. BBN-MLQc is seen to outperform all other frameworks. 102

Table 4.1: Characteristics of the datasets in the *LivDet2011* and *LivDet2013* competition. More details can be found in [135] and [37]. 122

Table 4.2: Establishing the **baseline performance** using conventional binary SVM (B-SVM) and one-class SVM (OC-SVM) using single feature set on the *LivDet2011* dataset. The listed combinations of training materials are only required by the B-SVM classifier, and the rest materials are used as "novel" materials to evaluate the CDR_N of both classifiers. 125

Table 4.3: Performance of the proposed ensemble of OC-SVMs compared to the automatic adaptation approach in [101] and conventional binary SVM (B-SVM) on the *LivDet2011* dataset. The correct detection rates tested on previously known materials (CDR_K) and on novel materials (CDR_N) are reported, respectively. It is notable that except the listed materials for training, the rest materials are tested as "novel materials". 127

Table 4.4: The correct detection accuracy on novel spoof materials (CDR_N) when different combinations of feature sets are used in the proposed ensemble of OC-SVMs (*LivDet2011* dataset). 130

Table 4.5: Performance of the proposed ensemble OC-SVM on the *LivDet2013* dataset. The Top 3 performed algorithms as reported in the competition are listed for a comparison. 134

Table 4.6: Performance of the proposed ensemble OC-SVM on on cross-sensor training. 134

Table 5.1: The spoof detection accuracy of the proposed BBN-AD fusion scheme on the LivDet 2011 fingerprint database. The true detection rates (TDRs) and the false detection rates (FDRs) are compared with two fusion schemes introduced in Chapter 3. Additionally, the accuracy of the original spoof detector is provided as a baseline. 146

Table 5.2: The overall acceptance accuracy of the proposed BBN-AD fusion scheme on the LivDet 2011 fingerprint database. The genuine acceptance rate rate (TDRs) under different overall false acceptance rates (OFARs) are compared with two fusion schemes introduced in Chapter 3. 147

LIST OF FIGURES

Figure 1.1: Illustration of a conventional fingerprint verification system. 2

Figure 1.2: Illustration of the proposed general framework for combining ancillary information with primary biometric traits. The design of a BBN is to model the relationship between ancillary factors and biometric scores. The design of GAM is to learn transformation functions and normalize the scores prior to being combined via the BBN. 4

Figure 1.3: Examples of fake fingerprint images (from *LivDet2011* database [134]) corresponding to the live finger (as the source fingerprint) and four different fabrication materials. (a) Live finger, (b) Latex, (c) EcoFlex, (d) Gelatin and (e) WoodGlue. 7

Figure 1.4: Examples of fingerprint and iris images exhibiting different sample quality values. The quality score of each image is obtained using the IQF freeware developed by MITRE (as seen in Chapter 5). Top row: Fingerprint images whose quality is impacted by different factors. Bottom row: Iris images exhibiting variations in gaze angle that impacts quality. The iris images are from Johnson et al. [50]. 8

Figure 2.1: An example scenario, involving a border control system, where the biometric traits and demographic attributes can be *potentially* combined to improve recognition accuracy. 15

Figure 2.2: Proposed fusion framework for combining demographic attributes with match scores. The raw match scores are transformed via a set of demographic-based score transformation functions which are learned using the proposed Generalized Additive Model during the training phase. The transformed scores are used to verify whether two samples are from the same identity. 17

Figure 2.3: Illustration of the partitioned score matrix from a conventional face matcher. The score matrix is partitioned into four quadrants according to different matching scenarios. For instance, "Q1" denotes the scenario where "Male" probe samples are compared against "Male" gallery samples. 20

Figure 2.4: Illustration of the stratified matching scheme. When the demographic characteristics from two samples are NOT the same, the stratified matching scheme simply rejects the probe sample without computing any match scores. On the other hand, if the characteristics are the same, the match scores from the conventional biometric matcher are used to render the final decision. The stratified matching scheme can be considered as a special case of demographic-based transformation. 24

Figure 2.5: Examples of ROC curves from the stratified matching scheme. The gender information of subjects in the WVU database are integrated with a commercial fingerprint matcher (which will be introduced in Section 2.5). It demonstrates that in order to achieve a consistent FMR for both male and female subjects, the stratified matching scheme requires different thresholds according to each strata. 26

Figure 2.6: An example of ROC curves from the stratified matching scheme, where the gender information of subjects are integrated with two conventional biometric matchers (i.e., Matcher 1 and Matcher 2), respectively. It demonstrates that in order to compare the accuracy of two matchers, the stratified matching scheme still need to exhibit a joint performance rather than the within-cohort performance. 27

Figure 2.7: Illustration of the decision-level fusion scheme. The decision from a demographic label matcher ("Same" or "Not Same") is combined with the decision from a conventional biometric matcher ("Match" or "Non-Match") to render the final decision ("Accept" or "Reject"). It can be considered as a special case of demographic-based transformation, where the match scores are transformed to zero and rejected regardless of the threshold, if demographic labels are "Not Same" for two samples. 28

Figure 2.8: An intuitive example of the transformation functions that can better separate the genuine and impostor score distributions and achieve a higher overall matching accuracy. 37

Figure 2.9: Examples of biometric images in the three datasets used in this work: a) Morph face database, b) LFW face database, and c) WVU multimodal dataset. 39

Figure 2.10: ROC curves before (marked as dashed lines) and after (marked as solid lines) integrating demographic attributes with the match scores generated by the COTS-B face matcher on the Morph face database. For instance, (a) face + gender, and (b) face + race 46

Figure 2.11: ROCs for integrating multiple demographic attributes, simultaneously. The left figure (a) is from the Morph face database, where the match scores are generated using the COTS-B face matcher. The right figure (b) is from the WVU FL1 fingerprint database, where the match scores are generated using the COTS-C fingerprint matcher. 47

Figure 2.12: ROC curves on the WVU face database before (dashed lines) and after (solid lines) integrating the gender labels which are generated using a built-in gender estimation module in COTS-A. In figure (a) and (b), the match scores are generated using the COTS-A and COTS-B face matcher, separately. 50

Figure 2.13: Transformation functions learned from the training set where the match scores from COTS-B are integrated with the gender information automated estimated via the gender estimation module in COTS-A. The automated gender estimation had an error rate around 12.0%. 53

Figure 3.1: Illustration of the fusion framework integrating match scores with quality scores and anti-spoofing measures from two fingerprint samples, and rendering a final accept/reject decision. 58

Figure 3.2: Taxonomy of existing fingerprint anti-spoofing algorithms. 60

Figure 3.3: Example of fake fingerprint images fabrication using latex, ecoflex and woodglue materials and the corresponding LBP-based anti-spoofing measures [85], in the LivDet 2011 database. 61

Figure 3.4: The match score distributions of the LSG and LLG state on samples acquired using **Biometrika** sensor in the LivDet 2011 database. 63

Figure 3.5: ROC Curves of the baseline performance of the fingerprint verification system under zero-effort impostors and spoof attacks. 66

Figure 3.6: Architecture of Method A. Here, the matcher is invoked before the spoof detector. The classifier in the first stage (classifier 1) is used to distinguish genuine from impostor based only on match scores. There are two pairs of classifiers in the spoof detection stage. One pair classifiers (classifier 2 and 3) that are invoked if the input samples are deemed by the matcher to belong to the Genuine (G) class and another pair (classifier 4 and 5) that is invoked if they are deemed to belong to the Impostor (I) class. This arrangement may be redundant (i.e., the use of four different spoof detectors may not be necessary). 68

Figure 3.7: Architecture of Method B. Here, the spoof detector is invoked before the matcher. Depending upon the output of classifier 1 and 2 (LL, LS, SL or SS), one of four classifiers in the verification stage is invoked. For example, classifier 3 operates only on input scores between gallery and probe samples that are both classified as Live, while classifier 6 operates only on scores between gallery and probe samples that are both classified as Spoof. 69

Figure 3.8: Architecture of Method C. Here, the classifier has three inputs: match score, spoof scores of gallery sample and spoof scores of probe sample. All 3 inputs are used simultaneously in order to determine the output class. 69

Figure 3.9: A simple example of Bayesian Network structure 71

Figure 3.10: Several possible BBNs for fusing fingerprint match scores with liveness and quality scores. BBN-MQ and BBN-ML are based on previous literature, while BBN-MLQ and BBN-MLQc are the proposed ones. 74

Figure 3.11: Boxplot of quality scores and probability distribution of the liveness scores for five different materials in the LivDet 2011 database when the **Biometrika** sensor is used. A similar observation can be made for Italdata, Sagem and Digital sensors as well. 80

Figure 3.12: The match score distributions of (a) LSG vs. LLG, (b) SSG vs. LLG, (c) LSI vs. LLG and (d) LLI vs. LLG on samples acquired using **Biometrika** sensor in the LivDet 2011 database. 89

Figure 3.13: ROC Curves of the baseline performance of the fingerprint verification system under zero-effort impostors and spoof attacks. 90

Figure 3.14: Spoof detection performance of the various BBN frameworks on the LivDet 2011 database. Note that the spoof detection accuracy of these frameworks is *not* the same as that of the LBP-based spoof detection algorithm used. This is because the interaction of liveness scores with match score and quality is taken into account when rendering the final decision. The results are from different sensors as: (a) Biometric, (b) Italdata, (c) Sagem and (d) Digital. 91

Figure 3.15: Boxplot of quality values and probability distribution of the liveness score for five different materials in **Biometrika** in the LivDet 2011 database. The same observation is made for Italdata, Sagem and Digital sensors as well. 93

Figure 3.16: Scatter plot and histogram of the liveness scores, before and after adaptation using the transformation function used in BBN-MLQc. It can be noticed that liveness values of the live samples are shifted towards one and those of spoof samples are shifted towards zero, leading to better spoof detection capability of BBN-MLQc over other frameworks. 94

Figure 3.17: Performance of the various frameworks when all eight events are considered for four sensors as (a) Biometrika, (b) Italdata, (c) Digital and (d) Sagem. It can be seen that BBN-MLQc outperforms all other frameworks. 96

Figure 3.18: Evaluation of the BBN-MLQ and BBN-MLQc across fabrication materials trained on only (a) Latex, (b) Gelatin, (c) EcoFlex and (d) Silgum tested on rest other four materials for the **Biometrika** sensor. 98

Figure 3.19: Evaluation of the BBN-MLQ and BBN-MLQc across fabrication materials trained using combination of (a) EcoFlex+ Latex and (b) EcoFlex+ Latex+ Gelatin and tested on rest other three and two materials, respectively, for the **Biometrika** sensor as an example. 99

Figure 3.20: Structural equations associated with the BBN-ML model as an example. 100

Figure 4.1: Schematic of the proposed ensemble framework that uses multiple OC-SVMs. Each OC-SVM utilizes a different set of features. While spoof fingerprints are not necessary for training the OC-SVMs, they are used to refine the decision boundary in the validation phase. 108

Figure 4.2: Proposed categorization for the study of image-based fingerprint spoof detection algorithms. The proposed ensemble OC-SVM classifier falls into the category of SVM-related classifiers that use multiple kinds of features extracted from only the live samples for training. 109

Figure 4.3: Categorization of current existing anti-spoofing approaches. We highlight the textual-based approaches and list several commonly used feature sets that provide comparable spoof detection accuracies. 111

Figure 4.4: Illustration of the support vector data description (SVDD) scheme. The figure on the left shows a simple dataset in the input feature space. The figure on the right shows the data projected to a higher dimensional space using SVM approaches. 114

Figure 4.5: Illustration of the proposed ensemble of OC-SVMs. Multiple OC-SVMs are built based on different feature sets, and their decision boundaries in the projected space are adjusted to minimize the volume of hypersphere that contains the training data. 116

Figure 4.6: The decisions changed by the different combinations of feature spaces that are used in the proposed ensemble of OC-SVMs in the *LivDet2011* dataset. 129

Figure 4.7: Performance of the ensemble OC-SVM after increasing the number of fake samples used in the validation phase. (a) CDR_L, (b) CDR_N and (c) when 200 spoof samples are used in validation phase. Training materials used here are as same as in Table 4.2 and 4.3. 133

Figure 5.1: Illustration of the general fusion framework integrating biometric match scores with ancillary information. It shows that the ancillary information of two samples, such as quality scores and liveness scores, are self-reliant and independent with each other. Meanwhile, only the biometric match scores are corresponding to both samples and the identities they belong. 137

Figure 5.2: Illustration of the fusion framework integrating match scores with quality scores and liveness scores from two fingerprint samples, and rendering a final accept/reject decision. 137

Figure 5.3: Illustration of the proposed general fusion framework. Ancillary information is categorized into direct variables (e.g. liveness scores) and latent variables (e.g. demographic attributes and quality scores), where the direct variables are involved into the BBN scheme as nodes and the latent variables are exploited to normalize the nodes of BBN prior to fusion. 139

Figure 5.4: The quality measures of the fingerprint samples from (a) live finger and fake fingerprints fabricated using (b) latex, (c) gelatin and (d) woodglue, using IQF measurement on the LivDet 2011 database. It can be noticed that quality of the spoof vary across fake fabrication materials. 140

Figure 5.5: Taxonomy of existing fusion frameworks incorporating match scores, liveness scores and image quality. 141

Figure 5.6: The performance of Spoof detection before and after updating the liveness scores via the GAM framework. The quality scores are used as the covariate of GAM. The samples are fabricated using a) silicone material and b) gelatin material. 144

Figure 5.7: The performance of biometric matching system before and after updating the match scores via the GAM framework. The quality scores are used as the covariate of GAM. The samples are fabricated using a) silicone material and b) gelatin material. 144

Figure 5.8: Example of spoof images in the LivDet 2009 (a-b) and 2013 (c-d) databases fabricated using consensual and non-consensual methods, respectively. These spoofs are acquired using Biometrika sensor. Note that the spoof images are either of very low quality (a-b) or partial (c-d). 145

CHAPTER 1

INTRODUCTION

1.1 Biometrics and Biometric Fusion

Biometrics is the science of recognizing individuals based on their physical (such as face, fingerprint, iris) and behavioral (such as speech and gait) traits [48]. A conventional biometric system can be viewed as an automatic pattern matching system that acquires biometric data from an individual (e.g. a fingerprint) using a sensor, extracts a set of discriminatory features from this data (e.g. minutia points), compares the extracted feature set with those stored in a database (referred to as a template), and results in a score indicating the similarity between the two feature sets [46]. This assessment of the similarity of the feature sets, referred to as a match score, may then be used to recognize the individual. Figure 1.1 illustrates such a process in the context of fingerprint verification.

Biometric systems that operate using a single biometric trait for human recognition are called unimodal biometric systems. Due to the diverse nature of biometric applications (e.g., ranging from mobile phone unlocking to international border crossing), no single biometric trait is likely to be optimal and satisfy the requirements of all applications [108]. Some of the limitations of a unimodal biometric system, such as noisy sensor data, non-universality of traits, lack of distinctiveness of traits and unacceptable error rates, can be alleviated by fusing multiple pieces of evidence from the same subject [47]. This kind of information fusion procedure, also referred to as the biometric fusion, typically can increase population coverage, provide better recognition accuracy compared to unimodal biometric systems, and meet the stringent performance requirements imposed by various applications [131, 81].

There are various sources of information that can be involved in a biometric fusion procedure. This information may be obtained from multiple biometric traits (e.g., face and fingerprint), or from the same biometric trait but with multiple representations (e.g., two

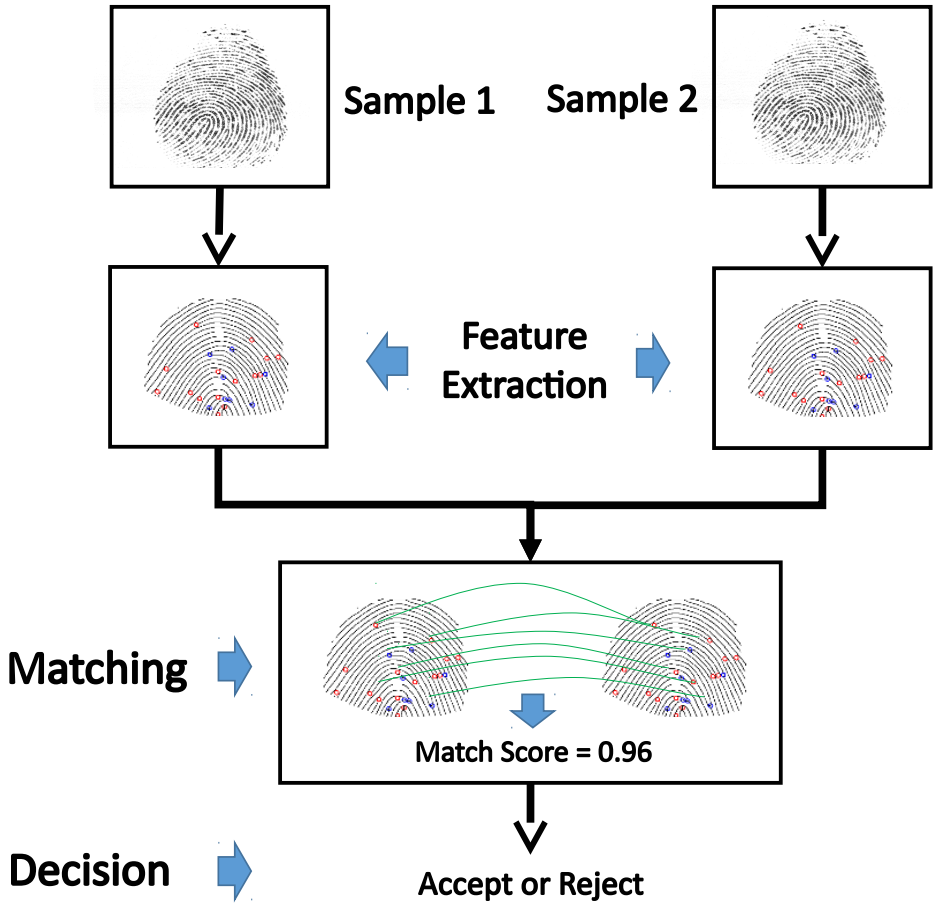


Figure 1.1: Illustration of a conventional fingerprint verification system.

facial images of an individual obtained at different pose angles), or multiple algorithms (e.g., two feature sets extracted using a texture-based algorithm and a minutiae-based algorithm, respectively). Ross et al. [108] described several major factors that impact the design and structure of a biometric fusion system as follows:

1. number of sensors deployed;
2. time taken to acquire the biometric data from users;
3. storage requirements;
4. processing time of the added algorithm;

5. perceived inconvenience experienced by the user.

In light of the listed factors, most of the existing fusion scenarios require additional sensors and storage space to process the additional biometric data acquired from the users, which may increase the response time of the system and decrease its usability.

Recently, another class of biometric fusion problems, that combine the ancillary information (such as the image quality of biometric samples, the gender of users, anti-spoofing measures, etc.) with primary biometric traits (e.g., fingerprints, facial images, etc.), has gained increasing attention. For instance, demographic attributes (such as age, gender and ethnicity) extracted from biometric data (as shown in Table 1.1) can be subsequently used to improve the recognition accuracy of primary biometric traits [45, 110]. Recent research has also sought to improve the resilience of biometric verification systems to spoof attacks by combining match scores with both anti-spoofing measurements and image quality [68, 23]. To contrast with the term **primary biometric traits**, the term **ancillary information** is used to indicate the fact that the ancillary information in themselves may not be suitable for human recognition, but can be judiciously used to improve recognition accuracy.

We focus on three categories of ancillary information in this dissertation: demographic attributes, the image quality of biometric samples and anti-spoofing measurements. Figure 1.2 provides an illustration of the proposed general framework for combining ancillary information with primary biometric traits. The challenge in incorporating such ancillary information into biometric matching framework are discussed in the following sections, respectively.

1.2 Ancillary Information

1.2.1 Demographic Attributes

In some biometric applications, several descriptive attributes of a user (such as gender, age, ethnicity, etc.) are requested at the time of enrollment and stored in the database along with the biometric data. For example, a border control biometric database may contain

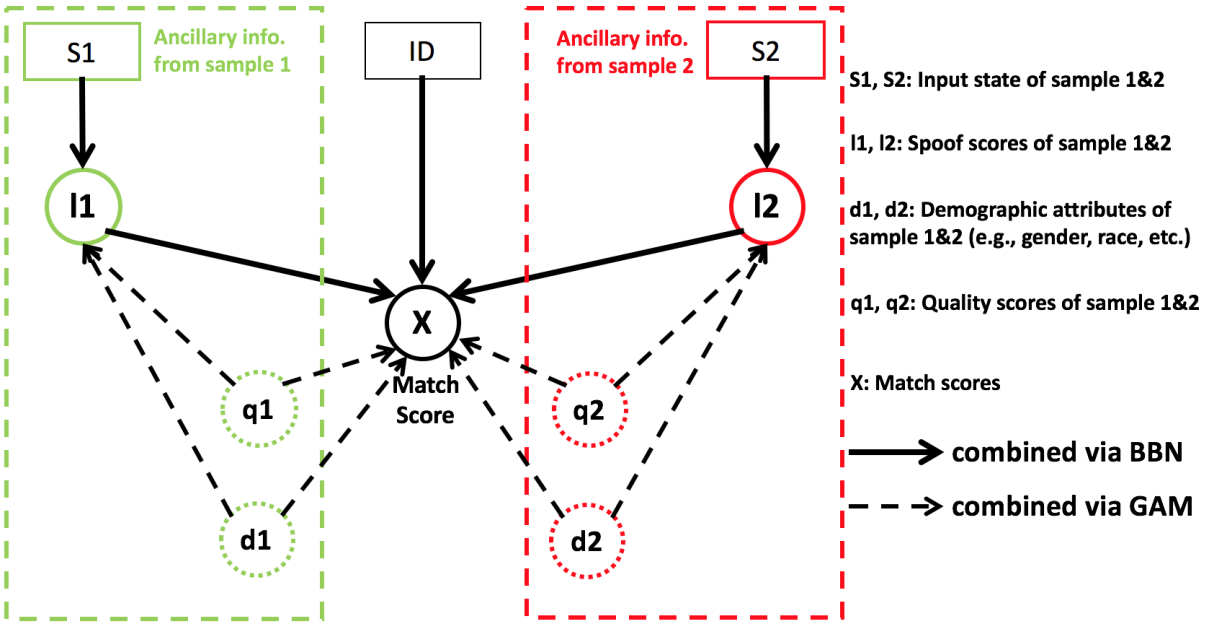


Figure 1.2: Illustration of the proposed general framework for combining ancillary information with primary biometric traits. The design of a BBN is to model the relationship between ancillary factors and biometric scores. The design of GAM is to learn transformation functions and normalize the scores prior to being combined via the BBN.

information such as the gender and ethnicity of users besides their fingerprints and facial images. These descriptive attributes are referred to as demographic attributes (as defined in [20]), which denotes the quantifiable characteristics of a population group. In some cases, it may be possible to automatically extract demographic attributes from the biometric data. For example, recent research has shown that a number of demographic attributes - sometimes referred to as soft biometrics - can be gleaned from biometric data using automated machine learning schemes (see Table 1.1). This raises the question of whether demographic attributes can be effectively combined with biometric match scores in order to improve the recognition accuracy of a system.

Several inherent characteristics of demographic attributes impact the design and structure of the fusion framework. First, most demographic attributes only contain a few discrete labels. For example, the gender attribute is usually considered as a binary variable with two labels: "male" and "female". The limited distinctiveness of the gender attribute means it

Table 1.1: Table of recent work on the automated extraction of demographic information from biometric data. For an elaborate treatment of the subject, see [20].

Biometric Traits	**Demographic Attributes**	**Authors**
Face	Gender&Race&Age, etc.	Han et al. 2017 [41]
Face	Gender&Race&Age, etc.	Liu et al. 2015 [62]
Face	Gender and Age	Bekios-Calfa et al. 2014 [6]
Face	Ethnicity	Fu et al. 2014 [31]
Face	Age&Others	Yi et al. 2014 [137]
Fingerprint	Gender	Rattani et al. 2014 [99]
Audio	Gender	El Shafey et al. 2014 [27]
Iris	Gender and Race	Lagree and Bowyer 2011 [56]
Iris	Age	Amanda et al. 2013 [114]
Periocular	Gender	Bobeldyk and Ross 2016 [8]
Face and Gait	Gender	Ng et al. 2012 [82]
Face and Finger	Gender	Li et al. 2010 [60]

unlikely to be useful for human recognition by itself. Moreover, the automatic extraction of demographic attributes may be less reliable compared to that of commonly used biometric traits. For example, the age of a subject cannot be precisely estimated by most of the existing age estimation algorithms [114, 32]. Furthermore, the demographic information directly collected from subjects may contain transcription errors (e.g. an error in entering the gender of a subject).

While previous literature has studied the impact of demographic factors on recognition performance (e.g., [52, 30, 40, 87, 91]), there is limited work on systematically incorporating them into the biometric matching framework. In fact, most of the current approaches primarily use demographic data in the context of biometric *identificaton*, by restricting the search to only those identities in the database having the same demographic characteristics as the input query. This is often referred to as biometric indexing or database filtering. For example, if the input face image is deemed to be a "Male Asian", then a face recognition system would constrain its search to only those identities in the database that are labeled as "Male" and "Asia". However, such an approach has one major problem: if the demographic

attribute is mislabeled, then it is possible for the input image to be never compared against the correct identity. This observation demonstrates the need for a framework that accounts for missing or inaccurate demographic labels when combining demographic attributes with biometric traits.

1.2.2 Anti-Spoofing Measurements

A presentation attack occurs when an attacker presents a fake or modified biometric trait to the sensor [113, 128]. For instance, it has been shown that some fingerprint systems can be fooled by using a finger-like object fabricated using easily available materials such as latex, glue and gelatin (as shown in Figure 1.3), with the fingerprint ridges of another person inscribed on it [73]. Fake biometric traits can also be used during the enrollment stage, especially in mobile applications where the enrollment process is not carefully monitored [23].

Spoofing is an example of a presentation attack, where the adversary uses a fake or altered biometric trait with the intention of masquerading as another individual [113]. Spoof detection refers to the ability of a system to correctly distinguish between a legitimate, live human biometric presentation and spoof artifacts [129]. An *anti-spoofing measure*, as the output of most anti-spoofing schemes discussed in the literature, is a numerical value indicating the probability that the input biometric sample corresponds to a live human biometric presentation (i.e., *liveness value*) or a spoof artifact (i.e., *spoof score*) [113]. In this thesis, the *spoof score*, which indicates how likely a biometric sample is to be a spoof, is preferred. The various anti-spoofing approaches proposed in the literature can be broadly classified into hardware-based and image-based solutions [70, 71]. Image-based spoof detection algorithms have the advantage over hardware-based systems of being (1) less expensive (as no extra device is needed) and (2) less intrusive for the user [70, 79].

Take fingerprint anti-spoofing as an example. Existing fingerprint anti-spoofing algorithms extract texture-based features [84, 36], anatomical features [28, 72] or physiological

Figure 1.3: Examples of fake fingerprint images (from *LivDet2011* database [134]) corresponding to the live finger (as the source fingerprint) and four different fabrication materials. (a) Live finger, (b) Latex, (c) EcoFlex, (d) Gelatin and (e) WoodGlue.

features [34, 67] from a fingerprint image (or sequence of images), and then train a binary classifier (such as a Support Vector Machine (SVM)) to distinguish between the features of "Live" and "Spoof" samples.

Most researchers [68, 113, 129] conjecture that the problem of confirming a live sample is a harder problem than that of deciding whether two samples are from the same identity. One of the main reasons is that as spoof attacks evolve, it is likely that new and more sophisticated materials and techniques will be used to create fake fingerprints thereby undermining existing learning-based anti-spoofing approaches (see Figure 1.3). In order to alleviate some of these concerns, this thesis proposes a One-Class Classification (OCC) approach that predominantly uses training samples from only a single class, i.e., the "live" class, to generate a hypersphere that encompasses most of the live samples and excludes all kinds of spoofs.

Anti-spoofing methods are designed to be incorporated into biometric systems in order to increase system security [70, 68]. This thesis proposes a novel fusion framework in which anti-spoofing algorithms are incorporated into conventional biometric systems using a Bayesian Belief Network (BBN) framework. Additionally, the fusion framework is extended by incorporating image quality, another ancillary attribute which is impacted by the choice of fabrication materials used, to further improve anti-spoofing performance.

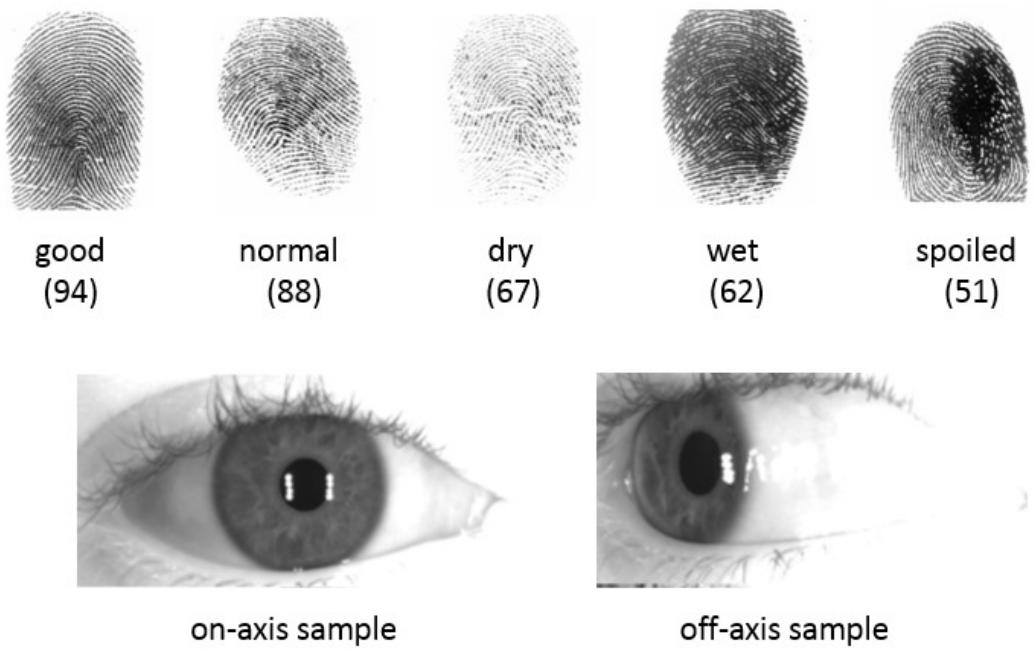


Figure 1.4: Examples of fingerprint and iris images exhibiting different sample quality values. The quality score of each image is obtained using the IQF freeware developed by MITRE (as seen in Chapter 5). Top row: Fingerprint images whose quality is impacted by different factors. Bottom row: Iris images exhibiting variations in gaze angle that impacts quality. The iris images are from Johnson et al. [50].

1.2.3 Sample Quality Assessment

The literature has shown that biometric recognition is affected by many factors related to the quality of biometric samples [127, 94, 80]. First, the acquisition process of biometric samples can be negatively affected by noise in the acquisition sensor. An example would be noisy fingerprints due to a malfunctioning sensor. Second, the inconsistent interaction between the human and sensing device can lead to the acquisition of suboptimal data. For example, the application of excess pressure on a fingerprint sensor may result in non-linear deformation of the ridges; similarly, turning one's head away from an iris camera can result in off-axis iris images. Third, environmental factors, such as dry fingerprints during the winter season, can lead to difficulty obtaining good quality samples (as seen in Figure 1.4). Consequently, these factors lead to a decrease in the quality of the biometric sample, and hence, compromise the

Table 1.2: Examples of schemes incorporating quality information in the biometric recognition process. This table is not intended to be exhaustive. It merely highlights a few examples of existing studies that use quality measures as ancillary information.

Authors	Main Contributions
Chen et al. [15]	Considered quality scores as **a predictor variable** in fingerprint matching performance
Wein and Baveja [132]	Improved the identification accuracy of the fingerprint system in U.S. VISIT program using **quality-dependent thresholds**
Fierrez-Aguilar et al. [92]	Implemented quality scores as **a weighting factor** in their multi-algorithm fingerprint score-level fusion
Nandakumar et al. [80]	Incorporated quality scores with match scores from fingerprint and face via **a GMM-based scheme**
Fierrez-Aguilar et al. [29]	Incorporated quality scores with match scores from fingerprint and voice via **a SVM-based method**
Maurer and Baker [74]	Proposed **a graphical model** to combine quality with match scores from fingerprint and voice
Abaza and Ross [1]	Proposed **a rank-level fusion scheme** to incorporate quality scores with match scores from fingerprint and face
Kryszczuk et al. [53]	Evaluated the impact of image quality **as a specific feature** in fingerprint recognition
Poh et al. [93]	Proposed **a Bayesian framework** for incorporating device-specific quality scores with match scores from fingerprint and face
Poh and Kittler [95]	Proposed **a Bayesian framework** for incorporating quality scores with match scores from fingerprint and face.

performance of the automated biometric system. In order to reduce the adverse effect of the above factors, recent research has explored the possibility of incorporating biometric sample quality in the biometric decision process (see Table 1.2).

As discussed by Poh and Kittler [95], the biometric sample quality can be defined in various ways, viz., i) the degree of extractability of the features used for recognition [132], ii) the degree of conformance of biometric samples to some predefined criteria known to influence the recognition performance [15, 39], and iii) the degree of richness of texture and other image characteristics, e.g., the sharpness, contrast, and detailed rendition of the image [15, 86]. Take the minutiae-based fingerprint matcher as an example. A fingerprint is deemed

to be of high quality if it contains sufficient number of reliable minutiae points that can be used by the automated matcher. This criterion may be different from the human perception of image quality, where high quality may indicate a fingerprint with clear ridges, low noise, and good contrast.

Quality Scores are commonly used for indicating how good the quality of a biometric sample is. These scores could be numerical values or categorical values depending on the definitions and metrics that are used. The lack of a uniform standard requires the design of a fusion framework that is resilient to inaccurate or uncertain quality measures when integrating them with biometric match scores.

1.3 Challenges and Possible Solution

As discussed in the previous section, the main challenges in combining ancillary information with biometric traits can be summarized as follows:

- **Lack of distinctiveness:** Most ancillary attributes, such as gender and ethnicity, can only distinguish population groups rather than individuals. As a result, utilizing ancillary information is not guaranteed to benefit the recognition performance. Therefore, the fusion framework should be able to predict in advance if fusing certain attributes with biometric match scores is beneficial with respect to a particular biometric matcher.

- **Lack of reliability:** Ancillary information is not as reliable as primary biometric traits. Automated extraction algorithms of ancillary information, such as anti-spoofing measurements and image quality, can have a large degree of uncertainty. Even the direct collection of demographic information, such as gender, from subjects may be susceptible to transcription errors. This lack of reliability of ancillary information will affect biometric matching accuracy if the fusion framework does not adequately account for such types of uncertainty.

- **Lack of unity and consistency:** Measurements of ancillary information are application-

specific and may have different numerical ranges and interpretations. For example, both the ANSI/NIST standard and the Electronic Fingerprint Transmission Specification (EFTS) lack any metrics or standards for image quality (as reported in [77]). Thus, the fusion framework should be able to accommodate diverse types of inputs.

- **Implicit relationship between ancillary information and match scores:** The relationship between match scores and ancillary information has not been systematically studied in the literature, primarily due to the lack of large datasets containing the relevant ancillary labels. Consequently, the assumption of independence between ancillary variables and biometric match scores may be presumptuous. One possible solution is to assume causal relationships based on domain knowledge, and then carefully validate the assumptions based on actual data.

1.4 Dissertation Contributions

This purpose of this thesis is to devise a principled framework to effectively combine ancillary information with primary biometric traits by addressing the aforementioned challenges. The main contributions of this thesis are summarized below:

- The primary purpose for combining demographic and biometric attributes is to improve biometric matching accuracy. In order to facilitate this, we first investigate existing attribute-based fusion schemes (here, the term “attributes" refers to ancillary information). This investigation inspired us to pose the problem as an exploration of optimal *transformation functions* on match scores based on ancillary measures that can maximize the matching accuracy. Based on this formulation, the rationale of several commonly used fusion schemes, such as the attribute-based indexing and decision-level fusion, are explained from both an intuitive and a mathematical perspective.

- We design a Generalized Additive Model (GAM) that uses spline functions to model the relation between match scores and demographic attributes resulting in a consistently

higher matching accuracy compared to other fusion schemes. It learns the optimal parameters of transformation functions. These model parameters can be used to predict in advance whether fusing demographic data with a certain biometric matcher is beneficial or not. Moreover, the proposed model is shown to be effective even in situations where the demographic data are incorrect or unreliable to some extent.

- We design a method to combine anti-spoofing measures with biometric match scores. In this regard, we employ a Bayesian Belief Network (BBN) that specifies the *relationship* between anti-spoofing measures and biometric match scores via a causal assumption. Further, the role of ancillary information on matching accuracy is carefully restricted via a conditional independence assumption. Experimental results demonstrate that the proposed BBN configuration can provide consistently better overall recognition performance than typical classifiers, such as naive Bayes, decision tree and neural networks.

- We propose the design of an ensemble of multiple One-Class SVM (OC-SVM) classifiers to address the problem of developing more *generalizable algorithms* for anti-spoofing. Experimental results on two public-domain LivDet datasets (2011 and 2013) demonstrate that the proposed ensemble approach can achieve competitive accuracy by predominately using training samples from only a single class, i.e., the live class. Several drawbacks of the single OC-SVM classifier are successfully overcome by the aggregation of decision boundaries from multiple independent OC-SVMs corresponding to different feature spaces. The proposed one-class classifier mitigates the concerns associated with the issue of "imbalanced training set" and "insufficient spoof samples" encountered by conventional anti-spoofing algorithms.

- Finally, this thesis proposes a general principled framework which is suitable for combining different types of ancillary information with biometric match scores. We use the quality score as an example to demonstrate that both the proposed BBN and

GAM scheme can be effectively extended to involve additional ancillary factors. Then, several different extensions of the simple BBN are compared, and the results show the advantage of utilizing a simple BBN configuration in which the match scores are updated via the GAM transformation functions.

The thesis is organized as follows: Chapter 2 introduces a Generalized Additive Model to model the correlation between match scores and demographic attributes, and normalize the match scores resulting in better verification performance. Chapter 3 compares the sequential and parallel scheme of combining anti-spoofing measures with match scores, and then presents the design of a Bayesian Belief Network to improve the overall recognition performance. Chapter 4 proposes an ensemble of one-class classifiers to improve spoof detection accuracy. Chapter 5 proposes a general fusion framework which can effectively combine different types of ancillary information with primary biometric traits. Chapter 6 summarizes the findings of this research work and outlines ideas for future research.

1.5 Notation

We use the following notation and symbols throughout this thesis.

X	:	Matrix of match scores
$x_i \quad i = 1, \ldots, n$	:	The original match score from the ith match
Y	:	Scores after fusion or transformation
$y_i \quad i = 1, \ldots, n$	:	The transformed match score from the ith match
$y_i' \quad i = 1, \ldots, n$	:	The decision of "Accept" or "Reject" of the ith match
δ	:	Operating thresholds
$\sum \mathbb{I}(x > \delta)$	:	The number of match scores that are greater than threshold δ
$x \in \mathbf{X}_{Q1}^{Gen}$	:	The genuine match scores in quadrant 1
Z	:	A coded demographic factor
$\mathbf{F}$	:	A set of transformation functions on original match scores
$f_{z_i}(x_i) \quad z_i = 1, \ldots, L$	:	The score transformation function corresponding to the z_i level of a demographic factor
$\mathbf{S}_t \quad t = \{1, 2\}$	:	Input states of sample 1 or 2
$\mathbf{a}^t \quad t = \{1, 2\}$	:	All the ancillary information of sample 1 or 2
$l_t \quad t = \{1, 2\}$	:	The spoof scores of sample 1 or 2
$q_t \quad t = \{1, 2\}$	:	The quality scores of sample 1 and 2
$Pr(x)$	:	The probability of a random variable x
α_0	:	Total interception
γ	:	Coefficients of interactions
ϵ	:	Residuals of the model

CHAPTER 2

COMBINING DEMOGRAPHIC ATTRIBUTES WITH BIOMETRIC TRAITS

2.1 Background

In some biometric applications, demographic attributes of a user (such as gender, age, ethnicity, etc.) are requested at the time of enrollment and stored in the database along with the biometric data. For example, a border control biometric database may contain information such as the name, gender, date of birth, nationality and ethnicity of subjects besides their facial images or fingerprints (see Figure 2.1). Further, recent research has established that demographic attributes - sometimes referred to as *soft biometrics* - can be deduced from biometric data using automated machine learning schemes [20]. Table 2.1 provides an overview of recent face-based and fingerprint-based gender estimation algorithms.

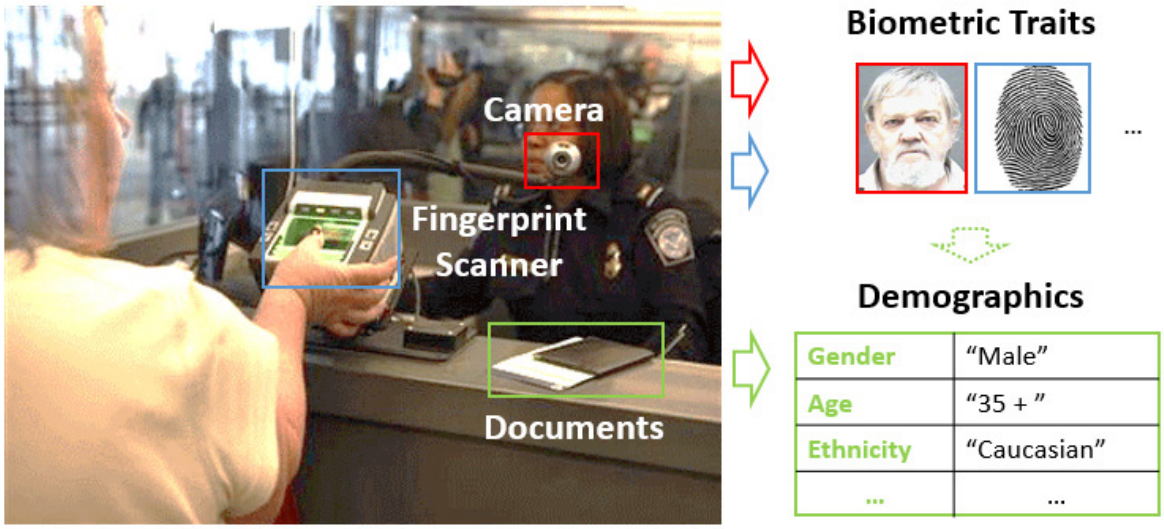


Figure 2.1: An example scenario, involving a border control system, where the biometric traits and demographic attributes can be *potentially* combined to improve recognition accuracy.

Table 2.1: Overview of recent face-based and fingerprint-based gender estimation algorithms using biometric data. Abbreviations used: a Deep Multi-Task Learning (DMTL), Principal Component Analysis (PCA), Support Vector Machines (SVM), Discrete Wavelet Transform (DWT), Convolutional Neural Networks (CNN).

Authors	**Classifiers**	**Performance**
Han et al. 2017 [41]	DMTL	95.45% on 62,566 face images
Liu et al. 2015 [62]	CNN	87.40% on 202,599 face images
Bekios-Calfa et al. 2014 [6]	PCA	88.04% on 337 face images
Yi et al. 2014 [137]	CNN	98.10% on 62,566 face images
Shan 2012 [115]	SVM	94.81% on 7,443 face images
Jia and cristianini 2015 [49]	C-Prgasos	96.86% on 4 million face images
Castrillón-Santana et al. 2017 [12]	CNN	94.20% on 28,220 face images
Gnanasivam and Muttan 2012 [38]	DWT	88.28% on 3,570 fingerprints
Rattani et al. 2014 [99]	SVM	71.70% on 948 fingerprints
Marasco et al. 2014 [69]	PCA	88.70% on 494 fingerprints

This raises the question of how such demographic attributes can be effectively combined with primary biometric traits for improving the recognition accuracy of the system.

In this chapter, we approach the problem of systematically combining demographic attributes with biometric match scores in a fusion framework. The proposed fusion scheme combines demographic data with biometric match scores via a Generalized Additive Model (GAM) that is applicable to the biometric verification scenario. The proposed GAM learns a set of penalized spline-based transformation functions that describe the relationship between match scores and demographic factors (as seen in Figure 2.2). The proposed framework has several advantages:

1. The model parameters obtained during the training phase can be used to *predict in advance* whether fusing demographic data with a certain biometric matcher is beneficial or not (section 2.5.5).

2. The proposed framework results in better verification accuracy than existing methods for combining demographic attributes with match scores (section 2.5.3).

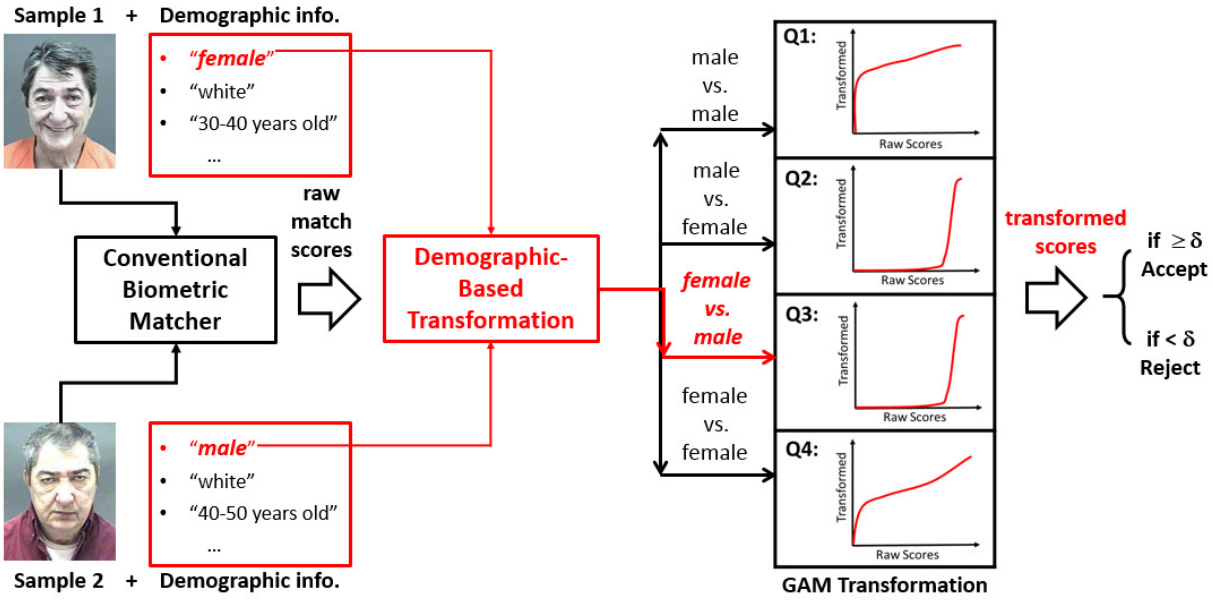


Figure 2.2: Proposed fusion framework for combining demographic attributes with match scores. The raw match scores are transformed via a set of demographic-based score transformation functions which are learned using the proposed Generalized Additive Model during the training phase. The transformed scores are used to verify whether two samples are from the same identity.

