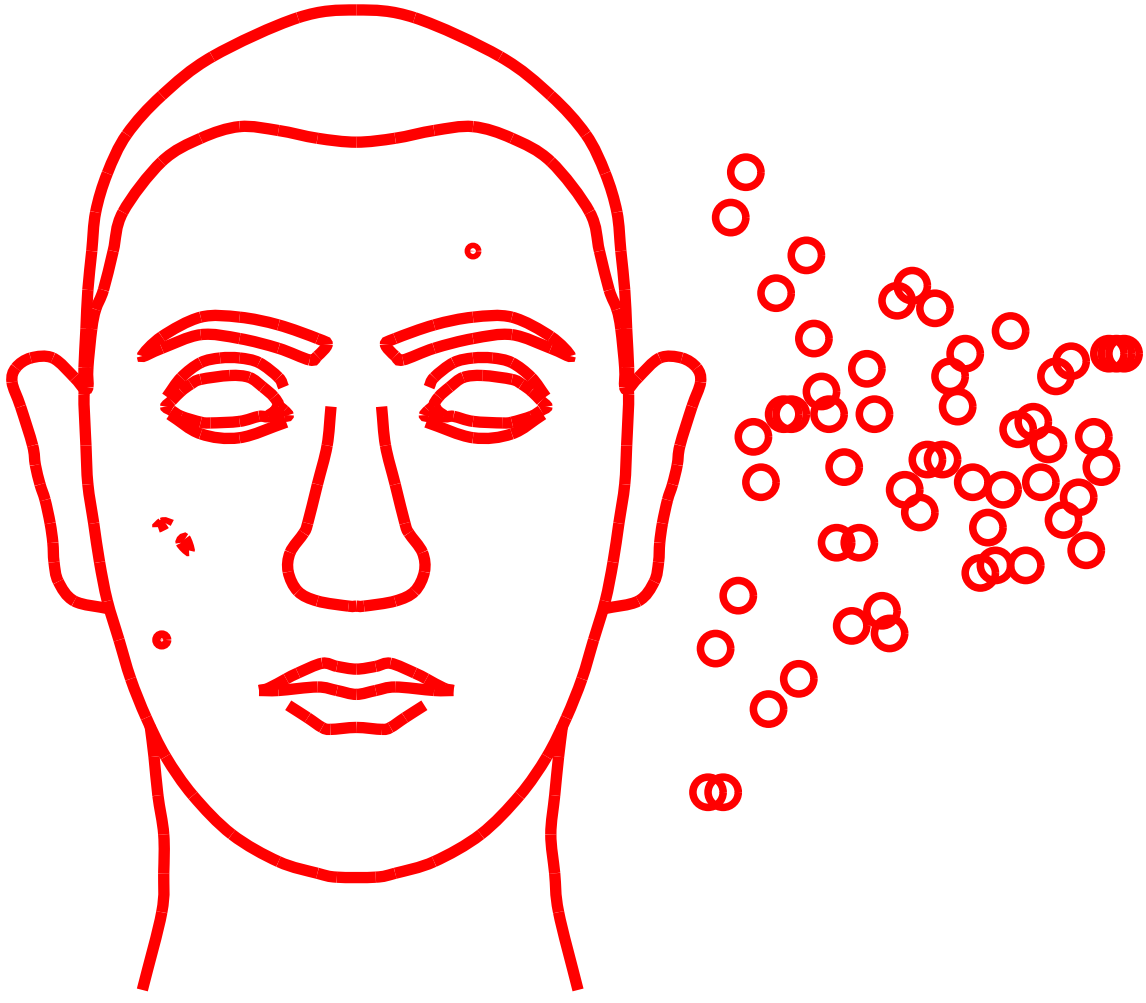
Forensic Face Recognition

From characteristic descriptors to strength of evidence



Chris Zeinstra

Forensic Face Recognition

From characteristic descriptors to strength of evidence

Chris Zeinstra

Composition of the Graduation Commitee:

prof.dr. P.M.G. Apers	University Twente, EWI
prof.dr.ir. R.N.J.Veldhuis	University Twente, EWI
dr.ir. L.J. Spreeuwers	University Twente, EWI
dr. A.C.C. Ruifrok	Netherlands Forensic Institute
prof.dr. D. Meuwly	University Twente, EWI
prof.dr.ir. G.J.A. Fox	University Twente, BMS
prof.dr. M. Tistarelli	Universita' degli Studi di Sassari, Italy
prof.dr. C.A.J Klaassen	University of Amsterdam

The doctoral research of C.G. Zeinstra was funded by the Netherlands Organisation for Scientific Research (NWO) Project Forensic Face Recognition, 727.011.008.

CTIT Ph.D. Thesis Series No. 17-439
Centre for Telematics and Information Technology
P.O. Box 217, 7500 AE
Enschede, The Netherlands

ISBN: 978-90-365-4375-0
ISSN: 1381-3617
DOI: 10.3990/1.9789036543750
URL: https://doi.org/10.3990/1.9789036543750

FORENSIC FACE RECOGNITION
FROM CHARACTERISTIC DESCRIPTORS TO STRENGTH OF EVIDENCE

DISSERTATION

to obtain
the degree of doctor at the University of Twente
on the authority of the rector magnificus,
prof.dr. T.T.M. Palstra
on account of the decision of the graduation committee,
to be publicly defended
on Friday the 3rd of November 2017 at 16:45.

by

Christopher Gerard Zeinstra
born on the 28th of August, 1971
in Geelong, Australia.

This dissertation has been approved by
Promotor: prof.dr.ir. R.N.J. Veldhuis
Co-promotor: dr.ir. L.J. Spreeuwers

Voor Mum, Dad en Hanneke

Je gaat het pas zien als je het doorhebt
Johan Cruijff

Contents

1 Introduction **1**
- 1.1 Forensic Face Recognition . . . 1
- 1.2 Research questions . . . 4
- 1.3 Contributions . . . 5
- 1.4 Overview of dissertation . . . 6
- 1.5 List of publications . . . 8

2 From biometric science and forensic science to Forensic Face Recognition **11**
- 2.1 Introduction . . . 11
- 2.2 Biometrics . . . 12
  - 2.2.1 Biometric characteristics . . . 12
  - 2.2.2 Biometric system architecture . . . 13
  - 2.2.3 Biometric use cases . . . 14
  - 2.2.4 Multi-modal, fusion, and soft biometrics . . . 15
  - 2.2.5 Performance: a biometric perspective . . . 15
- 2.3 Forensic science and forensic biometrics . . . 19
  - 2.3.1 Likelihood ratio paradigm: concept . . . 19
  - 2.3.2 Likelihood ratio paradigm: implementation . . . 20
  - 2.3.3 Performance: a forensic perspective . . . 22
- 2.4 Forensic Face Recognition as a means to determine strength of evidence: a survey . . . 23
  - 2.4.1 Abstract . . . 23
  - 2.4.2 Introduction . . . 23
  - 2.4.3 Operational level of FFR . . . 25
  - 2.4.4 Tactical and strategic levels of FFR . . . 28
  - 2.4.5 Criticism on FFR . . . 29
  - 2.4.6 FFR research directions . . . 30
  - 2.4.7 Conclusion and future directions . . . 33
- 2.5 FISWG characteristic descriptors and FFR classifiers . . . 34
  - 2.5.1 FISWG characteristic descriptors . . . 34
  - 2.5.2 FFR classifiers . . . 35
  - 2.5.3 Preprocessing: PCA and LDA . . . 36
  - 2.5.4 Neyman Pearson Lemma . . . 38
- 2.6 Chapter conclusion . . . 39

3 Human performance on an eyebrow verification task **41**
3.1 Introduction . . . 41
3.2 Examining the examiners: an on line eyebrow verification experiment inspired by FISWG . . . 41
3.2.1 Abstract . . . 41
3.2.2 Introduction . . . 42
3.2.3 Related work . . . 42
3.2.4 Quantification of FISWG characteristic descriptors . . . 43
3.2.5 Experimental setup . . . 44
3.2.6 Experimental results and discussion . . . 46
3.2.7 Conclusion . . . 49
3.2.8 Future work . . . 50
3.2.9 Acknowledgments . . . 50
3.3 Chapter conclusion . . . 50

4 Classifier performance on the periocular region **53**
4.1 Introduction . . . 53
4.2 Towards the automation of forensic facial individualisation: comparing forensic to non-forensic eyebrow features . . . 54
4.2.1 Abstract . . . 54
4.2.2 Introduction . . . 54
4.2.3 Related work . . . 55
4.2.4 Methods . . . 55
4.2.5 Experimental setup and results . . . 58
4.2.6 Conclusions and future work . . . 60
4.3 Beyond the eye of the beholder: on a forensic descriptor of the eye region . . . 60
4.3.1 Abstract . . . 60
4.3.2 Introduction . . . 60
4.3.3 Related work . . . 61
4.3.4 Methods . . . 62
4.3.5 Experiments . . . 64
4.3.6 Results . . . 66
4.3.7 Conclusion . . . 67
4.4 Chapter conclusion . . . 68

5 ForenFace dataset and toolset **69**
5.1 Introduction . . . 69
5.2 ForenFace: a unique annotated forensic facial image dataset and toolset . . . 69
5.2.1 Abstract . . . 69
5.2.2 Introduction . . . 70
5.2.3 Data . . . 74
5.2.4 Annotation . . . 75
5.2.5 Toolset . . . 79
5.2.6 Potential uses, evaluation protocols, and an example . . . 80
5.2.7 Conclusion . . . 83
5.2.8 Acknowledgments . . . 84

5.3 Chapter conclusion 84

6 FISWG characteristic descriptors under various forensic use cases 85
6.1 Introduction 85
6.2 Discriminating power of FISWG characteristic descriptors under different forensic use cases 86
6.2.1 Abstract 86
6.2.2 Introduction 86
6.2.3 Related work 87
6.2.4 FISWG characteristic descriptors 87
6.2.5 Forensic use cases and the ForenFace dataset 88
6.2.6 Experimental setup 89
6.2.7 Experimental results and discussion 92
6.2.8 Conclusion 95
6.2.9 Acknowledgment 95
6.3 Manually annotated characteristic descriptors: measurability and variability . 95
6.3.1 Abstract 95
6.3.2 Introduction 96
6.3.3 FISWG characteristic descriptors 96
6.3.4 Experimental setup 97
6.3.5 Experimental results and discussion 100
6.3.6 Conclusion 104
6.4 Chapter conclusion 105

7 Subject based: facial marks and a theoretical construction 107
7.1 Introduction 107
7.2 Grid based likelihood ratio classifiers for the comparison of facial marks . . . 108
7.2.1 Abstract 108
7.2.2 Introduction 108
7.2.3 Related Work 111
7.2.4 Methods 112
7.2.5 Experiments 117
7.2.6 Results and discussion 120
7.2.7 Conclusion and Future Work 127
7.3 Label specific versus general classifier performance: an extreme example . . 129
7.3.1 Abstract 129
7.3.2 Introduction 129
7.3.3 Main result 130
7.3.4 Conclusion 134
7.4 Chapter conclusion 134

8 Subject Based: framework and random versus non-random performance 137
8.1 Introduction 137
8.2 Mind the gap: a practical framework regarding classifiers for forensic evidence evaluation 138
8.2.1 Abstract 138

8.2.2 Introduction . . . 138
8.2.3 Framework . . . 140
8.2.4 Application 1: Balaclava . . . 145
8.2.5 Application 2: Grid Based Facial Mark Likelihood Ratio Classifiers . 151
8.2.6 Conclusion . . . 154
8.2.7 Acknowledgement . . . 155
8.3 How random is a classifier given its Area under Curve? . . . 155
8.3.1 Abstract . . . 155
8.3.2 Introduction . . . 155
8.3.3 Related Work . . . 156
8.3.4 Partition functions . . . 157
8.3.5 Exact Probabilities and an Approximation . . . 158
8.3.6 Examples . . . 159
8.3.7 Discussion . . . 161
8.3.8 Conclusion . . . 161
8.4 Chapter conclusion . . . 162

9 Conclusion and recommendations 165
9.1 Conclusions . . . 165
9.2 Final conclusion . . . 168
9.3 Recommendations for future research . . . 169
9.3.1 Recommendation 1 . . . 169
9.3.2 Recommendation 2 . . . 170
9.3.3 Recommendation 3 . . . 170
9.3.4 Recommendation 4 . . . 170
9.3.5 Recommendation 5 . . . 171

References 173

Summary 187

Samenvatting 189

Dankwoord 191

List of Figures

2.1 Examples of biometric modalities. 12
2.2 Essential stages in a biometric system. 13
2.3 Examples of features. 14
2.4 Examples of comparison score distributions and the corresponding ROC curves. 17
2.5 Examples Bertillonage system. 24
2.6 Holistic and detailed perspective on the face (characteristic descriptors). . . . 36
2.7 Upper, middle, and lower part of the face (characteristic descriptors). 37

3.1 A-E characteristic descriptors of the eyebrow. 43
3.2 Example eyebrow pair. 44
3.3 Interface Experiment A. 45
3.4 Interface Experiment B. 46
3.5 Performance at individual level with 95% credible interval (CI). 47
3.6 Performance as group. 49

4.1 Dong Woodard feature set. 55
4.2 A-E characteristic descriptors of the eyebrow. 56
4.3 Performance Experiment 1. 59
4.4 Performance Experiment 2. 60
4.5 Example appearance based features. 63
4.6 Example annotation lower eyelid. 64
4.7 Selection of performances of Experiments 1, 2, and 3. 65

5.1 Some eyebrow features. 70
5.2 Top view layout experiment. 73
5.3 Example CCTV footage from Camera 3. 73
5.4 Extracted stills. 76
5.5 Other images. 77
5.6 Holistic and detailed perspective on the face (annotation). 78
5.7 Upper, middle, and lower part of the face (annotation). 79
5.8 Examples of the four annotation types. 79
5.9 Example images that have been annotated. 80
5.10 The provided software tools. 81
5.11 ROC curves of baseline experiments. 82

6.1 Overview of our system. 87
6.2 Holistic and detailed perspective of the face (characteristic descriptors). . . . 88
6.3 Upper, middle, and lower part of the face (characteristic descriptors). 89
6.4 Available images. 90
6.5 Lowest EER of single and combined characteristic descriptors. 92
6.6 ROC curves of commercial systems versus characteristic descriptors. 94
6.7 Visualisation of pairwise difference. 98
6.8 Total standard deviation of landmarks. 101
6.9 Histogram of evidential values for same source and different source cases. . . 104

7.1 Considered facial mark types. 109
7.2 Addressed aspects that influence the evaluated biometric system. 110
7.3 Facial scars and marks in the Bertillonage system. 111
7.4 Example Grid. 113
7.5 Three example annotations. 118
7.6 Facial marks histograms. 120
7.7 Number of facial marks based on age and ethnicity. 121
7.8 Spatial pattern examples. 122
7.9 Binary: General and subject based EER evaluation. 123
7.10 Category: General and subject based EER evaluation. 124
7.11 Sampled ROC curves and percentage of subjects with EER=0. 125
7.12 Number of facial marks and EER. 126
7.13 General Cllr^{cal} evaluation as function of Δ. 127
7.14 Subject based Cllr^{cal} evaluation as function of Δ. 128
7.15 Correlation between Cllr^{cal} and number of facial marks. 129

8.1 Design Framework. 139
8.2 Five subjects with five observations in feature space. 144
8.3 Balaclava and FISWG characteristic descriptors. 146
8.4 Box plots of score and feature based likelihood ratio models. 150
8.5 Variation in performance of nine FISWG characteristic descriptors. 151
8.6 Facial mark grid, EER Hamming and example subject. 154
8.7 p(AUC) for $m = 1, \cdots, 15$ genuine and $n = 100$ imposter scores. 160
8.8 The upper limit of 95% and 99% confidence intervals of the approximation. . 161
8.9 Exact p(AUC) and approximation of p(AUC). 162

List of Tables

3.1 Number of statistically significant changes on experiments A and B. 46
3.2 Number of participants "guessing" on experiments A and B. 48
3.3 Accuracy on experiments A and B. 48
3.4 Accuracy given the highest confidence level on experiments A and B. 49
3.5 Correlation between aggregated correct judgments. 49
3.6 Optimal vote threshold for positive judgment, accuracy and (FAR, TAR). . . . 50

4.1 Sublist FISWG characteristic eye components and their descriptors. 62
4.2 Representation of non-appearance features. 63
4.3 Performance appearance features in terms of AUC. 66
4.4 Performance non-appearance features in terms of AUC. 67

5.1 Contents of datasets. 72
5.2 Surveillance cameras setup. 74
5.3 Positions A-D. 75
5.4 Surveillance camera types. 75
5.5 Available video sequences and extracted images. 76
5.6 Other images and 3D scans. 77
5.7 Annotated trace and reference images. 80

6.1 EER of score fusion. 93
6.2 Measurability of characteristic descriptors. 100
6.3 Total pairwise difference for some closed shapes. 102
6.4 Total pairwise difference for some open shapes. 102
6.5 Standard deviations of distances, the fissure angle and some counts. 103
6.6 Types of annotation variability influence on evidential value. 103
6.7 Influence of annotator variability on evidential value. 105

7.1 Overview features and classifiers 115
7.2 Overview demographics FRGCv2 120

Chapter 1

Introduction

1.1 Forensic Face Recognition

Forensic Face Recognition (FFR) is the use of biometric face recognition for several applications in forensic science. Biometric face recognition uses the face modality as a means to discriminate between human beings; forensic science is the application of science and technology to law enforcement. There are two image types involved in FFR. The trace image often captures a crime scene and is most of the time taken under uncontrolled conditions. The reference image is a photograph of a suspect and is taken under controlled conditions. In general, as described by Meuwly and Veldhuis [1], FFR includes scenarios of ID verification, identification, investigation and intelligence, and evaluation of strength of evidence. The evaluation of strength of evidence is commonly referred to as *forensic evidence evaluation*. The strength of evidence, in combination with prior assumptions, can be used by a court of law in its verdict whether a suspect is considered guilty or not. This dissertation is primarily concerned with topics related to forensic evidence evaluation in the domain of FFR.

The field of face recognition has made impressive improvements in the last two decades. State-of-the-art biometric face recognition can recognise faces with low error rates (e.g. a false-rejection probability of 1% at a false-acceptance probability of 0.1%) [2]. Although face recognition systems in principle can be used for investigation and intelligence purposes, forensic evidence evaluation is still largely a manual process performed by human FFR-examiners. They are able to amortise common influences on the quality of trace material during their assessment of trace and reference images. We refer to [3] for a study on (performance) differences between FFR-examiners and non-examiners. The influences include image compression artifacts, lens distortion, perspective effects, low resolution, interlacing, pose, illumination, and expression. Also, partial occlusion of the face is commonly encountered in trace images. These influences restrict the use of a standard face recognition system. An additional reason to be somewhat reluctant towards the use of face recognition systems is their use of abstract, general feature descriptors like SIFT [4] and LBP [5]. These descriptors are not endowed with any forensic meaning and are hardly understandable outside the technical computer vision domain, in particular in a court of law.

During the manual forensic evidence evaluation process, traces and references are assessed by the FFR-examiner who will pay attention to mostly shape like and potentially

highly discriminating facial features [6]. The Facial Identification Scientific Working Group (FISWG) [7] has published the Facial Image Comparison Feature List for Morphological Analysis [8]. It describes characteristic descriptors (facial features) that can be used during forensic evidence evaluation. Although this feature list is not a formal standard, similar forensic evidence evaluation procedures in The Netherlands and Sweden [9–11] indicate that it can be regarded as an informal standard, representative of those used throughout other countries as well [12].

The mere fact that the characteristic descriptors are documented in the FISWG Feature List does not automatically imply their suitability, in particular for their intended use under forensically relevant conditions. Actually, little research is done on this topic. The transfer from the Frye to the Daubert rule and the very critical report of the National Research Council of the National Academies on the state of forensic science in the USA, is an additional incentive to initiate such research on FISWG characteristic descriptors.

Prior to 2000, admissibility of expert evidence presented to a US trial court was governed by the Frye rule. This rule states that evidence is admissible as long its method is "(...) sufficiently established to have gained general acceptance in the particular field in which it belongs." [13]. In almost all jurisdictions, this rule has been superseded by the Daubert rule ("a trial judge must ensure that any and all scientific testimony or evidence admitted is not only relevant, but reliable") [13]. This rule puts more emphasis on the used methodology being scientific. This includes the use of peer reviewed methods, insight in known or potential error rates, the formulation of hypotheses, and the conduction of experiments to prove or to falsify hypotheses. In other words, there has been a shift from conclusions or opinions under the Frye rule to strength of evidence established in a scientific manner under the Daubert rule. A summary of forensic facial expert testimony illustrating the dire, non-scientific approach in some selected cases can be found in [14]. In 2009 the National Research Council of the National Academies published an elaborate and critical report [15] on the current state of forensic science in the USA. It includes an in depth discussion of the Frye and Daubert rules and its implications on current practice of forensic science. In total 13 recommendations have been formulated. Recommendation (3) is of particular interest: "Research is needed to address issues of accuracy, reliability, and validity in the forensic science disciplines. (...)".

Considering this discussion, we are interested in several aspects related either directly or indirectly to the FISWG characteristic descriptors. These aspects start in the vicinity of the current practice, the human FFR-examiner, and they gradually zoom out towards the presentation of a practical framework for forensic evidence evaluation that in principle also can be applied to research outside the FFR domain. These, in total eight, aspects in turn form the basis of the addressed research questions in this dissertation.

The first aspect is how well FFR-examiners and non-examiners perform on a comparison task when they use FISWG characteristic descriptors versus a best-effort approach. The results are indicative of the added value of characteristic descriptors over an alternative approach.

Starting from the second aspect, we set the human aside and focus on the design and usage of biometric classifiers. The previously mentioned face recognition systems are examples of biometric classifiers. In general, a classifier compares a trace (having a questioned label) and a reference (having a known label), outputs a comparison score that encapsulates how convinced the classifier is that trace and reference input have a common label, and given a

threshold, makes a decision[1]. If the comparison score exceeds this threshold, the decision is affirmative: trace and reference are assumed to have a common label, otherwise different labels are assumed. Although in this dissertation we use the term classifier, we are mostly interested in the produced comparison score. A biometric classifier is a classifier that uses biometric features as its input. In particular, we will primarily focus on biometric classifiers that use characteristic descriptors as their input. Furthermore, we are interested in comparison scores that are either modelled or converted to strength of evidence. The input and output of such classifiers have a clear forensic meaning and are understandable by a court of law, as opposed to the previously mentioned abstract, general feature descriptors like SIFT. Also, by using biometric classifiers that are specialised on a specific characteristic descriptor, we have by design the guarantee that only the descriptor is taken into account during the computation of strength of evidence.

Returning to the second aspect, it focuses on classifiers using FISWG characteristic descriptors as their input, producing strength of evidence, and how they perform in general in relation to other biometric classifiers that use non-forensic features, under relatively well-conditioned settings. General performance is measured by considering the comparison scores of a biometric classifier when it is offered a set of trace-reference pairs of multiple subjects whose ground truth (same source, different source) is known.

The third aspect extends the previous aspect by using trace images that are more representative of various forensic use cases. It considers the general performance of biometric classifiers using characteristic descriptors as their input, also in relation to face recognition systems.

The fourth aspect shifts the focus from the biometric classifier to mostly properties of the characteristic descriptors themselves. In particular, it considers (a) their measurability and (b) the influence of measurement variation on the value of characteristic descriptors and produced strength of evidence. Measurability refers to which extent characteristic descriptors can be extracted. Furthermore, in this dissertation, most characteristic descriptors have been extracted from manual annotation. This is due to the lower quality of trace images and the general difficulty of implementing a semantic definition of a characteristic descriptor in a robust extraction algorithm.

The fifth aspect considers differences between general and subject based performance. Subject based performance is measured by considering the comparison scores of a biometric classifier when it is offered a set of trace-reference pairs for which the traces only originate from the subject at hand, the references come from multiple subjects, and for each pair the ground truth (same source, different source) is known. The reason to consider this, is that a biometric classifier using a characteristic descriptor as its input might have poor general performance, whereas the subject based performance might be better or even good. We believe that this behaviour is exemplary for the face modality in a forensic context; looking into this matter seems warranted. Insight in the variation of subject based performance is indicative of the proportion of cases in which the characteristic descriptor could be used to discriminate a subject. Moreover, inspecting the appearance of a characteristic descriptor of a particular subject whose biometric classifier exhibits a good subject based performance connects its phenotype to that performance and is potentially beneficial for identifying discriminative characteristic descriptors in general. Finally, it shows the contribution of each characteristic

[1]The term label encompasses both enemy plane (class) and Joe Doe (individual).

descriptor but also their limits. This aspect is taken into account by considering empirical results and a theoretical construction creating a gap between perfect subject based and general random performance.

The sixth aspect considers the suitability of facial marks in forensic evidence evaluation and extends the previous subject based performance to a broader subject based approach. Facial marks are interesting as they are representative of FISWG characteristic descriptors that have a potential to be very discriminative. This aspect describes a proto-framework that contains possible choices during the design and evaluation of biometric classifiers that use features derived from facial mark locations. An example choice is whether to consider a classifier that is trained with subject based data. It also incorporates other, forensically relevant, performance characteristics that can be evaluated at a subject based level. The proto-framework is created as a response to existing facial mark classifier studies.

The seventh aspect extends the proto-framework of the previous aspect into a framework, applicable to the design and evaluation of biometric classifiers for forensic evidence evaluation in general, in principle even applicable outside the FFR domain, with a special emphasis on the subject based approach. Also, its applicability is shown by considering two relevant applications in the domain of FFR of which one extends the facial mark study.

The eighth, and final, aspect complements the previous aspects in an abstract manner. Although the subject based performance might be reasonable or even good in some cases, a large proportion of biometric classifiers will probably have a performance that is poor to the extent that it is unclear whether it could have been produced by a random classifier, that is, a classifier that essentially outputs random comparison scores without considering the trace and reference inputs. This aspect takes a particular performance measure, the Area Under the Curve (AUC), and quantifies the boundary between random and non-random performance.

Overall, we believe that by addressing these eight aspects in this dissertation, the FISWG characteristic descriptors are considered from relevant points of view and as such our approach does justice to the intention encapsulated in the Daubert rule.

1.2 Research questions

Given the discussion and presented aspects in the previous section, we address the following two main research questions and subordinate research questions in this dissertation. The two main research questions are only addressed by their subordinate research questions; in Chapter 9 we will revisit the main research questions.

1. What is the suitability of FISWG characteristic descriptors as a means to discriminate, taking human, classifier, feature, and forensic aspects into account?

    (a) Under relatively well-conditioned settings, what is the performance of FFR examiners in relation to non-examiners, both using FISWG characteristic descriptors and a best-effort approach in a verification task?

    (b) Under relatively well-conditioned settings, what is the general performance of biometric classifiers that use FISWG characteristic descriptors as their input and produce strength of evidence in relation to other non-forensic biometric classifiers?

(c) Under various forensic use cases, what is the general performance of biometric classifiers that use FISWG characteristic descriptors as their input and produce strength of evidence in relation to face recognition systems?

(d) Under various forensic use cases, what is (a) the measurability of FISWG characteristic descriptors and (b) the influence of annotation variation on characteristic descriptors and strength of evidence produced by biometric classifiers that use these characteristic descriptors?

2. What is the suitability of a subject based approach in forensic evidence evaluation, taking empirical results from specific applications, theoretical results, and a framework approach into account?

   (a) To which extent do we observe or can we construct differences in general and subject based performance?

   (b) How well can facial marks be used for forensic evaluation, also taking subject based data and subject based evaluation into account?

   (c) In which manner can a biometric approach to FISWG characteristic descriptors be generalised into a framework for forensic evidence evaluation that also incorporates a subject based approach?

   (d) What is a theoretical boundary between random and non-random behaviour of classifiers in a subject based performance evaluation based on AUC?

1.3 Contributions

This dissertation makes the following contributions to the field of Forensic Facial Recognition:

- The human performance on an eyebrow verification task.
- The general performance of biometric classifiers using FISWG characteristic descriptors compared to those that use non-forensic inspired features with respect to the periocular (eye and eyebrow) region.
- The ForenFace dataset (annotation, software and documentation).
- The general performance of biometric classifiers that use FISWG characteristic descriptors as their input and produce strength of evidence under various forensic use cases, also in comparison to face recognition systems.
- The measurability and variability of FISWG characteristic descriptors under various forensic use cases, including the variability effect on strength of evidence produced by biometric classifiers.
- A comparison of general and subject based performance (discriminating power and calibration) of various biometric classifiers that use features derived from facial mark locations.

Moreover, this dissertation contributes to the broader field of Forensic Biometrics:

- A practical framework for the design and evaluation of biometric classifiers for forensic evidence evaluation, with a specific emphasis on a subject based approach.

Finally, this dissertation also contributes to the field of Pattern Recognition:

- A theoretical construction showing that classifiers can exhibit perfect subject based performance, while the general performance is essentially random.
- The exact probability of AUC values produced by a random classifier and an approximation to them from which a boundary between random and non-random behaviour can be derived.

1.4 Overview of dissertation

This dissertation contains mostly published and submitted work. Each chapter contains an introduction that describes its structure and indicates which publications are included. Also, each chapter contains a reading guide for convenience. Each manuscript is added verbatim, apart from error corrections, changes to harmonise some terminology, and small clarifications. In particular:

- We use examiner or FFR-examiner as a neutral term instead of practitioner or expert;
- We sometimes use evidential value as a synonym for strength of evidence.

In some chapters, the connection to preceding and following chapters is also mentioned to reinforce the underlying narrative. In addition, Chapters 2 to 8 close with a "Chapter Conclusion" that describes the contribution of the contents of the chapter, and when applicable, its relation to one or more research questions.

Chapter 2 introduces some key concepts of biometrics, forensic science, and forensic biometrics in general. A large part of Chapter 2 presents Forensic Face Recognition at three levels (Operational, Tactical, and Strategic). Furthermore, it discusses criticism on FFR and past and current research directions related to FFR. This part has been submitted as "Forensic Face Recognition as a means to determine Strength of Evidence: a survey" [16]. The final part presents the FISWG characteristic descriptors and provides examples of biometric classifiers that produce a likelihood ratio, the quantity that represents strength of evidence. The goal of this chapter is to acquaint the reader with the context and the tools of this dissertation.

Chapter 3 contains a single study on an eyebrow verification task in which the extent of performance differences between (a) FFR-examiners and non-examiners and (b) FISWG characteristic descriptors and "best-effort" approaches are considered. It has been published as "Examining the examiners: an online eyebrow verification experiment inspired by FISWG" [17].

Chapter 4 studies the performance of biometric classifiers that use FISWG characteristic descriptors as their input in relation to those that use other non-forensic features as their input

with respect to the periocular region. This is the region around the eye and includes the eyebrow. This chapter consists of two parts. The first part has been published as "Towards the automation of forensic facial individualisation: Comparing forensic to non-forensic eyebrow features" [18]. It compares biometric classifiers that use the FISWG characteristic descriptors of eyebrows to those using non-forensic features introduced by a study of Dong and Woodard [19]. The second part is a small scale study that compares classifiers using FISWG characteristic descriptors of the eye to classifiers using non-forensic texture based approaches commonly encountered in periocular biometrics. It has been published as "Beyond the eye of the beholder: on a forensic descriptor of the eye region" [20].

Chapter 5 describes the ForenFace dataset. This dataset is introduced since an analysis showed that other datasets used in the realm of forensic research are not fully suitable for the study of FISWG characteristic descriptors under various forensic use cases. The main asset of the ForenFace dataset is the availability of manual annotation (landmarks, shapes, etc.) from which the characteristic descriptors can be derived. Also, the dataset contains various surveillance camera image types that correspond to representative forensic use cases. The chapter describes the acquisition, details of the annotation, and the available software tools. Also, it specifies evaluation protocols and compares the biometric performance of a baseline experiment using a face recognition system to what can be achieved with a specific characteristic descriptor. This chapter has been published as "ForenFace: a unique annotated forensic facial image dataset and toolset" [21].

Chapter 6 contains two studies, conducted using various forensic use cases introduced by the ForenFace dataset of Chapter 5. The first part of this chapter studies discriminating power in terms of EER of biometric classifiers using FISWG characteristic descriptors extracted from the ForenFace dataset. Four types of biometric classifiers are being used. Also, results acquired by combining results of either classifier type or facial category are presented. It has been published as "Discriminating power of FISWG characteristic descriptors under different forensic use cases" [22]. The second part of this chapter studies two other related properties of characteristic descriptors. The first property is measurability, that is, to which extent can characteristic descriptors be extracted on images representative of various forensic use cases. The second property is variability and studies the influence of annotator variability on landmark positions, shapes, etc. It also measures the influence of the annotator variability on the produced strength of evidence by biometric classifiers. It has been published as "Manually annotated characteristic descriptors: measurability and variability" [23].

Chapter 7 contains two studies. One study considers various biometric classifiers that use features derived from facial mark locations. This study identifies six, mostly forensic, aspects that are hardly considered in other studies on facial marks. These aspects include (a) the explicit use of subject based data, (b) the incorporation of subject based evaluation, and (c) the use of other, forensic, performance characteristics. It has been accepted for publication as "Grid Based Likelihood Ratio Classifiers for the Comparison of Facial Marks" [24]. The second, short, part presents a theoretical construction that shows that classifiers can exhibit perfect subject based[2] performance, while the general performance is essentially random. Its

[2]Coined label specific in the context of general classifiers.

aim is to complement the first part of this chapter that illustrates similar, but less extreme, behaviour. It has been published as "Label specific versus general classifier performance: an extreme example" [25].

Chapter 8 generalises the introduced aspects of Chapter 7 into a framework for forensic evidence evaluation. The first part of this chapter has been submitted as "Mind the Gap: A Practical Framework regarding Classifiers for Forensic Evidence Evaluation" [26]. It also includes two example applications. The first application is the use of nine simple characteristic descriptors, applicable in the case when a perpetrator wears a balaclava. The results show the large variation in discriminating power observed from a subject based evaluation. The second application extends the facial mark study of Chapter 7 by considering results on another forensically relevant dataset. The second part of Chapter 8 presents the exact probability of the Area Under the Curve (AUC) values produced by a random classifier and an approximation to them. The AUC measured on a finite set of scores is a random variable itself, and it is possible that the AUC is small to moderate, while the underlying biometric classifier is random. This is of relevance as the subject based evaluation introduced in Chapter 7 and the first part of this chapter typically uses a low number of genuine and imposter scores, exactly the situation in which this effect is the most apparent. This study has been accepted for publication as "How Random is a Classifier given its Area under Curve?" [27].

Chapter 9 is the closing chapter. It revisits the research questions and discusses how the work presented in this dissertation has addressed these questions. Also, recommendations for future research are presented.

1.5 List of publications

Chapters 2 to 8 are based on conference and journal papers, either submitted or published. We list them in order of appearance in this dissertation.

- [16] C. G. Zeinstra, D. Meuwly, A. C. C. Ruifrok, R. N. J. Veldhuis, and L. J. Spreeuwers. Forensic Face Recognition as a means to determine Strength of Evidence: a survey. *Submitted to Forensic Science Review*.

- [17] C. G. Zeinstra, R. N. J. Veldhuis, and L. J. Spreeuwers. Examining the examiners: an online eyebrow verification experiment inspired by FISWG. In *International Workshop on Biometrics and Forensics, IWBF 2015 , Glövik, Norway*, pages 1–6, USA, March 2015. IEEE Computer Society.

- [18] C. G. Zeinstra, R. N. J. Veldhuis, and L. J. Spreeuwers. Towards the automation of forensic facial individualisation: Comparing forensic to non-forensic eyebrow features. In *Proceedings of the 35th WIC Symposium on Information Theory in the Benelux, Eindhoven, Netherlands*, pages 73–80, Enschede, May 2014. Centre for Telematics and Information Technology, University of Twente.

- [20] C. G. Zeinstra, R. N. J. Veldhuis, and L. J. Spreeuwers. Beyond the eye of the beholder: on a forensic descriptor of the eye region. In *23rd European Signal Processing*

Conference, EUSIPCO 2015, Nice, pages 779–783. IEEE Signal Processing Society, September 2015.

- [21] Chris G. Zeinstra, Raymond N.J. Veldhuis, Luuk J. Spreeuwers, Arnout C.C. Ruifrok, and Didier Meuwly. Forenface: a unique annotated forensic facial image dataset and toolset. *IET Biometrics*, May 2017. http://digital-library.theiet.org/content/journals/10.1049/iet-bmt.2016.0160.

- [22] C. G. Zeinstra, R. N. J. Veldhuis, and L. J. Spreeuwers. Discriminating power of FISWG characteristic descriptors under different forensic use cases. In *BIOSIG 2016 - Proceedings of the 15th International Conference of the Biometrics Special Interest Group, 21.-23. September 2016, Darmstadt, Germany*, volume 260 of *LNI*, pages 171–182. GI, 2016.

- [23] Chris Zeinstra, Raymond Veldhuis, Luuk Spreeuwers, and Arnout Ruifrok. Manually annotated characteristic descriptors: measurability and variability. In *International Workshop on Biometrics and Forensics, IWBF 2017, Conventry, United Kingdom*.

- [24] Chris Zeinstra, Raymond Veldhuis, and Luuk Spreeuwers. Grid Based Likelihood Ratio Classifiers for the Comparison of Facial Marks. *Accepted for publication in IEEE Transactions on Information Forensics and Security*, 2017. http://dx.doi.org/10.1109/TIFS.2017.2746013.

- [25] Chris Zeinstra, Raymond Veldhuis, and Luuk Spreeuwers. Label specific versus general classifier performance: an extreme example. *University of Twente Students Journal of Biometrics and Computer Vision*. http://dx.doi.org/10.3990/3.utsjbcv.i0.25.

- [26] Chris Zeinstra, Raymond Veldhuis, Luuk Spreeuwers, and Didier Meuwly. Mind the Gap: A Practical Framework regarding Classifiers for Forensic Evidence Evaluation. *Submitted to Science & Justice*.

- [27] Chris Zeinstra, Raymond Veldhuis and Luuk Spreeuwers. How Random is a Classifier given its Area under Curve? *Accepted for publication in BIOSIG 2017*.

Moreover, the following publications are not related to the topic of this dissertation:

- [28] Aad Dijksma, Heinz Langer, Yuri Shondin, and Chris Zeinstra. Self-adjoint operators with inner singularities and Pontryagin spaces. In *Operator Theory and Related Topics*, pages 105–175. Springer, 2000.

- [29] M. A. Kaashoek and C. G. Zeinstra. The band method and generalized Carathéodory-Toeplitz interpolation at operator points. *Integral Equations and Operator Theory*, 33(2):175–210, 1999.

Chapter 2

From biometric science and forensic science to Forensic Face Recognition

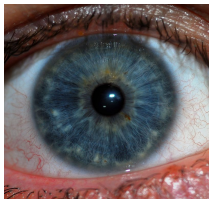
2.1 Introduction

In this chapter, we introduce some essential concepts underlying biometric science and biometric classifiers. We subsequently present an important concept within forensic science: the likelihood ratio as the bearer of strength of evidence, usable in a court of law. A large part of this chapter is devoted to FFR, with a particular emphasis on strength of evidence. It discusses the operational, tactical, and strategic levels of FFR. Also, criticism and research directions (past and current) are presented. The last section presents FISWG characteristic descriptors and includes two examples of biometric classifiers that produce strength of evidence. This section acts as a gateway from this chapter to the main contents of the dissertation.

Section 2.4 has been submitted as "Forensic Face Recognition as a means to determine strength of evidence: a survey" [16].

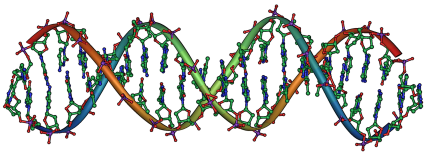
Reading Guide
Section 2.2. This section can be omitted by readers who already are familiar with the basic biometric concepts.
Section 2.3. This section can be omitted by readers who already are familiar with the role of the likelihood ratio in forensic science.
Section 2.4. This section should at least be browsed in order to see which aspects of FFR have been addressed in past and current research.
Section 2.5. This section should at least be browsed as it introduces the main ingredients of this dissertation. The last two subsections contain some mathematical aspects related to classifiers and can be omitted.

a) Ridge pattern finger tip



b) Iris

c) Face



d) DNA

e) Gait

Figure 2.1: Examples of biometric modalities. a) Taken from [31], b) Taken from [32], c) Image 02463d256 from [33], d) Taken from [34], and e) Taken from [35].

2.2 Biometrics

According to Jain et al. [30], biometrics is the science of establishing the identity of an individual based on the physical, chemical or behavioural attributes of a person. Typical examples of biometric modalities shown in Figure 2.1 are the ridge pattern on finger tips, the iris, and face (physical), DNA (chemical), and gait (behavioural).

2.2.1 Biometric characteristics

Jain et al. [30] describes seven characteristics a biometric modality should have in order to be usable.

- Universality: every individual should possess the modality.
- Distinctiveness: the ability to adequately discriminate between individuals of an entire population based on that particular modality.
- Permanence: how persistent an individual's biometric modality is over time with respect to the application and the matching algorithm used. If a modality does not possess sufficient permanence and thus changes dramatically over time, it is unsuitable for biometrics.
- Measurability: how possible it is to capture the biometric feature using a suitable device without causing harm or undue inconvenience via the capture procedure. The raw data captured must also allow for further processing, such as feature extraction.
- Performance: the recognition accuracy in terms of the resources required and the constraints imposed by the application.

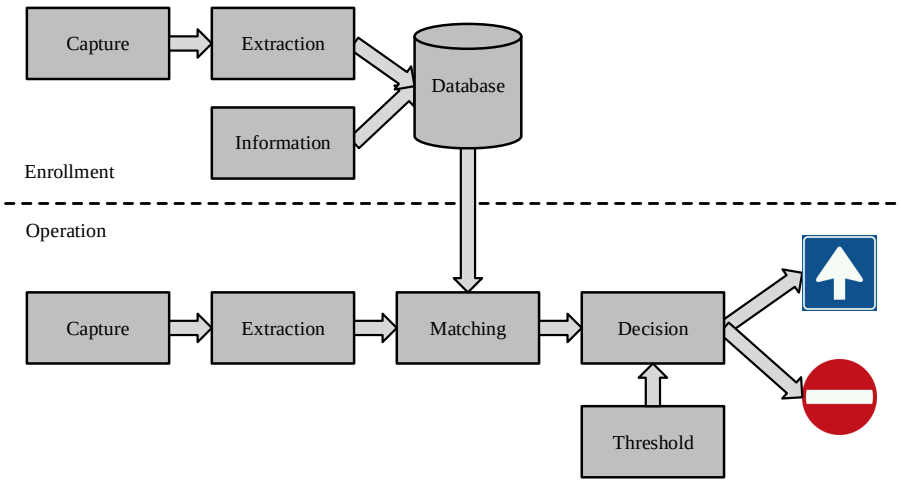


Figure 2.2: Essential stages in a biometric system. Top part shows the stages during the enrollment phase, bottom part shows the stages during the operational phase. Traffic sign images taken from [36] and [37].

- Acceptability: the acceptance of the biometric trait by target population and thus their willingness to use the modality.
- Circumvention: how easily an individual's physical or behavourial modality can be imitated by using artifacts or impersonation, respectively.

2.2.2 Biometric system architecture

A biometric system typically contains two phases (enrollment and operation) and four stages (capture, extraction, matching and decision) [30]. They are shown in Figure 2.2.

During the capture stage, a (dedicated) sensor captures a (digital) representation of the biometric modality. The quality of the representation is affected by a number of factors. If the sensor requires cooperation, for example a finger print sensor, any resistance by a criminal can induce a loss in quality. Also, if the sensor is not properly used, for example by applying too much pressure or it is not cleaned between captures, the quality may be too low. Finally, in an uncooperative setting, by definition, the operator of the biometric system cannot give instructions to a subject. This occurs for example in the case of surveillance cameras.

In the feature extraction phase, the captured representation is quality assessed and possibly pre-processed prior to the feature extraction. A feature is a representation of the biometric modality that is believed to contain discriminative information. For example, in the case of a finger print, minutiae are points where ridges start, end, and bifurcate. The feature representing the fingerprint are the minutiae locations and directions. Another example is the IrisCode. It consists of a binary sequence which describes the phase characteristics of the iris in a polar coordinate system. Both examples are shown in Figure 2.3.

As can be seen in Figure 2.2, both the enrollment and operational phases contain a capture

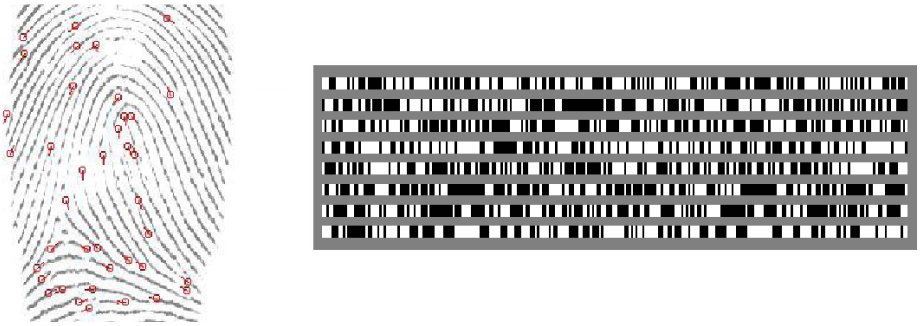


Figure 2.3: Examples of features. Left: minutiae locations and directions, taken from [38], right: IrisCode, taken from [39].

and extraction stage. With respect to the latter stage, extracted features are referred to as reference template during the enrollment and as test sample during the operational phase.

In the matching phase, the test sample and one of more reference templates who's identity are known are compared and a comparison score is calculated by a comparison score function.