3. The proposed model is shown to be effective even in scenarios where the demographic labels are incorrect or unreliable (section 2.5.6).

This chapter is organized as follows: Section 2.2 briefly discuss several commonly used combining schemes to integrate demographic data with the biometric matching framework. Section 2.3 explains the rationale for formulating this problem as an optimization of score transformation functions. Section 2.4 introduces the theory of Generalized Additive Model (GAM) and its extensions. Section 2.5 presents the advantages of the proposed fusion framework via experimental results conducted on multiple datasets. Section 2.6 summarizes the findings.

2.2 Related Work

The impact of demographic factors on recognition performance has been studied in the literature (e.g., [52, 30, 40, 87, 91]). These studies have shown that certain demographic cohorts are more susceptible to errors in the biometric matching process. For example, Klare et al. [52] pointed out that multiple face recognition algorithms consistently have lower matching accuracies on the same cohorts (Females, Blacks, and age group $18 - 30$). However, there is limited work that has been conducted on systematically incorporating demographic data into the biometric matching framework.

In the context of biometric *identification*, demographic data can be simply utilized as index values to restrict the search to only those identities in the database having the same demographic characteristics as the probe sample. However, as stated earlier, such an approach is heavily impacted by the mislabeling problem, where the probe image will never be compared against the correct identity. In the context of biometric *verification*, if the demographic characteristics from the probe sample and the claimed template are different, a conventional biometric system is likely to simply reject this probe without computing a match score. This scheme of integrating demographic data in a biometric system is referred to as the *stratified matching scheme* in this work, because it first partitions the biometric samples into multiple strata according to their demographic characteristics (the Male and Female strata, the Caucasian and Non-Caucasian strata, etc.) prior to the matching process. As investigated later in this paper, the stratified matching scheme cannot significantly increase the verification accuracy, especially when the demographic labels are erroneous.

Another way to combine biometric and demographic data is by utilizing decision-level fusion schemes. Decision-level fusion schemes first make a decision on whether the demographic labels of two samples are same. Then, this decision is merged with the decision that is independently rendered by a conventional biometric matcher. The final decision can be obtained by employing techniques like majority voting, or the logical AND/OR operators [108, 14]. However, these fusion schemes can still be heavily impacted by the mislabeling

problem.

Feature-level fusion [107] is another viable way of combining biometric and demographic data. Feature-level fusion schemes involve the concatenation of feature sets used for predicting demography (e.g., a texture-based feature set that is used for estimating the gender from irides) with the feature sets used in conventional biometric matchers (e.g., IrisCode used for iris recognition). Lu and Jain [63] proposed a face matching algorithm where the feature set used for ethnicity estimation was incorporated into a conventional face matcher. However, one of the challenges in such an approach is the low compatibility between feature sets, since this design heavily relies on the *nature* of feature sets used for biometric matching and demographic prediction. For example, reconciling minutiae points (used for fingerprint recognition) and BSIF-based feature vectors (used for gender prediction) may not be easily possible. Consequently, the generalizability of feature-level fusion schemes across different feature sets is limited. Moreover, feature-level fusion schemes require access to feature sets used by the biometric matcher as well as the demographic predictor, which are typically viewed as proprietary information and are, therefore, not easily accessible.

It must be mentioned here that other types of soft biometric attributes, besides demographic labels, have been successfully incorporated in biometric systems. For example, anthropometric attributes such as body height and face geometry, that are used in forensics, can be leveraged for use in a biometric system. As noted by Nixon in [90], a judicious combination of these attributes can result in a relatively high degree of distinctiveness for face recognition. Ramanathan and Wechsler [97] combined two appearance-based approaches (PCA and LDA) with anthropometric/geometric measurements (19 manually extracted geometric measurements of the head and shoulders) via a neural network, and the proposed algorithm was robust to occluded and disguised faces. Biographic information, such as name and address, has been utilized for the identity de-duplication of biometric databases [118]. As a summary, Dantcheva et al. [20] introduced a taxonomy of methods for utilizing these soft biometric information, which include biographics, anthropometrics and so on, in

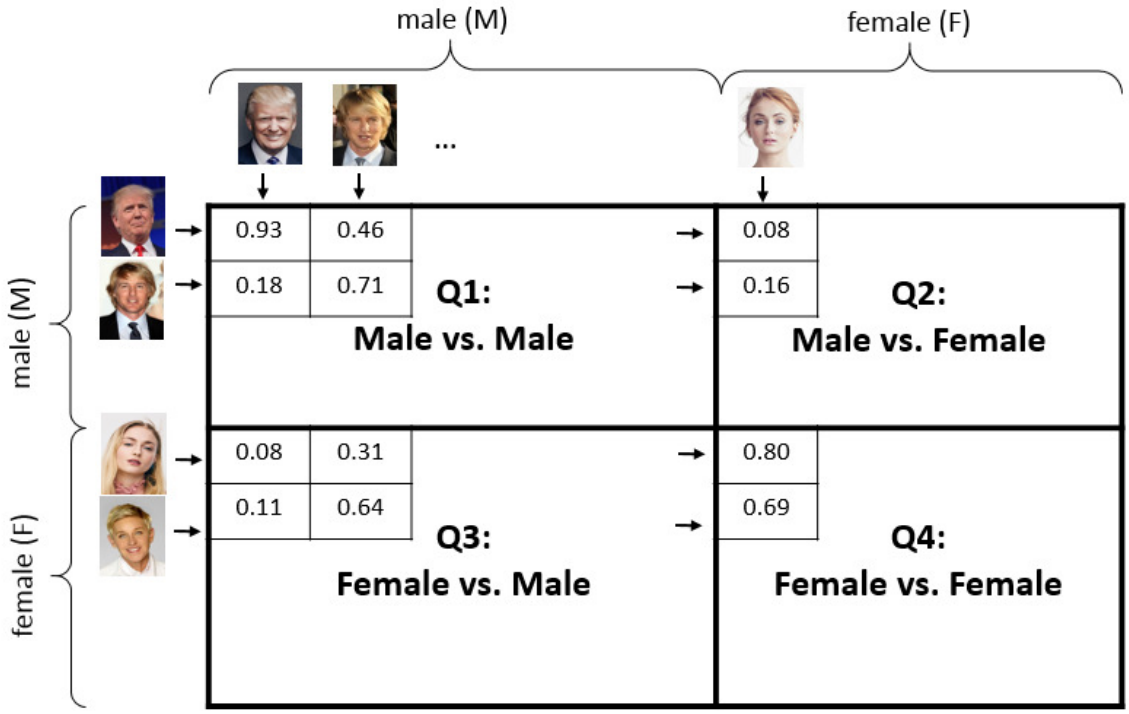


Figure 2.3: Illustration of the partitioned score matrix from a conventional face matcher. The score matrix is partitioned into four quadrants according to different matching scenarios. For instance, "Q1" denotes the scenario where "Male" probe samples are compared against "Male" gallery samples.

the context of biometric recognition systems. However, these methods cannot be trivially appropriated for use with demographic attributes. This is mainly because that most demographic attributes are even less distinctive across the population (e.g., gender) compared to other types of soft biometric information such an anthropometric attributes.

Moreover, the lack of reliability of demographic information can negatively affect biometric matching accuracy if the fusion framework does not adequately account for such types of uncertainty. As pointed out by a report from the Secure Flight Program in the U.S. [19], when travellers' name, gender and age information were used for comparing traveller identity against those on a FBI watch list, the rate of false rejection was significantly increased because of the unreliability of gender information.

2.3 Analytical Investigation on Fusion Schemes

In order to better motivate the proposed fusion approach, and to use a single formulation to explain other fusion schemes, we now turn our attention to the score matrix. The score matrix consists of match scores obtained when comparing every probe biometric sample against every gallery biometric sample. We partition the score matrix into multiple sections, where each section is a matrix of match scores when comparing probe samples with a certain demographic label against gallery samples with a certain demographic label (e.g., "Male vs Male" or "Female vs Male"). Such a partitioning also helps in explaining the rationale for formulating fusion frameworks as score transformation functions.

2.3.1 Partitioned Score Matrix

The biometric verification problem may be formally posed as follows: given a probe biometric sample and a claimed identity, determine whether this claim is true or false [48]. Typically, the probe sample is compared against the gallery sample corresponding to the claimed identity in order to generate a match score (typically a single number), which quantifies the degree of similarity or dissimilarity between these two samples. Consider the score matrix, $\mathbf{X}$, in Figure 2.3, where each entry, x (like 0.93, 0.46, ...), corresponds to the match score obtained when a probe sample is compared against a gallery sample. Hence, each row of this score matrix is a set of match scores generated when comparing an input probe sample against all gallery samples stored in the biometric database. The genuine scores can be denoted as $x \in \mathbf{X}^{Gen}$, while $x \in \mathbf{X}^{Imp}$ denotes the impostor scores. Typically, the score is compared against a threshold, δ, in order to render a decision.

Without loss of generality, the match scores can be assumed to be *similarity* scores. Thus, if the threshold, δ, is decreased to make the system more tolerant to input variations and noise, the False Match Rate (FMR) increases and the True Match Rate (TMR) increases. On the other hand, if δ is increased to make the system more secure, the FMR decreases

while the TMR increases. As a result, each {FMR, TMR} pair is a function of threshold δ:

$$FMR(\delta) = \int_{\delta}^{\infty} Pr(x|x \in \mathbf{X}^{Imp})dx;$$
$$TMR(\delta) = \int_{\delta}^{\infty} Pr(x|x \in \mathbf{X}^{Gen})dx,$$

where, $Pr(x|x \in \mathbf{X}^{Gen})$ denotes the probability density function of genuine scores.

The match score matrix, $\mathbf{X}$, can be partitioned by demographic attributes. Figure 2.3 illustrates a simple case where a face matcher is integrated with a binary gender attribute. There are four quadrants according to the following matching scenarios: "Male vs. Male (Q1)", "Male vs. Female (Q2)", "Female vs. Male (Q3)" and "Female vs. Female (Q4)". Consequently, $x \in \mathbf{X}_{Q1}^{Gen}$ and $x \in \mathbf{X}_{Q1}^{Imp}$ denote the genuine and impostor scores in Q1, separately, where male samples are matched against male samples. Given a threshold δ, the corresponding FMR and TMR can be calculated as :

$$\begin{aligned} FMR(\delta) &= \int_{\delta}^{\infty} Pr(x|x \in \mathbf{X}^{Imp})dx \\ &= \int_{\delta}^{\infty} Pr(x|x \in \mathbf{X}_{Q1}^{Imp})dx + \int_{\delta}^{\infty} Pr(x|x \in \mathbf{X}_{Q2}^{Imp})dx \\ &+ \int_{\delta}^{\infty} Pr(x|x \in \mathbf{X}_{Q3}^{Imp})dx + \int_{\delta}^{\infty} Pr(x|x \in \mathbf{X}_{Q4}^{Imp})dx; \end{aligned}$$

$$\begin{aligned} TMR(\delta) &= \int_{\delta}^{\infty} Pr(x|x \in \mathbf{X}^{Gen})dx \\ &= \int_{\delta}^{\infty} Pr(x|x \in \mathbf{X}_{Q1}^{Gen})dx + \int_{\delta}^{\infty} Pr(x|x \in \mathbf{X}_{Q2}^{Gen})dx \\ &+ \int_{\delta}^{\infty} Pr(x|x \in \mathbf{X}_{Q3}^{Gen})dx + \int_{\delta}^{\infty} Pr(x|x \in \mathbf{X}_{Q4}^{Gen})dx \end{aligned}$$

where $Pr(x|x \in \mathbf{X}_{Q1}^{Imp})$ denotes the probability density function of impostor scores in Q1.

In practice, this probability density can be estimated by counting the number of impostors scores corresponding to the scenario in Q1: $\sum \mathbb{I}(x|x \in \mathbf{X}_{Q1}^{Imp})$. Similarly, $\sum \mathbb{I}(x > \delta|x \in \mathbf{X}_{Q1}^{Imp})$ denotes the number of impostors scores in Q1 that are greater than the given threshold

δ. In summary, the empirical FMR and TMR can be calculated as:

$$\begin{aligned} FMR(\delta) &= \frac{\sum \mathbb{I}(x > \delta | x \in \mathbf{X}^{Imp})}{\sum \mathbb{I}(x | x \in \mathbf{X}^{Imp})} \\ &= \frac{\sum \mathbb{I}(x > \delta | x \in \mathbf{X}_{Q1}^{Imp}) + \ldots + \sum \mathbb{I}(x > \delta | x \in \mathbf{X}_{Q4}^{Imp})}{\sum \mathbb{I}(x | x \in \mathbf{X}_{Q1}^{Imp}) + \ldots + \sum \mathbb{I}(x | x \in \mathbf{X}_{Q14}^{Imp})}; \\ TMR(\delta) &= \frac{\sum \mathbb{I}(x > \delta | x \in \mathbf{X}^{Gen})}{\sum \mathbb{I}(x | x \in \mathbf{X}^{Gen})} \\ &= \frac{\sum \mathbb{I}(x > \delta | x \in \mathbf{X}_{Q1}^{Gen}) + \ldots + \sum \mathbb{I}(x > \delta | x \in \mathbf{X}_{Q4}^{Gen})}{\sum \mathbb{I}(x | x \in \mathbf{X}_{Q1}^{Gen}) + \ldots + \sum \mathbb{I}(x | x \in \mathbf{X}_{Q4}^{Gen})}. \end{aligned} \tag{2.1}$$

With the partitioned score matrix, it is able to investigate the difference of score distributions among matching scenarios. For example, if we assume that all the gender labels are accurate, then it is impossible to have any genuine scores in Q2 and Q3, which results in $\sum \mathbb{I}(x > \delta | x \in \mathbf{X}_{Q2}^{Gen} = \sum \mathbb{I}(x > \delta | x \in \mathbf{X}_{Q3}^{Gen} = 0$. Thus, the above TMR is rewritten as:

$$TMR(\delta) = \frac{\sum \mathbb{I}(x > \delta | x \in \mathbf{X}_{Q1}^{Gen}) + \sum \mathbb{I}(x > \delta | x \in \mathbf{X}_{Q4}^{Gen})}{\sum \mathbb{I}(x | x \in \mathbf{X}_{Q1}^{Gen}) + \ldots + \sum \mathbb{I}(x | x \in \mathbf{X}_{Q4}^{Gen})}.$$

2.3.2 Formulation of Stratified Matching Scheme

Figure 2.13 illustrates the stratified matching scheme as a special case of transformation on match scores. First, the demographic characteristics of the probe sample and the claimed identity are compared. If the demographic characteristics from two samples are same (similar to the proposed matching scenarios of Q1 and Q4 in Figure 2.3), these two biometric samples are compared by a conventional biometric matcher and a match score is generated for rendering the decision. On the other hand, if these characteristics are different, the system rejects the probe without computing a match score (denoted as N/A). The stratified matching scheme, therefore, reduces the computing time and speeds up the recognition process.

However, as is proved below, the verification accuracy cannot be significantly improved by the stratified matching scheme. According to the proposed formulation of verification

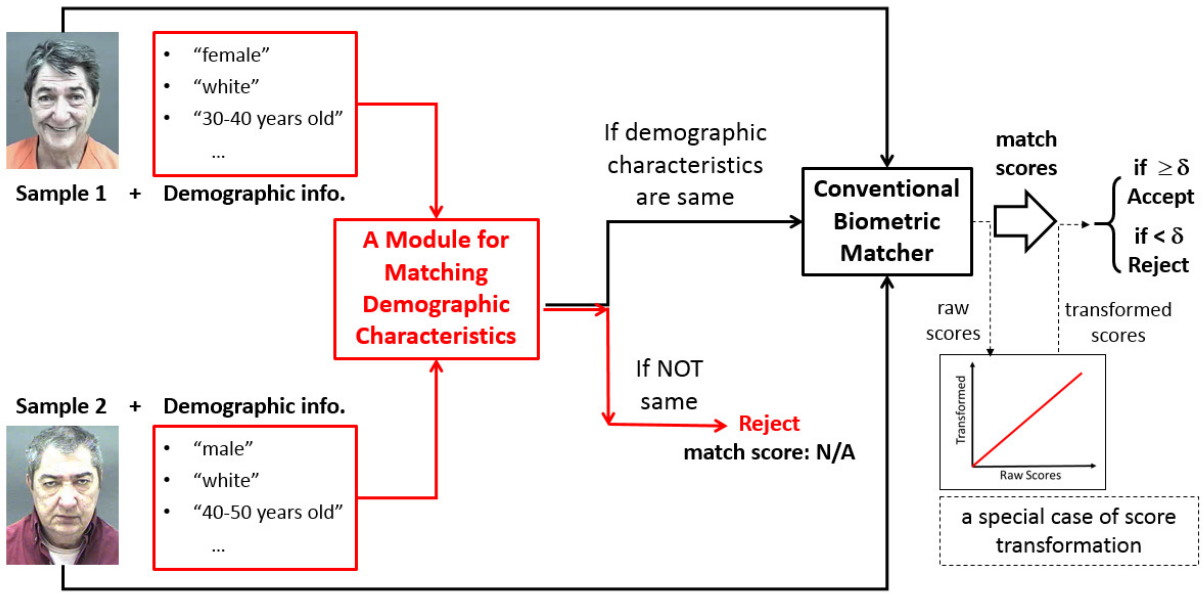


Figure 2.4: Illustration of the stratified matching scheme. When the demographic characteristics from two samples are NOT the same, the stratified matching scheme simply rejects the probe sample without computing any match scores. On the other hand, if the characteristics are the same, the match scores from the conventional biometric matcher are used to render the final decision. The stratified matching scheme can be considered as a special case of demographic-based transformation.

accuracy, as shown in Eqn (2.1), the stratified matching scheme only removes the entries $x \in \mathbf{X}_{Q2}$ and $x \in \mathbf{X}_{Q3}$ because they are not available. As a result, the FMR is reduced by removing the term $\sum \mathbb{I}(x > \delta | x \in \mathbf{X}^{Imp}_{Q2\&Q3})$ from the numerator and the term $\sum \mathbb{I}(x | x \in \mathbf{X}^{Imp}_{Q2\&Q3})$ from the denominator, while the TMR remains same. The FMR and TMR after using the stratified matching scheme are updated as follows:

$$
\begin{aligned}
&FMR_{strat}(\delta) = \\
&\quad \frac{\sum \mathbb{I}(x > \delta | x \in \mathbf{X}^{Imp}_{Q1}) + \sum \mathbb{I}(x > \delta | x \in \mathbf{X}^{Imp}_{Q4})}{\sum \mathbb{I}(x \in \mathbf{X}^{Imp}_{Q1}) + \sum \mathbb{I}(x \in \mathbf{X}^{Imp}_{Q4})}; \\
&TMR_{strat}(\delta) = \\
&\quad \frac{\sum \mathbb{I}(x > \delta | x \in \mathbf{X}^{Gen}_{Q1}) + \sum \mathbb{I}(x > \delta | x \in \mathbf{X}^{Gen}_{Q4})}{\sum \mathbb{I}(x \in \mathbf{X}^{Gen}_{Q1}) + \sum \mathbb{I}(x \in \mathbf{X}^{Gen}_{Q4})}.
\end{aligned}
\tag{2.2}
$$

Comparing Eqn (2.2) to (2.1), it can be safely concluded that the matching accuracy can

be significantly increased only if the following inequation is satisfied:

$$\begin{aligned} &\frac{\sum \mathbb{I}(x > \delta | x \in \mathbf{X}^{Imp}_{Q1\&Q4})}{\sum \mathbb{I}(x \in \mathbf{X}^{Imp}_{Q1\&Q4})} \ll \frac{\sum \mathbb{I}(x > \delta | x \in \mathbf{X}^{Imp}_{Q2\&Q3})}{\sum \mathbb{I}(x \in \mathbf{X}^{Imp}_{Q2\&Q3})} \\ \Longleftrightarrow &\int_{\delta}^{\infty} p(x \in \mathbf{X}^{Imp}_{Q1\&Q4}) dx \ll \int_{\delta}^{\infty} p(x \in \mathbf{X}^{Imp}_{Q2\&Q3}) dx. \end{aligned} \tag{2.3}$$

From an intuitive viewpoint, the original false matches (FM) consist of four parts:

$$\begin{aligned} \textit{Total FM} = &\textit{ FM from Q1} + \textit{FM from Q4} \\ &+ \textit{FM from Q2} + \textit{FM from Q3}. \end{aligned}$$

The stratified matching scheme reduces the total FMR by eliminating the false matches from Q2 and Q3. However, it cannot reduce the false matches within the same strata (Q1 and Q4), which results in limited accuracy improvement in practice.

Indeed, there are other practical concerns about the stratified matching scheme. First, it requires operating thresholds to be strata-specific. For example, in order to achieve a fixed FMR, the stratified matching scheme has to implement different thresholds for male and female subjects, separately. Figure 2.5 presents the ROC curves from a commercial fingerprint matcher (i.e., COTS-C as introduced in section 2.5) when integrating gender information via the stratified matching scheme. Here, if the same thresholds are used for both male and female subjects, the matcher may exhibit significantly different False Match Rates (FMRs) for male and female subjects (as shown in Figure 2.5). Moreover, as shown in Figure 2.6, it may be difficult to compare the accuracy of two matchers when the stratified matching scheme is implemented. Here, we observe that Matcher 1 constantly results in higher accuracy for male subjects while Matcher 2 performs better on female subjects. This is because different decision thresholds were used for male and female subjects and, hence, the matching accuracy of each strata had to be exhibited separately.

Furthermore, the stratified matching scheme could be negatively impacted by mislabeled demographic data. The process of automatically extracting demographic attributes from biometric data is vulnerable to errors. Even the direct collection of demographic information

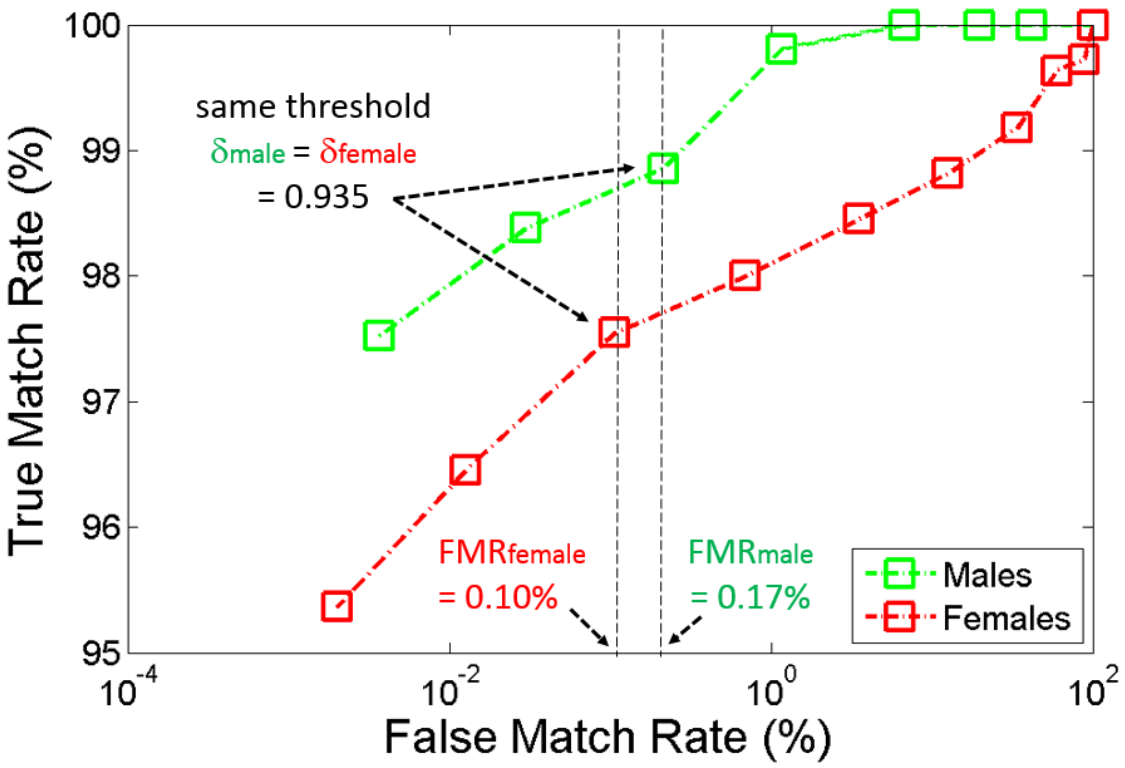


Figure 2.5: Examples of ROC curves from the stratified matching scheme. The gender information of subjects in the WVU database are integrated with a commercial fingerprint matcher (which will be introduced in Section 2.5). It demonstrates that in order to achieve a consistent FMR for both male and female subjects, the stratified matching scheme requires different thresholds according to each strata.

from subjects may be susceptible to transcription errors. As demonstrated by the experimental results in the latter sections, the matching accuracy from the stratified matching scheme sharply degrades when operating with mislabeled demographic information.

2.3.3 Formulation of Decision-Level Fusion Schemes

The decision-level fusion scheme is commonly used in the context of biometric fusion, where the outputs of the individual biometric sources are combined in order to generate the final decision. Fusion at the decision-level is bandwidth efficient because only the final decisions, often requiring just a single bit, are transmitted to the fusion engine [130]. Moreover, decision-level information is more easily accessible in proprietary systems compared to score-

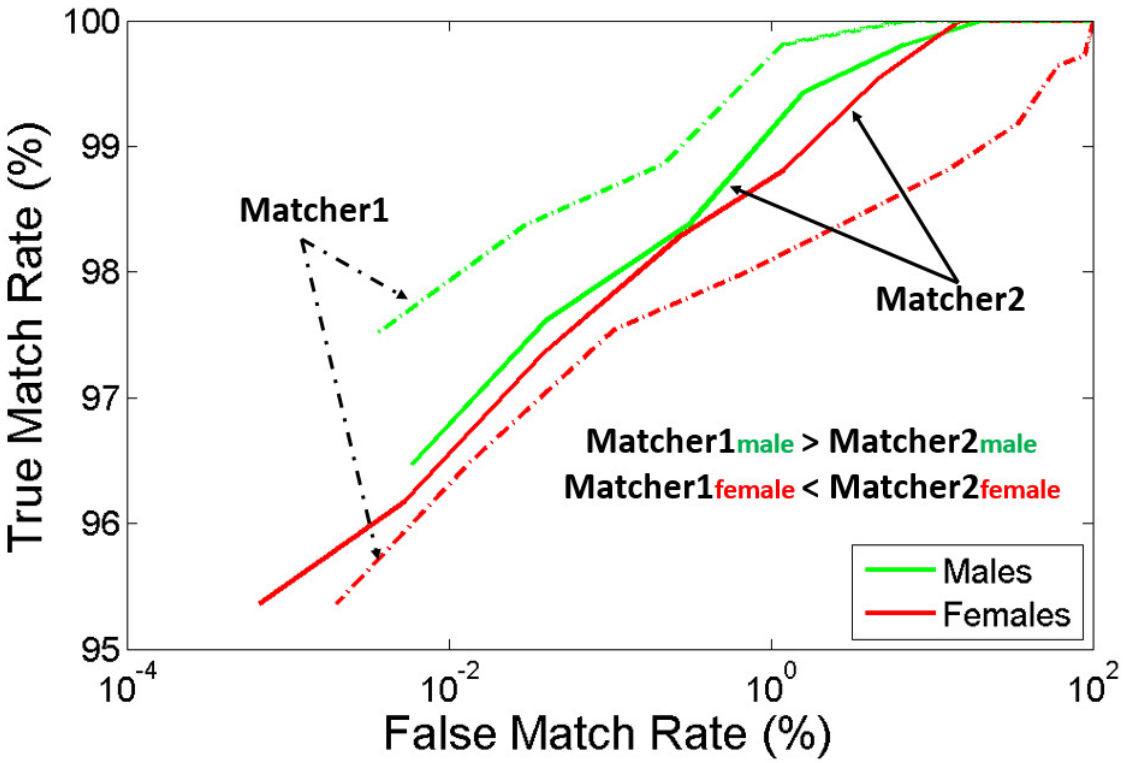


Figure 2.6: An example of ROC curves from the stratified matching scheme, where the gender information of subjects are integrated with two conventional biometric matchers (i.e., Matcher 1 and Matcher 2), respectively. It demonstrates that in order to compare the accuracy of two matchers, the stratified matching scheme still need to exhibit a joint performance rather than the within-cohort performance.

level or feature-level information [108, 14, 130].

In order to combine the demographic information via a decision-level fusion scheme, the demographic characteristics of the probe and the claimed gallery identity need to be compared to render a decision of "Match" or "Non-Match". Demographic-based and biometric-based decisions need to be merged in order to render the final decision. Various techniques are applicable in the biometric *verification* scenario such as majority voting [51], weighted majority voting [61] and naive-bayes combination [55]. We start by investigating the logical AND operator, which can be viewed as a specific case of majority voting, and then generalize the decision-level fusion scheme as a special case of match score transformation.

When a logical AND operator is implemented, as shown in Figure 2.7, the final decision is

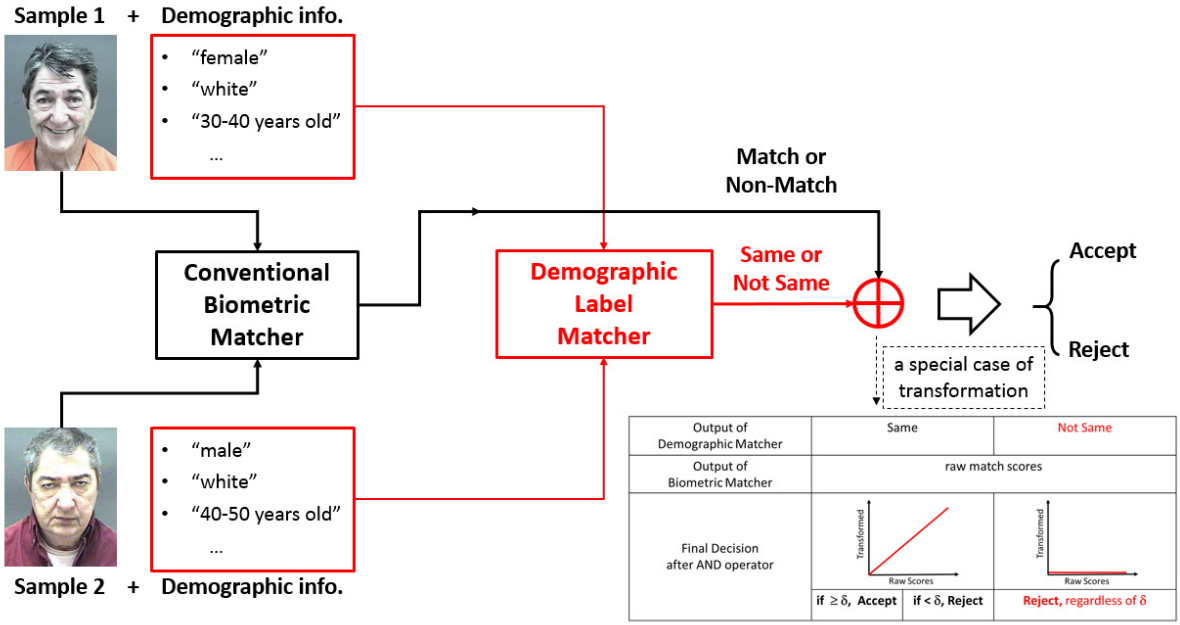


Figure 2.7: Illustration of the decision-level fusion scheme. The decision from a demographic label matcher ("Same" or "Not Same") is combined with the decision from a conventional biometric matcher ("Match" or "Non-Match") to render the final decision ("Accept" or "Reject"). It can be considered as a special case of demographic-based transformation, where the match scores are transformed to zero and rejected regardless of the threshold, if demographic labels are "Not Same" for two samples.

"Accept" only if the demographic-based decision is "Same" and the biometric-based decision is "Match". It is noted that the biometric-based decision relies on the operating threshold δ as well as the match score. According to the score matrix (as shown in Figure 2.3), when the gender labels of two samples are "Not Same" (as in the quadrant Q2 and Q3), the final decision is a "Non-Match" regardless of the threshold of the biometric matcher. This is equivalent to forcing all the match scores to zero resulting in a constant "Non-Match" decision irrespective of the threshold value. On the other hand, when the gender labels are the same for two samples (as in the quadrant Q1 and Q4), the final decision entirely depends on the decision of the biometric matcher. This is equivalent to forcing all the match scores to be the same non-zero value (as shown in Figure 2.7).

Accordingly, compared to Eqn (2.1), the FMR and TMR of the AND-based decision

fusion scheme can be rewritten as:

$$\begin{aligned} FMR_{Dcom}(\delta) &= \frac{\sum \mathbb{I}(x > \delta | x \in \mathbf{X}^{Imp})}{\sum \mathbb{I}(x | x \in \mathbf{X}^{Imp})} \\ &= \frac{\sum \mathbb{I}(x > \delta | x \in \mathbf{X}^{Imp}_{Q1\&Q4})}{\sum \mathbb{I}(x | x \in \mathbf{X}^{Imp}_{Q1\&Q4}) + \sum \mathbb{I}(x | x \in \mathbf{X}^{Imp}_{Q2\&Q3})}; \\ TMR_{Dcom}(\delta) &= \frac{\sum \mathbb{I}(x > \delta | x \in \mathbf{X}^{Gen})}{\sum \mathbb{I}(x | x \in \mathbf{X}^{Gen})} \\ &= \frac{\sum \mathbb{I}(x > \delta | x \in \mathbf{X}^{Gen}_{Q1\&Q4})}{\sum \mathbb{I}(x | x \in \mathbf{X}^{Gen}_{Q1\&Q4})}. \end{aligned} \tag{2.4}$$

Comparing Eqn (2.4) with Eqn (2.1), the denominators of FMR and TMR are both the same, and the only difference is the removal of the term $\sum \mathbb{I}(x > \delta | x \in \mathbf{X}^{Imp}_{Q2\&Q3})$ from the numerator for FMR. As a result, the FMR will be consistently reduced by this fusion scheme.

2.3.4 Generalization and Optimization

As stated earlier, the stratified matching scheme can be considered as a special case of demographic-based transformation of match scores. Let $\mathbf{F}$ denote such a transformation function, while x_i and y_i denote the i^{th} match score before and after the transformation, respectively. Here, $i = 1, \ldots, n$, where n is the total number of entries in the score matrix. Another input factor, z_i, is a coded demographic-based factor indicating which partition the i^{th} match score falls in. Suppose a score matrix is partitioned into four quadrants based on a binary gender factor, as illustrated in Figure 2.3, then z_i can take on one of 4 values, i.e., $z_i = \{1, \ldots, L\}$, where $L = 4$. Accordingly, the stratified matching scheme can be re-written as:

$$y_i = \mathbf{F}_{SM}(x_i, z_i) = \begin{cases} x_i, & z_i = 1 \quad (i.e., x_i \in Q1) \\ N/A, & z_i = 2 \quad (i.e., x_i \in Q2) \\ N/A, & z_i = 3 \quad (i.e., x_i \in Q3) \\ x_i, & z_i = 4 \quad (i.e., x_i \in Q4). \end{cases} \tag{2.5}$$

It shows that when the demographic characteristics from two samples are not the same, as in the case for $z_i = 2$ or 3, the transformation function $\mathbf{f}_{SM}(x_i, z_i)$ records the transformed score as N/A. On the other hand, if the demographic characteristics are the same, as in the case for $z_i = 1$ or 4, then $\mathbf{f}_{SM}(x_i, z_i)$ simply remains the original match score. Indeed, $\mathbf{f}_{SM}$ can be decomposed into L subordinate transformation functions according to different values of z_i. For instance, suppose $f_1(x_i)$ is the subordinate transformation function of the quadrant Q1, we actually have $y_i = f_1(x_i) = x_i$ in the stratified matching scheme.

As a summary of the above observations, we propose a general form of demographic-based transformation functions as:

$$y_i = \mathbf{F}_{general}(x_i, z_i) = \begin{cases} f_1(x_i), & z_i = 1 \\ f_2(x_i), & z_i = 2 \\ \dots & \\ f_L(x_i), & z_i = L. \end{cases} \tag{2.6}$$

The general transformation function, $\mathbf{F}_{general}(x_i, z_i)$, is decomposed into a set of transformation functions $f_l(x_i)$, where $l = 1, \dots, L$. The number of matching scenarios, L, relies on the number of demographic labels.

However, deriving such transformation functions is not easily possible. First, subordinate functions in each partition (i.e., $h_l(x_i)$) may be independent of each other. However, they need to be explored simultaneously, since it is important to improve the global verification accuracy and not just the within-partition verification accuracy. Further, there is no inherent constraint on the *form* of the subordinate functions. As shown in the stratified matching scheme in Eqn (2.5), one subordinate function is linear while another function always outputs N/A. The arbitrary form indeed enhances the difficulty of solving this problem analytically.

These issues inspire us to address the problem via the *additive model* (AM) with a con-

tinuous predictor and a factor-by-curve interaction, formulated as:

$$y_i = \mathbf{F}(x_i, z_i) + \epsilon_i = \alpha_0 + \sum_{z_i=1}^{L} f_{z_i}(x_i) + \gamma_{z_i} z_i + \epsilon_i,$$

where γ_{z_i} is the coefficient of the interaction, and ϵ_i is the residual. We will explain our rationale below.

2.4 Additive Model and Extension

2.4.1 Additive Model with Interaction

Suppose we have a set of observations $\{(\mathbf{x_1}, y_1), \ldots, (\mathbf{x_n}, y_n)\}$, where $\mathbf{x_i}$ is a vector of p continuous covariates and y_i is the continuous response of interest. The covariates, X, in our case, are original match scores from conventional biometric matchers, while the response, Y, denotes the transformed match scores which will be used to render the verification decision (as shown in Eqn (2.6)). As a widely used extension of traditional linear models, an additive model (AM) can represent the relationship between covariates and the response variable as the sum of low-dimensional transformation functions:

$$Y = \mathbf{F}(X) + \epsilon = \alpha_0 + \sum_{j=1}^{p} f_j(X) + \epsilon, \tag{2.7}$$

where α_0 is a constant and f_j are the smooth partial functions or effects associated with each continuous covariate in X. The AM is more flexible than the linear models since there is no assumption of a parametric form of the effects of the continuous covariates, X, but only assumes that these effects can be represented by unknown smooth functions, f_j. Without the restriction of linearity, additive models are more flexible than linear regression models. Besides flexibility and accuracy, a key promising point is the interpretability, as the additive predictors provide visual means for inspecting the models and identifying domain-specific relations between inputs and outputs [43].

As investigated in section 2.3, the effect of the original scores, X, on the transformed scores, Y, vary across groups defined by *levels* of a categorical *demographic factor* in our

case. Let us denote Z as a coded demographic factor with L levels and $p = 1$ indicating that X only includes the match scores from one biometric modality. Then, an extension of the additive model with factor-by-curve interactions included, which is proposed by [18], can better express the relationship between X, Y and Z as follows:

$$Y = \mathbf{F}(X, Z) + \epsilon = \alpha_0 + \sum_{k=1}^{L} f_k(X)\mathbb{I}_{Z=k} + \sum_{k=2}^{L} \gamma_k \mathbb{I}_{Z=k} + \epsilon, \tag{2.8}$$

where $\mathbb{I}_{Z=k}$ is the indicator function for the kth level of Z. The term, γ_k, is the coefficients of the factor-by-curve interaction. As pointed by Coull et al. [18], for the situations where the interaction term is statistically significant, the effect of the covariates on response can be expressed via different curves across levels of the categorical factor. Different curves, in our case, are different score transformation functions corresponding to different match scenarios, which can reduce the overall verification error rates.

2.4.2 Fitting AM via Penalized B-Splines

Various approaches has been developed for fitting the model in Eqn (2.8). Hastie and Tibshirani [42] discussed a number of approaches using smoothing splines. Coull et al. [18] implemented the penalized splines for fitting the additive models, and a difference penalty on coefficients of splines was used instead of using the integral of the squared second derivative. The term "spline" is used to refer to a wide class of functions that are used in applications requiring data interpolation and smoothing. The simplest spline is a piecewise polynomial function, with each polynomial having a single variable. If a spline is constructed of piecewise third-order polynomials which smoothly pass through a set of control points, also referred to as "knots", it becomes a so-called "natural" *cubic spline* [3].

Compared with the simple cubic spline, B-splines are more attractive for non-parametric modelling, where the optimal number and positions of knots is learned from the data. Equidistant knots can be used in B-splines as well, but their small and discrete number allows only limited control over smoothness and fit.

In order to avoid overfitting, a form of penalization is commonly required for learning splines. Eilers and Marx [26] first proposed to use a difference penalty on coefficients of adjacent B-splines. Compared to the familiar spline penalty on the integral of the squared second derivative, the computational complexity is sharply simplified, especially for the case of fitting additive models with factor-by-curve interactions [18].

For simplicity, we directly explain the factor-by-curve interactions used in this work, which is specified for one single covariate (i.e., the match scores, X, from one single biometric matcher), and one single categorical factor (i.e., the demographic factor Z). Consider the set of triples (x_i, y_i, z_i), where the x_i and y_i represent the i^{th} match score before and after the transformation, respectively, and z_i represents a coded demographic-based factor. The additive model for fitting is

$$y_i = \mathbf{F}(x_i, z_i) + \epsilon_i = \alpha_0 + \sum_{z_i=1}^{L} f_{z_i}(x_i) + \gamma_{z_i} x_i + \epsilon_i, \tag{2.9}$$

where $f_1, \ldots, f_L$ are L different subordinate transformation functions depending on the value of z_i, and ϵ_i i.i.d. $N(0, \sigma_\epsilon^2)$. It must be noted that the score after transformation, y_i, which is considered as the response variable in Eqn (2.9), is not the actual response in biometric verification study. Extra transformations are analyzed in section 2.4.3.

Let $\kappa_1, \ldots, \kappa_K$ be a set of distinct knots inside the range of the x_i's and let $x_+ = max(0, x)$. The knots are usually taken to be relatively dense among the observations in an attempt to capture the curvature in $f_l, l = 1, \ldots, L$. Ruppert and Carroll [11] described an algorithm for choosing the number of knots and demonstrated its effectiveness through simulation. Let us define:

$$z_{il} = \begin{cases} 1 & \text{if } z_i = l, \\ 0 & \text{otherwise.} \end{cases} \tag{2.10}$$

The linear (i.e., 1st order) penalized spline model for Eqn (2.9) is:

$$\begin{aligned} y_i =& \beta_0 + \beta_1 x_i + \sum_{k=1}^{K} b_k (x_i - \kappa_k)_+ + \sum_{l=2}^{L} z_{il}(\gamma_{0l} + \gamma_{1l} x_i) \\ &+ \sum_{l=1}^{L} z_{il} \Big\{ \sum_{k=1}^{K} c_k^l (x_i - \kappa_k)_+ \Big\} + \epsilon_i, \end{aligned} \tag{2.11}$$

subject to the constraints

$$\sum_{k=1}^{K} b_k^2 < B \text{ and } \sum_{k=1}^{K} (c_k^l)^2 < C_l, \quad l = 1 \dots L, \tag{2.12}$$

for some constant B and C_l. The term $\gamma_{0l} + \gamma_{1l} x_i$ models the linear deviation between f_1 and f_l, where $l = 2 \dots L$. The term $\sum_{k=1}^{K} b_k (x_i - \kappa_k)_+$ represents the overall smooth term. The term $\sum_{k=1}^{K} c_k^l (x_i - \kappa_k)_+$ represents deviations from the overall smooth term [11]. The penalty in Eqn (2.12) induces smoothness in the effect of our covariate variable X and Y. As pointed by Ruppert and Carroll [11], the exact number of knots is not a major concern.

Suppose the gender information with two different labels (i.e. "Male" and "Female") of subjects are integrated with a conventional face matcher. There are four different matching scenarios, indicated as $z_i \in \{1, 2, 3, 4\}$, and $L = 4$ subordinate transformation functions. Given an arbitrary match score in the score matrix X, x_i, Eqn (2.11) can be written as:

$$\begin{aligned} y_i = \beta_0 + \beta_1 x_i + \sum_{k=1}^{K} b_k (x_i - \kappa_k)_+ + \epsilon_i \\ + \begin{cases} \sum_{k=1}^{K} c_k^1 (x_i - \kappa_k)_+ & z_i = 1, \\ (\gamma_{02} + \gamma_{12} x_i) + \sum_{k=1}^{K} c_k^2 (x_i - \kappa_k)_+ & z_i = 2, \\ (\gamma_{03} + \gamma_{13} x_i) + \sum_{k=1}^{K} c_k^3 (x_i - \kappa_k)_+ & z_i = 3, \\ (\gamma_{04} + \gamma_{14} x_i) + \sum_{k=1}^{K} c_k^4 (x_i - \kappa_k)_+ & z_i = 4. \end{cases} \end{aligned} \tag{2.13}$$

Shively et al. [116] pointed out that, for given values of B and $C_l, l = 1 \dots L$, the model in Eqn (2.11), subject to the constraints in Eqn (2.12), yields fitted values equivalent to

those produced by the model:

$$\begin{aligned} y_i =& \beta_0 + \beta_1 x_i + \sum_{k=1}^{K} b_k (x_i - \kappa_k)_+ + \sum_{l=2}^{L} z_{il}(\gamma_{0l} + \gamma_{1l} x_i) \\ &+ \sum_{l=1}^{L} z_{il} \Big\{ \sum_{k=1}^{K} c_k^l (x_i - \kappa_k)_+ \Big\} + \epsilon_i, \end{aligned} \tag{2.14}$$

where, b_k i.i.d. $N(0, \sigma_b^2)$ and c_k^l i.i.d. $N(0, \sigma_{cl}^2)$ for appropriate values of σ_b and σ_{cl}. The above mixed model formulation of penalized spline models is used in the work. It can be rewritten in matrix notation as:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{z}\mathbf{u} + \boldsymbol{\epsilon}, \tag{2.15}$$

where

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 & z_{12} & \dots & z_{1L} & z_{12}x_1 & \dots & z_{1L}x_1 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & z_{n2} & \dots & z_{nL} & z_{n2}x_n & \dots & z_{nL}x_n \end{bmatrix},$$

$$\boldsymbol{\beta} = (\beta_0, \beta_1, \gamma_{02}, \dots, \gamma_{0L}, \gamma_{12}, \dots, \gamma_{1L})^T,$$

$$\mathbf{z} = \begin{bmatrix} (x_1 - \kappa_1)_+ & \dots & (x_n - \kappa_1)_+ \\ \vdots & \ddots & \vdots \\ (x_1 - \kappa_K)_+ & \dots & (x_n - \kappa_K)_+ \\ \vdots & \ddots & \vdots \\ z_{11}(x_1 - \kappa_1)_+ & \dots & z_{n1}(x_n - \kappa_1)_+ \\ \vdots & \ddots & \vdots \\ z_{11}(x_1 - \kappa_K)_+ & \dots & z_{n1}(x_n - \kappa_K)_+ \\ z_{12}(x_1 - \kappa_K)_+ & \dots & z_{n2}(x_n - \kappa_K)_+ \\ \vdots & \ddots & \vdots \\ z_{1L}(x_1 - \kappa_K)_+ & \dots & z_{nL}(x_n - \kappa_K)_+ \end{bmatrix}^T,$$

$$\mathbf{u} = (b_1, \dots, b_K, c_1^1, \dots, c_K^1, c_1^2, \dots, c_K^L)^T,$$

and

$$\begin{pmatrix} \mathbf{u} \\ \boldsymbol{\epsilon} \end{pmatrix} \sim N\Bigg(0, \begin{bmatrix} \mathbf{G} & 0 \\ 0 & \sigma_\epsilon^2 \mathbf{I} \end{bmatrix}\Bigg),$$

with $\mathbf{G} = diag(\sigma_b^2 \mathbf{1}_K, \sigma_{c1}^2 \mathbf{1}_K, \ldots, \sigma_{cL}^2 \mathbf{1}_K)$. Here, $\mathbf{1}_K$ is the $K \times 1$ vector of ones. Thus, the penalized spline model in Eqn (2.14) falls within the linear mixed model framework with $\mathbf{X} \in \mathbb{R}^{n \times (2L+1)}$ and $\mathbf{z} \in \mathbb{R}^{n \times K(L+1)}$, and there is a well-developed body of methodology for this broad class of models that can be used to estimate the parameters [116]. In particular, the best linear unbiased predictor (BLUP) proposed by Robinson [106] is used for the estimation:

$$\hat{\boldsymbol{\beta}} = \Big\{ \mathbf{X}^T (\mathbf{z}\mathbf{G}\mathbf{z}^T + \sigma_\epsilon^2 \mathbf{I})^{-1} \Big\}^{-1} \mathbf{X}^T \Big(\mathbf{z}\mathbf{G}\mathbf{z}^T + \sigma_\epsilon^2 \mathbf{I} \Big)^{-1} \mathbf{y} \tag{2.16}$$

and

$$\hat{\mathbf{u}} = \sigma_\epsilon^2 \Big(\sigma_\epsilon^2 \mathbf{z}^T \mathbf{z} + \mathbf{G}^{-1} \Big)^{-1} \mathbf{z}^T (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}). \tag{2.17}$$

Extension to models of higher order polynomials $(x_i - \kappa_K)^m$ with $m > 1$ is straightforward. Specifically, this study implements the 2nd order penalized B-spline model for solving Eqn (2.9):

$$\begin{aligned} y_i =& \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \sum_{k=1}^{K} b_k (x_i - \kappa_k)_+^2 \\ &+ \sum_{l=2}^{L} z_{il} (\gamma_{0l} + \gamma_{1l} x_i + \gamma_{2l} x_i^2) \\ &+ \sum_{l=1}^{L} z_{il} \Big\{ \sum_{k=1}^{K} c_k^l (x_i - \kappa_k)_+^2 \Big\} + \epsilon_i, \end{aligned} \tag{2.18}$$

where, again, b_k i.i.d. $N(0, \sigma_b^2)$, and c_k^l i.i.d. $N(0, \sigma_{cl}^2)$, $l = 1, \ldots, L$.