An example of a comparison score function that compares aligned IrisCode test sample X and reference template Y is:

$$s(X,Y) = -\frac{\|(X \oplus Y) \wedge \text{Mask}(X) \wedge \text{Mask}(Y)\|}{\|\text{Mask}(X) \wedge \text{Mask}(Y)\|}. \tag{2.1}$$

Here, $\oplus$ denotes the bitwise exclusive or operator, $\wedge$ is the bitwise and operator, Mask is the operator that for every bit indicates whether it is visible (1) or occluded by for example the eyelid (0), and $\|\cdot\|$ counts the number of 1's in a binary sequence.

The comparison score defined by (2.1) is always non-positive, and we expect that two IrisCodes from the same person have a higher comparison score (close to zero) than two IrisCodes from two different persons (more negative). This is a general assumption throughout this dissertation: the higher the comparison score value, the more the biometric system is "convinced" that the test sample and reference template originate from the same person.

In the decision phase, the system compares the comparison score to a predefined threshold. The system decides that the test sample and reference template are from the same person or different persons if it exceeds or falls short of the threshold, respectively. Which considerations play a role in the choice of such a threshold is discussed later.

The final phase is the enrollment phase. In this phase, one or more reference templates are extracted from a subject for the purpose of storage in a reference database, alongside with identification information (for example name and identification number) and other relevant information (for example acquisition date and location). An example of the enrollment phase is the acquisition of fingerprints for a biometric passport.

2.2.3 Biometric use cases

There are two major operation modes of a biometric system.

- The first mode is the *identification* mode. The biometric system is given a test sample whose identity is unknown and the system is requested to return a list of the identities of the matching reference templates with the highest comparison scores.

- The second mode is the *verification* mode. The biometric system is given an unknown test sample and is requested to report whether or to which extent the test sample and a particular reference template match.

2.2.4 Multi-modal, fusion, and soft biometrics

Multi-modal biometrics, the combination of different biometric modalities, is commonly used. The primary reason is increased robustness against noise and other factors that influence the capture, extraction, and matching and decision making processes. A notable example of the use of multi-modal biometrics is given by the Aadhaar project [40]. This project aims to collect 10 fingerprints, two iris scans, and a facial image from each Indian citizen. Its primary aim is provide a uniform verification method during interaction with government agencies and banks.

The combination of different modalities is called fusion and its operation can be applied at several levels. The Handbook of Multibiometrics [41] describes several levels; the most common levels are the feature level, score level, and decision level. At feature level, multiple feature representations are concatenated and possibly post-processed by a dimensionality reduction step that retains most of the information in the data. At score level, scores are combined; several strategies exist, ranging from pre-scaling and adding (z-normalisation and sum-rule) to modeling dependency structures between scores of different modalities. At decision level, the binary decisions can be combined by using for example a majority voting scheme.

In recent years, so called soft biometrics have been studied extensively. They are mostly used to augment hard biometrics in a multi-modal setting. Examples of soft biometrics are gender, race, but also include for example the angle of the eye fissure. A soft biometric on its own sometimes helps to exclude a person. In some cases, it might even discriminate a person within a group, for example when that person is the only one within that group with protruding ears.

2.2.5 Performance: a biometric perspective

Given ground truth, using a fixed threshold on a comparison score gives a decision that always falls exactly in one of four classes:

- True Match: positive decision, test sample and reference templates are from the same source.

- True Non Match: negative decision, test sample and reference templates are from different sources.

- False Match: positive decision, test sample and reference templates are from different sources.

- False Non Match: negative decision, test sample and reference templates are from the same source.

An ideal biometric system does not make any mistake. In general, we can empirically assess the performance of a biometric system as follows. Given a fixed value for the threshold τ, we present the biometric system n pairs of test-reference pairs with known ground truth. If the ground truth is positive or negative, that is, test and reference have a common or a different source, the score is called genuine or imposter, respectively.

Based on the outcome, we can calculate four related measures[1]:

$$\begin{aligned} \mathrm{TNMR}(\tau) &= \frac{\#(s < \tau \wedge \mathrm{GT} = N)}{\#(\mathrm{GT} = N)} \\ \mathrm{TMR}(\tau) &= \frac{\#(s \geq \tau \wedge \mathrm{GT} = P)}{\#(\mathrm{GT} = P)} \\ \mathrm{FNMR}(\tau) &= \frac{\#(s < \tau \wedge \mathrm{GT} = P)}{\#(\mathrm{GT} = P)} \\ \mathrm{FMR}(\tau) &= \frac{\#(s \geq \tau \wedge \mathrm{GT} = N)}{\#(\mathrm{GT} = N)}. \end{aligned} \tag{2.2}$$

Here s denotes the score, and GT is the ground truth which is known to be positive (P) or negative (N). TNMR is True Negative Match Rate, and the rate refers to the measurement of the number of true negatives with respect to the total number of test-reference pairs with a negative ground truth.

Since the equalities

$$\mathrm{TNMR} + \mathrm{FMR} = \mathrm{TMR} + \mathrm{FNMR} = 1, \tag{2.3}$$

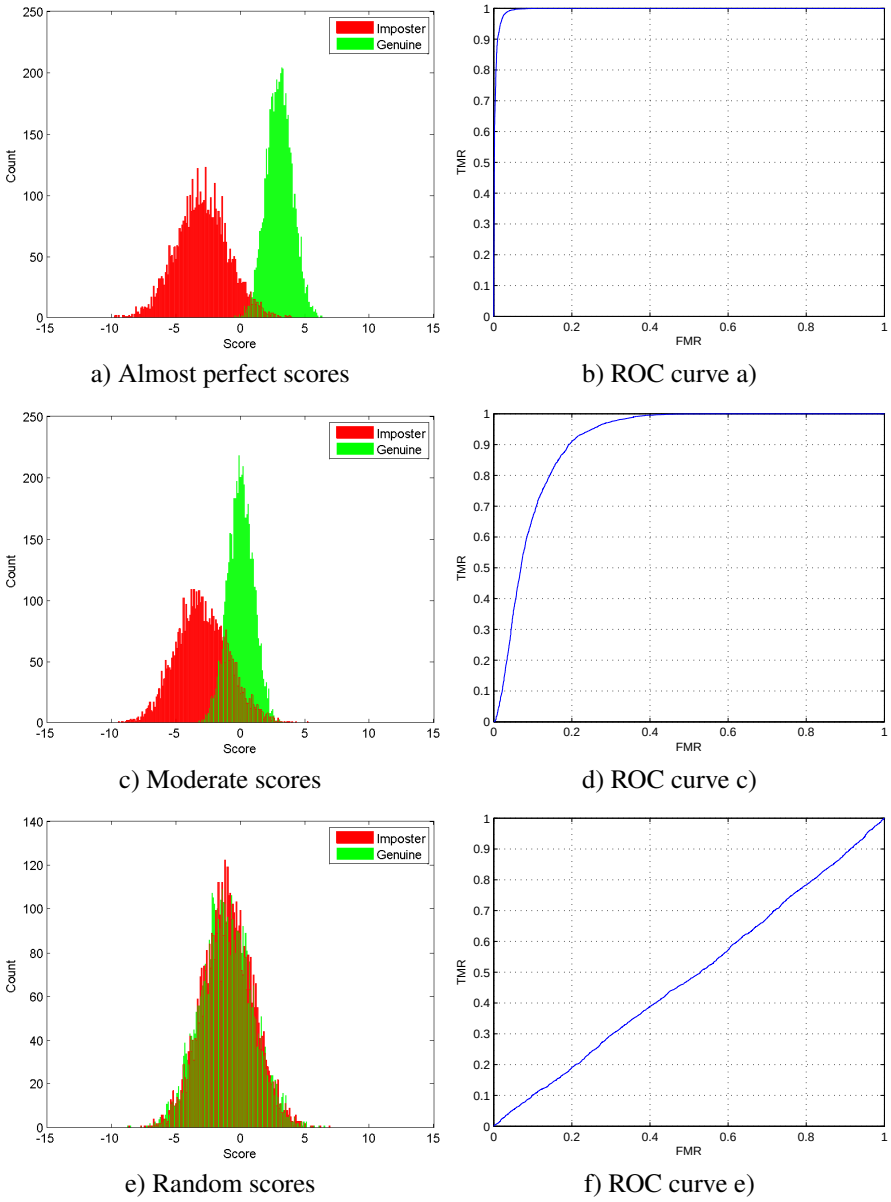
hold, it suffices to consider the common choice FMR and TMR only. An ideal biometric system does not make any errors, that is, $\mathrm{FMR} = 0$ and $\mathrm{TMR} = 1$. We make several key observations.

- FMR and TMR depend on the treshold τ, therefore for the perfect biometric system there exists a threshold τ or a range of thresholds $\tau \in \mathrm{T}$, for which $\mathrm{FMR}(\tau) = 0$ and $\mathrm{TMR}(\tau) = 1$.
- *Not all* biometric systems are created equal, so there might not exist any threshold τ for which $\mathrm{FMR}(\tau) = 0$ and $\mathrm{TMR}(\tau) = 1$.
- *Every* biometric system has the same behaviour at $\tau = -\infty$ and $\tau = +\infty$. If we set the threshold infinitely low, every decision is positive, so $\mathrm{TMR}(-\infty) = 1$ (virtue) but $\mathrm{FMR}(-\infty) = 1$ (vice). Similarly, $\mathrm{FMR}(+\infty) = 0$ (virtue) but $\mathrm{TMR}(+\infty) = 0$ (vice).

The Receiver Operator Characteristic (ROC) curve[2] is a standard method to visualise the performance of a biometric system in terms of FMR and TMR when the threshold is varied

[1] Several synonyms are commonly used, for example True Accept Rate (TAR), equal to TMR and False Accept Rate (FAR), equal to FMR, see for example Section 3.2.

[2] Sometimes the Detection Error Trade off or DET curve is used, showing the FNMR (or False Reject Rate (FRR)) as a function of FMR (or FAR), possibly with the horizontal axis warped to an inverse cumulative normal distribution. The advantage of such warping is that the DET curve of a system who's genuine and imposter scores are drawn from a normal distribution is plotted as a straight line, see Section 4.2.



a) Almost perfect scores b) ROC curve a)

c) Moderate scores d) ROC curve c)

e) Random scores f) ROC curve e)

Figure 2.4: Three examples of the empirical genuine and imposter comparison score distribution and the corresponding Receiver Operator Characteristic Curve. a) & b): almost perfect system, c) & d): moderate system, e) & f): random system.

from $\tau = -\infty$ to $\tau = +\infty$. The horizontal ordinate is the FMR, and the vertical oordinate is the TMR. The threshold $\tau = -\infty$ corresponds to $(1,1)$. As we increase the threshold, the ROC curve travels to $(0,0)$, the point corresponding to $\tau = \infty$. Also, observe that if we apply

a strictly increasing transformation on the set of scores, due to the order preserving nature of such a transformation, we obtain exactly the same ROC curve, but just reparameterised in terms of the thresholds.

Three examples of empirical genuine and imposter comparison score distributions with their ROC curves are shown in Figure 2.4. In particular, if the genuine and imposter comparison scores fully overlap, the ROC curve resembles the diagonal line TMR = FMR.

The choice of a threshold fixes the (FMR, TMR) pair. This point is called the *operating point* of the biometric system. Depending on the context, often a fixed value for FMR or FNMR is chosen, from which the threshold is derived. For example, if safety is important the FMR is set at for example 0.1%, whereas throughput or subject satisfaction is important, the FNMR is specified.

We have considered finite sets of genuine and imposter scores. Under the assumption that the genuine and imposter scores are drawn from unknown distributions with probability densities p_g and p_i respectively, we can give the continuous analog to (2.2) in terms of integrals:

$$\begin{aligned}
\text{TNMR}(\tau) &= \int_{-\infty}^{\tau} p_i(s)ds \\
\text{TMR}(\tau) &= \int_{\tau}^{\infty} p_g(s)ds \\
\text{FNMR}(\tau) &= \int_{-\infty}^{\tau} p_g(s)ds \\
\text{FMR}(\tau) &= \int_{\tau}^{\infty} p_i(s)ds.
\end{aligned} \tag{2.4}$$

The ROC curve shows the performance of a biometric system as the threshold is varied. There exist several metrics that summarise the performance of the biometric system in a single number. A selection is

- AUC: Area under the Curve. Measures what it claims to measure. A perfect system has AUC $= 1.0$, a random system has AUC $= 0.50$, see for example Figure 2.4f.
- EER: Equal Error Rate. Measures which FMR is equal to FNMR (or $1 - \text{TMR}$). Intersection between the ROC curve and the line $\text{TMR} = 1 - \text{FMR}$.

The AUC can be interpreted as the probability that a randomly chosen genuine score is larger than a randomly chosen imposter score [42]. The AUC is a nonlinear performance measure. For biometric systems with poor (0.50) to moderate (0.80) values for the AUC this is not so obvious. However, if for example the ROC curve is constructed by using genuine and imposter sets of 10 scores each, it is possible to have AUC $= 0.99$ but EER $= 10\%$! Especially when considering almost fully discriminative systems, the AUC should be used with care. Also, if the number of genuine and imposter scores is low, the AUC of a random biometric system (a system that draws comparison scores from a single probability distribution) can deviate significantly from the expected AUC $= 0.50$. This is further explored in Chapter 8 of this dissertation.

Apart from the fact that the EER can be used to measure the performance of a biometric system, it can also be used to determine the operating point, since EER can be interpreted as the trade off between the FMR and FNMR.

2.3 Forensic science and forensic biometrics

Forensic science can loosely be described as the application of science and technology to law enforcement. In particular, the interpretation and analysis of traces is a central activity. According to [1], there are four distinct inferences: identification, individualisation, association, and reconstruction. They are studied at three levels: source level (origin of the trace), activity level (which activity led to the trace), and the offence level (is the activity an offence).

Biometrics plays a pivotal role in some of these inferences. Applications include scenarios of ID verification and open-set identification, investigation and intelligence, and evaluation of the strength of evidence used in a court of law. The latter is also referred to as forensic evidence evaluation. The collection of these applications form the domain of Forensic Biometrics.

In this dissertation, but also throughout the whole domain of forensic science and forensic biometrics, strength of evidence is commonly represented by the likelihood ratio. This ratio essentially measures the probability of the occurrence of evidence relative to the typicality of occurrence. In which manner this ratio is used is explained in the next section.

2.3.1 Likelihood ratio paradigm: concept

Strength of evidence is commonly expressed as a likelihood ratio in modern forensic science[3]:

$$\mathrm{LR}(E) = \frac{\mathrm{p}(E|\mathscr{H}_\mathrm{s},I)}{\mathrm{p}(E|\mathscr{H}_\mathrm{d},I)}. \tag{2.5}$$

Here E denotes evidence, $\mathscr{H}_\mathrm{s}$ is the same source hypothesis, $\mathscr{H}_\mathrm{d}$ is the different source hypothesis, and I is background information. In the previous section, we used test sample and reference templates to refer to the extracted features during the operational phase and enrollment phase, respectively. Given the forensic context of this dissertation, from now on, we will use trace and reference to refer to these features as it is more common in forensic science. The term trace emphasises the fact that what we first referred to as test sample is often found at or depicts a crime scene. The same source hypothesis states that the trace x and reference y originate from a common donor, the different source hypothesis states the trace x and reference y do not have a common donor. An alternative formulation will be presented later. We also use *same source* to denote genuine scores and *different source* to denote imposter scores in a forensic context.

As described in Jackson et al. [44], the forensic examiner is responsible for the calculation of $\mathrm{LR}(E)$, whereas a court of law determines the prior odds $\frac{\mathrm{p}(\mathscr{H}_\mathrm{s}|I)}{\mathrm{p}(\mathscr{H}_\mathrm{d}|I)}$ and ultimately the posterior odds $\frac{\mathrm{p}(\mathscr{H}_\mathrm{s}|E,I)}{\mathrm{p}(\mathscr{H}_\mathrm{d}|E,I)}$:

$$\frac{\mathrm{p}(\mathscr{H}_\mathrm{s}|E,I)}{\mathrm{p}(\mathscr{H}_\mathrm{d}|E,I)} = \mathrm{LR}(E) \times \frac{\mathrm{p}(\mathscr{H}_\mathrm{s}|I)}{\mathrm{p}(\mathscr{H}_\mathrm{d}|I)}. \tag{2.6}$$

Finally, often (2.5) is used in its $\log_{10}$ form:

$$\mathrm{LLR}(E) = \log_{10}\left(\frac{\mathrm{p}(E|\mathscr{H}_\mathrm{s},I)}{\mathrm{p}(E|\mathscr{H}_\mathrm{d},I)}\right). \tag{2.7}$$

[3] Although Darboux, Appell, and Poincaré suggested its use already in 1906 for the appeal in the Dreyfus case [43], mostly during the last decade it has seen a mainstream acceptance.

The advantage of using (2.7) over (2.5) is the emphasis on the magnitude of the likelihood ratio rather than its exact value.

2.3.2 Likelihood ratio paradigm: implementation

As such, (2.7) is not directly usable. The background information I is case dependent and typically involves auxiliary information like the model of the jacket worn by the perpetrator as seen in trace material. For example, during a forensic workshop held at the Netherlands Forensic Institute in 2016, a forensic case involving a perpetrator wearing a specific green jacket sold in Sweden was given. It helped to limit the group of suspects significantly. We exclude the background information, since we aim to describe a generic, case independent, approach.

Both the features and a biometric comparison score can be considered as evidence.

If the evidence E is the simultaneous occurrence of trace x and reference y, we obtain the *feature based log-likelihood ratio*:

$$\mathrm{LLR}(x,y) = \log_{10}\left(\frac{\mathrm{p}(x,y|\mathscr{H}_\mathrm{s})}{\mathrm{p}(x,y|\mathscr{H}_\mathrm{d})}\right). \tag{2.8}$$

In this dissertation, we employ parametric models for $\mathrm{p}(x,y|\mathscr{H}_\mathrm{s})$ and $\mathrm{p}(x,y|\mathscr{H}_\mathrm{d})$ in (2.8) such that it reverts into an analytic formula. Of course, other approaches exist as well, including for example the use of copula models that relate joint probability distributions to their marginal distributions. We refer to the dissertation of Susyanto [45] in which a copula approach is studied in the context of score fusion.

If the evidence E is a biometric comparison score s computed on a trace x and a reference y, then (2.7) reverts to the *score based log-likelihood ratio*:

$$\mathrm{LLR}(s) = \log_{10}\left(\frac{\mathrm{p}(s|\mathscr{H}_\mathrm{s})}{\mathrm{p}(s|\mathscr{H}_\mathrm{d})}\right). \tag{2.9}$$

Note that the numerator and denominator in (2.9) is equal to p_g and p_i as introduced in (2.4), respectively. As in the previous feature based log-likelihood ratio case, there exist several methods to estimate $\mathrm{p}(s|\mathscr{H}_\mathrm{s}$ and $\mathrm{p}(s|\mathscr{H}_\mathrm{d}$ and consequently the strength of evidence. Examples include the assumption of a parametric model (for example a normal distribution) or non-parametric (for example Parzen window [46]). Another approach is the use of the Pool of Adjacent Violators algorithm [47]. This algorithm creates the convex hull of a ROC curve[4] by estimating $\mathrm{p}(\mathscr{H}_\mathrm{s}|s)$ from which the likelihood $\mathrm{LLR}(s)$ can be derived:

$$\mathrm{LLR}(s) = \mathrm{logit}(\mathrm{p}(\mathscr{H}_\mathrm{s}|s)) - \mathrm{logit}(\mathrm{p}(\mathscr{H}_\mathrm{s})), \tag{2.10}$$

with $\mathrm{logit}(x) = \log_{10}\left(\frac{x}{1-x}\right)$. Note that the prior $\mathrm{p}(\mathscr{H}_\mathrm{s})$ in (2.10) is the fraction of same source pairs in the set, and is not the same as a prior $\mathrm{p}(\mathscr{H}_\mathrm{s})$ set by a court of law. This process is referred to as *score calibration*. Loosely speaking, a score is calibrated as if it is interpretable as a likelihood ratio. A property of a calibrated score is that recalibration yields the same score, or rephrased, the likelihood ratio of a likelihood ratio is the likelihood ratio. There

[4] Actually, it creates the *concave* hull, but we follow the accepted terminology.

exists an interesting relationship between $\mathrm{LR}(\tau) = \frac{\mathrm{p}(\tau|\mathscr{H}_\mathrm{s})}{\mathrm{p}(\tau|\mathscr{H}_\mathrm{d})}$ and the ROC curve that also gives a visual interpretation of calibration. This relationship is

$$\frac{\mathrm{dTMR}}{\mathrm{dFMR}}(\tau) = \mathrm{LR}(\tau), \tag{2.11}$$

that is, the slope of the tangent line at a specific point on the ROC is the likelihood ratio of the corresponding threshold. The proof of (2.11) is straightforward, using the definitions of in (2.4):

$$\begin{aligned}\frac{\mathrm{dTMR}}{\mathrm{dFMR}}(\tau) &= \frac{\mathrm{dTMR}}{\mathrm{d}\tau}(\tau)\left(\frac{\mathrm{dFMR}}{\mathrm{d}\tau}(\tau)\right)^{-1} = \frac{\frac{\mathrm{d}}{\mathrm{d}\tau}\int_\tau^\infty \mathrm{p}(s|\mathscr{H}_\mathrm{s})ds}{\frac{\mathrm{d}}{\mathrm{d}\tau}\int_\tau^\infty \mathrm{p}(s|\mathscr{H}_\mathrm{d})ds} \\ &= \frac{-\frac{\mathrm{d}}{\mathrm{d}\tau}\int_\infty^\tau \mathrm{p}(s|\mathscr{H}_\mathrm{s})ds}{-\frac{\mathrm{d}}{\mathrm{d}\tau}\int_\infty^\tau \mathrm{p}(s|\mathscr{H}_\mathrm{d})ds} = \frac{\mathrm{p}(\tau|\mathscr{H}_\mathrm{s})}{\mathrm{p}(\tau|\mathscr{H}_\mathrm{d})} = \mathrm{LR}(\tau).\end{aligned} \tag{2.12}$$

We can interpret this result as "scores are calibrated when they are equal to the slope of the tangent line of the ROC curve at the threshold they define".

In Section 2.5 we provide two concrete examples of biometric classifiers in the context of this dissertation based either on (2.8) or (2.9).

The final topic of this section is *how* the same source $\mathscr{H}_\mathrm{s}$ and different source $\mathscr{H}_\mathrm{d}$ hypotheses are formulated, as it influences the procedure for training and evaluation of biometric classifiers using either (2.8) or (2.9). We define two distinct formulations. The general formulation is

- $\mathscr{H}_\mathrm{s} = \mathscr{H}_\mathrm{s}^\mathrm{g}$: the trace x and reference y originate from a common donor.
- $\mathscr{H}_\mathrm{d} = \mathscr{H}_\mathrm{d}^\mathrm{g}$: the trace x and reference y do not have a common donor.

The subject based formulation is

- $\mathscr{H}_\mathrm{s} = \mathscr{H}_\mathrm{s}^\mathrm{s}$: the trace x and reference y originate from the same specific donor.
- $\mathscr{H}_\mathrm{d} = \mathscr{H}_\mathrm{d}^\mathrm{s}$: the trace x and reference y do not have the same specific donor.

Since the subject based formulation is tailored towards a specific subject (the suspect), one could argue that the subject based formulation should be favoured over the general formulation. However, the subject based formulation also has a clear drawback related to the general formulation. In the general formulation, training and evaluation use same and different source pairs of a collection of subjects. In the subject based formulation, same source pairs consist of a trace of the specific subject and a reference of the same subject; different source pairs consist of a trace of the specific subject and a reference of another subject. This implies that the number of training and evaluation pairs in the subject based formulation is limited compared to those available in the general formulation; this might hamper the robustness of the training and evaluation of a subject based classifier. Notwithstanding this observation, we use the subject based formulation in Chapters 7 and 8.

2.3.3 Performance: a forensic perspective

In Section 2.2.5 we presented the ROC curve, the AUC, and EER as commonly used measures of performance of biometric classifiers. Although these are important from a forensic perspective, there are actually more performance characteristics with a forensic relevance. According to "A Guideline for the validation of likelihood ratio methods used for forensic evidence evaluation" [48], there exist several primary and secondary performance characteristics and metrics. The primary performance characteristics are

- Accuracy: Closeness of agreement between a likelihood ratio computed by a given method and the ground truth status of the proposition in a decision-theoretical inference model;
- Discriminating Power: Performance property representing the capability of a given method to distinguish amongst forensic comparisons where different propositions are true;
- Calibration: A property of a set of likelihood ratios. Perfect calibrations imply that the likelihood ratio is exactly as big or small as is warranted by the data (...).

Accuracy is measured in terms of the cost of log-likelihood ratio [49]. Given a set $\mathscr{S}$ of n_s same source and a set $\mathscr{D}$ of n_d different source scores under the same source hypothesis $\mathscr{H}_\mathrm{s}$ and the different source hypothesis $\mathscr{H}_\mathrm{d}$) respectively, the cost of log-likelihood ratio is:

$$\mathrm{Cllr} = \frac{1}{2}\left(\frac{1}{n_s}\sum_{s\in\mathscr{S}}\log_2(1+\mathrm{e}^{-s}) + \frac{1}{n_d}\sum_{s\in\mathscr{D}}\log_2(1+\mathrm{e}^{s})\right). \tag{2.13}$$

Accuracy can be interpreted as the combination of discriminating power and calibration. We use ROC, AUC, and EER to explore discriminating power. The Pool of Adjacent Violators algorithm as a calibration method was already presented in Section 2.3.2. Calibration is typically measured in terms of calibration loss and can be calculated as follows. If we apply the PAV algorithm to the set of scores and reapply (2.13), we obtain the minimal achievable cost of likelihood ratio $\mathrm{Cllr}^{\mathrm{min}}$. This quantity is an alternative measure for discriminating power. The difference

$$\mathrm{Cllr}^{\mathrm{cal}} = \mathrm{Cllr} - \mathrm{Cllr}^{\mathrm{min}} \tag{2.14}$$

is calibration loss and it measures how well calibrated the original scores were.

The secondary performance characteristics are

- Robustness: The ability of the method to maintain a performance characteristic when a measurable property in the data changes.
- Coherence: The ability of the method to yield likelihood ratio values with better performance with the increase of intrinsic quantity/quality of the information present in the data.
- Generalisation: Property of a given method to maintain its performance under dataset shift.

According to [48], during the method development the focus is on the primary performance characteristics whereas during the validation stage all performance characteristics are considered. In this dissertation, all but one concrete research activities are confined to one or more primary performance characteristics. One notable exception is Chapter 8 in which the generalisation of a biometric classifier using facial marks is studied. Also, Chapter 8 describes a practical framework that includes these six performance characteristics.

2.4 Forensic Face Recognition as a means to determine strength of evidence: a survey

2.4.1 Abstract

This paper presents a survey of Forensic Face Recognition (FFR), with a particular focus on the strength of evidence as used in a court of law. FFR is the use of biometric face recognition for several applications in forensic science. It includes scenarios of ID verification and open-set identification, investigation and intelligence, and evaluation of the strength of evidence. We present FFR from an operational, tactical and strategic perspective. We discuss criticism on FFR and we provide an overview of research efforts from multiple perspectives that relate to domain of FFR. Finally, we sketch possible future directions of FFR.

2.4.2 Introduction

In this survey paper, we present different aspects of Forensic Face Recognition (FFR), with a particular emphasis on strength of evidence. The aim of this paper is to present the breath of FFR, its many aspects and connections to related domains.

FFR is the use of biometric face recognition for several applications in forensic science. Biometric face recognition uses the face modality as a means to discriminate between human beings; forensic science is the application of science and technology to law enforcement.

In general, FFR includes scenarios of ID verification (1:1) and open-set identification (1:N+1), investigation and intelligence (M:N+1), and evaluation of the strength of evidence as described in Meuwly and Veldhuis [1]. There are two image types involved in FFR. The trace image often captures a crime scene and is most of the time taken under uncontrolled conditions. The reference image is a photograph of a suspect and is taken under controlled conditions. Concrete FFR use cases are given in Zeinstra et al [22].

A use case in which FFR is frequently employed is to investigate criminal activities which are carried out in places monitored by surveillance cameras, like shops or gas stations. Extracted stills from CCTV recordings that contain the face of the perpetrator are used as trace images. Another example is the withdrawal of money using a stolen debit card. In this case trace images are recorded by a small camera in the ATM and they typically exhibit perspective image distortion. These use cases are examples of investigation (M:N+1) or, in the case of a concrete suspect, examples in which the strength of evidence against that suspect is evaluated. Another example is when an immigration officer might be convinced that the used identity document is genuine, but that it does not correspond to the person who is presenting it. If the immigration officer forbids the person to enter, the subsequent investigation is an

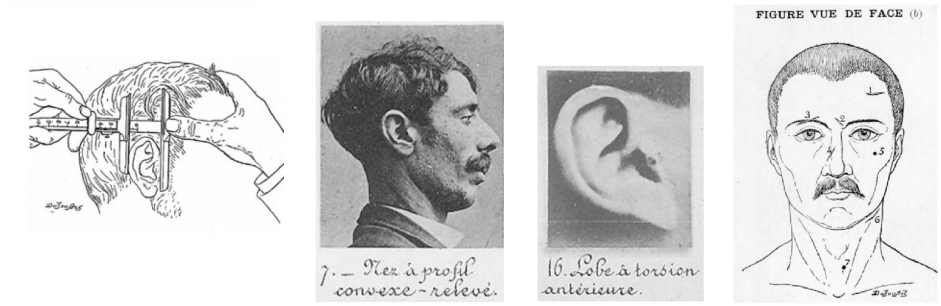


Figure 2.5: Example measurements and categorisation of the Bertillonage system. From left to right: ear size measurement, nose and ear profile categories and the location of scars and marks. Taken from [53].

example of evaluation of strength of evidence in which the passport photograph serves as a trace image.

A survey by Jain et al. [50] discusses additional open-set investigation (M:N+1) use cases: (a) mug shot search that is robust to facial aging , (b) matching forensic (composite) sketches to face photograph databases, and (c) retrieval using facial scars and marks. Case (b) is an example in which trace images consists of a representation (sketch) by an image, instead of a captured image.

A final, very noteworthy but rather extreme, example of an FFR use case (M:N+1) are super recognisers [51] at the Metropolitan Police. Super recognisers are claimed to be able to identify persons from CCTV footage, based on an exceptional memory for discriminating facial features in previously seen low quality images. Super recognisers were used for example during the London Riots of 2011 [52].

FFR has its modern genesis in the Bertillonage system [53]. Bertillonage systematically uses facial and body features to describe criminal individuals. It features anthropometric measurements, as well as categorisations of facial features; for example, it recognises 16 different ear shape types. Also highly discriminating features like facial marks can be described. Figure 2.5 depicts some examples. Bertillon particularly advocates for a mugshot from face and profile, enhancing that the profile contains information that is in the same time more distinctive and less subject to intra variability (ear, upper profile), that has been forgotten in the modern mugshot process. Finn [54] gives a historical account of Bertillonage; in particular the use and acceptance of photography (“the criminal image”) as a means to represent information and evidence. Bertillonage as such has been superseded as a means to individualise persons by fingerprinting (and DNA profiling in the last 25 years) [55]. However, with the proliferation of cameras, either CCTV cameras or digital cameras in mobile phones, potential trace images are omnipresent; it is hardly a surprise that FFR in general is very actively used today.

Despite the advances in automatic face recognition systems, significant parts of FFR are still done manually, notably the evaluation of the strength of evidence. We believe that this can be attributed to several factors that are inherent to FFR and that influence the performance of such automated face recognition systems. Although one can classify these factors into for example subject and imaging conditions, there might be some overlap in this classification.

Therefore we use an example to illustrate the cascading effect of these factors.

Assume a perpetrator uses a stolen debit card to withdraw money from an ATM. He has a hurried glance (expression), does not look straight into the camera (pose), and wears a scarf (occlusion). Natural light (illumination) comes from behind the perpetrator. He stands close to the ATM (perspective effect). The scene is captured through a lens (distortion and aberration) by a recording device, either analog (interlacing) or digital (sensor thermal noise/bleeding). The resulting raw material is then possibly converted (analog to digital, resolution, compression artifacts) and extracted (motion blur). The pose, illumination and expression (PIE) of the person are a common problem for face recognition systems [30].

In the remainder of this paper we will focus mainly on strength of evidence scenario. The paper is structured as follows. In Section 2.4.3 we describe the operational level of FFR structured in terms of the ACE-V protocol. This protocol consists of an Analysis, Comparison, Evaluation and Verification phase and is commonly used for source level inference in forensic science (Langenburg [13] and Finding 22 of Prince [12]). In Section 2.4.4 we present FFR at a tactical and strategic level. During the last decade criticisms on forensic science in general and FFR in particular became more visible at the public level, but they are much older than the last decade within the profession, see [56] for a discussion. In Section 2.4.5 we address some of these criticisms, and in Section 2.4.6 we explore research efforts pertaining to FFR. Finally, in Section 2.4.7 we present our conclusion and sketch future directions for FFR.

As a final note, in this paper we consistently use the general term *FFR-examiner* or *examiner* for short to refer to any individual that undertakes FFR activities, disregarding their level of proficiency; we use *non-examiner* to refer to individuals that do not undertake FFR activities.

2.4.3 Operational level of FFR

Prince [12] gives an overview of the use of facial recognition systems and facial image comparison procedures at several forensic institutions in Israel, The Netherlands, UK, USA, and Canada. Although differences exists, the institutions have a significant collective approach. Spaun [57] specifically describes the approach taken at the Federal Bureau of Investigation and is in line with Prince.

Analysis and comparison

During the Analysis phase, the trace is investigated for its usability and its evidential content. During the Comparison phase the trace and reference images are compared. In this process the examiner takes similarities and dissimilarities between trace and reference material into account. According to Spaun [57], the examiner distinguishes between class and individual characteristics. However, one could argue there are characteristics that are distinctive, contain information to differentiate people and characteristics who are not, and the degree of distinctiveness varies; hence the class/individual distinction is too strict.

FISWG, the Facial Identification Scientific Working Group [7], is a working group in which scientific knowledge and case work experience is organised. According to their Guidelines [6], mainly four FFR methods exist that can be used during the Analysis and Comparison phase:

1. Holistic Comparison: Assessment in which all facial features are considered at once. This is a pure perceptual approach without real analysis.

2. Morphological Analysis: Assessment of correspondence of the shape, appearance, presence and/or location of facial features.

3. Photo-anthropometry: Assessment of correspondence of dimensions and angles of landmarks and other facial features.

4. Superimposition: Assessment in which images are aligned and analysed using image transitions.

FISWG recommends morphological analysis by trained examiners as the primary method of comparison; superimposition is known to be inefficient and should only be used in conjunction with morphological analysis and is to be confined to rotations, scaling and translation.

By definition, in holistic comparison and superimposition trace and reference images are considered simultaneously, so the Analysis and Comparison phase partly overlap. This might lead to bias and is further discussed in Section 2.4.6. In contrast, when applying morphological analysis or photo-anthropometry, trace and reference images do not have to be considered simultaneously.

FISWG has also published the Facial Image Comparison Feature List for Morphological Analysis [8]. This feature list describes for each facial part a number of characteristic descriptors in considerable detail that can be used during forensic case work. As such, it is not a standard but rather representative of considered facial features of different institutes. Feature lists used at the NFI (Netherlands) and NFC (Sweden) are similar but differ on the type and number of considered details [10, 11]. They include

- Face shape
- Forehead
- Eyebrows
- Eyes
- Cheek area
- Nose
- Mouth and mouth area
- Jawline
- Chin
- Scars, Marks, Tattoos

Evaluation and verification

One notable variation between forensic institutions is the used opinion scale (Finding 31, Prince [12]). In an ideal situation this opinion scale is the strength of evidence that is determined during the Evaluation phase. The strength of evidence is constructed from multiple comparisons between features found in trace and reference images. Essentially one could adopt either a numerical or a verbal formulation of strength of evidence. An example of the former is "the trace and reference specimens are 1.0×10^6 more likely under the hypothesis that the suspect is the donor of both, than under the hypothesis that the suspect is not the donor of both". An example of the latter is "there is strong support for the hypothesis that the suspect is the donor of both the trace and reference specimens against the hypothesis the suspect is not the donor of both the trace and reference specimens".

The strength of evidence should be expressed in terms of the (magnitude of the) *likelihood ratio*. Loosely speaking, the likelihood ratio measures the probability of observing *(dis)similarities* between a particular feature found in trace and reference image, scaled by the probability of the *typicality* or *rarity* of observing this feature in trace and reference images in general. The problem with the verbal description is that it can be understood as a verbal scale of posterior probabilities, and it is often the case because of lack of knowledge from both the examiners and the requesters.

An element to take into account is how to *combine* the likelihood ratios of several facial features into *one* likelihood ratio that is reported as the overall strength of evidence. One obvious approach is to assume statistical independence between facial features. In that case the corresponding overall likelihood ratio reverts to the product of the underlying likelihood ratios. However, the independence assumption is most of the time wrong and will be immediately pointed at in court. Approaches that capture or model the dependency structure between features are copula or Bayesian Belief Networks (BBN) and can be used for FFR. A copula describes the relationship between the multivariate distribution of (in our case facial) features and their marginal distributions; Sklar [58] proves that this relationship holds for any multivariate distribution. Another popular approach are BBN (see Barber [59]) in which domain experts model dependency structures. The major advantage of BNN is the significant reduction of the dimensionality of the feature space, making inferences feasible with the limited quantity of data available in a forensic case.

In the verbal formulation the strength of evidence is expressed in a standardised verbal description. The advantage is that a court of law can understand the outcome of the evaluation in principle; however, it must be clear that this evaluation is not a posterior odds and as such it can be easily misunderstood. The verbal approach is either a result mapped from the numerical strength of evidence or a protocol that uses criteria to determine the verbal strength of evidence directly. We refer to "Guidance for Evaluating Levels of Support" [60] and [61] for examples.

During the Verification phase, one or more of the ACE steps are repeated independently in order to reduce the human factor. We refer to Section 2.4.6 for a further discussion on this topic. According to a private communication with an FFR-examiner, the ACE steps can be performed independently by three examiners, after which the final evaluation is determined by a consensus model [10].

The final, verified, evaluation outcome is reported to a court of law. In some cases an FFR-examiner will witness during a court session when additional information or clarification is

needed.

2.4.4 Tactical and strategic levels of FFR

Recommendations and working groups

European institutes are organised in the European Network of Forensic Science Institutes (ENFSI) [62]. This organisation has published a general guideline regarding evaluative reporting [61]. Forensic facial expertise is organised in the Digital Image Working Group (ENFSI-DIWG).

As mentioned previously, the Facial Identification Scientific Working Group (FISWG) [7] is an organisation in which the FBI and several other forensic institutes from the USA and other countries participate. They have published recommendations on facial comparison [6] and which features should be considered during casework [8]. Recent additions (in draft status) are image processing steps for the improvement of automated facial recognition search [63] and the physical stability of adult facial features [64].

Levels of expertise, training, and proficiency tests

Most agencies considered in Prince [12] have three proficiency levels. The *foundation level* means that the examiner has had a basic training and can only do verification (ACE-**V**). The *advanced level* means that the examiner has had more training and experience to do the full ACE-V process. The *expert level* means that the examiner operated at an advanced level and may give an examiner testimony in court. It is remarkable that, at least with respect to the Dutch situation, FFR-examiners are *not* yet registered in a national register of forensic experts [65] that does include already for example forensic psychiatrists.

Training differs per institute but according to [12, 66] may involve

- Knowledge of relevant recommendations.
- Competence in quality assessment (interlacing, codecs, compression, lens distortion, etc.) of trace material.
- Competence in extraction of facial images from CCTV.
- Competence in Adobe Photoshop or similar software.
- Knowledge of and competence in image processing techniques like image enhancement.
- Competence in facial comparison, notably anatomical knowledge and considered facial features in morphological analysis.
- Knowledge of standarised evaluation and reporting.
- Awareness of possible bias and other human errors.
- Competence in statistical concepts, notably Bayesian statistics.
- Basic knowledge of legal aspects and competence in expert testimony.

Apart from initial training to obtain the competences to practise, examiners should participate in proficiency tests on a regular basis [67]. In a recent ENFSI-DIWG Facial Image Comparison Proficiency Test, FFR-examiners had to compare CCTV footage with 10 reference images. For 17 comparisons a single conclusion had to be reported, whereas in one case a full report using the ENFSI evaluative reporting guideline [61] had to be handed in; results have been discussed with peers within the ENFSI-DIWG.

2.4.5 Criticism on FFR

FFR has been criticised for its lack of scientific rigor. According to Evison [55], little research is done on the validation of FFR ("there is no reported error rate (...)" both for human FFR and in assessing the claimed ability of "super recognisers"); as a field it is mostly not scientifically founded yet. This is reiterated by the FISWG Guidelines [6] in which morphological analysis is coined as the primary comparison method, but "only limited studies have been done on accuracy or reproducibility". Only in recent years, some validation studies have appeared and indicate that examiners are better than non-examiners; see Section 2.4.6 for some examples. Formally, human based methods are not validated/accredited on basis of performance but of competence and proficiency; this provides some safeguards but less than a method validated/accredited on basis of performance. We refer to [48] for a full discussion of this topic and to Section 2.4.6 for a discussion on the FFR-examiner as an expert.

Humans are subjective and it is partly mitigated by the verification step in the ACE-V protocol. However, notably the protocol for assigning strength of evidence is subjective (Mallett and Evison, [68]) and the strength of evidence not necessarily represents a likelihood ratio; that is, the strength of evidence.

Edmond et al. [14] contains a complete and very critical review of examiner identification evidence based on trace images. Their study presents several examples of FFR-examiner testimony that illustrate the nonscientific approach and the examiner as the single bearer of absolute truth. One poignant example is "(...) used photo-anthropometry, morphology and photo superimposition to make a positive identification (...). (...) unwilling to disclose her techniques, particularly the points she relied upon (...), because of concerns about her intellectual property rights". Another example is "during cross-examination (...) rejected the suggestion that there was a degree of subjectivity in her assessment, i.e. morphological comparison".

These examples exactly show the lack of fundamental understanding of what inference in forensic science is. According to Saks and Koehler [69] "In normal science, (...) students receive four (..) years of doctoral training where much of the socialisation into the culture of science takes place. This culture emphasises methodological rigor, openness, and cautious interpretation of data. In forensic science, 96% of positions are held by persons with bachelor's degrees (or less), 3% master's degrees, and 1% PhD.s"

The criticism can be placed in the context of the elaborate and critical report [15] on the current state of forensic science in the USA by the National Research Council of the National Academies. One of their recommendations states that "research is needed to address issues of accuracy, reliability, and validity in the forensic science disciplines. (...)".

2.4.6 FFR research directions

In this section we describe several research directions related to FFR. First we discuss human and expert aspects of FFR-examiners. Another branch of FFR research is concerned with the use of anthropometry. Some FFR datasets are available for research purposes. Finally, several studies have considered using more or less distinctive features.

Human aspect of the FFR-examiner

The examiner has a pivotal role in FFR. In O'Toole et al. [70] and [71] the human aspect in FFR is described as being underestimated.

Recent experiments by Papesh and Goldinfer [72] on face matching indicate that under realistic viewing conditions (for example at an airport) infrequently occurring identity mismatches go undetected. Results that relate to trace images taken under uncontrolled conditions are summarised by Sinha et al. [73] as "people can recognise familiar faces in very low-resolution images" and "the ability to tolerate degradations increases with familiarity". In particular, the study by Burton et al. [74] shows that even under severely distorted CCTV footage familiar faces can be recognised, but that does not hold for unfamiliar faces. This might explain why super recognisers have such high success rates, since they "just" recognise a familiar face that they have seen before in other CCTV recordings. In Bruce et al. [75, 76] it is shown that recognition of unfamiliar faces is very error-prone, but this can be claimed of any perceptive intelligence. Megreya and Burton [77] show that there are large individual differences on unfamiliar face matching. A recent study of Gold et al. [78] states that familiarity has a quantitative rather than a qualitative effect on the efficiency with which information is extracted from individual features.

Another well studied negative effect in psychology and forensic science is that of confirmation bias and contextual information. A proper implementation of the ACE-V protocol, with the shield of the examiner from the unnecessary information during the A and C phase helps to limit this effect. An overview by Pronin [79] describes that people can recognise and estimate the operation of bias in human judgment of other persons, except when it is their own bias. The study by Dror et al. [80] show the risks of contextual information and bias with respect to fingerprint examination, which could easily be extended to any other forensic modality, in particular FFR.

Expert aspect of the FFR-examiner

The study of Norell et al. [3] shows that on a set of image pairs examiners reached their conclusions with a significantly lower number of errors than non-examiners. Also if the quality of the trace was lowered, it led to more careful conclusions by examiners, but not for non-examiners. We believe that both findings stem from the fact that the proper methodology is used.

Work with similar findings is White et al [81]. They administered several challenging face matching tests to examiners and non-examiners and concluded that examiners not only outperformed untrained participants, but also computer algorithms, thereby providing the evidence that these examiners are experts at this task.

Zeinstra et al. [17] describes an on line experiment in which examiners and non-examiners participate. Their task is to compare isolated eyebrow pairs using either a "best-effort" ap-

proach versus an approach that uses FISWG characteristic descriptors of the eyebrow. It was found that there are no significant differences in accuracy, however the group of examiners performed significantly better than the non-examiners when they used FISWG.

These results indicate that experts (a) are more aware of fallacies in their judgment and (b) have a better judgment than untrained participants.

Anthropometry

Anthropometry is the science of measuring body or facial dimensions, notably distances and angles. Anthropometry is a key ingredient of the Bertillonage system.