The constraints on exploring transformation functions are discussed as aforementioned. It is notable that there are plenty of alternative forms available for the transformation, such as the LOESS function, the simple polynomial function, etc. In the following section, we need to connect the transformation function exploration with the global verification accuracy.

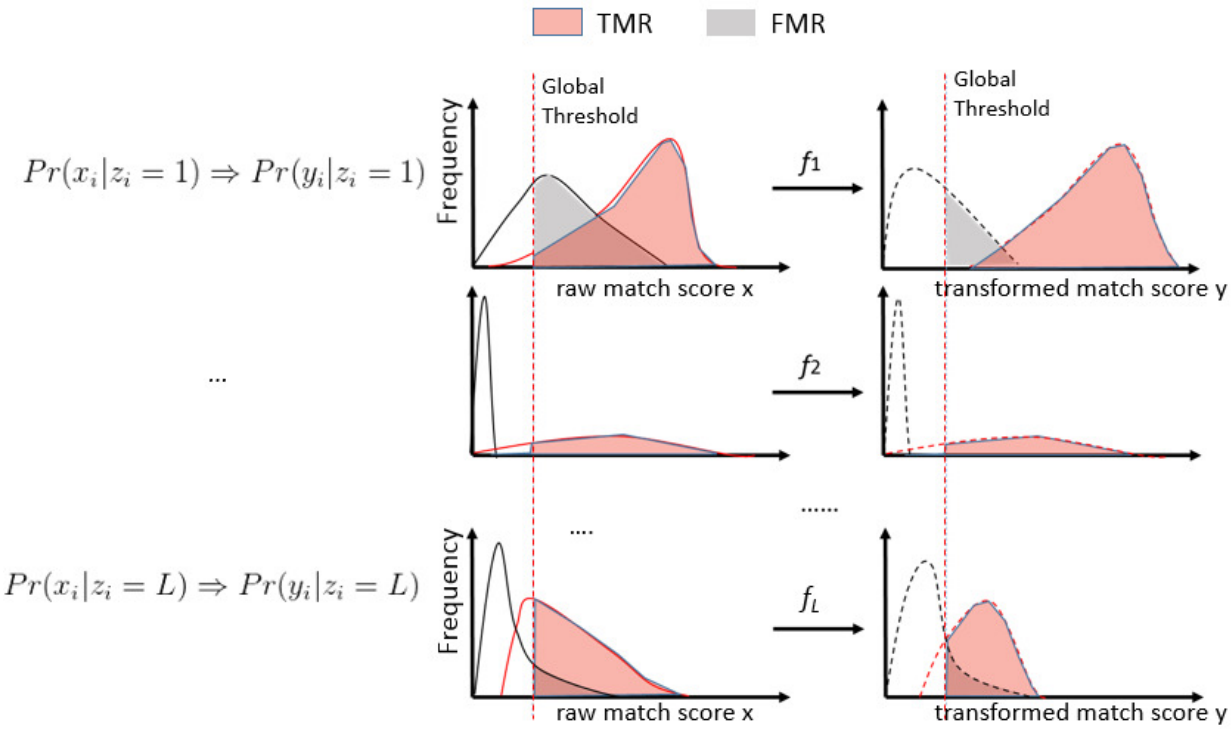


Figure 2.8: An intuitive example of the transformation functions that can better separate the genuine and impostor score distributions and achieve a higher overall matching accuracy.

Fortunately, we have an extension of the additive model, the Generalized Additive Model, that has been commonly implemented in typical classification problems and will now be used in the biometric fusion scenario.

2.4.3 Generalized Additive Model

In a biometric verification study, the decision of "Accept" or "Reject", rather than the transformed score, Y, is considered as the response variable. The Generalized Additive Model (GAM) techniques, which can be used to predict the mean of a response variable, depending on the values of other explicative covariates, allow us for a further extension to include categorical response variables in the AM, that is really essential in a biometric verification problem. It is worth noticing that the GAMs avoid the curse of dimensionality by restricting the non-parametric regression problem to an additive model [43]. In other words, a GAM

can inherit the interpretability from the AM, and its additive components simply describe the influence of each covariate, separately.

Explicitly, we are interested in predicting the biometric verification decision using a GAM for binary response, Y', with two levels , "1/0", corresponding to "Accept/Reject". The match scores, X, are generated using one single biometric matcher, which results in $p = 1$ in Eqn (2.7). A link function, L, is used to convert the continuous variable, Y, which denotes the scores after the transformation in Eqn (2.7), into the binary variable, Y'. This response variable Y' follows a Bernoulli distribution, where $E(Y'|X) = Pr(Y' = 1|X)$. Thus, the corresponding link function L can be written as:

$$\mathsf{L}(Pr(Y')) = \ln \frac{Pr(Y')}{1 - Pr(Y')}.$$

As a typical logistic conversion, the GAM framework takes the form:

$$\mathsf{L}(E(Y'|X)) = \mathbf{F}(X, Z) + \epsilon. \tag{2.19}$$

According to Eqn (2.7) and the investigation in section 2.3.4, the link function directly connects the transformed match scores with the verification accuracy. Intuitively speaking, a transformation function would improve the verification accuracy only if it better separates the score distributions of genuine and impostor scores compared to the original match scores. Figure 2.8 exhibits the distributions of match scores before and after implementing a GAM-based score transformation.

Eqn (2.20) summarizes the formula of GAM used in this work. The transformation functions, $\mathbf{F}(x_i, z_i)$, are estimated based on the penalized B-spline models (as seen in Eqn (2.9)), which is a special form of a piecewise function that can be simply implemented without limitations on the number and location of knots.

$$\mathsf{L}(E(y'_i|x_i, z_i)) = ln(\frac{Pr(y'_i|x_i)}{1 - Pr(y'_i|x_i)}) = \mathbf{F}(x_i, z_i) + \epsilon_i. \tag{2.20}$$

In summary, the generalized additive model is a logistic transformation of additive models (as shown in Eqn (2.20)), where binary responses are used in order to fit the biometric

Figure 2.9: Examples of biometric images in the three datasets used in this work: a) Morph face database, b) LFW face database, and c) WVU multimodal dataset.

verification scenario. The mean of the binary response is related to the predictors using a link function L. The use of the link function is one of the central ideas of generalized linear models.

In this paper, we use the methodology proposed by Wood [133] for fitting GAMs in the form of Eqn (2.20). The existence of standard software in R, such as Wood's *mgcv* package, makes it easy to fit models of this type in practice. The main idea is to implement a penalized iteratively re-weighted least squares scheme (P-IRLS), and more details can be found in [133]. Besides, the *mgcv* package offers an option of *tensor product (te)* which produces spline functions of multiple predictors. Compared to the *isotropic (s)* model, the tensor product model is better for modelling interactions of quantities measured in different units, or where very different degrees of smoothness appropriate relative to different levels in a factor [133]. In this study, the tensor product model is applicable to test whether the transformation functions corresponding to different levels in Z are significantly different (will be seen in section 2.5.5).

2.5 Experimental Results

2.5.1 Databases and Tools

Extensive experiments were conducted to investigate whether the proposed GAM-based fusion scheme can effectively integrate the demographic information into the biometric matching framework. Three main databases, along with two commercial face matchers (COTS-A and COTS-B) and one fingerprint matcher (COTS-C), are used to evaluate the universality of the proposed scheme. Table 2.4 summarizes the purpose of each set of experiments, along with the corresponding databases and tools which have been used. The examples of biometric sample images from the databases used in this work are presented in Figure 2.9.

(1) **Morph face database**: The Morph face database was collected over two sessions, and in each session different number of face samples were collected. Further, there are different number of samples available for each subject. Subjects with only one sample were not used in our work. Still, more than 11,000 face images of 3,500 subjects were retained. A 5-fold cross-validation protocol was used to reduce the potential over-fitting problem, and the average accuracy is presented.

To investigate whether the proposed scheme is affected by unevenly distributed demographic labels, the subjects in each fold are intentionally organized. Table 2.2 gives an example of how the gender labels are distributed in an arbitrary fold of the cross-validation protocol. On the other hand, if the target is to evaluate the performance of combining the race attribute with face matchers, the subjects would be re-organized according to the distribution of the race attribute within each fold. Indeed, experimental results did not demonstrate a significant effect associated with the imbalance issue.

As aforementioned, the proposed GAM scheme is a learning-based method, whose parameters highly related with the biometric matcher which is used for generating match scores. Generally speaking, the COTS-B performs better than the COTS-A. However, the COTS-A provides a built-in gender estimation module which can automatically extract the gender

Table 2.2: A demonstration of how the demographic labels are distributed in one fold of the 5-fold cross-validation protocol that was executed on the MOR face database. Subjects are organized according to their gender information to retain class balance.

		Female	Male	Total
Training Sets	Black	789	778	1,567 sub
	White	656	667	1,323 sub
	Total	1,445	1,445	2,890 sub
Test Sets	Black	200	197	397 sub
	White	178	181	359 sub
	Total	378	378	756 sub

information from the facial images which were collected for the recognition purpose. By comparing with the gender labels manually annotated, the gender estimation results from the COTS-A may consist of a mislabeling rate around 12.0%.

(2) **LFW face database**: The LFW database, which stands for Labeled Faces in the Wild, is designed for studying the problem of unconstrained face recognition. The database consists of more than 13,000 images of faces collected from the web, which are varied in many factors, such as background, pose, illumination, etc.

In order to adhere to a publicly available benchmark, the design of our experiments carefully followed the protocol defined under the category of "Image-Restricted, No Outside Data Results" in LFW's official website[1]. Regarding the 10-fold cross-validation as required by this benchmark, 300 matched pairs (leading to genuine scores) and 300 mismatched pairs (leading to impostor scores) are fixed in each fold. As noted, the sample size in each fold is much smaller compared to the Morph face database. However, the experimental results, where the manually annotated gender attribute is integrated with the match scores generated using the COTS-B face matcher, indicate a significant improvement in the matching accuracy.

The LFW database does not include the ground truth of subjects' demographic labels. As a result, the gender attribute and race attribute of 1,665 individuals were labeled manually. Both attributes are highly imbalanced across the classes. For example, there are 1,231 male

[1] `http://vis-www.cs.umass.edu/lfw/results.html`

Table 2.3: A demonstration of how the demographic labels are distributed in one fold of the 5-fold cross-validation which was performed using the WVU multimodal dataset. In this example, the distribution of race labels is intentionally kept balanced for the two categories (i.e., Caucasian and Non-Caucasian), while the gender distribution may not be balanced at the same time.

		Female	Male	Total
Training Sets	Caucasian	39	66	105 sub
	Non-Caucasian	22	88	110 sub
	Total	61	154	215 sub
Test Sets	Caucasian	16	27	43 sub
	Non-Caucasian	12	30	42 sub
	Total	28	57	85 sub

subjects and 434 female subjects, among which 953 individuals are labeled as "White" and 712 subjects are labeled as "Not White." This manual annotation was compared with the result from Kumar et al.'s automated estimation algorithm [54]. It is observed that around 10% of the subjects are differently labeled, which illustrates a practical scenario where the mislabeling issue exists. As we show later, if the mislabeling rate is less than 20%, the performance of the GAM fusion scheme is not adversely impacted.

(3) **WVU multimodal dataset**: In the WVU Multimodal database, each subject has five samples of fingerprints corresponding to the left index (marked as "FL1"), five samples of the fingerprint corresponding to left thumb (marked as "FL2"), and five frontal facial images (marked as "Face"). The match scores of fingerprint samples are generated using a commercial fingerprint matcher, COTS-C, while both COTS-A and COTS-B face matchers are used to generate match scores from the facial images (as shown in Table. 2.4).

Both gender and race information is directly collected from each of 240 subjects during the enrollment, whose gender is labeled as "Male" or "Female", while the race information is recorded as "Caucasian", "Asian-Indian", "Asian" or "Others". The latter three categories are combined and labeled as "Non-Caucasian" in this work. Table 2.3 demonstrates how the demographic labels are distributed in one fold of the 5-fold cross-validation in this database.

Table 2.4: A summary of the experimental design in this work. An extensive experiments were carried on three biometric databases with three biometric modalities. The match scores are generated using three commercial biometric matchers. The demographic attributes are labelled by: i) a direct collection (marked as "D") from subjects, ii) a manual annotation (marked as "M"), or iii) a machine learning based gender estimation module from COTS-A (marked as "L").

Purpose	**Database**	**Biometric Matcher**	**Demographics & Source**	**Results**
Accuracy	LFW Face	*COTS-B*	gender (M)	Tab. 2.5
			race (M)	
	Morph Face	*COTS-B*	gender (M)	Fig. 2.10 (a)&(b)
			race (M)	
Scalability	Morph Face	*COTS-B*	gender (M) + race (D)	Fig. 2.11 (a)
			gender (L) with 3 levels	Tab.2.6
	WVU FL1	*COTS-C*	gender (M) + race (D)	Fig. 2.11 (b)
	WVU Face	*COTS-B*	gender (L) with 3 levels	Tab.2.6
Predicting	WVU Face	*COTS-A*	gender (L)	Fig. 2.12 (a)&(b)
		COTS-B	gender (L)	
	Morph Face	*COTS-A*	gender (L)	Fig. 2.11 (a)
		COTS-B	gender (M)	
	WVU	*COTS-C*	race (D)	Tab. 2.7
Robustness	Morph Face	*COTS-A*	No gender	Tab. 2.8
			gender (M)	
			10% Mislabeled	
			gender (L)	
			20% Mislabeled	

2.5.2 Experimental Design

The experimental design consists of *Four* major parts:

1. *Improvement in Matching Accuracy:* This set of experiments are designed to investigate the most fundamental question that whether the matching accuracy is benefited by incorporating a single binary demographic attribute into the biometric matching framework. The experimental results from the LFW face databases and the WVU mutlimodal database, which consist of match scores from three different biometric matchers, are exhibited to convey the benefits.

2. *Fusion Scalability:* The experiment is to investigate the scalability of the proposed GAM scheme by combining multiple demographic attributes with match scores, simultaneously. The experimental results inspire us to predict the accuracy improvement in advance by conducting a model diagnostic.

3. *Model Diagnostic and Predictive Metric:* The purpose of this experiment is to propose a metric which can predict in advance if integrating match scores with particular demographic information is beneficial in the context of a specific biometric matcher. The predictive metric relies on the linear model diagnostic process, which is applicable here because of certain inherent properties of the proposed GAM scheme.

4. *Robustness in Inaccurate Labels:* This set of experiments demonstrate that the proposed GAM scheme is robust to missing or inaccurate demographic labels when these labels are used in conjunction with biometric traits. Rather than only simulating the "mislabeling cases" in the test set, a proportion of subjects in the training set are assumed to contain "reversed" demographic labels. The purpose of this design is to simulate the scenario where demographic labels are gleaned from biometric data using automated machine learning schemes.

Table 2.5: The matching accuracy of the proposed GAM fusion scheme on the LFW face database. The true match rates (TMRs) and standard errors are reported under the category of "Image-Restricted, No Outside Data" on the LFW face database. The performance is compared with multiple existing algorithms reported under the same protocol.

Algorithms	Average TMR ± SE
MRF-Fusion-CSKDA [4]	0.9589 ± 0.0194
POP-PEP [58]	0.9110 ± 0.0147
Eigen-PEP [59]	0.8897 ± 0.0132
RSF [109]	0.8881 ± 0.0078
COTS-B (as baseline)	0.8777 ± 0.0052
COTS-B + gender via proposed GAM	0.9280 ± 0.0099
COTS-B + race via proposed GAM	0.8989 ± 0.0105

2.5.3 Experiment 1. Matching Accuracy

The main purpose of integrating demographic information with a biometric matching framework is to improve the human recognition accuracy. The matching accuracy is commonly compared via the true match rates (TMRs) and false match rates (FMRs) corresponding to given operating thresholds (as shown in Eqn (2.1)).

Our experimental results on the LFW face database are reported in the last 3 rows of Table 2.5. As can be seen here, without integrating any demographic attributes, the match scores generated using the commercial face matcher COTS-B can provide a 87.78% true match rate under the required protocol [44], which is comparable with the best existing algorithms reported by the LFW official website (as shown in the Table 2.5. It is noted that the benchmark of LFW required a strict 10-fold cross-validation, and all the FMRs reported here are an average accuracy over all folds. Hence, it can be seen that the COTS-B matcher performs stably over the 10 folds since the standard error of FMRs is small. The same match scores from the COTS-B matcher are then transformed via the proposed GAM scheme, where four gender-based transformation functions are learned from the training sets, respectively. As shown in the $6th$ row of Table 2.5, the averaged TMR is increased to 92.80% at the same FMR level. If the race attributes are integrated instead of gender, the TMR achieves

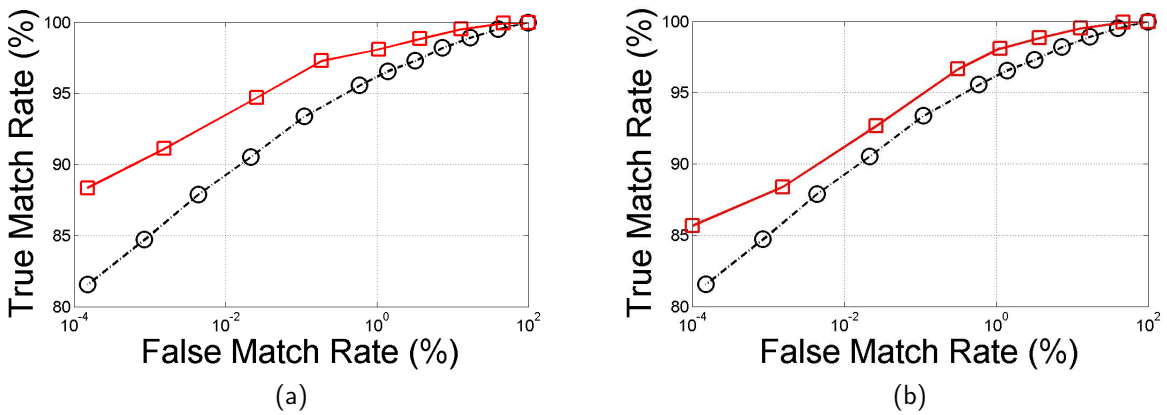


Figure 2.10: ROC curves before (marked as dashed lines) and after (marked as solid lines) integrating demographic attributes with the match scores generated by the COTS-B face matcher on the Morph face database. For instance, (a) face + gender, and (b) face + race

89.89%. Both cases demonstrate a significant improvement in the matching accuracy due to the proposed fusion scheme.

The relatively high standard errors on TMRs (i.e., 0.0099 and 0.0105) suggest that in certain folds of the cross-validation, this GAM-based combination has a comparable performance with the top face matching algorithms which are listed in the first 4 rows of Table 2.5. This high variance among folds is mainly due to the limited training data in each fold. In each fold, there are only 300 genuine scores and 300 impostor scores available for estimating parameters of the GAM corresponding to that fold. Compared to linear models, the training of additive models requires more samples. Moreover, folds in the cross-validation are randomly selected without regarding how the demographic labels are distributed across the classes. Suppose the training set in a fold only consists of few match scores conforming to the scenario of "Female vs. Female", then the subordinate transformation function corresponding to this scenario may have a very low degree of freedom, which leads to inferior performance on the test set.

Figure 2.10 demonstrates the experimental results on the Morph face database. The match scores are generated using the COTS-B face matcher. Both the gender and race labels in Figure 2.10 are manually annotated. The ROCs convey the improvement in the

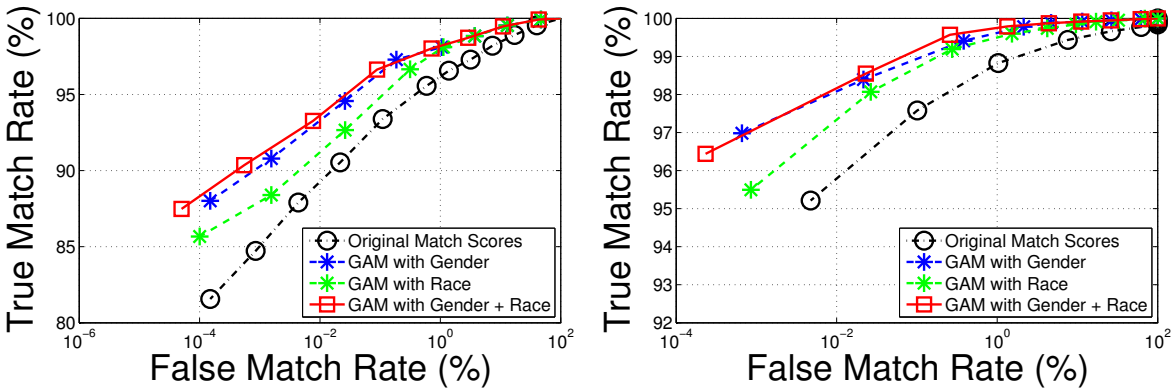


Figure 2.11: ROCs for integrating multiple demographic attributes, simultaneously. The left figure (a) is from the Morph face database, where the match scores are generated using the COTS-B face matcher. The right figure (b) is from the WVU FL1 fingerprint database, where the match scores are generated using the COTS-C fingerprint matcher.

matching accuracy after incorporating demographic attributes into the matching framework via the proposed GAM scheme. For example, when the FMR was fixed at 0.01%, the TMRs were increased from 88.2% to 92.7% by integrating the gender attribute, and to 91.7% by integrating the race attribute.

As a summary, it is evident that the proposed GAM scheme can effectively combine the demographic information with conventional biometric matcher and improve the verification performance.

2.5.4 Experiment 2. Scalability to Multiple Attributes

So far, the gender and race are integrated with match scores via the proposed GAM scheme, separately. Both attributes are binary factors with only two levels, which results in four different transformation functions learned during the training phase. In this set of experiments, both two available demographic attributes are incorporated into the matching framework, simultaneously, where more GAM-based transformation functions need to be learned for the purpose of scalability investigation.

Take the Morph face database as an example. The subjects' gender information is man-

ually annotated as "Male" or "Female", while the race attribute has labels like "White" and "Non-White". Hence, each face image could be labeled as one of these four demographic characteristics: "Male&White", "Female&White", "Male&Non-White" and "Female&Non-White". After matching, there were 16 possible matching scenarios across the score matrix. Regarding each scenario, the proposed GAM scheme learned one transformation function according to the original match scores, x_i, and the corresponding recognition decision, "y_i'".

Figure 2.11 exhibits the matching accuracies before and after combining demographic attributes on the Morph face database and the WVU FL1 fingerprint database. The gender and race were first integrated via the GAM scheme, separately and then, jointly. It can be seen from both figures that, when the gender and race were separately combined with match scores, either of them was beneficial to the matching accuracy (marked as green and blue lines). When both attributes were jointly incorporated (marked as red lines), the matching accuracy were significantly improved compared to the baseline where only original match scores were used (marked as black lines). However, if we compared the combination of " face + gender" with the combination of "face + gender + race" (i.e., blue lines vs. red lines), the matching accuracy was not improved further by adding the race attribute. This observation inspires us to investigate the reason why fusing more demographic attributes is not guaranteed to improve the recognition accuracy via the GAM scheme (will be discussed in section 2.5.5).

Moreover, the scalability of the proposed GAM fusion scheme was investigated by integrating the demographic attribute with a "Uncertain" level. As mentioned, there is a built-in module in the face matcher COTS-A which can estimate the gender from the face image using machine learning algorithms. Additional to the gender estimation result, the module outputs a confidence value between 0 and 100 which indicates how much confidence to making this estimation. In this set of experiments, the estimated gender labels were categorized into 3 groups instead of binary classes. For instance, if a face image was estimated as "male" or "female" with a confidence value below 65, it would be labeled as "Uncertain". Then,

Table 2.6: The true match rates (TMRs) on Morph face and WVU face databases before and after integrating the gender attribute via the proposed GAM scheme. The match scores are generated using the COTS-B face matcher. The gender label is from: i) a direct collection (marked as "D") from subjects, ii) a manual annotation (marked as "M"), or iii) a built-in gender estimation module in COTS-A with binary outputs (marked as "2L") or 3-level outputs (marked as "3L").

Databases	Gender	FMR = 0.01%	FMR = 0.1%	FMR = 1%
Morph	no	88.2	94.0	96.8
	2L	92.7	97.2	97.5
	3L	92.9	97.2	97.5
WVU	no	96.6	98.4	99.0
	2L	98.0	99.2	99.5
	3L	97.9	99.2	99.4

the gender attribute with 3 levels (e.g., 32.9% "Male", 29.4% "Female" and 37.7% "Uncertain" for the Morph face database) was incorporated into the matching framework. There were 9 subordinate transformation functions learned from the training phase, and Table 2.6 summarizes the matching accuracy on the Morph face database and WVU face database before and after integrating this 3-level gender attribute via the proposed GAM scheme. The corresponding TMRs are calculated when the FMRs are fixed at multiple levels, i.e. $FMR = 0.01\%, 0.1\%, 1\%$.

It is evident that the GAM-based incorporation of gender attribute increased the matching accuracy on both databases, regardless of whether a demographic attribute was refined into more levels. The refinement of gender attribute has realistic scenarios, such as mitigating the challenging issue about LBGT people. However, compared the results from 2-level and 3-level gender attribute, the matching accuracy did not vary much for the proposed fusion scheme. One possible reason may relate to the sources of gender labels used in this experiment. It is noted that the "Uncertain" label relied on the confidence values provided by the built-in module in COTS-A, rather than a realistic gender collection procedure.

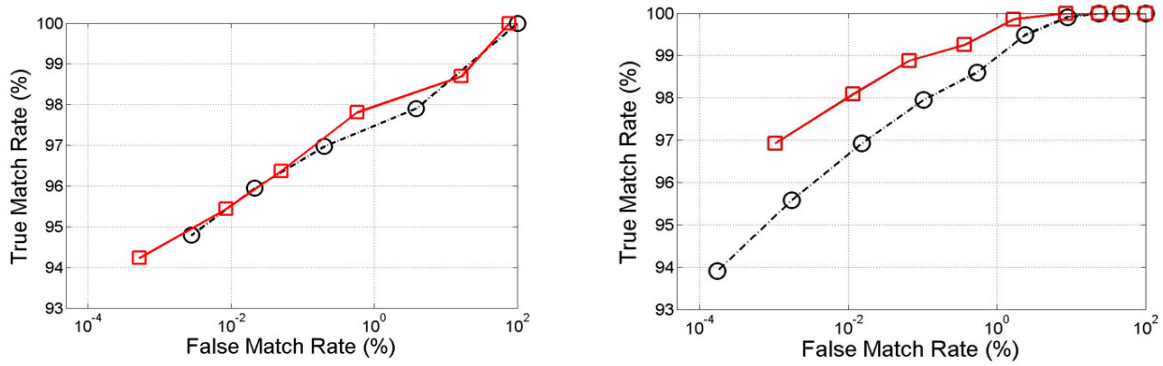


Figure 2.12: ROC curves on the WVU face database before (dashed lines) and after (solid lines) integrating the gender labels which are generated using a built-in gender estimation module in COTS-A. In figure (a) and (b), the match scores are generated using the COTS-A and COTS-B face matcher, separately.

2.5.5 Experiment 3. Predicting Gain

As discussed earlier about Figure 2.11, fusing more demographic attributes may not be able to improve the recognition accuracy via the GAM scheme. Besides, one more failure case is shown in Figure 2.12 (a), where the gender estimated by a built-in gender estimation module of the COTS-A is integrated with the match scores from the COTS-A. Both two cases demonstrate that it is not guaranteed that integrating demographic attributes can always improve the recognition accuracy. Hence, it will be beneficial to propose a metric that be able to predict whether the demographic labels can be used to the recognition accuracy of the biometric matcher before running the experiments on the entire test dataset.

The parameters of the GAM model offer some insights into how such a prediction metric can be derived. As a regression model, the GAM scheme provides certain diagnostic procedure which can be implemented to analyze the covariate/factor effects and the interaction effect between them.

The interaction effect is critical in the proposed GAM scheme. As expressed in the Eqn 2.20, if the interaction term, $\gamma_{z_i} z_i$, is not significantly different across the demographic classes, it is not reasonable to have different transformation functions for each matching

scenario and hence, the transformed scores cannot better separate the "Accept" and "Reject" classes and provide a better global matching accuracy. According to the R package *mgcv* [133] used for estimating the parameters in GAM, it provides a powerful tool which can test the interaction effect: tensor product model, $te(X, z)$. In this work, the tensor product model is implemented as:

```
Formula:
y  ~ s(X) + s(z) + te(X, z) ,
```

where $s(X)$ and $s(z)$ are isotropic smooths that produce spline functions, marginally. As pointed by Wood [133], the tensor product model would be preferred for modelling the interactions where very different degrees of smoothness are related to different levels in a factor. In other words, a significant interaction effect via the tensor product model can indicate the existence of significantly different score transformation functions among different levels of a demographic factor. Because of the space limitation, more details about the test of the interaction effect in a GAM scheme can be found in [133]. The package calculates the approximate significance of the above three smooth terms, and the P-Value of $te(X, z)$ is reported in Table 2.7. Compared with the results in the 3rd and 4th column of Table 2.7, it is noted that there is a clear connection between the performance improvement and the diagnostics of the interaction effect in GAM. Take the 1st row of Table 2.7 as an example. When the match scores from COTS-A were integrated with the race on the WVU face database, a significant interaction effect (P-Value is $1.70e-5$) was observed in the learned GAM model, which indicates that the parameters of the transformation functions among different race levels are quite different with each others. Therefore, the matching accuracy is increased (AUC was increased by 8.5%) when this learned GAM model is used for fusing the race and match scores.

It is observed that both two failure cases at improving accuracy (the 2nd row and 5th row) were using the gender labels automatically estimated via a gender estimation module in

Table 2.7: The P-Values generated from a statistical analysis on interaction effects in the GAM scheme. The highlighted P-Values denote the interaction effects between demographic factors and match scores are significant at the significance level 0.001.

Database	**Biometric Matcher + Demographics**	**Gain in AUC**	**Interaction (P-Values)**
WVU Face	COTS-A + race (D)	+ **8.5%**	**1.70e-5** < 0.001
	COTS-A + gender (L)	+ 0.4%	$0.156 > 0.001$
	COTS-B + gender (L)	+ **3.4%**	**1.96e-4** < 0.001
Morph Face	COTS-B + race (M)	+ **7.9%**	**1.07e-4** < 0.001
	COTS-A + gender (L)	+ 0.3%	$2.68 > 0.001$
WVU FL2	COTS-C + gender (D)	+ **6.5%**	**2.03e-5** < 0.001
	COTS-C + gender (L)	+ **6.3%**	**0.0001** < 0.001

COTS-A. Meanwhile, the match scores were also generated using the COTS-A. It is possible that the internal design of COTS-A caused the both failure cases of integration. Suppose the match scores from the COTS-A matcher has already reflected the gender estimation result of its built-in module, then the proposed GAM scheme cannot further improve the matching accuracy by adding the same estimated gender information. The observation has a realistic scenario when any arbitrary vendor of a biometric matcher wondered if their matching accuracy can be improved by integrating demographic attributes. They can simply implement the proposed GAM scheme on a training set with the match scores from their specific matcher, and the interaction effect in the learned GAM would indicate whether the integration with particular demographic attributes is beneficial or not.

2.5.6 Experiment 4. Robustness to Mislabeling Problem

Table 2.8 demonstrates the verification performance of the proposed GAM scheme on the Morph face database where 10% and 20% of the probe subjects' demographic labels were intentionally mislabeled. The performance is compared against the stratified matching scheme. In the stratified matching scheme, the verification performance would decrease sharply if the demographic labels of subjects are incorrect. For example, if the demographic label of the

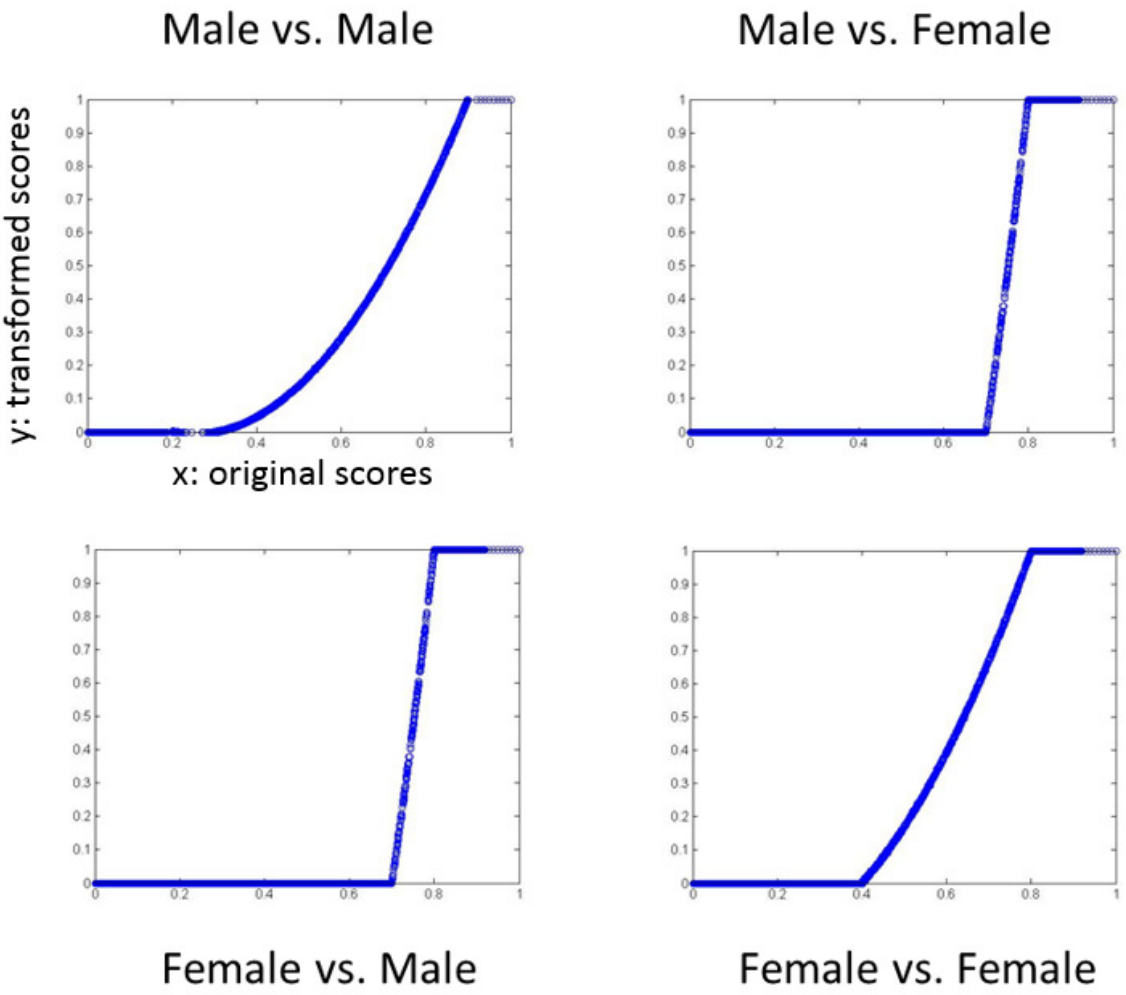


Figure 2.13: Transformation functions learned from the training set where the match scores from COTS-B are integrated with the gender information automated estimated via the gender estimation module in COTS-A. The automated gender estimation had an error rate around 12.0%.

Table 2.8: Matching accuracy of the proposed GAM when the demographic labels are incorrect. The proportion of mislabeled data in indicated in the left-most column.

State of gender Labels	TMR (%) at FMR = 0.01%	
	via Stratified Matching	via GAM
from manual annotation	88.5	92.6
10% are mislabeled intentionally	61.8 (-26.7)	92.5 (-0.1)
20% are mislabeled intentionally	44.2 (-44.3)	90.6 (-2.0)
from an automated estimation with $\approx 12.0\%$ errors	50.5 (-38.0)	92.2 (-0.4)

probe and the claimed template sample are different, the stratified matching scheme is likely to simply reject this probe without computing a match score. In that case, a mislabeled subject has no opportunity to generate a match score that may be high enough to overcome the threshold.

From the results in Table 2.8, it is noted that when the mislabeling rate is below 20%, the recognition accuracy proposed GAM scheme won't significantly decrease. In fact, when the mislabeling rate is above 25% (not been shown here), the recognition accuracy of proposed GAM scheme significantly decreases as well. Although the 25% mislabeling rate is not tolerable, the error rates of current demographic estimation algorithms could be much lower in practice. As pointed out by Sun et al.'s survey paper [119], although a number of challenging issues continue to inhibit its full potential, the error rates of recent demographic estimation algorithms can be controlled below 10%. Therefore, the proposed GAM provides a sufficient robustness in situations where the demographic data are incorrect or unreliable, especially for incorporating the demographic data generated by automated estimation algorithms. Figure 2.13 illustrates the learning-based transformation functions that were learned from the training set of the Morph face database with gender labels estimated by COTS-A gender estimation module. It is noted that for two samples with different demographic labels, their match score is generated and transformed into a new score according to the matching scenario (e.g., "Male vs. Male"). Plus, if their original match score is high enough, it can be retained at a relatively high value by the corresponding transformation function,

which avoid a potential false non-match case.

2.6 Summary and Future Work

Demographic attributes (such as gender, age, race), are potential to improve the performance of biometric matchers. While previous literature has studied the impact of these demographic factors on recognition performance, this work develops a principled approach to combine demographic data with biometric match scores that is applicable to the biometric verification scenario.

In this chapter, the GAM approach uses spline functions to model the relationship between match scores and demographic factors via a learning-based process. Compared to other fusion methods, the parameters of the transformation functions are optimized with respect, the matching accuracy, which results in a consistently better recognition performance (7.5% on average).

As a regression curve based approach, the resulting GAM framework can also be used to predict in advance if fusing match scores with certain demographic attributes is beneficial in the context of a particular biometric matcher. This advantage of GAM mitigates the concern associated with the issue of "lack of distinctiveness" encountered by integrating ancillary attributes which only contain a few discrete labels.

Moreover, experimental results conducted on databases where the demographic information are extracted using erroneous automated estimation algorithms, indicate that the resulting GAM framework pertain continues to be effective even in situations where the demographic labels are incorrect or unreliable. The experimental results on the MORPH face database and LFW face database suggest that the learned transformation functions are useful until the mislabeled training samples becomes greater than 30% of the entire training set. This suggests the reliability of the model even in situations when the damographic or ancillary labels are incorrect.

As future work, we plan to pursue this learning-based combining scheme in the following

ways:

- When the number of labels increases, the number of transformation functions in GAM increases rapidly, which impacts the computational complexity of the algorithm. One possible solution is to bring in a diagnostic procedure to the GAM learning stage. The main effect and interaction effect between demographic attributes and match scores need to be carefully analyzed before embodying them into the predictive model.

- When it comes to the fusion of match scores from multiple biometric traits with ancillary factors, the interactions among all the covariates can be incorporated in the GAM model.

CHAPTER 3

COMBINING ANTI-SPOOFING MEASUREMENTS WITH BIOMETRIC MATCH SCORES

3.1 Background

In the field of biometrics, a presentation attack occurs when an attacker presents a fake or modified biometric trait to the sensor [113, 128]. For instance, it has been shown that some fingerprint systems can be fooled by using a finger-like object fabricated using easily available materials such as latex, glue and gelatin (as shown in Figure 1.3), with the fingerprint ridges of another person inscribed on it [73].

Spoofing is an example of a presentation attack, where the adversary uses a fake or altered biometric trait with the intention of masquerading as another individual [113]. Such attacks pose a direct threat because they leverage commonly available materials and do not require any knowledge of the internal functionality of the underlying biometric authentication system. Fake biometric traits can also be used during the enrollment stage, especially in mobile applications where the enrollment process is not carefully monitored [23].

Spoof detection refers to the ability of a system to correctly distinguish between a legitimate, live human biometric presentation and spoof artifacts [129]. An *anti-spoofing measure*, as the output of most anti-spoofing schemes discussed in the literature, is a numerical value indicating the probability that the input biometric sample corresponds to a live human biometric presentation (i.e., *a liveness value*) or a spoof artifact (i.e., *a spoof score*) [113]. In this thesis, the *spoof score*, which indicates how likely a biometric sample is to be a spoof, is preferred. Specifically, biometric samples that are assigned less spoof score are less likely to be a spoof, and vice-versa.

The various anti-spoofing approaches proposed in the literature can be broadly classified into sensor-based and image-based solutions [70, 71]. Image-based spoof detection algorithms

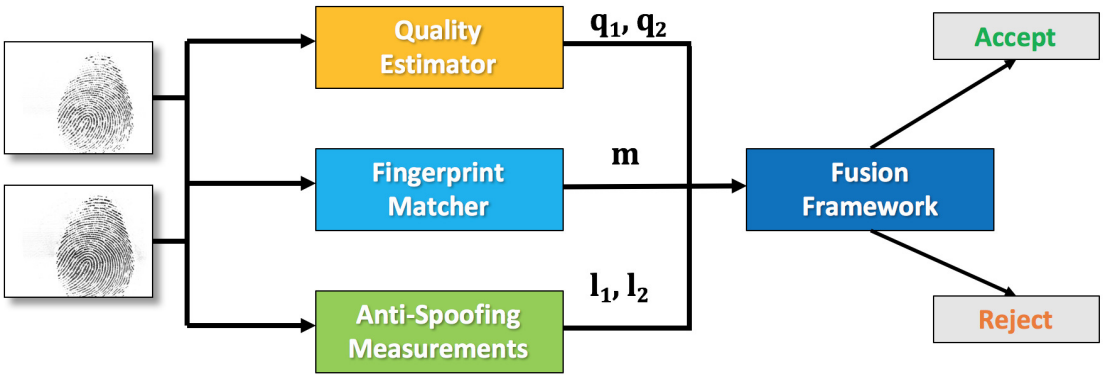


Figure 3.1: Illustration of the fusion framework integrating match scores with quality scores and anti-spoofing measures from two fingerprint samples, and rendering a final accept/reject decision.

have the advantage over sensor-based systems of being (1) less expensive (as no extra device is needed) and (2) less intrusive for the user [70, 79].

It must be noted that, as an inherent demand of system security, anti-spoofing methods are designed to be incorporated into biometric systems [68, 70]. The major contributions of this thesis is to design a novel fusion framework in which anti-spoofing approaches are incorporated into conventional biometric systems using a Bayesian Belief Network (BBN) framework. Additionally, the fusion framework is extended by incorporating image quality, another ancillary attribute which is impacted by the choice of fabrication materials used, to further improve anti-spoofing performance.

In this chapter, we first compare two commonly used fusion frameworks: sequential and parallel frameworks. The experimental results from three different methods, which do not explicitly model the interaction between match scores and anti-spoofing measures (i.e., spoof scores), are reported. Then, we propose a framework for combining match scores and the corresponding spoof scores based on a Bayesian Belief Network (BBN) model that assumes a certain influence of the spoof scores on match scores. Further, we investigate if the proposed BBN framework can improve the verification performance by adding more ancillary information, such as image quality of biometric samples. Figure 3.1 shows a block

diagram where image quality, anti-spoofing measures and match scores extracted from a pair of fingerprint images are integrated together in a fusion framework to render the final accept/reject decision.

3.2 Related Work

3.2.1 Feature Extraction for Anti-Spoofing

Liveness Detection Competitions (LivDet), which are aimed at comparing biometric spoof detection methodologies using a standardized testing protocol and large quantities of spoof and live samples, have been hosted in 2009, 2011, 2013 and 2015. The competitions are open to all academic and industrial institutions that have a software-based or system-based biometric spoof detection solution. They are shown themselves to provide a crucial look at the current state of the art in detection schemes [70, 85, 34, 98].

We take the reported fingerprint anti-spoofing algorithms in the Fingerprint LivDet as an example to review the related literature (as seen in Figure 3.2). Image-based spoof detection algorithms, in this work, have received more attention since they do not require the use of additional hardware and are based only on the images that are subsequently used by the fingerprint matcher. Generally speaking, existing fingerprint spoof detection algorithms extract **textural**, **coarseness**, **anatomical** or **physiological** attributes from live and fake fingerprint samples (as seen in Table 3.1).

In comparative evaluations on LivDet 2013 database [34], local textural features (such as LBP, LPQ, and BSIF) have been shown to outperform other competing anti-spoofing measures based on anatomical (such as pore detection [72]) and perspiration [2] as well as the algorithms submitted to the second liveness detection competition (LivDet 2011) held in 2011 whose error rates were in the range [20%, 40%].