In the dissertation of Kleinberg [82], a series of experiments using locations of anatomically defined facial landmarks is conducted and it is concluded that "using high resolution images to compare video images with photographic images, (...) anthropometry (...) does not generate the results necessary for use as evidence in a court of law".

In a large scale study by Evison and Vorder Bruegge [83] concerning landmark-based analysis of 3D landmarks of more than 3000 persons, it was found that "the 3D distribution of anthropometric landmarks (...) is unlikely to be sufficient to allow for identification of individuals (...)" .

The study by Davis [84] presents a software-assisted photo-anthropometric facial landmark identification system that uses 37 distance and 25 angular measures. Based on a set of 70 subjects adhering to a similar description that "Identification verification was found to be unreliable unless multiple distance and angular measurements from both profile and full-face images were included in an analysis." Here verification refers to ID verification rather than strength of evidence.

Two other studies on statistics of anthropometric measures (one on South African males [85] and one on three European populations [86]) show that although differences might exist between populations, mostly "Matching these rare features on facial photographs will be useful during cases of disputed identification".

We conclude that anthropometry either in 2D or 3D, and either photographs or in vivo, yields in general limited evidential value, unless a rare or extreme valued feature is observed.

Automatic face recognition systems

The last 25 years have seen the development of automated face recognition systems into a mature and active area of research, with some use in FFR [12]. Although some initial work predates it, the Eigenfaces paper [87] can be regarded as the work that successfully sparked a whole new research area. Eigenfaces is an example of a global appearance model. Later methods either use a hybrid (global and local appearances) or a local appearance approach to facial features. The underlying concept is that faces reside in a highly nonlinear manifold of the linear space of images [30], so a linear approach should be locally confined. Local appearance methods can use general feature descriptors like Scale Invariant Feature Transform (SIFT) [4], Local Binary Patterns (LBP) [5], and Histogram of Oriented Gradients (HOG) [88]. By combining multiple regions represented by these features types, a compact representation of the face can be constructed and used.

A recent development in face recognition - and more broadly in artificial intelligence and computer vision - is deep learning, also referred to as deep neural networks or convolutional

neural networks [89]. An archetypal example in which deep learning has shown impressive results is the DeepFace system [90] developed by Facebook, but it is questionable whether the used images are representative of those found in forensic casework. Neural networks are computational structures that contain adaptable parameters. Neural networks as such are not new, their topology was already known and used 30 years ago. Their resurgence is mainly enabled by the availability of (a) massive amounts of training data and (b) sheer parallel computing power provided by Graphical Processor Units (GPU), making the training of the parameters of a deep neural network with many layers feasible. A key difference between these neural networks and other local appearance methods is that they *train* which features are used instead of using features *designed* by a human.

As described in the Introduction, automatic face recognition systems are applied in the FFR domain, but mainly for investigation and intelligence purposes. Additional reasons to rely on human FFR-examiners are the liability and repercussion issues rendered by a misjudgement, irrespectively of it is in favour or against a suspect. According to Prince [12] "facial recognition systems presently lack good integration into forensic facial comparison procedures." Also automated face recognition systems produce a score that (a) is based on abstract features and (b) is a relative measure and does represent the strength of evidence. However, "score calibration" methods convert scores determined by a biometric system into what can be interpreted as strength of evidence, see for example the dissertation of Ali [91] for an overview of several of such methods.

FFR datasets

We believe that one of the factors that hampers FFR research is the low number of publicly available forensically relevant datasets, especially in relation to what is available for face biometrics (either controlled or "in-the-wild"). Also, particularly datasets that contain images from surveillance cameras are limited in the number of subjects. The curious situation is that CCTV are primary designed to monitor the activity of people. But when these activities are recognised as criminal, then the question of the source becomes immediately obvious without that the technology being able to capture the relevant features for the source level inference; this situation exists at least for a decade. Finally, all, except one dataset (ForenFace), lack an elaborate set of forensically relevant annotation.

The SCFace dataset [92] has been used in numerous publications on low resolution face recognition. It contains only frontal surveillance camera images of 130 subjects. The ChokePoint dataset [93] is designed for "person identification/verification under real-world surveillance conditions" and contains 29 subjects. Since it does not contain reference images, it is not suitable enough for research within a forensic context. The NIST [94] (1573 subjects) and Morph [95] (13.618 subjects) datasets contains mugshots, and are very well suited for longitudinal research. The ATVS Forensic DB [96] only contains high resolution mugshots of 50 subjects.

Two recent additions are the Quis-Campi [97] and ForenFace [98] (97 subjects) datasets. The former uses stills from a PTZ camera, showing subjects possibly non-frontal, partly occluded, blurred, or overexposed. The images are representative of modern CCTV cameras; notably higher resolution. A subset of Quis-Campi has been used in an ICB 2016 Challenge [99] on Biometric Recognition in the wild. The unique property of ForenFace is the availability of manual annotation from which a large subset of the FISWG characteristic de-

scriptors can be extracted.

Computational forensic approaches

Face recognition systems use a constellation of abstract features and as such, the outcome of a facial comparison is difficult to understand outside the broader technical domain of computer vision. There also exist approaches in which either more emphasis is laid upon the forensic relevance while still using general feature descriptors or features are used that have a clear forensic semantic meaning.

Examples of the first approach are Tome et al. [100] in which the biometric performance of linear SVM classifiers on 15 forensic facial regions is investigated. This study uses the SCFace and a subset of Morph. They conclude that "... depending on the acquisition distance, the discriminative power of regions change, having in some cases better performance than the full face". Other examples are facial marks. They are interesting from a forensic perspective as they can be very discriminating. They have been the subject of several studies, notably Park et al. [101] and recently Srinivas et al [102]. Related work is that of Lee et al. [103] that uses SIFT descriptors for the description of tattoos for search purposes in mugshot databases.

Examples of the second approach include another work by Tome et al [104]. Here the performance of continuous and discrete soft biometric features are evaluated on the Morph and ATVS Forensic DB datasets. Experimental results show high discrimination power and good recognition performance for some specific cases. However, these cases correspond to relatively good quality images. In some studies all features are extracted manually. Two small studies concerning the eyebrow [18] and the periocular region [20] both show that FISWG characteristic descriptors are comparable to their nonforensic counterparts under good image quality. A much larger study by Zeinstra et al. [22] extends this work and investigates discriminating power of many FISWG characteristic descriptors in four representative FFR use cases in [98]. According to [105] and a forensic guideline [48] currently used as a basis for the part 8 of 19795 ISO standard "Methodology and tools for the validation of biometric methods for forensic evaluation and identification application" under development, discriminating power is one of six aspects that should be taken into account during the validation of a forensic evaluation method. They train and evaluate biometric classifiers that are specialised on single and combined characteristic descriptors. They found that in all but one use case, commercial systems clearly outperform single and combined characteristic descriptors. In the use case with the lowest quality trace images (11px interpupillary distance, severe image compression) they found that (a) the combination of visibility features and (b) the hairline perform better than a commercial system.

Finally, there is another development that can be mentioned. Landmark detection is important for the automatic detection and extraction of certain facial features, especially the shape-like ones. Recent work by Kazemi et al. [106] and Milborrow [107] show that it is now possible to locate to a certain extent landmarks in even uncontrolled scenarios. These results can be used to extract forensically relevant facial features in an automatic manner.

2.4.7 Conclusion and future directions

In this survey, we have presented several aspects of FFR: the historical context, use cases, the three operational levels of FFR, criticism on FFR, and research efforts pertaining to FFR.

We observe several positive developments. Some recent validation studies indicate that the FFR-examiner is "doing better", in particular with respect to the non-examiners. Although anthropometry is closely tied to FFR, especially in the minds of members of the general public, multiple studies reinforce the conclusion that it is limited in its ability to produce meaningful strength of evidence. We recognise the potential of automated face recognition systems as an instrument to help the examiners to assess the strength of evidence and complement the human-based approach. Furthermore, recent advances in fast and accurate automatic detection of facial features could aid the work of the FFR-examiner. Examiners can assess features that are difficult to describe statistically but can only be validated mostly on basis of competence and proficiency and not performance. Automatic approaches use a reduced set of features that can be described statistically but can be validated empirically, extensively and can be improved.

Despite the recent progress, challenges remain.

At a higher, general forensic level, we think the community should (a) better understand the goal of being able to assign/compute the strength of evidence, (b) be able to validate analysis, comparison and interpretation methods, and (c) be able to combine the human and computer-based approaches to generate the most correct strength of evidence. All these goals are not easy to reach.

At the level of FFR, there are other problems that need to be addressed. Since large publicly available forensically relevant datasets are lacking, descriptive statistics of facial features extracted from images representative of forensic use cases are not available. This is important as it would have helped to determine strength of evidence in a more scientific manner. The use of automatic detection of facial features can aid this process. Moreover, current datasets lack the broad variation and use cases needed to systematically investigate the influence of multiple factors found in real forensic casework.

We therefore advocate the collection of a large scale dataset of images grounded in clear forensic use cases, employing forensically relevant parameters. An alternative approach is the development of a large synthetic dataset for the study of the effect of those forensically relevant parameters.

2.5 FISWG characteristic descriptors and FFR classifiers

In this chapter, we have presented the key concepts of biometrics, forensic science and forensic biometrics, in particular the likelihood ratio as the bearer of strength of evidence in a court of law, and many other aspects of Forensic Face Recognition. It has set the stage, and time has come to introduce the two main characters of the play: FISWG characteristic descriptors [8] and FFR classifiers that use the descriptors as their input and produce strength of evidence, either in a direct or indirect manner. The last two sections contain two feature dimension reduction algorithms and a property of likelihood ratio classifiers. Due to their more mathematical nature, these sections can easily be omitted.

2.5.1 FISWG characteristic descriptors

As mentioned before, FISWG characteristic descriptors are facial features that can be used during forensic case work. Although their number may well exceed 250, in this dissertation

we study a large subset of them in a systematic manner. We introduce them visually in Figures 2.6 and 2.7. Figure 2.6a shows the face from a holistic perspective and shows landmark positions that give the rough outline of facial parts. It also shows some large scale structures like the facial shape. Figure 2.6b gives a more detailed perspective and mostly shows different facial lines and potentially discriminating features like facial marks and tattoos. The three subfigures in Figures 2.7 show detailed characteristics of facial parts contained in the upper, middle, and lower parts of the face.

In most of our included work we assume that each descriptor falls exactly in one of four classes:

- Low dimensional: either one or two dimensional and real valued. Examples are the angles of the fissure openings (one dimensional) or the landmark positions of the outer corners of the eyes (two dimensional);
- Availability or visibility: binary indicator. Example is the availability/visibility of the cheekbone;
- Count: non-negative integer. Example is the number of forehead creases;
- Shape: point cloud of two dimensional points. Example is the chin shape.

2.5.2 FFR classifiers

We present two examples of classifiers that use a characteristic descriptor extracted from a trace x and a reference y and produce strength of evidence.

Example 1: Landmark location. We assume a normal distribution (after subtraction of the mean) with same source $\Sigma_s = \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{xy}^T & \Sigma_{yy} \end{pmatrix}$ and different source $\Sigma_d = \begin{pmatrix} \Sigma_{xx} & 0 \\ 0 & \Sigma_{yy} \end{pmatrix}$:

$$\begin{pmatrix} x \\ y \end{pmatrix} | \mathscr{H}_\mathrm{s} \sim \mathscr{N}(0, \Sigma_s) \text{ and } \begin{pmatrix} x \\ y \end{pmatrix} | \mathscr{H}_\mathrm{d} \sim \mathscr{N}(0, \Sigma_d). \tag{2.15}$$

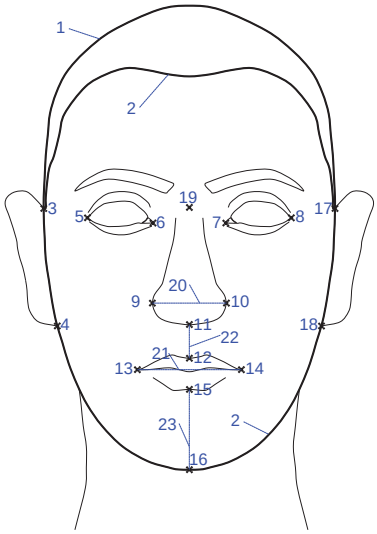
In this case (2.8) has a closed form, with $\Delta = \Sigma_d^{-1} - \Sigma_s^{-1}$:

$$l_N(x,y) = \frac{1}{2}\left(\log|\,\Sigma_d| - \log|\Sigma_s| + (x^T y^T)\Delta \begin{pmatrix} x \\ y \end{pmatrix}\right). \tag{2.16}$$

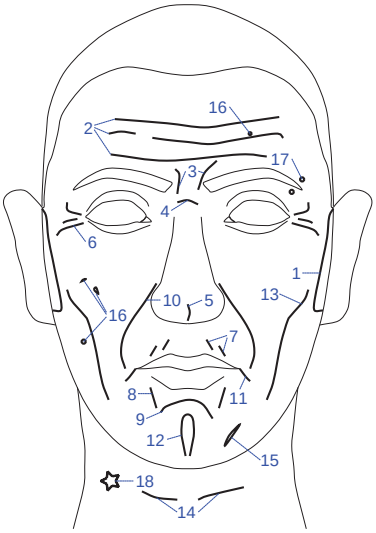
Example 2: Eye fissure shape The eye fissure shape is represented in terms of a point cloud. If $X = \{\mathbf{x}_i \in \mathbb{R}^2 | i = 1, \ldots, N_x\}$ and $Y = \{\mathbf{y}_i \in \mathbb{R}^2 | i = 1, \ldots, N_y\}$, then the shape similarity score function is defined by

$$s_{Shape}(X,Y) = -\frac{1}{N_x}\sum_{i=1}^{N_x} d_{pc}^2(\mathbf{x}_i, Y) - \frac{1}{N_y}\sum_{i=1}^{N_y} d_{pc}^2(\mathbf{y}_i, X), \tag{2.17}$$

where d_{pc} measures the minimal distance between a point $\mathbf{w} \in \mathbb{R}^2$ and a point cloud $Z = \{\mathbf{z}_i \in \mathbb{R}^2 | i = 1, \ldots, N\}$: $d_{pc}(\mathbf{w}, Z) = \min_{i=1,\ldots,N} \|\mathbf{w} - \mathbf{z}_i\|$. Scores are PAV calibrated and converted into a likelihood ratio $\mathrm{LR}(s)$ (2.9), using (2.10).



a) Face from a holistic perspective



b) Face from detailed perspective

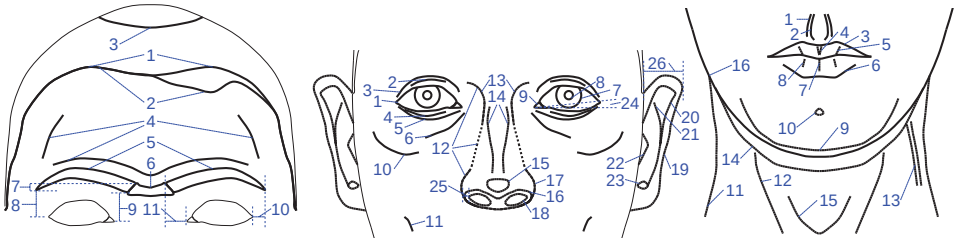
Figure 2.6: Face from a) holistic and b) detailed perspective. Prefixes H and D refer to holistic and detailed perspective, respectively. (H1) Cranial Vault shape/availability, (H2) Facial shape, (H3) location of 17 landmarks (upper/lower connection ears to face (H3, H4, H17, H18), inner/outer corners eyes (H5-H8), nose (H9-H11), mouth (H12-H15), chin (H16), and nasal root (H19)), (H20) width of nose, (H21) width of mouth, (H22) nose-mouth distance, and (H23) mouth-chin distance. (D1) Facial Hair shape/symmetry/availability, (D2) Forehead Creases shape/size/availability/count, (D3) Vertical Glabellar shape/size/availability/count, (D4) Nasion Crease shape/availability/count, (D5) Bifid Nose Crease shape/availability/count, (D6) Periorbital Creases shape/size/availability/count, (D7) Upper Circumoral Striae shape/size/availability/count, (D8) Lower Circumoral Striae shape/size/availability/count, (D9) Mentolabial Sulcus shape/size/availability, (D10) Nasolabial Creases shape/size/availability, (D11) Marionette Lines shape/size/availability, (D12) Cleft Chin shape/size/availability, (D13) Buccal Creases shape/size/availability, (D14) Neck wrinkles shape/size/availability/count, (D15) Scars shape/availability/count, (D16) Facial Marks shape/availability/count, (D17) Piercing shape/availability/count, and (D18) Tattoo shape/availability/count.

2.5.3 Preprocessing: PCA and LDA

Although in general the dimension of considered characteristic descriptors is not very large, we used Principle Component Analysis (PCA) followed by Linear Discriminant Analysis (LDA) as a dimensionality preprocessing step in Zeinstra et al. [18] prior to the application of the classifier. The PCA step compactifies the data while retaining the essential variation in the data; the LDA step is a trade off between enlarging between variation and lowering within variation. Note that PCA and LDA can applied independently. An additional intermediate step is data whitening: under the assumption that the data is normally distributed, this whitening step makes the data independent.

PCA

Suppose X contains training data in a column wise manner: $X = [x_1 \cdots x_m] \in \mathbb{R}^{n \times m}$, and let $\mu_X \in \mathbb{R}^n$ be its mean. Then $X^0 = X - [\mu_X \cdots \mu_X]$ has zero mean. The SVD of X^0 is given



a) Upper facial parts b) Middle facial parts c) Lower facial parts

Figure 2.7: Upper a), middle b), and lower c) parts of the face. Prefixes U, M, and L refer to the upper, middle and lower parts, respectively. (U1) Forehead hairline shape/symmetry/size, (U2) Hair/Forehead boundary shape, (U3) Cranial baldness shape/availability, (U4) Ridge structures shape/availability, (U5) Eyebrows shape/size/symmetry, (U6) Unibrow shape/availability, (U7-U11) Relative positions A-E. The relative positions are measured on both eyebrows. (M1) Fissure shape/size/symmetry, (M2) Upper Folds shape/availability/count, (M3) Superior Palpebral Furrow shape/availability, (M4) Lower Folds shape/availability/count, (M5) Inferior Palpebral Furrow shape/availability, (M6) Infraorbital Furrow shape/availability, (M7) Iris shape, (M8) Pupil shape, (M9) Caruncle shape, (M10) Cheekbone shape/availability, (M11) Dimple Cheek shape/availability, (M12) Nose shape/size/symmetry, (M13) Nasal Root shape/size, (M14) Nasal Body shape/size/symmetry, (M15) Nasal Tip shape/symmetry, (M16) Nasal Base size/deviation, (M17) Alae shape, (M18) Nostrils shape/size/symmetry, (M19) Outer Helix shape/symmetry/size, (M20) Inner Helix shape/size, (M21) Anti-Helix shape/size, (M22) Tragus shape/size, (M23) Anti-Tragus shape/size, (M24) Fissure angle, (M25) Nostril thickness, and (M26) Ear Protrusion. (L1) Philtrum Ridges width/symmetry, (L2) Philtrum Furrow width/symmetry, (L3) Upper Lip shape/symmetry, (L4) Upper Lip Tubercle shape, (L5) Upper Lip Creases shape, (L6) Lower Lip Outline shape/symmetry, (L7) Lower Lip Median Sulcus shape, (L8) Lower Lip Creases shape, (L9) Chin shape/size/symmetry, (L10) Chin Dimple shape/availability, (L11) Neck Boundaries size, (L12) Musculature shape/availability, (L13) Veins shape/availability, (L14) Double chin shape/availability, (L15) Laryngeal shape/size/availability, (L16) Jawline shape.

by $X^0 = U\Sigma V^T$, where $U \in \mathbb{R}^{n\times n}$ and $V \in \mathbb{R}^{m\times m}$ are orthonormal matrices and $\Sigma \in \mathbb{R}^{n\times m}$ is a diagonal matrix containing the singular values $\sigma_1 \geq \cdots \geq \sigma_r > 0$, with $r = \text{rank}(X_0)$, on its diagonal. By selecting the first $p \leq r$ columns of U and V, denoted by $U_p \in \mathbb{R}^{n\times p}$ and $V_p \in \mathbb{R}^{m\times p}$, and the corresponding $\Sigma_p \in \mathbb{R}^{p\times p}$ sub matrix of Σ, it can be shown that the reduced matrix $X_p = U_p\Sigma_p V_p^T \in \mathbb{R}^{n\times m}$ is the closest (in terms of the sum of squared column differences or equivalently the Frobenius norm) to the matrix X^0 of all matrices having rank p. Hence, X_p can be seen as a good approximation that captures the variation in the data well. Define the transformation $T_p = \sqrt{m-1}\Sigma_p^{-1}U_p^T \in \mathbb{R}^{p\times n}$ and $Y_p = T_pX_p \in \mathbb{R}^{p\times m}$, then

$$\text{Cov}(Y_p) = T_p\text{Cov}(X_p)T_p^T = \frac{m-1}{m-1}\Sigma_p^{-1}U_p^TU_p\Sigma_pV_p^TV_p\Sigma_p^TU_p^TU_p\Sigma_p^{-T} = I_p.$$

Therefore, under the assumption that the training and test data have the same distribution, the transformation T_p (a) reduces the dimension of the training and test data retaining as much as variation possible, and (b) whitens it.

LDA

Each data point y_i contained in Y_p belongs to some class $c \in \{1,\ldots,C\}$, so the dataset can be partitioned into C different subsets $\mathscr{C}_c$, each containing m_c data points having mean μ_c. For

the full dataset we have $\sum_{c=1}^{C} m_c = m$ and $\mu = \frac{1}{m}\sum_{i=1}^{m} y_i = \frac{1}{m}\sum_{c=1}^{C} m_c \mu_c$. The next dimensionality reduction step, LDA, projects these data points to a subspace such that the variation within classes is low and variation between classes is high. Formulated differently, given a dimension $l < p$, we search for an orthonormal basis $W_l \in \mathbb{R}^{p\times l}$ of a subspace (or equivalently, a projection W_l^T onto that subspace) such that the projected data $Z_l = W_l^T Y_p \in \mathbb{R}^{l\times m}$ has the desired low within and high between variation. The projection W_l^T is found by maximising the Rayleigh quotient:

$$W_l = \text{argmax}_{W\in\mathbb{R}^{p\times l}} \frac{|W_l^T S_B W_l|}{|W_l^T S_W W_l|}. \tag{2.18}$$

Here the matrix S_B is the between scatter matrix and S_W is the within scatter matrix defined by

$$S_B = \sum_{c=1}^{C} m_c(\mu_c - \mu)(\mu_c - \mu)^T,\ S_W = \sum_{c=1}^{C}\sum_{y\in\mathscr{C}_c} (y-\mu_c)(y-\mu_c)^T. \tag{2.19}$$

Since any scalar multiple of W_l would yield the same value for the Rayleigh quotient, the maximisation of (2.18) can be rephrased as a Lagrangian problem: maximise $|W_l^T S_B W_l|$ under the assumption that $|W_l^T S_W W_l| = 1$. This is equivalent to finding l eigenvectors w_i of the generalised eigenvalue problem:

$$S_B w_i = \lambda_i S_W w_i \tag{2.20}$$

such that the eigenvalues λ_i are as large as possible. It can be shown that (a) these l eigenvectors w_i exist, (b) span an l dimensional space, and (c) the eigenvalues λ_i are non-negative. The total scatter matrix S_T is defined as $S_T = (m-1)\text{Cov}(Y_p) = \sum_{i=1}^{m}(y_i-\mu)(y_i-\mu)^T = (m-1)I_p$, and it is straightforward to show that $S_T = S_B + S_W$. Hence, (2.20) can be rewritten as

$$S_B w_i = ((m-1)I_p - S_W)w_i = \lambda_i S_W w_i \Leftrightarrow S_W w_i = \frac{m-1}{\lambda_i + 1} w_i. \tag{2.21}$$

Therefore, finding the l eigenvectors w_i corresponding to the l *smallest* eigenvalues of S_W yields the projection W_l^T. This solution is not unique, any other orthonormal basis of the space spanned by the columns of W_l is also a solution.

A common assumption is that every class shares the same covariance and only differs in the mean μ_c; in that case S_W can be written as

$$S_W = \sum_{c=1}^{C}\sum_{y\in\mathscr{C}_c} (y-\mu_c)(y-\mu_c)^T = (m-C)\text{Cov}(Y_p). \tag{2.22}$$

2.5.4 Neyman Pearson Lemma

In this final section we state a well known theorem (or rather lemma) that emphasises an important property of classifiers that produce a likelihood ratio.

The Neyman-Pearson lemma states that given an $x \in \mathbb{R}^p$, when deciding between two hypotheses $x \in \mathscr{H}_0$ (negative class) or $x \in \mathscr{H}_1$ (positive class), a positive decision that uses a threshold T on the likelihood ratio

$$\text{LR}(x) = \frac{\text{p}(x|\mathscr{H}_1)}{\text{p}(x|\mathscr{H}_0)} > T \tag{2.23}$$

is optimal in the sense that it gives the maximum TMR, given fixed FMR (determined by T) or minimum FMR, given fixed TMR (determined by T).

2.6 Chapter conclusion

In this chapter, we have presented some of the key ingredients of biometrics and forensic science; the field of Forensic Face Recognition from various perspectives; the FISWG characteristic descriptors and we have provided examples of biometric classifiers that use these descriptors to produce strength of evidence. We did not address any research question in this chapter.

As mentioned in the survey (Section 2.4), there have been some positive results with respect to the performance of FFR-examiners. Also, the potential of automated face recognitions was recognised; however, we expected that future forensic case work still should and will involve FFR-examiners.

In the remainder of this dissertation, we will investigate the use of biometric classifiers in a forensic context, trained on FISWG characteristic descriptors, producing strength of evidence in the manner presented in Section 2.5, and evaluated on images representative of forensic use cases. The purpose of such an investigation is twofold: it gives insight into the suitability of such classifiers and the FISWG characteristic descriptors upon which they act. Such an investigation tries to address the criticism on FFR as put forward in Section 2.4, notably the aim for scientific rigor; and it is in line with the essence of the Daubert rule discussed in the previous chapter.

Prior to the exploration of such classifiers, we first remain close to current practice as we consider the human performance on an eyebrow verification experiment that involves characteristic descriptors. It makes sense to include the human as the results are indicative of the added value of characteristic descriptors. This study is presented in Chapter 3.

Chapter 3

Human performance on an eyebrow verification task

3.1 Introduction

This chapter addresses research question 1a: *Under relatively well-conditioned settings, what is the performance of FFR-examiners in relation to non-examiners, both using FISWG characteristic descriptors and a best-effort approach in a verification task?*

It includes one study that investigates the performance of humans in an eyebrow comparison task. In this work, performance differences between (a) FFR-examiners and non-examiners and (b) FISWG characteristic descriptors and "best-effort" approaches are studied.

Section 3.2 has been published as "Examining the examiners: an online eyebrow verification experiment inspired by FISWG" [17].

Reading Guide
Section 3.2. The Abstract and Introduction can be omitted as it is mostly based on previously presented material.

3.2 Examining the examiners: an on line eyebrow verification experiment inspired by FISWG

3.2.1 Abstract

In forensic face comparison, one of the features taken into account are the eyebrows. In this paper, we investigate human performance on an eyebrow verification task. This task is executed twice by participants: a "best-effort" approach and an approach using FISWG characteristic descriptors: forensically relevant facial features. The group of participants is divided into FFR-examiners and non-examiners. The rationale behind this experiment is to determine whether there exist differences between (a) FFR-examiners and non-examiners and (b) the "best-effort" and a forensic approach. It is shown empirically on a specially

constructed dataset that there do not exist major differences between FFR-examiners and non-examiners, however, non-examiners do perform worse when using the forensic approach instead of the "best-effort" approach.

3.2.2 Introduction

FFR-examiners compare crime scene images and reference images taken from a suspect and formulate a (descriptive) estimation of the strength of evidence that the images depict the same person. A judge can incorporate the strength of evidence in the verdict whether the suspect is considered guilty or not. The comparison protocol typically involves the assessment of (dis)similarities found during a morphological analysis; its details may vary between forensic organisations [12]. Since the comparison process is to some extent subjective, insight into decision making and efforts to objectification are important. FISWG [7], a scientific working group in which facial identification knowledge and experience is organised, has published recommendations for different levels of the comparison process. Their facial comparison list [8] describes overall and detailed FISWG characteristic descriptors. Apart from using this list as a mnemonic during case work, it can also be seen as a complete set of descriptors of the facial region that are considered to be important for the comparison process. We adopt the latter interpretation, since one can argue that using these descriptors (a) makes the comparison process more transparent, (b) might lead to consistent results among examiners, and (c) helps to (semi-)automate the comparison process.

The aim of this paper is twofold. We compare the performance of a "best-effort" approach to a "quantified-FISWG" approach in a verification setting. Moreover, we compare the performance of examiners and non-examiners. In the "best-effort" approach, best means in the capacity of being a non-examiner or an FFR-examiner, without using the FISWG characteristic descriptors explicitly. We limit ourselves to the eyebrow modality in this paper. The choice for this modality is motivated by the facts that (a) eyebrows are relatively often observable in real crime scene images when other parts of the face are occluded and (b) previous studies have shown that the eyebrows can be exploited as a soft biometric modality. However, it should be noted that this modality is also subject to change, for example due to eyebrow plucking, aging or medical conditions. The work presented in this paper is part of a project in which we aim to semi-automate forensic facial comparisons based on other modalities as well and compare its performance to that of humans.

This paper is organised as follows. In Section 3.2.3 we review related work. Section 3.2.4 discusses the FISWG characteristic descriptors used in this experiment. Section 3.2.5 describes the experimental setup. Section 3.2.6 contains the results and discussion, where as in Section 3.2.7 the conclusions are given. Finally, Section 3.2.8 describes planned future work on extended versions of this experiment and similar research on different facial features.

3.2.3 Related work

In previous research, the eyebrow is shown to be a rich container of information. This is true for humans [108], as well for automated systems. For example [109] reports "compared with the full face, the eyebrow region has a drop of $\frac{5}{6}$ in size, but only a $\frac{1}{6}$ drop in rank-1 identification". This and some other studies [110, 111] use spatial (LBP), frequency (Fourier/DCT), and hybrid spatial/frequency (LBP after Fourier/DCT or wavelet) information

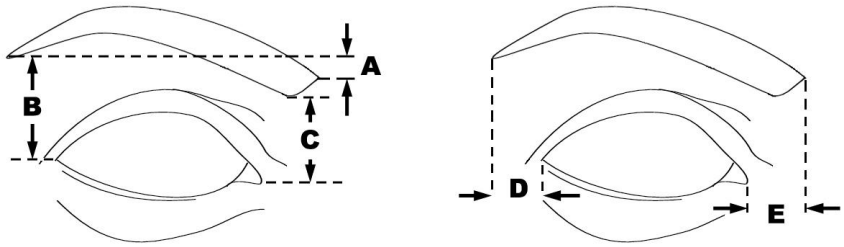


Figure 3.1: A-E characteristic descriptors of the eyebrow, taken from [8].

in their feature descriptors. In a study by [19], only general shape characteristics are used as a feature descriptor. Reference [18] is the first study that uses features from a forensic viewpoint and searches for optimal combinations of FISWG characteristic descriptors and Dong Woodard [19] features. It is found that FISWG and Dong Woodard perform equally well. A related biometric modality is the periocular region which roughly can be defined as the area around the eye, including the eye, and possibly including the eyebrow. Initially conceived as a soft biometric complementing low quality iris images, it has gained attention on its own right. Compared to the eyebrow research, work on this modality has utilised many other feature descriptors such as SIFT [112] and advanced versions of feature descriptors such as 3P-LBP, and hierarchical 3P-LBP [113] as well. Also, some papers compare human [114] and machine performance [115]. In other papers as [100], the parts-based approach to face recognition is used. This paper extends the work of [18] by comparing FFR-examiner and non-examiner performances.

3.2.4 Quantification of FISWG characteristic descriptors

The set of FISWG characteristic descriptors of the eyebrow [8] contains qualitative and quantitative elements. The advantage of a quantitative over a qualitative feature is (a) its unambiguous definition, (b) an almost unambiguous application, (c) the ability to use match/non-match statistics, (d) possible semi-automation of detection/verification, and (e) transparency and repeatability. However, qualitative features are not used without a reason: they utilise the power and flexibility of human recognition, sometimes yielding better performance, but always worse transparency. Since this work is part of a larger framework in which the feasibility of a (semi)-automated system that can be used by forensic examiners is investigated, we choose to quantify the FISWG features as much as possible. This enables future performance comparisons between human and semi-automated systems.

The set of FISWG characteristic descriptors of the eyebrow essentially consists of four feature clusters: a shape description (SH), relative bounding box size (BB), five specific relative positions (AE), and a description of the hair distribution throughout the eyebrow (HD). We refer to [8] and [18] for more details. In an automated system, (SH) can be represented by a 2D Fourier Shape Descriptor. This representation is not suitable for human comparison, therefore a related low frequency reconstruction is used visually to capture the main proper-

Figure 3.2: An eyebrow pair taken from the dataset.

ties of the shape. (BB) is the smallest box that encompasses the eyebrow. The (AE) features are shown in Figure 3.1. Both (BB) and (AE) are measured relative to the size of the eye. To determine (HD), the eyebrow is segmented into 4 equiangular sectors, emanating from the midpoint between the inner and outer eye corner. For each sector, the relative number of hair (=nonskin) pixels within the eyebrow is determined. The skin detection problem has many, often heuristic, solutions, based on membership of colour space subsets [116]. In this paper, a subset of the YCrCb colour space is used, after histogram equalisation in the Y channel.

3.2.5 Experimental setup

The main goal of the experiment is to investigate the human ability to determine whether two eyebrow images are the same. The underlying goals are to investigate whether (a) differences exists between FFR-examiners and non-examiners and (b) differences exist between using a "best-effort" versus a "quantified-FISWG" approach.

Dataset

In order to assess the performance of the comparison process under realistic circumstances, we created a dataset of 100 eyebrow pairs (50 match pairs, 50 non-match pairs) that contains a number of hard cases. Using a steered random selection of actors and actresses having a Wikipedia article, a set of 200 facial images of somewhat varying quality (variation in resolution, illumination, pose, expression, and age) was manually selected and rectified such that the eyes are horizontally aligned. Each individual is used once in a pair. In each image, the eye corners are manually landmarked and the eyebrow is manually segmented. The segmentation is randomly altered into a somewhat larger segmentation. This is a trade off between having inherent shape information (eyebrow segmentation) and having skin patch information of the periocular region (rectangular eyebrow region segmentation) during the comparison. The larger segmentation is applied to the facial images, and scaled such that the eye corner distance is 100 pixels. The eye corner positions are marked by coloured crosses. An eyebrow pair is shown in Figure 3.2.

Experimental design and interfaces

The main task of the experiment is to compare two eyebrow images, state whether they are from the same person or not, and indicate the level of confidence in the judgment. The ex-

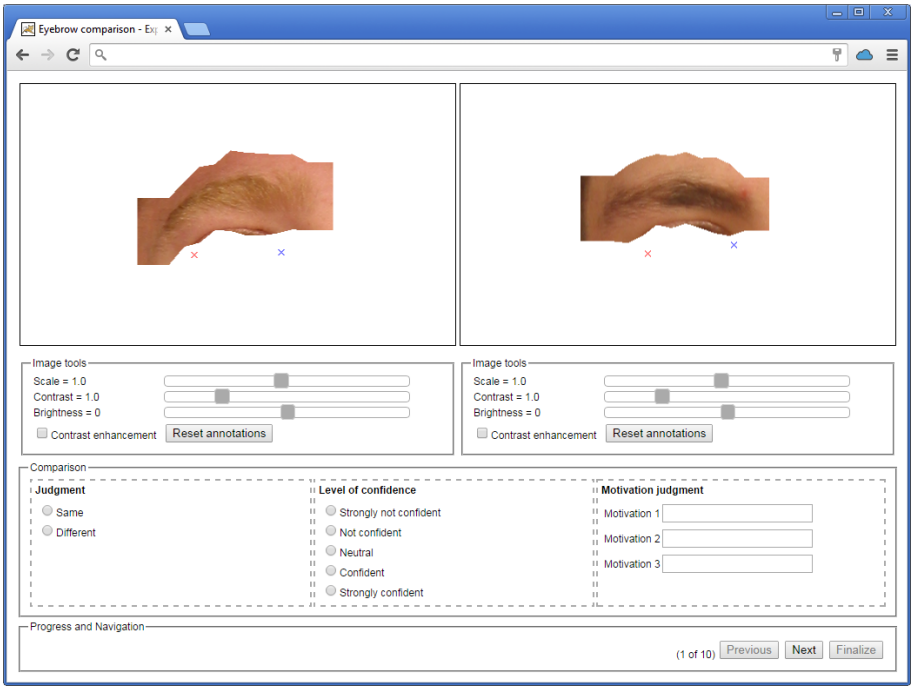


Figure 3.3: Interface Experiment A.

periment is divided into two components. In Experiment A, a best-effort approach has to be taken, whereas in Experiment B FISWG characteristic descriptors must be used. The experiment is offered as an on line application, so participants are not restricted by time and location constraints. A user guide and match/non-match statistics for FISWG features on a separate dataset is provided. Both experiments use the same dataset, and the order and position in the interface in which the eyebrow pairs are presented to participants is random. In Figures 3.3 and 3.4, the interfaces of Experiment A and B are shown. The interfaces are similar: the eyebrow pairs are shown, a small set of image processing tools (scaling, contrast, and brightness correction) is provided, and participants are free to navigate through the eyebrow pairs during the experiment. Although in both experiments the judgment and level of confidence must be stated, there are differences in the comparison. In Experiment A at least one motivation for the judgment must be provided. The comparison procedure in Experiment B is more elaborate. Participants are required to draw an outline of the eyebrow shapes, after which quantitative FISWG characteristic descriptors are shown. Based on the values and a visual inspection, the participant is required to state whether the FISWG characteristic descriptors are the same, and whether using the FISWG characteristic descriptors changed the judgment at the beginning of this comparison. Participants are able to practice with a fully functional demo version of the application in which the dataset is replaced by an easier set.

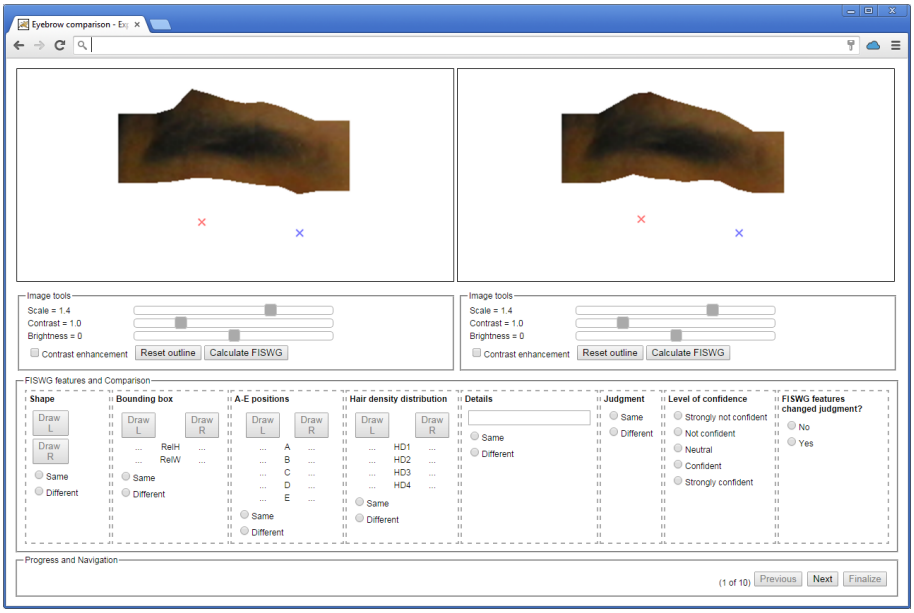


Figure 3.4: Interface Experiment B.

Table 3.1: Number of statistically significant changes on experiments A and B, F=FAR, T=TAR.

	F≈/T↓	F↑/T≈	F↑/T↑	F↓/T≈
Examiner	1	0	1	1
Non-Examiner	4	2	1	0

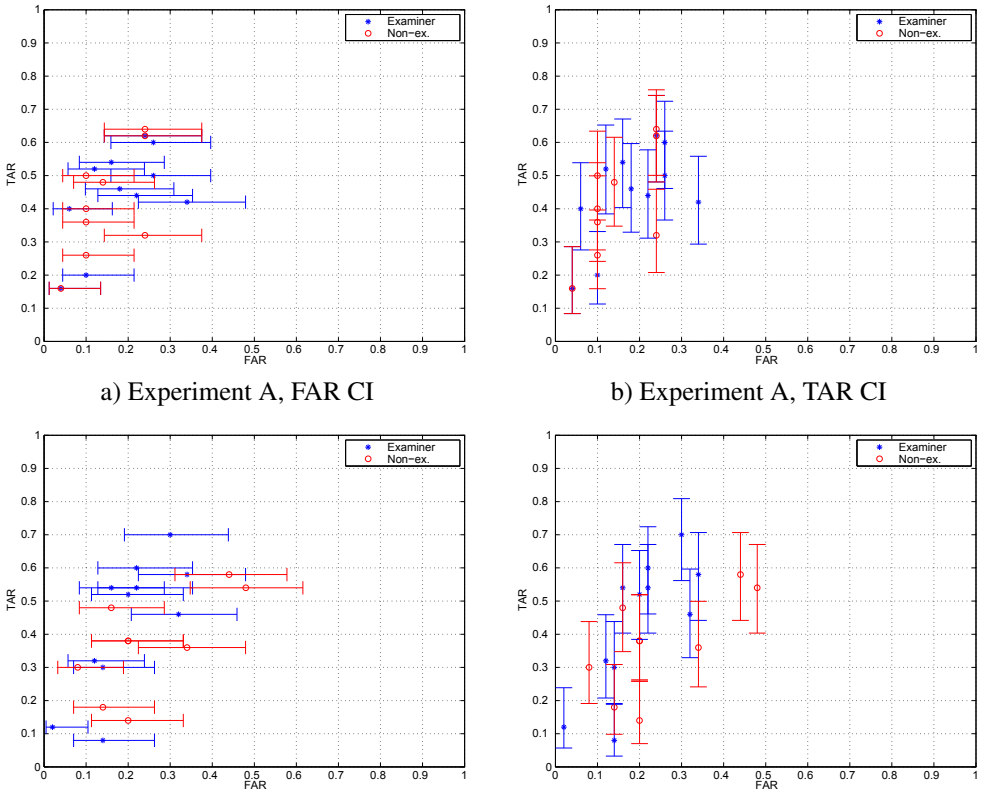
Recruitment of participants

The participants of this experiment have been recruited in two ways. The experiment was announced through a network of biometric and forensic examiners interested in the FISWG standarisation. In total 11 FFR-examiners agreed to participate. The non-examiners were recruited from a student group and other people affiliated with the authors. In total 9 non-examiners participated. It was estimated that Experiment A would take 60 to 90 minutes, were as Experiment B might take up to twice as much time. This large time investment especially explains the lower number of non-examiner participants.

3.2.6 Experimental results and discussion

Individual performance

In this section, we report and discuss the performance of individual participants in terms of FAR, TAR and accuracy. In Figure 3.5, the (FAR,TAR) points of the participants are drawn in ROC space. By interpreting the number of FP and TP cases as binomial random variables



a) Experiment A, FAR CI

b) Experiment A, TAR CI

c) Experiment B, FAR CI

d) Experiment B, TAR CI

Figure 3.5: Performance at individual level with 95% credible interval (CI).

and assuming a non-informative uniform prior on the values of FAR and TAR, we can derive a 95% credible interval on the (FAR, TAR) values. A credible interval is the Bayesian equivalent of a frequentist confidence interval [117]. To avoid cluttered images, the intervals for FAR and TAR are drawn in separate graphs. As is visually clear from Figure 3.5, there does not exist a clear overall distinction between the (FAR, TAR) values between FFR-examiners and non-examiners. It seems that the (FAR, TAR) values tend to spread more in Experiment B. Statistically significant individual changes (95% credible intervals on FAR and TAR separately) are summarised in Table 3.1. These changes indicate that 6 (out of 9) non-examiners perform worse on Experiment B, whereas only one (out of 11) FFR-examiner performs less. Only one examiner performs better on Experiment B. Also, the results indicate that some participants do not perform better than a randomly guessing classifier. Indeed, by using a 95% credible interval around the main diagonal in ROC space, Table 3.2 shows that especially the number of non-examiners "guessing" during Experiment B increases. In Table 3.3, the accuracies on Experiments A and B are shown. The difference in accuracy of non-examiners between Experiments A and B is significant ($p < 0.01$). This change might be caused by the fact that there is a difference in prior knowledge of FISWG between FFR-

Table 3.2: Number of participants "guessing" on experiments A and B.

	Examiner	Non-Examiner
Experiment A	2 (out of 11)	1 (out of 9)
Experiment B	1 (out of 11)	4 (out of 9)

Table 3.3: Accuracy on experiments A and B.

	Average	Stddev.
Examiner Experiment A	0.631	0.060
Examiner Experiment B	0.617	0.071
Non-Examiner Experiment A	0.635	0.062
Non-Examiner Experiment B	0.561	0.059
Combined Experiment A	0.633	0.059
Combined Experiment B	0.592	0.070

examiners and non-examiners, despite an explanation in the user guide. This decline in accuracy also causes the difference in accuracy of the combination of examiners and non-examiners to change significantly ($p < 0.01$). The differences between FFR-examiners and non-examiners on Experiments A and B are not significant. For each judgment, participants also stated their level of confidence. One would expect that the accuracy is high given the highest level of confidence: especially an FFR-examiner should not make mistakes when the stated level of confidence is high. In Table 3.4, the accuracy results are only given at the highest level of confidence. There are no significant differences between Experiments A and B of FFR-examiners, non-examiners and the combined group. Also, there are no significant differences between FFR-examiners and non-examiners on experiments A and B. Please note the larger standard deviations compared to Table 3.3. This is probably caused by different internal thresholds for the category "very confident", yielding a mixture of a few maximum accuracies and "average" accuracies.

Group performance

In this section, we focus on the performance of non-examiners and FFR-examiners acting as a group. There exists a strong correlation (Table 3.5) between the number of aggregated correct judgments for each comparison between Experiments A and B, and the FFR-examiners and non-examiners. This indicates that participants overall make the same kind of judgments.

Typically, during casework, members of a group of FFR-examiners perform the comparison task independently from each other. A consensus or majority model is used to arrive at a conclusion. This approach can be simulated. By thresholding the number of positive individual judgments, we can find (a) the number of votes that induces the highest accuracy (Table 3.6) and (b) compare performance in terms of the ROC curve, instead of on (FAR, TAR) points in ROC space. The shape of the ROC curves in Figure 3.6 confirm the results of Table 3.4 and it seems that examiners maintain their performance better than non-examiners. However, the sudden drop in TAR when FAR < 0.1 for the examiner group on Experiment B

Table 3.4: Accuracy given the highest confidence level on experiments A and B.

	Average	Stddev.
Examiner Experiment A	0.815	0.198
Examiner Experiment B	0.822	0.207
Non-Examiner Experiment A	0.743	0.154
Non-Examiner Experiment B	0.706	0.166
Combined Experiment A	0.781	0.177
Combined Experiment B	0.767	0.193

Table 3.5: Correlation between aggregated correct judgments of examiners, non-examiners, and combined.

	Correlation	Between
Exp. A	0.86	Examiner/Non-Examiner
Exp. B	0.84	Examiner/Non-Examiner
Examiner	0.85	Exp. A/B
Non-Examiner	0.78	Exp. A/B
Combined	0.89	Exp. A/B

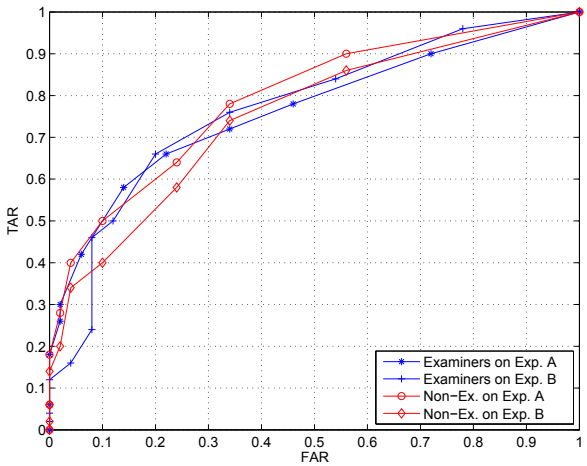


Figure 3.6: Performance as group.

is for FAR=0.04 and FAR=0.08 statistically significant.

3.2.7 Conclusion

In this study, we show that on this particular dataset there exists no differences in individual performance between examiners and non-examiners, when asked to compare "best-effort" (Experiment A) and "quantified-FIWSG" (Experiment B). However, the non-examiners per-

Table 3.6: Optimal vote threshold for positive judgment, accuracy and (FAR, TAR).

	Votes	Accuracy	FAR	TAR
Examiners Exp. A	5 (out of 11)	0.72	0.14	0.58
Examiners Exp. B	4 (out of 11)	0.73	0.20	0.66
Non-Examiners Exp. A	5 (out of 9)	0.72	0.34	0.78
Non-Examiners Exp. B	3 (out of 9)	0.61	0.24	0.58

formance in terms of (FAR, TAR) and accuracy decreases significantly on Experiment B. Also, the number of non-examiners performing not better than a random classifier increases on Experiment B. This suggests that some non-examiners have difficulty in applying the FISWG characteristic descriptors in their decision process, maybe caused by a lack of prior knowledge. The usage of FISWG characteristic descriptors does not induce better accuracy results given the highest confidence level. Finally, when voting as a group, FFR-examiners maintain their performance better than non-examiners, but suffer from a large TAR decline in the FAR region < 0.1 in Experiment B.

3.2.8 Future work

The work presented in this paper will be continued at several levels. Firstly, outlines drawn by participants can be used as an input to a semi-automatic classifier. The performance of such an classifier can be compared to the human performance. Secondly, we can extend this "best-effort" versus FISWG approach to other forensically relevant modalities like the periocular region or scars, marks, and tattoos in the facial region.

3.2.9 Acknowledgments

The authors would like to thank Arnout Ruifrok (Netherlands Forensic Institute) for fruitful discussions, valuable feedback, and the opportunity to use his network of examiners during this project. Participants are thanked for their valuable time investment.

3.3 Chapter conclusion

In this chapter, one study has been presented. It addressed research question 1a: *Under relatively well-conditioned settings, what is the performance of FFR-examiners in relation to non-examiners, both using FISWG characteristic descriptors and a best-effort approach in a verification task?*

The included study considered the human performance on an eyebrow comparison task. It was shown on a particular, prepared dataset that the performance measured in accuracy was relatively poor in general. Its average ranged between 0.56 and 0.64. Also, there did not exist significant differences in individual performance between FRR-examiners and non-examiners when asked to compare "best-effort" (Experiment A) and "quantified-FIWSG" (Experiment B). However, the non-examiner performance in terms of (FAR, TAR) and accuracy decreased

significantly on Experiment B. Also, the number of non-examiners performing not better than a random classifier increased from 1 to 4 out of 9 on Experiment B.

With respect to research question 1a, we find that this study in essence shows that non-examiners should not use characteristic descriptors and the FFR-examiners do not perform better or worse when using characteristic descriptors. The former conclusion does not have any real repercussion, however the second conclusion can lead to the questioning of their added value. This question is further explored in Chapter 4, where classifiers instead of humans are the main object of study.

Chapter 4

Classifier performance on the periocular region

4.1 Introduction

This chapter addresses research question 1b: *Under relatively well-conditioned settings, what is the general performance of biometric classifiers that use FISWG characteristic descriptors as their input and produce strength of evidence in relation to other non-forensic biometric classifiers?*

It combines two studies on the periocular region, that is, the region of the eye and eyebrow. The first study compares biometric classifiers that use the eyebrow FISWG characteristic descriptors as their input to classifiers that utilise non-forensic features that have been introduced by a study of Dong and Woodard [19]. The second study is a small scale study that compares classifiers using FISWG characteristic descriptors of the eye to a non-forensic texture based approach commonly used in periocular biometrics.

Section 4.2 has been published as "Towards the automation of forensic facial individualisation: Comparing forensic to non-forensic eyebrow features" [18].

Section 4.3 has been published as "Beyond the eye of the beholder: on a forensic descriptor of the eye region" [20].

Reading Guide
Section 4.2. The Abstract, Introduction and Related Work can be omitted as it is based on previously presented material.
Section 4.3. The Abstract and Introduction can be omitted as it is mostly based on previously presented material.

4.2 Towards the automation of forensic facial individualisation: comparing forensic to non-forensic eyebrow features

4.2.1 Abstract

The Facial Identification Scientific Working Group (FISWG) publishes recommendations regarding one-to-one facial comparisons. At this moment, a draft version of a facial image comparison feature list for morphological analysis has been published. This feature list is based on casework experience by FFR-examiners. This paper investigates whether the performance of the FISWG characteristic descriptors (forensically relevant facial features) of the eyebrow can be considered as being "state-of-the-art". We compare the recognition performance of one particular state-of-the-art non-forensic eyebrow feature set to a semi-automated version of the FISWG characteristic descriptor of the eyebrow. The recognition performance is measured in terms of the forensically relevant log-likelihood-ratio cost metric Cllr. It is shown there exists a collection of feature sets that have similar performance.

4.2.2 Introduction

When comparing a facial image from a crime scene with a police photograph, FFR-examiners pay attention to morphological-anthropological features, following a prescribed one to one facial comparison protocol. For example, at the Netherlands Forensic Institute (NFI), a list of facial feature comparisons may be independently scored by three examiners. A consensus model is used to arrive at a verbal description of the strength of evidence that the crime scene image and the police photograph have the same origin. A judge combines this description with other evidence to arrive at a verdict.

This approach has some acknowledged issues such as latent examiner bias and inter examiner differences. Automating this process might mitigate the impact of these issues. Also, the comparison protocol is not standarised between law enforcement agencies. The Facial Identification Scientific Working Group (FISWG) publishes recommendations regarding one-to-one facial comparisons. A draft version of a facial image comparison feature list for morphological analysis [118] has been published by this organisation. Although the FISWG list can be regarded as a mnemonic tool for the forensic facial examiner, it is also possible to interpret it as a definition of facial features. This paves the way for (semi-)automation of the facial comparison process.

The FISWG feature list is based on case work experience by FFR-examiners. In this paper, we evaluate the recognition performance of the FISWG eyebrow modality in a semi-automatic setting. To our knowledge, this is the first work to evaluate a FISWG feature description. The choice for the eyebrow modality is additionally motivated by the recent attention from the biometric community for soft biometric modalities in general and the eyebrow in particular. This makes a comparison with a non-forensic feature set possible. Also, whether a more optimal feature set can be found by combining non-forensic with forensic features will be investigated.

4.2.3 Related work

Some studies have shown that the eyebrow is a compact and rich container of information, both for humans [108] and for automatic recognition [109]. Early work of [119] based on a Hidden Markov Model reports recognition rates of 92.6% on a set of 54 high quality images. [110] automatically segments eyebrows and uses a Euclidian distance measure to compare contours of eyebrows. On a set of 200 high quality images a recognition rate of 88.1% is reported. The work of [109] is the first to use a substantial dataset (FRGCv2 Experiment 4 protocol) [33]. LBP is applied on spatial and frequency transformed images of the eyebrow strip. In general, around a 10-20% TAR is reported at 1% FAR, depending on parameter settings and frequency representations. At first glance this might not seem impressive, but "compared with the full face, the eyebrow region has a drop of $\frac{5}{6}$ in size, but only a $\frac{1}{6}$ drop in rank-1 identification". [19] selects shape-based eyebrow features for biometric recognition and gender classification. On a subset of the FRGCv2 dataset a rank-1 recognition rate of approximately 75% on the eyebrow is achieved. [120] combines dimensionality reduction techniques with a Radon transform and reports a recognition rate of approximately 87% on the high quality BJUT dataset [121]. [111] uses cross correlation for eyebrow detection and transforms the region of interest into the frequency domain. Recognition rates vary between 96.4% and 98.6% on the BJUT dataset, depending on parameter settings and distance measures.

Although most of the reported performances are impressive, they were obtained using good quality images in which individual hairs can be recognised. This is not representative of the forensic situation where the quality (visibility, pose, illumination, expression, resolution) of the trace material is in general less than the reference material. Under these limiting circumstances, the Dong Woodard feature set [19] can be considered as "state-of-the-art". Moreover, it contains features that could, in principle, be determined by an FFR-examiner.

4.2.4 Methods

Dong Woodard feature set

The Dong Woodard feature set [19] contains three feature clusters: global (GL), local (LO) and critical (CR). The global cluster contains three general shape measures: rectangularity, eccentricity and isoperimetrical quotient. A bounding box is divided into four equal horizontally

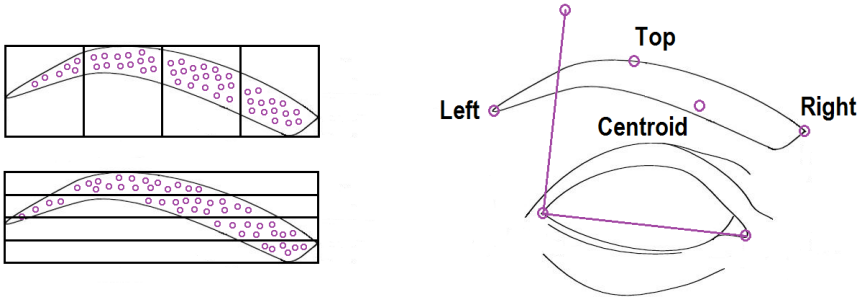


Figure 4.1: The local (left) and critical (right) features of the Dong Woodard feature set.

and vertically adjacent subregions. The local feature consists of the relative percentage of eyebrow area in these eight boxes. The critical features are the coordinates of the left, right, top and centroid point of the shape, expressed in a local coordinate system relative to the eye corners. The local and critical features are shown in Figure 4.1.

FISWG characteristic descriptorsfeature set

In essence, the FISWG characteristic descriptors for the eyebrow [118] consists of four feature clusters: a shape description (SH), a relative bounding box size (BB), five specific relative distances (AE) and a description of hair distribution throughout the eyebrow (HD). The shape description and the hair distribution are formulated in a qualitative manner, implying the need for a quantitative interpretation of these features. We experiment with different implementations of these features.

Shape Initial experiments indicate that the 2D Fourier Shape Descriptor yields the most promising recognition results. This descriptor interprets the n points of the shape as a periodic signal in $\mathbb{C}$. Suppose $c_0, \cdots, c_{n-1}$ are its Fourier coefficients, then the k dimensional Fourier Descriptor is given by $(|\frac{c_2}{c_1}|, \cdots, |\frac{c_{k+1}}{c_1}|)$. This shape descriptor is invariant under translation, rotation, and scaling [122]. Based on additional experiments, we choose equidistant sampling of $n = 512$ points on the original shape and the subsequent Fourier Descriptor representation on $k = 15$ coefficients.

Bounding box and A-E measures The second and third feature cluster have an anthropometric nature. The bounding box size (BB) is measured relative to the eye size. In our implementation, the horizontal distance between the inner and outer eye corner is used. Furthermore, five special measures (A-E) are shown in Figure 4.2. In our implementation, these five measures are measured relative to the size of the eye.

Hair distribution The eyebrow is segmented into 4 equiangular sectors, emanating from the midpoint between the inner and outer eye corner. For each sector, the relative number of hair pixels within the eyebrow is determined. A pixel is considered to be hair if the probability being a skin colour falls below a threshold. This probability is determined empirically in the same image on a skin patch above the eyebrow. A hue saturation bin of size 64×128 with a threshold of 0.01 is chosen.

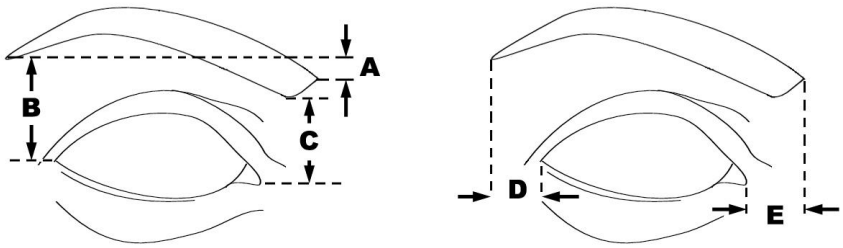


Figure 4.2: A-E characteristic descriptors of the eyebrow, taken from [118].

Likelihood ratio paradigm

The task of the FFR-examiner is to estimate strength of evidence in term of a likelihood ratio. Trace material from a crime scene (e.g. CCTV still image) and reference material (e.g. frontal image of suspect) form the basis for two hypotheses: the same source hypothesis $\mathscr{H}_\mathrm{s}$ (”trace and reference originate from a common donor”) and the defence hypothesis $\mathscr{H}_\mathrm{d}$ (”trace and reference do not have a common donor”). Given the evidence E, the forensic examiner estimates the likelihood ratio $\mathrm{LR}(E) = \frac{\mathrm{p}(E|\mathscr{H}_\mathrm{s})}{\mathrm{p}(E|\mathscr{H}_\mathrm{d})}$. Based on prior odds $\frac{\mathrm{p}(\mathscr{H}_\mathrm{s})}{\mathrm{p}(\mathscr{H}_\mathrm{d})}$ and the likelihood ratio value $\mathrm{LR}(E)$, the judge uses the posterior odds $\frac{\mathrm{p}(\mathscr{H}_\mathrm{s}|E)}{\mathrm{p}(\mathscr{H}_\mathrm{d}|E)}$ to arrive at a verdict.

Likelihood ratio calculation in a (semi-)automatic setting

To determine $\mathrm{LR}(E)$ in a (semi-)automatic setting, a score function $s(\cdot,\cdot)$ is applied on a training set containing pairs of feature vectors whose labels are known. This yields $\mathrm{p}(s|\mathscr{H}_\mathrm{s})$ (“same source”) and $\mathrm{p}(s|\mathscr{H}_\mathrm{d})$ (“different source”). Given a score value s^* from the case at hand, $\mathrm{LR}(s^*) = \frac{\mathrm{p}(s^*|\mathscr{H}_\mathrm{s})}{\mathrm{p}(s^*|\mathscr{H}_\mathrm{d})}$ is interpreted as $\mathrm{LR}(E)$. We adopt the approach from [123] where the score function $s(\cdot,\cdot)$ is directly modeled as a log-likelihood ratio:

$$s(x_1,x_2) = -\frac{1}{2}\log(|\Lambda|) + \frac{1}{2}(x_1^T x_2^T)(I-\Lambda^{-1})\begin{pmatrix}x_1\\x_2\end{pmatrix},$$

with $x_1,\ x_2 \in \mathbb{R}^k$. We assume that the feature vectors have zero mean and unit variance and individuals share a diagonal within variance $\Lambda \in \mathbb{R}^{k\times k}$. Given a score value s^* from the case at hand, we now may interpret this as an estimate for $\log(\mathrm{LR}(E))$.

Training, testing, and PAV calibration phase

The score function only acts on whitened data and requires a value for Λ. The training phase takes care of this. We sketch the procedure given in [123]. Given a training set $X = [X_1 \cdots X_n] \in \mathbb{R}^{m\times n}$, we subtract the mean μ_X from all feature vectors in the training set. Next, we select two dimensionality reduction parameters p and l, $m \geq p > l \geq 1$. The transformation $M \in \mathbb{R}^{l\times m}$ is a composition of a PCA projection from m onto p dimensions, whiting, individual mean subtraction, and an LDA projection from p onto l dimensions. The within variation Λ is estimated from the transformed data $Y = M(X-\mu_X)$.

During the testing phase μ_X, M, and Λ are known. The query X_q and target X_{tar} datasets are transformed into $Y_q = M(X_q-\mu_X)$ and $Y_{tar} = M(X_{tar}-\mu_X)$, after which the log-likelihood ratio score function is applied. Since we use small datasets, it can be beneficial to calculate the optimal classifier belonging to the convex hull of the ROC by means of the Pool of Adjacent Violaters (PAV) algorithm [47]. Moreover, the PAV algorithm also converts scores into log-likelihood ratios [124], a process known as calibration. The output of the testing phase is a calibrated same source score set $\mathscr{S}$ and a calibrated different source score set $\mathscr{D}$.

The Cllr performance measure

Cllr is a measure that captures both the discriminative power of a classifier and how well the scores are calibrated [49]. Since we use calibrated scores, it will solely measure the

discriminative power. It is defined as

$$\mathrm{Cllr} = \frac{1}{2}\left(\frac{1}{|\mathscr{S}|}\sum_{s\in\mathscr{S}}\log_2(1+\mathrm{e}^{-s}) + \frac{1}{|\mathscr{D}|}\sum_{d\in\mathscr{D}}\log_2(1+\mathrm{e}^{d})\right),$$

where $\mathscr{S}$ and $\mathscr{D}$ are the same source and different source score sets, respectively.

4.2.5 Experimental setup and results

Dataset and preprocessing

We select three datasets for our experiments. The first set, denoted by Sel1, consists of 500 images from 125 distinct persons taken from a selection of the FRGCv2 dataset. Each person is represented by two good quality and two lesser quality images. The second set, denoted by Sel2, consists of 400 good quality images from 100 distinct persons again taken from another selection of the FRGCv2 dataset. The final set is a subset of the high quality PUT [125] dataset, approximately 2200 images from 100 distinct persons. In every dataset, the right and left eyebrow are manually segmented after which the Dong Woodard and FISWG features are automatically determined.

Experiments

We conduct two experiments. The purpose of Experiment 1 is twofold. First, we measure the recognition performance of the separate feature clusters of FISWG. Next, we search for a small collection of feature cluster sets that have a promising recognition performance. By varying all possible dimensionality reduction parameters p and l a set of 37472 classifiers is obtained. Experiment 1 uses a 5 fold cross validation scheme and is repeated six times (3 datasets, left/right eyebrows).

Experiment 2 builds upon the first experiment. The purpose of Experiment 2 is to assess the performance of the Dong Woodard, FISWG and a small collection of promising feature cluster sets. We train in total 3093 classifiers using these feature combinations on the Sel2 dataset and test the recognition performance on the Sel1 and PUT datasets. This experiment is repeated twice (left/right eyebrows).

Results Experiment 1

In this experiment, the performance of the separate feature clusters of FISWG is measured. Also, we search for promising feature cluster sets. The best classifiers on a given feature set are shown in Figure 4.3. For the purpose of comparison, the classifiers using the Dong Woodard and FISWG feature sets are also provided. In general, the results on right and left eyebrows are consistent within a dataset. On the Sel1 and Sel2 datasets, the recognition performance of the underlying feature clusters is in decreasing order `AE-SH-BB-HD`, on the PUT dataset `SH-AE-HD-BB`. Two differences are noteworthy. The `AE-SH` difference might be explained by the difference of detailed variation in the original eyebrow shapes. The improved performance of the hair feature on the PUT dataset is explained by a higher quality in terms of resolution and illumination, yielding a clearer distinction between hair and skin pixels.

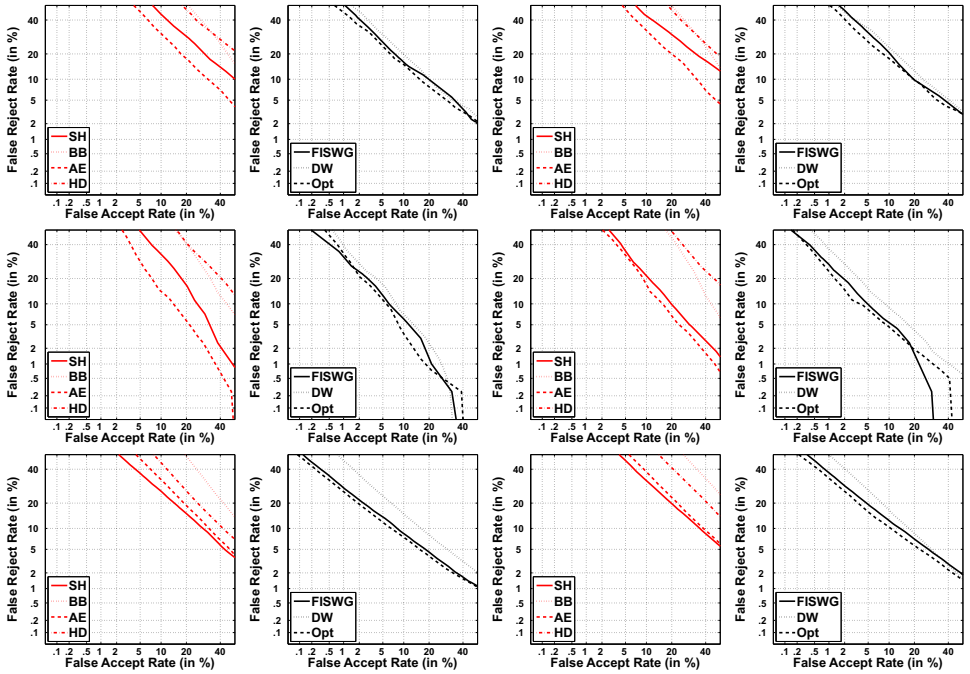


Figure 4.3: DET curves for Experiment 1. The columns from left to right are FISWG clusters right, Dong Woodard/FISWG/Optimal features right, FISWG clusters left, Dong Woodard/FISWG/Optimal features left; the rows from top to bottom are Sel1, Sel2, and PUT.

On the Sel1 dataset, the optimal classifier operates on the feature set {AE, SH, CR}. On the other two datasets, the feature set on which the optimal classifier operates differs between the right and left eyebrow. On the Sel2 dataset, the best classifier on the right eyebrow is the same as on Sel1. The set {HD, AE, SH, CR} is optimal for the left eyebrow of Sel2 and for the right eyebrow of PUT. Finally, the set {HD, AE, SH, LO} is optimal for the left eyebrow of PUT. This indicates that there does not exist a unique optimal feature set but rather a small collection of optimal feature sets.

When comparing the Dong Woodard and FISWG feature set performances in Figure 4.3, only on the PUT dataset there seems to be a consistent difference in favour of the FISWG feature set. As mentioned earlier, the FISWG feature set uses texture information, so it is expected to perform better than the Dong Woodard feature set on good quality eyebrow images.

Results Experiment 2

In this experiment, a limited set of classifiers are trained on the Sel1 dataset and tested on the Sel2 and PUT datasets. In Figure 4.4, the best classifiers on the Dong Woodard, FISWG and optimal feature cluster set are shown. The performance of the Dong Woodard and FISWG feature sets are comparable. Also, the performance of the optimal feature cluster set is not significantly better than these feature sets.

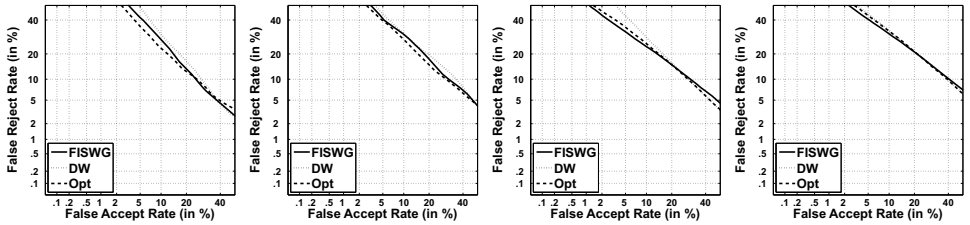


Figure 4.4: DET curves for Experiment 2. From left to right: Sel1/Right, Sel1/Left, PUT/Right, PUT/Left.

4.2.6 Conclusions and future work

In our study, we have implemented the FISWG eyebrow feature and investigated its performance. The components of FISWG ordered in increasing performance are `{AE, SH}` and `{BB, HD}`, the order within the sets depends on the used dataset. Our study shows that the performance of the FISWG feature set is comparable to the Dong Woodard feature set, in terms of the Cllr performance measure. Also, the performance of optimal feature cluster sets do not differ significantly from the FISWG feature set, emphasising the existence of a small collection of good feature cluster sets.

For future work, we intend to measure the performance of FFR-examiners and compare their performance with our semi-automatic system.

4.3 Beyond the eye of the beholder: on a forensic descriptor of the eye region

4.3.1 Abstract

The task of an FFR-examiner is to assess the likelihood ratio whether a suspect is depicted on crime scene images. They typically (a) do a morphological analysis when comparing parts of the facial region, and (b) combine partial evidence into a final judgment. Facial parts can be considered as soft biometric modalities and have been studied by the biometric community in recent years. In this paper, we focus on the region around the eye from a forensic perspective by applying the FISWG feature list to the eye modality. We compare our approach with existing work on a texture descriptor that represents the soft biometric perspective.

4.3.2 Introduction

The biometric community traditionally has focused on highly discriminative modalities such as the fingerprint and the iris. The discriminative property of fingerprints and DNA trace material is utilised in forensic case work for (a) inclusion/exclusion of suspects and (b) assessment of the strength of evidence whether a suspect is the source of the trace material. However, in practice, trace material might only consist of for example CCTV footage. In that case, FFR-examiners can use a morphological analysis on parts of the facial region and combine the outcome of this analysis into a final assessment whether a suspect is depicted

on the crime scene images. The Facial Identification Scientific Working Group (FISWG) [7] has published several recommendations for this comparison process, including a one-to-one checklist [8] that summarises properties of facial parts. Irrespectively of the used facial comparison procedure, its core feature is that it combines a multitude of so-called soft biometric modalities instead of one highly discriminating biometric modality. In recent years, soft biometric modalities have been studied extensively in the biometric community.

This paper focuses on the eye region and compares the forensic approach as described in [8] to a texture based approach found in biometric literature on the periocular region. A formal anatomical definition of the periocular region does not exist. Often the area around the eye (possibly including the eyebrow and the eyeball) is meant. The goal of this paper is to assess (a) the feasibility of FISWG eye features for verification purposes and (b) how their performance relates to existing texture based feature representation. Although real forensic casework typically involves low quality trace material, the feasibility assessment is done on a limited subset of the FRGCv2 dataset. Therefore, this investigation is preliminary and its results should be considered indicative. Annotation of forensic characteristic details can be an elaborate process, but is not unrealistic in a forensic setting.

This paper is organised as follows. In Section 4.3.3, we discuss a selection of related work on the periocular region and describe the FISWG eye features. The methodology is discussed in Section 4.3.4, whereas in Section 4.3.5 the data preparation, FISWG descriptor extraction, and experiments are described. Section 4.3.6 discusses the results and finally in Section 4.3.7, conclusions are drawn.

4.3.3 Related work

Before the emergence of the periocular region as a soft biometric modality, the meticulously detailed eye region model [126] was introduced as a generative model that is "capable of detailed analysis (...) in terms of the position of the iris, (...) eyelid opening, and the shape, complexity, and texture of the eyelids." Mainstream interest in the periocular region as a soft biometric modality was sparked by [112]. This paper combines a local (SIFT) and global approach (HOG/LBP texture description on an array of image patches) into a periocular feature set. Tests were conducted on a specially constructed dataset of 30 subjects and approximately 900 images and on a subset of FRGCv2 consisting of 1704 facial images. It was found that manually selected periocular regions that include the eyebrow area and the eyeball give the highest rank-one accuracy rates. Subsequent research has mainly focused on performance under nonideal conditions, alternative texture descriptors and recognition by humans. For example, [127, 128] investigate the performance of uniform LBP (ULBP) and the influence of image quality on the performance, [129] uses LBP after a (frequency) transformation of the periocular region, and [130] uses the GIST descriptor. Advanced versions of LBP like 3P-LBP, and hierarchical 3P-LBP [113] are also utilised, yielding a rank-one accuracy of 98% on the challenging Notre Dame twins dataset [131]. The studies on identifying useful recognition features [114] and the performance of human recognition [115] are particularly interesting as they give insight into what clues humans use during their recognition process. In this paper, we choose the ULBP texture descriptor as a representative of texture descriptors that work well under ideal conditions.

The FISWG description of the eye contains an extensive list of characteristic features. In our work we have identified a large subset of these features. Some of the features have been

Table 4.1: Sublist FISWG characteristic eye components and their descriptors. R/L denotes right/left eye. The prefix in the enumeration refers to (D)erived or (A)nnotated Characteristic Descriptors.

Component characteristic	Characteristic Descriptors
Inter-eye distance	(D1) Distance R/L eye
R/L Fissure Opening	(A1) Shape
	(D2) Angle
R/L Upper Eyelid	(A2) Superior palperal fold
	(A3) Folds
	(A4) Epicanthic fold
	(A5) Lashes
R/L Lower Eyelid	(A6) Lashes
	(A7) Folds
	(A8) Inferior palperal fold
	(A9) Infraorbital furrow
R/L Sclera	(A10) Blood
	(A11) Defects
	(D3) Colour
R/L Iris	(A12) Shape
	(D4) Position, diameter
	(D5) Colour
	(A13) Shape pupil
	(D6) Pupil pos., diameter
R/L Medial canthus	(A14) Shape caruncle
	(D7) Angle inner eye
R/L Lateral canthus	(D8) Angle outer eye

dropped, because they overlap with other features or follow implicitly from other features. The subset is shown in Table 4.1. Each feature or *characteristic descriptor* is either annotated or derived from annotation. We refer to [8] for the complete list.

4.3.4 Methods

Each annotated characteristic descriptor (A1)-(A14) listed in Table 4.1 is represented by a 2D point cloud.

These point clouds use the same coordinate system in which the right and left medial canthi are mapped to (-1, 0) and (1,0) respectively, rectifying the face representation. This is advantageous for the calculation of the derived characteristic descriptors since some of them mandate a rectified face representation. Some of the point clouds represent a shape, while other designate noticeable artifacts. Although parametric or more general shape descriptors such as Fourier Descriptors can principally be used for the former case (see for example our work [18] on the eyebrow modality), initial experiments yielded unsatisfactory results. We instead adopt an appearance based approach for all the annotated characteristic descriptors. An example of these appearance based features is shown in Figure 4.5. Instead of directly

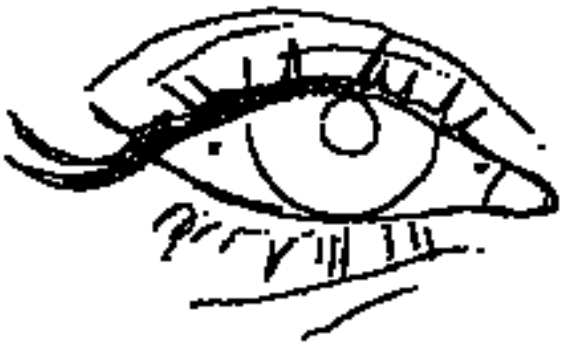


Figure 4.5: Example appearance based features.

Table 4.2: Representation of non-appearance features.

Non-appearance feature	Representation
(D1) Distance R/L eye	$\mathbb{R}_+$
(D2) Angle eye	$[0,360]$ (°)
(D3) Colour sclera	$[0,255]^3$ (RGB)
(D4) Position, diameter iris	$\mathbb{R}^2 \times \mathbb{R}_+$
(D5) Colour Iris	$[0,255]^3$ (RGB)
(D6) Position, diameter pupil	$\mathbb{R}^2 \times \mathbb{R}_+$
(D7) Angle inner eye	$[0,360]$ (°)
(D8) Angle outer eye	$[0,360]$ (°)

superimposing the point clouds of two images, we define a clipping region $[-4,0] \times [-2,2]$ for the right eye and $[0,4] \times [-2,2]$ for the left eye. By using a 2D binary bin of size 45×45 for each eye region, every annotated point is assigned to a bin. These values are found empirically: the bin size is a trade-off between precision and robustness. In order to determine the influence of the 14 constituent annotated characteristic descriptors the same approach is used.

The derived characteristic descriptors form the non-appearance based features and their representation is shown in Table 4.2. In order to compare the verification performance with existing periocular literature, we use two highly related ULBP approaches. In the first approach we use the original 7×5 grid arranged around the iris as described in [112] for 35 ULBP histograms of 59 bins each. The advantage is that we can compare a basic version of the original descriptor with our approach. The disadvantage is that it contains the eyebrow area, an area with characteristic features that are not taken into account in our approach. We therefore also use a version that operates on a region of the same size, but shifted downwards with 1.5 bins above and 3.5 bins below the vertical position of the iris center. The original images are gray scaled and rectified based on their medial canthi using bicubic interpolation before ULBP is applied.

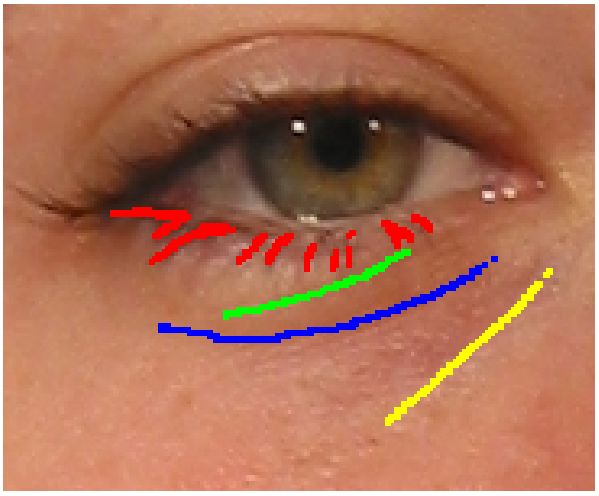
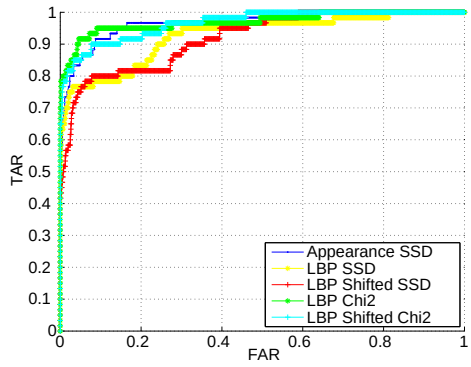


Figure 4.6: Example annotation lower eyelid. Annotation from top to bottom are lashes, folds, inferior palperal fold, and infraorbital furrow.

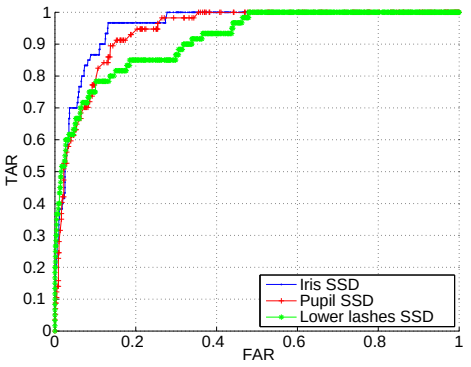
4.3.5 Experiments

Data preparation and descriptor extraction

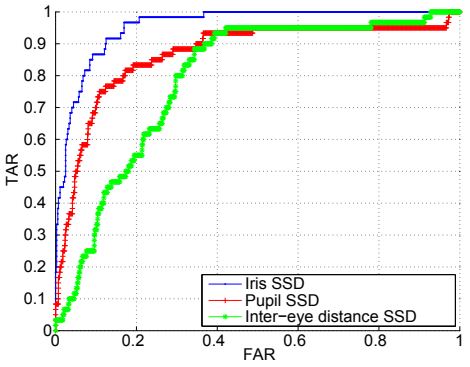
For this paper, we randomly selected a subset of the FRGCv2 dataset consisting of 10 persons, 4 images per person from the Spring 2003 session. All images were taken under conditioned illumination and neutral expression. A dedicated annotation tool has been developed in Java. The characteristic descriptors (A1)-(A14) can be annotated in this tool. The user can select and zoom in on parts of the face. An example annotation is shown is Figure 4.6. The user also manually selects six landmarks on the face: (1) right earlobe connection to the head, (2) right lateral canthus, (3) right medial canthus, (4) left medial canthus, (5) left lateral canthus, and (6) left earlobe connection to the head. The annotation is stored as a collection of points together with the characteristic descriptor type. The remaining characteristic descriptors (D1)-(D8) are calculated based on the landmarks and annotation. The width of the face is estimated by the intra earlobe landmark distance. The inter-eye distance (D1) is calculated as the distance between the right and left medial canthus in terms of the width of the face. The angle of the eyes (D2) is estimated from the medial and lateral canthi positions. The position and diameter of the iris (D4) are calculated by minimising the geometric distance [132] to the annotation (A12), the estimation of (D6) is similarly based on (A13). The iris colour (D5) is determined by averaging the colours of the pixels that lie (a) in the fissure opening (A1), (b) in the estimated iris (D4), and (c) outside the estimated pupil (D6). An additional automatic post processing step removes bright artifacts in the iris caused by studio lighting. The colour of the sclera is also determined by averaging the colours of pixels that lie (a) inside the fissure opening (A1), (b) outside the estimated iris (D5), and (c) outside the caruncle (A14). The inner (D7) angle is estimated by the following procedure: the fissure shape is partitioned into two sets that lie above and below the line segment between the medial and lateral canthus. From those two sets, the points that lie outside the proximity of the medial canthus (radius



a) App. vs LBP, SSD and χ^2



b) Best appearance, SSD



c) Best non-appearance, SSD

Figure 4.7: Selection of performances of Experiments 1, 2, and 3. The SSD and χ^2 refer to the sum of squared differences and χ^2 score, respectively.

of 10% of the eye width) are removed. The remaining two sets are least squares linearly interpolated and the resulting slopes determine the inner angle. The estimation for the outer angle (D8) is similar.

Experiments

We define three experiments. In Experiment 1, we assess the verification performance of the appearance based features as a group and compare that with the texture based ULBP approaches. In this experiment we use the sum of squared difference and the χ^2 score function. The latter score function is often employed in the LBP case and is equivalent to the sum of squared differences score function in the (binary) appearance based features approach. In Experiment 2, we determine the verification performance of the appearance based features separately. In Experiment 3 we focus on the verification performance of the non-appearance based features. In Experiment 2 and 3 we use the sum of squared differences score function. All experiments are conducted on the right and left eye. All presented results are fused at score level by using min-max scaling and sum fusion.

Table 4.3: Performance appearance features in terms of AUC.

Appearance feature	AUC
(A1) Fissure Shape	0.851
(A2) Superior palperal fold	0.820
(A3) Upper Folds	0.758
(A4) Epicanthic fold	0.500
(A5) Upper Lashes	0.808
(A6) Lower Lashes	0.915
(A7) Lower Folds	0.754
(A8) Inferior palperal fold	0.882
(A9) Infraorbital furrow	0.824
(A10) Sclera Blood	0.500
(A11) Sclera Defects	0.787
(A12) Shape Iris	0.957
(A13) Shape Pupil	0.941
(A14) Shape Caruncle	0.762

4.3.6 Results

In Experiment 1, we compare the verification performance of the appearance approach versus the ULBP and shifted ULBP measured in terms of AUC, resulting in 0.966, 0.926, and 0.919 in the case of the sum of squared differences score function, respectively. We notice that the appearance based method seems to perform slightly better than the texture based methods. However, this difference disappears when the three approaches use the χ^2 score function. In that case the AUC values are 0.966, 0.970, and 0.966, respectively. In both cases one might have expected a lower AUC value for the shifted ULBP relative to the AUC value of the original UBLP as the former contains "less" information than the latter one. However, the measured differences seem insignificant. The ROC curves of Experiment 1 are shown in Figure 4.7a.
The results of Experiment 2 (separate appearance based features) are shown in Table 4.3. Apart from two trivial outcomes (0.500 in the cases of (A4) Epicanthic fold and (A10) Sclera Blood, caused by absence of annotation), the AUC values vary between 0.754 and 0.957.

Surprisingly, the shapes of the iris and the pupil yield the top two AUC values. The shapes reveal the position and size of these modalities and apparently represent a discriminating property. Every (non-)appearance based feature is easily changed, especially the iris and pupil positions can change instantaneously just by gazing away from the camera. Also the lower lashes, in comparison to the upper lashes, are performing well. This difference might be explained by the fact that upper lashes (a) tend to cover the whole upper fissure opening and (b) are quite dense in their distribution. In contrast, lower lashes when traversing from the medial to the lateral canthus (a) often start around the projection of the iris on the lower fissure opening and (b) exhibit a more sparse distribution. The latter property implies that the lower lashes have a potential to be more "unique". These observations are also illustrated by the annotation example in Figure 4.6. A similar difference can be observed between the two outlines of the eyelids (superior palperal and iInferior palperal folds): again, the lower part

Table 4.4: Performance non-appearance features in terms of AUC.

Non-appearance feature	AUC
(D1) Distance R/L eye	0.789
(D2) Angle eye	0.700
(D3) Colour sclera	0.726
(D4) Position, diameter iris	0.956
(D5) Colour Iris	0.714
(D6) Position, diameter pupil	0.868
(D7) Angle inner eye	0.773
(D8) Angle outer eye	0.623

seems more "unique". Apart from the shape of the fissure opening, the AUC of the remaining four appearance based features fall below 0.800. The study presented in [114] indicates that humans found the eyelashes, tear duct (caruncle), eye shape (fissure opening) and the eyelids most helpful in their identity decision making process while using near-infrared images.
This result is clearly reproduced in this study, with one noticeable exception: the caruncle. This might be explained by the difference in the representation of this modality. In our approach, only the shape drawn on a limited resolution image exhibiting reflection artifacts (see Figure 4.6) is taken into account, whereas in [114] the use of the near-infrared spectrum and high resolution demarcates the caruncle very well. In Figure 4.7b, ROC curves of the three best performing appearance based features (iris, pupil, and lower lashes) are shown.

Finally, the verification performance of the non-appearance based features are measured in Experiment 3 and are shown in Table 4.4. The top two performing modalities are the iris and pupil positions and confirm the performance of their appearance based counterparts. It also indicates that the implicit location and size information contained in annotation data yield this performance, while other implicit information like fissure opening contributes less. The next best performing feature is the inter-eye distance. Although a very simple measure, it performs relatively well. This is caused by the fact that the localisation of the medial canthi is very robust. However, its value expressed as a relative measure can be hampered by an approximate localisation of the earlobe positions. The inner angle is much more stable than the outer angle, as the localisation of the lateral canthi is less robust. This also explains the lower AUC value for the angle of the eyes. Finally, the iris and sclera colours are not very convincing non-appearance features. Overall, the verification performance of the non-appearance features is generally inferior to the appearance based features. ROC curves of the three best performing non-appearance based features are shown in Figure 4.7c.

4.3.7 Conclusion

In this paper, we have studied the feasibility of FISWG characteristics descriptors for verification purposes and how their performance relates to a representative of a texture based feature representation. We find that some of the FISWG features work well (iris, pupil position, either appearance of non-appearance based, lashes, fissure shape), while others are less convincing. Especially the non-appearance features that measure an angle or colour do not perform well. This is in line with a FISWG recommendation "(photo-)anthropometry has

limited discriminating power". Using semantically important information encapsulated in the FISWG characteristic descriptors can yield verification performances comparable to texture based methods. Finally, we are very aware that this study has been conducted on a limited subset, so its results should be considered to be indicative.