Rattani and Poh [98] demonstrated the influence of fabrication materials on the obtained anti-spoofing measures of the fake fingerprints. Specifically, the probability distribution of the LBP-based anti-spoofing measures varied for five different fabrication materials: latex,

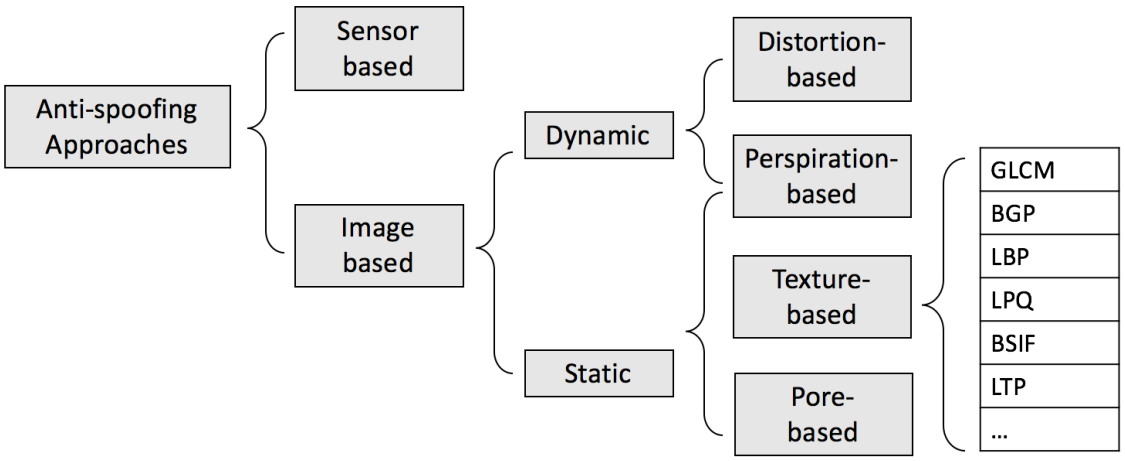


Figure 3.2: Taxonomy of existing fingerprint anti-spoofing algorithms.

Table 3.1: Examples of features that have been proposed for fingerprint spoof detection. A more detailed review can be found in [70].

Features	**Associated Studies**
Textural	Nikam and Agarwal's grey level co-occurence matrix (GLCM) [84]
	Ghiani et al.'s Local phase quantization (LPQ) [36]
	Ghiani et al.'s binary statistical image features (BSIF) [35]
	Nikam and Agarwal's local binary patterns (LBP) [83]
	Zhang et al.'s Binary gabor pattern (BGP) [140]
Anatomical	Espinoza and Champod's pore analysis for spoof detection [28]
	Marcialis et al.'s statistics related to fingerprint pore analysis [72]
	Tan and Schukers's fusion of ridge signal & valley noise analysis [121]
Perspiration	Marasco and Sansone's fusion of morphological and perspiration [67]
	Abhyankar and Schukers's perspiration analysis using wavelets [2]
Coarseness	Moon et al.'s coarseness analysis using noise residue [76]
	Coli et al.'s power spectrum analysis [17]
	Tan and Schukers's wavelet based statistics [120]

woodglue, silicone, gelatin and ecoflex. Figure 3.3 show variation in the spoof scores of live fingerprint sample (0.02), and the fake fingerprint samples fabricated using latex (0.22), ecoflex (0.45) and woodglue (0.69) for a subject in LivDet 2013. These anti-spoofing measures are obtained using LBP-based spoof detector [85]. Remind these studies mentioned above did not combine spoof scores obtained by the spoof detector with the fingerprint verification system.

Figure 3.3: Example of fake fingerprint images fabrication using latex, ecoflex and woodglue materials and the corresponding LBP-based anti-spoofing measures [85], in the LivDet 2011 database.

3.2.2 Compromised Templates

As each biometric system operation involve template-query pair for the comparison and decision making, *eight* possible events can occur during the system operation. In Table 3.2, these events are described based on the properties of template and query samples (i.e., $\mathbf{S}_i = \{\mathbf{L}, \mathbf{S}\}$ for $i \in \{1, 2\}$), whether they are from the same (genuine) or different identity (impostor) ($\mathbf{K} = \{\mathbf{G}, \mathbf{I}\}$) and the desirable classification decisions (i.e., Accept or Reject).

These cases are listed in detail as follows:

- The first 2 cases (**LLG** and **LLI**) show the property of genuine and impostor access considered in the traditional system. Among them **LLG** denote that template and query samples are live and belong to the genuine subject (i.e., genuine access). **LLG**

Table 3.2: Eight possible events during the biometric system operation for a pair of enrolled and input fingerprint image. These events are distinguished by the input state of the pair of fingerprint images, which can be live or spoof, and whether they are from the claimed identity or not. The desirable classification decisions are provided as well.

Case No.	Template State	Query State	Genuine or Impostor	Summary of State	Desirable Classification
1	**L**ive	**L**ive	**G**enuine	**LLG**	**A**ccept
2	**L**ive	**L**ive	**I**mpostor	**LLI**	**R**eject
3	**L**ive	**S**poof	**G**enuine	**LSG**	**R**eject
4	**S**poof	**L**ive	**G**enuine	**SLG**	**R**eject
5	**S**poof	**S**poof	**G**enuine	**SSG**	**R**eject
6	**L**ive	**S**poof	**I**mpostor	**LSI**	**R**eject
7	**S**poof	**L**ive	**I**mpostor	**SLI**	**R**eject
8	**S**poof	**S**poof	**I**mpostor	**SSI**	**R**eject

is the only desirable *accept* case. **LLI** denote the case that template and query sample are live but belong to different subjects (impostor access).

- The case 3 (**LSG**) illustrate the most common case of spoofing attacks where fake probe samples are compared against live enrollment sample of the claimed identity. These fake probe samples are the replica of the original fingerprint of the claimed identity. The cases 4 and 5 (**SLG** and **SSG**) are the most hazardous cases, where fake artifacts may be used to enroll the identity. Further, these fake fingerprint may be delegated to multiple individuals and then the system may be accessed using the live and fake fingerprint sample of the claimed identity.

- The last three cases (**LSI**, **SLI**, and **SSI**) consider the possibility of matching live and fake fingerprint samples belonging to different identities (**LSI** and **SLI**). The case **SSI** consider the possibility of matching a pair of fake fingerprint images belonging to different identities. All these cases do not adhere to the traditional definition of spoofing where the fake artifact is the replica of the original fingerprint image of the claimed identity. However, the likelihood of occurrence of these cases cannot be undermined.

Figures 3.4 show the example match score distributions of the LSG against LLG, from

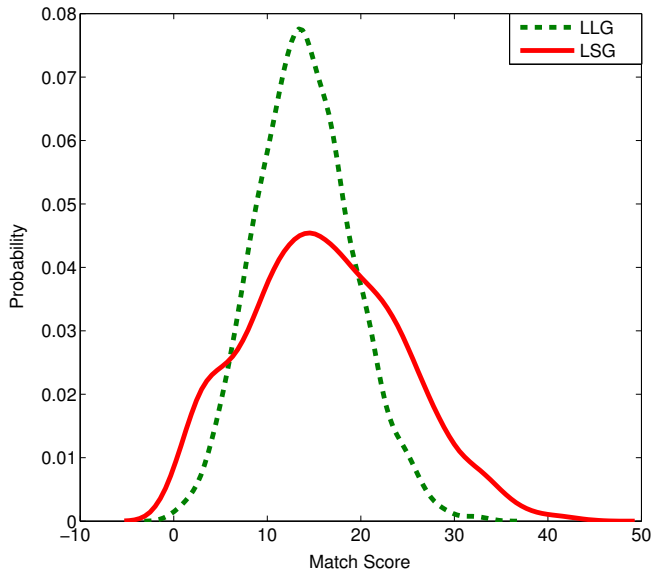


Figure 3.4: The match score distributions of the LSG and LLG state on samples acquired using **Biometrika** sensor in the LivDet 2011 database.

live and fake fingerprint images acquired using Biometrika, from the LivDet 2011 database. The high overlap in the match score distributions corresponding to LLG vs. LSG suggests that fingerprints can be effectively spoofed to gain illegitimate access to the system. Further, the match score distributions corresponding to the case SSG is quite similar to that of LLG (not shown here). Furthermore, the match score distribution of LSI is similar to that of LLI (can be seen in Figure 3.12). The same trend is observed for Italdata, Sagem, and DigitalPersona sensors.

3.2.3 Performance Evaluation Metrics

When distinguishing spoofs from live samples, the LivDets proposed the following performance metrics to evaluate the various anti-spoofing algorithms submitted to the competitions [85, 34]:

- Ferrlive: Percentage of misclassified live fingerprints.

- Ferrfake: Percentage of misclassified spoof fingerprints.

Further, EER of the spoof detection (indicated as S-EER) is the rate at which FerrLive is equal to Ferrfake.

However, in this work, because the compromised templates are considered, the LivDet proposed performance metrics cannot sufficiently evaluate all eight possible spoof detection scenarios. In stead, we proposed the following evaluation metrics from both spoof detection and global verification perspectives as:

- *Global verification*: When distinguishing genuine user from zero-effort impostors and spoof attacks. Accordingly, the errors of the system can be described as follows:
  - **False Reject Rate (FRR)**: Proportion of samples belonging to class LLG that are incorrectly classified as belonging to LLI, LSG, LSI, SLG, SLI, SSG, SSI. Genuine Acceptance Rate (GAR) is calculated as 1 - FRR.
  - **False Acceptance Rate (FAR)**: Proportion of samples belonging to class LLI that are incorrectly classified as belonging to LLG.
  - **Spoof False Acceptance Rate (SFAR)**: Proportion of samples belonging to classes LSG, SLG and SSG that are incorrectly classified as belonging to class LLG. Note that the classes LSI, SSI and SLI do not constitute spoof attacks according to the basic definition, they may be considered to evaluate the overall performance of the system.
- *Spoof detection*: When distinguishing spoofs from live samples.
  - Live Detection Rate (LDR): Percentage of correctly detected live samples. It is equivalent to 1 - Ferrlive for the cases where only the anti-spoofing performance were evaluated.

– Spoof Detection Rate (SDR): Percentage of correctly detected spoof samples. It is equivalent to 1 - Ferrfake for the cases where only the anti-spoofing performance were evaluated.

As a result, EER of the spoof detection (remained as S-EER) indicates the rate at which LDR is equal to SDR.

As the fingerprint verification system operates under both zero-effort impostor and spoof attacks, the overall error rates can be defined as follows:

- Genuine Acceptance Rate (GAR): Proportion of the **LLG** class that are incorrectly classified as genuine and accepted by the system.
- Overall False Acceptance Rate (OFAR): Proportion of zero-effort impostor and spoof samples that are incorrectly classified as the **LLG** class.
- Overall Equal Error Rate (O-EER): The rate at which OFAR equals 1 minus the Genuine Acceptance Rate (GAR). The **O-EER** of each fusion scheme is shown in the ROC curves.

Figure 3.13 shows the ROC Curves of the baseline performance of the fingerprint verification system under zero-effort impostors and spoof attacks. It is notable that:

- The EER of the baseline systems under zero-effort impostors (i.e., LLG vs. LLI) are in the range [2.2%, 5.1%] for the Biometrika, Italdata, Sagem and DigitalPersona sensors, respectively.
- The EER for the cases when the spoof artifact is the replica of the original fingerprint image of the claimed identity (i.e., LLG vs. LSG (SSG)) are in the range [29.4%, 54.1%] for the Biometrika, Italdata, Sagem and DigitalPersona sensors, respectively. Thus, demonstrating the hazard of the spoof attacks to the biometric system security.

- The case LLG vs. SSI obtain higher error rate than LLG vs. LSI, this is due to variation in the quality of spoof samples. Consequently, leading to high error rate when a pair of poor quality spoof images, belonging to different identities (SSI), are matched. This experiment emphasizes the urgent need for enhancing the security of the fingerprint verification system against spoof attacks.

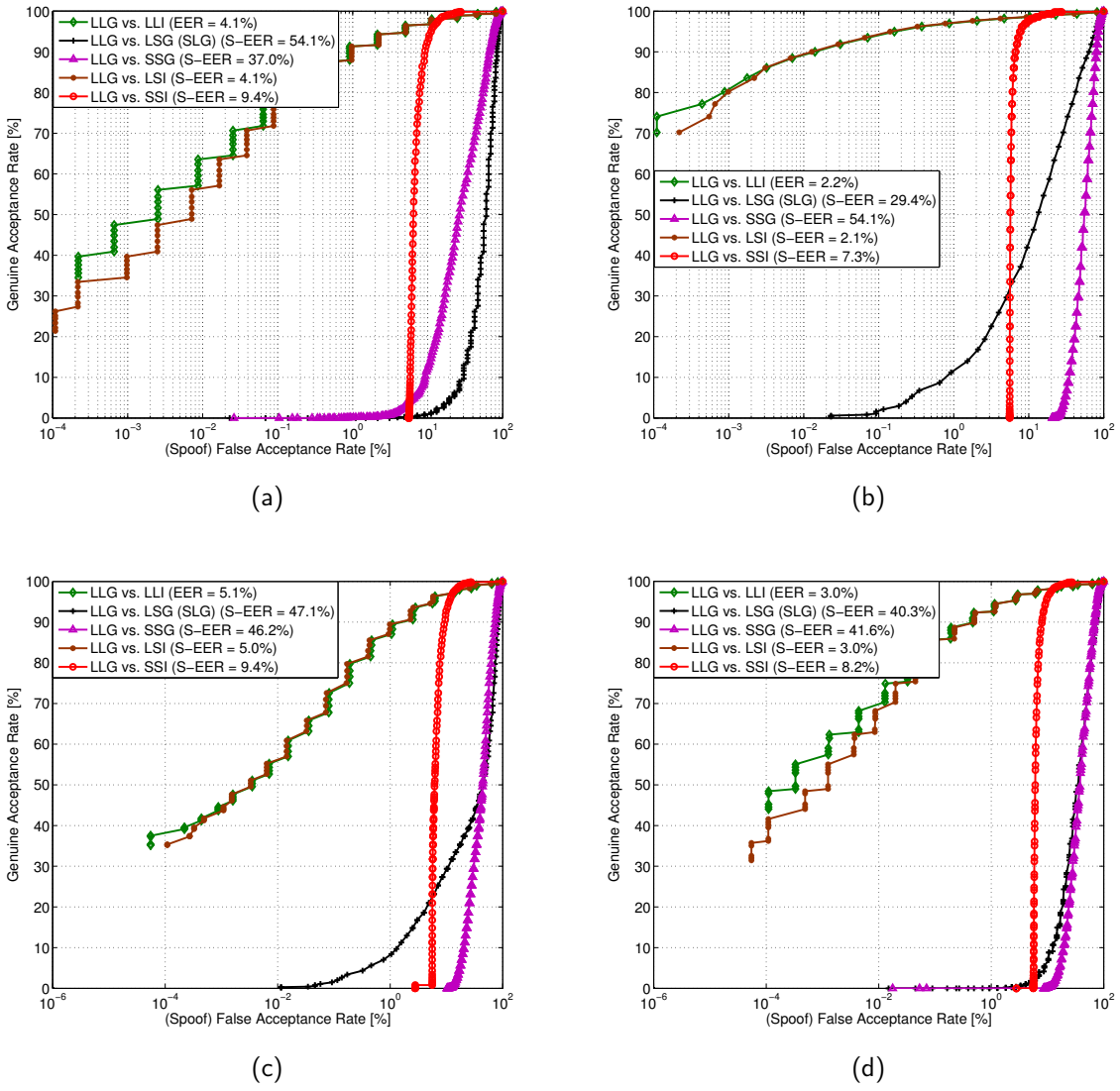


Figure 3.5: ROC Curves of the baseline performance of the fingerprint verification system under zero-effort impostors and spoof attacks.

3.3 Fusion Schemes: Sequential vs. Parallel

In this work, we assume that the matcher and spoof detector are "classifiers". The inputs to the matcher are two fingerprint samples (e.g., gallery and probe images). The output is a match score that indicates the proximity of the two samples. A threshold is applied to this match score to determine if the samples correspond to the same identity ("Genuine (G)") or different identities ("Impostor (I)"). Thus, the verification stage has two output classes: G and I. The input to the spoof detector is a fingerprint sample (e.g., gallery or probe image). The output is a spoof score indicating the degree of spoofness of the sample. A threshold is applied to this spoof score to determine if the sample is "Live (L)" or "Spoof (S)". Since there are two samples, spoof detection stage has four output classes: LL, LS, SL, SS. We consider various arrangements of the matcher and the spoof detector modules. Some configurations may not be operationally tenable - however, these have been considered only for completeness sake.

- In Method A, the classifier is invoked before the spoof detector as seen in Figure 3.6. The matcher in the first stage is used to distinguish genuine from impostor based only on match scores. In the spoof detection stage, there are two pairs of classifiers: one pair that is invoked if the input samples are deemed to belong to the Genuine (G) class and another that is invoked if they are deemed to belong to the Impostor (I) class. This arrangement may be redundant (i.e., the use of four different spoof detectors may not be necessary).

- In Method B, the spoof detector is invoked before the matcher as seen in Figure 3.7. Depending upon the output of the two spoof detectors in the first stage (LL, LS, SL or SS), one of four matchers in the verification stage is invoked. For example, the first matcher (Classifier 3) operates only on gallery and probe samples that are both classified as Live, while the fourth matcher (Classifier 6) operates only on the gallery and probe samples that are both classified as Spoof.

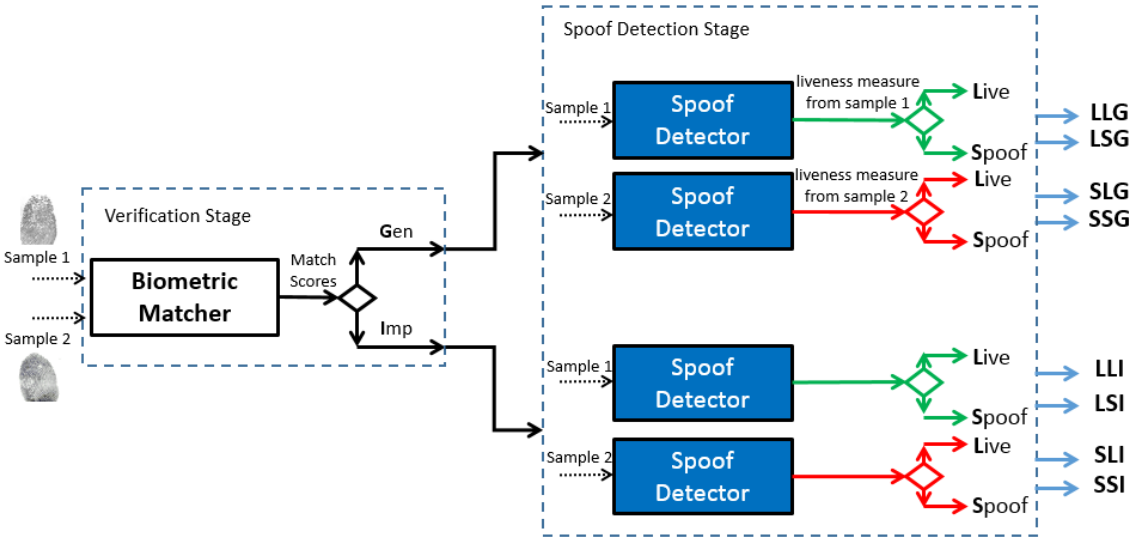


Figure 3.6: Architecture of Method A. Here, the matcher is invoked before the spoof detector. The classifier in the first stage (classifier 1) is used to distinguish genuine from impostor based only on match scores. There are two pairs of classifiers in the spoof detection stage. One pair classifiers (classifier 2 and 3) that are invoked if the input samples are deemed by the matcher to belong to the Genuine (G) class and another pair (classifier 4 and 5) that is invoked if they are deemed to belong to the Impostor (I) class. This arrangement may be redundant (i.e., the use of four different spoof detectors may not be necessary).

In Method C (see Figure 3.8, the match score, and the spoof scores are provided as inputs to a single classifier. This classifier has one of eight possible outputs: LLG, LSG, SLG, SSG, LLI, LSI, SLI, SSI. It can be considered as a multi-label problem. For each class label, the first two letters denote the input state of the samples ("Live" or "Spoof"), while the third letter denotes whether the samples correspond to the Genuine or Impostor class. In this method, no explicit assumption is made regarding a possible relationship between spoof scores and match scores.

The three methods described above do not explicitly model the relationship between spoof scores and match scores. A powerful framework for modeling causal relationships among a set of variables X is offered by graphical models such as Bayesian Belief Networks.

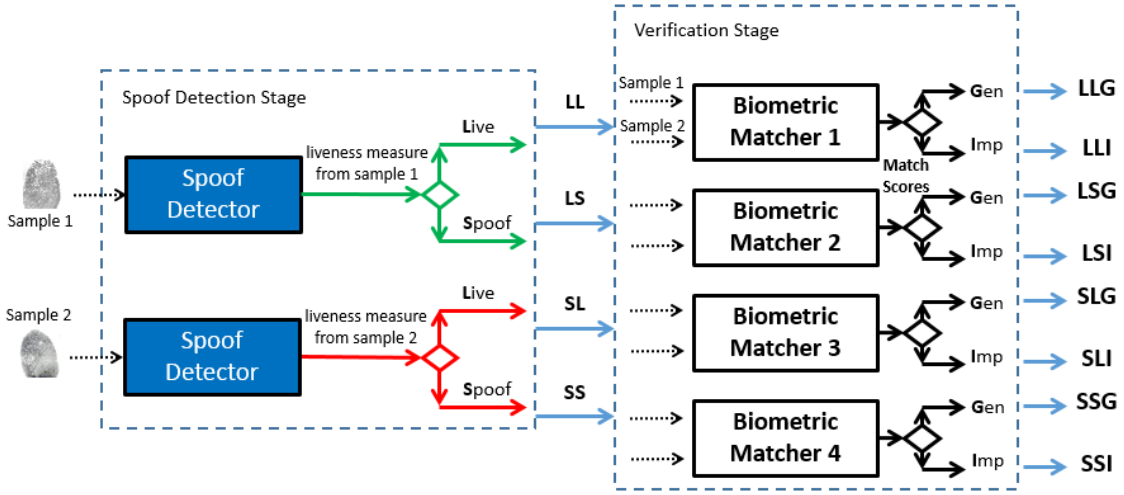


Figure 3.7: Architecture of Method B. Here, the spoof detector is invoked before the matcher. Depending upon the output of classifier 1 and 2 (LL, LS, SL or SS), one of four classifiers in the verification stage is invoked. For example, classifier 3 operates only on input scores between gallery and probe samples that are both classified as Live, while classifier 6 operates only on scores between gallery and probe samples that are both classified as Spoof.

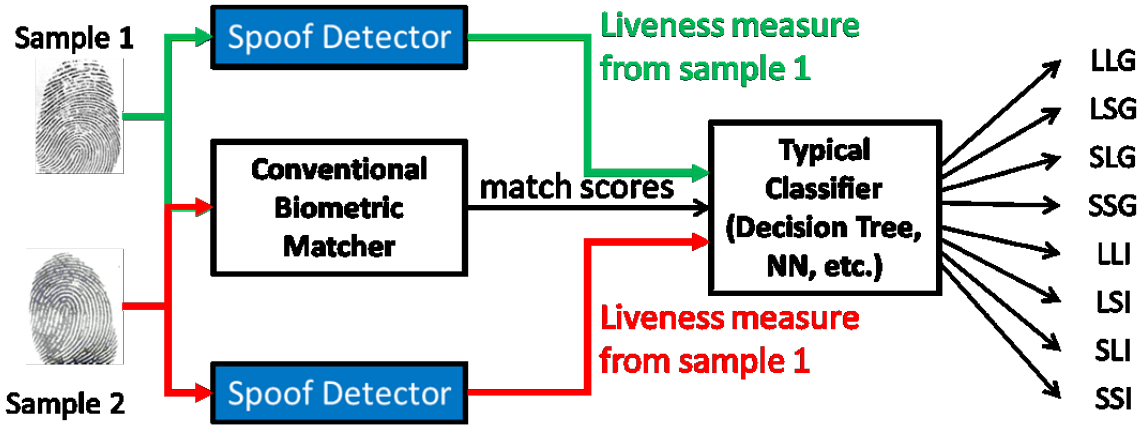


Figure 3.8: Architecture of Method C. Here, the classifier has three inputs: match score, spoof scores of gallery sample and spoof scores of probe sample. All 3 inputs are used simultaneously in order to determine the output class.

3.4 Bayesian Belief Networks in Biometrics

Classification is a basic task in data analysis and pattern recognition that requires the construction of a classifier, that is, a function that assigns a class label to instances described by a set of attributes. The induction of classifiers from data sets of pre-classified instances is a central problem in machine learning. Numerous approaches to this problem are based on various functional representations such as decision trees, decision lists, neural networks, decision graphs, and rules. One of the most effective classifiers, in the sense that its predictive performance is competitive with state-of-the-art classifiers, is the so-called naive Bayesian classifier described, for example, by Duda and Hart [24] and by Langley et al. [57]. This classifier learns from training data the conditional probability of each node X_i given the class label C as seen in Figure 3.9.

Classification is then done by applying Bayes rule to compute the probability of C given the particular instance of $X_1, \ldots, X_n$, and then predicting the class with the highest posterior probability. This computation is rendered feasible by making a strong independence assumption: all the attributes X_i are conditionally independent given the value of the class C. By independence we mean probabilistic independence, that is, A is independent of B given C whenever P(A|B,C) = P(A|C) for all possible values of A, B and C, whenever $P(C) > 0$.

Compared to the naïve Bayes classifier, the Bayesian belief network (BBN) classifier can often offer better performance by avoiding unwarranted (by the data) assumptions about independence. A BBN is a probabilistic graphical model that represents a set of random variables and their conditional dependencies via a directed acyclic graph (DAG). For example, a Bayesian network could represent the probabilistic relationships between diseases and symptoms. Given symptoms, the network can be used to compute the probabilities of the presence of various diseases. Formally, Bayesian networks are DAGs whose nodes represent random variables in the Bayesian sense: they may be observable quantities, latent variables, unknown parameters or hypotheses. Edges represent conditional dependencies; nodes that

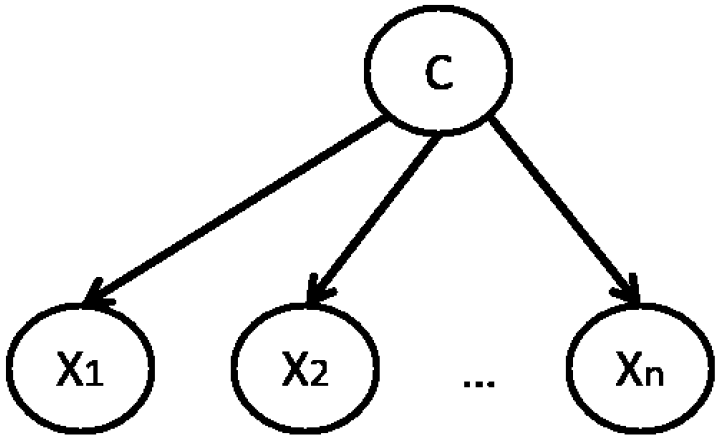


Figure 3.9: A simple example of Bayesian Network structure

are not connected (there is no path from one of the variables to the other in the Bayesian belief network) represent variables that are conditionally independent of each other. Each node is associated with a probability function that takes, as input, a particular set of values for the node's parent variables, and gives (as output) the probability (or probability distribution, if applicable) of the variable represented by the node.

Why BBN can provide a better performance than naïve Bayes classifier? It is mainly because that using the independence statements encoded in the network, the joint distribution is uniquely determined by these local conditional distributions. Consider a finite set $U = \{X_1, \ldots, X_n\}$ of discrete random variables where each variable X_i may take on values from a finite set. Formally, a BBN for U includes two components: the graph encoding the independence assumptions, and the set of parameters that quantifies the network. The joint probability function of U can be specified as:

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i | parents(X_i))$$

And from the configuration of Figure 3.9, the joint probability function is:

$$P(X_1, \ldots, X_n, C) = P(C) \prod_{i=1}^{n} P(X_i | C)$$

In this section, we first introduce the existing BBNs which have been implemented in the context of biometrics. Then, three extended BBN configurations are proposed and compared, theoretically. The experimental results of these BBNs are presented in the later sections. The notation used in this chapter is listed as follows:

Notation: Let the observation be $\mathbf{x} = [m, l_1, l_2, q_1, q_2]$ where $m \in \mathbb{R}$ is a fingerprint match score, $l_1 \in \mathbb{R}$ ($l_2 \in \mathbb{R}$) denotes liveness scores of the gallery sample (probe sample), and $q_1 \in \mathbb{R}$ ($q_2 \in \mathbb{R}$) is the quality value of the gallery sample (probe sample). Let $\mathbf{K} = \{\mathbf{G}, \mathbf{I}\}$ denote two possible outputs: genuine (two fingerprint samples are from the same finger) and impostor (two fingerprint samples are from different fingers). Note that $\mathbf{K}$ does not include any assumptions about whether the pair of matched samples are live or fake. Further, let S_1 and S_2 denote the liveness states of the gallery and probe samples, which can be either **L**ive or **S**poof, i.e., $\mathbf{S}_i = \{\mathbf{L}, \mathbf{S}\}$ for $i \in \{1, 2\}$. Thus, the output of a fingerprint matcher working in conjunction with a spoof detector can result in 8 possible events $\{\mathbf{S}_1, \mathbf{S}_2, \mathbf{K}\}$: **LLG**, **LLI**, **LSG**, **LSI**, **SLG**, **SLI**, **SSG**, **SSI**.

3.4.1 Existing Bayesian Belief Networks

In the context of biometrics, a conventional generative classifier attempts to model the match scores (m) conditioned on the ground truth of the image pair being compared ($\mathbf{K}$), i.e., $p(m|\mathbf{K})$. The BBN model representing this conventional classifier is denoted as $\mathbf{K} \rightarrow m$. This conventional classifier can be extended to include all eight events and can be effectively realized using likelihood ratio-based test statistics as in Eqn (3.1). This conventional classifier, referred to as BBN-M, is considered as one of the baseline classifiers in this work.

$$f^{llr} = \frac{p(\mathbf{LLG}|m)}{p(\sim \mathbf{LLG}|m)}. \tag{3.1}$$

Model (a) **BBN-MQ**: Figure 3.10 (a) show the BBN model proposed in [74] that combined fingerprint match score (m) with the image quality (q_1 and q_2). The model is based on the

following assumption: Assumption: *Quality measure of a sample influences the corresponding match score.*

One advantage of BBN is to explicitly depict the dependence between predictor variables, such as the match scores and the quality scores from two fingerprint samples, by the prior knowledge or the causal understanding from a human perspective rather than just the data. Take the quality scores of the gallery and probe sample (i.e., q_1 and q_2) as an example. Firstly, the variable q_1 and q_2 are supposed to be independent (denoted as $q_1 \perp\!\!\!\perp q_2$), because two fingerprint samples can be arbitrary. Moreover, they can be assumed to influence the match score m (denoted as $q_1 \rightarrow m$ and $q_2 \rightarrow m$) from a causal understanding, but they are expected to be independent with the ground truth $\mathbf{K}$. This is because the ground truth of two fingerprint samples being from the same finger or from two different fingers cannot be influenced by the quality scores of these samples [1] This advantage is further discussed regarding the calculation of likelihood ratio-based test statistics below.

As a summary, the assumption is shown as $q_i \rightarrow m$ for $i \in \{1, 2\}$ in Figure 3.10 (a), and the joint density represented by the **BBN-MQ** model can be directly calculated as,

$$p(\mathbf{K}, q_1, q_2, m) = p(m|\mathbf{K}, q_1, q_2)p(q_1)p(q_2)p(\mathbf{K}). \tag{3.2}$$

Since this model does not consider spoof attacks, the conditional probability of $\mathbf{K} = \{\mathbf{G}, \mathbf{I}\}$ does not include the liveness states ($\mathbf{S}_1$ and $\mathbf{S}_2$). The final decision $\mathbf{K} = \{\mathbf{G}, \mathbf{I}\}$ is made based on the likelihood ratio-based test statistic (f^{llr}) as follows:

$$\begin{aligned} f^{llr} &= \frac{p(\mathbf{K}{=}\boldsymbol{G}|m, q_1, q_2)}{p(\mathbf{K}{=}\boldsymbol{I}|m, q_1, q_2)} = \frac{p(\mathbf{K}{=}\boldsymbol{G}, m, q_1, q_2)}{p(\mathbf{K}{=}\boldsymbol{I}, m, q_1, q_2)} \\ &\quad \textit{(based on Eqn (3.2) and since } \boldsymbol{K} \perp\!\!\!\perp q_1 \perp\!\!\!\perp q_2\textit{)} \\ &= \frac{p(\mathbf{K}{=}\boldsymbol{G})p(m, q_1, q_2|\mathbf{K}{=}\boldsymbol{G})}{p(\mathbf{K}{=}\boldsymbol{I})p(m, q_1, q_2|\mathbf{K}{=}\boldsymbol{I})}. \end{aligned} \tag{3.3}$$

Assuming the prior probability of $p(\mathbf{K} = \mathbf{G})$ and $p(\mathbf{G} = \mathbf{I})$ are equal, the above f^{llr} can be obtained by estimating the joint probability of $\{m, q_1, q_2\}$ given the target class $\mathbf{K}$.

[1] Of course, there could be cases where a person's fingerprint is consistently poor due to implicit skin issues.

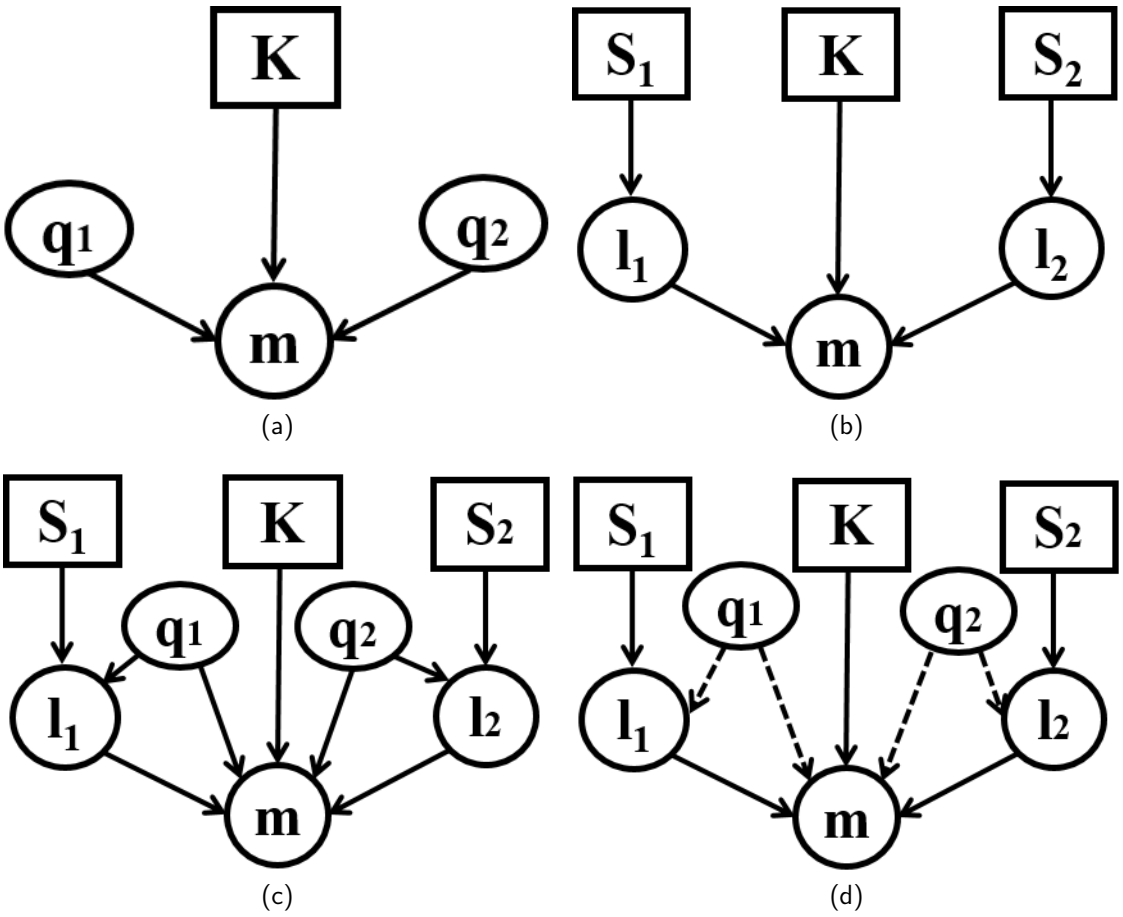


Figure 3.10: Several possible BBNs for fusing fingerprint match scores with liveness and quality scores. BBN-MQ and BBN-ML are based on previous literature, while BBN-MLQ and BBN-MLQc are the proposed ones.

3.4.2 Proposed Bayesian Belief Networks

Model (b) **BBN-ML**: Figure 3.10 (b) show the BBN model proposed by Marasco et al. [68] for combining match scores (m) with the corresponding liveness scores (l_1 and l_2). BBN-ML is based on the following assumption:

Assumption: *liveness scores of a sample influences the corresponding match score.*

As mentioned before, for spoof detection, the variables $\mathbf{S}_1$ and $\mathbf{S}_2$ represent the ground truth of the states of liveness of the two fingerprint samples. If $\mathbf{S}_i$ is a spoof, then it will likely result in a lower liveness value l_i. The BBN model representing the relationship

between liveness state $\mathbf{S}_i$ and the liveness scores l_i is denoted as $\mathbf{K} \rightarrow m$ and $\mathbf{S}_i \rightarrow l_i$. This assumption is shown as two directional arrows (i.e., $l_i \rightarrow m$) in Figure 3.10 (b). The joint densities represented by the **BBN-ML** model can be written as:

$$\begin{aligned} & p(\mathbf{K}, \mathbf{S}_1, \mathbf{S}_2, m, l_1, l_2) \\ & = p(m|\mathbf{K}, l_1, l_2)p(l_1|\mathbf{S}_1)p(l_2|\mathbf{S}_2)p(\mathbf{K})p(\mathbf{S}_1)p(\mathbf{S}_2). \end{aligned} \tag{3.4}$$

This BBN-ML representation is also referred to as Method D corresponding to the methods discussed in Section 3.3. For spoof detection, the variables S_1 and S_2 represent the ground truth of the input states of two fingerprint samples. If S_i is a spoofing, then it will be more likely to obtain higher spoof scores l_i. The BBN model representing the relationship between state S_i and the anti-spoofing measure l_i are shown in Figure 3.10 (b). These basic causal relationships are shown as $K \rightarrow m$ and $S_i \rightarrow l_i$ in all the BBNs discussed in this section.

For example, a match score between two samples of different individuals ($K = 1$) is likely to be lower than that of samples coming from the same individual ($K = 0$). The variables S_0 and S_1 represent the events related to the presence of a spoof biometric presentation at enrollment and verification times, respectively. The variables l_1 and l_2 denote the spoof scores of the gallery and probe samples, respectively. In the proposed method, we assume that the spoof scores l_1 and l_2 influence the corresponding match score, m. The interactions among the involved variables are based on the idea that the events S_1, S_2 and K influence a common effect, i.e., the decision made by the biometric system, through variables l_1, l_2 and m. We study how the impact of the event K on the final decision depends on the other events S_1 and S_2. This approach has one of eight possible outputs: LLG, LSG, SLG, SSG, LLI, LSI, SLI, SSI.

The computational paradigm of Bayesian Networks is based on probabilistic evidence where new evidence has to be propagated to other parts of the network. When performing Bayesian inference, a combination of observed data with prior knowledge is required. In

our study, we seek to integrate the biometric matcher, the spoof detector, and prior of the three distributions $P(K)$, $P(S_1)$ and $P(S_2)$. In the Bayesian Network model, all the conditional probabilities are given and the goal is to determine the maximum posterior value of the unknown variables in the network, through careful application of the Bayes rule [25, 136]. The joint probability distribution, represented as $P(K, S_1, S_2, x, l_1, l_2)$, is factorized according to the structure of the network, as follows:

$$
\begin{aligned}
p(\boldsymbol{LLG}|m, l_1, l_2) &\rightarrow \frac{p(\mathbf{K}{=}\boldsymbol{G}, \mathbf{S}_1{=}\boldsymbol{L}, \mathbf{S}_2{=}\boldsymbol{L}, m, l_1, l_2)}{p(m, l_1, l_2)} \\
&\textit{(from Eqn (3.4))} \\
&= \frac{p(m|\mathbf{K}, l_1, l_2)p(l_1|\mathbf{S}_1)p(l_2|\mathbf{S}_2)p(\mathbf{K})p(\mathbf{S}_1)p(\mathbf{S}_2)}{p(m|l_1, l_2)p(l_1, l_2)} \\
&\textit{(since } l_1 \perp\!\!\!\perp l_2\textit{)} \\
&= \frac{p(\mathbf{S}_1)p(l_1|\mathbf{S}_1)}{p(l_1)} \; \frac{p(\mathbf{S}_2)p(l_2|\mathbf{S}_2)}{p(l_2)} \; \frac{p(\mathbf{K})p(m|\mathbf{K}, l_1, l_2)}{p(m|l_1, l_2)} \\
&\textit{(since } \boldsymbol{K} \perp\!\!\!\perp l_1 \textit{ and } \boldsymbol{K} \perp\!\!\!\perp l_2\textit{)} \\
&\rightarrow p(\mathbf{S}_1{=}\boldsymbol{L}|l_1)\; p(\mathbf{S}_2{=}\boldsymbol{L}|l_2)\; p(\mathbf{K}{=}\boldsymbol{G}|m, l_1, l_2). \qquad (3.5)
\end{aligned}
$$

The final decision is made using the likelihood ratio based test statistic (f^{llr}) of the conditional probability of eight possible events (classes) given the match score (m) and liveness scores (l_1 and l_2). Taking the only acceptance case[2] It must be noted that the above mathematical derivation can simplify the calculation of the likelihood ratio (f^{llr}) between the classes **LLG** and $\sim$ **LLG**.

The above equation shows that the proposed BBN can be considered as being composed of three independent components. The first two terms indicate that both the gallery and probe samples are classified as being live or spoof based only on their spoof scores. The third term indicates that the input biometric presentation is classified as being genuine or impostor based on both match scores and spoof scores.

[2]The *acceptance case* indicates the event where the two samples are live and they originate from the same finger.

As discussed earlier, the proposed Bayesian Belief Network (BBN) based fusion frameworks overcomes multiple conventional directly modeled classifiers when combining anti-spoofing measures with match scores. In the following two method, we extended the BBN-ML with more input variables, which are the image quality measurements from the probe sample and claimed template sample, referred to as BBN-MLQ and BBN-MLQc. **Quality Scores** are commonly used for indicating how good the quality of a biometric sample is. These scores could be numerical values or categorical values depending on the definitions and metrics that are used. The lack of a uniform standard requires the design of a fusion framework that is resilient to inaccurate or uncertain quality measures when integrating them with biometric match scores. In this section, we proposed two different configurations of BBN to combine the quality scores. Experimental results show that the BBN-MLQc where clustering the continuous quality scores prior to fusion consistently obtain lower error rates over existing frameworks from two perspectives: (i) anti-spoofing capability, and (ii) verification of an identity.

Model (c) **BBN-MLQ**: Figure 3.10 (c) shows one of the proposed BBN model that combines match scores with quality and liveness scores. This model is based on the following three assumptions:

Assumption 1: *Quality measure of a sample influences the corresponding match score, i.e.,* $q_i \rightarrow m$

Assumption 2: *liveness scores of a sample influences the corresponding match score, i.e.,* $l_i \rightarrow m$

Assumption 3: *Quality measure of a sample influences the corresponding liveness scores, i.e.,* $q_i \rightarrow l_i$.

The joint probabilities represented by **BBN-MLQ** are factorized as:

$$\begin{aligned} & p(\mathbf{K}, \mathbf{S}_1, \mathbf{S}_2, m, l_1, l_2, q_1, q_2) \\ & \quad = p(m|\mathbf{K}, l_1, l_2, q_1, q_2) p(l_1|\mathbf{S}_1, q_1) p(l_2|\mathbf{S}_2, q_2) \\ & \quad p(\mathbf{K}) p(\mathbf{S}_1) p(\mathbf{S}_2) p(q_1) p(q_2). \end{aligned} \tag{3.6}$$

BBN-MLQ can be realized using the likelihood ratio based test statistic (f^{llr}) as follows:

$$\begin{aligned} f^{llr} &= \frac{p(\mathbf{LLG}|m, l_1, l_2, q_1, q_2)}{p(\sim \mathbf{LLG}|m, l_1, l_2, q_1, q_2)} \\ & \textit{(from Eqn (3.6) and since } \boldsymbol{K} \perp\!\!\!\perp \boldsymbol{S}_1, \boldsymbol{S}_2\textit{)} \\ &= \frac{p(m|\mathbf{K}{=}\boldsymbol{G}, l_1, l_2, q_1, q_2) p(l_1|\mathbf{S}_1{=}\boldsymbol{L}, q_1) p(l_2|\mathbf{S}_2{=}\boldsymbol{L}, q_2) p(\boldsymbol{LLG})}{\sum_{\sim \boldsymbol{LLG}} p(m|\mathbf{K}, l_1, l_2, q_1, q_2) p(l_1|\mathbf{S}_1, q_1) p(l_2|\mathbf{S}_2, q_2) p(\sim \boldsymbol{LLG})} \\ & \textit{(since } \boldsymbol{K} \perp\!\!\!\perp l_1, l_2, q_1, q_2\textit{)} \\ &= \frac{p(m, l_1, l_2, q_1, q_2|\mathbf{K}{=}\boldsymbol{G}) p(l_1, q_1|\mathbf{S}_1{=}\boldsymbol{L}) p(l_2, q_2|\mathbf{S}_2{=}\boldsymbol{L}) p(\boldsymbol{LLG})}{p(m, l_1, l_2, q_1, q_2|\mathbf{K}) p(l_1, q_1|\mathbf{S}_1) p(l_2, q_2|\mathbf{S}_2) p(\sim \boldsymbol{LLG})}. \end{aligned} \tag{3.7}$$

The configuration of this BBN model can be considered as a direct extension of BBN-ML by adding quality scores as new predictor variables. Although the inference of the model is straightforward, the influence of latent factors has not been considered. As a result, we propose another configuration of the BBN model to utilize the quality scores in a more effective way.