4.4 Chapter conclusion

In this chapter, two studies on the periocular region have been combined. They addressed research question 1b: *Under relatively well-conditioned settings, what is the general performance of biometric classifiers that use FISWG characteristic descriptors as their input and produce strength of evidence in relation to other non-forensic biometric classifiers?*

The first study showed that the performance of classifiers using FISWG characteristic descriptors were comparable to ones that use the Dong Woodard feature set, in terms of the Cllr performance measure. Moreover, several combinations of characteristic descriptor components of the eyebrow had a similar performance compared to the full eyebrow characteristic descriptor, indicating that the descriptor could be made more compact. The second study considered the feasibility of FISWG characteristic descriptors of the eye for verification purposes and how their performance related to a representative of a texture based feature representation. Some of the FISWG characteristic descriptors worked well (iris, pupil position, either appearance or non-appearance based, lashes, fissure shape), while others were less convincing. The study concluded that using semantically important information encapsulated in the FISWG characteristic descriptors can yield verification performances that are comparable to texture based methods.

We find the following with respect to the addressed research question. Both studies indicate that at least the biometric classifiers are comparable to their non-forensic counterparts. Also, both studies show that not every part of a characteristic descriptor contributes equally well, so it seems that a more compact representation is possible. Despite their comparable performance, it raises the question regarding their added value in relation to their non-forensic counterparts. This question was also raised in Chapter 3 in an experiment involving humans. This is especially relevant when non-technical, non-forensic features (like Dong Woodard) can be understood by a court of law. Moreover, in both studies the characteristic descriptors have been manually extracted, while in the second study the ULBP features are determined automatically; classifiers that use them have a similar performance. Although some characteristic descriptors can be determined in an automatic fashion, due to the sometimes intricate definition of characteristic descriptors (for example facial line types), we expect that a complete detection is not feasible, let alone under forensic conditions. Our conclusion with respect to research question 1b is that classifiers using characteristic descriptors in terms of performance are mostly comparable to those that use non-forensic features under relatively well-conditioned settings, but that their added value can also be questioned.

These conclusions are restricted to the periocular region, but we assume that the outcomes are representative of classifiers that use other characteristic descriptors. However, both studies were conducted using images taken under relatively well-conditioned settings. Therefore, the conclusions drawn might not hold for trace images that are representative of various forensic use cases. We explore the usage of characteristic descriptors in those use cases in Chapter 6, whereas the next chapter presents the dataset used for that assessment.

Chapter 5

ForenFace dataset and toolset

5.1 Introduction

The previous two chapters showed that the performance of humans and classifiers using characteristic descriptors of the periocular region were comparable to humans and classifiers that use other non-forensic features. One limitation of these studies is that they have been conducted under relatively well-conditioned settings.

Therefore, in subsequent research we focus on more realistic, forensic use cases. In the next chapter, the performance of characteristic descriptors under those use cases is studied. The aim of this chapter is to be instrumental for that chapter by presenting the ForenFace dataset. This dataset is introduced since an analysis shows that other datasets used in the realm of forensic research are not fully suitable for the study of FISWG characteristic descriptors under these forensic use cases. This chapter as such does not directly address any research question.

Section 5.2 has been published as "ForenFace: a unique annotated forensic facial image dataset and toolset" [21].

Reading Guide
Section 5.2. This section should at least be browsed, in particular the reader should get acquainted with the various image types and the results of the baseline experiment.

5.2 ForenFace: a unique annotated forensic facial image dataset and toolset

5.2.1 Abstract

Few facial image datasets are suitable for forensic research. In this paper, we present ForenFace, a facial image and video dataset. It contains video sequences and extracted images of 97 subjects recorded with six different surveillance camera of various types. Moreover, it also contains high-resolution images and 3D scans. The novelty of this dataset lies in two aspects: (a) a subset of 435 images (87 subjects, five images per subject) has been manually

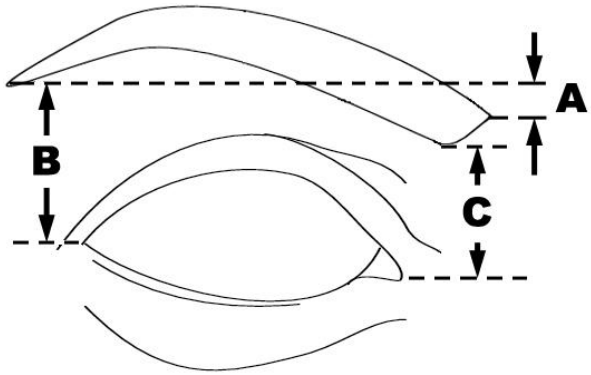


Figure 5.1: Some FISWG eyebrow characteristic descriptors that capture the vertical position of the eyebrow, from [8].

annotated, yielding a very rich forensically relevant annotation of almost 19.000 facial parts, and (b) making available a toolset to create, view, and extract the annotation. We present protocols and the result of a baseline experiment in which two commercial software packages and an annotated facial feature contained in this dataset are compared. The dataset, the annotation and tools are available under a usage license.

5.2.2 Introduction

In forensic evaluation, trace material may for example consist of facial images extracted from CCTV footage taken at a crime scene and reference material may be (high quality) mugshots or 3D scans. During the comparison process, the FFR-examiner particularly pays attention to shape like features [12, 57]. The Facial Identification Scientific Working Group (FISWG) [7] has published recommendations for this process [6]. In particular, their Image Comparison Feature List for Morphological Analysis [8] contains characteristic descriptors (facial features) that can be taken into account. The nature of these descriptors ranges from broad and qualitative to narrow and quantitative. As an example, we show in Figure 5.1 three vertical differences. In this example, A refers to the difference between inner and outer eyebrow tips, B refers to the difference between the outer eyebrow tip and outer eye corner, and C refers to the difference between the inner eye corner and the lowest point on the eyebrow outline in the vicinity of the inner eye corner.

The comparison process is largely manual, making it worthwhile to investigate whether biometric classifiers can assist the practitioner. However, quality of trace material is typically limited by technological and subject factors. Technological factors include image compression artifacts, perspective effects, low resolution, and interlacing. Subject factors include pose, illumination, expression, and partial occlusion of the face by hoodies or balaclavas. Therefore, it is not always possible to use "off-the-shelf" classifiers that have been developed for a specific biometric non-forensic application. A different approach is to use classifiers (or rather the evidential value derived from their comparison score) that are specialised in a particular facial part or a set of characteristic descriptors. For example, such classifiers using the FISWG characteristic descriptors of the eyebrow and eye region have been studied by

Zeinstra et al [18, 20]. In Tome et al. [100, 104] results on automatic classifiers on forensic regions and shapes are presented.

For the development and testing of such classifiers, the availability of datasets that are representative of forensic trace and reference material is of paramount importance. This can be observed from the related field of automatic face recognition. This field has grown from initial work by Kanade [133] and Turk and Pentland [87] into a well-established, mature, and wide field of research. Many face recognition systems have been developed and successfully deployed in real world use case scenarios. A key success factor has been the availability of public facial image datasets (for example FRGC [33]) and vendor challenges using those datasets [2, 134, 135]. Initially, those datasets mainly consisted of images acquired under controlled conditions, but gradually there has been a shift towards sets acquired under uncontrolled, more realistic "in the wild", circumstances. Examples are Labeled Faces in the Wild [136], HELEN [137], and Quis-Campi [97].

However, to date, the number of facial image datasets that are suitable for forensic research is limited. Even within the group of forensic type datasets, not every dataset is suitable for forensic evaluation of trace and reference material as described before. We identify three criteria that in our opinion determine the suitability of such a specific dataset. They are:

1. Representativeness of trace material;
2. Representativeness of reference material;
3. Availability of forensic features.

Representativeness refers to being typical of images encountered in forensic evaluation. With respect to the first criterion, real trace material typically consists of CCTV video footage and extracted stills of subjects that may have occluded parts of their face. The quality of trace material can vary between cases and depends for example on resolution and physical placement of the camera. Representativeness of reference material, the second criterion, are often high-resolution frontal, quarter profile, and profile images, and sometimes 3D scans are employed as well. Finally, the third criterion, which we believe is the most important one in the context of forensic evaluation, is the availability of forensic features that are typically used by an FFR-examiner. Exactly these features can be used to train and test specialised biometric classifiers.

In this paper, we introduce ForenFace, a forensic dataset designed with these three criteria in mind. It includes very rich manual annotation from which forensically relevant features can be extracted.

We note that some of the included annotations in some use cases might have been obtained by a computer vision algorithm. However, in general the poor image quality of trace material restricts the usability of such approaches. Moreover, facial part definitions are not always easily captured in an algorithm. For example, the proper detection of facial lines can be difficult.

Manual annotation is a very resource intensive process. Therefore, we restrict the annotation to three different forensic use cases that as a whole are representative of forensic case work. We define a forensic use case as a criminal act whose traces consist of distinct facial image types. The first very common forensic use case is a money robbery at a bank, shop or gas station. At those premises often CCTV surveillance cameras are mounted on a wall or ceiling. Since this is such an important use case, we have annotated two images of different

Table 5.1: Contents of datasets.

Dataset	Subjects	Material	Forensic Features
SCFace [92]	130	Traces: 7 surveillance cameras, 3 distances, 1 close-up surveillance References: 5 images	Four landmarks
Chokepoint [93]	29	Footage of three surveillance cameras	Eye coordinates (per frame)
NIST Mugshot [94]	1573	Gray scale mugshot	None
Morph (Academic) [95]	13.618	Scanned and digital mugshots	Eye coordinates
ATVS Forensic DB [96]	50	High-resolution, three distances	21 landmarks
FRGC [33]	568	Frontal images and 2.5D scans, taken under (un)controlled conditions with neutral or smiling subjects	4 landmarks
Labeled Faces in the Wild [136]	5749	Unconstrained face images	Identity label
HELEN [137]	2330	Unconstrained face images	199 landmarks
Quis-Campi [97]	320	Traces: videos and images References: registration images, gait, 3D model face	Eye coordinates (per frame)
ForenFace	97	Traces: CCTV video and stills from 6 surveillance cameras of visible and partially occluded subjects References: 5 images, 3D scan	Annotated facial parts

resolution and illumination. Another use case is money withdrawal from an ATM using a stolen debit card. Here, the trace material is recorded by a small camera, often mounted near the keypad. The final use case that we address is when a customs or immigration officer suspects that the used identity document has not been tampered with, but does not correspond to the person who is presenting it. These forensic use cases correspond to specific images. In each use case, the reference material consists of a high-resolution frontal image and its annotation is compared to annotation on trace images. We refer to this as the annotation scenario.

Although the annotation scenario forms the main raison d'être of this dataset, there are several other research scenarios possible in which the annotation is not employed, but for which this dataset is still of interest. We mention two of them here, other uses are discussed in Section 5.2.6. The first scenario is one in which stills extracted from video sequences are compared to a 3D image for forensic investigation. In the second scenario, two video sequences are compared to investigate whether the videos contain the same person. We present evaluation protocols for all three scenarios in Section 5.2.6.

In Table 5.1, we compare the ForenFace dataset with nine publicly available image datasets that can be used in forensic research. Although the SCFace dataset has its merits and has been used in numerous publications on low resolution face recognition, traces only consist of frontal surveillance camera images. The ChokePoint dataset is designed for "person identification/verification under real-world surveillance conditions". Since it does not contain reference images, it is not suitable enough for research within a forensic context. The NIST and Morph datasets only contain mugshots, and are mainly suited for longitudinal research. The ATVS Forensic DB only contains high-resolution mugshots. FRGC has been used in numerous face biometric studies, but it is somewhat limited in its forensic relevance. Labeled Faces in the Wild is widely used to evaluate modern face recognition algorithms that can cope with uncontrolled settings. HELEN is mainly used for the training and evaluation of facial feature localisation algorithms on images taken under non-ideal conditions. Finally,

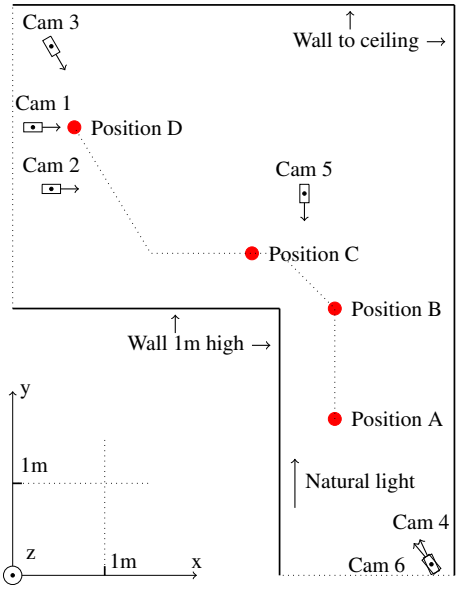


Figure 5.2: Abstract top view of the layout of the experiment showing camera locations, subject positions, and subject paths. Objects present at the physical setup like tables and a closet are omitted for clarity.

Figure 5.3: CCTV footage from Camera 3 when subject is standing at position B wearing a baseball cap. Camera 2 can be seen at the lower right corner and the pole on which Cameras 4 and 6 are attached is visible at the top, slightly left from the middle.

Quis-Campi contains videos and stills taken from modern surveillance systems that typically have a higher resolution than those acquired by traditional systems. However, Quis-Campi lacks a good set of reference images.

The ForenFace dataset also comes with a set of three software tools. One tool can be used for the viewing and creation of manual annotation. Another tool can be used to setup a new dataset for annotation, and the third tool can be used to extract annotation in a flexible manner. The software tools are usable on any platform for which a JVM (Java Virtual Machine) is available.

Table 5.2: Surveillance cameras setup.

Camera	3D coordinates	Pose (°)
Camera 1	(0.22, 4.87, 1.00)	(90, 45)
Camera 2	(0.42, 4.20, 1.60)	(90, 0)
Camera 3	(0.42, 5.75, 2.40)	(150, -25)
Camera 4	(4.55, 0.12, 2.00)	(335, -25)
Camera 5	(3.17, 4.15, 1.60)	(180, 0)
Camera 6	(4.55, 0.12, 2.60)	(320, -50)

The dataset, annotation, and the toolset are available under a usage license. This usage license encompasses a privacy policy/statement, the right to use this dataset for research purposes, and the requirement to cite this paper whenever it is used in published research. A User Guide is provided that contains any necessary details. More information can be found at [98].

The structure of this paper follows the three constituent parts of the ForenFace dataset. In Section 5.2.3 we present the data, in Section 5.2.4 we discuss the included annotation, and in Section 5.2.5 we show the accompanying toolset. In Section 5.2.6, we discuss potential uses, propose evaluation protocols, and give a baseline result. Finally, in Section 5.2.7 we present our conclusion.

5.2.3 Data

Video sequence acquisition

Data acquisition took place at the Netherlands Forensic Institute in The Hague, Netherlands, over a period of four days. The location is an open space between the staircase, offices, and a corridor. The arrangement of the six surveillance cameras is depicted in Figure 5.2. Figure 5.3 is a still image extracted from Camera 3 footage showing a subject on position C. The location and pose parameters (p_1, p_2) of the cameras are shown in Table 5.2. Here, the parameter p_1 denotes the compass bearing (clockwise from North=0°) and assumes that the positive y-axis corresponds to the Northern bearing. The parameter p_2 is the angle with the constant z-plane. A positive or negative angle means the camera is pointing upwards or downwards, respectively. The camera types are shown in Table 5.4. Additionally, the location of positions A-D are shown in Table 5.3.

The location was artificially illuminated. Using the compass bearing described in Table 5.2, natural light was coming out of the 180° direction. Subjects were asked to (1) stand at position A facing 0°, look around, (2) stand at position B facing 0°, (3) stand at position C facing 270°, look around, (4) stand at position D facing 270°, look down into camera 1, look around, look up into camera 3, (5) stand at position C facing 90°, look around, (6) stand at position B facing 180°, look around, and finally (7) stand at position A facing 180°, look around. This procedure was executed twice (with/without baseball cap) and leads to 12 video sequences. Frontal facial images were extracted and a selection is shown in Figure 5.4. A selection of reference data is shown in Figure 5.5.

Table 5.3: Positions A-D.

Position	2D coordinates
Position A	(3.50, 1.70)
Position B	(3.50, 2.90)
Position C	(2.60, 3.50)
Position D	(0.67, 4.87)

Table 5.4: Surveillance camera types.

Camera	Model	Type
1	Watec WAT-230A	BW Pinhole
2	Sanyo VCC-6580P	Narrow Angle
3	Panasonic WVP480	Wide Angle
4	Vista VEC30H-DN	Low Light
5	Sony SSC-D372	Narrow Angle
6	Dallmeier DDF3000A	Dome

3D scan and other image acquisition

In an adjacent room three (half profile/half frontal left and right, and frontal) 3D scans were taken using a Minolta VIVID910 scanner, after which they were merged into one collection of polygons (ply format). Five (profile left and right, half profile/half frontal left and right, and frontal) reference images were acquired by a Canon EOS 10D. Finally, the identity document type images were taken from employee cards. These passport style photographs were taken several months or years before.

Image and scan contents

For each subject a number of video sequences, images, and 3D scans are available. Details are given in Tables 5.5 and 5.6. In Table 5.5, ≪sid≫ refers to the subject id, and ≪camera≫ to the camera number $\in \{1,\ldots,6\}$. Also, a and b refers to footage and images in which the subject does not wear and does wear a baseball cap, respectively. Finally, IPD is the interpupillary distance measured in pixels. Additionally, in Table 5.6 in the Canon EOS 10D entry, f refers to frontal, p to profile (right/left), and q to quarter profile (right/left).

The video sequences were converted from a Dallmeier proprietary format to MPEG4 by using the PStream Convert conversion tool [138]. The Dallmeier SMAVIA viewer software [138] was used to manually select and extract still images from the CCTV footage. The 3D scans can be viewed with several open source software packages, such as MeshLab [139].

5.2.4 Annotation

Forensic features

As indicated in the Introduction, FFR-examiners use shape like features during forensic case work. We have selected a large set of these features to be included in this dataset, presented

Figure 5.4: From top left clockwise, stills from camera (subject position): 1 (D), 2 (B), 4 (C), 3 (D), 6 (A), 5 (B), 4 (B), and 3 (B).

Table 5.5: Available video sequences and extracted images.

Source	Description	Format	Avg. IPD (px)	# Wearing No cap/Cap
Camera 1-6	Video sequence	≪sid≫c≪camera≫{a,b}.mpeg	N/A	97/97
Camera 1	Position D	≪sid≫c1{a,b}7.bmp	65	89/86
Camera 2	Position B	≪sid≫c2{a,b}3.bmp	27	97/96
Camera 3	Position B	≪sid≫c3{a,b}3.bmp	11	97/97
	Position C	≪sid≫c3{a,b}8.bmp	15	97/97
	Position D	≪sid≫c3{a,b}16.bmp	38	90/90
Camera 4	Position C	≪sid≫c4{a,b}2.bmp	13	97/97
	Position B	≪sid≫c4{a,b}7.bmp	15	97/96
	Position A	≪sid≫c4{a,b}12.bmp	23	95/95
Camera 5	Position B	≪sid≫c5{a,b}3.bmp	68	97/97
Camera 6	Position A	≪sid≫c6{a,b}3.bmp	38	93/94

in Figures 5.6 and 5.7. Most FISWG characteristic descriptors are contained in or can be determined from this set. For example, with respect to Figure 5.1, the eyebrow shape is contained in the set, the A position can be determined from the eyebrow shape, whereas the B and C positions can be determined from the eyebrow and fissure outline.

Manual Annotation

In our dataset, the forensic features can be extracted from the manual annotation. We use four annotation types. Examples are shown in Figure 5.8.

The **first** annotation type is the landmark type, which is a well-defined, fiducial point on the facial image. We identified in total 21 landmarks, which can be used to determine the overall face/head composition (Figure 5.6a, (H3)-(H23)).

The **second** and **third** annotation types are used to annotate shapes. Often shapes are represented by a polygon defined by the collection of landmarks. A disadvantage of this approach is that parts of the shape with a high curvature need significantly more landmarks

Figure 5.5: From top left clockwise: identity document, frontal reference, 3D scan, and half profile reference.

Table 5.6: Other images and 3D scans.

Source	Description	Format	Avg. IPD (px)	#
Canon EOS 10D	Reference	≪sid≫{f, lp, lq, rp, rq}.jpg	370	97
Unknown camera	Identity document	≪sid≫a.jpg	35	97
Minolta	3D scan	≪sid≫.ply	N/A	93

than almost linear parts of the shape. Therefore, we propose a more flexible and compact solution using Hermite splines. A Hermite spline is a piecewise third order polynomial parametric curve [140]. It is defined by the interpolation of the landmarks, and, in our work, by assuming that the tangent at a landmark is given by the directional vector that interpolates the neighbouring landmarks. This approach has several advantages over the polygon approach. First, it needs a reduced number of landmarks to capture a rich variation in shapes. Second, if needed, it can be subsampled into a set of landmarks of arbitrary resolution. More details on the subsampling process are given in Section 5.2.5. The second and third types are the open and closed facial shape respectively, in both cases represented by a Hermite spline. The nose outline is an example of an open shape, whereas the eyebrow shape is an example of a closed shape.

The **fourth** annotation type is the point cloud type, which describes multiple points belonging to the same feature, without performing an interpolation. Although these points could also have been represented by open or closed curves, it is more efficient to utilise this type. Typical examples are eye lashes or lip creases.

The trace and reference image names and annotation properties are summarised in Table 5.7. As expected, the average number of annotated facial parts depends inversely on the interpupillary distance. This is caused by the fact that a significant proportion of the considered forensic features are detailed to very detailed, and therefore are only discernible in good quality images with a relative large interpupillary distance. Example annotations are shown

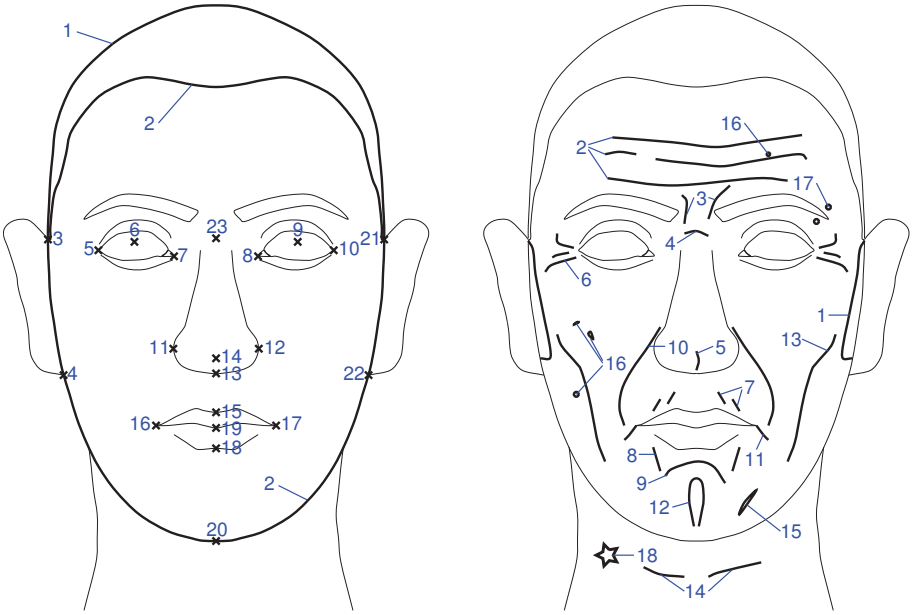


Figure 5.6: Left: Face from a holistic perspective. Right: face from detailed perspective. Prefixes H and D refer to holistic and detailed perspective, respectively. (H1) Cranial Vault, (H2) Shape of Face, (H3-H23) 21 landmarks (upper/lower connection ears to face (H3, H4, H21, H22), inner/outer corners eyes (H5, H7, H8, H10), pupils (H6, H9), alae (H11, H12), below nose (H13), nose tip (H14), upper/lower lip (H15, H18), mouth corner (H16, H17), mouth (H19), chin (H20), and nasal root (H23). (D1) Facial Hair Outline, (D2) Forehead Creases, (D3) Vertical Glabellar, (D4) Nasion Crease, (D5) Bifid Nose Crease, (D6) Periorbital Creases, (D7) Upper Circumoral Striae, (D8) Lower Circumoral Striae, (D9) Mentolabial Sulcus, (D10) Nasolabial Creases, (D11) Marionette Lines, (D12) Cleft Chin, (D13) Buccal Creases, (D14) Neck wrinkles, (D15) Scars, (D16) Facial Marks, (D17) Piercing, and (D18) Tattoo.

in Figure 5.9. Out of 97 subjects, 87 subjects have all five images available, yielding in total 435 annotated images.

Annotation acquisition

Three paid participants were recruited for the annotation. The participants had prior knowledge and experience, as they had participated in another forensic annotation experiment. Prior to the instruction, the instructor discussed instruction details with the NFI. The instruction was given in a single session of a day. Image sets were prepared such that participants annotated a subject exactly once in a session of 87 images. After a week the annotations were evaluated, and feedback was given to the annotators. The duration of an annotation session varied between two to four weeks. The annotation was facilitated by an application that provides basic drawing tools and a visualiser that gives feedback to the participant. Also, the participant could either specify that he/she could not determine an annotation, or state his/her confidence in the annotation on a five-point scale (very unconfident, unconfident, neutral, confident, very confident). The annotation was assessed and approved by the instructor.

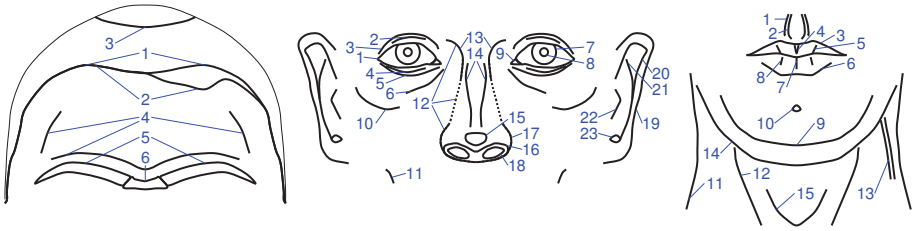


Figure 5.7: Left: Upper facial parts. Middle: Middle facial parts. Right: Lower facial parts. Prefix U, M, and L refer to the upper, middle, and lower parts, respectively. (U1) Forehead hairline, (U2) Hairline/Forehead boundary, (U3) Cranial baldness, (U4) Ridge structures, (U5) Eyebrows Outline, and (U6) Unibrow. (M1) Fissure Outline, (M2) Upper Folds, (M3) Superior Palpebral Furrow, (M4) Lower Folds, (M5) Inferior Palpebral Furrow, (M6) Infraorbital Furrow, (M7) Iris, (M8) Pupil, (M9) Caruncle Outline, (M10) Cheekbone, (M11) Dimple Cheek, (M12) Nose Outline, (M13) Nasal Root, (M14) Nasal Body, (M15) Nasal Tip, (M16) Nasal Base, (M17) Alae, (M18) Nostrils, (M19) Outer Helix, (M20) Inner Helix, (M21) Anti-Helix, (M22) Tragus, and (M23) Anti-Tragus. (L1) Philtrum Ridges, (L2) Philtrum Furrow, (L3) Upper Lip Outline, (L4) Upper Lip Tubercle, (L5) Upper Lip Creases, (L6) Lower Lip Outline, (L7) Lower Lip Median Sulcus, (L8) Lower Lip Creases, (L9) Chin Outline, (L10) Chin Dimple, (L11) Neck Boundaries, (L12) Musculature, (L13) Veins, (L14) Double chin, and (L15) Laryngeal.

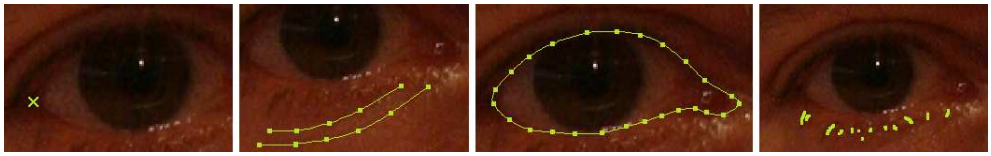


Figure 5.8: Examples of the four annotation types. From left to right: Outer eye (landmark), Lower folds (open shape), Fissure (closed shape), and Lower eye lashes (point cloud type).

5.2.5 Toolset

Three graphical applications bundled in a toolset are made available. The ScratchPadTool is primarily used to create the annotation, and it is also possible to view or create the annotation with this tool. Moreover, the ScratchPadTool can be used to annotate any other dataset. For this, the ScratchPadEnroller must be used to prepare a dataset for use by the ScratchPadTool.

The final tool, ScratchPadExtractor, has two functions. Since the annotation uses the coordinate system of the image it belongs to, the annotation must be registered to a common coordinate system prior to use. Therefore, ScratchPadExtractor provides a function to register the annotation on pupil coordinates. The tool is also used to extract points from the annotation. As discussed in Section 5.2.4, the annotation is stored as a collection of points that define a Hermite spline. The tool can sample these Hermite splines to create a dense collection of points that represent a shape. The user can provide some parameters, such as the number of sampling points and the manner in which the sampling is performed. As an example, two sets of sampled Hermite splines are provided. These points can be used directly as a feature (for example eyebrow shape) or indirectly in a feature (for example the angle of the eye fissure).

The tools are shown in Figure 5.10 and are described in more detail in the User Guide.

Table 5.7: Annotated trace and reference images.

Forensic Use Case Illustration	Short Image Name	Trace Material	Avg. IPD (px)	Avg. # Annotated Facial Parts
ID card	mid-res	≪sid≫a.jpg	35	51
Debit Card	bw-down	≪sid≫c1a.bmp	65	50
Robbery 1	near-frontal low-res 1	≪sid≫c4a12.bmp	23	27
Robbery 2	near-frontal low-res 2	≪sid≫c3a3.bmp	11	19
Reference	high-res	≪sid≫f.jpg	370	74

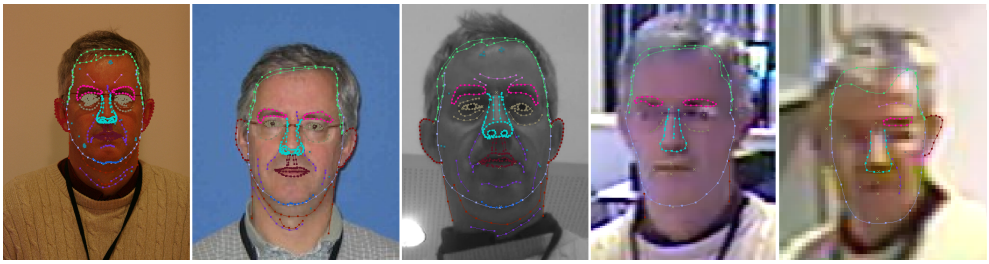


Figure 5.9: Example images that have been annotated. From left to right: Reference image high-res and four trace images: mid-res, bw-down, near-frontal low-res 1, and near-frontal low-res 2. Short image names are defined in Table 5.7.

5.2.6 Potential uses, evaluation protocols, and an example

Potential uses

We envision that this dataset is particularly useful for forensic research. The chosen forensic features are in line with the characteristic descriptors found in [8]. We can extract a large number of these characteristic descriptors from the annotation. For example, from the eyebrow and eye fissure annotation we can derive multiple characteristic descriptors: the eyebrow shape, eye fissure shape and angle and for example three particular relative positions A, B, and C as shown in Figure 5.1 in the Introduction. These features can then in turn be used by biometric classifiers. Other uses include the matching of 3D images with video sequences and video sequences versus video sequences, respectively.

We realise that the size of this dataset is small from a biometric perspective. For example, the FRGC dataset [33] contains more than 39000 images of 568 subjects. Still, we believe that the availability of a rich annotation could aid research on facial parts that have had little attention before. This is in line with a growing interest of the biometric community to fuse soft biometric facial features with highly discriminating biometric features to enhance performance in non-ideal situations. A typical example is the periocular region complementing iris images. An earlier study by Zeinstra et al. [20] has shown that using information captured in annotated images of the periocular region performs comparably to more texture based approaches described in for example the work by Park et al [112].

Another potential use of this dataset is the evaluation of computer vision algorithms for the extraction of facial features. In this respect, the annotation contained in this dataset can serve as ground truth for the evaluation of such algorithms.

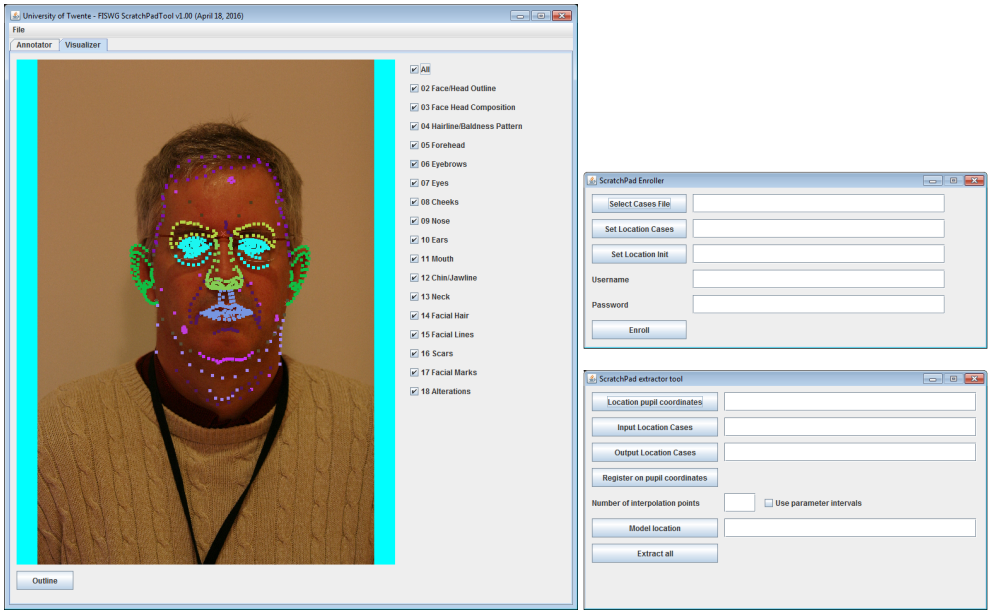


Figure 5.10: The provided software tools. Left: ScratchPadTool for viewing and creation of annotation. Top Right: ScratchPadEnroller for the preparation of a new dataset. Bottom Right: ScratchPadExtractor for the extraction of features from the annotation.

Evaluation protocol 1: annotation scenario

We propose the following evaluation protocol for the annotation scenario. Since the number of subjects is limited, we provide 50 random different partitions of 67 training and 20 test subjects. This particular choice is a trade-off between having enough training and test subjects. The test results of each of the 50 partitions are collected in a single test result set. The performance of a system using this scenario should be reported on this aggregated set. Note that the test results are those of a family of very related classifiers, rather than a single classifier. More details can be found in the User Guide.

Evaluation protocol 2: video versus 3D scenario

We propose the following two evaluation protocols for the video versus 3D scenario.

The first evaluation protocol (2A) is that this dataset is only used for the evaluation of video versus 3D algorithms that are trained on other datasets. Both identification and verification modes of operation are possible. Any camera sequence (12=6 cameras $\times$ wearing no cap/cap) can be matched against all 3D reference shapes.

The second evaluation protocol (2B) is similar to evaluation protocol 1. We provide 50 random different partitions of 73 train and 20 test subjects. For each of the 12 camera sequences, the test results of each of the 50 partitions are collected in a single test result set. The performance of a system using this scenario should be reported on this aggregated set.

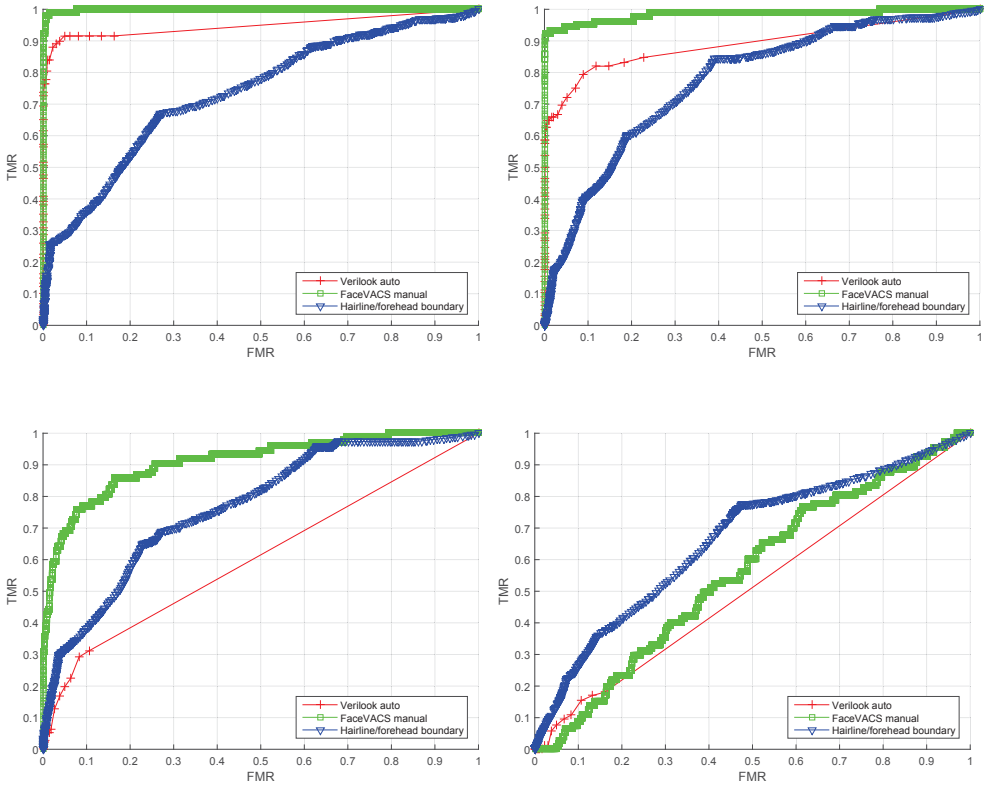


Figure 5.11: Receiver Operator Characteristic curves for Verilook, FaceVACS, and the hairline/forehead boundary in the four use cases. Top row, left: mid-res versus high-res, right: bw-down versus high-res. Bottom row, left: near-frontal low-res 1 versus high-res, right: near-frontal low-res 2 versus high-res.

Evaluation protocol 3: video versus video scenario

We propose the following two evaluation protocols for the video versus video scenario.

The first evaluation protocol (3A) is that this dataset is only used for the evaluation of video versus video algorithms that are trained on other datasets. Both identification and verification modes of operation are possible. Any camera sequence (12=6 cameras $\times$ wearing no cap/cap) can be matched with any other camera sequence, giving a total of 12$\times$ 11/2=66 possible combinations.

The second evaluation protocol (3B) is similar to evaluation protocols 1 and 2B. We provide 50 random different partitions of 77 train and 20 test subjects. For each of the 66 camera combinations, the test results of each of the 50 partitions are collected in a single test result set. The performance of a system using this scenario should be reported on this aggregated set.

Example protocol 1: baseline versus hairline/forehead boundary

In this section, we present baseline results and compare those with what can be achieved by using only the Hairline/Forehead boundary. All experiments use the proposed evaluation protocol of Section 5.2.6.

For the baseline experiment, we use Neurotec Verilook 6.0 [141] and Cognitec FaceVACS 9.1 [142]. These systems use the full face. Prior to the experiment, we let Neurotec Verilook 6.0 automatically determine the pupil coordinates and we provide FaceVACS with manually determined pupil coordinates. We consider four cases: mid-res versus high-res, bw-down versus high-res, near-frontal low-res 1, and near-frontal low-res 2. The results are shown in Figure 5.11.

We compare these results with what can be achieved by only using the Hairline/Forehead boundary. Prior to comparison, we register the annotation on pupil coordinates in order to introduce a common coordinate system. We subsample the Hairline/Forehead boundary with 100 equidistant points. We use a shape similarity score function to compare two shapes. If $X = \{\mathbf{x}_i \in \mathbb{R}^2 | i = 1, \ldots, N_x\}$ and $Y = \{\mathbf{y}_i \in \mathbb{R}^2 | i = 1, \ldots, N_y\}$ are two shapes, we define the shape similarity score function $s_s : \mathbb{R}^{2\times N_x} \times \mathbb{R}^{2\times N_y} \to \mathbb{R}$ as:

$$s_s(X,Y) = -\frac{1}{N_x}\sum_{i=1}^{N_x} d_{pc}^2(\mathbf{x}_i, Y) - \frac{1}{N_y}\sum_{i=1}^{N_y} d_{pc}^2(\mathbf{y}_i, X), \tag{5.1}$$

where d_{pc} measures the minimal distance between a point $\mathbf{w} \in \mathbb{R}^2$ and a point cloud $Z = \{\mathbf{z}_i \in \mathbb{R}^2 | i = 1, \ldots, N\}$: $d_{pc}(\mathbf{w}, Z) = \min_{i=1,\ldots,N} \|\mathbf{w} - \mathbf{z}_i\|$.

We compare the results with the Hairline/Forehead boundary shape as an illustration of its relative robustness against severe image degradation.

The results are shown in Figure 5.11. We can make several observations. First of all, both commercial systems clearly outperform the hairline/forehead boundary shape-based system in the comparison of mid-res versus high-res and bw-down versus high-res. The commercial systems are typically designed to cope with these image types and conditions. The situation changes when we consider the near-frontal low-res 1 and 2 images. In the case of near-frontal low-res 1 versus high-res we notice that Verilook is performing worse than the hairline/forehead boundary shape, but FaceVACS still has the best performance. However, we observe that the hairline/forehead boundary shape performs better than both commercial systems in near-frontal low-res 2 versus high-res. Another observation is that the hairline/forehead boundary shape is not very discriminating, but has some robustness under different comparisons. Note, however, that an FFR-examiner takes all available comparison results into account during an assessment of evidential value.

5.2.7 Conclusion

In this paper, we have presented ForenFace, a novel forensic facial video and image dataset. It contains CCTV footage, extracted still images, reference images, and 3D scans. Its novelty with respect to other forensic facial datasets in the forensic domain is twofold. Inspired by the FISWG characteristic descriptors, it is the first dataset that includes a rich forensically relevant annotation of almost 19.000 facial parts on 435 images of five different image types of varying quality. Moreover, it comes with a toolset of three complementary software tools

that can be used on other datasets as well. We believe that these two factors lead to a dataset that has an added value in the field of forensic face datasets.

We proposed evaluation protocols and showed in the annotation scenario that the baseline performance of commercial systems on the severest case is less than a system that is only using the hairline/forehead boundary shape.

By making this dataset available to the research community, we hope to encourage research especially in the forensic domain. As can be seen from the baseline experimental results, face recognition in a realistic forensic setting is still not a solved issue.

5.2.8 Acknowledgments

We want to thank the volunteers at the Netherlands Forensic Institute for their participation in the creation of the dataset, and the annotators for their time investment. Finally, we would like to thank Neurotechology and Cognitec Systems GmbH. for supporting our research by providing the VeriLook and FaceVACS software. Results obtained for VeriLook and FaceVACS were produced in experiments conducted by the University of Twente, and should therefore not be construed as a vendor's maximum effort full capability result.

5.3 Chapter conclusion

In this chapter, we have presented the ForenFace dataset. No research question has been addressed by this chapter. The ForenFace dataset can be used in forensic research as it contains forensic type videos, images, and annotation from which the FISWG characteristic descriptors can be extracted. The results of a small baseline experiment indicated that at least in low quality surveillance camera footage, a hairline shape classifier was somewhat better than a system that uses the full face. This implies that it is indeed worthwhile to study classifiers and characteristic descriptors themselves under various forensic settings. This is the topic of the next chapter.

Chapter 6

FISWG characteristic descriptors under various forensic use cases

6.1 Introduction

As indicated in the chapter conclusion of Chapter 4, we aim to explore biometric classifiers that use characteristic descriptors under various forensic use cases. The ForenFace dataset introduced in the previous chapter is used in this chapter, in particular its manual annotation from which the characteristic descriptors are extracted. This chapter addresses research question 1c: *Under various forensic use cases, what is the general performance of biometric classifiers that use FISWG characteristic descriptors as their input and produce strength of evidence in relation to face recognition systems?* and research question 1d: *Under various forensic use cases, what is (a) the measurability of FISWG characteristic descriptors and (b) the influence of annotation variation on characteristic descriptors and strength of evidence produced by biometric classifiers that use these characteristic descriptors?*

This chapter contains two studies. The first study considers discriminating power (measured in EER) of biometric classifiers using FISWG characteristic descriptors as their input under various forensic use cases; in total four classifier types are being studied. Score fusion results based on classifier type and facial category are also presented. The second study complements the first study by considering two related properties of characteristic descriptors. The first property is measurability, that is, to which extent can characteristic descriptors be extracted under various forensic use cases. The second property is variability and studies the influence of annotator variability on landmark positions, shapes, and ultimately on the strength of evidence produced by the biometric classifiers.

Section 6.2 has been published as "Discriminating power of FISWG characteristic descriptors under different forensic use cases" [22].

Section 6.3 has been published as "Manually annotated characteristic descriptors: measurability and variability" [23].

Reading Guide
Section 6.2. The Abstract, Introduction and FISWG characteristic descriptors can be omitted as it is mostly based on previously presented material.
Section 6.3. The Abstract, Introduction and FISWG characteristic descriptors can be omitted as it is mostly based on previously presented material.

6.2 Discriminating power of FISWG characteristic descriptors under different forensic use cases

6.2.1 Abstract

FISWG characteristic descriptors are facial features that can be used for evidence evaluation during forensic case work. In this paper, we investigate the discriminating power of a biometric system that uses these characteristic descriptors as features under various forensic use cases. We show that in every forensic use case we can find characteristic descriptors that exhibit moderate to low discriminating power. In all but one use cases, a commercial face recognition system outperforms the characteristic descriptors. However, in low resolution surveillance camera images, some (combination of) characteristic descriptors yield better results than commercial systems.

6.2.2 Introduction

One of the tasks of a forensic facial examiner is to compare trace images to reference images taken from a suspect in order to determine evidential value. This process is referred to as forensic face verification. Although there does not exist a de jure or de facto international standard for forensic face verification, a standardisation effort is done by FISWG [7]. FISWG has published several recommendations on facial identification, including a one-to-one comparison list describing characteristic descriptors [8] that can be used during forensic case work.