Model (d) **BBN-MLQc**: Figure 3.10 (d) shows another configuration of the BBN model. This model is based on the fact that a simpler BBN configuration with fewer assumptions is more likely to generalize over unseen data. This is because additional assumptions of causal relationships can lead to a more complex joint probability function (such as in Eqn (3.6)) which may be difficult to estimate and interpret. Therefore, this model incorporates quality scores into the existing **BBN-ML** model without making any additional assumptions, while the match scores and liveness scores are calibrated/normalized based on the quality measure. The model is referred to as **BBN-MLQc** in this work, and the assumption made in this model is as same as the one made in the **BBN-ML** model:

Assumption: *liveness scores of a sample influences the corresponding match score.*

The conditional probability can be estimated in a manner similar to the **BBN-ML** model:

$$p(\mathbf{K}{=}\boldsymbol{G}, \mathbf{S}_1{=}\boldsymbol{L}, \mathbf{S}_2{=}\boldsymbol{L} \mid m^{norm}, l_1^{norm}, l_2^{norm}) \tag{3.8}$$
$$= p(\mathbf{S}_1{=}\boldsymbol{L}|l_1^{norm})\ p(\mathbf{S}_2{=}\boldsymbol{L}|l_2^{norm})\ p(\mathbf{K}{=}\boldsymbol{G}|m^{norm}, l_1^{norm}, l_2^{norm})$$

where m^{norm} and l_i^{norm} for $i \in \{1, 2\}$ are the quality-normalized match scores and liveness scores, respectively. The proposed quality-based calibration is based on the following observations:

1. Similar quality scores are likely to share a similar combination of factors, such as image resolution, noise level, clarity of ridges/valley structures, or fabrication materials used. Quality categorization can, therefore, capture these latent factors.

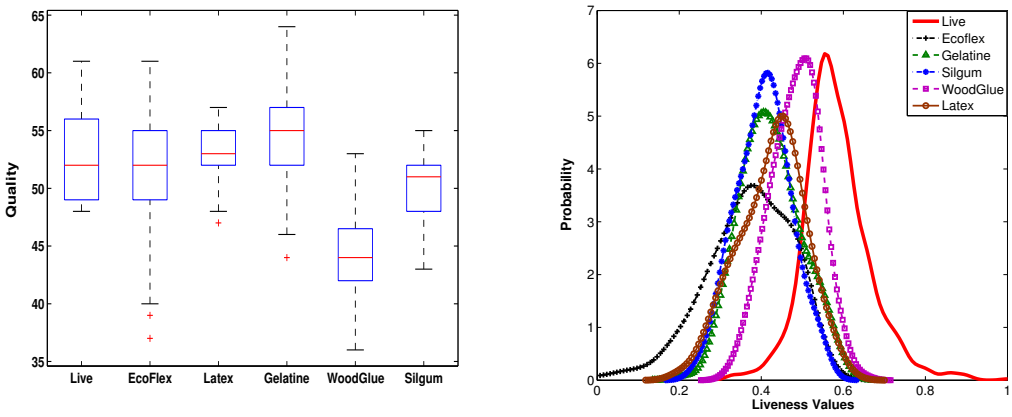
2. Certain liveness scores may result in higher spoof detection accuracy than others. In such cases, the quality measure of the biometric samples can be ignored by the spoof detector. This suggests the use of a piecewise function to calibrate liveness scores by the quality measure only over certain ranges.

The rationale behind the proposed **BBN-MLQc** model is to categorize the quality measure into discrete states, and then apply different calibration functions for each quality based on the spoof detection accuracy.

The categorization (or discretization) of continuous quality scores is achieved using the Minimum Optimal Description Length (MODL) algorithm based on the minimal description length (MDL) principle [9]. The class entropy of a set of quality scores q is defined as:

$$Ent(q) = -\sum_{i=1}^{Z} p(c_i, q) log(p(c_i, q))), \tag{3.9}$$

where $p(c_i, q)$ is the proportion of samples lying in category c_i, and Z is the total number of categories. Suppose the first bin B_1 is added as a cut-off point and the set q is partitioned



(a) (b)

Figure 3.11: Boxplot of quality scores and probability distribution of the liveness scores for five different materials in the LivDet 2011 database when the **Biometrika** sensor is used. A similar observation can be made for Italdata, Sagem and Digital sensors as well.

into subsets q_{c_1} and q_{c_2}, then the entropy of the partition is:

$$Ent(q, B_1) = \frac{|q_{c_1}|}{|q|} Ent(q_{c_1}) + \frac{|q_{c_2}|}{|q|} Ent(q_{c_2}), \tag{3.10}$$

where $|q|$ denotes the number of samples in the set q. There could be $Z-1$ bins. The original MODL algorithm in [9] scores all possible categorization possibilities and selects the one with the lowest entropy, and is also employed to decide the number of categories Z in this work.

The quality categorization is followed by an exploration of optimal calibration functions for liveness scores. There are multiple ways to transform the liveness scores using quality. In this work, the basic Fisher's linear discriminant analysis (LDA) is employed. The calibrated liveness scores can be considered as a linear combination of variables (l, q).

$$l_i^{norm} = \begin{cases} l_i & i \in c_1, \\ f_{LDA}^{c_i}(l_i, q_i) & i \in c_{\{2,\ldots Z\}}. \end{cases} \tag{3.11}$$

Basically, Eqn (3.11) indicates that if the samples lie in the quality state c_1 - corresponding to the quality state obtaining the highest spoof detection accuracy - the liveness scores

do not need to be calibrated by any image quality. However, if the samples lie in other quality states, the liveness scores are calibrated using $f_{LDA}^{c_i}$, where c_i denotes the corresponding quality state. The output classes used for training the LDA functions are **L**ive or **S**poof, i.e.., $\mathbf{S}_i = \{\mathbf{L}, \mathbf{S}\}$ for $i \in \{1, 2\}$.

It should be noted that the above quality-based calibration is non-linear with respect to the liveness scores, and the estimation of the joint probability function represented by the proposed BBN is greatly simplified by the calibration process. In this chapter, we focus on the configurations of multiple Bayesian Belief Networks. Practical scenarios of combining quality states with conventional biometric systems can be found in Section 5.2.1.

3.5 Databases and Protocol

There are two major parts of experiments conducted on two database, separately. First, the performance of the sequential methods (Method A and B), parallel methods (Method B with options), and Bayesian Belief Network (Method D with options) was evaluated on a subset of the CrossMatch database taken from the Fingerprint Liveness Detection Competition 2009 [34]. It is made up of live and spoof fingerprint samples imaged using a CrossMatch optical scanner with a resolution factor of 500 dpi and an image size of 480x640 pixels. Two spoof materials were considered in our experiments: gelatin and silicone. Match scores were extracted using the VeriFinger SDK software by matching all pairs of images across all subjects. The scores, therefore, correspond to four different matching scenarios: Live vs Live, Live vs Spoof, Spoof vs Live, and Spoof vs Spoof. For each image, the spoof scores was extracted by using an algorithm which combines morphological and perspiration-based characteristics [34].

The verification performance of the fingerprint recognition system is analyzed using the Receiving Operating Characteristic (ROC) curves. ROC curves are obtained for both spoof materials under the four different matching scenarios (live-live, live-spoof, spoof-live and spoof-spoof). On the CrossMatch database, the spoof scores seems to be better in detecting

spoof samples made with gelatin and poor in detecting spoof samples made with silicone (as shown in Figure 3.4. So the spoof detection has higher reliability in the case of gelatin and lower reliability in the case of silicone.

Besides of the ROC curves, the comparison of four methods are conducted at a practical operating threshold (as shown in Table 3.4). The sequential methods (Method A and Method B) require a threshold, i.e., the classifiers seen in Figure 3.6 and Figure 3.7.are threshold-based. In order to determine a practical operating threshold, a training set is needed for each classifier (Biometric matchers and spoof detectors).

The training set is composed by randomly sampling the data set of subjects at 3 different rates: 25%, 50% and 75%. In order to avoid overfitting the training sets, a 10-fold cross validation is used. In each fold, the threshold that yields the minimum total error rates is determined. In some cases, two or more thresholds may have the same minimum value. To resolve such a tie, the threshold corresponding to the lowest FMR for biometric matchers is selected. Once the threshold is determined for every training fold, the average threshold of all 10 folds is used as the final threshold. The performance is then evaluated on all the test folds using this average threshold. The evaluation of Method C was carried out by implementing four different classifiers and choosing the one that resulted in the best performance. Classifiers were trained at different rates (25%, 50% and 75%) as well. The Neural Network (NN) presented the lowest FMR, compared to the Decision Tree (DT), the Naive Bayes (NB) and the K Nearest Neighbor (KNN). For Method C, we report results obtained by using the NN since it provides the highest accuracy. The NN method was then employed in Method D as well, as an estimator to compute the conditional probability obtained by the mathematical deviation expressed in Eqn. 3.5. The classifiers were implemented by using the Matlab Version 7.6.0.324 (R2008a) software.

The second part of experimental analysis is conducted on the LivDet 2011 [134] database. It consists of 1,000 live and 1,000 fake fingerprint samples in the training set, and the same number of samples from different subjects in the test set. The spoof artifacts in the LivDet

2011 database are fabricated using five materials, viz., gelatine, silicone, woodglue, ecoflex, and latex. For each material, 200 fingerprints were fabricated from 20 fingers using the consensual method (i.e., with the consent and collaboration of the user). Both live fingers and spoof artifacts were obtained using *four* different sensors, i.e., Biometrika, Italdata, Sagem and DigitalPersona (as shown in table 3.3). In this part of experiments, the proposed Bayesian Belief Networks, such as BBN-ML, BBN-MLQ and BBN-MLQc, are compared according to the following perspectives:

1. *Analysis of the match score distribution and baseline performance of the fingerprint verification system under zero-effort impostors and spoof attacks:* This experiment is designed to demonstrate the hazard due to spoof attacks by performing comparative analysis of the match score distribution and baseline performance of the system under spoof attacks with respect to impostor access. In particular, the match score distribution and the baseline performance of the fingerprint verification system is assessed for the following cases (a) LLG vs. LLI, (b) LLG vs. LSG, (c) LLG vs. SSG, (d) LLG vs. LSI and (e) LLG vs. SSI.

2. *A comparison of spoof detection performance of the fusion frameworks:* This experiment evaluate the spoof detection performance of the proposed BBN-MLQc and BBN-MLQ. Comparative assessment is made with the existing BBN-ML and GMM-based direct modelling scheme (DM-GMM) based on quality, liveness and match scores. The spoof detection accuracy of these frameworks is estimated by calculating the proportion of enrolled and input spoof images correctly classified into one of these classes i.e., LSG, SLG, SSG, LSI, SLI and SSI. Note that the spoof detection accuracy of these frameworks will not be equal to that of the baseline spoof detection algorithm used (LBP in this study), due to interaction of liveness scores with match score and quality values in rendering the final classification. The aim is to validate the assumption that appropriate modelling of relationship between quality and liveness score enhance the

spoof detection accuracy of the proposed models (BBN-MLQ and BBN-MLQc) over existing frameworks and the baseline spoof detection algorithm used.

3. *A comparison of performance of various frameworks against spoof attacks:* This experiment evaluate the performance of the proposed frameworks against spoof attacks. The trained frameworks are evaluated against the events LSG, SLG and SSG against LLG and the performance is reported. We considered only LSG, SLG and SSG against LLG in this case, adhering to the basic definition of the spoofing attacks.

4. *A comparison of overall performance of various frameworks:* As these fusion frameworks are developed to operate under both zero-effort impostors and spoof attacks. This experiment evaluate the overall performance of the proposed frameworks against all possible *eight* operations listed in Table 3.2, including also the cases when spoof artifact other than those of the claimed identity is used to access the system (LSI, SLI and SSI).

5. *Robustness of the BBN models across fabrication materials:* This experiment evaluate the robustness of the proposed BBN-MLQ and BBN-MLQc on new fabrication materials. To this aim, these models are tested using spoofs generated using fabrication materials not used during the training stage. The aim is to validate the assumption that appropriate modelling of relationship between quality and liveness score also enhance the interoperability of the proposed models (BBN-MLQ and BBN-MLQc) across fabrication materials over existing frameworks. This is due to accounting for the material specific characteristics (i.e, change in the quality) of the fake fingerprint images generated using different fabrication materials. Comparative assessment is made with BBN-ML model that does not consider quality values.

Table 3.3: Number of match scores, liveness scores and quality values corresponding to different states based on 5-fold cross validation. These scores are used for used for training and testing the fusion frameworks against spoof attacks.

Sensors	No. of Samples	Spoof Materials	Scores	
			Training	Test
Biometrika	Live: 200 fingers (5 live samples each) Spoof: 100 fingers (10 spoof samples each)	Ecoflex, Gelatine, Latex, Silgum and WoodGlue	No. of match scores: $\approx 1,600,000$	$\approx 400,000$
			No. of LLS: $\approx 1,600$	≈ 400
Italdata	Live: 200 fingers (5 live samples each) Spoof: 100 fingers (10 spoof samples each)		No. of LFS and FLS: $\approx 2,000$	≈ 500
			No. of FFS: $\approx 3,600$	≈ 900
Digital	Live: 200 fingers (5 live samples each) Spoof: 100 fingers (10 spoof samples each)	Gelatine, Latex, PlayDoh, Silicone, and WoodGlue	No. of LLD: $\approx 400,000$	$\approx 100,000$
			No. of LFD and SFD: $\approx 800,000$	$\approx 200,000$
Sagem	Total 1000 live samples and 1000 spoof samples		No. of FFD: $\approx 400,000$	$\approx 100,000$
			No. of liveness values and quality scores: 2,000	

3.6 Experimental Results on LivDet 2009 Database

In order to compare the performance of different fusion frameworks in a practical scenario, we report the error rates at specific operating points where all the four proposed methods have comparable false acceptance error rates. In the case of Method C, since the false acceptance obtained by the Neural Network was not comparable with the other three proposed methods, we also report the error rates of the Full Bayesian classifier which showed a comparable performance. The results are summarized in Table 3.4 and 3.5).

- Tables 3.4 and 3.5) indicate that the best verification performance is achieved by Method D. This outcome suggests that combining anti-spoofing information with match scores leads to a verification performance improvement compared to the case where the spoof scores are not used (see the error rates of stage 1 of Method A). For example, at a training rate of 25%, FMR is 0.11% for Method D and 0.18% when spoof scores are not used.

- In the presence of a reliable anti-spoofing measure (see Table 3.5 which corresponds to gelatin spoof), the best spoof detection performance is achieved by Method C, while when dealing with a less reliable anti-spoofing measure (see Table 3.4 which corresponds to silicone spoof) it is achieved by Method A.

- The best global performance is achieved by Method D. This result demonstrates that

Table 3.4: Comparison of all the methods from verification, spoof detection and global error perspective (silicone samples)

		Verification		**Spoof Detection**		**Global Error**	
Rates	Method	FAR	FRR	1 - SDR	1 - LDR	OFAR	1 - GAR
25%	A	0.1752	5.2609	1.6190	12.4870	0.5428	13.7391
	B	0.3228	5.4928	10.4675	12.4025	0.7014	14.8986
	C(NN)	0.3360	10.1014	3.9093	12.7438	0.2529	25.6522
	C(FB)	0.6245	5.0000	12.0664	12.4756	0.5052	13.3913
	D(NN)	**0.0020**	5.5275	5.6879	12.5440	0.2478	15.2899
	D(FB)	0.1086	5.0580	5.2720	12.5142	0.3262	14.7246
50%	A	0.2342	5.0217	1.4669	12.4918	0.5026	14.4348
	B	0.5474	5.0435	10.2243	12.3301	0.8619	14.3478
	C(NN)	0.3449	9.5000	3.7694	12.7703	**0.2169**	26.3478
	C(FB)	0.6374	4.9565	12.0126	12.4711	0.4797	13.1304
	D(NN)	**0.0020**	5.1870	5.2968	12.5576	0.2408	15.6957
	D(FB)	0.1030	5.0435	5.2439	12.5158	0.3077	14.2174
75%	A	0.1443	4.7826	0.6604	12.5168	0.3172	18.7826
	B	0.4822	4.7826	7.3144	12.3459	0.6899	17.0435
	C(NN)	0.3597	11.4348	**4.0000**	12.9944	**0.1918**	30.0782
	C(FB)	0.5929	**4.4783**	11.8765	12.4824	0.5351	14.4348
	D(NN)	**0.0020**	5.0652	5.3384	12.5610	0.2217	19.3043
	D(FB)	0.1119	**4.5652**	5.3510	12.5011	0.3130	18.0435

the configuration of the Bayesian Network is effective and the assumption that spoof scores influence match scores works well. Lowest global error rates are observed in the presence of a reliable anti-spoofing measurement (see Table 3.5).

According to the above observation, the design of Bayesian Belief Network (BBN) is proposed to further improve the overall security performance. As a graphical model based parallel scheme, the proposed BBN does not overwhelm other parallel classifiers (e.g., neural network and decision tree) from the spoof detection perspective. However, the overall matching accuracy of the BBN is consistently better than other classifiers. One possible reason is that the configuration of BBN assumes that the match score would not affect the spoof detection accuracy. Compared to the equivalence assumption of match scores and anti-spoofing measures which is used in other classifiers, the BBN assumes a one-way influence which is more practical and in accordance with a causal conception. This observation

Table 3.5: Comparison of all the methods from verification, spoof detection and global error perspective (gelatin samples)

		Verification		**Spoof Detection**		**Global Error**	
Rates	Method	FAR	FRR	1 - SDR	1 - CDR	OFAR	1 - GAR
25%	A	0.1913	15.7073	0.1943	12.4577	0.0453	4.4228
	B	0.2179	16.0081	0.5687	12.4577	0.0469	4.5854
	C(NN)	0.0392	26.8943	0.6475	12.5021	0.0744	13.1707
	C(FB)	0.2646	15.2846	0.0082	12.4369	0.0862	5.3089
	D(NN)	0.0112	16.6341	0.3331	12.4786	0.0056	5.7236
	D(FB)	0.1420	15.6016	0.4852	12.4338	0.0396	4.6179
50%	A	0.2506	15.9146	0.1844	12.4550	0.0512	4.4878
	B	0.1762	16.4634	0.6520	12.4595	0.0467	4.4878
	C(NN)	0.4021	26.9390	1.3361	12.7726	0.0512	19.3845
	C(FB)	0.2712	15.5366	**0.0020**	12.4507	0.0563	5.8537
	D(NN)	**0.0110**	16.4146	0.3200	12.4750	**0.0081**	4.6341
	D(FB)	0.0893	15.8659	0.3161	12.4340	0.0459	4.1951
75%	A	0.1884	15.6341	0.1137	12.4614	0.0419	4.4878
	B	0.2232	15.5366	0.5719	12.4614	0.0419	4.4878
	C(NN)	0.4018	28.1707	1.6498	12.7726	0.0391	21.4533
	C(FB)	0.2841	15.7561	**0.0003**	12.4486	0.0616	6.0488
	D(NN)	**0.0098**	16.9512	0.3017	12.4744	**0.0054**	**4.2927**
	D(FB)	0.1171	15.1463	0.3062	12.4322	0.0289	**4.2927**

inspires us to implementing the similar causal assumptions in the future work on combining ancillary factors which do not have an evident relationship with biometric match scores.

3.7 Experimental Results on LivDet 2011 Database

3.7.1 EXP1. Baseline

Figures 3.12 show the example match score distributions of the LSG, SSG, LSI and LLI against LLG, from live and fake fingerprint images acquired using Biometrika, from the LivDet 2011 database. The high overlap in the match score distributions corresponding to LLG vs. LSG suggest that fingerprints can be effectively spoofed to gain illegitimate access to the system. Further, the match score distributions corresponding to the case SSG is quite similar to that of LLG. Furthermore, the match score distribution of LSI is similar to that of LLI. The same trend is observed for Italdata, Sagem and DigitalPersona sensors.

Figure3.13 shows the ROC Curves of the baseline performance of the fingerprint verification system under zero-effort impostors and spoof attacks. It can be seen that:

- The EER of the baseline systems under zero-effort impostors (i.e., LLG vs. LLI) are in the range [2.2%, 5.1%] for the Biometrika, Italdata, Sagem and DigitalPersona sensors, respectively.
- The EER of the fingerprint verification system under spoof attacks, for the cases when the spoof artifact is the replica of the original fingerprint image of the claimed identity (i.e., LLG vs. LSG (SSG)) are in the range [29.4%, 54.1%] for the Biometrika, Italdata, Sagem and DigitalPersona sensors, respectively. Thus, demonstrating the hazard of the spoof attacks to the biometric system security.
- The case LLG vs. SSI obtain higher error rate than LLG vs. LSI, this is due to variation in the quality of spoof samples. Consequently, leading to high error rate when a pair of poor quality spoof images, belonging to different identities (SSI), are matched. This experiment emphasize on the urgent need of enhancing the security of the fingerprint verification system against spoof attacks.

3.7.2 EXP2. Performance Under Zero-Effort Impostors

Figure 3.14 show the ROC curves for the spoof detection accuracy of the BBN-MLQc, BBN-MLQ, BBN-ML, BBN-MQ and GMM-based direct modelling scheme (DM-GMM). Comparative assessment is made with BBN-M based only on match scores. It can be seen that the proposed BBN-MLQ and BBN-MLQc obtained better spoof detection performance in comparison to the existing frameworks and the baseline LBP-based anti-spoofing algorithm. This is due to appropriate modelling of quality with the liveness score. Further, BBN-MLQc always outperformed BBN-MLQ. The BBN-M and BBN-MQ do not incorporate the liveness score, hence, they are not evaluated in this experiment.

It can be seen that (Figure 3.14)

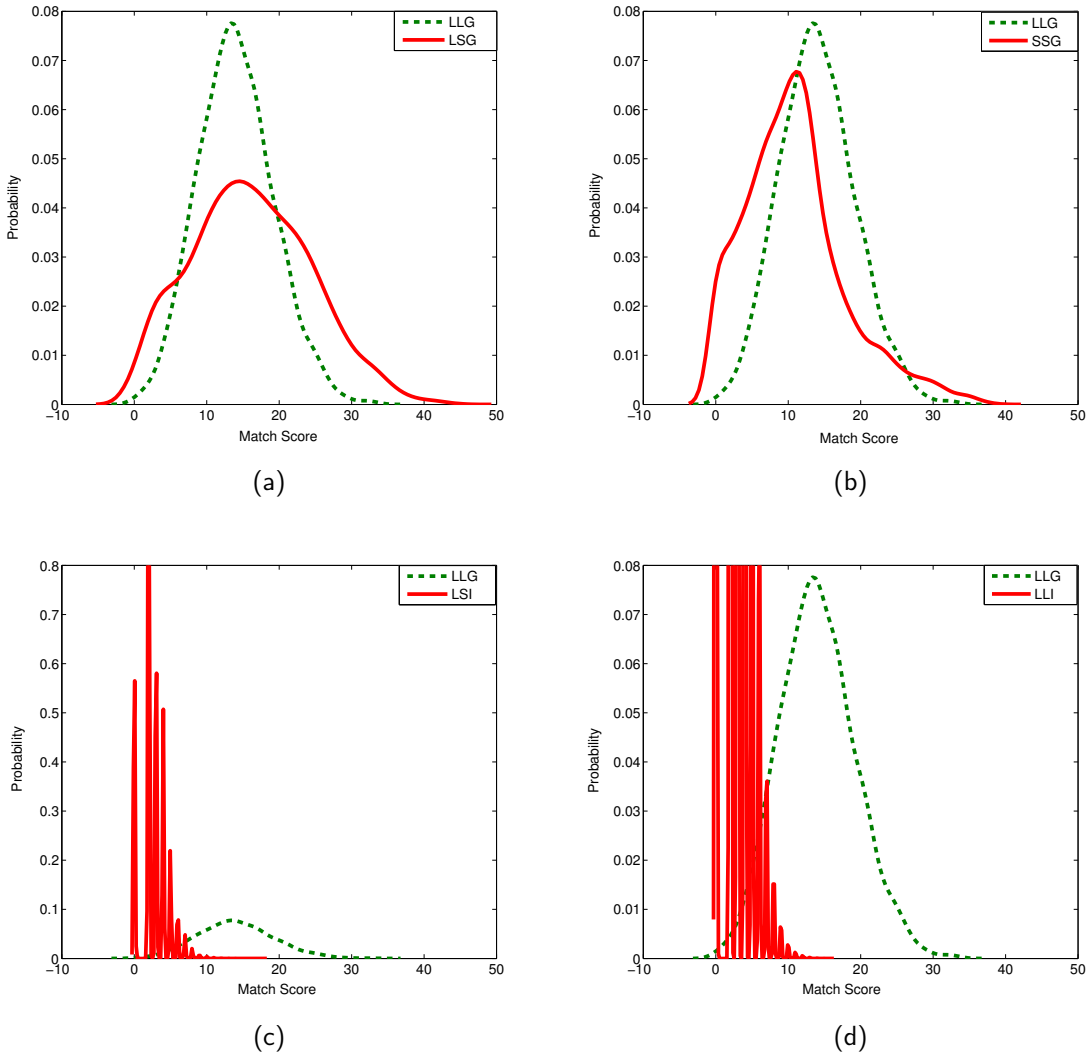


Figure 3.12: The match score distributions of (a) LSG vs. LLG, (b) SSG vs. LLG, (c) LSI vs. LLG and (d) LLI vs. LLG on samples acquired using **Biometrika** sensor in the LivDet 2011 database.

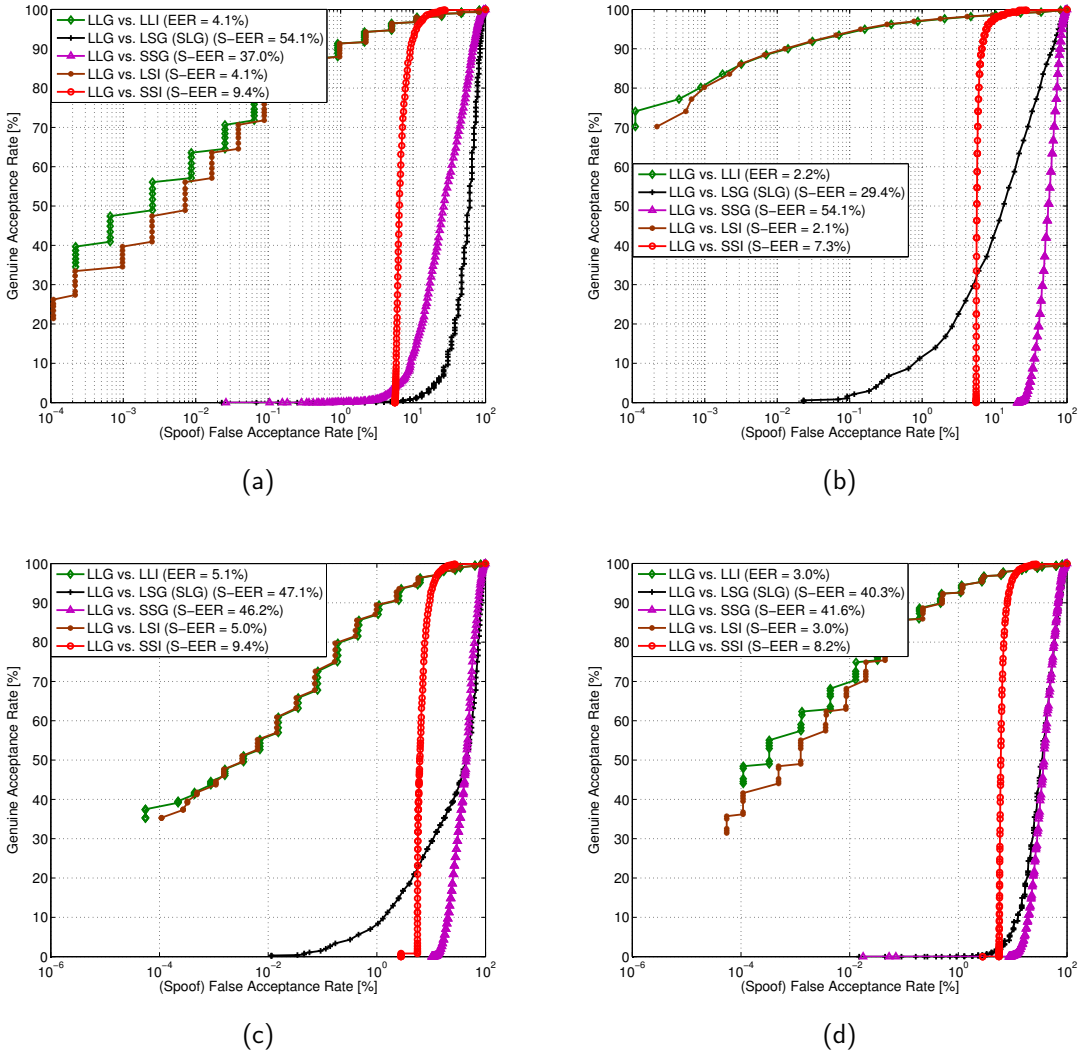


Figure 3.13: ROC Curves of the baseline performance of the fingerprint verification system under zero-effort impostors and spoof attacks.

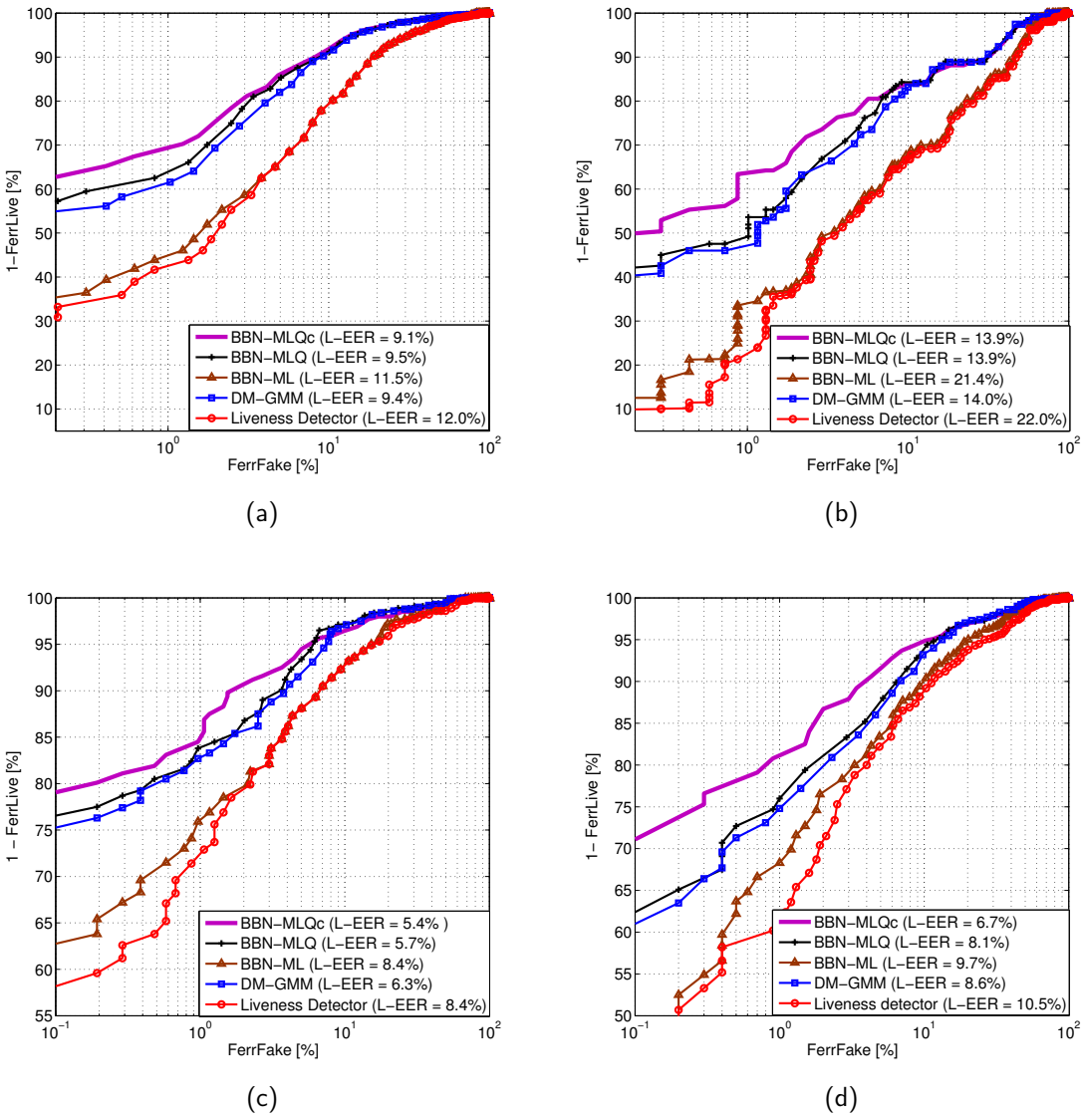


Figure 3.14: Spoof detection performance of the various BBN frameworks on the LivDet 2011 database. Note that the spoof detection accuracy of these frameworks is *not* the same as that of the LBP-based spoof detection algorithm used. This is because the interaction of liveness scores with match score and quality is taken into account when rendering the final decision. The results are from different sensors as: (a) Biometric, (b) Italdata, (c) Sagem and (d) Digital.

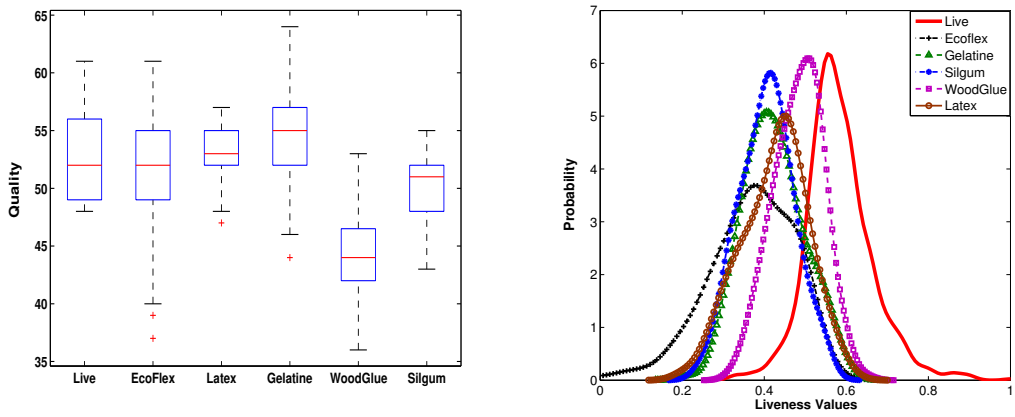
- The EER of the BBN-MLQc reduced by 24.2%, 36.8%, 35.7% and 36.2% (range [24.2%,36.8%]) over the baseline LBP-based spoof detection scheme for the Biometrika, Italdata, Sagem and DigitalPersona sensors, respectively. For instance, EER of the BBN-MLQc is 13.9% whereas the EER of the spoof detection is 22.0% for the Italdata sensor. The spoof detection rate (SDR) increased from 56% to 79% using BBN-MLQc over LBP-based spoof detection at a fixed 99.9% live detection rate (LDR).

- Further, EER of the BBN-MLQ reduced by 20.8%, 36.8%, 32.1% and 22.8% (range [20.8%,36.8%]) over the baseline LBP-based spoof detection scheme for all the four sensors, respectively.

- DM-GMM always outperformed BBN-ML because BBN-ML, this is due to consideration of quality values in DM-GMM. The spoof detection accuracy of the BBN-MLQ is slightly better than DM-GMM.

Figure 3.15 shows the variation in the quality and variation in the distribution of the liveness score as a function of the five fabrication materials used to generate spoofs in LivDet 2011 (Biometrika sensor). This experiment show the efficacy of normalizing the liveness score based on the quality of the fake fingerprint samples in the proposed models. Thus, reducing the impact of variation in the quality of the fake fabrication materials on the performance of the spoof detection.

Further, Figure 3.16 show the liveness score before and after adaptation using the transformation function of BBN-MLQc (as in Eqn. (5.1). Specifically, liveness values of the samples corresponding to normalized quality range [0.3,0.5] are transformed. As a consequence, liveness values of the live samples are shifted towards one and those of spoof samples are shifted towards zero, leading to better spoof detection capability of BBN-MLQc over other frameworks.



(a) (b)

Figure 3.15: Boxplot of quality values and probability distribution of the liveness score for five different materials in **Biometrika** in the LivDet 2011 database. The same observation is made for Italdata, Sagem and Digital sensors as well.

3.7.3 EXP3. Spoof Detection Accuracy

This experiment evaluates the performance of the proposed frameworks against spoof attacks. As discussed earlier, *Live Detection Rate (LDR)* indicates the percentage of correctly detected live samples, while *Spoof Detection Rate (SDR)* indicates the percentage of correctly detected spoof samples. EER of the spoof detection (remained as S-EER) indicates the rate at which LDR is equal to SDR.

Table 5.1 shows the spoof detection performance of the BBN-MLQc, BBN-MLQ, BBN-ML, BBN-MQ and GMM-based direct modelling scheme (DM-GMM) on four sensors.

It can be seen that the proposed BBN-MLQ and BBN-MLQc obtained better spoof detection performance in comparison to the existing frameworks and the baseline LBP-based spoof detector. This is due to appropriate modeling of quality with the liveness scores.

- When spoof detection error (1 - SDR) was fixed at 1%, the live detection rate (LDR) of the BBN-MLQc is increased by 28.0%, 41.0%, 19.3% and 13.5% over the baseline LBP-based spoof detection scheme for the Biometrika, Italdata, DigitalPersona and Sagem sensors, respectively. It demonstrates the advantage of incorporating the quality of

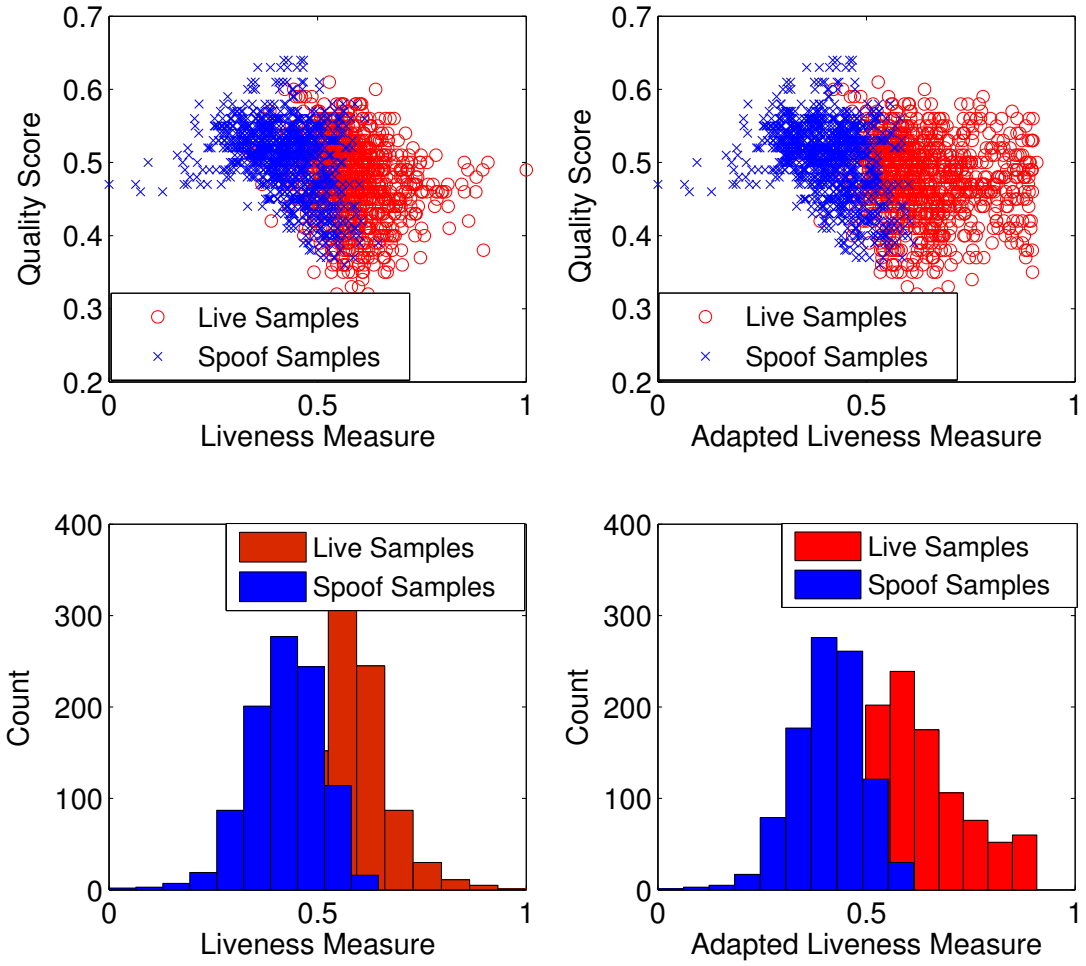


Figure 3.16: Scatter plot and histogram of the liveness scores, before and after adaptation using the transformation function used in BBN-MLQc. It can be noticed that liveness values of the live samples are shifted towards one and those of spoof samples are shifted towards zero, leading to better spoof detection capability of BBN-MLQc over other frameworks.

images for spoof detection.

- The spoof detection error (1 - SDR) of the BBN-MLQc is significantly lower than the DM-GMM direct modelling scheme. However, the BBN-MLQ is only slightly better than DM-GMM. It indicates that the benefits of the graphical modelling algorithm depends on the *configuration* of networks.

- Furthermore, when the spoof detection error (1 - SDR) was fixed at 1%, the live detection rate (LDR) of the BBN-MLQc model increased by 24.8%, 19.8%, 14.1% and 9.0% over the existing BBN-ML model, although both of them share the same causal assumptions. It demonstrates the benefits of utilizing the proposed quality-based calibration scheme.

The ROC curves of spoof detection performance on data from the Biometrika and Italdata sensors are shown in Figure 3.14. The above observations are consistent across all the four sensors.

3.7.4 EXP4. Overall Recognition Accuracy

This experiment evaluates the overall performance of the proposed BBN-MLQc and BBN-MLQ frameworks under all possible spoof attack scenarios. Comparative assessment is made with the existing Bayesian Networks and GMM-based direct modelling scheme (DM-GMM).

As the fingerprint verification system operates under both zero-effort impostor and spoof attacks, the overall performance rates can be defined as follows:

- **G**enuine Acceptance Rate (GAR): Proportion of the **LLG** class that are incorrectly classified as genuine and accepted by the system.

- **O**verall False Acceptance Rate (OFAR): Proportion of zero-effort impostor and spoof samples that are incorrectly classified as the **LLG** class.

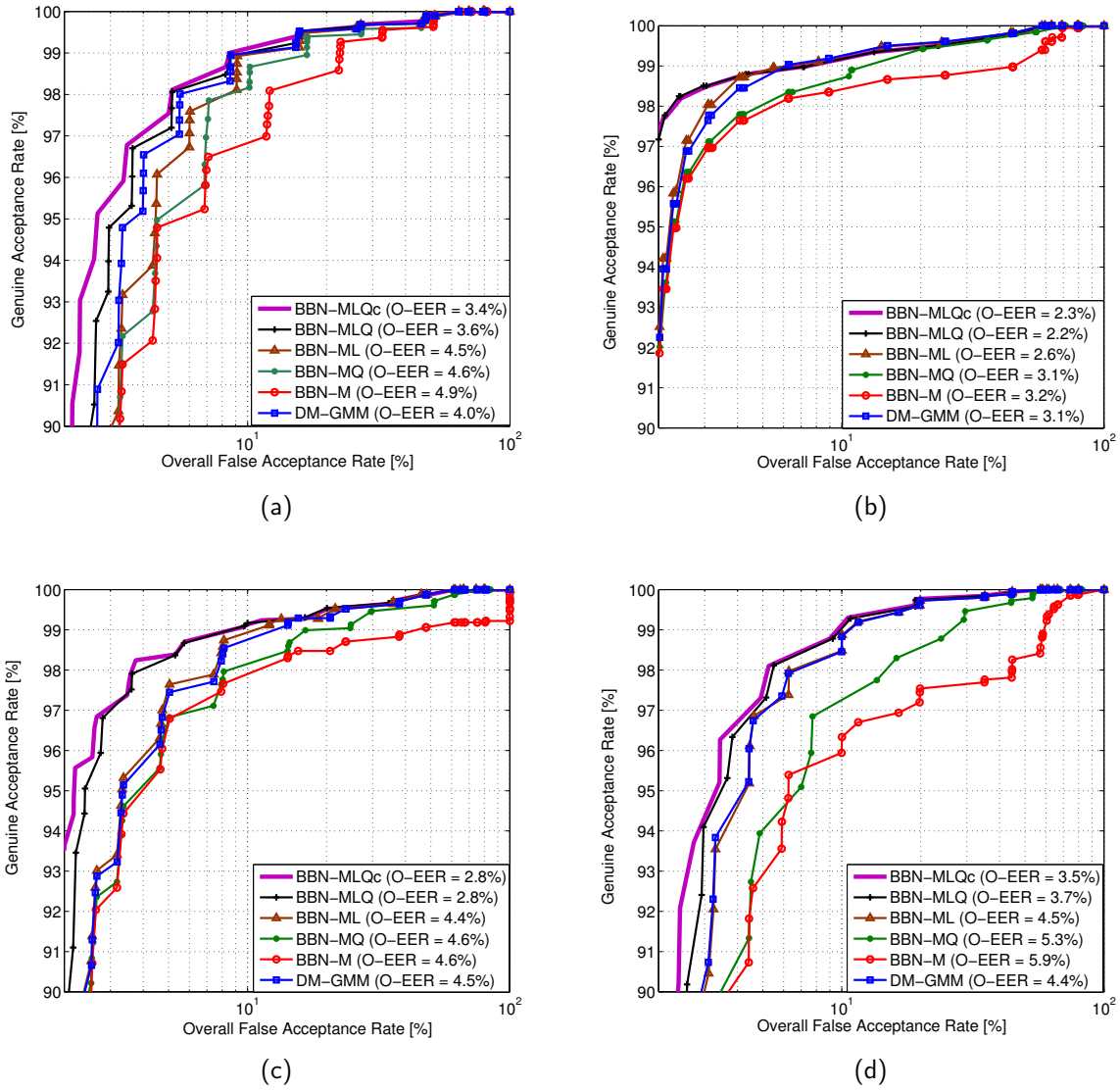


Figure 3.17: Performance of the various frameworks when all eight events are considered for four sensors as (a) Biometrika, (b) Italdata, (c) Digital and (d) Sagem. It can be seen that BBN-MLQc outperforms all other frameworks.

- **O**verall Equal Error Rate (O-EER): The rate at which OFAR equals 1 minus the Genuine Acceptance Rate (GAR). The **O-EER** of each fusion scheme is shown in the ROC curves.

Table 5.2 demonstrates that BBN-MLQc performs much better than all the existing frameworks and the baseline BBN-M. This is due to its high spoof detection capability and better performance under spoof attacks (see Experiment 1). The ROC curves of each fusion

scheme are shown in Figure 3.17.

- At a fixed 1% OFAR, the GAR of the BBN-MLQc increased by 17.0%, 5.77%, 9.49% and 6.66% (range [5.77%, 17.0%]) over the BBN-M for the Biometrika, Italdata, Sagem and DigitalPersona sensors, respectively. For instance, the GAR of the BBN-MLQc is 95.5% whereas GAR of the BBN-M is 81.7% at a 1% OFAR for the Biometrika sensor.

- At a fixed 1% OFAR, the GAR of the BBN-MLQ increased by 16.5%, 5.13%, 9.03% and 6.02% (range [16.5%,5.13%]) over the BBN-M for all the four sensors, respectively. The GARs of the BBN-ML increased in the range [13.7%,5.17%], and are similar to the GARs of the DM-GMM that increased in the range [13.7%,5.14%]). Further, BBN-MQ performed just a little better than BBN-M by 1.47%, 0.22%, 0.14% and 0.13% (range [1.47%,0.13%]), respectively.