We envision a forensic facial evaluation system that receives input from a forensic facial examiner and computes evidential value. According to [105] and a recently proposed forensic guideline [48], discriminating power is one of six aspects that should be taken into account during the validation of such a forensic evaluation method. Therefore, the aim of this paper is not to present classifiers that have state-of-the-art results, but rather to investigate the discriminating power of biometric classifiers using FISWG characteristic descriptors as features under various forensic use cases.

Note that we present biometric results, that is, averaged results. In particular, same source scores stem from comparisons of different subjects. In an ideal situation sufficient trace and reference material of one subject is available, making a specific subject based comparison possible. This will be investigated in future work.

Typically trace and reference images are wholly visible in casework, whereas in this work we extract FISWG characteristic descriptors prior to the comparison. Its advantage is that we can investigate the truly isolated FISWG characteristic descriptor, and we do not create a bias

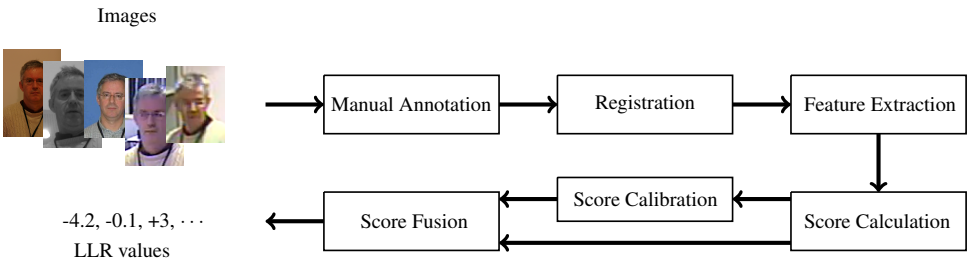


Figure 6.1: Overview of our system. LLR refers to log-likelihood ratio as a means to represent evidential value.

by seeing the features simultaneously in the trace and reference images. An overview of our system is given in Figure 6.1.

This paper is organised as follows. In Section 2 we discuss related work, in Section 3 we introduce the FISWG characteristic descriptors, and in Section 4 we present forensic use cases. In Section 5 the experimental setup is discussed. In Section 6 we discuss results of single and specific combinations of characteristic descriptors. We also compare our results to two commercial face recognition systems. Finally, in Section 7 we present the conclusion.

6.2.3 Related work

There exist numerous studies [82–84] showing that in general anthropometric measurements are not suitable for evidential evaluation. Therefore, forensic face verification typically involves the examination of (dis)similarities of shape like facial features. It is remarkable that the FISWG one-to-one comparison list includes some anthropometric measures.

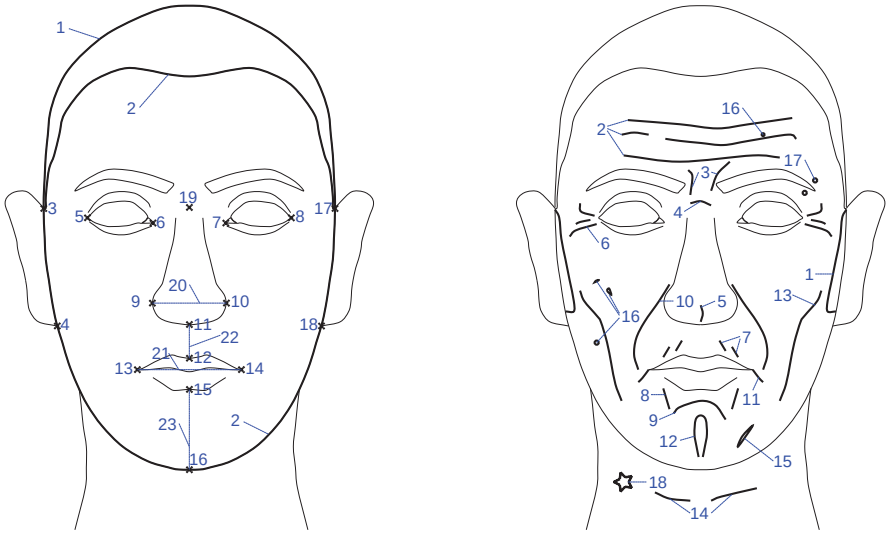
Two studies by Tome et al. are closely related to our work. In [100], the biometric performance of linear SVM classifiers on 15 forensic facial regions is investigated. Here the SCFace [92] and subset of Morph [95] are used. They conclude that "... depending on the acquisition distance, the discriminative power of regions change, having in some cases better performance than the full face". In [104], the performance of continuous and discrete soft biometric features are tested on the Morph [95] and ATVS Forensic DB [96] datasets. Experimental results show high discrimination power and good recognition performance for some specific cases. However, these cases correspond to relatively good quality images.

There exist some smaller scale studies by Zeinstra et al. that also consider FISWG from an extract feature first based approach, on eyebrows [17, 18], and the periocular region [20].

Other research efforts focus on somewhat different aspects of forensic face recognition: facial aging, forensic sketch recognition, and facial mark based matching and retrieval [50, 143].

6.2.4 FISWG characteristic descriptors

Characteristic descriptors capture information that is considered important during forensic casework. In most cases multiple characteristic descriptors are extracted from one facial trait. For example, from the eyebrow the shape, size, hair density, symmetry, and specific relative positions can be extracted. We present the characteristic descriptors only visually in Figures



a) Face from a holistic perspective

b) Face from detailed perspective

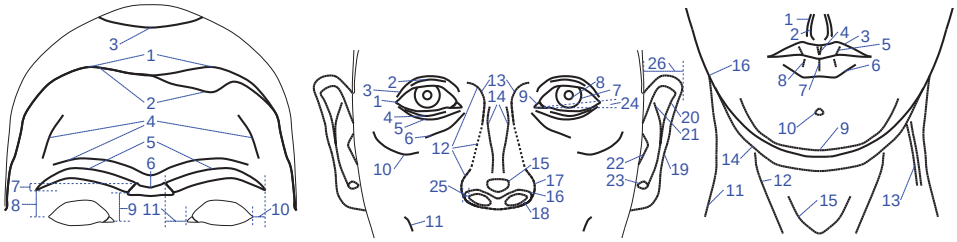
Figure 6.2: Face from a) holistic and b) detailed perspective. Prefixes H and D refer to holistic and detailed perspective, respectively. (H1) Cranial Vault shape/availability, (H2) Facial shape, (H3) location of 17 landmarks (upper/lower connection ears to face (H3, H4, H17, H18), inner/outer corners eyes (H5-H8), nose (H9-H11), mouth (H12-H15), chin (H16), and nasal root (H19)), (H20) width of nose, (H21) width of mouth, (H22) nose-mouth distance, and (H23) mouth-chin distance. (D1) Facial Hair shape/symmetry/availability, (D2) Forehead Creases shape/size/availability/count, (D3) Vertical Glabellar shape/size/availability/count, (D4) Nasion Crease shape/availability/count, (D5) Bifid Nose Crease shape/availability/count, (D6) Periorbital Creases shape/size/availability/count, (D7) Upper Circumoral Striae shape/size/availability/count, (D8) Lower Circumoral Striae shape/size/availability/count, (D9) Mentolabial Sulcus shape/size/availability, (D10) Nasolabial Creases shape/size/availability, (D11) Marionette Lines shape/size/availability, (D12) Cleft Chin shape/size/availability, (D13) Buccal Creases shape/size/availability, (D14) Neck wrinkles shape/size/availability/count, (D15) Scars shape/availability/count, (D16) Facial Marks shape/availability/count, (D17) Piercing shape/availability/count, and (D18) Tattoo shape/availability/count.

6.2 and 6.3 due to the sheer number of them (250). In general, most characteristic descriptors fall into classes as landmark, shape, width, size, etc. Also, some very specific characteristic descriptors are defined. For example, the B position of the eyebrow is defined as the vertical position of the outer tip of the eyebrow with respect to the outer eye corner (Figure 6.3 U8).

6.2.5 Forensic use cases and the ForenFace dataset

A forensic use case refers to a criminal act whose traces consist of distinct facial image types. In our work we use the ForenFace dataset [98]. This dataset contains manually annotated images of 87 subjects that are representative of three forensic use cases. Moreover, the FISWG characteristic descriptors can automatically be derived from the annotation.

The ID Card use case occurs for example when a customs or immigration officer suspects that the used identity document has not been tampered with, but does not correspond to the person who is presenting it. The Debit Card use case is the withdrawal of money using a



a) Upper facial parts b) Middle facial parts c) Lower facial parts

Figure 6.3: Upper a), middle b), and lower c) parts of the face. Prefixes U, M, and L refer to the upper, middle, and lower parts, respectively. (U1) Forehead hairline shape/symmetry/size, (U2) Hair/Forehead boundary shape, (U3) Cranial baldness shape/availability, (U4) Ridge structures shape/availability, (U5) Eyebrows shape/size/symmetry, (U6) Unibrow shape/availability, (U7-U11) Relative positions A-E. The relative positions are measured on both eyebrows. (M1) Fissure shape/size/symmetry, (M2) Upper Folds shape/availability/count, (M3) Superior Palpebral Furrow shape/availability, (M4) Lower Folds shape/availability/count, (M5) Inferior Palpebral Furrow shape/availability, (M6) Infraorbital Furrow shape/availability, (M7) Iris shape, (M8) Pupil shape, (M9) Caruncle shape, (M10) Cheekbone shape/availability, (M11) Dimple Cheek shape/availability, (M12) Nose shape/size/symmetry, (M13) Nasal Root shape/size, (M14) Nasal Body shape/size/symmetry, (M15) Nasal Tip shape/symmetry, (M16) Nasal Base size/deviation, (M17) Alae shape, (M18) Nostrils shape/size/symmetry, (M19) Outer Helix shape/symmetry/size, (M20) Inner Helix shape/size, (M21) Anti-Helix shape/size, (M22) Tragus shape/size, (M23) Anti-Tragus shape/size, (M24) Fissure angle, (M25) Nostril thickness, and (M26) Ear Protrusion. (L1) Philtrum Ridges width/symmetry, (L2) Philtrum Furrow width/symmetry, (L3) Upper Lip shape/symmetry, (L4) Upper Lip Tubercle shape, (L5) Upper Lip Creases shape, (L6) Lower Lip Outline shape/symmetry, (L7) Lower Lip Median Sulcus shape, (L8) Lower Lip Creases shape, (L9) Chin shape/size/symmetry, (L10) Chin Dimple shape/availability, (L11) Neck Boundaries size, (L12) Musculature shape/availability, (L13) Veins shape/availability, (L14) Double chin shape/availability, (L15) Laryngeal shape/size/availability, (L16) Jawline shape.

stolen debit card. In this case, trace material is recorded by a small camera in the ATM. The Robbery use case is a robbery on a bank, shop or gas station. At those premises, often CCTV surveillance cameras are mounted on a wall or ceiling.

The images and forensic use cases are shown for one subject in Figure 6.4. In particular, with the average interpupillary distance (IPD) in pixels, for each subject we have one annotated reference image (370px), and four annotated trace images: ID Card (35px), Debit Card (65px), and two for the Robbery use case: Robbery 1 (23px), and Robbery 2 (11px). The first two images are acquired by a photo camera, the latter three are extracted from CCTV footage. All images are colour images, except Debit Card images.

6.2.6 Experimental setup

Annotation, registration and extraction of characteristic descriptors

The annotation in the ForenFace dataset contains landmarks and shapes. The latter are represented by Hermite splines. A Hermite spline is a piecewise third order polynomial parametric curve [140]. It is defined by the interpolation of the annotated points and, in the case of the ForenFace dataset, by assuming that the tangent at an annotated point is equal to the vector connecting the neighboring points.

Since the raw manual annotation data lacks a common coordinate system, we apply an

a) Reference b) ID Card c) Debit Card d) Roberry 1 e) Roberry 2

Figure 6.4: Available images with average IPD: a) Reference image (370px), b) ID Card (35px), c) Debit Card (65px), d) Robbery 1 (23px), and e) Robbery 2 (11px).

affine transformation as a registration step. This registration maps pupil coordinates to fixed locations.

Characteristic descriptors are then extracted from the registered annotation. Shape descriptors are equidistantly subsampled from the corresponding Hermite spline. All other descriptors can be derived from Hermite splines and landmarks. Other descriptors like the B position of the eyebrow (Figure 6.3 U8) are derived from two types of annotation, in this case the eye fissure and eyebrow shapes.

Similarity score functions, score calibration, and score fusion

We use four different similarity score functions. Given the forensic context of our work, scores should be interpretable as evidential value. In some cases we can directly model the similarity score function as a log-likelihood ratio:

$$s(x,y) = \log(\mathrm{LR}(x,y)) = \log\left(\frac{\mathrm{p}\left(\begin{pmatrix}x\\y\end{pmatrix}|\mathscr{H}_\mathrm{s}\right)}{\mathrm{p}\left(\begin{pmatrix}x\\y\end{pmatrix}|\mathscr{H}_\mathrm{d}\right)}\right). \tag{6.1}$$

Here, $\mathrm{p}\left(\begin{pmatrix}x\\y\end{pmatrix}|\mathscr{H}_\mathrm{s}\right)$ and $\mathrm{p}\left(\begin{pmatrix}x\\y\end{pmatrix}|\mathscr{H}_\mathrm{d}\right)$) model the joint probability of trace x and reference y under the same source and different source hypothesis, respectively. The same source hypothesis states that trace and reference originate from a common donor, whereas the different source hypothesis states that trace and reference do not have a common donor.

For low dimensional continuous descriptors like width, size, etc. we have $(x,y) \in \mathbb{R}^k \times \mathbb{R}^l$, $k,l \leq 2$. We assume a normal distribution (after subtraction of the mean) with $\Sigma_s = \begin{pmatrix}\Sigma_{xx} & \Sigma_{xy}\\ \Sigma_{xy}^T & \Sigma_{yy}\end{pmatrix}$ and $\Sigma_d = \begin{pmatrix}\Sigma_{xx} & 0\\ 0 & \Sigma_{yy}\end{pmatrix}$:

$$\begin{pmatrix}x\\y\end{pmatrix}|\mathscr{H}_\mathrm{s} \sim \mathscr{N}(0,\Sigma_s) \text{ and } \begin{pmatrix}x\\y\end{pmatrix}|\mathscr{H}_\mathrm{d} \sim \mathscr{N}(0,\Sigma_d). \tag{6.2}$$

In this case (6.1) has a closed form, with $\Delta = \Sigma_d^{-1} - \Sigma_s^{-1}$:

$$l_N(x,y) = \frac{1}{2}\left(\log|\,\Sigma_d| - \log|\Sigma_s| + (x^T y^T)\Delta\begin{pmatrix}x\\y\end{pmatrix}\right). \tag{6.3}$$

For availability descriptors, that is, $(x,y) \in \{0,1\}\times\{0,1\}$, we assume a bivariate Bernoulli distribution:

$$\begin{pmatrix}x\\y\end{pmatrix}|\mathscr{H}_\mathrm{s} \sim \mathrm{Bern}(p_{00},p_{10},p_{01},p_{11}) \text{ and } \begin{pmatrix}x\\y\end{pmatrix}|\mathscr{H}_\mathrm{d} \sim \begin{pmatrix}\mathrm{Bern}(q_x)\\\mathrm{Bern}(q_y)\end{pmatrix}. \tag{6.4}$$

Under this assumption, (6.1) reverts to

$$l_B(x,y) = \log\left(\frac{p_{xy}}{q_x^x(1-q_x)^{(1-x)}q_y^y(1-q_y)^{(1-y)}}\right). \tag{6.5}$$

We also define two similarity score functions that are not modeled as a log-likelihood ratio. This is necessary when (6.1) of a characteristic descriptor either cannot be easily modeled or its parameters cannot be reliably estimated.

The count similarity score function is applied on count descriptors and is given by

$$s_C(x,y) = -|x-y|. \tag{6.6}$$

We represent shapes in terms of point clouds, so if $X = \{\mathbf{x}_i \in \mathbb{R}^2 | i = 1,\ldots,N_x\}$ and $Y = \{\mathbf{y}_i \in \mathbb{R}^2 | i = 1,\ldots,N_y\}$, then the shape similarity score function is defined by

$$s_{Shape}(X,Y) = -\frac{1}{N_x}\sum_{i=1}^{N_x} d_{pc}^2(\mathbf{x}_i,Y) - \frac{1}{N_y}\sum_{i=1}^{N_y} d_{pc}^2(\mathbf{y}_i,X), \tag{6.7}$$

where d_{pc} measures the minimal distance between a point $\mathbf{w} \in \mathbb{R}^2$ and a point cloud $Z = \{\mathbf{z}_i \in \mathbb{R}^2 | i = 1,\ldots,N\}$: $d_{pc}(\mathbf{w},Z) = \min_{i=1,\ldots,N}\|\mathbf{w}-\mathbf{z}_i\|$.

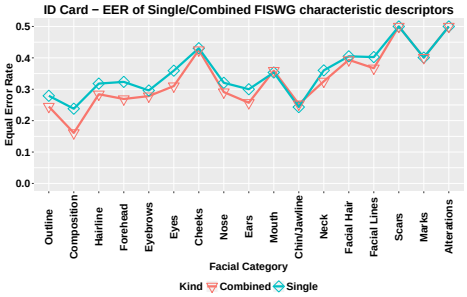
By assumption, the scores obtained from (6.3) and (6.5) are log-likelihood ratios. The scores obtained by (6.6) and (6.7) are converted into log-likelihood ratios by using the Pool of Adjacent Violators algorithm [47] on the set of scores. This algorithm constructs a monotonic transformation such that a similarity score s is mapped to an a posteriori probability $\mathrm{p}(\mathscr{H}_\mathrm{s}|s)$ from which the log-likelihood ratio $l(s)$ can be derived:

$$l(s) = \log(\mathrm{LR}(s)) = \mathrm{logit}(\mathrm{p}(\mathscr{H}_\mathrm{s}|s)) - \mathrm{logit}(\mathrm{p}(\mathscr{H}_\mathrm{s})). \tag{6.8}$$

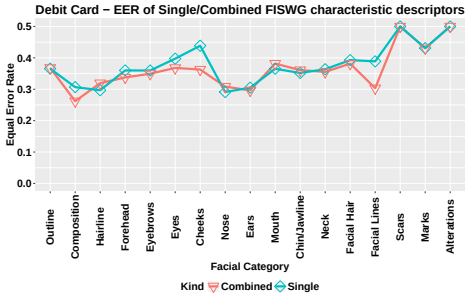
If we assume independence of facial features, then score fusion by adding scores corresponds to the log-likelihood ratio of the combined characteristic descriptors.

Experimental protocol

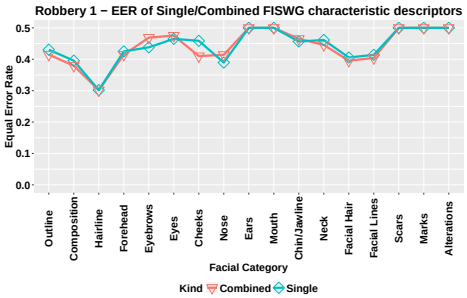
We use the train-test protocol specified by ForenFace. It specifies 50 randomly generated splits of 87 subjects into 67 training and 20 test subjects. Test scores for each round are aggregated. Parameters for (6.3) and (6.5) are estimated during the training phase. For the characteristic descriptors on which (6.6) or (6.7) are applied, during the training phase the transformation (6.8) is estimated, which is then applied to test scores for conversion into log-likelihood values.



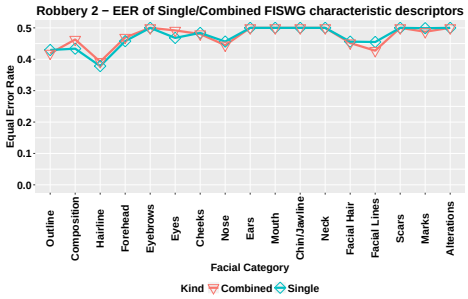
a) ID Card use case



b) Debit Card use case



c) Robbery 1 use case



d) Robbery 2 use case

Figure 6.5: Lowest EER of single and combined characteristic descriptors within each facial category under different forensic use cases. Combined refers to score fusion within a facial category.

6.2.7 Experimental results and discussion

Since the number of characteristic descriptors is large, we will restrict the presentation of results. In particular, we graphically present characteristic descriptors with the lowest EER within a facial category. This measure of discriminating power is chosen in accordance with [48]. Also, results of fusion based on the used similarity score function outside the facial categories are presented. Finally, results are compared to commercial face recognition systems.

Single and combined characteristic descriptors

Figure 6.5 a) shows the results in the ID Card use case. We find that for single characteristic descriptors two modalities have a similar lowest EER=0.24: the position of the lower ear landmark, and the shape of the jaw. Furthermore, the remainder of the characteristic descriptors have EER ≥ 0.28. The single best performing facial category is the composition of the face. This category encompasses all landmark positions and 4 distances, see Figure 6.2 a) H3-H23. When we combine all normal (l_N) scores within that category, we obtain EER=0.16. In most other facial categories a combination based on normal scores also yields the lowest EER.

Table 6.1: EER of score fusion outside a facial category and under different forensic use cases.

	ID Card	Debit Card	Robbery 1	Robbery 2
Normal (l_N)	0.13	0.24	0.37	0.46
Bernoulli (l_B)	0.19	0.23	0.29	0.40
Count (s_C)	0.42	0.46	0.50	0.50
Shape (s_{Shape})	0.18	0.33	0.38	0.45
All	0.13	0.22	0.31	0.43

In the Debit Card use case (Figure 6.5 b)) we observe that some single characteristic descriptors have EER=0.29 to 0.31. They are in order of EER: the size of the nose, the hairline/forehead boundary, and the size of the ears. The combination of the face composition normal scores yields the lowest EER (0.26) for any considered combination.

Apart from one specific landmark, in the ID Card use case mostly shape based features yield the best results. In the Debit Card use case simple measures like sizes have the highest discriminating power. This is probably caused by a low contrast, making a majority of shapes more difficult to discern. In both use cases, the composition of the face yields the best combined results within a facial category. As indicated before, anthropometry in general has limited use in forensic evaluation. However, it seems that flexibility in the model underlying (6.3) helps to capture the relation between the trace and reference descriptors.

Figures 6.6 c) and d) indicate a significant reduction in discriminating power. We find for Robbery 1 and 2 one single characteristic descriptor performing relatively well: the hairline/forehead boundary (EER=0.31 and EER=0.38), respectively. The shape of the hair/ forehead boundary seems actually somewhat resilient to harsh image conditions. This can be explained by the fact that even under challenging conditions this boundary is still visible due to its length and clear colour or gray scale difference. In the work of Tome et al. [100], it is reported that in a similar situation the forehead area is the best performing facial area. We think that the discriminative nature of this area actually stems from the inclusion of this boundary.

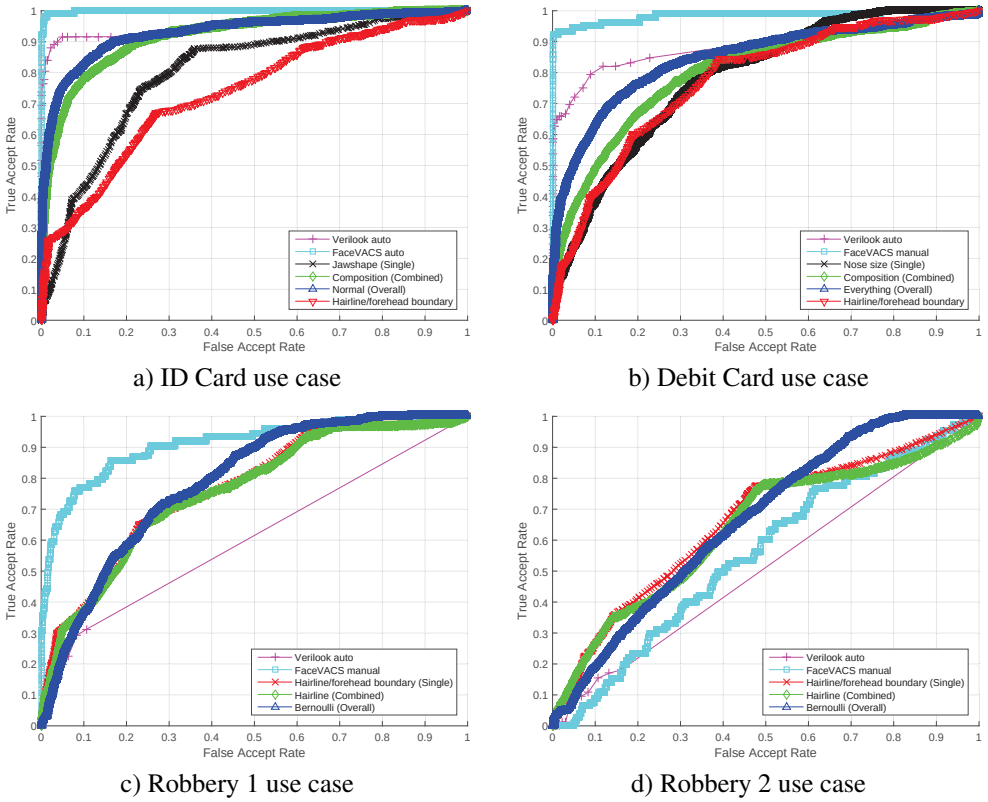
The EER of the characteristic descriptors rapidly pass the EER=0.40 mark. Other simple descriptors like the availability of facial hair, the size and availability of facial lines (especially the nasolabial lines, see Figure 6.2 D10), and size of the nose start to emerge as single and combined characteristic descriptors with the lowest EER.

Fusion outside a facial category

In Table 6.1 results of score fusion based on similarity score function are presented. This table also includes the fusion of all scores.

First of all, we observe that the fusion of count scores does not yield any satisfactory results. This has several causes. Upon inspection, we observed a mismatch between counts in trace and reference images, as counts seemingly are very sensitive to image size and conditions. Moreover, the score function s_C only measures count difference and does not take the count itself into account.

We observe that in the ID Card use case, fusion of all normal scores is marginally better



a) ID Card use case

b) Debit Card use case

c) Robbery 1 use case

d) Robbery 2 use case

Figure 6.6: ROC's of commercial systems versus characteristic descriptors under different forensic use cases.

(EER=0.13) than that of the composition of the face (EER=0.16). This indicates that already a part of the discriminative power is contained in the composition of the face. A similar effect can be seen in the Debit Card use case.

Fusion of normal scores yields the highest discriminating power in the ID use case. This prevalence shifts towards fusion based on availability (l_B) features in the Debit Card and Robbery use case. This effect is to be expected for the latter use case, as only the availability descriptors are robust to severe image quality degradation.

Commercial systems

We also test performance of two commercial systems using the same evaluation protocol. The Neurotec Verilook 6.0 [141] (with automatic eye coordinate detection) and Cognitec FaceVACS 9.1 [142] (with manually provided eye coordinates) systems are used for this purpose. Both systems use proprietary algorithms. The results are shown in Figure 6.6 for each forensic use case.

The characteristic descriptors are both in the ID Card and Debit Card use cases outperformed by commercial systems. The situation gets more interesting in the Robbery use cases.

In both use cases, Verilook generates a large number of zero same source and different source scores, causing the large linear part in the ROC. In the Robbery 2 case, the discriminative power of both commercial systems are now essentially random and the shown characteristic descriptors have lower EER than the commercial systems. One could argue that the hairline/forehead boundary is not a very good biometric since its permanence and measurability properties can be challenged. An alternative biometric modality is the fusion of availability features. However, the reported EERs are not satisfactory from a biometric perspective.

6.2.8 Conclusion

In this work, we have investigated the discriminating power of a biometric system that uses FISWG characteristic descriptors as features under various forensic use cases. In every forensic use case we can find characteristic descriptors, either single or combined, that yield moderate to low discriminating power. However, in all but one forensic use case, the characteristic descriptors are outperformed by a commercial face recognition system. In the use case with low interpupillary distance (11px), we found that the hairline forehead boundary as a single characteristic descriptor and the combination of availability features perform somewhat better than both commercial systems. However, their discriminating power is low.

The presented results are averaged biometric results. In an ideal situation sufficient trace material of one subject is available, making a specific subject based comparison possible. Further research is needed to investigate the discriminative power of specific FISWG characteristic descriptors in that situation.

6.2.9 Acknowledgment

We would like to thank Cognitec Systems GmbH. for supporting our research by providing the FaceVACS software. Results obtained for FaceVACS were produced in experiments conducted by University of Twente, and should therefore not be construed as a vendor's maximum effort full capability result.

6.3 Manually annotated characteristic descriptors: measurability and variability

6.3.1 Abstract

In this paper, we study the measurability and variability of manually annotated characteristic descriptors on a forensically relevant face dataset. Characteristic descriptors are facial features (landmarks, shapes, etc.) that can be used during forensic case work. With respect to measurability, we observe that a significant proportion of characteristic descriptors cannot be determined in images representative of forensic case work. Landmarks, closed and open shapes, and other facial features mostly show that their variability depends on the image quality. Up to 50% of all considered evidential values are either positively or negatively influenced by annotator variability. However, when considering images with the lowest quality, we found that more than 70% of the evidential value intervals in principle could yield the wrong conclusion.

6.3.2 Introduction

When a person robs a shop, recordings of that person made by a CCTV camera might be usable as trace material. If a suspect is caught, the FFR-examiner will compare the, often low quality, trace image(s) to high quality mugshot reference images taken from the suspect when the suspect is in custody. There exist numerous studies ([6, 82–85]) that show that in general anthropometric measurements are not suitable for forensic evaluation. Therefore, typically the general composition of the face, shape like features (for example the shape of the jaw), and when possible, highly discriminating features like facial marks are taken into account. The outcome of the process is evidential value.

Different forensic institutes use similar but not identical comparison procedures, see Spaun [57] for the operating procedures at the FBI and Prince [12] for other institutes as well. The Facial Identification Scientific Working Group [7] has published recommendations on the comparison process [6]. Their FISWG Facial Image Comparison Feature List for Morphological Analysis [8], FISWG Feature List for short, contains characteristic descriptors, that is, facial features, that can be used during forensic case work.

The comparison process itself is largely manual. We assume that the examiner endows the trace and reference images with manual annotation from which the characteristic descriptors are derived. Corresponding characteristic descriptors in trace and reference images are then compared. In this work, we investigate the measurability and variability of manually annotated characteristic descriptors.

According to Jain et al. [30], measurability refers to *"how possible it is to capture the biometric feature using a suitable device (...). The raw data captured must also allow for (...) feature extraction."* Here, the biometric feature is the characteristic descriptor, and the device is the annotator who creates annotation from which the characteristic descriptors are to be derived. The dynamic range of characteristic features is large, that is, they range from large scale features (for example the outline of the face) to small scale features (for example lip wrinkles). This suggests there is a relationship between the measurability of characteristic descriptors and for example the resolution of a trace image.

Since taking these measurements is an inherently subjective process, they will exhibit variability. We will investigate this with respect to the placement of landmarks, open and closed shapes, and a selection of other characteristic descriptors. The variability of landmarks in a forensic context has been studied before by Tome et al [96]. Since the ultimate output of a forensic facial examiner is evidential value, we present four evidential value models, and study the variability of evidential value caused by annotation variability.

The remainder of this paper is structured as follows. In Section 6.3.3 we introduce characteristic descriptors. In Section 6.3.4 we describe the experimental setup. In Section 6.3.5 we present and discuss the experimental results and in Section 6.3.6 we formulate our conclusion.

6.3.3 FISWG characteristic descriptors

Since there are more than 250 FISWG characteristic descriptors, we refer to Section 6.2.4 and [8] for an overview. Figure 6.2a shows the face from a holistic perspective that contains large scale structures like the outline of the face and landmarks that indicate the position of facial parts within the face. Figure 6.3b shows the characteristic descriptors that reside in the

middle part of the face.

Although the number of characteristic descriptors is large, each of them falls into one of four feature types. The first feature type is "low-dimensional" $\mathbb{R}^k$, that is either $k = 1$ (for example the eye fissure angle), or $k = 2$ (for example a landmark position). The second type is the visual occurrence of a facial feature (for example the cheekbone), expressed as a binary value. The third type is count (for example the number of upper eye folds). The final type are shapes. An example is the shape of the outer ear helix. The shape feature type is the most frequent type in the set of characteristic descriptors.

6.3.4 Experimental setup

Dataset

In this study, we employ the ForenFace dataset [98]. This dataset contains a reference image and four different trace images for 87 subjects. The trace images are chosen such that they are representative of particular forensic use cases: (a) ID Card refers to the use of a valid identity document of another person, (b) Debit Card refers to the use of a stolen Debit Card and (c) Robbery refers to a robbery on for example a bank or shop. Two images having different resolution and illumination properties represent the latter use case. Example images with their average interpupillary distance (IPD) are shown in Figure 6.4. ForenFace also contains annotation from which all characteristic descriptors can be derived.

Extraction of characteristic descriptors

The annotation of ForenFace consists of either landmark positions, see Figure 6.2a, or points that collectively constitute a Hermite spline, representing a shape. A Hermite spline [140] is a piecewise third order polynomial that defines a smooth open or closed curve that can be subsampled into an arbitrary dense point cloud. Most point clouds are directly usable as a characteristic descriptor, for example the eye fissure shape (Figure 6.3b, item 1). Other point clouds, possibly in conjunction with other point clouds, can be used to derive other characteristic descriptors. For example, the eye fissure angle (Figure 6.3b, item 24) is derived from the eye fissure shape, whereas the nostril thickness (Figure 6.3b, item 25) is derived from the alae and nostril shapes. The visual occurrence and count type features are extracted by counting the number of distinct shapes that constitute a characteristic descriptor, for example the number of upper eye fold shapes.

Measurability

For each characteristic descriptor and forensic use case, we calculate the percentage of subjects for which we found a measurement in the ForenFace dataset. Due to the large amount of characteristic descriptors, we average over each forensic use case and each of the 18 facial categories defined in the FISWG Feature List.

Variability

Annotation variability refers to the variability of multiple annotations of a single facial feature in a single image. The term variability is chosen instead of for example variation or standard

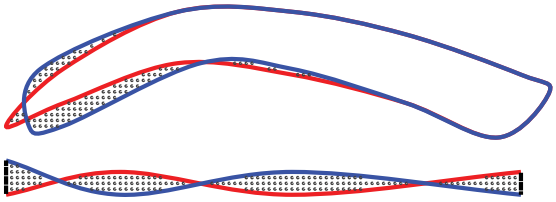


Figure 6.7: Visualisation of pairwise difference. Given blue and red closed (top) and open (bottom) shapes, the pairwise difference is indicated by dots.

deviation, as we use related but different measures for different characteristic descriptors.

We select five subjects (ids 1,4,19,82, and 101) from the ForenFace dataset for which five images are annotated three times by three trained annotators, yielding in total 25 annotated images with 9 annotations each. At least one week between every session is taken into account. Although we can identify three types of variability (the variability within an annotator, the variability between annotators, and the total variability), due to similar results, we only report total variability.

Landmarks. We investigate the variability of landmarks by measuring the standard deviation with respect to their mean and report the results relative to interpupillary distance. This enables the comparison between different forensic use cases.

Open and closed shapes. There does not exist a point to point correspondence between two point clouds sampled from two Hermite splines. This implies that a mean shape cannot be defined. We measure shape variability in terms of pairwise difference between two shapes instead. For closed shapes, this is defined as the area constituted by all points that are inside of one shape and outside of the other shape. For open shapes it is defined as the area of the (possibly self intersecting) closed shape resulting from concatenating corresponding begin and end points of the shapes. Both definitions are visualised in Figure 6.7.

In order to make comparisons between the variability of different closed and open shapes possible, as in the landmark case we scale our results. The scaling factor is not interpupillary distance but its two dimensional analog: the interpupillary area (IPA). It is defined as the area covered by a square with sides equal to the interpupillary distance. Note that with this approach small and large shapes will inherently have small and large variability, respectively. An alternative is to scale to the relative size of the shape. It is straightforward to calculate and to interpret for closed shapes, but a similar approach does not exist for open shapes.

Other characteristic descriptors. Other characteristic descriptors are derived from landmarks and/or shapes, and we will present the variability of a selection.

Evidential value. Evidential value is commonly expressed in a log-likelihood ratio. In the next section, we present four complementary models that are used to calculate evidential value. We report the variability of evidential value in terms of the standard deviation.

Models for evidential value

This section briefly introduces four different models to calculate evidential value. It follows the models presented in [22], in which more details on the derivation are given. In all cases,

x is a trace and y a reference. The same source hypothesis $\mathscr{H}_\mathrm{s}$ states that trace and reference originate from a common donor, whereas the different source hypothesis $\mathscr{H}_\mathrm{d}$ states that trace and reference do not have a common donor. The log is taken with respect to base 10.

Low dimensional features If we assume

$$\mathrm{p}\left(\begin{pmatrix}x\\y\end{pmatrix}|\mathscr{H}_\mathrm{s}\right)\sim N(0,\Sigma_s) \text{ and } \mathrm{p}\left(\begin{pmatrix}x\\y\end{pmatrix}|\mathscr{H}_\mathrm{d}\right)\sim N(0,\Sigma_d) \tag{6.9}$$

then (with $\Delta=\Sigma_d^{-1}-\Sigma_s^{-1}$)

$$l_N(x,y)=\frac{1}{2}\left(\log|\,\Sigma_d|-\log|\Sigma_s|+(x^T y^T)\Delta\begin{pmatrix}x\\y\end{pmatrix}\right) \tag{6.10}$$

yields evidential value. We apply (6.10) only when $(x,y)\in\mathbb{R}^k\times\mathbb{R}^l$, $k,l\leq 2$.

Visual occurrence features If we assume a bivariate Bernoulli distribution for $(x,y)\in\{0,1\}\times\{0,1\}$

$$\begin{pmatrix}x\\y\end{pmatrix}|\mathscr{H}_\mathrm{s}\sim p_{xy} \text{ and } \begin{pmatrix}x\\y\end{pmatrix}|\mathscr{H}_\mathrm{d}\sim\begin{pmatrix}\mathrm{Bern}(q_x)\\\mathrm{Bern}(q_y)\end{pmatrix}, \tag{6.11}$$

then

$$l_B(x,y)=\log\left(\frac{p_{xy}}{q_x^x(1-q_x)^{(1-x)}q_y^y(1-q_y)^{(1-y)}}\right) \tag{6.12}$$

yields evidential value.

Count features The count comparison score function is applied to count descriptors and is given by

$$s_C(x,y)=-|x-y|. \tag{6.13}$$

Shape features We represent shapes in terms of point clouds, so if $X=\{\mathbf{x}_i\in\mathbb{R}^2|i=1,\ldots,N_x\}$ and $Y=\{\mathbf{y}_i\in\mathbb{R}^2|i=1,\ldots,N_y\}$, then the shape comparison score function is defined by

$$s_{Shape}(X,Y)=-\frac{1}{N_x}\sum_{i=1}^{N_x}d_{pc}^2(\mathbf{x}_i,Y)-\frac{1}{N_y}\sum_{i=1}^{N_y}d_{pc}^2(\mathbf{y}_i,X), \tag{6.14}$$

where d_{pc} measures the minimal distance between a point $\mathbf{w}\in\mathbb{R}^2$ and a point cloud $Z=\{\mathbf{z}_i\in\mathbb{R}^2|i=1,\ldots,N\}$: $d_{pc}(\mathbf{w},Z)=\min_{i=1,\ldots,N}\|\mathbf{w}-\mathbf{z}_i\|$.

The scores obtained by (6.13) and (6.14) are converted to log-likelihood ratios by an application of the Pool of Adjacent Violators algorithm [47]. This algorithm outputs a monotonic transformation from which the mapping

$$s\mapsto\ell(s)=\log\left(\frac{\mathrm{p}(s|\mathscr{H}_\mathrm{s})}{\mathrm{p}(s|\mathscr{H}_\mathrm{d})}\right) \tag{6.15}$$

can be constructed.

Table 6.2: Measurability of FISWG characteristic descriptors in percentage of subjects, averaged over characteristic descriptors within a facial category.

	Reference	ID Card	Debit Card	Robbery 1	Robbery 2
Face Head Outline	70.9%	71.3%	71.3%	71.3%	70.9%
Face/Head Composition	94.4%	93.8%	91.0%	86.0%	74.1%
Hairline/Baldness pattern	77.9%	75.6%	75.9%	76.7%	79.3%
Forehead	72.0%	81.6%	83.9%	71.3%	80.1%
Eyebrows	90.9%	87.1%	78.6%	38.8%	9.9%
Eyes	78.2%	41.5%	35.6%	11.3%	7.0%
Cheeks	7.5%	11.2%	8.3%	4.0%	2.6%
Nose	93.6%	69.1%	87.5%	27.5%	8.1%
Ears	50.1%	16.9%	14.8%	5.7%	3.8%
Mouth	92.9%	80.3%	85.7%	7.1%	0.8%
Chin/Jawline	79.8%	76.6%	75.2%	52.4%	26.7%
Neck	41.5%	47.0%	23.6%	17.6%	6.2%
Facial Hair	20.7%	18.4%	19.5%	17.2%	8.0%
Facial Lines	42.0%	30.2%	27.9%	12.4%	4.0%
Scars	0.0%	0.0%	0.0%	0.0%	0.0%
Facial Marks	92.0%	55.2%	58.6%	37.9%	6.9%
Alterations	1.1%	0.0%	0.0%	0.0%	0.0%

6.3.5 Experimental results and discussion

Measurability

In Table 6.2, we see that the measurability of the face/head outline and hairline/baldness pattern categories remain constant over the various forensic use cases. These categories can be considered as large scale facial categories, explaining their constant value.

For most other lower scale facial categories, the percentages decrease as the image quality decreases. Probably, either features are considered to be too bad (for example artifacts by image compression) or too small (feature size compared to pixel size) to be annotated. A special case are scars which were apparently not found in this dataset.

In some cases, the forehead brow structures are more visible in the Robbery 1 and 2 cases, compared to the Reference image. This is probably caused by differences in illumination. The reference images have frontal illumination, whereas the Robbery 1 and 2 images have an artificial illumination component from the ceiling, emphasising brow structures in the forehead.

When looking at the facial categories with the highest feature measurability, we find that reference images favour facial marks, nose, and eyebrows. For the ID Card, the categories are eyebrows, face/head composition, forehead, and mouth; for the Debit Card, nose, mouth, forehead and eyebrows. We see a shift towards larger scale structures when we consider Robbery 1 and 2 images: Hairline/Baldness pattern, Face Head/Outline, and Forehead.

Overall, with decreasing quality and image size, the larger scale facial categories seem to be the relatively highest measurable features. However, the percentage of cases for which these (and most other) features are found decreases. There are two expected effects here

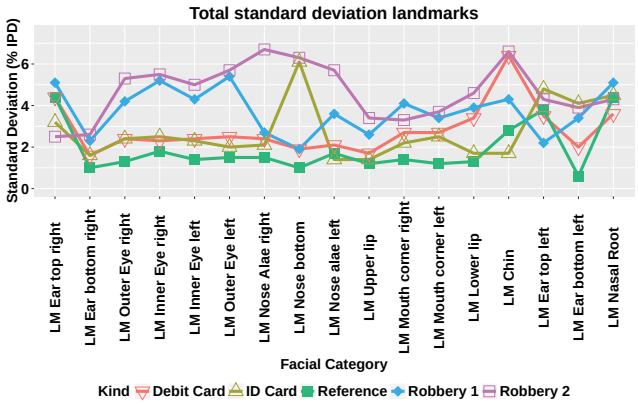


Figure 6.8: Total standard deviation of landmarks, expressed as percentage of average IPD.

caused by alternating between higher and lower quality images.

We conclude that the FISWG Feature List contains a large number of features that are not measurable in realistic forensic use cases, notably the Robbery use cases.

Variability of landmarks

We report the total standard deviations of the landmarks in Figure 6.8.

We notice that the ear upper landmark exhibits more variability than the ear lower landmark. This is probably caused by hair occlusion effecting the annotation of the ear top landmark. Also, we observe that the eye landmarks variability increases in the Robbery 1 and 2 use cases. As can be seen in Figures 6.4d) and 6.4e), the integrity of the eye region in those two use cases has been significantly deteriorated. Moreover, in the same images we see that the nose and mouth region seems somewhat more robust to image degradation in the Robbery 1 and 2 cases. This can also be seen in the reduced variability of those landmarks. The chin landmark variability increases in Robbery 1, Robbery 2, and Debit Card use cases. The former two are caused by the blurring of the chin and neck areas, causing an almost indiscernible chin landmark position. The Debit Card variability has another reason. Since not all subjects look straight into the camera, some images are taken somewhat in an upright position, causing some annotator dubiety on the location of the chin landmark. Finally, the location of the nasal root exhibits variability, even in the reference images. We think that this is caused by the inherent difficulty of locating this landmark in a frontal view.

Variability of closed and open shapes

In Tables 6.3 and 6.4, we present the total pairwise differences in terms of IPA for some closed and open shapes. For most open and closed shapes we observe a trend that the variability increases with decreasing image quality. We notice that larger scale structures (for example the face outline) exhibit a larger variability than smaller scale structures like the eye fissure, as described in Section 6.3.4. In some cases, the variability increases in the Debit Card case and/or (partially) decreases in the Robbery case(s). We think there are two explanations for

Table 6.3: Total pairwise difference for some closed shapes, expressed as percentage of the IPA.

	Reference	ID Card	Debit Card	Robbery 1	Robbery 2
Eyebrow	1.3%	1.4%	1.4%	1.9%	1.2%
Eye Fissure	0.3%	0.5%	0.5%	0.4%	N/A
Mouth	0.9%	1.1%	2.0%	2.9%	N/A

Table 6.4: Total pairwise difference for some open shapes, expressed as percentage of the IPA.