3.7.5 EXP5. Performance Across Fabrication Materials

Figure 3.18 show the performance of the BBN-MLQc and BBN-MLQ in comparison to BBN-ML across fabrication materials for the Biometrika sensor. These models are trained using a single kind of material say Latex and tested on the rest four materials say Gelatin, Latex, EcoFlex, Silgum and WoodGlue. It can be seen that performance of all the frameworks dropped significantly across materials. This is due to different characteristics of different fabrication materials. However, the proposed BBN-MLQc and BBN-MLQ always outperformed BBN-ML.

It can be seen that (Figure 3.18)

- The EER of the BBN-MLQc reduced by 41.4%, 22.6%, 71.9%, 39.4% and 41.4% over BBN-ML when trained using latex, gelatine, ecoflex, silgum and woodglue, respectively, for the Biometrika sensor.

- Further, The EER of the BBN-MLQ reduced by 34.3%, 5.7%, 17.5%, 2.6% and 33.9% over BBN-ML when trained using latex, gelatine, ecoflex, silgum and woodglue, re-

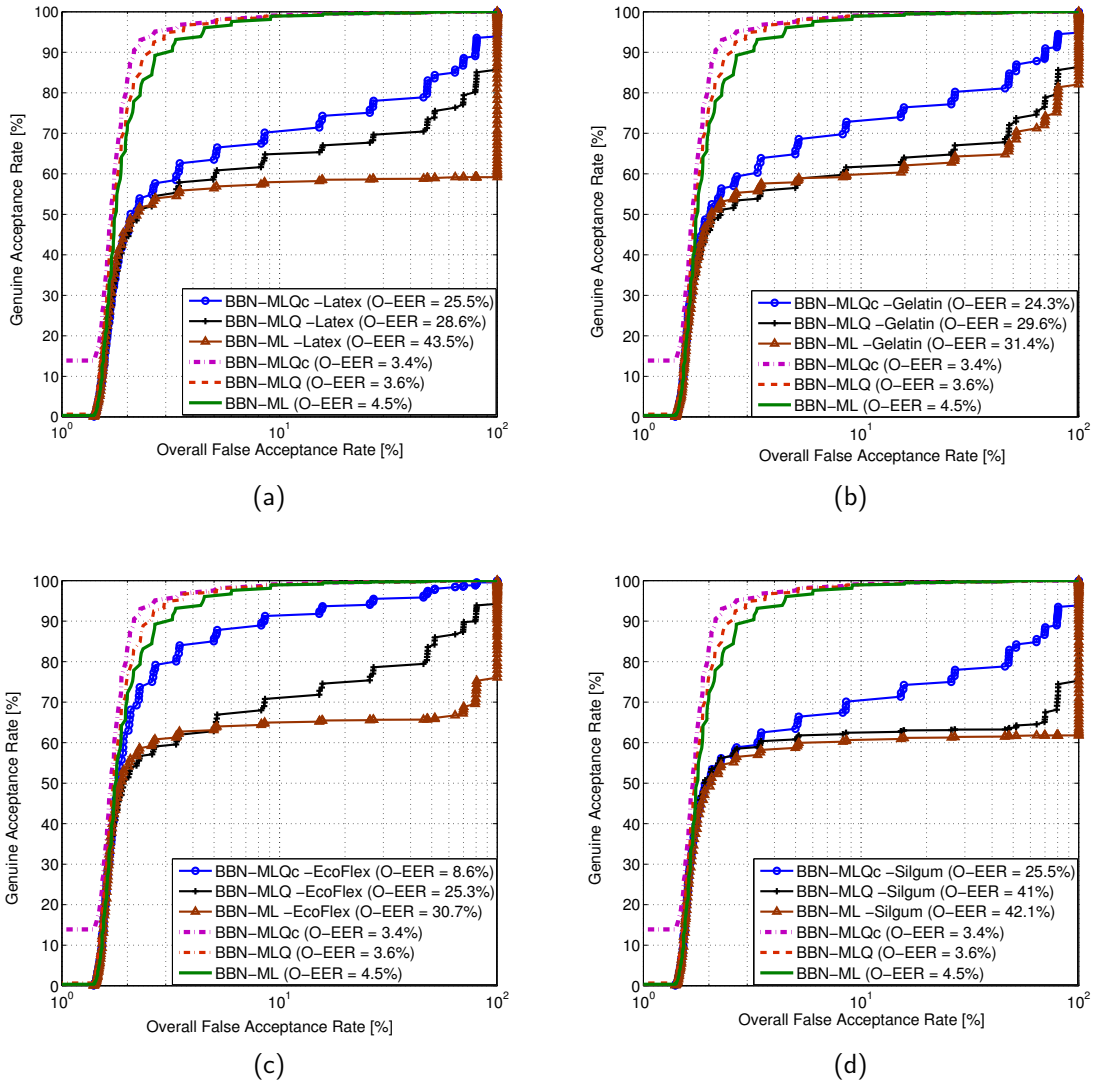


Figure 3.18: Evaluation of the BBN-MLQ and BBN-MLQc across fabrication materials trained on only (a) Latex, (b) Gelatin, (c) EcoFlex and (d) Silgum tested on rest other four materials for the **Biometrika** sensor.

spectively. However, the drop is the performance is significant in this experiment. For instance, EER of the BBN-MLQc increased from 3.4% to 25.5% when trained using all the available materials over the one trained using only latex. This is because of the worst case assumption of training using single kind of material.

The same observation is made for different combination of one, two and three training materials for Italdata, Sagem and Digital sensor.

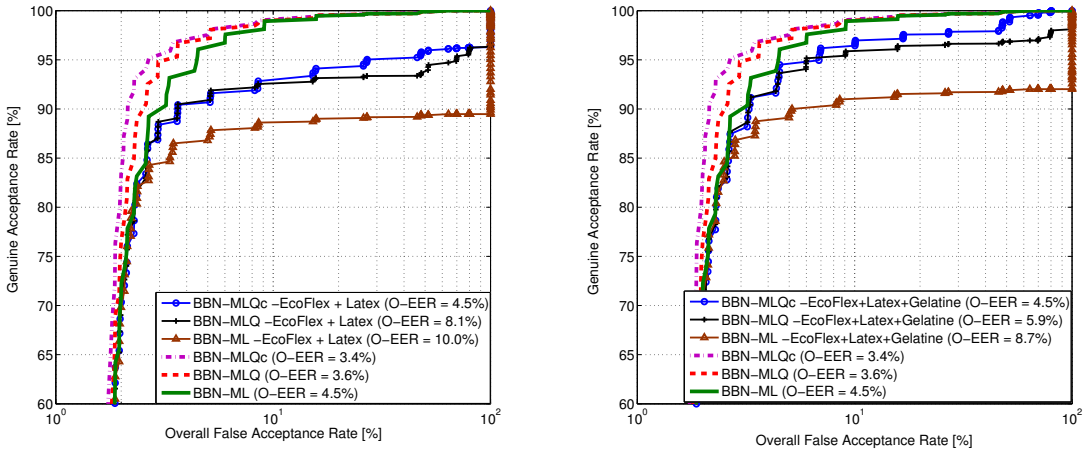


Figure 3.19: Evaluation of the BBN-MLQ and BBN-MLQc across fabrication materials trained using combination of (a) EcoFlex+ Latex and (b) EcoFlex+ Latex+ Gelatin and tested on rest other three and two materials, respectively, for the **Biometrika** sensor as an example.

3.7.6 EXP6. BBN-Based Validation

The assumptions in the existing and proposed BBN models are statistically validated using Structural equation modeling (SEM). Structural equation modeling (SEM) is a causal modeling approach that combine cause−effect information with statistical data to provide a quantitative assessment of relationships among the studied variables. If the relationships are significant, the theoretical construction is considered valid and can be used to provide guidelines for the application of the model in practice. Druzdzel and Simon[3]examined the conditions under which one can reasonably interpret the structure of a Bayesian network as a causal graph of the system. The authors also proposed a method, referred to as causal ordering, to link BBN to structural equation modeling (SEM) in order to test the causal relationships between variables.

Basically, the proposed causal ordering method is a mechanical procedure that transform the dependency structure of an acyclicity causal graph (such as BBN) into a set of

[3]Marek Druzdzel and Herbert Simon, "Causality in Bayesian Belief Networks", *The Ninth Annual Conference on Uncertainty in Artificial Intelligence*, 1993, page 3–11

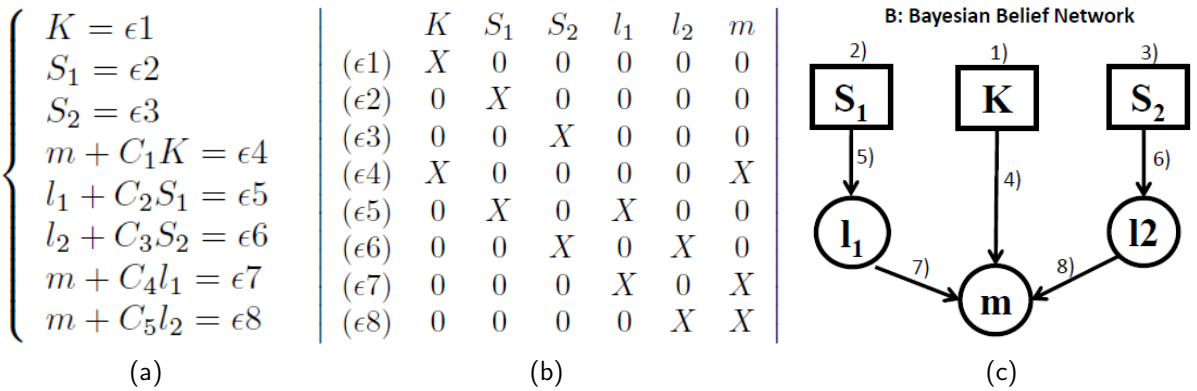


Figure 3.20: Structural equations associated with the BBN-ML model as an example.

simultaneous equations. After the set of equations are obtained for a particular BBN, it is straightforward to build an equation model and test the goodness-of-fit using existing SEM software or packages. Next, the procedure of causal ordering and equation extraction is briefly described as follows:

1. Let `B` be a BBN model. The acyclicity assumption in a causal structure of `B` ensure that there exists an equation model `S`, which involves all variables in `B`, and the joint probability functions from `B` and `S` are equivalent with respect to all variables.

2. For a structural equation model, a mechanism (M) can be described as:

$$F_M(x_1, x_2, \ldots, x_n, \epsilon) = 0$$

   The presence of a variable x_i means that the system's element that is denoted by x_i directly participates in the mechanism M. The structural relationship of a equation model `S` with n simultaneous structural equations $\epsilon_1, \epsilon_2, \ldots \epsilon_n$ can be denoted by a matrix with X and zero entries. As an example, Figure 3.20 shows the structural equations and matrix (with an entry (marked with X corresponding to each variable and the relationship between variables) associated with BBN-ML.

3. The causal ordering theorem also states that, the structural model `S` reflects the causal structure of a Bayesian Belief Network `B` if and only if (1) each node of `B` and all its

direct predecessors describe variables involved in a separate mechanism in the system and (2) each node with no predecessors represent an exogenous variable.

In this work, set of structural equations are retrieved for each BBN model the above mentioned approach, and then input to a R package called "SEM" to obtain the goodness-of-fit. The following is an example of the obtained outputs on testing the **BBN-ML** model using R package called "SEM".

```
Model Chisquare =  117.30
Df =  2 Pr(>Chisq) = 0.9811e-33
Chisquare (null model) =  323.03   Df =  12
Goodness-of-fit index =  0.9521
```

The goodness-of-fit value is calculated using chi-square test, which assumes that the ratio between the variance from the proposed model with observations, and the variance from the theoretical saturated model follows a chi-square distribution with a certain degree of freedom. Recall that BBN with fewer assumptions and simpler configurations have higher goodness-of-fit. Note that even if the goodness-of-fit value is high, it can only concluded that the model fits the training data and cannot be used to predict the performance of the trained BBN.

Table 3.6: Spoof detection performance of the various BBN frameworks on the LivDet 2011 database.

Various Frameworks	Biometrika		Italdata		Digital		Sagem	
	CDR at 1% 1- SDR	CDR at 10% 1- SDR	CDR at 1% 1- SDR	CDR at 10% 1- SDR	CDR at 1% 1- SDR	CDR at 10% 1- SDR	CDR at 1% 1- SDR	CDR at 10% 1- SDR
BBN-MLQc	**70.1**	**91.1**	52.6	**84.8**	**81.2**	**95.8**	**85.6**	**97.2**
BBN-MLQ	62.3	**91.1**	49.8	83.2	77.1	**95.8**	84.1	**97.2**
BBN-ML	45.3	80.3	34.1	67.8	67.1	91.4	76.6	92.5
DM-GMM	61.7	91.0	46.0	82.1	75.3	93.3	83.0	95.6
Spoof Detector	42.0	80.0	22.9	66.9	61.9	88.0	72.1	92.5

Table 3.7: Performance of the various frameworks when all eight events are considered for the Biometrika and Italdata sensors. BBN-MLQc is seen to outperform all other frameworks.

Various Frameworks	Biometrika		Italdata		Digital		Sagem	
	GAR [%] at OFAR = 1%	GAR [%] at OFAR = 5%	GAR [%] at OFAR = 1%	GAR [%] at OFAR = 5%	GAR [%] at OFAR = 1%	GAR [%] at OFAR = 5%	GAR [%] at OFAR = 1%	GAR [%] at OFAR = 5%
BBN-MLQc	**80.5**	**88.3**	**72.5**	**89.0**	**84.8**	**88.6**	**75.3**	**88.3**
BBN-MLQ	75.6	85.2	72.1	88.6	83.0	87.6	72.3	87.2
BBN-ML	74.2	86.7	**72.5**	88.7	83.0	87.7	73.3	87.4
BBN-MQ	77.9	85.5	72.0	88.2	77.5	87.1	68.1	84.4
BBN-M	76.7	85.1	71.8	88.1	77.5	86.9	68.1	83.5
DM-GMM	79.8	87.1	**72.5**	**89.0**	82.9	87.5	72.5	87.4

3.8 Summary and Future Work

While the primary purpose of a biometric recognition system is to ensure a reliable and accurate user authentication, the security of the recognition system itself can be jeopardized by spoof attacks. Anti-spoofing approaches, although is still under developing, are designed for incorporation with biometric systems to increase the security with an inherent demand. In this chapter, we firstly investigate two fundamental combining scheme, sequential and parallel schemes, for combining anti-spoofing measures with biometric match scores. The experimental results on two public-domain LivDet databases (2009 and 2011) show that the parallel scheme overwhelms the sequential scheme from both the spoof detection perspective and overall security perspective. It also evident the potential that anti-spoofing measures can improve the human recognition performance by combining with biometric traits.

It is notable that this work also presents a novel viewpoint on the attacking scenario by considering compromising templates. The distorted distributions of match scores clearly demonstrate the risk if the enrolled templates are spoofed. Additionally, we point out that from a security perspective, the ability of locating the potentially aimed templates of a spoof attack is essential if the anti-spoofing algorithms are still under developing and erroneous.

According to the above observation, the design of Bayesian Belief Network (BBN) is proposed to further improve the overall security performance. As a graphical model based parallel scheme, the proposed BBN does not overwhelm other parallel classifiers (e.g., neural network and decision tree) from the spoof detection perspective. However, the overall matching accuracy of the BBN is consistently better than other classifiers. One possible reason is that the configuration of BBN assumes that the match score would not affect the spoof detection accuracy. Compared to the equivalence assumption of match scores and anti-spoofing measures which is used in other classifiers, the BBN assumes a one-way influence which is more practical and in accordance with a causal conception. This observation inspires us to implementing the similar causal assumptions in the future work on combining ancillary factors which do not have an evident relationship with biometric match scores.

Besides, we proposed two Bayesian Belief Network (BBN) models that can effectively integrate liveness scores with quality scores and match score. The proposed BBN models have two different configurations distinguished on the basis of how the quality scores are incorporated. This study also compares the proposed BBN models with existing fusion frameworks against spoof attacks. Comprehensive experiments are conducted on the LivDet 2011 dataset. Results indicate that the proposed BBN-MLQ and BBN-MLQc methods consistently outperform existing fusion frameworks. Based on the experiments, the following conclusions can be drawn:

- **Causal relationship:** Fusion frameworks that model the appropriate relationship between the considered variables, such as the influence of the quality on liveness scores, obtain better performance.

- **Benefits of quality:** Incorporating image quality is beneficial in the fusion framework (BBN-MLQ and BBN-MLQc). This is because quality scores can take into account the material-specific characteristics of spoof fabrication materials. Further, the models incorporating quality also have benefits (better performance) when evaluated on novel spoof fabrication materials [102].

- **The role of quality:** These quality scores can be incorporated as features (as in BBN-MLQ) or used as a normalization parameter (as in BBN-MLQc). Experimental results suggest the efficacy of quality when used as a normalization parameter rather than a feature, since the latter makes the Bayesian Belief Network more complicated to be interpreted and calculated.

- **The role of latent variables**: The consistently better performance of BBN-MLQc over existing frameworks show the efficacy of quality-based clusters in adapting the liveness scores and match scores against the sample quality. Even for a single acquisition device, clusters of quality values (Q_{ci}) can be obtained corresponding to image

resolution, ridge and valley clarity, noise level, and spoof fabrication materials. Further, quality and liveness scores are also influence by the acquisition sensor used. As a part of future work, these latent variables i.e., quality clusters and sensor information will be incorporated in the BBN models. Further, the role of these latent variables will be analyzed for novel sensors and fabrication materials.

- **Semi-supervised learning in BBN models**: Our experimental results suggest that performance of all the BBN models drops across materials. Hence, automatically adapting these BBN models to novel spoof materials is another research avenue. In other words, models will be incorporated with the learning ability to automatically detect and adapt themselves to spoof samples generated using novel materials.

- **Effect of the baseline anti-spoofing algorithms:** Continuous efforts are being directed towards developing spoof detection schemes which offer lower error rates, also evident by three spoof detection competitions (LivDet) conducted between 2009 and 2011. The performance of existing and proposed fusion frameworks will be evaluated on incorporating liveness scores obtained using novel spoof detection schemes and comparative assessment will be drawn with respect to existing ones.

- Empirical analysis would be implemented to validate the causal assumptions made by the proposed BBN model, especially when the framework involves match scores from multiple biometric traits and the causal relationship become more implicit. Further more, when other kinds of ancillary information is involved with the proposed BBN model, it is essential to extend the configuration of BBN in a reasonable way.

CHAPTER 4

COMBINING ONE-CLASS SVMS FOR ANTI-SPOOFING

4.1 Background

As mentioned in the previous chapter, a biometric system is vulnerable to spoof attacks, where a fake fingerprint is used to circumvent the system. In order to detect or deflect fingerprint spoof attacks, a number of sensor-based and image-based anti-spoofing solutions have been proposed [70, 79]. Image-based solutions, in particular, have received plenty of attention in the literature since they do not require the use of additional hardware and are based only on the images that are subsequently used by the fingerprint matcher. Such algorithms typically extract texture-based features [84, 36], anatomical features [28, 72] or physiological features [34, 67] from a fingerprint image (or sequence of images), and then train a binary classifier (such as a Support Vector Machine) that distinguishes the features of "Live" and "Spoof" samples.

However, there are some concerns associated with the use of binary classifiers in the context of spoof detection. In practice, it is easy to obtain training samples pertaining to the "Live" class but difficult to obtain samples for the "Spoof" class, thereby leading to imbalanced training sets where the latter has substantially fewer training samples. Further, the training set for the spoof class may not have data corresponding to all possible types of fabrication materials. This makes it difficult for the classifier to reliably learn the concept of a spoof. In fact, it has been shown that spoof detection accuracy degrades sharply, when the test set has fake samples fabricated using materials that were previously "unseen" in the training set (as reported in [135, 37]). As spoof attacks evolve, it is likely that new and more sophisticated materials will be used to create fake fingerprints thereby undermining existing learning-based spoof detectors.

To generalize the effectiveness of spoof detectors across fabrication materials - even those

that are not encountered during training - recent work has formulated spoof detection as an open-set problem [101, 102]. Others utilize quality-based measures to minimize the impact of fabrication materials [33, 23]. While such methods have demonstrated success, they still require a large number of training samples from the spoof class. Menotti et al. [75] proposed a convolutional neural network (CNN) whose performance exceeded that of many fingerprint spoof detection benchmarks. However, just like other CNN-based methods, it requires a large number of training samples. Further, its robustness across fabrication materials was not evaluated.

The aforementioned concerns (related to interoperability across fabrication materials and limited spoof training samples) motivated us to consider approaching spoof detection as a *one-class* problem. The one-class classification paradigm differs from the multi-class paradigm in that only data from a single class (e.g., the live class) is used for training the classifier [112]. The task in one-class classification is to derive a decision boundary around samples of the live class that accepts as many samples as possible from that class while excluding other samples. Take the one-class support vector machine classifier as an example. The idea is to minimize the volume of the decision hypersphere containing the training data from a single class. However, this makes the problem harder than two-class classification because it is difficult to determine the tightness of the hypersphere enclosing the training samples. Moreover, it is difficult to determine what type of features extracted from a sample would effectively model the samples from the “Live" class.

In this chapter, we present a unified view of the general one-class classification approaches based on i) the features/descriptors used, ii) the availability of training data, and iii) the classifiers used in the context of fingerprint spoof detection. Besides, an ensemble of multiple one-class SVM (OC-SVM) classifiers is proposed, where each OC-SVM uses a different feature set to find the smallest possible hypersphere around the majority of training samples pertaining to the “Live" class. Then, a Least Square Estimation (LSE) based weighting algorithm is proposed to aggregate those independently trained OC-SVMs by assigning the

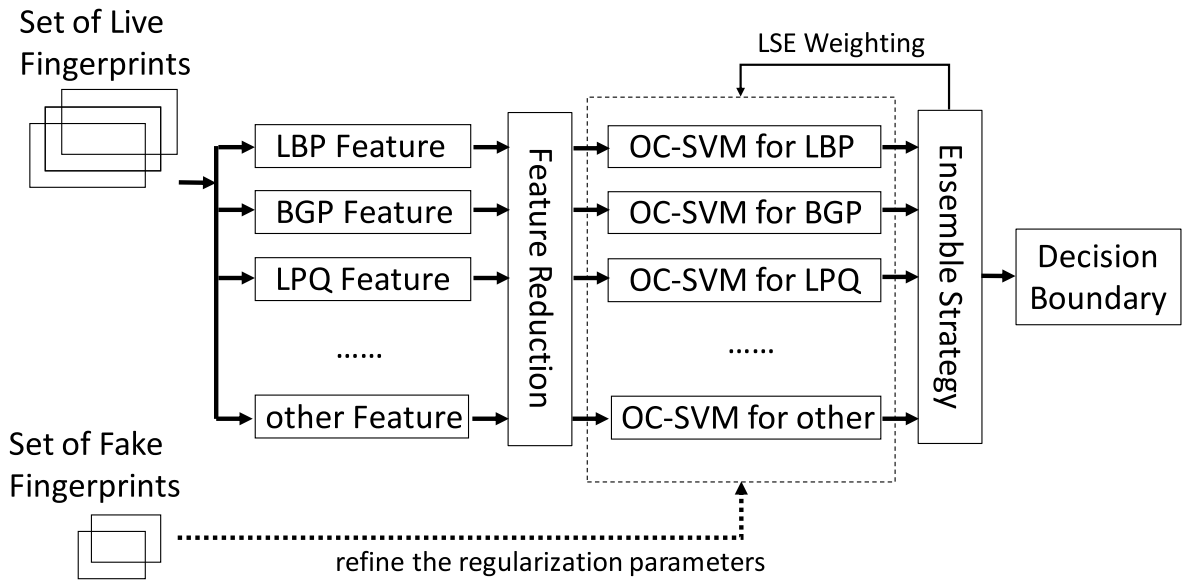


Figure 4.1: Schematic of the proposed ensemble framework that uses multiple OC-SVMs. Each OC-SVM utilizes a different set of features. While spoof fingerprints are not necessary for training the OC-SVMs, they are used to refine the decision boundary in the validation phase.

higher weight to the one with higher accuracy. Furthermore, the boundary of the hypersphere is further refined using a small number of spoof samples (as shown in Figure 4.1).

The experimental results show three significant advantages of the proposed ensemble of OC-SVMs:

1. The detection accuracy is comparable with state-of-art spoof detection algorithms; however, only a smaller number of spoof samples is required for training.

2. The spoof detection accuracy remains consistent, regardless of what fabrication material is used to forge spoofs and what fingerprint sensor is used to collect them.

3. The detection accuracy can be further improved by increasing the number of spoof fingerprint samples utilized as training samples, without suffering from the imbalanced class problem encountered by conventional binary SVM (B-SVM) classifiers.

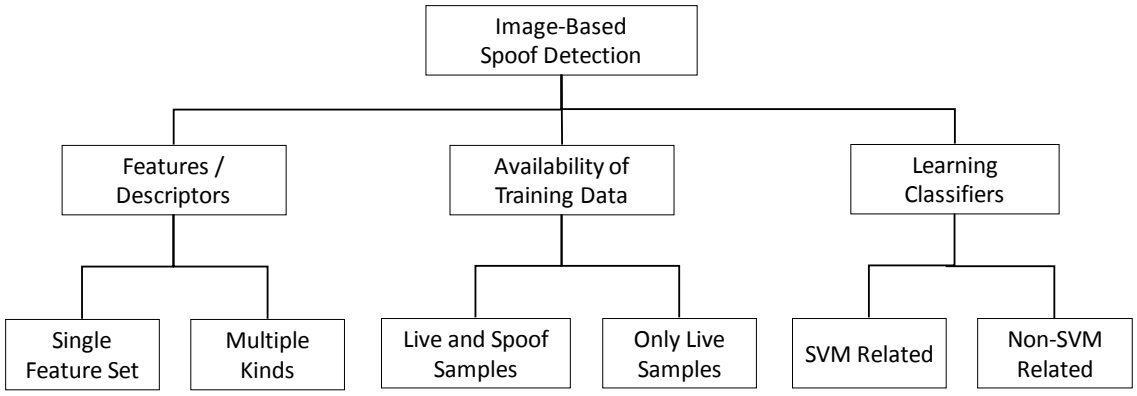


Figure 4.2: Proposed categorization for the study of image-based fingerprint spoof detection algorithms. The proposed ensemble OC-SVM classifier falls into the category of SVM-related classifiers that use multiple kinds of features extracted from only the live samples for training.

This chapter is organized as follows: Section 4.2 categorizes the current state of the art research on image-based fingerprint spoof detection from a new perspective. Section 4.3 focuses on one particular category under the proposed categorization, the one-class classification based SVM classifiers using multiple kinds features, and proposes renovations that can effectively combine multiple OC-SVM classifiers to achieve an optimal decision boundary. Section 4.4 presents the experimental protocol and analyzes the results based on commonly used performance metrics. Section 4.5 summarizes the findings of this work.

4.2 An Overview of Image-Based Spoof Detection

Spoof detection approaches represent a common countermeasure to address the issue of spoofing and can be sensor-based or image-based. Sensor-based solutions exploit characteristics of vitality such as pulse oximetry, finger temperature, the electrical conductivity of the skin, and skin resistance [88, 103, 104]. These methods require additional hardware in conjunction with the biometric sensor, which makes the device expensive. This work focuses on the image-based approaches that commonly use machine-learning algorithms to deal with the problem.

Based on reviewing past research that has been carried out in the field of image-based fingerprint spoof detection, we propose a categorization (as shown in Figure 4.2) based on three broad families for the study. The categorization can be summarized as:

- **Features**: the use of different kinds of feature sets has a significant impact on the spoof detection performance.
- **Availability of Training Data**: a spoof detection approach can learn learn with both live and spoof samples, or with live samples only.
- **Learning Classifiers**: the learning classifiers may base on the Support Vector Machines (SVM) or other methodologies.

4.2.1 Feature Extraction for Spoof Detection

An image-based fingerprint spoof detector aims to disambiguate live fingers from fake (spoof) artifacts by exploiting their differences in dynamic behaviors of live fingertips (e.g., ridge distortion, perspiration) or static characteristics (e.g., textural characteristics, ridge frequencies, elastic properties of the skin). Thus far, four fingerprint liveness detection competitions (*LivDet*) have been conducted between 2009 and 2015. The static features that were extracted from a single fingerprint impression, have been widely used for the contestants. It is mainly because that compared to other approaches based on multiple impressions (i.e., dynamic features), the static features are much cheaper and more user-friendly (as shown in Figure 4.3).

Static features may concern textural characteristics, ridge frequencies, elastic properties of the skin, or a combination of these. From the results reported in these competitions [135, 37, 79], the local texture-based features have been shown to outperform other competing techniques based on anatomical (such as pore detection [72]) or perspiration [2] features. Hence, the experiments in this work are conducted using local textural features as shown in Table 3.1. Briefly, Grey Level Co-occurrence Matrix (GLCM) characterizes the texture of

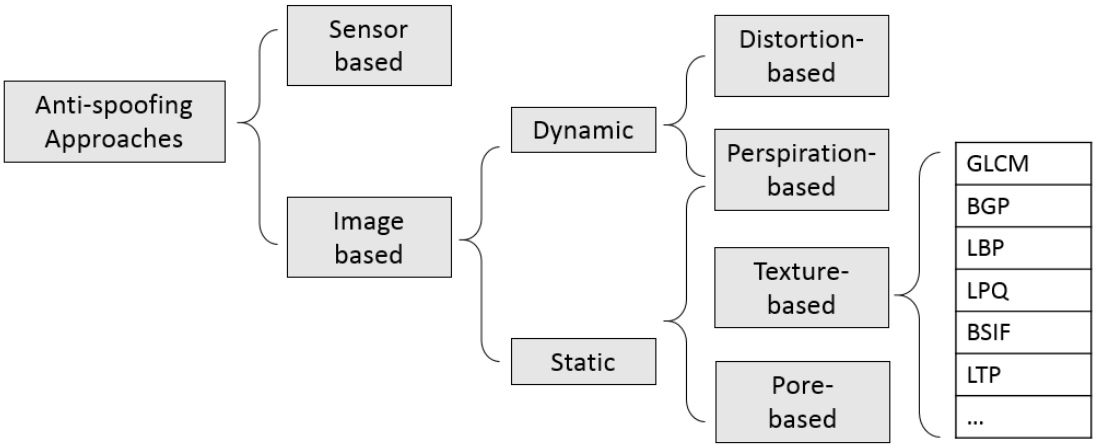


Figure 4.3: Categorization of current existing anti-spoofing approaches. We highlight the textual-based approaches and list several commonly used feature sets that provide comparable spoof detection accuracies.

an image by calculating the frequency of occurrence of pairs of pixels with specific values and in a specified spatial relationship; statistical measures are then extracted from this matrix [84]. Local Phase Quantization (LPQ) utilizes phase information computed locally in a window [36]. The phases of the four low-frequency coefficients are decorrelated and uniformly quantized. Binary Statistical Image Features (BSIF) encode texture information as a binary code for each pixel by linearly projecting local image patches onto a subspace, whose basis vectors are from natural images [35]. The Local Binary Pattern (LBP) operator compares a pixel with its neighbours, thresholds the ensuing results into a decimal value, and converts this value into a binary code [85]. Binary Gabor Patterns (BGP) encode textual information by convolving the image with Gabor filters and binarizing the responses [140].

A more detailed discussion of features in spoof detection can be found in Marasco and Ross's survey paper [70]. The authors also pointed out an open issue in the field of anti-spoofing is to develop interoperable approach for detecting spoofs under more complex attack scenarios, such as across different fingerprint sensors and across multiple fabrication materials. Since each of the above texture descriptors are expected to capture different attributes of live and spoof samples, one possible solution is to fuse the above texture descriptors and

adapt them to generalize over multiple spoof materials and fingerprint sensors.

4.2.2 Availability of Training Data

As mentioned earlier, most spoof detector adopt a machine learning approach where a classifier is trained to capture the concept of the "spoof" class and "live" class. The training samples of the "spoof" class needs to be created in laboratory by giving certain source fingerprints, and the fake prints can be obtained via the consensual method (i.e., with the collaboration of the user) or the non-consensual method (e.g., from a latent fingerprint) [37].

A variety of readily available materials such as latex, gelatin, woodglue, etc., have been used to fabricate fake fingerprints. Figure 5.4 shows example of fake fingerprint samples corresponding to four different fabrication materials and their source finger (from LivDet 2011 database [135]). As reported by Nixon et al. [89], there are more than fifty seven materials and material variants that have been identified for fake fingerprint fabrication. The flexibility in material choice leads to several concerns associated with the use of spoof samples in the design of spoof detection algorithms.

In practice, it is easy to obtain training samples pertaining to the "Live" class but difficult to obtain samples for the "Spoof" class, thereby leading to imbalanced training sets where the latter has substantially fewer training samples. Further, the training set for the spoof class may not have data corresponding to all possible types of fabrication materials. This makes it difficult for the classifier to reliably learn the concept of a spoof. In fact, it has been shown that spoof detection accuracy degrades sharply, when the test set has fake samples fabricated using materials that were previously "unseen" in the training set (as reported in [135, 37]). As spoof attacks evolve, it is likely that new and more sophisticated materials will be used to create fake fingerprints [5] thereby undermining existing learning-based spoof detectors.

The aforementioned concerns (related to interoperability across fabrication materials and limited spoof training samples) motivated us to consider approaching spoof detection via the

second category of frameworks, the *one-class classification (OCC)*, where only the samples from live fingers are used for training the classifier.

There are a great mount of existing works available for a deep investigation on the OCC frameworks. Tax and Duin [123, 124] and Schölkopf et al [112] have developed algorithms based on support vector machines (SVM) to tackle the problem of OCC using positive examples only (refer to Section 4.3.1). The main idea behind these strategies is to construct a decision boundary around samples of the live class so as to differentiate other sample.

However, this makes the problem harder than two-class classification because the decision boundary is determined by the data from one side of the boundary rather than the both sides. In the proposed ensemble of one-class SVM classifier, the decision boundary is further refined by using a relatively small number of spoof fingerprints in a validation phase. As discussed by Tax and Duin [126], the rationale behind the validation phase is to adjust the decision boundary to better classify the points that are in the vicinity of the boundary by utilizing negative examples (spoof fingerprints).

4.2.3 Learning Classifiers

Most spoof (or liveness) detection algorithms proposed in the literature are learning based, that is, they learn a decision policy to distinguish real fingerprints from fake ones based on a set of training samples. As mentioned earlier, this work focuses on the spoof detection approaches that only use the samples from the "Live" class, so this section pays more attention on the available OCC learning classifiers. An OCC learning classifier can be arguably categorized into two families according to whether it utilize the support vector machines (SVM) to tackle the problem.

- **Non SVM-Related OCC Algorithms:** Ridder et al. [105] conduct an experimental comparison of various OCC algorithms, including: (a) Global Gaussian approximation, (b) Parzen density estimation, (c) 1-Nearest Neighbor method, and (d) Gaussian approximation (combines aspects of (a) and (b)). Manevitz and Yousef [64] proposed a

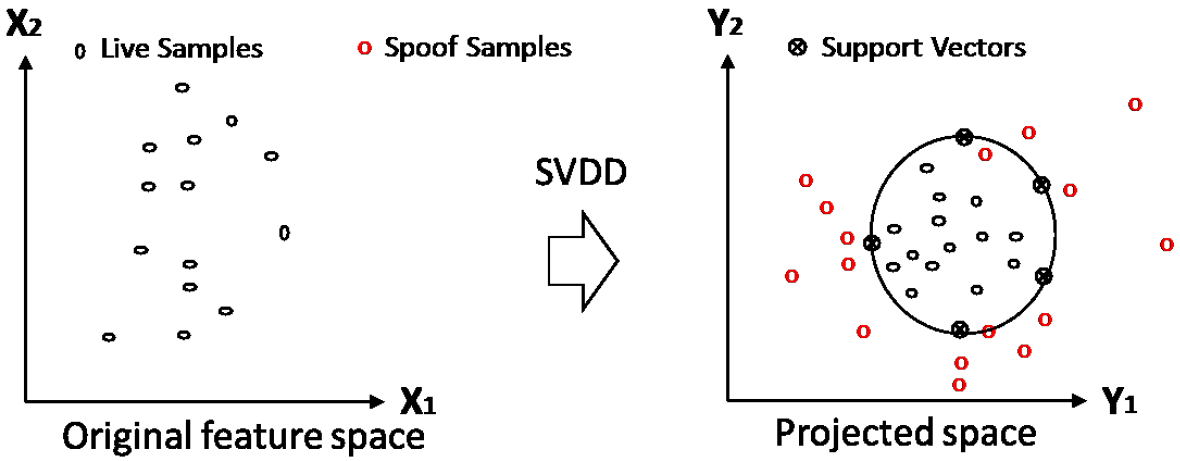


Figure 4.4: Illustration of the support vector data description (SVDD) scheme. The figure on the left shows a simple dataset in the input feature space. The figure on the right shows the data projected to a higher dimensional space using SVM approaches.

three-level feed-forward neural network to filter documents when only positive information is available. DeComite et al. [21] modify the C4.5 decision tree algorithm [96] to develop an algorithm that takes as input a set of labelled examples, a set of positive examples, and a set of unlabelled data, and then use these three sets to construct the decision tree. Letouzey et al. [22] design an algorithm which is based on positive statistical queries (estimates for probabilities over the set of positive instances)

- **SVM-Related OCC Algorithms:** The one-class classification problem is often solved by estimating the target density [78], or by fitting a model to the data support vector classifier [10]. Tax and Duin [123, 124] seek to solve the problem of OCC by distinguishing the positive class from all other possible patterns in the pattern space. Instead of using a hyperplane to distinguish between two classes, a hypersphere is found around the positive class data that encompasses almost all points in the data set with the minimum radius. This method is called the Support Vector Data De-

scription (SVDD), which also be used in this work. Thus training this model has the possibility of rejecting some fraction of the positively-labeled training samples, when this sufficiently decreases the volume of the hypersphere. Furthermore, the hypersphere model of the SVDD can be made more flexible by introducing kernel functions. Tax [122] considers a Polynomial and a Gaussian kernel and found that the Gaussian kernel works better for most data sets. A drawback of this technique is that they often require a large data set; in particular, in high dimensional feature spaces, it becomes very inefficient. Also, problems may arise when large differences in density exist. Samples in low-density areas will be rejected although they are legitimate objects.

Schölkopf et al. [112, 111] present an alternative approach to the work mentioned above of Tax and Duin on OCC using a separating hyperplane. The difference between theirs and Tax and Duin's approach is that instead of trying to find a hypersphere with minimal radius to fit the data, they try to separate the surface region containing data from the region containing no data. This is achieved by constructing a hyperplane which is maximally distant from the origin, with all data points lying on the opposite side from the origin and such that the margin is positive. Their paper proposes an algorithm that computes a binary function that returns $+1$ in small regions (subspaces) that contain data and -1 elsewhere. The data is mapped into the feature space corresponding to the kernel and is separated from the origin with maximum margin. They evaluate the efficacy of their method on the US Postal Services data set of handwritten digits and show that the algorithm is able to extract patterns which are very hard to assign to their respective classes and a number of outliers were identified. Figure 4.5 intuitively demonstrates this approach when multiple kinds of feature sets are involved.

Manevitz and Yousef [65] propose a different version of the one class SVM which is based on identifying outlier data as representative of the second class. The idea of this methodology is to work first in the feature space, and assume that not only is the origin the second class, but also that all data points close enough to the origin are to

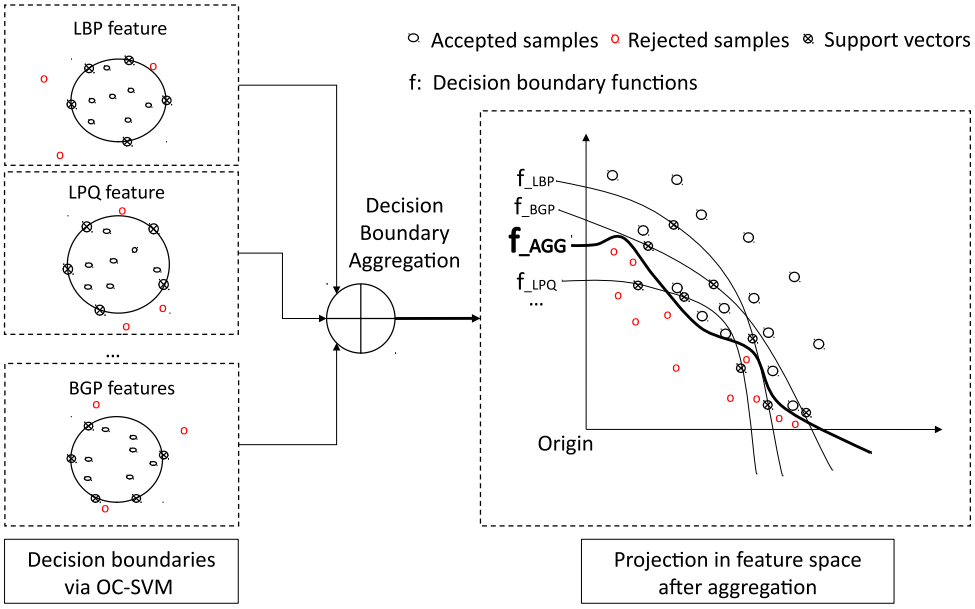


Figure 4.5: Illustration of the proposed ensemble of OC-SVMs. Multiple OC-SVMs are built based on different feature sets, and their decision boundaries in the projected space are adjusted to minimize the volume of hypersphere that contains the training data.

be considered as noise or outliers. The vectors lying on standard sub-spaces of small dimension (i.e. axes, faces, etc.) are treated as outliers.

Classifiers are commonly ensembled to provide a combined decision by averaging the estimated posterior probabilities. In this work, we also implement a Sum combination rule to ensemble mulitple OC-SVMs. Besides, when Bayes theorem is used for the combination of different classifiers, under the assumption of independence, a product combination rule can be used to create classifier ensemble. The outputs of the individual classifiers are multiplied and then normalized (also called the logarithmic opinion pool [7]). In OCC, as the information on the non-positive data is not available, in most cases, the outliers are assumed to be uniformly distributed and the posterior probability can be estimated. Tax [122] mentions that in some OCC methods, distance is estimated instead of probability for one class classifier ensembling. Tax observes that the use of ensembles in OCC improves performance, especially when the product rule

is used to combine the probability estimates.

Yu [138] proposes an OCC algorithm with SVMs using positive and unlabeled data, and without labeled negative data, and discusses some of the limitations of other OCC algorithms [125, 139, 112, 65]. Yu comments that in the absence of negative examples, OC-SVM requires a much larger amount of positive training data to induce an accurate class boundary.

4.3 Proposed Ensemble of OC-SVMs Approach

4.3.1 Conventional OC-SVM

The one-class paradigm, also known as single-class classification or anomaly/novelty detection, is a learning scheme developed by Schölkopf et al [112]. One-class paradigm allows for the modeling of just a single class of patterns (e.g., real live fingerprints), and distinguishing them from all other possible patterns (e.g., spoof fingerprints fabricated by different materials). Tax and Duin [126] constructed a hypersphere with radius $R > 0$ and center a around the positive class data, which encompasses almost all points in the data set, while allowing for some samples to be excluded as outliers. This method is called Support Vector Data Description (SVDD), and the hypersphere formulation involves solving the following quadratic programming optimization problem:

$$\begin{aligned} &\arg\min_{\mathbf{a},R,\xi} \left\{ R^2 + \frac{1}{N\nu} \sum_i \xi_i \right\}, \\ &\text{subject to} \quad ||\phi(\mathbf{x_i}) - \mathbf{a}||^2 \leq R^2 + \xi_i \quad \xi_i \geq 0. \end{aligned} \tag{4.1}$$

Here, the training set is denoted as $\{\mathbf{x_i}\}, i = 1 \ldots N$, where $\mathbf{x_i}$ are column vectors. The term $\phi(\mathbf{x_i})$ is a non-linear mapping function that maps each input feature vector to a higher dimensional space. ν is a predefined regularisation parameter that governs the trade-off between the size of the hypersphere and the fraction of data points falling outside the hypersphere, i.e., the fraction of training examples that can be classified as outliers. The ξ_i terms

are the slack variables that allow some of the data points to lie outside the hypersphere. The Lagrange multipliers $\alpha_i \geq 0$ and $\gamma_i \geq 0$ are used to solve Eqn. (4.1):

$$\begin{aligned} L(\mathbf{a}, R, \xi, \alpha_i, \gamma_i) &= R^2 + \frac{1}{N\nu}\sum_i \xi_i - \sum_i \alpha_i \{R^2 \\ &+ \xi_i - (||\phi(\mathbf{x_i})||^2 - 2\mathbf{a}\cdot\phi(\mathbf{x_i}) + ||\mathbf{a}||^2)\} - \sum_i \gamma_i \xi_i. \end{aligned} \tag{4.2}$$

L should be minimized with respect to $\mathbf{a}, R$ and ξ, and maximized with respect to α_i and γ_i. When L's partial derivatives w.r.t $\mathbf{a}$ and ξ_i are set to zero, it results in the following constraints:

$$\begin{aligned} \frac{\partial L}{\partial \mathbf{a}} : &\quad \mathbf{a} = \frac{\sum_i \alpha_i \phi(\mathbf{x_i})}{\sum_i \alpha_i} = \sum_i \alpha_i \phi(\mathbf{x_i}). \\ \frac{\partial L}{\partial \xi_i} : &\quad \frac{1}{N\nu} - \alpha_i - \gamma_i = 0. \end{aligned} \tag{4.3}$$

Eqn. (4.3) suggests that the center of the hypersphere is a linear combination of the input vectors. Further, because $\alpha_i \geq 0$ and $\gamma_i \geq 0$, the Lagrange multiplier γ_i can be removed when we require that $0 \leq a_i \leq \frac{1}{N\nu}$. As a result, the dual problem for Eqn. (4.1) can be written as:

$$\begin{aligned} \arg\max_{\alpha_i} &\left\{ \sum_i \alpha_i (\phi(\mathbf{x_i}) \cdot \phi(\mathbf{x_i})) - \sum_{i,j} \alpha_i \alpha_j (\phi(\mathbf{x_i}) \cdot \phi(\mathbf{x_j})) \right\}, \\ &\text{subject to} \quad 0 \leq \alpha_i \leq \frac{1}{N\nu}. \end{aligned} \tag{4.4}$$

When a training sample $\mathbf{x_i}$ satisfies the inequality $||\phi(\mathbf{x_i}) - \mathbf{a}||^2 < R^2 + \xi_i$, the constraint in Eqn. (4.4) is satisfied and the corresponding Lagrange multiplier α_i will be zero. For training samples that satisfy the equality $||\phi(\mathbf{x_i}) - \mathbf{a}||^2 = R^2 + \xi_i$, the constraints have to be enforced and the Lagrange multiplier will become greater than zero. This can be summarized as:

$$\begin{aligned} ||\phi(\mathbf{x_i}) - \mathbf{a}||^2 < R^2 + \xi_i \rightarrow &\qquad \alpha_i = 0 \quad \text{(inlier)} \\ ||\phi(\mathbf{x_i}) - \mathbf{a}||^2 = R^2 + \xi_i \rightarrow &\quad 0 < \alpha_i < \tfrac{1}{N\nu} \quad \text{(border SVs)} \\ ||\phi(\mathbf{x_i}) - \mathbf{a}||^2 > R^2 + \xi_i \rightarrow &\qquad \alpha_i = \tfrac{1}{N\nu} \quad \text{(outlier)}. \end{aligned}$$

After the center $\mathbf{a}$ and the radius R of the hypersphere are deduced, a test sample $\mathbf{z}$ can be detected as an outlier, i.e., assigned to the spoof class, if its distance to the center of the hypersphere is greater than the radius:

$$||\phi(\mathbf{z}) - \mathbf{a}||^2 = (\phi(\mathbf{z}) \cdot \phi(\mathbf{z})) - 2\sum_i \alpha_i(\phi(\mathbf{z}) \cdot \phi(\mathbf{x_i})) + \sum_{i,j} \alpha_i\alpha_j(\phi(\mathbf{x_i}) \cdot \phi(\mathbf{x_j})) > R^2.$$

In this work, the LIBSVM package [13] (ver 3.18) was used to solve the above optimization problem.

4.3.2 Proposed Ensemble of OC-SVMs

Juszczak and Duin, Biggio et al. and Medina-Perez et al. aimed to improve the accuracy of classification by employing ensembles based on several instances of the same base classifiers. The techniques used in for feature subspace partition included fixed combining rules, RSM and bagging. In general, the ensembles exhibit robustness and diversity, which allow them to obtain better classification accuracy.

In the context of spoof detection, if the training data resides in a single feature space (e.g., LPQ feature space), the use of a single OC-SVM classifier can easily lead to overfitting problems. This is because the hypersphere attempts to tightly encompass live fingerprints and so a single feature space may not adequately capture the concept of a “live" class. To overcome this drawback, diversity is intuitively induced by combining several OC-SVMs that are based on descriptions of live fingerprint patterns in different feature spaces. Two different combination methods, the majority voting and the LSE-based weighting approach, are used here for combining the outputs of multiple OC-SVMs.

Majority voting is the simplest method for combining multiple classifiers. Multiple OC-SVMs, pertaining to different feature sets but derived from the same training samples, will result in multiple hyperspheres as decision functions, $f_j(\mathbf{x})$, $j = 1 \ldots L$. Here, L is the number of feature sets (OC-SVMs) considered. Let $\mathbf{y_i}$ denote the class label.

While $\mathbf{y_i}$ is always $+1$ for the training data (i.e., the live class), $\hat{\mathbf{y}}_\mathbf{i}$, which denotes the output label of an OC-SVM classifier, could be -1 (i.e., the spoof class) or $+1$. Let $N_k(\mathbf{x}) = \sum_{j=1}^{L} \mathbb{I}(\hat{\mathbf{y}} = k | f_j(\mathbf{x}))$ where $k \in \{+1, -1\}$, denote the number of OC-SVMs that assign the input sample to the live or spoof class. Then the final decision of the OC-SVM ensemble via majority voting, $f_{MV}(\mathbf{z})$, for a test sample, $\mathbf{z}$, is determined by:

$$f_{MV}(\mathbf{z}) = \arg\max_{k}(N_k(\mathbf{z})) \quad k \in \{+1, -1\}. \tag{4.5}$$

An alternative to majority voting is the LSE-based weighting approach. The LSE-based weighting technique assigns different weights to individual OC-SVMs based on their classification accuracy. In the training phase, the weight vector $\mathbf{w}$ is estimated as $\hat{\mathbf{w}} = \mathbf{A}^{-1}\mathbf{y}$, where $\mathbf{A} = (f_j(\mathbf{x}_i))_{N \times L}$ consists of the estimated class label of each OC-SVM on training samples, and $\mathbf{y} = (\mathbf{y_i})_{N \times 1}$. The final decision of the OC-SVM ensemble for a given input sample $\mathbf{z}$ due to the LSE-based weighting is determined by:

$$f_{LSE}(\mathbf{z}) = sign\{\hat{\mathbf{w}} \cdot (f_j(\mathbf{z}))_{L \times 1}\}. \tag{4.6}$$

Since the performance of the LSE-based weighting approach was consistently better than the majority voting approach, only results from the LSE-based weighting are reported.

As stated earlier, one of the challenges in one-class classification is to determine how tightly the boundary should fit the training data. We propose two adjustments to the proposed ensemble OC-SVM scheme to address this concern. Firstly, the global regularisation parameter ν, that governs the trade-off between the radius of each hypersphere and the fraction of training data falling outside of the hypersphere, is gradually adjusted in the interval $[0.1\%, 10\%]$ in increments of 0.001. The LSE-based weights are also adjusted to optimize the detection accuracy during the training phase. In order to evaluate the detection accuracy, the Correct Detection Rate (CDR) on live fingers is defined as follows:

- CDR of "Live" fingers (CDR_L): the proportion of live samples that are correctly classified as "Live".

The rationale behind the adjustment is for the decision hypersphere to better fit the training data in every feature space rather than on a single feature space.

Secondly, the hypersphere is further refined by using a relatively small number of spoof fingerprints in a **v**alidation phase. As discussed by Tax and Duin [126], the rationale behind the validation phase is to adjust the decision hypersphere to better classify the points that are in the vicinity of the hypersphere of any one of the L OC-SVMs by utilizing negative examples (spoof fingerprints). The available negative examples are labelled as outliers. Hence, they decrease the fraction of positive training samples that are classified as outliers, which leads to a readjustment of the global regularisation parameter ν. Hence, the following performance metric is defined to validate the detection accuracy on spoof fingerprints.

- CDR of "Spoof" fingers ($\text{CDR}_{\mathcal{S}}$): the proportion of fake samples that are correctly classified as "Spoof".

4.4 Experimental Results

4.4.1 Database and Protocol

We used the *LivDet2011* [135] and *LivDet2013* [37] dataset for performance assessment of the proposed ensemble of one-class SVMs (as shown in Table 4.1). The *LivDet2011* dataset comprises images from 4 different sensors. Corresponding to each sensor, there are 1,000 live and 1,000 fake fingerprint samples in the training set, and the same number of samples, but from different subjects, in the test set. The spoof materials used for Biometrika and ItalData sensors were gelatine, latex, ecoflex (platinum-catalysed silicone), silgum and wood glue (400 each), while the fake fingerprints for Digital Persona and Sagem sensors were made of gelatine, latex, playdoh, silicone and wood glue (400 each).

The *LivDet2013* dataset consists of images from four different sensors as well. The spoof materials used for spoof samples were Body Double, latex, Play-Doh and wood glue for Crossmatch and Swipe sensors, and gelatine, latex, ecoflex, modasil and wood glue for

Table 4.1: Characteristics of the datasets in the *LivDet2011* and *LivDet2013* competition. More details can be found in [135] and [37].

LivDet2011 DATASET

	#1	#2	#3	#4
Sensor	Biometrika	Italdata	Persona	Sagem
Resolution(dpi)	500	500	500	500
Image Size	312*372	640*480	352*384	355*391
Live Samples	2000	2000	2000	2000
Live Subjects	200	200	20	52
Fake Samples	2000	2000	2000	2000
Fake Subjects	34	34	68	42

LivDet2013 DATASET

	#1	#2	#3	#4
Sensor	Biometrika	Italdata	Crossmatch	Swipe
Resolution(dpi)	569	500	500	96
Image Size	315*372	640*480	800*750	208*1500
Live Samples	2000	2000	2250	2250
Live Subjects	50	50	94	100
Fake Samples	2000	2000	2250	2250
Fake Subjects	15	15	45	45

Biometrika and Italdata sensors. The images were divided into two equal datasets, training and testing. Live images came from 300 fingers from 50 subjects for Biometrika and Italdata datasets, 940 fingers representing 94 subjects for Crossmatch dataset, and 1000 fingers from 100 subjects for Swipe dataset. Spoof images come from approximately 225 fingers representing 45 people for the Crossmatch and Swipe Datasets and 100 fingers representing 15 subjects for the Biometrika and Italdata datasets.

It can be noted that the following experiments have placed different emphasis on these two datasets. To compare the proposed method against state-of-the-art spoof detection algorithms that exhibit interoperability across fabrication materials, the experimental protocol described in [101] is carefully followed in this work. Rattani and Ross [101] divided the test set of *LivDet2011* dataset into two non-overlapping subsets according to the fabrication materials used. Each subset consists of 500 live samples and 500 fake fingerprints,

where 200 fake fingerprints correspond to two fabrication materials that are used during the training stage (these are the “known" materials) and 300 fake fingerprints correspond to the rest three fabrication materials that are *not* used during the training stage (these are the “novel" materials). Although those fake fingerprints in the training set were useless for the proposed ensemble of OC-SVMs classifier, fake samples in the test set are used to evaluate the detection accuracy. As noted, the detection accuracies on all ten possible combinations of known materials (and ten combinations of novel materials as well) are reported to prove the consistency.

Although seven different fabrication materials are available in the *LivDet2013* dataset, not every sensor has corresponding images for the complete set. This is the possible reason that most of the current literature utilized the dataset to assess the detection accuracy rather than analyze the impact of fabrication materials. In this work, the detection accuracy of the proposed ensemble OC-SVM approach is reported to compare with multiple contestants in the competition. Furthermore, the performance improvement of the proposed approach brought by increasing the number of spoof samples in the validation is analyzed on the same dataset as well.

As pointed in the competition report of *LivDet2013* [37], the live images collected with the Crossmatch sensor turned to be especially difficult to recognize for most of the algorithms, which leads to a further investigation. Moreover, the samples from Swipe sensor have a significantly lower resolution compared to all the other sensors in the two dataset (as seen in Table 4.1). In order to generate unbiased results, the experiments about detecting spoofs collected across different fingerprint sensors are only implemented on the *LivDet2011* dataset.

To show the advantage of the proposed ensemble OC-SVM on the detection of novel fabrication materials, CDR_S is intuitively divided into two parts:

- CDR of “Known” fake samples (CDR_K): the proportion of fake samples generated using known materials (i.e., materials encountered in the training set) that are correctly classified as “Spoof”;

- CDR of "Novel" fake samples (CDR_N): the proportion of fake samples generated using novel materials (i.e., materials *not* encountered in the training set) that are correctly classified as "Spoof".

In the following, the known materials are also noted as training materials. Although they were *not* used to train any one-class classifiers in this work, the rest of materials are used as "Novel" materials to evaluate CDR_N.

Table 4.2: Establishing the **baseline performance** using conventional binary SVM (B-SVM) and one-class SVM (OC-SVM) using single feature set on the *LivDet2011* dataset. The listed combinations of training materials are only required by the B-SVM classifier, and the rest materials are used as "novel" materials to evaluate the CDR_N of both classifiers.

	Training Materials	GLCM Feature		BSIF Feature		LPQ Feature		LBP Feature		BGP Feature	
	(Only Used by B-SVM)	B-SVM	OC-SVM	B-SVM	OC-SVM	B-SVM	OC-SVM	B-SVM	OC-SVM	B-SVM	OC-SVM
1	Latex + EcoFlex	47.4	40.2	51.1	37.4	56.2	35.6	53.5	28.5	58.2	40.4
2	WoodGlue + Latex	55.0	38.0	53.9	37.7	52.5	38.3	58.2	30.2	60.0	37.5
3	Gelatine + Latex	55.7	42.2	53.1	39.8	58.6	35.3	53.5	28.3	55.0	40.4
4	Silgum + Latex	50.2	30.2	48.8	29.9	46.9	44.0	47.4	27.5	53.4	33.3
5	EcoFlex + Silgum	50.2	28.9	51.9	39.8	58.6	34.2	49.7	33.9	55.0	33.3
6	Gelatine + EcoFlex	47.0	37.9	41.4	33.5	56.3	43.3	40.0	33.3	50.2	40.2
7	Silgum + Gelatine	53.9	40.4	48.4	33.3	52.1	40.8	47.0	38.0	53.5	37.5
8	WoodGlue + Silgum	47.4	42.2	42.1	35.4	54.2	43.3	40.9	33.3	47.4	38.0
9	Gelatine + WoodGlue	50.2	31.2	49.3	33.3	52.1	39.9	47.9	31.2	47.9	37.2
10	WoodGlue + EcoFlex	50.4	42.2	51.9	39.8	54.2	43.3	40.0	33.9	53.4	38.0
	Average CDR_N	**50.7**	**37.3**	**49.2**	**36.0**	**54.2**	**39.8**	**47.8**	**31.8**	**53.4**	**37.6**

4.4.2 Conventional B-SVM and OC-SVM

This section evaluates the performance of conventional binary SVM (B-SVM) and conventional one-class SVM classifiers. This provides a baseline for the experiments in the subsequent sections.

Five different kinds of texture descriptors were used in this work, and their dimensionalities were 40, 516, 256, 54 and 216 for GLCM, BSIF, LPQ, LBP and BGP, respectively. The training set used in this experiment consists of 400 live samples and 400 fake fingerprints made using two fabrication materials (200 each). Note that only B-SVM classifiers use fake fingerprints for training. In this experiment, no validation phase for the OC-SVM is implemented and the parameters of both classifiers are tuned following a conventional estimation procedure. Table 4.2 shows the correct detection rates on "novel" fake samples (CDR_N) using conventional B-SVM and conventional OC-SVM (in parentheses) in the *LivDet2011* dataset. Note that all the accuracy rates reported here are carried out on the exact same test set that was stated earlier. Since similar trends were observed across all 4 sensors, only results from the Biometrika sensor are reported.

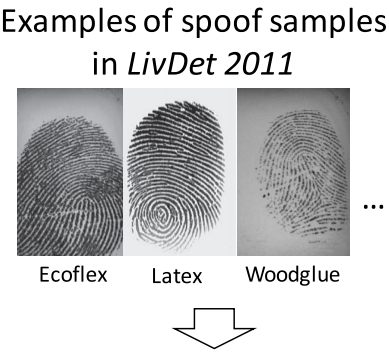
It can be seen that both conventional classifiers *do not* provide an acceptable correct detection accuracy on the fake samples manufactured using novel materials. The conventional OC-SVM classifier performed worse than the conventional binary SVM. However, we noted that the conventional OC-SVM provides higher correct detection rates on live fingers (CDR_L) than conventional B-SVM in some cases (results not shown here). These results are not surprising because the conventional OC-SVM is unable to find a tight enough decision boundary when using only the live fingerprints for training, leading to a higher CDR_L but a much lower CDR_N compared to B-SVM.

Table 4.3: Performance of the proposed ensemble of OC-SVMs compared to the automatic adaptation approach in [101] and conventional binary SVM (B-SVM) on the *LivDet2011* dataset. The correct detection rates tested on previously known materials (CDR_K) and on novel materials (CDR_N) are reported, respectively. It is notable that except the listed materials for training, the rest materials are tested as "novel materials".

Part A. Performance of Biometricka-Based Spoof Detectors.							
	Training Materials	Proposed Ensemble of OC-SVMs		Automatic Adaptation (LBP)		B-SVM (Esemble of Features)	
		CDR_K	CDR_N	CDR_K	CDR_N	CDR_K	CDR_N
1	Latex + EcoFlex	92.8	89.2	95.0	86.6	77.2	63.8
2	WoodGlue + Latex	94.0	92.8	94.0	90.4	78.0	65.0
3	Gelatine + Latex	91.8	91.0	92.2	86.6	75.8	61.8
4	Silgum + Latex	91.0	90.4	91.0	86.0	69.8	61.6
5	EcoFlex + Silgum	92.8	89.2	91.0	82.0	77.8	67.8
6	Gelatine + EcoFlex	91.0	91.0	92.8	85.8	77.8	66.2
7	Silgum + Gelatine	92.8	90.8	90.0	84.2	77.8	66.2
8	WoodGlue + Silgum	92.8	92.8	90.8	85.6	78.0	66.0
9	Gelatine + WoodGlue	90.0	89.2	91.8	89.2	72.8	64.0
10	WoodGlue + EcoFlex	91.0	89.2	94.0	87.2	72.0	66.4
Average and Std. Dev.		92.0 ± 1.2	90.6 ± 1.3	92.3 ± 1.6	86.4 ± 2.2	75.7 ± 2.9	64.9 ± 1.9
Part B. Performance of Italdata-Based Spoof Detectors.							
	Training Materials	Proposed Ensemble of OC-SVMs		Automatic Adaptation (LPQ)		B-SVM (Esemble of Features)	
		CDR_K	CDR_N	CDR_K	CDR_N	CDR_K	CDR_N
1	Latex + EcoFlex	82.2	81.6	82.8	83.0	71.4	69.6
2	WoodGlue + Latex	84.2	83.4	85.1	85.9	72.0	69.6
3	Gelatine + Latex	82.8	82.8	84.9	83.7	73.2	69.8
4	Silgum + Latex	82.4	81.6	85.0	83.9	66.8	66.2
5	EcoFlex + Silgum	82.8	82.0	81.2	74.3	68.8	63.8
6	Gelatine + EcoFlex	82.2	82.2	80.7	78.1	76.8	72.1
7	Silgum + Gelatine	83.6	82.0	82.9	83.0	71.4	70.0
8	WoodGlue + Silgum	84.6	84.6	85.6	82.0	71.4	70.0
9	Gelatine + WoodGlue	84.6	83.6	82.9	83.4	69.6	69.6
10	WoodGlue + EcoFlex	82.0	81.6	83.2	79.8	72.2	71.0
Average and Std. Dev.		83.1 ± 1.0	82.5 ± 1.0	83.4 ± 1.6	81.7 ± 3.2	71.4 ± 2.5	69.2 ± 2.3

Table 4.3 (cont'd)

Part C. Performance of Digital Persona-Based Spoof Detectors.							
	Training Materials	Proposed Ensemble of OC-SVMs		Automatic Adaptation (LPQ)		B-SVM (Esemble of Features)	
		CDR_K	CDR_N	CDR_K	CDR_N	CDR_K	CDR_N
1	Latex + PlayDoh	89.6	88.8	89.9	82.6	74.2	67.6
2	WoodGlue + Latex	88.8	88.2	88.1	81.0	76.0	67.6
3	Gelatine + Latex	88.8	88.8	90.0	82.9	75.6	68.2
4	Silicone + Latex	89.8	89.2	91.9	81.1	69.8	64.6
5	PlayDoh + Silicone	90.0	90.0	91.2	81.1	74.8	69.8
6	Gelatine + PlayDoh	89.6	88.8	84.6	76.2	77.8	70.4
7	Silicone + Gelatine	90.0	89.2	85.9	75.9	73.2	67.6
8	WoodGlue + Silicone	89.0	89.0	91.0	79.7	73.2	70.4
9	Gelatine + WoodGlue	89.4	88.8	88.8	79.5	70.2	68.2
10	PlayDoh + WoodGlue	86.8	86.2	90.5	83.6	71.2	69.2
Average and Std. Dev.		89.2 ± 0.9	88.7 ± 0.9	89.2 ± 2.3	80.4 ± 2.5	73.6 ± 2.5	66.8 ± 1.6
Part D. Performance of Sagem-Based Spoof Detectors.							
	Training Materials	Proposed Ensemble of OC-SVMs		Automatic Adaptation (LBP)		B-SVM (Esemble of Features)	
		CDR_K	CDR_N	CDR_K	CDR_N	CDR_K	CDR_N
1	Latex + PlayDoh	82.8	81.0	82.1	82.0	69.6	66.2
2	WoodGlue + Latex	82.2	82.2	80.6	79.0	70.2	65.4
3	Gelatine + Latex	82.8	82.8	87.0	83.4	69.0	67.8
4	Silicone + Latex	83.2	82.1	80.1	79.0	62.2	60.2
5	PlayDoh + Silicone	83.2	83.2	78.1	83.7	70.0	63.4
6	Gelatine + PlayDoh	82.8	82.6	81.4	79.8	71.4	66.1
7	Silicone + Gelatine	83.2	82.4	87.6	83.4	66.9	60.4
8	WoodGlue + Silicone	82.9	82.9	83.5	80.5	70.2	67.8
9	Gelatine + WoodGlue	82.8	82.8	87.3	84.7	65.2	62.4
10	PlayDoh + WoodGlue	81.2	81.0	80.7	83.2	68.4	63.4
Average and Std. Dev.		82.7 ± 0.6	82.3 ± 0.7	82.8 ± 3.2	81.9 ± 2.0	68.3 ± 2.7	64.3 ± 2.6



Feature Sets Used in the Proposed Ensemble of OC-SVMs	Decision on Above Examples	Accuracy on Entire Test Sets:
LBP + BGP	Live (X)	60.4 %
GLCM + LBP + BGP	Live (X)	64.0 %
GLCM + LBP + BGP + BSIF	Live (X)	73.5 %
GLCM + LBP + BGP + LPQ	Live (X)	79.7 %
GLCM + LBP + BGP + BSIF + LPQ	**Spoof (√)**	**83.2 %**

Figure 4.6: The decisions changed by the different combinations of feature spaces that are used in the proposed ensemble of OC-SVMs in the *LivDet2011* dataset.

4.4.3 Analysis of Proposed Ensemble Strategy

In order to investigate the impact of the proposed ensemble strategy, Table. 4.4 compared the detection accuracy on novel spoof materials (CDR_N) from different combinations of feature sets used in the proposed ensemble of OC-SVMs. It it noted that the proposed combination of GLCM, LBP, BGP, BSIF and LPQ overcame the other combinations of feature spaces.

Further, Fig. 4.6 provides the values of optimized regularisation parameters, ν, and the corresponding CDR_S on different combination of feature sets. It is noted that although the CDRs on spoof fingers are consistently increased by adding more feature sets, the regularisation parameters and the weight vectors ($\mathbf{w}$, which has not been shown here) are fluctuant upon different combinations.

Take the detection result on S1, a single spoof sample fabricated using Silgum (sample

Table 4.4: The correct detection accuracy on novel spoof materials (CDR_N) when different combinations of feature sets are used in the proposed ensemble of OC-SVMs (*LivDet2011* dataset).

Used Feature Sets or Combinations	**Average CDR_N on Different Sensors**			
	Biometrika	Italdata	Persona	Sagem
GLCM feature [84]	46.5	40.1	47.3	47.4
LPQ feature [36]	55.2	53.2	46.3	50.1
BSIF feature [35]	56.1	53.0	55.3	51.7
LBP feature [83]	56.7	52.7	53.3	57.3
BGP feature [140]	58.5	53.1	55.3	49.9
LBP+BGP (the best two)	67.2	61.9	64.8	63.5
GLCM+LBP+BGP	69.9	62.2	66.6	64.9
GLCM+LBP+BGP+BSIF	73.5	73.0	76.4	76.4
GLCM+LBP+BGP+LPQ	79.6	77.2	79.6	79.7
GLCM+LBP+BGP+BSIF + LPQ (Proposed)	**83.9**	**83.0**	**84.1**	**84.7**

ID 76_7), as an example. It was correctly classified as spoof by the ensemble of LBP and BGP features. However, by adding GLCM and BSIF features, the detection result on this particular sample flopped although the CDRs on the entire test set increased. It is eventually detected as spoof when all five feature sets are involved in the proposed ensemble approach. The fluctuant results on this random sample indicate the important role played by the proposed ensemble procedure in some extents.

4.4.4 Proposed Ensemble of OC-SVMs

This section evaluates the performance of the ensemble OC-SVM classifier, especially on novel materials. To achieve a fair comparison, two variations of the conventional B-SVM were used as baselines:

- A feature-level fusion of B-SVM (referred to as B-SVM-F): The feature sets are concatenated into a single feature vector and the concatenated feature vector is used to train the conventional B-SVM and generate the binary outputs.

- A decision-level fusion of B-SVM (referred to as B-SVM-D): Several B-SVM classifiers are trained, and each of them is trained on a different feature set to generate binary outputs, then those outputs are combined using the majority vote rule.

Table 4.3 reports the performance of the proposed ensemble OC-SVM compared to an adaptive approach (referred to as Automatic Adaptation) proposed earlier by Rattani and Ross [101], which was shown to significantly increase the correct detection rate on novel spoof materials (CDR_N).

As described earlier, the proposed ensemble OC-SVM utilizes the live fingerprint samples in the training set to generate the decision hypersphere. Although the spoof samples are not used by the learning procedure, they are used to readjust the decision boundary. In order to demonstrate the impact of this readjustment, the table reports the CDRs before and after the validation phase in the same cell. For example, the average CDR_N of the proposed OC-SVM is reported as $83.8 + 2.4\%$; this means the correct detection rate before the validation phase was 83.8%, and it increased by 2.4% after the validation. It must be noted that the number of fake samples used for validation is relatively small (50 spoof samples) compared to the larger training set (400 spoof samples) used by other approaches.

From Table 4.3, it can be seen that the ensemble OC-SVM provides significantly higher correct detection rates than the other two SVM-based fusion schemes. One possible reason for the poor performance of the feature-level B-SVM (B-SVM-F) is the curse of dimensionality. A similar feature-level fusion was implemented for the conventional one-class SVM (OCSVM-F) as well. However, the poor performance (as shown in Figure 4.7) on both live samples (60.0%) and spoof samples (50.4%) indicates that multiple feature sets need to be aggregated more carefully to avoid potential issues such like the curse of dimensionality. It must be noted that the decision-level fusion of B-SVM results in an improvement in accuracy for detecting novel materials (as evidenced by the CDR_N for B-SVM-D). This result substantiates our previous conjecture that the use of different feature sets can better characterize the concept of "live" fingerprints to some extent.

Without considering the impact of fabrication materials, the proposed ensemble OC-SVM is comparable with the best reported algorithm in *LivDet2011* (89% CDR_L and 81% CDR_S on the Biometrika sensor as shown in [135]). However, it does not exceed the performance of the automatic adaptation approach [101], which has the lowest error rates on the same database so far. Further, along with the accuracy improvement on detecting fake samples (as evidenced by CDR_N and CDR_K), the validation phase decreased the accuracy on detecting live samples (CDR_L is reduced by 0.5%). We address both issues in the next experiment.

4.4.5 Validation Using Spoof Samples

This section evaluates the performance of the proposed ensemble OC-SVM by increasing the number of fake samples used in the validation phase. Although fake samples are not required for training the classifier, they can be used to improve the overall accuracy by tuning the decision hypersphere (i.e., the global regularisation parameter ν). Figure 4.7 presents two bar plots of the average CDR on live and spoof samples under this experimental design.

Figure 4.7(b) indicates that when increasing the number of fake samples in the validation phase (from 0 to 400), the proposed ensemble OC-SVM provides consistently higher CDRs on spoof samples than the binary SVM classifier with feature level fusion (B-SVM-F) and decision-level fusion (B-SVM-D). Figure 4.7(a) suggests that the proposed ensemble OC-SVM can provide similar detection rates as the state-of-art detector in [101], although the former only needs half the number of spoof samples as the latter (200/400). Moreover, the detection rates of the proposed method are more stable with a smaller standard deviation across different fabricated materials. This suggests that the proposed method is not unduly impacted by the choice of fabrication material used for generating the spoof fingerprint. The average CDRs on live samples are presented in Figure 4.7(a). Similar to the results in Table 4.3, the CDR_L marginally decreased by 0.5% to 0.8% when the number of fake samples is increased during validation. This demonstrates the trade-off between the misclassification of live samples and the size of the decision hypersphere. However, compared to the performance

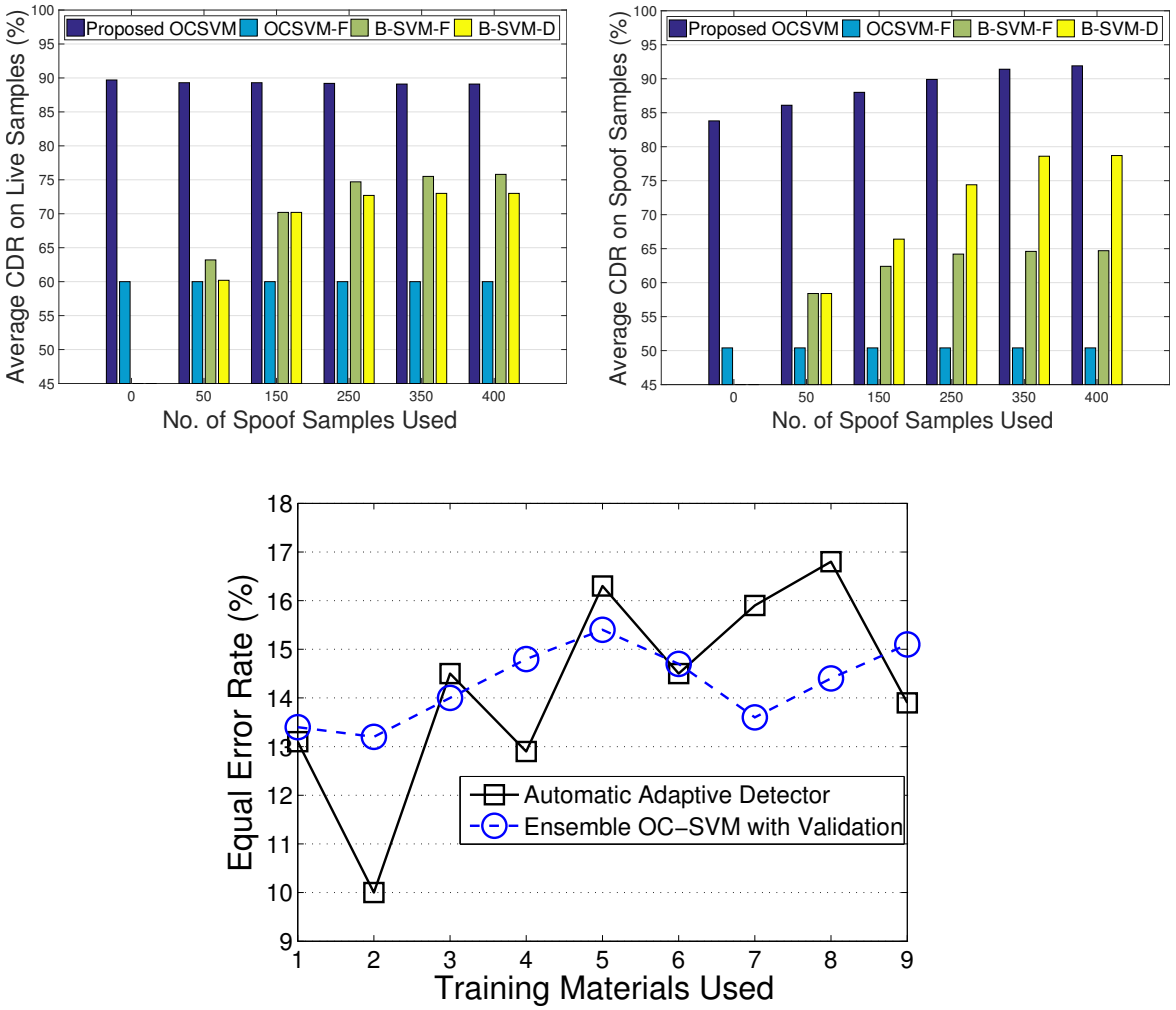


Figure 4.7: Performance of the ensemble OC-SVM after increasing the number of fake samples used in the validation phase. (a) CDR_L, (b) CDR_N and (c) when 200 spoof samples are used in validation phase. Training materials used here are as same as in Table 4.2 and 4.3.

Table 4.5: Performance of the proposed ensemble OC-SVM on the *LivDet2013* dataset. The Top 3 performed algorithms as reported in the competition are listed for a comparison.

Algorithms		CDR_L	CDR_K	CDR_N	Average
LivDet2013	Dermalog	73.1%	98.9%	84.6%	85.9%
	UniNap1	88.0%	85.4%	86.6%	87.0%
	Anonum3	74.4%	94.7%	83.3%	84.5%
Proposed Ensemble OC-SVM		80.1%	94.9%	84.6%	**87.6%**
Proposed Ensemble OC-SVM with Spoofs for Validation		88.0%	94.7%	88.0%	**90.2%**

Table 4.6: Performance of the proposed ensemble OC-SVM on on cross-sensor training.

Correct Detection Rates (CDR_L and CDR_S)		Average CDR ± STDERROR			
		Biometrika	Digital	Italdata	Sagem
Same Sensor	CDR_L	89.9 ± 0.1	88.3 ± 0.2	80.1 ± 0.2	77.9 ± 0.1
	CDR_S	83.0 ± 0.1	85.2 ± 0.1	73.2 ± 0.1	70.4 ± 0.2
Cross Sensors	CDR_L	87.6 ± 0.4	86.9 ± 0.2	77.1 ± 0.4	74.9 ± 0.2
	CDR_S	73.1 ± 0.4	74.1 ± 0.2	64.0 ± 0.3	62.9 ± 0.2

gain on spoof detection (an increase from 83.0% to 89.7%), the modest degradation in CDR_L is acceptable.

4.4.6 Performance on Cross-Sensor Training

In order to generalize the effectiveness of the proposed ensemble of OC-SVM classifier cross different fingerprint sensors, the classifier is trained only using the live fingerprint samples from one sensor (e.g. Biometrika) then tested on both live and spoof samples from the other three sensors (as shown in Table 4.6). The correct detection rates using the training samples from the same sensor are also reported as the baseline performance.

It is noted that when the training phase includes samples from different sensors, the correct detection accuracy on live samples (CDR_L) are degraded while the correct detection accuracy on spoof samples keep consistent. One possible reason is that, when the validation

sets include the spoof samples captured by different sensors, it is harder to approach a tight enough boundary which leads to a higher detection accuracy on the live class.

4.5 Summary and Future Work

In this work, the problem of spoof detection is posed as a one-class problem where the classifier learns the concept of a “live" fingerprint sample and uses this to reject spoof samples. It was shown that the accuracy of a conventional one-class SVM (OC-SVM) could be significantly improved by fusing multiple kinds of features and optimizing the decision functions across these features. Experimental analysis conducted on the LivDet2011 database show that the proposed ensemble OC-SVM outperforms Binary SVMs, and its performance is comparable with state-of-art spoof detection algorithms that are interoperable across fabrication materials. However, the proposed method requires much fewer spoof training samples than competing techniques. Further, the performance of the proposed method is observed to be stable across different fabrication materials. Thus, the proposed approach successfully mitigates some of the concerns associated with the issue of “imbalanced training sets” and “insufficient spoof samples” encountered by conventional spoof detection algorithms.

Bolded results shows cases in which a proposed diversity measure was significantly better than results obtained by a single classifier. In most cases the diversity measures were not worse than a single classi- fier, even often outperforming it. This is caused by the selection of mutually complementary classifiers to the pool. Therefore using more than one classi- fier lead to a better decision boundary, when a single model generated too generic solution. In several cases an ensemble with pool consisting of classi- fiers selected by diversity measured was not as good as a single classifier. This is caused by a fact that diversity measure itself is not the sole determinant of the accuracy. Probably in such cases classifiers with high diversity but low quality were chosen to an ensemble.

CHAPTER 5

PROPOSED FRAMEWORK FOR COMBINING ANCILLARY INFORMATION WITH BIOMETRIC TRAITS

5.1 Background

The term of "ancillary information", as discussed earlier in this dissertation, is used to contrast with the "primary biometric traits" and describes the fact that the ancillary information in themselves may not be suitable for the purpose of human recognition. However, ancillary information such as the biographic and demographic information of a user (e.g., name, gender, age, ethnicity), or the image quality of the biometric sample, anti-spoofing measurements, etc. are potential to be beneficial to the biometric system. The aim of this work is to design fusion frameworks that can *mitigate the limitations of existing frameworks* by simultaneously incorporating ancillary information in a biometric verification system. Figure 5.1 illustrates such a fusion framework integrating biometric match scores with ancillary information by taking the fingerprint verification system as an example.

The Generalized Additive Model (GAM), as was explored in Chapter 2, is devised to combine demographic attributes with biometric match scores and improve the matching accuracy. The Bayesian Belief Network proposed in Chapter 3 can effectively combine anti-spoofing measurements in the design of a biometric recognition system under spoof attacks. These works inspire us to integrate the GAM and BBN design in a general way which can retain all their hallmarks. However, the current public-domain anti-spoofing databases did not collect the corresponding demographic attributes of subjects. Instead, experiments of the proposed general fusion framework are conducted by integrating match scores with quality scores and liveness scores to render a final accept/reject decision. Figure 5.2 illustrates the general fusion framework by taking the fingerprint recognition system as the example.

To realize a fingerprint system capable of handling variations in the image quality as

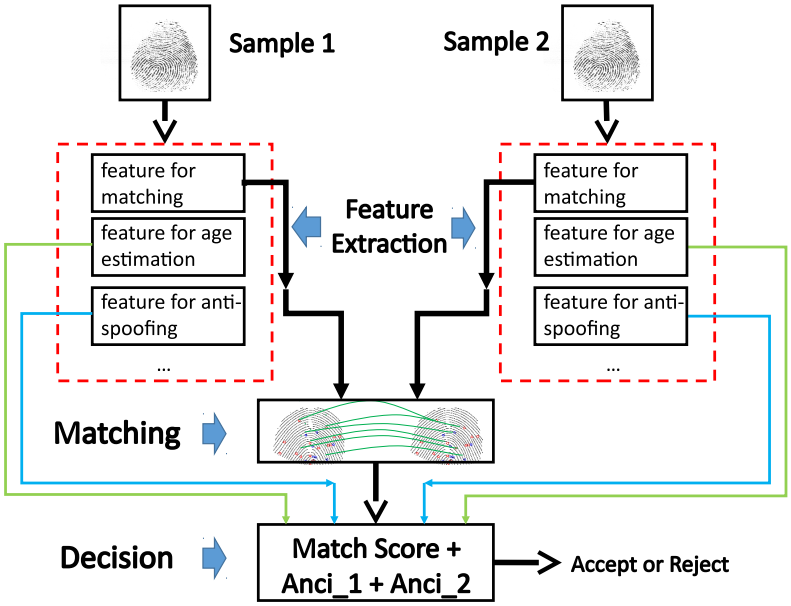


Figure 5.1: Illustration of the general fusion framework integrating biometric match scores with ancillary information. It shows that the ancillary information of two samples, such as quality scores and liveness scores, are self-reliant and independent with each other. Meanwhile, only the biometric match scores are corresponding to both samples and the identities they belong.

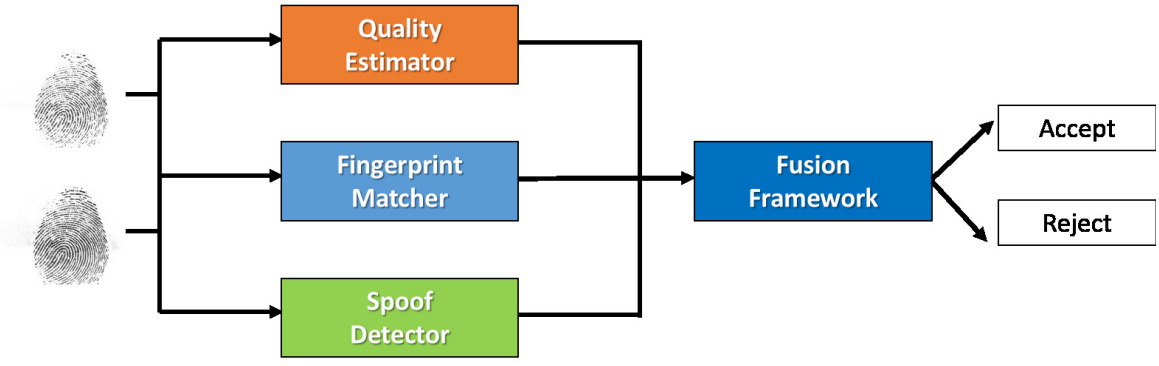


Figure 5.2: Illustration of the fusion framework integrating match scores with quality scores and liveness scores from two fingerprint samples, and rendering a final accept/reject decision.

well as robustness against spoof attacks, three major components are required: (a) image quality estimator yielding *quality scores* to indicate how good the image quality is [39, 132], (b) spoof detector yielding *liveness scores* to indicate how likely the fingerprint is from a live finger [66, 117], and (c) an effective *fusion framework* capable of incorporating quality scores and liveness scores with the fingerprint match scores to make an optimal accept/reject decision. Figure 5.2 shows a block diagram where image quality scores, liveness scores and match scores extracted from a pair of fingerprint images are integrated together in a fusion framework to render the final accept/reject decision.

In this chapter, we first categorize existing fusion frameworks incorporating ancillary information into two categories: (a) direct modelling, and (b) graphical modelling, based on the relationship assumed between the variables (i.e., match scores, liveness scores and quality scores). Then, these fusion frameworks are generalized to incorporate ancillary attributes with biometric match scores by categorizing ancillary attributes into direct variables and latent variables (as shown in Figure 5.3). The direct variables (such as liveness scores) which explicitly affect the system targets (e.g., anti-spoofing capability or verification of an identity) are exploited as nodes via a BBN design, while the latent variables (e.g., demographic attributes, quality scores or confidence measure) that do not directly influence the system targets are exploited to update the scores of nodes in the BBN design via the GAM method. Experiments are conducted with three variables, match scores, liveness scores and quality scores, and experimental results are analyzed according to the proposed performance metrics in Chapter 3, followed by a summarized finding of this work.

5.2 Related Literature

5.2.1 Introduction of Fingerprint Sample Quality

A fingerprint is a pattern of friction ridges on the surface of a fingertip. A good quality fingerprint have distinguishable patterns and features that provide more useful information for subsequent applications, i.e., verification or spoof detection. Several definitions [95] have

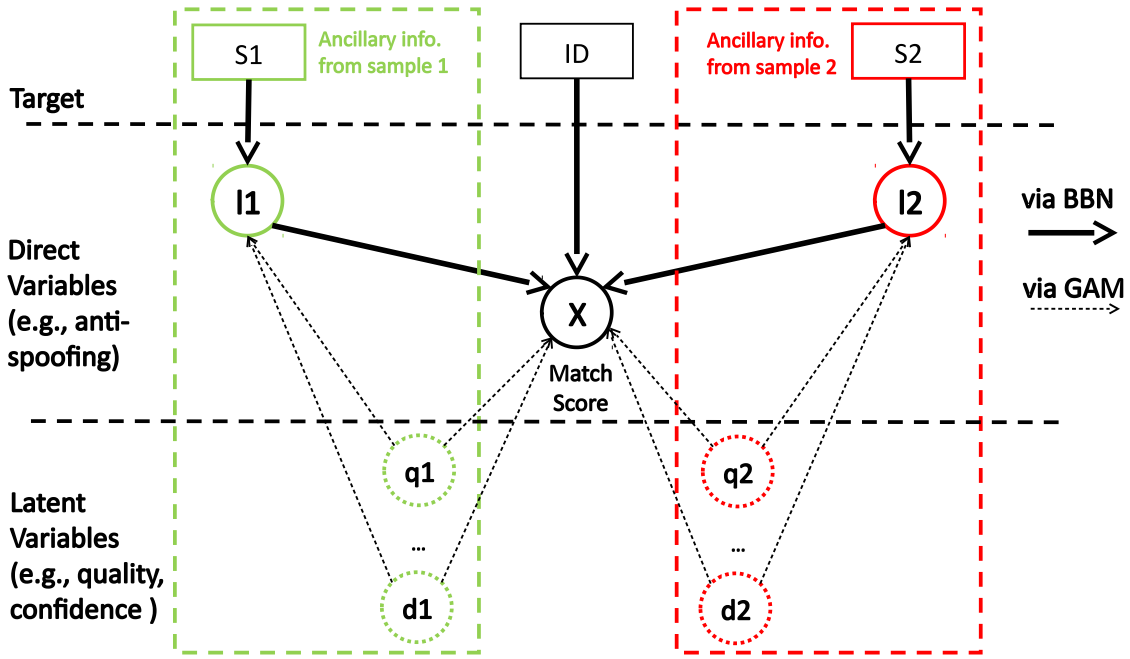


Figure 5.3: Illustration of the proposed general fusion framework. Ancillary information is categorized into direct variables (e.g. liveness scores) and latent variables (e.g. demographic attributes and quality scores), where the direct variables are involved into the BBN scheme as nodes and the latent variables are exploited to normalize the nodes of BBN prior to fusion.

been given for quality measures as (a) the degree of extractability of the features used for recognition, (b) the degree of conformance of fingerprint samples to some predefined criteria known to influence the recognition performance [39, 132], and (c) the degree of texture richness and general quality information, e.g., the sharpness, contrast, and detail rendition of the image [15, 86]. A quality detector is an algorithm designed to assess the quality of a fingerprint sample.

Figure 5.4 show the quality of the live and fake fingerprints fabricated using silicone and playdoh materials, estimated using Image Quality of Fingerprint (IQF) freeware developed by MITRE[1]. It can be observed that fake fingerprints fabricated using different materials obtained different quality values of the same finger. This is due to difference in the noise component in the fake fingerprint samples fabricated using different materials. As

[1] `http://www2.mitre.org/tech/mtf/`

a consequence, quality of the fake fingerprint samples usually vary across fabrication materials [99, 101]. Thus, emphasizing on the need of modelling influence of the sample quality of the fake fingerprints across fabrication materials in a framework against spoof attacks.

Figure 5.4: The quality measures of the fingerprint samples from (a) live finger and fake fingerprints fabricated using (b) latex, (c) gelatin and (d) woodglue, using IQF measurement on the LivDet 2011 database. It can be noticed that quality of the spoof vary across fake fabrication materials.

5.2.2 Taxonomy on Fusion Frameworks against Spoof Attacks

In this work, we categorize the existing fusion frameworks that combine match scores with liveness scores and image quality into: (i) *Direct modelling* or (ii) *Graphical modelling* (as shown in Figure 5.5). This taxonomy is based on whether the dependence between the variables involved is purely learned from the data or assumed via causal understandings.

(i) Direct modelling: Direct modelling based schemes attempt to favor an equivalent impact from each involved variable, and the relationship among variables are purely learned from the data.

Marasco et al. [68] proposed and compared different schemes for combining liveness scores with match scores. Compared to sequential schemes that invoked the spoof detector and the fingerprint matcher sequentially, parallel schemes that combined liveness scores and match scores as a two-dimensional input variable to classifiers such as Decision Trees, Naive Bayes and Neural Networks, were observed to result in a consistently higher accuracy. The

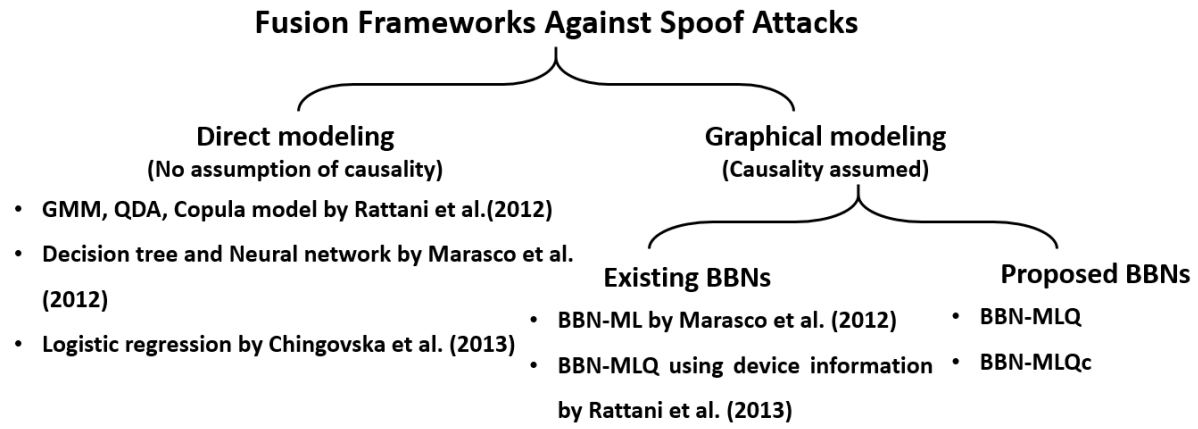


Figure 5.5: Taxonomy of existing fusion frameworks incorporating match scores, liveness scores and image quality.

authors remark that existing spoof detectors cannot be used for automated rejection of biometric samples until their detection accuracies are substantially improved.