	Reference	ID Card	Debit Card	Robbery 1	Robbery 2
Face outline	2.8%	2.9%	9.1%	6.6%	11.2%
Hairline boundary	1.4%	1.1%	1.2%	2.2%	3.5%
Cheekbone	1.9%	8.0%	N/A	2.0%	2.0%
Nose outline	0.9%	1.4%	1.7%	2.5%	4.2%
Ear outline	0.3%	0.3%	1.3%	1.7%	3.7%
Chin outline	0.9%	0.7%	3.7%	2.6%	1.0%
Neck outline	0.8%	0.6%	1.1%	4.5%	3.9%

this observation. The former effect is due to a lack of contrast in the black and white images, causing annotator ambiguity; this effect is noticeable in the face outline. The latter effect is caused by the image quality degradation such that some facial features are so distorted that (a) they are still visible, but (b) that there is almost no room for interpretation. In some cases, the image degradation is so severe (for example of the eye fissure), that there exist no measurements at all.

The cheekbone (its location and visual presence) is sometimes difficult to observe in well conditioned images. Also, its precise location is subject to interpretation variability. If we compare the variability with respect to average feature area, we find that the eyebrow exhibits the largest pairwise difference. This might be caused by (a) interpretation issues regarding the location of the exterior eyebrow tip and (b) sensitivity of the eyebrow/skin boundary visibility to contrast and blur. One can argue that the eye fissure and mouth are less sensitive to these factors as they have better discernible boundaries in terms of colour.

Variability of other characteristic descriptors

In Table 6.5, we list a selection of characteristic descriptors derived from landmarks and shapes that represent the encountered variability. In the first five rows, we list four distinct distance measures and the fissure angle. As expected, generally the standard deviation increases as the image quality degrades. In the last three rows, we report three count descriptors. Notice the relationship between image quality and variability, caused by the lack of data in the Robbery cases. Also, we mention that facial marks exhibit the largest variability of all considered count like characteristic descriptors.

Table 6.5: Standard deviations of distances, the fissure angle and some counts. Distances are reported with respect to IPD, the angle is in degrees, and count is dimensionless.

	Reference	ID Card	Debit Card	Robbery 1	Robbery 2
Width nose	0.6%	2.2%	1.7%	2.9%	7.1%
Width mouth	2.1%	4.2%	4.3%	6.1%	6.0%
Nose-mouth distance	1.2%	2.1%	2.6%	1.9%	4.1%
Mouth-chin distance	1.9%	1.9%	8.4%	4.3%	6.8%
Eye fissure angle	1.5°	2.0°	2.4°	2.9°	4.4°
Eye lower folds count	3.3	0.1	0.2	0.0	0.0
Eye upper folds count	1.2	0.0	0.1	0.0	0.0
Facial marks count	5.4	1.2	2.8	0.4	0.0

Table 6.6: Types of annotation variability influence on evidential value.

Influence	Description
$\mathscr{I}_+$	Interval I completely in correct region
$\mathscr{I}_\downarrow$	$\hat{\mu}$ in correct region, I partly in correct region
$\mathscr{I}_\uparrow$	$\hat{\mu}$ in incorrect region, I partly in correct region
$\mathscr{I}_-$	$\hat{\mu}$ in incorrect region, I completely in incorrect region

Evidential value and its variability

For each forensic use case and characteristic descriptor, we construct the set of same source and different source evidential values, using the method explained in Section 6.3.4. In Figure 6.9, we present histograms of the empirical means $\hat{\mu}$ of these two sets for each of the four forensic use cases. We observe that the evidential value of a single characteristic descriptor in general is limited, especially in the Robbery 2 use case. This is in line with the findings presented in [22]. However, in general (a) the combination of characteristic descriptors yields larger evidential value or (b) a single extreme feature value yields high discriminating power.

To each set of same source and different source evidential values having empirical mean $\hat{\mu}$ and standard deviation $\hat{\sigma}$, we associate an interval $I = [\hat{\mu} - 2\hat{\sigma}, \hat{\mu} + 2\hat{\sigma}]$ that captures approximately 95% of the range of evidential values. In order to quantify the influence of annotation variability on the evidential value, we use I and $\hat{\mu}$ to define four distinct influences (Table 6.6). In particular, "$\hat{\mu}$ in correct region" means that the same source and different source evidential value are positive and negative, respectively.

Table 6.7 shows the annotator influence on evidential value. The influences are similar for same source and different source cases. When lowering the image quality, we observe that in general the neutral influence I_+ reduces from around 20% to 7%, whereas the negative influence $I_\downarrow$ and positive influence $I_\uparrow$ reduce from 30% to 10% and from 20% to 10%, respectively. The percentage of I_- rises over 70% in the Robbery 2 use case, or rephrased, 70% of the evidential value intervals yields the wrong conclusion. However, note that these

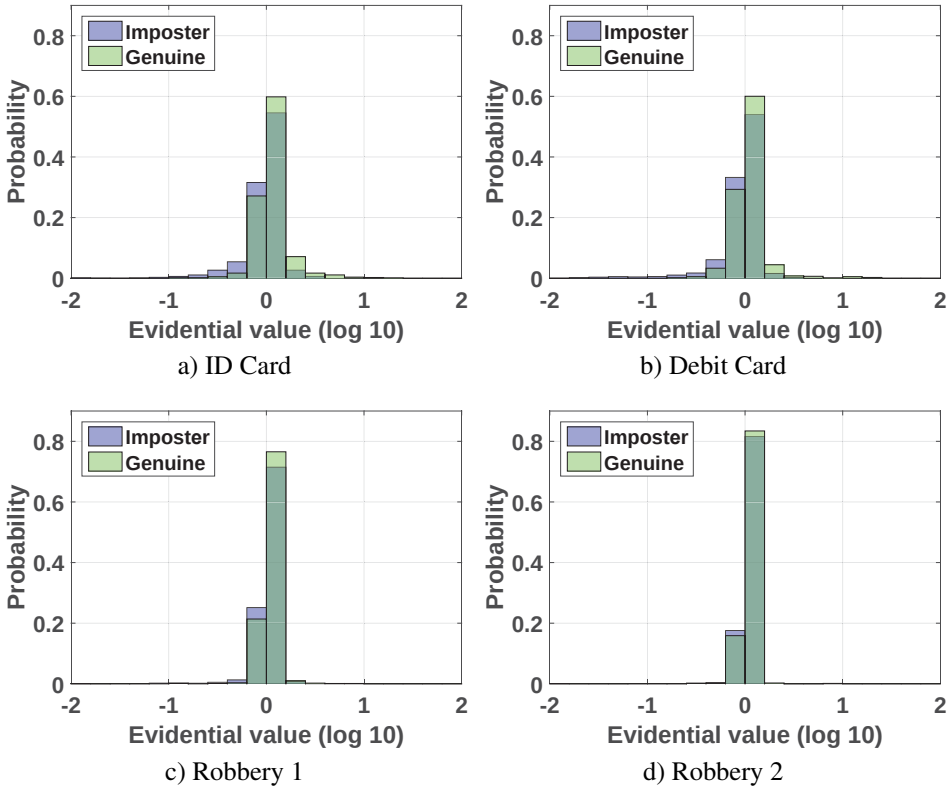


Figure 6.9: Histogram of attained evidential values for same source and different source cases: (a) ID Card, (b) Debit Card, (c) Robbery 1, and (d) Robbery 2.

evidential values are too low to be used in real forensic case work and are caused by using evidential value models that already were shown to have low discriminating power [22].

6.3.6 Conclusion

In this work, we have presented the results of two related experiments using manual annotation from which characteristic descriptors can be derived. With respect to measurability, we found that a large number of characteristic descriptors cannot be determined in images representative of forensic case work and therefore we question its detailed nature.

In general, the variability of landmarks, closed and open shapes (in terms of IPA), and other characteristic descriptors increases when the image quality decreases. In other cases we can explain slightly different variability dependence on image quality.

We found that the evidential value of single characteristic descriptors in general is very limited and this reiterates the results of a related study on discriminating power of FISWG characteristic descriptors. The annotation variability of the characteristic descriptors influences up to 50% of all considered evidential values. In the severest use case, we found that

Table 6.7: Influence of annotator variability on evidential value in percentage total number of characteristic descriptors for each forensic use case.

Type	ID Card	Debit Card	Robbery 1	Robbery 2
Same Source $\mathscr{I}_{+}$	17.5%	20.7%	14.3%	6.6%
Same Source $\mathscr{I}_{\downarrow}$	29.1%	31.1%	18.7%	12.1%
Same Source $\mathscr{I}_{\uparrow}$	21.4%	20.3%	15.6%	9.5%
Same Source $\mathscr{I}_{-}$	32.0%	27.9%	51.5%	71.9%
Same Source Total	100%	100%	100%	100%
Different Source $\mathscr{I}_{+}$	13.5%	14.5%	7.5%	7.4%
Different Source $\mathscr{I}_{\downarrow}$	30.5%	29.7%	20.2%	10.9%
Different Source $\mathscr{I}_{\uparrow}$	19.5%	22.5%	14.9%	9.9%
Different Source $\mathscr{I}_{-}$	36.5%	33.2%	57.3%	71.8%
Different Source Total	100%	100%	100%	100%

more than 70% of the evidential value intervals in principle could yield the wrong conclusion.

6.4 Chapter conclusion

In this chapter, two studies have been combined. They addressed research question 1c: *Under various forensic use cases, what is the general performance of biometric classifiers that use FISWG characteristic descriptors as their input and produce strength of evidence in relation to face recognition systems?* and research question 1d: *Under various forensic use cases, what is (a) the measurability of FISWG characteristic descriptors and (b) the influence of annotation variation on characteristic descriptors and strength of evidence produced by biometric classifiers that use these characteristic descriptors?*

In the first study, we have investigated the discriminating power of biometric classifiers that use FISWG characteristic descriptors as their input under various forensic use cases. In every forensic use case we found characteristic descriptors, either single or combined, that yield moderate to poor discriminating power. In all forensic use cases, the characteristic descriptors were outperformed by a face recognition system, except in the severest case (IPD 11px). In that case, the hairline/forehead boundary as a single characteristic descriptor and the combination of availability features performed somewhat better than both considered face recognition systems. Nonetheless, their discriminating power was poor. Also, their strength of evidence found by inspecting their ROC curves was not convincing.

The second study presented the results of two related experiments regarding manual annotation from which characteristic descriptors can be derived. With respect to measurability, there was a tendency to have lower measurability when the image quality decreased, with some exceptions due to for example large scale structures (Face Head Outline, Hairline/Baldness pattern) and differences in illumination (Forehead). In general, we found that a large number of characteristic descriptors cannot be determined in images representative of forensic case work. The variability of landmarks, closed and open shapes (in terms of

Interpupillary Area[1]), and other characteristic descriptors increased when the image quality decreased. As in the measurability case, some explainable exceptions exist. They include decreased variability in the Robbery 2 case, probably caused by a lower, but more stable, number of annotations. We found that the strength of evidence of single characteristic descriptors in general is very limited and it reiterated the results of the first study of this chapter. The influence of annotator variability on the strength of evidence generally decreased with decreasing image quality, but this is certainly not an asset. In fact, it was shown that up to 70% of the evidential value intervals *completely* lie in the wrong region.

With respect to the addressed research questions, we find the following.

The results of the first study shows that in many cases, in terms of discriminating power, it is better to use a face recognition system. This goes against the desire to use features with forensic semantics instead of abstract and general features. An exception to these results occurs in the severest case in which the strength of evidence produced by a biometric classifier is poor, but is still better than the used face recognition systems. We overall conclude that in various forensic use cases, classifiers that use FISWG characteristic descriptors exhibit moderate to poor discriminating power and are mostly outperformed by a face recognition system. Even if they are better than such a system, their produced strength of evidence is poor.

The results of the second study show, especially in the severest cases, that the number of characteristic descriptors that can be measured is low. Their usability and especially their detailed nature as described in the FISWG Feature List [8] can be questioned. Although one could argue that characteristic descriptors can be extracted on faces that are partially occluded, the results of Chapter 4 also indicate that other, non-forensic, features and classifiers might be applicable in that situation as well. The results of the second study also show the general negative influence of annotator variability on the variability of landmarks, shapes, and other characteristic descriptors when the quality of the trace images decreases. This effect is particularly visible in the Robbery Use Cases. A decreasing effect in variability of strength of evidence is observable when the quality of the trace images decreases, but the increased percentage of evidential value intervals that fully lie in the wrong region overshadows this result.

One could argue that a general performance evaluation does not fully do justice to the fact that some subjects might be discriminated from others based on an extreme value or occurrence of specific biometric traits, while in general such biometric traits have limited discriminating power. Examples are an extreme angle for the eye fissure opening or protruding ears. An alternative approach is to accommodate for a subject based approach. In such an approach, biometric classifiers are evaluated based on how they perform when traces only originate from a single subject. Also, it can be taken even a step further by training classifiers only on data of a single subject. In Chapter 7, this subject based approach is studied as a response to existing facial mark classifier studies and it yields a proto-framework that includes subject based training and evaluation. This chapter also includes a theoretical construction of an extreme discrepancy between general and subject based performance. In Chapter 8, the proto-framework is further expanded into a general framework that considers this subject based approach from a broader perspective.

[1] The 2-dimensional analog to Interpupillary Distance

Chapter 7

Subject based: facial marks and a theoretical construction

7.1 Introduction

We argued in the conclusion of Chapter 6 that only performing a general evaluation of discriminating power does not do justice to the fact that some subjects might be discriminated on biometric traits that in general have moderate to poor performance. This idea is further explored in this chapter in (a) the practical context of facial marks and (b) the theoretical context of a construction that shows extreme general versus subject based performance. Facial marks are particularly interesting from a forensic perspective, since they are representative of a class of potentially highly discriminating features. This chapter addresses research question 2a: *To which extent do we observe or can we construct differences in general and subject based performance?* and research question 2b: *How well can facial marks be used for forensic evaluation, also taking subject based data and subject based evaluation into account?*

The first part of this chapter contains a comparative study of six biometric classifiers types that use features based on facial mark spatial patterns. The study identifies six, mostly forensic, aspects that are hardly considered in other studies on facial marks, and these aspects are summarised in a proto-framework. The aspects include (a) the explicit use of subject based data in classifier training, (b) the incorporation of a subject based evaluation, and (c) the use of forensically relevant performance characteristics and metrics. The influence of several of those aspects on two performance characteristics (discriminating power and calibration) is systematically investigated. The second part complements the facial mark results and advocates a broader subject based performance evaluation by presenting a theoretical construction.

Section 7.2 has been accepted for publication as "Grid Based Likelihood Ratio Classifiers for the Comparison of Facial Marks" [24].

Section 7.3 has been published as "Label specific versus general classifier performance: an extreme example" [25].

Reading Guide
Section 7.2. This section should at least be browsed, in particular the reader should get acquainted with the six aspects and the differences between general and subject based performance.
Section 7.3. The theorem contained in this section should at least be read, the remainder can be omitted.

7.2 Grid based likelihood ratio classifiers for the comparison of facial marks

7.2.1 Abstract

Facial marks have been studied before, either as a complement to face recognition systems or for their suitability as a single biometric modality. In this work, we use a subset of the FRGCv2 dataset (12307 images, 568 subjects) to study properties of facial marks, their spatial patterns, and classifiers acting upon these patterns. We observe differences between age and ethnic groups in the number of facial marks. Also, facial marks tend to be clustered. We present six forensically relevant aspects with respect to the design and evaluation of classifiers. These aspects help to systematically study factors that influence performance characteristics (discriminating power and calibration loss) of these classifiers. Calibration loss is of particular forensic importance; it essentially measures how well the classifier output can be used as strength of evidence in a court of law. We use various facial mark grids to which the facial mark spatial patterns are assigned. We find that a classifier that utilises the facial mark grid of a specific subject outperforms all other classifiers. We also observe that the calibration loss of such subject based classifier indicates that that small grid cell sizes should be avoided.

7.2.2 Introduction

Facial marks and the spatial patterns they form have been studied before as a soft biometric [144] and as a biometric modality on their own [102]. Mugshot databases can be queried for matches to a facial mark spatial pattern of a perpetrator [50]. Facial marks can also be used to establish strength of evidence during forensic casework.

Strength of evidence is the outcome of a forensic evidence evaluation process in which crime scene trace images and reference images are compared by the FFR-examiner. Strength of evidence is commonly expressed as a log-likelihood ratio[1]:

$$\mathrm{LLR}(E) = \log_{10}\left(\frac{\mathrm{p}(E|\mathscr{H}_\mathrm{s})}{\mathrm{p}(E|\mathscr{H}_\mathrm{d})}\right). \tag{7.1}$$

Here $\mathrm{p}(E|\mathscr{H}_\mathrm{s})$ is the computed or estimated probability to observe evidence E under the same source hypothesis $\mathscr{H}_\mathrm{s}$ (“trace and reference originate from a common donor”). In the likelihood ratio, $\mathrm{p}(E|\mathscr{H}_\mathrm{s})$ is taken relatively to $\mathrm{p}(E|\mathscr{H}_\mathrm{d})$: the probability to observe evidence E under the different source hypothesis $\mathscr{H}_\mathrm{d}$ (“trace and reference do not have a common

[1]The non-log version is referred to as $\mathrm{LR}(E)$. The advantage of the log is that it emphasizes the magnitude of the LR, rather than its exact value.

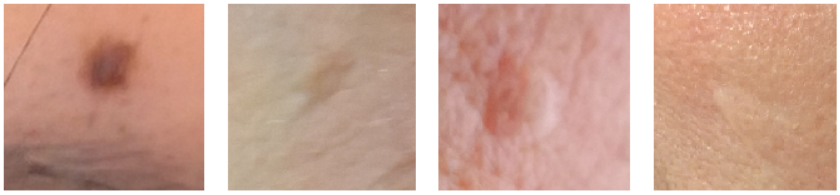


Figure 7.1: From left to right: mole, pockmark, raised skin, and scar. Images taken by first author.

donor"). Evidence E might be the similarity of facial mark spatial patterns in trace and reference images, possibly expressed as a biometric comparison score. The forensic face examiner computes $\mathrm{LR}(E)$; a court of law sets the prior odds $\frac{\mathrm{p}(\mathscr{H}_\mathrm{s})}{\mathrm{p}(\mathscr{H}_\mathrm{d})}$ and determines the posterior odds $\frac{\mathrm{p}(\mathscr{H}_\mathrm{s}|E)}{\mathrm{p}(\mathscr{H}_\mathrm{d}|E)}$ based on the prior odds and the likelihood ratio according to Bayes' rule:

$$\begin{aligned} \frac{\mathrm{p}(\mathscr{H}_\mathrm{s}|E)}{\mathrm{p}(\mathscr{H}_\mathrm{d}|E)} &= \frac{\mathrm{p}(E|\mathscr{H}_\mathrm{s})}{\mathrm{p}(E|\mathscr{H}_\mathrm{d})} \times \frac{\mathrm{p}(\mathscr{H}_\mathrm{s})}{\mathrm{p}(\mathscr{H}_\mathrm{d})} \\ &= \mathrm{LR}(E) \times \frac{\mathrm{p}(\mathscr{H}_\mathrm{s})}{\mathrm{p}(\mathscr{H}_\mathrm{d})}. \end{aligned} \tag{7.2}$$

Although in general face recognition systems can be used for investigation and intelligence purposes, forensic evidence evaluation is still largely a manual process [12]. Moreover, face recognition systems use abstract features like LBP [5]. These features are not endowed with any forensic meaning and require careful explanation outside a technical domain, in particular in a court of law.

In this paper, we (a) investigate properties of facial marks including their spatial patterns and (b) systematically compare classifiers that use the location of facial marks to produce a comparison score that can be interpreted as strength of evidence (7.1). The major advantage of these classifiers is that both their input and output can be understood by a court of law. This work is part of a series of studies in which we investigate classifiers that use FISWG characteristic descriptors (forensic facial features) [8] and which produce strength of evidence [22, 23].

The design and evaluation of classifiers is the outcome of a process that addresses six aspects with a forensic relevance. Figure 7.2 visually presents these aspects. They can be grouped into three groups of two aspects; the groups influence the feature, score, and the performance evaluation, respectively.

The six aspects in this paper are:

(i) **Aspect 1: Which facial mark types to consider?** A facial mark is a patch of skin that does not resemble the skin in its neighborhood. Srinivas et al. [102] and related studies identify approximately ten facial mark types. Some types are sensitive to (a) time lapse (acne or seasonal dependence of freckles), (b) image illumination (light or dark patches), and (c) image resolution and blurring (small facial marks). We consider

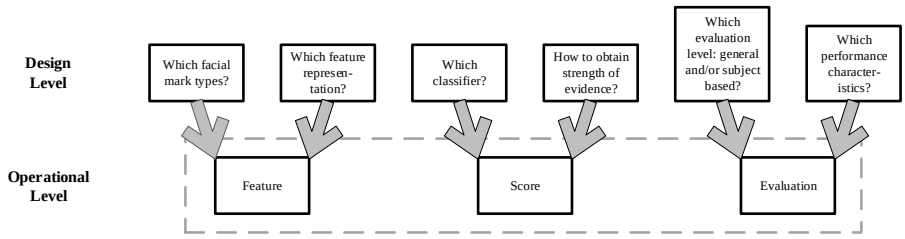


Figure 7.2: Addressed aspects ("design level") that influence the components of an evaluated biometric system ("operational level"). An arrow denotes a unidirectional "influences" relationship.

prominent and permanent facial marks such as moles, pockmarks, raised skin, and scars (Figure 7.1).

(ii) **Aspect 2: Which facial mark feature representations to consider?** The facial mark spatial pattern itself can be used as a feature. However, a face without any facial marks cannot be represented. In this study, we superimpose grids with various grid cell sizes on a facial image. We consider various feature representations (*facial mark grids*) based on the number of facial marks every grid cell contains.

(iii) **Aspect 3: Which classifier types to consider?** If classifiers require training, they can be trained using statistical properties of either facial mark grids in general (*general classifier*) or facial mark grids of a single subject (*subject based classifier*). Their application is different: general and subject based classifiers output whether trace and reference originate from the same and this specific subject, respectively. In this study, we consider both types, as well as classifiers that do not require training.

(iv) **Aspect 4: How to obtain strength of evidence?** Classifiers produce scores and typically do not produce likelihood ratios. Therefore, a score cannot be used in a court of law as strength of evidence. We refer to Ramos and Gonzalez-Rodriguez [145] for more details. We consider differences between classifiers that produce scores which either are converted into a likelihood ratio or can be interpreted as a likelihood ratio.

(v) **Aspect 5: Which evaluation levels to consider?** Apart from doing a *general evaluation* (using traces and references of multiple subjects), we also evaluate at a subject level. We refer to this as *subject based evaluation*. At this level, performance is measured using trace-reference pairs for which the traces only originate from a specific subject and the references come from multiple subjects. This can show that some subjects can be discriminated well based on their facial marks grid, whereas a general evaluation indicates a moderate performance. Any classifier can be evaluated at subject level.

(vi) **Aspect 6: Which performance characteristics to consider?** A forensic guideline by Meuwly et al. [48], to form a basis for an ISO standard, presents "performance characteristics" for the validation of likelihood ratio methods for *forensic* evaluation.

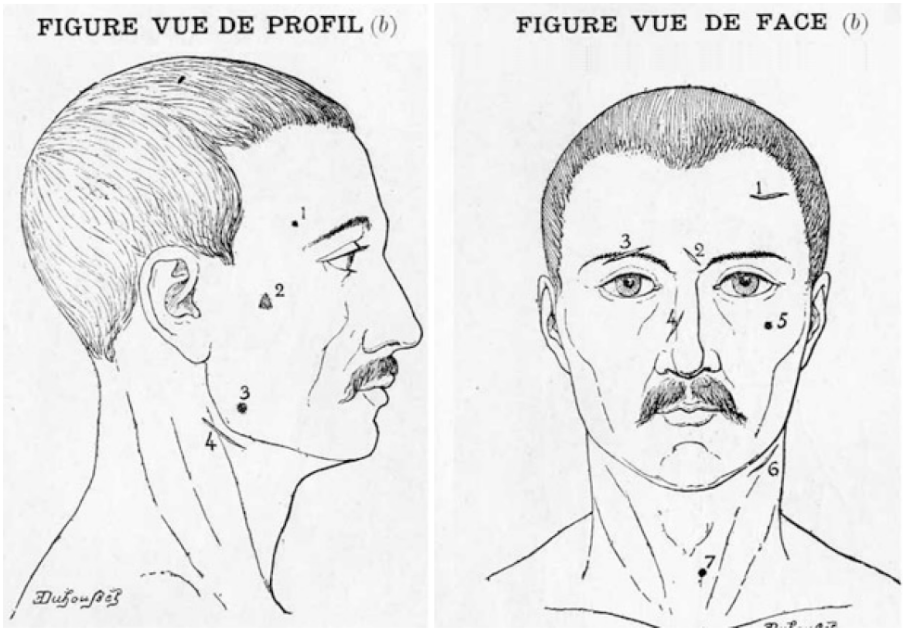


Figure 7.3: Examples of facial scars and marks in the Bertillonage system. Left: Profile View, Right: Frontal View. Taken from [53].

In this paper, we investigate discriminating power (EER) and calibration loss ($Cllr^{cal}$) of the computed likelihood ratio. Calibration loss essentially measures how well the computed likelihood ratio can be used as strength of evidence in a court of law.

We address three research questions in this work.

RQ1 What are demographic and spatial properties of facial marks?
RQ2 What is the influence of choices addressed in Aspects 2 to 4 and the number of facial marks on the discriminating power of facial mark grid based facial mark classifiers?
RQ3 What is the influence of choices addressed in Aspects 2 and 3 and the number of facial marks on the calibration loss of facial mark grid based facial mark classifiers?

The paper is structured as follows. In Section 7.2.3 we discuss related work. In Section 7.2.4 we present the methods and Section 7.2.5 describes the experiments. The results are presented and discussed in Section 7.2.6. The conclusion and future work are presented in Section 7.2.7.

7.2.3 Related Work

The Bertillonage system [53,54] is the first modern system that uses facial features to describe criminals. It includes the description of facial scars and marks; Figure 7.3 shows an example. We refer to Evison [55] for other examples.

FISWG [7] recommends the use of shape like features [6], possibly in conjunction with superposition, during a facial comparison. Several studies [82–85] show that *in general* facial measurements, either 2D or 3D, either photometric or in vivo, are not suitable for forensic evaluation. FISWG characteristic descriptors include facial marks. More details on operational procedures used at different forensic institutes are given in Spaun [57], Prince [12], and ENFSI [61].

The modern forensic relevance of facial marks is presented by a survey by Jain et al [50]. They discuss (a) robustness to facial aging, (b) matching forensic (composite) sketches to face photograph databases, and (c) image retrieval using facial scars and marks. Facial marks fit in a facial taxonomy proposed by Jain and Klare [146], comparable to a taxonomy of fingerprints. At the first level the holistic face is considered, at the second level facial parts are taken into account, and at the third level facial marks can be used. A study by Lin and Tang [147] follows this multi-level approach. They use an adapted version of Linear Discriminant Analysis [117] as a global feature descriptor and SIFT [4] as a local feature descriptor.

In several studies by Srinivas et al. [148,149], facial marks are used to distinguish between mono zygotic twins; in Shalin et al. [150] this problem is analysed using other biometric modalities as well. In Biswas et al. [151], the ability of humans to distinguish between mono zygotic twins is studied.

In Park and Jain [101], an automatic facial mark detection system is presented that uses an Active Appearance Model [152] and a Laplacian of Gaussian blob detector. They show, using several face databases, that the incorporation of facial marks can improve face-recognition performance of a state-of-the-art face recognition system. In another study by Srinivas et al. [102], the Fast Radial Symmetry Transform [153] is used for the detection of facial marks. They compare their automatically detected facial marks with those found by a system in which the facial mark locations and types have been manually post-processed. On their High Resolution Face Database (HRFD), they report an EER of 12%. In general, detection of facial marks is a challenging problem as it tends to produce a large number of false positives.

Some studies restrict the number of facial mark types. In Pierrard and Vetter [154], a framework for the detection of prominent moles (nevi) is presented. It uses a combination of Grabcut like segmentation and a saliency measure that uses the uniqueness of the detected mark in relation to its neighborhood. A series of studies by Nurhudatiana et al. [155–157] consider Relatively Permanent Pigmented or Vascular Skin Marks (RPPVSM) as a means to individualise in for example cases of child sexual abuse. Their studies are confined to the back torso and they (a) find that middle to low density patterns tend to be uniformly distributed, whereas high density patterns tend to be more clustered, (b) present a model for which error rates tend to be very small, (c) present an automatic RPPVSM detector and extractor, and (d) find that dermatology physicians have the capacity to act as an RPPVSM examiner.

7.2.4 Methods

Feature

We restrict our measurements to the facial mark spatial pattern, in line with the studies by Nurhudatiana et al [155]. As indicated in the Introduction, the spatial pattern itself is less practical as a feature, as images that do not contain facial marks cannot be represented. We

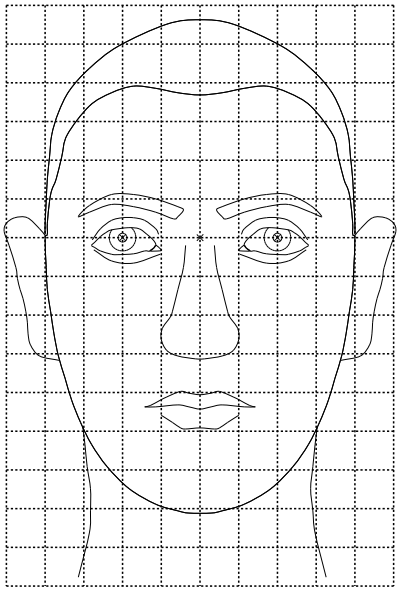


Figure 7.4: Grid shown for $\Delta = 0.25$ IPD.

explore various feature representations based on the facial mark spatial pattern.

We assume that the facial mark spatial pattern is given in a coordinate system in which the pupils coordinates are $(-\frac{1}{2},0)$ and $(\frac{1}{2},0)$, making the interpupillary distance (IPD) always 1. We superimpose a grid with square cells having size Δ on this coordinate system. Figure 7.4 shows the grid for $\Delta = 0.25$ IPD. The neck is also covered by this grid, motivated by its inclusion in [8]. Given parameters $x_{\min}$, $x_{\max}$, $y_{\min}$, and $y_{\max}$ that mark the boundaries of the grid, and a value for Δ, the number of grid cells in the horizontal direction I and vertical direction J is given by:

$$I = \lceil \frac{x_{\max} - x_{\min}}{\Delta} \rceil \text{ (resp. } J = \lceil \frac{y_{\max} - y_{\min}}{\Delta} \rceil). \tag{7.3}$$

For $i \in \{1,\ldots,I\}$ and $j \in \{1,\ldots,J\}$, we define the grid cell g_{ij} as the square:

$$g^{ij} = [x_{\min} + (i-1)\Delta, x_{\min} + i\Delta] \times [y_{\min} + (j-1)\Delta, y_{\min} + j\Delta] \tag{7.4}$$

and G_Δ as the collection of grid cells:

$$G_\Delta = \{g^{ij} \mid i \in \{1,\ldots,I\},\ j \in \{1,\ldots,J\}\}. \tag{7.5}$$

Given a facial mark spatial pattern $F = \{(x_k, y_k)\}$ and G_Δ, we define the category facial mark grid feature $c = (c^{ij})$ as:

$$c^{ij} = \text{Category}(\#(F \cap g^{ij})). \tag{7.6}$$

The category is a mapping from $\mathbb{N}$ to $\mathbb{N}$ that assigns to every number of facial marks in a grid cell a category. A special case of (7.6) is the binary facial mark grid feature $b = (b^{ij})$ in

which the mapping is given by:

$$\text{Category}(x) = \begin{cases} 1, \text{ if } x \geq 1 \\ 0, \text{ otherwise.} \end{cases} \tag{7.7}$$

Score

As indicated in the Introduction, strength of evidence is commonly expressed as a log-likelihood ratio:

$$\text{LLR}(E) = \log_{10}\left(\frac{\text{p}(E|\mathscr{H}_\text{s})}{\text{p}(E|\mathscr{H}_\text{d})}\right). \tag{7.8}$$

If the evidence E is a biometric comparison score $s = s(x,y)$ computed on trace x and reference y, (7.8) reverts to a *score based log-likelihood ratio*:

$$\text{LLR}(s) = \log_{10}\left(\frac{\text{p}(s|\mathscr{H}_\text{s})}{\text{p}(s|\mathscr{H}_\text{d})}\right). \tag{7.9}$$

There exist several techniques to estimate (7.9) through the estimation of the numerator and denominator. They include parametric, semi-parametric [45], and non-parametric [46] approaches. The Pool of Adjacent Violators (PAV) algorithm [47] is a commonly used approach to directly determine (7.9). This algorithm estimates $\text{p}(\mathscr{H}_\text{s}|s)$ from which the log-likelihood $\text{LLR}(s)$ can be computed:

$$\text{LLR}(s) = \text{logit}(\text{p}(\mathscr{H}_\text{s}|s)) - \text{logit}(\text{p}(\mathscr{H}_\text{s})), \tag{7.10}$$

where $\text{logit}(x) = \log_{10}\left(\frac{x}{1-x}\right)$ is the log odds function. In general, the conversion from a score to the (log) likelihood ratio of the score is referred to as score calibration [91]. We present two classifiers whose scores are calibrated using (7.10) in Section 7.2.5.

If the evidence E is the simultaneous occurrence of trace x and reference y, we obtain the *feature based log-likelihood ratio*:

$$\text{LLR}(x,y) = \log_{10}\left(\frac{\text{p}(x,y|\mathscr{H}_\text{s})}{\text{p}(x,y|\mathscr{H}_\text{d})}\right). \tag{7.11}$$

In addition to using either (7.9) or (7.11) to compute strength of evidence, we are also interested in the likelihood ratio due to a well known theoretical result. The Neyman-Pearson lemma states that a likelihood ratio based classifier has the largest possible True Match Rate for a given False Match Rate.

In the context of facial marks, we assume that the number of possible (trace x, reference y) value combinations is finite for every grid cell g^{ij}. We define p_{xy}^{ij} and q_{xy}^{ij} as the probabilities of observing the (trace x, reference y) value pair in grid cell g^{ij} under the same source and different source hypothesis, respectively. By definition, these states follow a categorical distribution[2] that lists the probability for every (trace x, reference y) value pair. This distribution is the multi-state generalisation of the Bernoulli distribution.

[2] Some authors would refer to this distribution as multinomial. This is not correct, since a multinomial distribution involves a *series* of draws from a categorical distribution. This distinction is a generalisation of the distinction between a Bernoulli and binomial distribution.

Table 7.1: Overview features and classifiers

Values Grid Cell	Classifier	Score Function	Trained
$0,1$	Hamming	$-\sum_{i,j} \lvert x_1^{ij} - y_2^{ij} \rvert$	No
$0,1$	Bin LLR Gen	$\sum_{i,j} \log_{10}(p^{ij}_{x_{ij}y_{ij}}) - \log_{10}(q^{ij}_{x_{ij}y_{ij}})$.	Yes, general data
$0,1$	Bin LLR Subject	$\sum_{i,j} \log_{10}(p^{ij}_{x_{ij}y_{ij}}) - \log_{10}(q^{ij}_{x_{ij}y_{ij}})$	Yes, subject based data
# Facial Marks	χ^2	$-\sum_{i,j} \frac{(x^{ij}-y^{ij})^2}{x^{ij}+y^{ij}}$.	No
$0,1,2-3,\geq 4$	Cat LLR Gen	$\sum_{i,j} \log_{10}(p^{ij}_{x_{ij}y_{ij}}) - \log_{10}(q^{ij}_{x_{ij}y_{ij}})$.	Yes, general data
$0,1,2-3,\geq 4$	Cat LLR Subject	$\sum_{i,j} \log_{10}(p^{ij}_{x_{ij}y_{ij}}) - \log_{10}(q^{ij}_{x_{ij}y_{ij}})$	Yes, subject based data

Under the assumption that the number of facial marks in a grid cell is independent of the number of facial marks in any other grid cell, (7.11) reverts to:

$$\begin{aligned}\mathrm{LLR}(x,y) &= \log_{10}\left(\frac{\mathrm{p}(x,y|\mathscr{H}_\mathrm{s})}{\mathrm{p}(x,y|\mathscr{H}_\mathrm{d})}\right)\\ &= \log_{10}\left(\frac{\prod_{i,j} p^{ij}_{x_{ij}y_{ij}}}{\prod_{i,j} q^{ij}_{x_{ij}y_{ij}}}\right) = \log_{10}\left(\prod_{i,j} \frac{p^{ij}_{x_{ij}y_{ij}}}{q^{ij}_{x_{ij}y_{ij}}}\right)\\ &= \sum_{i,j} \log_{10}(p^{ij}_{x_{ij}y_{ij}}) - \log_{10}(q^{ij}_{x_{ij}y_{ij}}).\end{aligned} \tag{7.12}$$

Evaluation

We use primary forensic performance characteristics as presented in Meuwly et al. [48] to evaluate the classifiers.

Discriminating power is a "property representing the capability of a given method to distinguish amongst forensic comparisons where different propositions are true" [48]. In this work, we use the EER to explore discriminating power.

Given a set $\mathscr{S}$ of n_s same source and a set $\mathscr{D}$ of n_d different source) scores under the same source hypothesis $\mathscr{H}_s$ and different source hypothesis $\mathscr{H}_d$ respectively, the cost of log-likelihood ratio [49] is defined by:

$$\mathrm{Cllr} = \frac{1}{2}\left(\frac{1}{n_s}\sum_{s\in\mathscr{S}} \log_2(1+\mathrm{e}^{-s}) + \frac{1}{n_d}\sum_{s\in\mathscr{D}} \log_2(1+\mathrm{e}^{s})\right). \tag{7.13}$$

This performance characteristic measures both discriminating power and calibration. Calibration is a "property of a set of LRs (...)" [48]. Perfect calibration implies that scores can be interpreted as strength of evidence. A commonly used measure for calibration is calibration loss $\mathrm{Cllr}^{\mathrm{cal}}$. If we apply the PAV algorithm to the set of scores and reapply (7.13), we obtain the minimal achievable cost of likelihood ratio $\mathrm{Cllr}^{\mathrm{min}}$. The difference $\mathrm{Cllr}^{\mathrm{cal}} = \mathrm{Cllr} - \mathrm{Cllr}^{\mathrm{min}}$ is the calibration loss and it measures how well calibrated the original scores are. A lower value for $\mathrm{Cllr}^{\mathrm{cal}}$ means better calibration. We refer to [48] for a discussion of other forensically relevant characteristics.

Evaluation characteristics have graphical representations. Discriminating power is typically represented by an ROC curve or DET plot [158]. Calibration can be visualised by a

Tippett Plot or ECE plot [145, 159], but since we are interested in a subject based evaluation, we use Tukey boxplots [160] to visualise the range of EER and Cllr^{cal} values.

Facial mark spatial pattern analysis

Every subject S_i has n_i facial mark spatial patterns S_i^j containing $\ell_i^j \geq 0$ facial marks, $j = 1,\ldots,n_i$:

$$\begin{aligned} S_i^1 &= \{(x_{i,k}^1, y_{i,k}^1) \mid k = 1,\ldots,\ell_i^1\}, \\ \ldots &\quad \ldots \\ S_i^{n_i} &= \{(x_{i,k}^{n_i}, y_{i,k}^{n_i}) \mid k = 1,\ldots,\ell_i^{n_i}\}. \end{aligned} \tag{7.14}$$

For every subject S_i, we randomly select indices g_i and s_i such that:

$$g_i, s_i = 1,\ldots,n_i \text{ and } g_i \neq s_i. \tag{7.15}$$

We define the general facial mark spatial pattern $\mathscr{G}$ as

$$\mathscr{G} = \cup_i S_i^{g_i} \tag{7.16}$$

and a subject based facial mark spatial pattern $\mathscr{S}_i$ as

$$\mathscr{S}_i = S_i^{s_i}. \tag{7.17}$$

The uniform pattern $\mathscr{U}_f$ is any spatial pattern of f facial marks, randomly sampled from a uniform spatial probability distribution on $[x_{\min}, x_{\max}] \times [y_{\min}, y_{\max}]$. Finally, $\mathscr{G}_f$ is any spatial pattern of f facial marks, randomly sampled from $\mathscr{G}$.

In order to measure the spatial pattern properties mentioned in RQ1, we empirically show that the following three hypotheses are true:

- H_{GU}: The general facial mark spatial pattern $\mathscr{G}$ is not sampled from a uniform pattern $\mathscr{U}_f$, where f is the number of facial marks in $\mathscr{G}$.
- H_{SU}: The subject based facial mark spatial patterns $\mathscr{S}_i$ with f facial marks are more clustered than $\mathscr{U}_f$.
- H_{SG}: The subject based facial mark spatial patterns $\mathscr{S}_i$ with f facial marks are more clustered than $\mathscr{G}_f$.

H_{GU} is a member of the Complete Spatial Randomness Hypothesis; we test this hypothesis by the quadrat counts method (QCM) [155]. The QCM is essentially a Pearson Goodness-of-fit test in which a face containing f marks is partitioned into N regions of equal size (quadrats). The number of facial marks N_i in every quadrat is compared to the expected number of facial marks $E_i = E$ under the assumption that the facial marks are drawn from $\mathscr{U}_f$. The χ^2 statistic with $N-1$ degrees of freedom

$$\chi^2 = \sum_{i=1}^{N} \frac{(N_i - E_i)^2}{E_i} \tag{7.18}$$

is used to test this assumption.

The other hypotheses (H_{SU} and H_{SG}) use clustering. Given a facial mark spatial pattern $F = \{(x_k, y_k) \mid k = 1, \ldots, f\}$, having $f \geq 2$ facial marks, we define its clustering as:

$$\text{Clustering}(F) = \frac{1}{f} \sum_{k=1}^{f} d_{min}\left((x_k, y_k), F_{\setminus k}\right), \tag{7.19}$$

where d_{min} measures the minimum Euclidian distance between a facial mark location (x_k, y_k) and its neighbors $F_{\setminus k}$. Given the number of facial marks $f \geq 2$, we compute for every subject based facial mark spatial pattern $\mathscr{S}_i$ having f facial marks its clustering. We randomly sample spatial patterns $\mathscr{U}_f$ and $\mathscr{G}_f$ and compute the empirical probability of their clusterings. Using (7.18), we can test the assumption whether the facial mark spatial patterns $\mathscr{S}_i$ with f facial marks are more clustered than $\mathscr{U}_f$ (H_{SU}) and $\mathscr{G}_f$ (H_{SG}) facial mark spatial patterns.

7.2.5 Experiments

Dataset

In this study, we employ a subset of the FRGCv2 dataset [33]. This dataset has been used in many face recognition studies and algorithm evaluations. In this work, we only use 2D images taken under controlled conditions, showing subjects with a neutral expression. This yields a set of 12307 images of 568 subjects with an average of 360px IPD. The reason to use this dataset is the trade off between image quality, the number of subjects, and the number of available images for (most) subjects.

Experiment 1-Facial Mark Properties

In Experiment 1, we acquire the facial mark spatial patterns and study their spatial and demographic properties.

Prior to the acquisition, we crop every image into a window defined by $[x_{\text{left}} - \text{IPD}, x_{\text{right}} + \text{IPD}] \times [y_{\text{top}} - 1.5\ \text{IPD}, y_{\text{bottom}} + 2.5\ \text{IPD}]$. The IPD is based on pupil coordinates provided by the FRGCv2 dataset.

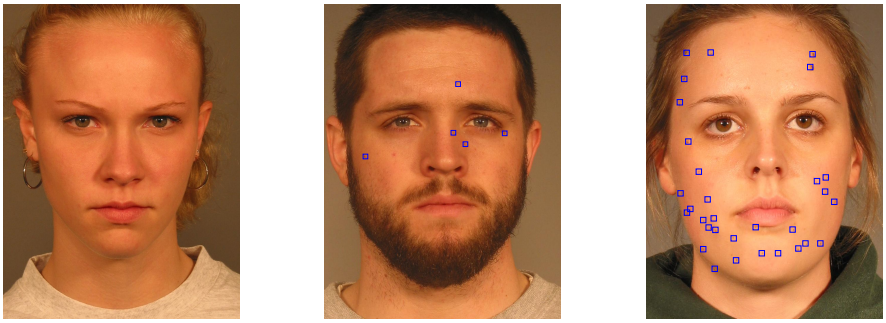
The facial mark spatial patterns are selected manually. We explored the approaches taken in [102] (Fast Radial Symmetry Transform) and [101] (Laplacian of Gaussian), but found that the automatic detection gave an unsatisfactorily large number of false positives. A software application presents the images in random order. A single participant inspects every image and indicates the facial mark spatial pattern.

We test hypothesis H_{GU} on a 6×4 quadrat grid. Hypotheses H_{SU} and H_{SG} are tested for different numbers of facial marks $f \geq 2$ by randomly sampling 5000 $\mathscr{U}_f$ and 5000 $\mathscr{G}_f$) facial mark spatial patterns, respectively.

Experiment 2-Discriminating power

In Experiment 2, we study the influence of choices in Aspects 2 to 4 and the number of facial marks on discriminating power.

We transform the facial mark spatial patterns such that the pupil coordinates are mapped to $(-\frac{1}{2}, 0)$ and $(\frac{1}{2}, 0)$. We set $x_{\min} = -1.5$, $x_{\max} = 1.5$, $y_{\min} = -3$, and $y_{\max} = 6$ to mark the boundaries of the grid. Furthermore, we consider 20 grid cell sizes $\Delta = 0.05, 0.10, \ldots, 1.0$ IPD.



a) Subject 04379 b) Subject 04919 c) Subject 04300

Figure 7.5: Following the subject identification scheme of FRGCv2, from left to right: a) Subject 04379, no facial marks, b) Subject 04919, average number of facial marks, c) Subject 04350, large number of facial marks.