Rattani and Poh [98] proposed a fusion framework that combined biometric sample quality and liveness scores with fingerprint match scores. The framework was implemented using three generative classifiers based on Gaussian Mixture Model (GMM), Gaussian Copula and Quadratic Discriminant Analysis (QDA). The results indicated that the GMM classifier provided the lowest overall error rate. The authors also established the benefit of fusing *both* quality and liveness scores in a fingerprint verification system. Chingovska et al. [16] proposed a fusion framework that incorporated LBP-based liveness scores with face match scores using logistic regression.

(ii) Graphical modelling: Graphical modelling based schemes assume a causal relationship between the variables. These schemes are often more accurate than direct modelling based schemes because the estimation of conditional probabilities is often simplified by such assumptions. Based on the assumptions about the relationship between the involved variables, different configurations of graphical models may be designed.

Marasco et al. [68] proposed a Bayesian Belief Network (BBN) model that combined match scores with liveness scores. This BBN (referred to as BBN-ML in this work) assumed

a one-directional influence of match scores on liveness scores. Based on this configuration, the conditional probability of an input fingerprint sample being from a genuine user, given its liveness scores and match score, was inferred. *The authors also demonstrated the effectiveness of the proposed BBN over direct modelling schemes that did not explicitly assume any relationship between match scores and liveness scores.*

However, the image quality was not incorporated by Marasco et al. in their proposed framework. *Thus, the variation in the match score and liveness scores as a function of the change in the sample quality was not taken into account.* Further, the framework also did not take into account the influence of latent factors - such as the type of sensor and fake fabrication material (i.e., material-specific characteristics) - on the liveness scores. Note that the fabrication materials can influence the quality of the fabricated spoofs and the liveness scores as pointed in [102].

Rattani et al. [100] proposed a fusion framework that fused the match scores, quality and liveness scores, while also accounting for the sensor influence, using a Bayesian framework. Although the model was not further generalized to consider the influence of other latent variables, it provided a good insight into the advantage of graphical modelling. The results indicated that the performance of the proposed model in a multi-sensor scenario, was comparable to a fusion framework that was trained and tested using fingerprint images from the same sensor. As Rattani et al. 's model is based on modelling a specific factor (i.e.. the *sensor* influence) on match scores, it is not further discussed in this manuscript.

(iii) A brief introduction of the proposed modelling: It is noticeable that the quality-based calibration approach in **BBN-MLQc** exploits quality scores to normalize the match scores and liveness scores prior to integrating them into a BBN framework. An alternative normalization method is to apply the GAM scheme introduced in Chapter 2 and normalize match scores and liveness scores via a set of spline transformation functions (as shown in Eqn. 2.7). Similar to the gender attribute combined with match scores in the Eqn. 2.7, the quality scores after the categorization by Eqn. 3.9 and 3.10 are now used to divide matching

scenarios into multiple cohorts.

$$y = \mathbf{f}(x, q) = \alpha_0 + \sum_{j=1}^{p^m} \beta_j (x|q = j) + \gamma * d + \epsilon. \tag{5.1}$$

The additive model integrating match scores with quality scores relies on the combination of discrete quality levels from both two samples (as shown in Eqn. 5.1). The number of cohorts is denoted as p^m. Suppose the quality scores have two levels, such as high and low, then $p^m = 4$ and four cohorts are listed as: "high vs. high", "high vs. low", "low vs. high" and "low vs. low". Consequentially, four different transformation functions are trained to normalize the match scores before involving them into the BBN framework. Similar to the match scores, the liveness score of each sample is normalized by the corresponding quality scores via GAM (as shown in Eqn. 5.2) .

$$l_i^{norm} = \mathbf{f^l}(l_i, q_i) = \alpha_0^l + \sum_{j=1}^{p^l} \beta_j^l (l_i|q_i = j) + \gamma^l * d^l + \epsilon^l. \tag{5.2}$$

To evaluate the effectiveness of the proposed GAM based normalization of liveness scores and match scores, we generate the ROC curves of spoof detection accuracy and matching accuracy before and after the normalization (as shown in Figure 5.6 and 5.7).

5.3 Experimental Results

The spoof images in the LivDet 2011 database are fabricated using the consensual method which are supposed to be more difficult to detect than the non-consensual method. A consensual procedure [134] (i.e., with the consent and collaboration of the user) for fake fingerprint fabrication consists of the following steps: (a) a user is asked to press his finger against a soft material, such as wax, play-doh or plaster, to create a mould that holds a negative impression of the fingerprint; (b) a casting (fabrication) material such as liquid silicon, wax, gelatin, or clay is poured on the mould; and (c) after the liquid solidifies, the cast is lifted from the mould and is used as a fingerprint replica or fake finger. The casting (i.e.,

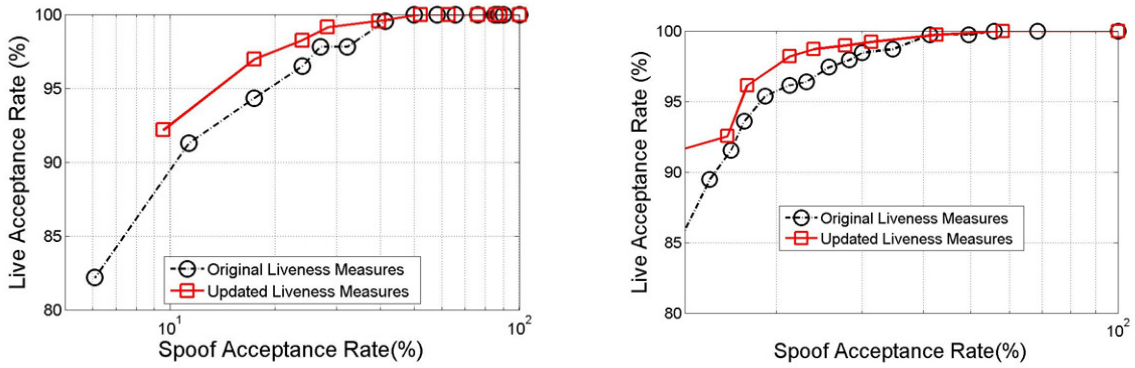


Figure 5.6: The performance of Spoof detection before and after updating the liveness scores via the GAM framework. The quality scores are used as the covariate of GAM. The samples are fabricated using a) silicone material and b) gelatin material.

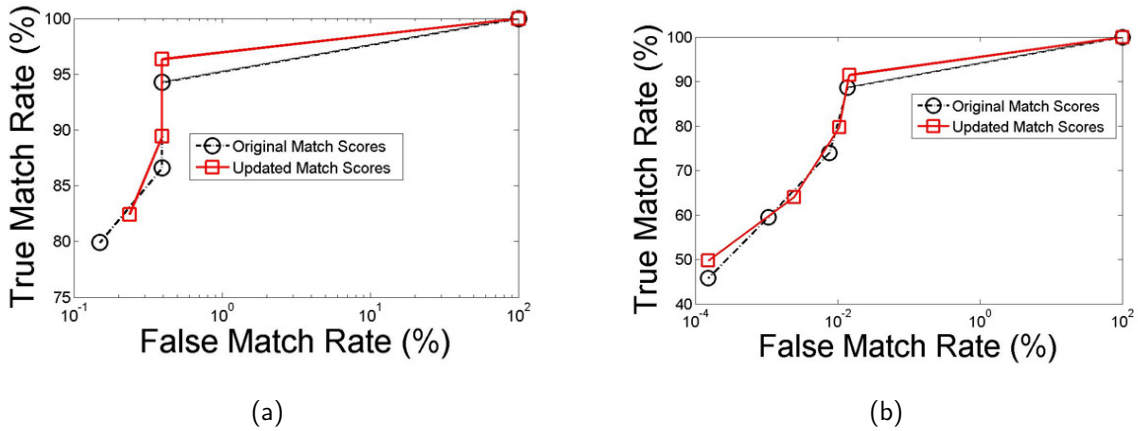


Figure 5.7: The performance of biometric matching system before and after updating the match scores via the GAM framework. The quality scores are used as the covariate of GAM. The samples are fabricated using a) silicone material and b) gelatin material.

Figure 5.8: Example of spoof images in the LivDet 2009 (a-b) and 2013 (c-d) databases fabricated using consensual and non-consensual methods, respectively. These spoofs are acquired using Biometrika sensor. Note that the spoof images are either of very low quality (a-b) or partial (c-d).

fabrication) material should have high elasticity and very low shrinkage to avoid reduction in volume as the cast cools and solidifies.

Figure 5.8 show the fake images acquired using Biometrika sensor in LivDet 2009 and 2013. It can be seen that the images are either of poor quality or partial owing to non-consensual approach to fake fingerprint fabrication (LivDet 2013).

The VeriFinger SDK[2] is used to generate match scores by matching all pairs of images within and across all subjects for live and spoof impressions. The quality of live and spoof impressions was obtained using the IQF freeware developed by MITRE[3]. The quality measure ranges between 0 and 100, with 0 being the lowest and 100 being the highest quality. Finally, fingerprint liveness was assessed using the recently proposed spoof detection algorithm based on local binary patterns (LBP) [85]. A two class Support Vector Machine (SVM) (implemented using LIBSVM package) was trained using LBP features extracted from live and fake images in the training set. The output score (probability estimate) of the SVM was then used as a liveness scores. The LBP-SVM spoof detector provides a better spoof detection accuracy over existing techniques as reported in [79].

The evaluation of the various BBN frameworks is conducted in terms of the spoof detec-

[2]`http://www.neurotechnology.com/vf_sdk.html`
[3]`http://www.mitre.org/tech/mtf/`

Table 5.1: The spoof detection accuracy of the proposed BBN-AD fusion scheme on the LivDet 2011 fingerprint database. The true detection rates (TDRs) and the false detection rates (FDRs) are compared with two fusion schemes introduced in Chapter 3. Additionally, the accuracy of the original spoof detector is provided as a baseline.

Various Frameworks	Biometrika		Italdata		Digital		Sagem	
	TDR at 1% FDR	TDR at 10% FDR	TDR at 1% FDR	TDR at 10% FDR	TDR at 1% FDR	TDR at 10% FDR	TDR at 1% FDR	TDR at 10% FDR
BBN-AD	78.2	91.1	77.1	88.8	77.1	91.1	82.6	95.8
BBN-MLQc	70.1	91.1	52.6	84.8	81.2	95.8	85.6	97.2
BBN-MLQ	62.3	91.1	49.8	83.2	77.1	95.8	84.1	97.2
Spoof Detector	42.0	80.0	22.9	66.9	61.9	88.0	72.1	92.5

tion accuracy and overall performance. We used scores from the training set (see Table 3.3) to train the fusion frameworks against spoof attacks and the scores in the testing part were used for the performance evaluation. Specifically, the match score (m) and a pair of quality values (q_1, q_2) as well as liveness scores (l_1, l_2) extracted from a pair of *training* images - the input and the template which can be live as well as fake. This observation vector (m, l_1, l_2, q_1, q_2) is mapped to one of eight classes: LLG, LLI, LSG, LSI, SLG, SLI, SSG, SSI (see Table 3.2) and used for training various fusion frameworks; BBN-MQ, BBN-ML, BBN-MLQ, BBN-MLQc and the GMM based direct modelling (referred to as DM-GMM) based scheme based on joint density estimation of match scores, quality and liveness scores. Comparative assessment of the various frameworks is done with BBN-M based only on the match scores and trained using all the eight possible events during the biometric system operation (see Table 3.2). Similarly, the observation (m, l_1, l_2, q_1, q_2) extracted from a pair of *testing* images - the input image and the template sample - is assigned to one of the eight classes and the error rates of these frameworks are evaluated. The detailed performance evaluation metrics used in this work were discussed in section 3.2.3.

5.4 Summary and Future Work

In this work, we proposed two Bayesian Belief Network (BBN) models that can effectively integrate liveness scores with quality scores and match score. The proposed BBN models have two different configurations distinguished on the basis of how the quality scores are

Table 5.2: The overall acceptance accuracy of the proposed BBN-AD fusion scheme on the LivDet 2011 fingerprint database. The genuine acceptance rate rate (TDRs) under different overall false acceptance rates (OFARs) are compared with two fusion schemes introduced in Chapter 3.

Various Frameworks	Biometrika		Italdata		Digital		Sagem	
	GAR [%] at OFAR = 1%	GAR [%] at OFAR = 5%	GAR [%] at OFAR = 1%	GAR [%] at OFAR = 5%	GAR [%] at OFAR = 1%	GAR [%] at OFAR = 5%	GAR [%] at OFAR = 1%	GAR [%] at OFAR = 5%
BBN-AD	81.2	88.3	79.2	89.0	82.6	90.2	80.2	88.6
BBN-MLQc	80.5	88.3	72.5	89.0	84.8	88.6	75.3	88.3
BBN-MLQ	75.6	85.2	72.1	88.6	83.0	87.6	72.3	87.2

incorporated. This study also compares the proposed BBN models with existing fusion frameworks against spoof attacks. Comprehensive experiments are conducted on the LivDet 2011 dataset. Results indicate that the proposed BBN-MLQ and BBN-MLQc methods consistently outperform existing fusion frameworks. Based on the experiments, the following conclusions can be drawn:

- **Causal relationship:** Fusion frameworks that model the appropriate relationship between the considered variables, such as the influence of the quality on liveness scores, obtain better performance.

- **Benefits of quality:** Incorporating image quality is beneficial in the fusion framework (BBN-MLQ and BBN-MLQc). This is because quality scores can take into account the material-specific characteristics of spoof fabrication materials. Further, the models incorporating quality also have benefits (better performance) when evaluated on novel spoof fabrication materials [102].

- **The role of quality:** These quality scores can be incorporated as features (as in BBN-MLQ) or used as a normalization parameter (as in BBN-MLQc). Experimental results suggest the efficacy of quality when used as a normalization parameter rather than a feature, since the latter makes the Bayesian Belief Network more complicated to be interpreted and calculated.

As a part of future work, the following experiments and analysis will be done:

- **The role of latent variables**: The consistently better performance of BBN-MLQc over existing frameworks show the efficacy of quality-based clusters in adapting the liveness scores and match scores against the sample quality. Even for a single acquisition device, clusters of quality values (Q_{ci}) can be obtained corresponding to image resolution, ridge and valley clarity, noise level, and spoof fabrication materials. Further, quality and liveness scores are also influence by the acquisition sensor used. As a part of future work, these latent variables i.e., quality clusters and sensor information will be incorporated in the BBN models. Further, the role of these latent variables will be analyzed for novel sensors and fabrication materials.

- **Semi-supervised learning in BBN models**: Our experimental results suggest that performance of all the BBN models drops across materials. Hence, automatically adapting these BBN models to novel spoof materials is another research avenue. In other words, models will be incorporated with the learning ability to automatically detect and adapt themselves to spoof samples generated using novel materials.

- **Effect of the baseline anti-spoofing algorithms:** Continuous efforts are being directed towards developing spoof detection schemes which offer lower error rates, also evident by three spoof detection competitions (LivDet) conducted between 2009 and 2011. The performance of existing and proposed fusion frameworks will be evaluated on incorporating liveness scores obtained using novel spoof detection schemes and comparative assessment will be drawn with respect to existing ones.

- **Cross database matching:** For the real time applications, these learning-based fusion frameworks against spoof attacks should be able to generalize well on cross database matching i.e., training using one database (say LivDet 2009) and testing using other (say LivDet 2011). As a part of future work, we will develop more robust models that can generalize well across databases.

CHAPTER 6

SUMMARY

While the primary purpose of a biometric recognition system is to ensure reliable and accurate human recognition, ancillary information may be available in most biometric application scenarios. They may be collected from official documents (such as demographic information of a user) or deduced from the collected biometric data itself (such as the image quality of a biometric sample, and the spoof measures). This raises the research question of whether this ancillary information can be effectively combined with biometric match scores to improve the recognition accuracy of a system. This dissertation attempts to investigate this question and addresses several challenging issues at the same time. A summary of the contribution is listed below:

- We design a Generalized Additive Model (GAM) that learns an optimal transformation function to normalize the match scores according to demographic attributes prior to fusion. The empirical analysis shows that the resulting framework can be used to predict in advance if exploiting match scores with certain demographic attributes is beneficial in the context of a specific biometric matcher. Experimental results conducted on multiple databases indicate that the resulting framework proves to be effective even in the situation where the attributes are unreliable or incorrect to some extent. These advantages of GAM mitigate the concerns associated with issues of "lack of distinctiveness" and "lack of reliability" encountered when integrating ancillary information.

- We design a Bayesian Belief Network (BBN) to appropriately model the relationship between biometric scores and ancillary factors, and exploit the ensuing structure in a fusion framework. As a graphical model, the BBN can utilize causal assumptions to reduce the computational complexity of estimating the joint probability of a fusion framework with multiple covariates. More important, by assigning different weights

to match scores and other ancillary factors (e.g. spoof scores) from a biometric recognition perspective, the overall matching accuracy after combining ancillary factors is consistently better than other typical classifiers.

- We design an ensemble of one-class classifiers to improve the classification performance in the context of biometric anti-spoofing. We adopt a One Class Support Vector Machine (OC-SVM) approach that predominantly uses training samples from only a single class, i.e., the live class, to generate a hypersphere that encompasses most of the live samples. The goal is to learn the concept of a "live" biometric sample. The boundary of the hypersphere is refined using a small number of spoof samples. The proposed method uses an ensemble of such OC-SVMs based on different feature sets. Experimental results show the advantages of the proposed ensemble of OC-SVMs for detecting spoofs generated from previously "unseen" materials, or collected via previously "unknown" sensors.

- We design a general fusion framework to combine ancillary information via the aforementioned GAM and BBN schemes. We utilize the quality measure of biometric samples as an example to test the scalability of the proposed fusion framework. Experimental results show that a consistent performance improvement is obtained using the proposed framework, and a significant accuracy benefit (2.5% to 10.5%) is observed compared to other commonly used direct modeling frameworks.

In conducting the studies on a general fusion framework as proposed in this dissertation, a number of areas for future work can be explored by researchers in ancillary information extraction. One obvious direction for future work is to incorporate extensive ancillary information via the proposed fusion framework, such as the confidence of age estimation algorithms, the uncertainty measurements from anti-spoofing algorithms, and so on. Similar to the ancillary attributes discussed in this thesis, many of these attributes are reliant

of a single biometric sample rather than a pair of samples. Therefore, they can be simply included as latent variables via the GAM architecture.

Moreover, the proposed framework can be extended by combining match scores from multiple biometric modalities. Along with the usage of additional biometric modalities, it is possible to independently extract extra ancillary information from each of them, which can lead to significant performance improvement. However, directly applying the proposed framework to a large-scale database may result in degraded performance. The possible reason is that both the GAM and BBN models depend on assumptions of relationships between ancillary attributes and match scores. In other words, when these relationships become complex, a validation of the assumptions is required before implementing the proposed framework.

The usefulness of the proposed one-class classification approach can be further advanced with the development of feature engineering, such as deep neural network based feature selection. Recent research has pointed out that the architecture of deep neural networks is a promising technique for learning robust features. It is possible to further improve the anti-spoofing accuracy by training an unsupervised deep neural network and extracting generic underlying features, and then applying the proposed ensemble of one-class SVMs on the feature sets learned from the networks. Alternately, convolutional auto-encoders can be used to formulate this as a one-class problem. Because the proposed OCC approach is scalable and computationally efficient, it is a promising framework to exploiting more robust and sophisticated features and eventually addressing the performance degradation issue under cross-database and cross-attack scenarios.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] Abaza, Ayman & Arun Ross. 2009. Quality based rank-level fusion in multibiometric systems. *IEEE International Conference on Biometrics: Theory, Applications, and Systems (BTAS)* 1–6.

[2] Abhyankar, A. & S. Schuckers. 2009. Integrating a wavelet based perspiration liveness check with fingerprint recognition. *Pattern Recognition* 42. 452–464.

[3] Ahlberg, J Harold, Edwin Norman Nilson & Joseph Leonard Walsh. 2016. *The theory of splines and their applications: Mathematics in science and engineering: A series of monographs and textbooks*, vol. 38. Elsevier.

[4] Arashloo, Shervin Rahimzadeh & Josef Kittler. 2014. Class-specific kernel fusion of multiple descriptors for face verification using multiscale binarised statistical image features. *IEEE Transactions on Information Forensics and Security (TIFS)* 9(12). 2100–2109.

[5] Arora, Sunpreet S, Kai Cao, Anil K Jain & Nicholas G Paulter. 2014. 3d fingerprint phantoms. In *22nd international conference on pattern recognition (icpr)*, 684–689.

[6] Bekios-Calfa, Juan, José M Buenaposada & Luis Baumela. 2014. Robust gender recognition by exploiting facial attributes dependencies. *Pattern Recognition Letters* 36. 228–234.

[7] Benediktsson, Jon Atli & Philip H Swain. 1992. Consensus theoretic classification methods. *IEEE transactions on Systems, Man, and Cybernetics* 22(4). 688–704.

[8] Bobeldyk, Denton & Arun Ross. 2016. Iris or periocular? exploring sex prediction from near infrared ocular images. In *Ieee international conference of the biometrics special interest group (biosig)*, 1–7.

[9] Boulle, Marc. 2006. MODL: A Bayes optimal discretization method for continuous attributes. *Machine Learning* 65(1). 131–165.

[10] Burges, Christopher JC. 1998. A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery* 2(2). 121–167.

[11] Carroll, Raymond J & David Ruppert. 1988. *Transformation and weighting in regression*, vol. 30. CRC Press.

[12] Castrillón-Santana, M, J Lorenzo-Navarro & E Ramón-Balmaseda. 2017. Descriptors and regions of interest fusion for in-and cross-database gender classification in the wild. *Image and Vision Computing* 57. 15–24.

[13] Chang, Chih Chung & Chih Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* 2. 1–27. Software available at `http://www.csie.ntu.edu.tw/cjlin/libsvm`.

[14] Chatzis, Vassilios, Adrian G Bors & Ioannis Pitas. 1999. Multimodal decision-level fusion for person authentication. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans* 29(6). 674–680.

[15] Chen, Yi, Sarat C Dass & Anil K Jain. 2005. Fingerprint quality indices for predicting authentication performance. In *Audio-and video-based biometric person authentication*, 160–170. Springer.

[16] Chingovska, I., A. Anjos & S. Marcel. 2013. Anti-spoofing in action: joint operation with a verification system. In *Ieee conference on computer vision and pattern recognition workshops (cvprw)*, 1–8. USA.

[17] Coli, P., G.L. Marcialis & F. Roli. 2007. Power spectrum-based fingerprint vitality detection. In *Ieee international workshop on automatic identification advanced technologies autoid*, 169–173. Alghero, Italy.

[18] Coull, Brent A, David Ruppert & MP Wand. 2001. Simple incorporation of interactions into additive models. *Biometrics* 57(2). 539–545.

[19] Currah, Paisley & Tara Mulqueen. 2011. Securitizing gender: Identity, biometrics, and transgender bodies at the airport. *Social Research* 78(2). 557–582.

[20] Dantcheva, Antitza, Petros Elia & Arun Ross. 2016. What else does your biometric data reveal? a survey on soft biometrics. *IEEE Transactions on Information Forensics and Security (TIFS)* 11(3). 441–467.

[21] De Comité, Francesco, François Denis, Rémi Gilleron & Fabien Letouzey. 1999. Positive and unlabeled examples help learning. In *International conference on algorithmic learning theory*, 219–230. Springer.

[22] Denis, François, Rémi Gilleron & Fabien Letouzey. 2005. Learning from positive and unlabeled examples. *Theoretical Computer Science* 348(1). 70–83.

[23] Ding, Yaohui, Ajita Rattani & Arun Ross. 2016. Bayesian belief models for integrating match scores with liveness and quality measures in a fingerprint verification system. In *Ieee conference on international conference on biometrics (icb)*, 1–8.

[24] Duda, Richard O & Peter E Hart. 1973. *Pattern elesslfication and scene analysis*. Wiley.

[25] Duda, Richard O, Peter E Hart & David G Stork. 2012. *Pattern classification*. John Wiley & Sons.

[26] Eilers, Paul HC & Brian D Marx. 1996. Flexible smoothing with b-splines and penalties. *Statistical Science* 89–102.

[27] El Shafey, Laurent, Elie Khoury & Sébastien Marcel. 2014. Audio-visual gender recognition in uncontrolled environment using variability modeling techniques. In *Ieee international joint conference on biometrics (ijcb)*, 1–8.

[28] Espinoza, M. & C. Champod. 2011. Using the number of pores on fingerprint images to detect spoofing attacks. In *Intl. conf. on hand-based biometrics*, 1–5. Hong Kong, China.

[29] Fierrez-Aguilar, J., J. Ortega-Garcia, J. Gonzalez-Rodriguez & J. Bigun. 2004. Kernel-based multimodal biometric verification using quality signals. In *Spie workshop on biometric technology for human identification*, 544–554.

[30] Frick, M, Shimon K Modi, S Elliott & Eric P Kukula. 2008. Impact of gender on fingerprint recognition systems. In *International conference on information technology and applications, cairns, australia*, 717–721.

[31] Fu, Siyao, Haibo He & Zeng-Guang Hou. 2014. Learning race from face: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36(12). 2483–2509.

[32] Fu, Yun, Guodong Guo & Thomas S Huang. 2010. Age synthesis and estimation via faces: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32(11). 1955–1976.

[33] Galbally, Javier, Fernando Alonso-Fernandez, Julian Fierrez & Javier Ortega-Garcia. 2012. A high performance fingerprint liveness detection method based on quality related features. *Future Generation Computer Systems* 28(1). 311–321.

[34] Ghiani, L., P. Denti & G. L. Marcialis. 2012. Experimental results on fingerprint liveness detection. In *Ieee international conference on articulated motion and deformable objects*, 210–218. Spain.

[35] Ghiani, L., A. Hadid, G.L. Marcialis & F. Roli. 2013. Fingerprint liveness detection using binarized statistical image features. In *Ieee international conference on biometrics: Theory, applications and systems (btas)*, 1–6.

[36] Ghiani, L., G.L. Marcialis & F. Roli. 2012. Fingerprint liveness detection by local phase quantization. In *International conference on pattern recognition (icpr)*, 537–540.

[37] Ghiani, Luca, David Yambay, Valerio Mura, Simona Tocco, Gian Luca Marcialis, Fabio Roli & Stephanie Schuckcrs. 2013. Livdet 2013 fingerprint liveness detection competition. In *International conference on biometrics (icb)*, 1–6.

[38] Gnanasivam, P & Dr S Muttan. 2012. Fingerprint gender classification using wavelet transform and singular value decomposition. *arXiv preprint arXiv:1205.6745* .

[39] Grother, Patrick & Elham Tabassi. 2007. Performance of biometric quality measures. *IEEE transactions on pattern analysis and machine intelligence* 29(4). 531–43.

[40] Grother, Patrick J, George W Quinn & P Jonathon Phillips. 2010. Report on the evaluation of 2d still-image face recognition algorithms. *NIST Interagency Report* 7709. 106.

[41] Han, H., A. K. Jain, F. Wang, S. Shan & X. Chen. 2017. Heterogeneous face attribute estimation: A deep multi-task learning approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence* .

[42] Hastie, Trevor, Robert Tibshirani & Jerome Friedman. 2002. The elements of statistical learning: Data mining, inference, and prediction. *Biometrics* .

[43] Hastie, Trevor J & Robert J Tibshirani. 1990. *Generalized additive models*, vol. 43. CRC Press.

[44] Huang, Gary B & Erik Learned-Miller. 2014. Labeled faces in the wild: Updates and new reporting procedures. *Department of Computer Science, University of Massachusetts Amherst, USA, Technique Report* 14–003.

[45] Jain, Anil K, Sarat C Dass & Karthik Nandakumar. 2004. Can soft biometric traits assist user recognition? In *Defense and security*, 561–572. International Society for Optics and Photonics.

[46] Jain, Anil K, Karthik Nandakumar & Arun Ross. 2016. 50 years of biometric research: Accomplishments, challenges, and opportunities. *Pattern Recognition Letters* 79. 80–105.

[47] Jain, Anil K & Arun Ross. 2004. Multibiometric systems. *Communications of the ACM* 47(1). 34–40.

[48] Jain, Anil K, Arun Ross & Salil Prabhakar. 2004. An introduction to biometric recognition. *IEEE Transactions on Circuits and Systems for Video Technology* 14(1). 4–20.

[49] Jia, Sen & Nello Cristianini. 2015. Learning to classify gender from four million images. *Pattern Recognition Letters* 58. 35–41.

[50] Johnson, Peter A, Paulo Lopez-Meyer, Nadezhda Sazonova, F Hua & S Schuckers. 2010. Quality in face and iris research ensemble (q-fire). In *Ieee international conference on biometrics: Theory applications and systems (btas)*, 1–6.

[51] Kittler, Josef, Mohamad Hatef, Robert PW Duin & Jiri Matas. 1998. On combining classifiers. *IEEE transactions on pattern analysis and machine intelligence* 20(3). 226–239.

[52] Klare, Brendan F, Mark J Burge, Joshua C Klontz, Richard W Vorder Bruegge & Anil K Jain. 2012. Face recognition performance: Role of demographic information. *IEEE Transactions on Information Forensics and Security (TIFS)* 7(6). 1789–1801.

[53] Kryszczuk, K., J. Richiardi & A. Drygajlo. 2009. Impact of combining quality measures on biometric sample matching. In *Ieee international conference on biometrics: Theory, applications, and systems (btas)*, 1–6.

[54] Kumar, Neeraj, Alexander C Berg, Peter N Belhumeur & Shree K Nayar. 2009. Attribute and simile classifiers for face verification. In *Ieee international conference on computer vision (iccv)*, 365–372.

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

[56] Lagree, Stephen & Kevin W Bowyer. 2011. Predicting ethnicity and gender from iris texture. In *Ieee international conference on technologies for homeland security (hst)*, 440–445.

[57] Langley, Pat & Stephanie Sage. 1994. Induction of selective bayesian classifiers. In *The10th international conference on uncertainty in artificial intelligence*, 399–406. Morgan Kaufmann Publishers Inc.

[58] Li, Haoxiang & Gang Hua. 2015. Hierarchical-pep model for real-world face recognition. In *Ieee conference on computer vision and pattern recognition (cvpr)*, 4055–4064.

[59] Li, Haoxiang, Gang Hua, Xiaohui Shen, Zhe Lin & Jonathan Brandt. 2014. Eigen-pep for video face recognition. In *Asian conference on computer vision*, 17–33. Springer.

[60] Li, Xiong, Xu Zhao, Yun Fu & Yuncai Liu. 2010. Bimodal gender recognition from face and fingerprint. In *Ieee conference on computer vision and pattern recognition (cvpr)*, 2590–2597.

[61] Littlestone, Nick & Manfred K Warmuth. 1994. The weighted majority algorithm. *Information and computation* 108(2). 212–261.

[62] Liu, Ziwei, Ping Luo, Xiaogang Wang & Xiaoou Tang. 2015. Deep learning face attributes in the wild. In *Ieee international conference on computer vision (iccv)*, 3730–3738.

[63] Lu, Xiaoguang & Anil K Jain. 2004. Ethnicity identification from face images. In *Spie defense, security, and sensing*, 114–123. International Society for Optics and Photonics.

[64] Manevitz, Larry M & Malik Yousef. 2000. Document classification on neural networks using only positive examples (poster session). In *The 23rd annual international acm sigir conference on research and development in information retrieval*, 304–306. ACM.

[65] Manevitz, Larry M & Malik Yousef. 2001. One-class svms for document classification. *Journal of Machine Learning Research* 2(Dec). 139–154.

[66] Marasco, E. & A. Ross. 2014. A survey on anti-spoofing schemes for fingerprint recognition systems. *ACM Computing Surveys* 47(2). 1–36.

[67] Marasco, E. & C. Sansone. 2012. Combining perspiration- and morphology-based static features for fingerprint liveness detection. *Pattern Recognition Letters* 33. 1148–1156.

[68] Marasco, Emanuela, Yaohui Ding & Arun Ross. 2012. Combining match scores with liveness values in a fingerprint verification system. *Biometrics: Theory, Applications and Systems (BTAS)* .

[69] Marasco, Emanuela, Luca Lugini & Bojan Cukic. 2014. Exploiting quality and texture features to estimate age and gender from fingerprints. *SPIE Defense+ Security* 90750F–90750F.

[70] Marasco, Emanuela & Arun Ross. 2015. A survey on antispoofing schemes for fingerprint recognition systems. *ACM Computing Surveys (CSUR)* 47(2). 28.

[71] Marcel, Sébastien, Mark S Nixon & Stan Z Li. 2014. *Handbook of biometric anti-spoofing*. Springer.

[72] Marcialis, G.L., F. Roli & A. Tidu. 2010. Analysis of fingerprint pores for vitality detection. In *International conference on pattern recognition (icpr)*, 1289–1292.

[73] Matsumoto, T., H. Matsumoto, K. Yamada & S. Hoshino. 2002. Impact of Artificial Gummy Fingers on Fingerprint Systems. *SPIE* .

[74] Maurer, Donald E. & John P. Baker. 2008. Fusing multimodal biometrics with quality estimates via a Bayesian belief network. *Pattern Recognition* 41(3). 821–832.

[75] Menotti, David, Giovani Chiachia, Allan Pinto, William Robson Schwartz, Helio Pedrini, Alexandre Xavier Falcao & Anderson Rocha. 2015. Deep representations for iris, face, and fingerprint spoofing detection. *IEEE Transactions on Information Forensics and Security* 10(4). 864–879.

[76] Moon, Y.S., J.S. Chen, K.C. Chan, K. So & K.S. Woo. 2005. Wavelet based fingerprint liveness detection. *Electronic Letters* 41. 1112–1113.

[77] Moses, Kenneth R, P Higgins, M McCabe, S Prabhakar & S Swann. 2011. Automated fingerprint identification system (afis). *Scientific Working Group on Friction Ridge Analysis Study and Technology and National institute of Justice (eds.) SWGFAST-The fingerprint sourcebook* 1–33.

[78] Moya, Mary M, Mark W Koch & Larry D Hostetler. 1993. One-class classifier networks for target recognition applications. Tech. rep. Sandia National Labs., Albuquerque, NM (United States).

[79] Mura, Valerio, Luca Ghiani, Gian Luca Marcialis, Fabio Roli, David A Yambay & Stephanie A Schuckers. 2015. Fingerprint liveness detection competition. In *Ieee 7th international conference on biometrics: Theory, applications and systems (btas)*, 1–6.

[80] Nandakumar, K., Yi Chen, Anil K. Jain & Sarat C. Dass. 2006. Quality based score level fusion in multibiometric systems. In *International conference on pattern recognition (icpr)*, 473–476.

[81] Nandakumar, Karthik. 2008. *Multibiometric systems: Fusion strategies and template security*. ProQuest.

[82] Ng, Choon Boon, Yong Haur Tay & Bok-Min Goi. 2012. Recognizing human gender in computer vision: a survey. In *Pacific rim international conference on artificial intelligence*, 335–346. Springer.

[83] Nikam, S.B. & S. Aggarwal. 2008. Local binary pattern and wavelet-based spoof fingerprint detection. *Intl. Journal of Biometrics* 1(2). 141–159.

[84] Nikam, S.B. & S. Aggarwal. 2008. Wavelet energy signature and glcm features-based fingerprint anti-spoofing. In *Ieee international conference on wavelet analysis and pattern recognition*, 717 –723. Hong Kong, China.

[85] Nikam, Shankar Bhausaheb & Suneeta Agarwal. 2008. Local binary pattern and wavelet-based spoof fingerprint detection. *International Journal of Biometrics* 1(2). 141–159.

[86] Nill, NB. 2007. IQF (Image Quality of Fingerprint) Software Application. `http://wwwsrv2.mitre.org/work/tech\_papers/tech\_papers\_07/07\_0580/07\_0580.pdf`.

[87] Nithin, MD, BM Balaraj, B Manjunatha & Shashidhar C Mestri. 2009. Study of fingerprint classification and their gender distribution among south indian population. *Journal of Forensic and Legal Medicine* 16(8). 460–463.

[88] Nixon, Kristin A, Robert K Rowe, Jeffrey Allen, Steve Corcoran, Lu Fang, David Gabel, Damien Gonzales, Robert Harbour, Sarah Love, Rick McCaskill et al. 2004. Novel spectroscopy-based technology for biometric and liveness verification. In *Defense and security*, 287–295. International Society for Optics and Photonics.

[89] Nixon, Kristin Adair, Valerio Aimale & Robert K Rowe. 2008. Spoof detection schemes. In *Handbook of biometrics*, 403–423. Springer.

[90] Nixon, Mark. 1985. Eye spacing measurement for facial recognition. In *29th annual technical symposium*, 279–285. International Society for Optics and Photonics.

[91] Phillips, P Jonathon, J Ross Beveridge, Bruce A Draper, Geof Givens, Alice J O'Toole, David S Bolme, Joseph Dunlop, Yui Man Lui, Hassan Sahibzada & Samuel Weimer. 2011. An introduction to the good, the bad, & the ugly face recognition challenge problem. In *Ieee international conference on automatic face & gesture recognition and workshops*, 346–353.

[92] Poh, N. & J. Kittler. 2008. A family of methods for quality-based multimodal biometric fusion using generative classi fiers. In *Ieee international conference on control, automation, robotics and vision (icarcv)*, 1162–1167.

[93] Poh, N., J. Kittler & T. Bourlai. 2010. Quality-based score normalization with device qualitative information for multimodal biometric fusion. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans* 40(3). 539–554.

[94] Poh, Norman, Thirimachos Bourlai, Josef Kittler, Lorene Allano, Fernando Alonso-Fernandez, Onkar Ambekar, John Baker, Bernadette Dorizzi, Omolara Fatukasi, Julian Fierrez et al. 2009. Benchmarking quality-dependent and cost-sensitive score-level multimodal biometric fusion algorithms. *IEEE Transactions on Information Forensics and Security* 4(4). 849–866.

[95] Poh, Norman & Josef Kittler. 2012. A unified framework for biometric expert fusion incorporating quality measures. *IEEE Transactions on pattern analysis and machine intelligence* 34(1). 3–18.

[96] Quinlan, J Ross. 2014. *C4. 5: programs for machine learning*. Elsevier.

[97] Ramanathan, Venkatesh & Harry Wechsler. 2010. Robust human authentication using appearance and holistic anthropometric features. *Pattern Recognition Letters* 31(15). 2425–2435.

[98] Rattani, A. & N. Poh. 2013. Biometric system design under zero and non-zero effort attacks. In *Ieee international conference on biometrics*, 1–8. Madrid, Spain.

[99] Rattani, Ajita, Cunjian Chen & Arun Ross. Evaluation of texture descriptors for automated gender estimation from fingerprints. In *2014 european conference on computer vision (eccv) workshops*, 764–777. Springer.

[100] Rattani, Ajita, Norman Poh & Arun Ross. 2013. A bayesian approach for modeling sensor influence on quality, liveness and match score values in fingerprint verification. In *Ieee international workshop on information forensics and security (wifs)*, 37–42.

[101] Rattani, Ajita & Arun Ross. 2014. Automatic adaptation of fingerprint liveness detector to new spoof materials. In *International joint conference on biometrics (ijcb)*, 1–8.

[102] Rattani, Ajita, Walter J Scheirer & Arun Ross. 2015. Open set fingerprint spoof detection across novel fabrication materials. *IEEE Transactions on Information Forensics and Security (TIFS)* 10(11). 2447–2460.

[103] Reddy, P Venkata, Ajay Kumar, SMK Rahman & Tanvir Singh Mundra. 2007. A new method for fingerprint antispoofing using pulse oxiometry. In *Biometrics: Theory, applications, and systems, 2007. btas 2007. first ieee international conference on*, 1–6. IEEE.

[104] Reddy, P Venkata, Ajay Kumar, SMK Rahman & Tanvir Singh Mundra. 2008. A new antispoofing approach for biometric devices. *IEEE Transactions on Biomedical Circuits and Systems* 2(4). 328–337.

[105] de Ridder, Dick, D Tax & R Duin. 1998. An experimental comparison of one-class classification methods. In *The 4th annual conference of the advacned school for computing and imaging, delft*, .

[106] Robinson, George K. 1991. That blup is a good thing: the estimation of random effects. *Statistical Science* 15–32.

[107] Ross, Arun & Rohin Govindarajan. 2005. Feature level fusion using hand and face biometrics. In *Spie conference on biometric technology for human identification ii*, vol. 5779, 196–204.

[108] Ross, Arun A, Karthik Nandakumar & Anil K Jain. 2006. *Handbook of multibiometrics*, vol. 6. Springer.

[109] Sagonas, Christos, Yannis Panagakis, Stefanos Zafeiriou & Maja Pantic. 2017. Robust statistical frontalization of human and animal faces. *International journal of computer vision* 122(2). 270–291.

[110] Scheirer, Walter J, Neeraj Kumar, Karl Ricanek, Peter N Belhumeur & Terrance E Boult. 2011. Fusing with context: a bayesian approach to combining descriptive attributes. In *Ieee international joint conference on biometrics (ijcb)*, 1–8.

[111] Schölkopf, Bernhard, R Williamson, Alex Smola & John Shawe-Taylor. 1999. Sv estimation of a distribution's support. *Advances in neural information processing systems* 12.

[112] Schölkopf, Bernhard, Robert C Williamson, Alexander J Smola, John Shawe-Taylor, John C Platt et al. 1999. Support vector method for novelty detection. In *Nips*, vol. 12, 582–588. Citeseer.

[113] Schuckers, SAC. 2002. Spoofing and anti-spoofing measures. *Information Security Technical Report* 7(4). 56–62.

[114] Sgroi, Amanda, Kevin W Bowyer & Patrick J Flynn. 2013. The prediction of old and young subjects from iris texture. In *Ieee international conference on biometrics (icb)*, 1–5.

[115] Shan, Caifeng. 2012. Learning local binary patterns for gender classification on real-world face images. *Pattern Recognition Letters* 33(4). 431–437.

[116] Shively, Thomas S, Robert Kohn & Sally Wood. 1999. Variable selection and function estimation in additive nonparametric regression using a data-based prior. *Journal of the American Statistical Association* 94(447). 777–794.

[117] Sousedik, C. & C. Busch. 2014. Presentation attack detection methods for fingerprint recognition systems: a survey. *IET Biometrics* .

[118] Sudhish, Prem Sewak, Anil K Jain & Kai Cao. 2016. Adaptive fusion of biometric and biographic information for identity de-duplication. *Pattern Recognition Letters* 84. 199–207.

[119] Sun, Yunlian, Man Zhang, Zhenan Sun & Tieniu Tan. 2017. Demographic analysis from biometric data: Achievements, challenges, and new frontiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)* .

[120] Tan, B. & S. Schuckers. 2006. Liveness detection for fingerprint scanners based on the statistics of wavelet signal processing. In *Workshop on biometrics in computer vision and pattern recognition*, 26–26.

[121] Tan, Bozhao & Stephanie Schuckers. 2010. Spoofing protection for fingerprint scanner by fusing ridge signal and valley noise. *Pattern Recognition* 43(8). 2845–2857.

[122] Tax, David MJ. 2001. *One-class classification.* PhD Thesis, Delft University of Technology.

[123] Tax, David MJ & Robert PW Duin. 1999. Data domain description using support vectors. In *Esann*, vol. 99, 251–256.

[124] Tax, David MJ & Robert PW Duin. 1999. Support vector domain description. *Pattern recognition letters* 20(11). 1191–1199.

[125] Tax, David MJ & Robert PW Duin. 2001. Uniform object generation for optimizing one-class classifiers. *Journal of Machine Learning Research* 2(Dec). 155–173.

[126] Tax, David MJ & Robert PW Duin. 2004. Support vector data description. *Machine learning* 54(1). 45–66.

[127] Toh, Kar-Ann. 2004. Personalized learning and decision for multimodal biometrics. In *Ieee conference on cybernetics and intelligent systems*, vol. 2, 1112–1117.

[128] Toth, Bori. 2005. Biometric liveness detection. *Information Security Bulletin* 10(8). 291–297.

[129] Toth, Bori. 2005. Introduction to Biometric Liveness Detection. *Information Security* 10(October). 291–298.

[130] Veeramachaneni, Kalyan, Lisa Osadciw, Arun Ross & Nisha Srinivas. 2008. Decision-level fusion strategies for correlated biometric classifiers. In *Ieee computer society conference on computer vision and pattern recognition workshops (cvprw)*, 1–6.

[131] Wayman, James, Anil Jain, Davide Maltoni & Dario Maio. 2005. *An introduction to biometric authentication systems.* Springer.

[132] Wein, L. & M. Baveja. 2005. Using fingerprint image quality to improve the identification performance of the u.s. visit program. In *the national academy of sciences*, vol. 102, 7772–7775.

[133] Wood, Simon N. 2017. *Generalized additive models: an introduction with r.* CRC press.

[134] Yambay, D., L. Ghiani, P. Denti, G. L. Marcialis, F. Roli & S. Schuckers. 2012. LivDet 2011 - fingerprint liveness detection competition. In *Ieee international conference on biometrics (icb)*, 208–215. Delhi, India.

[135] Yambay, David, Luca Ghiani, Paolo Denti, Gian Luca Marcialis, Fabio Roli & S Schuckers. 2012. Livdet 2011 - fingerprint liveness detection competition. In *5th iapr international conference on biometrics (icb)*, 208–215.

[136] Yanushkevich, Svetlana N. 2011. Belief network design for biometric systems. In *Computational intelligence in biometrics and identity management (cibim), 2011 ieee workshop on*, 1–10. IEEE.

[137] Yi, Dong, Zhen Lei & Stan Z Li. 2014. Age estimation by multi-scale convolutional network. In *Asian conference on computer vision (accv)*, 144–158. Springer.

[138] Yu, Hwanjo. 2005. Single-class classification with mapping convergence. *Machine Learning* 61(1-3). 49–69.

[139] Yu, Hwanjo, Jiawei Han & Kevin Chen-Chuan Chang. 2002. Pebl: positive example based learning for web page classification using svm. In *The 8th acm sigkdd international conference on knowledge discovery and data mining*, 239–248. ACM.

[140] Zhang, Lin, Zhiqiang Zhou & Hongyu Li. 2012. Binary gabor pattern: An efficient and robust descriptor for texture classification. In *19th international conference on image processing (icip)*, 81–84.