Table 7.1 lists the feature representations (Aspect 2) and the classifiers (Aspect 3). We use binary features and two types of category features (Table 7.1, first column). We found after experimentation with several category mappings (7.6), that the 4 categories $0, 1, 2-3$ and ≥ 4 yielded good results while the number of categories was kept low. We use the feature based log-likelihood ratio (LLR) described by (7.12) as an LLR classifier, both in the binary and category cases. Finally, we also use the Hamming and χ^2 classifiers that do not require any training. We refer to Sections 7.2.5 and 7.2.5 for details on training and testing.

Aspect 4 addresses how we obtain strength of evidence. As described in Section 7.2.4, we use PAV calibration and the transformation (7.10) on the Hamming and χ^2 scores, and assume that the LLR classifiers in Table 7.1 produce a likelihood ratio. This assumption is tested in Experiment 3. We use the Validation Toolbox beta v1.07 release [161] to calibrate the Hamming and χ^2 scores and to compute EER.

Experiment 3-Calibration Loss

Experiment 3 is the same as Experiment 2 except for two differences. First, we consider the calibration loss and to what extent it is influenced by Aspects 2, 3, and the number of facial marks. Second, we do not consider the Hamming and χ^2 classifiers as their scores have been calibrated. We use the Validation Toolbox to compute Cllr^{cal}.

Classifier Training

The FRGCv2 subset is partitioned into a training and a test set so that we can compare the classifiers in the same manner, including the subject based classifiers.

The general training set G^{TR} is formed by the facial mark patterns of 376 subjects who have less than 25 facial mark spatial patterns.

For the remaining 192 subjects S_i having $n_i \geq 25$ facial mark patterns S_i^j, we define a

training and a test set:

$$S_i^{TR} = \{S_i^j \mid j = 1,\ldots,10\}, \tag{7.20}$$

$$S_i^{TE} = \{S_i^j \mid j = 11,\ldots,n_i\}. \tag{7.21}$$

Given a grid cell size Δ, prior to training and testing, we determine for every facial mark spatial pattern in G^{TR}, S_i^{TR}, and S_i^{TE} its facial mark grid. Since we assume independence in (7.12), it suffices to demonstrate how parameters p_{xy} and q_{xy} belonging to a single grid cell are estimated. These parameters describe the probability of the (trace x, reference y) combination under the same source and the different source hypotheses. The number of such combinations is either $k = 2 \times 2 = 4$ (binary classifier) or $k = 4 \times 4 = 16$ (category classifier). We use the general notation p_i to refer to the probability of the i^{th} combination.

With respect to the general classifiers, we determine the frequencies $n_1,\ldots,n_k$ of every possible combination under the same source and the different source hypotheses from G^{TR}. We follow a Bayesian approach for the estimation of parameters. We choose a uniform, non-informative, Dirichlet distribution as the conjugate prior to the categorical distribution on k (trace x, reference y) combinations:

$$(p_1,\cdots,p_k) \sim \text{Dirichlet}(1,\cdots,1). \tag{7.22}$$

If the total number of observations is n, then the posterior distribution is [59]:

$$(p_1,\cdots,p_k) \sim \text{Dirichlet}(n_1+1,\cdots,n_k+1). \tag{7.23}$$

The marginal distribution of p_i is [59]:

$$p_i \sim \text{Beta}(n_i+1, n+k-n_i-1). \tag{7.24}$$

We estimate parameter values by the expected value of their posterior value. We do not use the MAP estimate of the posterior value as it is equal to the ML estimate, thereby nullifying the prior assumption (7.22). The expected value of the posterior p_i is equal to:

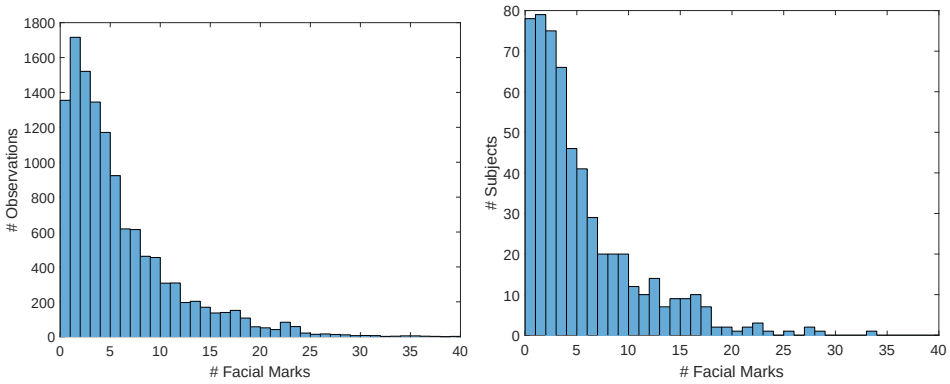
$$\mathbb{E}(p_i) = \frac{n_i+1}{n+k}. \tag{7.25}$$

The procedure to estimate the parameters of subject based classifiers is very similar to what is described above. We construct the same subject pairs from S_i^{TR} and different subject pairs by combining samples from S_i^{TR} with samples from G^{TR}.

Classifier Evaluation

For every subject S_i, $i = 1,\cdots,192$, we construct same subject pairs from S_i^{TE} and different subject pairs from S_i^{TE} and S_j^{TE}, $j \neq i$. Every classifier is then applied to these pairs, resulting in six distinct subject based score sets $\mathscr{S}_i^1,\ldots,\mathscr{S}_i^6$. The subject based evaluation uses these score sets.

Although a general evaluation of classifier $c = 1,\ldots,6$ can be determined by combining all subject based scores $\mathscr{S}_i^c$ of classifier c, this set is very large and is biased towards subjects having more facial mark spatial patterns. Prior to any general evaluation, we randomly select once for every subject the positions of 100 same source and 100 different source scores within $\mathscr{S}_i^c$. We use these positions to create a general evaluation set consisting of 19200 same source and 19200 different source scores.



a) #Facial Marks/Image b) Avg #Facial Marks/Subject

Figure 7.6: a) Histogram of the number of images with a given number of facial marks, b) Histogram of the number of subjects with a given average number of facial marks.

Table 7.2: Overview demographics FRGCv2

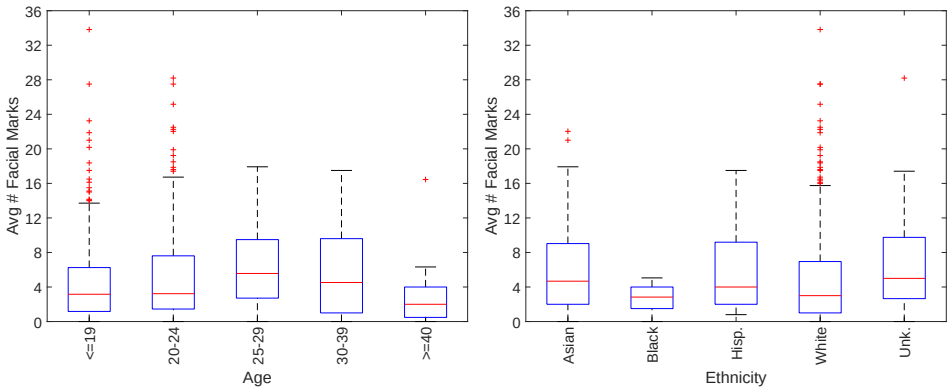
Partition	Counts
Age	$\leq$19=208, 20-24=230, 25-29=77, 30-39=33, $\geq$40=20
Ethnicity	Asian=134, Black=10, Hispanic=15, White=386, Unknown=23
Gender	Male=327, Female=241

7.2.6 Results and discussion

Experiment 1-Facial Mark Properties

The annotation software has been used over a period of two weeks, for approximately six hours every day. Figure 7.5 shows three subjects with their annotated facial mark spatial pattern. Figure 7.6 contains a histogram of the number of facial marks for every image and a histogram of the average number of facial marks for every subject. The average number of facial marks of subjects is 5. Table 7.2 summarises some demographic properties of the FRGCv2 dataset. It is clear that in terms of (a) age, (b) ethnicity, and (c) gender, this dataset is somewhat skewed towards (a) younger people, (b) white and Asian people, and (c) males.

There is no significant difference in the average number of facial marks between males and females ($p = 40\%$). Figure 7.7 shows the number of facial marks partitioned into age and ethnic groups. It can be shown that the age group ≥ 40 has a significantly lower average number of facial marks than all other age groups ($p < 5\%$). Also, people of Asian descent have a significantly higher number of facial marks ($p < 1\%$) than white people. Remarkably, this result is the exact opposite of what was found in [155] with respect to RPPVSM on the back torso.



a) #Facial marks by age b) #Facial marks by ethnicity

Figure 7.7: Average number of facial marks based on a) age, b) ethnicity.

Figure 7.8a shows the general facial mark spatial pattern. We observe that, when measured on a frontal image, (a) the general facial mark spatial pattern appears to be symmetric, (b) the cheeks have a high density of facial marks, and (c) the periocular region, forehead, nose, mouth, and neck have a low density of facial marks. We find that that H_{GU} is true ($p < 0.1\%$).

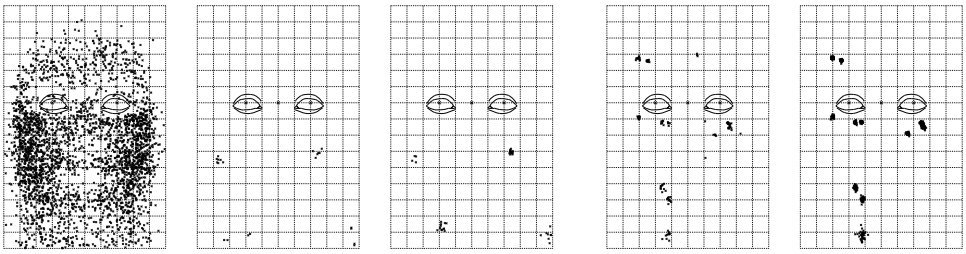
Figures 7.8b to 7.8e show the facial mark spatial patterns contained in the training set S_i^{TR} and the test set S_i^{TR} of two subjects. Facial mark spatial patterns of subjects are sparse and tend to be clustered (Figures 7.8d and 7.8e). Since the number of subjects with more than 10 facial marks is generally too low for a reliable χ^2 test (7.18), we test H_{SU} and H_{SG} for subjects who have between 2 and 10 facial marks, which accounts for approximately 60% of the subjects. Hypotheses H_{SU} and H_{SG} are found to be true ($p < 0.1\%$)[3]. This result does not correspond with the validity of Complete Spatial Randomness hypothesis reported in [155] regarding RPPVSM on the back torso. We assume that this can be attributed to the fact that (a) the face contains some areas (lips, eyes, eyebrows) where it is highly unlikely that it contains facial marks and (b) the face is very curved towards its sides, causing an increased number of observed facial marks in the region between the cheekbone and the ear in a frontal image.

Experiment 2-Discriminating power

Figures 7.9 and 7.10 present discriminating power in terms of EER as a function of the grid cell size Δ, both for a general (7.9a and 7.10a) and a subject based (7.9b to 7.9d and 7.10b to 7.10b) evaluation. In Figures 7.11a and 7.11b, the sampled ROC curves are shown for $\Delta = 0.50$ IPD.

From a general evaluation perspective, we observe in Figure 7.9a and Figure 7.10a that classifiers that use binary or category features behave in a similar manner. The EERs of the untrained Hamming and χ^2 classifiers increase as the grid cell size decreases; the χ^2 classifier performs better than the Hamming classifier. Towards smaller grid cell sizes their

[3] In one case out of 18 cases, we have $p < 1.6\%$.



a) General b) Train 04300 c) Test 04300 d) Train 04350 e) Test 04350

Figure 7.8: From left to right: a) General Facial Mark Spatial Pattern, Subject based Facial Mark Spatial Patterns for b) Training of 04300, c) Testing of 04300, d) Training of 04350, and e) Testing of 04350.

EERs converge: the feature representations tend to resemble each other for smaller grid cell sizes.

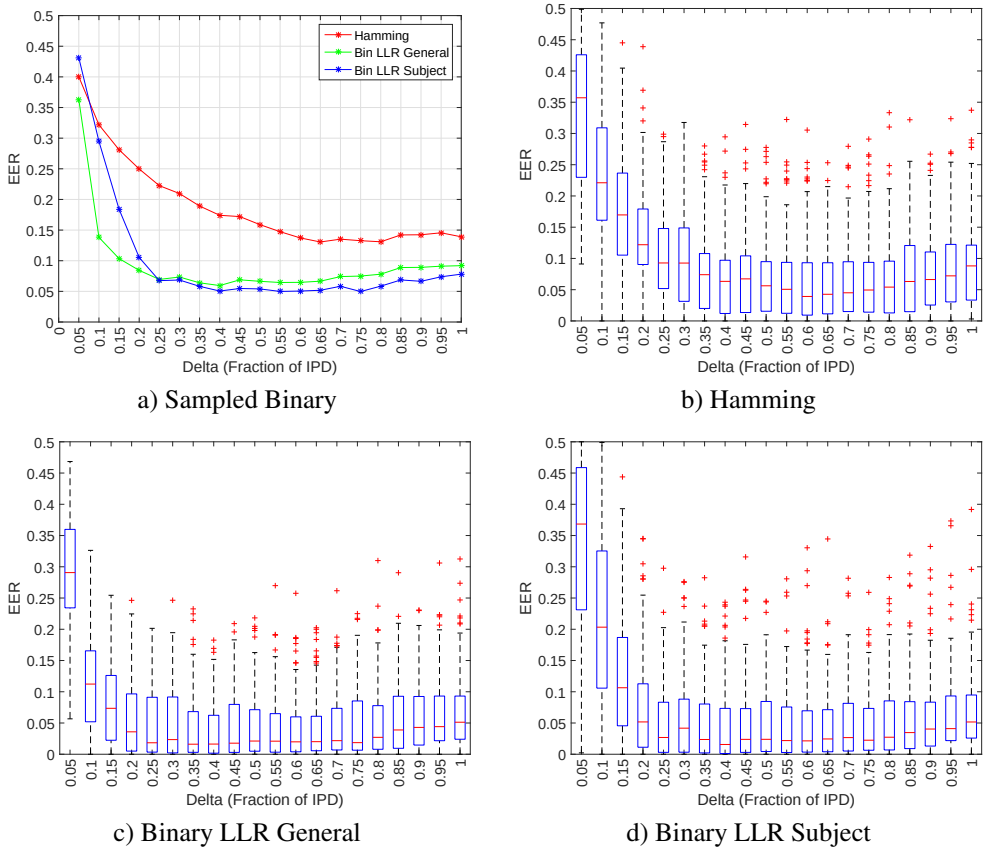
The untrained classifiers are outperformed by trained classifiers. Initially, the EERs of the trained classifiers decrease as the grid cell size decreases and they start to increase again when the grid cell size falls below $\Delta = 0.40$ IPD. For larger grid cell sizes, the subject based classifiers have a lower EER than the general classifiers. However, for smaller grid cell sizes ($\Delta \leq 0.25$ IPD), this behaviour is reversed. This can be explained by two factors. First, we expect that a facial mark grid of a subject is a better description of that subject than a general description. Second, for smaller grid cell sizes, we think that the training and evaluation of a subject based classifier is more sensitive to measurement variation due to the sparsity of facial marks within their facial mark grid.

The trained binary classifiers mostly have similar EERs; the same is true for general classifiers. The Category LLR subject based classifier has the best performance.

A subject based evaluation reiterates most of these findings. For example, it can be shown that at subject level, for almost every grid cell size, the Hamming classifier is significantly worse ($p < 0.1\%$) than any other classifier. The Category LLR subject based classifier is significantly better ($p < 0.1\%$) than any other classifier.

In the box plots of Figures 7.9b to 7.9d and Figures 7.10b to 7.10d, we observe a considerable variation between subjects in terms of EER. It can be shown that for a fixed grid cell size, some subjects are a negative outlier for every classifier. A number of them (for example subject 04300) even occur at different grid cell sizes. Their poor classifier performance can explained by the common location of their facial marks ("not very distinctive") and, as mentioned before, the measurement variation that influences both the training of a subject based classifier and its evaluation (see for example Figures 7.8b and 7.8c). Also, although not directly shown as a positive outlier, subject 04350 has a low EER for grid cell sizes up to $\Delta = 0.50$ IPD. This can be attributed to the rare facial mark locations within the facial mark grid and the stability of the observations shown in Figures 7.8d and 7.8e.

For almost every classifier and grid cell size, there are subjects that have EER=0, even if the sampled ROC curve indicates that the considered classifier has moderate discriminating power. An example is the Hamming classifier for $\Delta = 0.50$ IPD shown in Figure 7.11a. According to Figure 7.11c, more than 6% of the subjects can perfectly be discriminated by



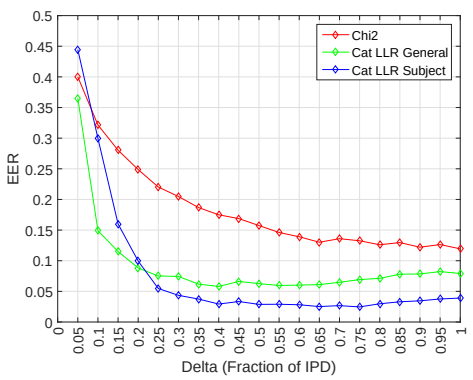
a) Sampled Binary b) Hamming

c) Binary LLR General d) Binary LLR Subject

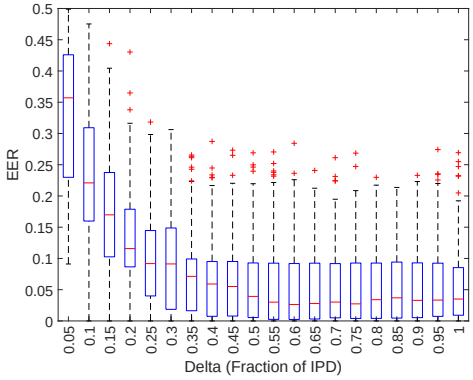
Figure 7.9: Binary: General and subject based evaluation of EER as a function of the grid cell size Δ: a) Sampled Binary, b) Hamming, c) Binary LLR General, d) Binary LLR Subject.

this classifier on this dataset. In fact, Figure 7.11c shows that the Category LLR subject based classifier can perfectly discriminate more than 20% of the subjects over a range of grid cell sizes on this dataset. These results depend on several factors, including the used dataset and the number of subjects. We found no significant correlation between the number of test samples and the attained EER.

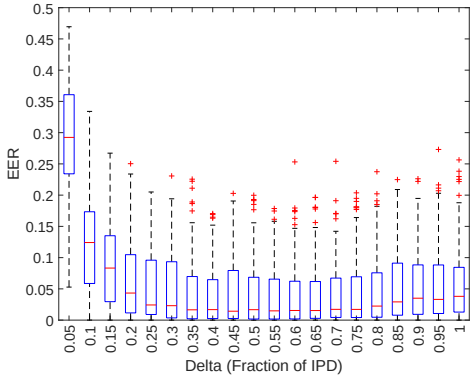
The number of facial marks is the final factor that we consider. Figure 7.12a shows the correlation ρ between the number of facial marks and the EER of the six classifiers as a function of Δ. Apart from the lower values of Δ, we find that all classifiers converge to a negative correlation, hence a larger number of facial marks can somewhat positively influence the EER. A scatter plot (Figure 7.12b) illustrates this dependency. For low values of Δ, this dependency becomes stronger. This can be attributed to the fact that having more facial marks makes the face more distinctive in relation to the grid cell size. Distinctiveness is rewarded by the structure of a trained classifier and may outweigh errors caused by measurement variation. However, for the two untrained classifiers and for lower values of Δ, the dependency is actually strongly reversed: the number of facial marks negatively influences the EER. A



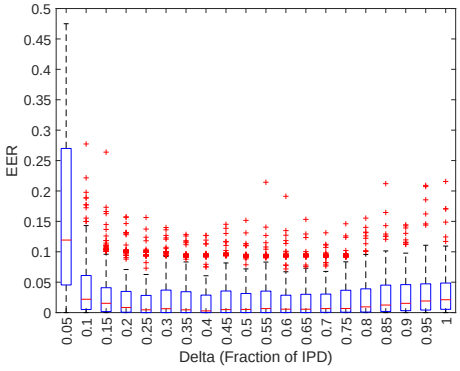
a) Sampled Category



b) χ^2



c) Category LLR General



d) Category LLR Subject

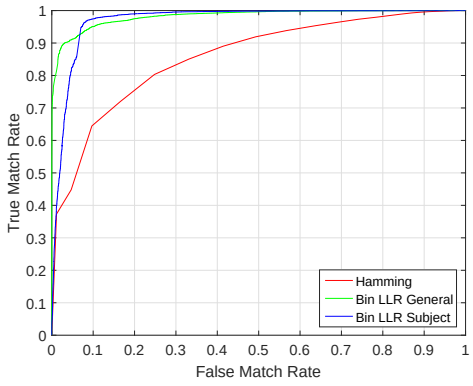
Figure 7.10: Category: General and subject based evaluation of EER as a function of the grid cell size Δ: a) Sampled Category, b) χ^2, c) Category LLR General, and d) Category LLR Subject.

scatter plot (Figure 7.12c) illustrates this. The remarkable conclusion is that having *no* facial marks actually yields the best performance in these combinations, caused by the fact that the absence of observations implies the absence of within variation. In all cases we observe that having no facial marks yields an EER≈8-10%.

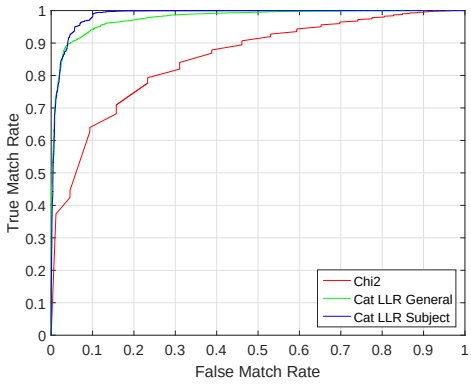
Experiment 3-Calibration loss

Figures 7.13 and 7.14 show the calibration loss of the trained classifiers, both from a general and subject based evaluation perspective, respectively. We observe two effects that depend on the grid cell size. First, the general classifiers exhibit a relatively constant calibration loss, until the lowest value for Δ. Second, the subject based classifiers show a major calibration loss when $\Delta \leq 0.50$ IPD.

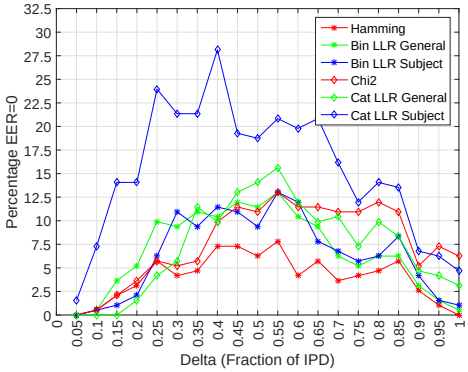
Several related factors might cause these calibration losses.



a) Sampled Binary ROC's, $\Delta = 0.50$ IPD



b) Sampled Category ROC's, $\Delta = 0.50$ IPD



c) Percentage EER=0

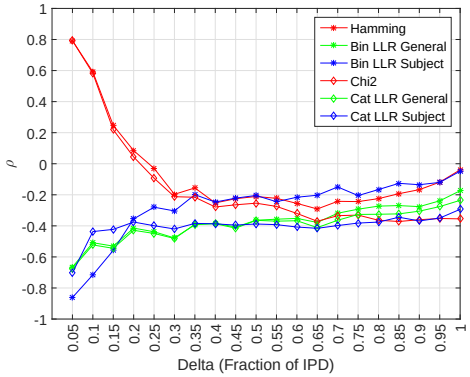
Figure 7.11: From left to right: a) Sampled Binary ROC curves for $\Delta = 0.50$ IPD, b) Sampled Category ROC curves for $\Delta = 0.50$ IPD, and c) percentage of subjects with EER=0 as a function of Δ.

The first factor is the difference in the number of training samples used to estimate parameters of the general and subject based classifiers, leading to a less robust estimation in the latter case.

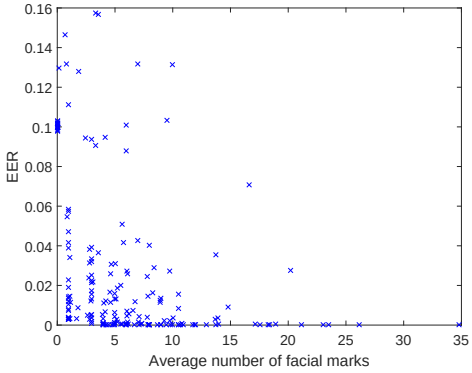
The second factor is the measurement variation in facial mark spatial patterns. This especially negatively impacts the parameter estimation of a subject based classifier. Together with the sparse subject based facial mark grids, this often results in a bad model that may produce extreme likelihood ratio values. Some subjects whose facial mark spatial patterns are similar to those shown in Figures 7.8b and 7.8c have a high calibration loss, even for large grid cell sizes.

The third factor is the grid cell size. For smaller values, the negative effect of measurement variation on the produced wrong likelihood ratio values is reinforced.

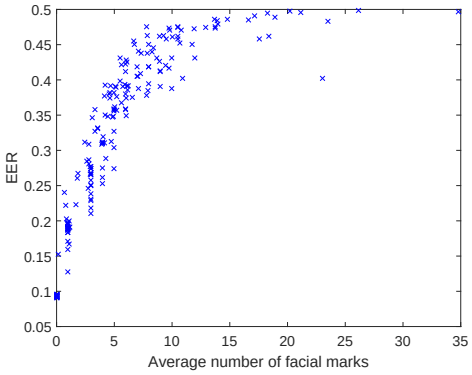
The fourth factor is the structure of the classifiers. Due to the independence assumption (7.12) and the fact that the number of considered grid cells depends inverse quadratically on the grid cell size, the sum of multiple extreme likelihood ratio values results in a meaningless



a) Correlation EER/#Facial Marks



b) Cat. LLR Subject, $\Delta = 0.40$ IPD



c) Hamming, $\Delta = 0.10$ IPD

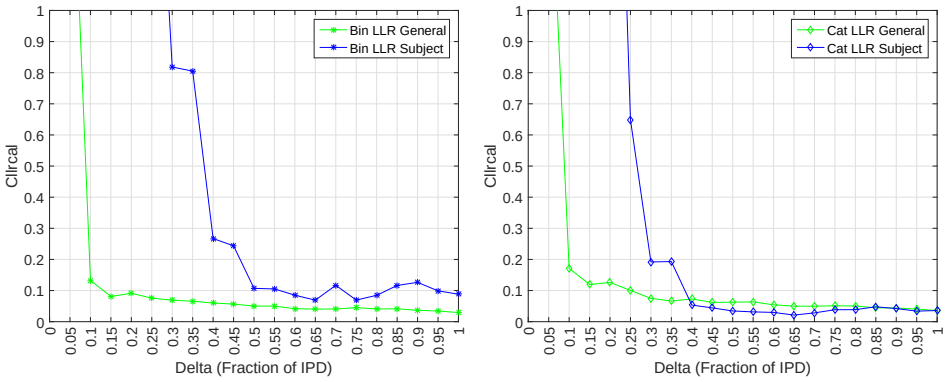
Figure 7.12: From left to right: a) the correlation between EER and number of facial marks as a function of Δ, b) scatter plot for Category LLR Subject based and $\Delta = 0.40$ IPD, and c) scatter plot for Hamming and $\Delta = 0.10$ IPD.

value.

The fifth factor is the number of facial marks. In Figure 7.15, the correlation between the number of facial marks and calibration loss is shown as a function of the grid cell size.

For general classifiers, calibration loss is negatively correlated to large grid cell sizes; for very small grid sizes the correlation is reversed. A possible explanation is that such a classifier produces (too) moderate likelihood values for large grid cell sizes, and as discussed before, for small grid cell sizes tends to produce larger incorrect values. If the grid cell size is large, then the number of facial marks should be higher to obtain a lower calibration loss (negative correlation). If the grid cell size is small, the likelihood ratio of observing facial marks in particular grid cells increases and might be too extreme; having less facial marks leads to a lower calibration loss (positive correlation).

For subject based classifiers, the correlation is reversed. One could argue that in the case of large grid cell sizes, the model is correct and having less facial marks could lead to less error in the calculated likelihood ratio for some subjects. When the grid cell size is small,



a) Sampled Binary b) Sampled Category

Figure 7.13: Cllr^{cal} as a function of the grid cell size Δ: a) Sampled Binary, b) Sampled Category.

although more facial marks might reduce the effect of a wrong model (see the discussion on Figure 7.12a), the overall calibration loss indicates that the produced likelihood ratio cannot be used as strength of evidence.

There is no significant correlation between the number of test observations and Cllr^{cal}.

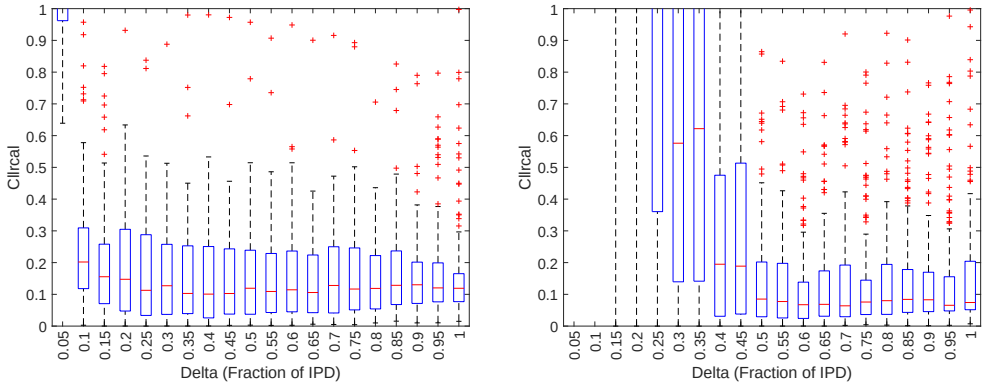
7.2.7 Conclusion and Future Work

In this paper, we have considered properties of facial marks, their spatial patterns, and classifiers acting upon facial mark grids.

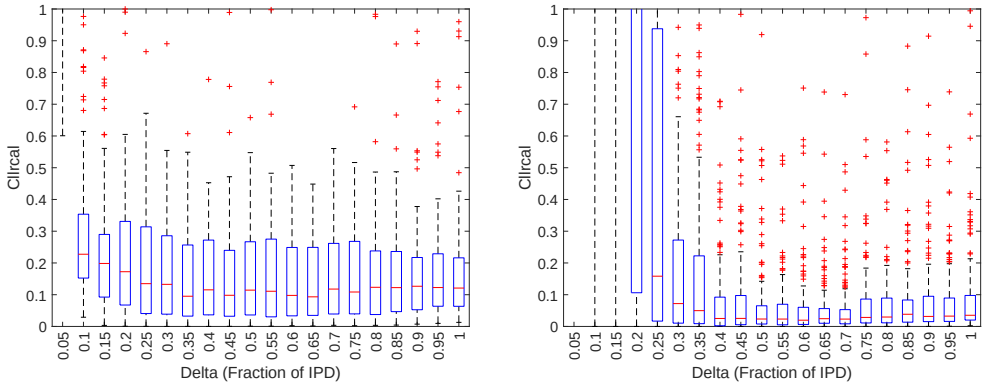
The first research question considered demographic and spatial properties of facial marks. We found differences in the number of facial marks between age groups. People of Asian descent had a significantly higher number of facial marks ($p < 1\%$) than white people. The general facial mark spatial pattern was shown not to be uniformly distributed. Up 60% of the subjects, the facial marks lay significantly closer ($p < 0.1\%$) than sampling from the general facial mark spatial pattern or a uniform spatial pattern would predict.

The second research question considered the influence of Aspects 2 to 4 and the number of facial marks on discriminating power. In general, the grid cell size influenced the discriminating power. In most cases, category features yielded better results than binary features. Trained classifiers outperformed untrained ones and the subject based category classifier was shown to be the best classifier until the grid cell size became too small. A subject based evaluation showed that some subjects exhibited EER=0, even if the classifier had poor general discriminating power. We also observed correlations between the number of facial marks and the EER as a function of the grid cell size.

The third research question considered the influence of Aspects 2, 3, and the number of facial marks on calibration loss of trained classifiers. Subject based facial grid classifiers were susceptible to large calibration loss; in that case, grid sizes below $\Delta = 0.50$ IPD should be avoided. Identified factors that influenced the calibration loss were parameter estimation, measurement variation, grid cell size, and classifier structure. Another factor was the number



a) Binary LLR General b) Binary LLR Subject



c) Category LLR General d) Category LLR Subject

Figure 7.14: Cllr^{cal} as a function of the grid cell size Δ: a) Binary LLR General, b) Binary LLR Subject, c) Category LLR General, and d) Category LLR Subject.

of facial marks for which it was found to be correlated to Cllr^{cal}.

We conclude that the subject based category classifiers are superior to all other considered classifiers. However, we note that the presented discriminating power and calibration loss as function of the grid cell size depends on the size, resolution, and other characteristics of the used images.

Future work includes a study whether the automated system for body marks described in Nurhudatiana et al. [156] can be applied to facial marks. When robust methods for the automatic detection of facial marks are available, the true potential of facial marks as a means to speed up search and to compactly represent a facial image can be investigated, possibly in the context of modern approaches to face recognition such as deep learning.

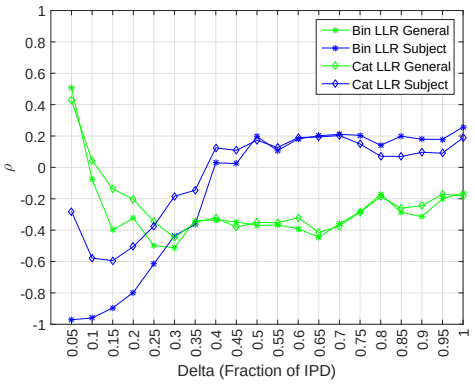


Figure 7.15: Correlation Cllr^{cal}/#Facial Marks

7.3 Label specific versus general classifier performance: an extreme example

7.3.1 Abstract

In this paper, we present an example that compares the label specific performance of a classifier to its general performance. Given a number of labels, the label specific performance is computed using test samples confined to a single label, whereas during the general performance evaluation, test samples from all labels are used. In both evaluation schemes, reference templates are taken from all labels. We choose Area Under the Curve (AUC) as the performance measure. The example is extreme in the sense that it shows that the label specific performance can be perfect (AUC $= 1$ for each label), while we can approach random general performance (that is AUC $= \frac{1}{2}$) arbitrarily close. Although this demonstration is purely theoretical, the effect it describes does play a role in for example the domain of forensic biometrics. In this domain, labels correspond to subjects and poor or moderate general performance does not automatically imply unsuitability in forensic case work, as long as a feature helps to discriminate the subject at hand. Also, more generally, label specific performance could lead to more insight into feature properties that influence such performance.

7.3.2 Introduction

The procedure to assess the performance of a classifier can be summarised as follows. For each pair in a collection of test-reference pairs with known ground truth (either genuine or imposter), the classifier calculates the comparison score. A score below and above a fixed decision threshold $\tau \in (-\infty, \infty)$ is considered as a non-match and match decision, respectively. Given τ, the classifier can make two types of mistakes: false non-match and false match. Given this collection of comparison scores and their ground truth, the amount of false non-matches and false matches is scaled to the number of genuine and imposter pairs respectively, from which the commonly used true match rate (TMR) and false match rate (FMR) are derived. The receiver operator characteristic (ROC) curve is the parametric curve $(\text{FMR}(\tau), \text{TMR}(\tau))$ for $\tau \in (-\infty, \infty)$. Every ROC curve has at least two points in common:

$(1,1)$ ($\tau = -\infty$) and $(0,0)$ ($\tau = \infty$). One dimensional performance metrics that summarise the performance of the classifier can be derived from the ROC curve. In this paper, we will use Area Under the Curve (AUC). A perfect classifier has AUC $= 1$, whereas a random classifier has AUC $= \frac{1}{2}$.

Sometimes a classifier is linked to a specific label, and as a consequence, the previously described evaluation procedure involves test-reference pairs in which the test sample is restricted to that label. We refer to this as *label specific evaluation*, whereas the evaluation procedure that involves test samples of different labels is referred to as *general evaluation*. In both schemes, reference templates are taken from all labels. There is no principled reason against using this label specific evaluation on generally applicable classifiers, complementing the result of a general evaluation.

The last point is of specific relevance to forensic biometrics. Although a certain feature and corresponding classifier might have a moderate or even poor general AUC, there might be labels, or subjects in this context, for which actually this feature yields a high AUC. The latter fact can then be used in a court of law. Previous work that studied label specific performance in the context of forensic biometrics includes a study on facial marks [24] and a practical framework for forensic evidence evaluation [26]. These papers empirically show that there can exist a significant variation in performance between different labels. For example, an included example in [26] shows that facial marks are not suitable for forensic evaluation in general, while still some subjects can be perfectly discriminated. This label specific perspective has not been addressed in for example a recent guideline for the validation of likelihood ratio methods for forensic evidence evaluation [48]. A label specific evaluation can lead to more insight into feature properties that influence that performance.

However, one can argue that the AUC found in a general evaluation leaves room for variation at label specific level and this variation is expected to be (somewhat) centered around the general evaluation. Or in other words, this label specific variation is inherent to and determined by the general evaluation, and thereby nullifying any added value of looking into this matter.

The aim of this paper is to present a counterintuitive, *theoretically constructed* example that complements previous *empirical* work mentioned here above. The example actually shows extreme behaviour: a label specific evaluation that shows AUC $= 1$ for each label, while at a general level we can approach AUC arbitrarily close to $\frac{1}{2}$. In all results we use the Euclidean score distance function $s(x,y) = -|x-y|$. Despite its theoretical construction, this example shows there is added value in considering label specific performance, and not only in the context of forensic biometrics. In the remainder of this paper, we present and prove the main theorem of this paper.

7.3.3 Main result

The main result of this paper is the following theorem.

Theorem 7.1. *Given any $\delta > 0$ and Euclidean score distance function $s(x,y) = -|x-y|$, there exists a two dimensional configuration of test samples and reference templates, such that for each label i we have* $\mathrm{AUC}_i = 1$ *while* $\mathrm{AUC}_{gen} \leq \frac{1}{2} + \delta$. *Here* AUC_i *is the* AUC *attained by label i and* AUC_{gen} *is the general* AUC.

We postpone the proof of this theorem to the end of this section. The proof relies on several lemmas. The first lemma is presented below and is also used in a subsequent example.

Lemma 7.1. *Suppose each of $n \geq 2$ labels, referred to as $i = 0, \ldots, n-1$, has a collection G_i of G genuine scores and a collection I_i of I imposter scores, such that*

$$I_{n-1} < G_{n-1} < \cdots < I_0 < G_0, \tag{7.26}$$

then $\mathrm{AUC}_i = 1.0$ *and* $\mathrm{AUC}_{gen} = \frac{1}{2} + \frac{1}{2n}$.

Proof. The label specific evaluation of label i only uses I_i and G_i. According to the ordering in (7.26), these sets can be perfectly separated, hence $\mathrm{AUC}_i = 1$. The threshold τ travels from $-\infty$ to $+\infty$, and when it moves through the first collection I_{n-1} of imposter scores, the ROC curve moves from $(1,1)$ to $(1-\frac{1}{n},1)$; when it moves through the first collection G_{n-1} of genuine scores, the ROC curve moves from $(1-\frac{1}{n},1)$ to $(1-\frac{1}{n},1-\frac{1}{n})$. This behaviour is repeated until all score collections have been passed, resulting in a staircase like ROC curve with n equal stairs. The area between the ROC curve and the main diagonal is $\frac{1}{2}\frac{n}{n^2} = \frac{1}{2n}$, hence $\mathrm{AUC}_{gen} = \frac{1}{2} + \frac{1}{2n}$. □

We now present an trivial example configuration for which (7.26) holds.

Example. Assume we have 2 labels, both having a single test sample t and single reference template r in $\mathbb{R}$. Assume that $r_0 = 1,\ r_1 = -1, t_0 = 2, t_1 = -6$. Then $s(t_0, r_0) = -1,\ s(t_0, r_1) = -3,\ s(t_1, r_1) = -5$ and $s(t_1, r_0) = -7$.

Although this example meets (7.26), the FMR and TMR can only attain 0 or 1, making the AUC either 0 or 1. Furthermore, their is no variation in the genuine scores. However, this example serves as a template for a two dimensional configuration for n labels. Before we present the other lemma that uses this configuration, we introduce additional notation and two auxiliary results.

Definition 7.1. *Given nonempty compact sets A and B in $\mathbb{R}^2$, we define*

$$m(A,B) = \min_{a\in A, b\in B} |a-b|, \tag{7.27}$$

$$M(A,B) = \max_{a\in A, b\in B} |a-b|. \tag{7.28}$$

Lemma 7.2. *Given two closed disks D_1 and D_2 of radius R, and whose centers are a distance a away from each other ($R < \frac{a}{2}$), then $m(D_1, D_2) = a - 2R$ and $M(D_1, D_2) = a + 2R$.*

Proof. The extreme values are attained on the boundaries ∂D_i of the disks. Suppose otherwise, then since the interior $\mathrm{int}(D_i)$ of the disk D_i is open, around any point $p \in \mathrm{int}(D_i)$ that is claimed to correspond to an extreme value, there exists an open disk $D_p \subset \mathrm{int}(D_i)$. This means that by simply extending or shortening the line segment of which p is an end point while remaining in the disk D_p, we obtain a better extreme value. Hence, the extreme value must be attained on ∂D_i.

Let M_i be the center of disk D_i, and let $P_i \in \partial D_i$. By the triangle inequality we have

$$\begin{aligned} |P_1 - P_2| &\leq |P_1 - M_1| + |M_1 - M_2| + |M_2 - P_2| \\ &= a + 2R. \end{aligned} \tag{7.29}$$

We also have

$$\begin{aligned} |M_1 - M_2| &\leq |M_1 - P_1| + |P_1 - P_2| + |P_2 - M_2| \\ &= |P_1 - P_2| + 2R, \end{aligned} \tag{7.30}$$

hence $|P_1 - P_2| \geq |M_1 - M_2| - 2R = a - 2R$. The inequality in (7.29) reverts to an equality when the concatenation of line segments $[P_1, M_1]$, $[M_1, M_2]$ and $[M_2, P_2]$ is equal to the line segment $[P_1, P_2]$, the inequality in (7.30) reverts to an equality when the concatenation of line segments $[M_1, P_1]$, $[P_1, P_2]$ and $[P_2, M_2]$ is equal to the line segment $[M_1, M_2]$. From this it follows that $m(D_1, D_2) = a - 2R$ and $M(D_1, D_2) = a + 2R$. □

Lemma 7.3. *Given $n \geq 2$, then*

$$\frac{1 - \cos(\frac{2\pi}{n})}{32n} < \sqrt{\frac{1}{32}(1 - \cos\left(\frac{2\pi}{n}\right))}, \tag{7.31}$$

$$\frac{1 - \cos(\frac{2\pi}{n})}{32n} < \frac{1}{4}. \tag{7.32}$$

Proof. $0 < \frac{1-\cos(\frac{2\pi}{n})}{32n} < \frac{1-\cos(\frac{2\pi}{n})}{32} \leq \frac{1}{16} < 1$, so (7.31) holds. Furthermore, $1 - \cos(\frac{2\pi}{n}) \leq 2 < 8n$, implying (7.32). □

Lemma 7.4. *Given $n \geq 2$, $i = 0, \ldots, n-1$, define*

- *Radius $R = \frac{1-\cos(\frac{2\pi}{n})}{32n}$,*
- *Disk R_i with radius R and center $r_i = (\cos(\frac{2\pi i}{n}), \sin(\frac{2\pi i}{n}))$,*
- *Disk T_i with radius R and center $t_i = ((4i+2)\cos(\frac{2\pi i}{n}), (4i+2)\sin(\frac{2\pi i}{n}))$,*

then for every $i = 0, \ldots, n-1$ and $j = 1, \ldots, n-1$ we have

$$M(T_i, R_i) < m(T_i, R_{i+j}) \tag{7.33}$$

and for every $i = 0, \ldots, n-2$ and $j = 1, \ldots, n-1$ we have

$$M(T_i, R_{i+j}) < m(T_{i+1}, R_{i+1}). \tag{7.34}$$

We assume that the indices are taken modulo n.

Proof. **Proof of (7.33)** Due to rotational symmetry, we may assume that t_i is placed at $(4i + 1, 0)$, r_i at $(0, 0)$ and $r_{i+j} = (\cos(\frac{2\pi j}{n}) - 1, \sin(\frac{2\pi j}{n})) = (-x_j, y_j)$. Note that $x_j > 0$, since $j \neq 0$.

Now select $u_i \in T_i$, $v_i \in R_i$, and $v_{i+j} \in R_{i+j}$. It suffices to prove that the angle between the vectors $u_i - v_i$ and $v_{i+j} - v_i$ is larger than $\frac{\pi}{2}$, since the law of cosines implies that the length of $u_i - v_i$ is smaller than $u_i - v_{i+j}$, leading to $M(T_i, R_i) < m(T_i, R_{i+j})$ for every $i = 0, \ldots, n-1$ and $j = 1, \ldots, n-1$. To prove this angle property, we show that the inner product $(u_i - v_i, v_{i+j} - v_i)$ is negative.

In principle, any of the points $u_i \in T_i$, $v_i \in R_i$, and $v_{i+j} \in R_{i+j}$ should be considered to produce an angle between the vectors $u_i - v_i$ and $v_{i+j} - v_i$, but it suffices to only consider

boundary points of T_i and R_{i+j} as an angle produced by internal points could also been produced by boundary points. Therefore, the following parameterisations of the three points are capable of producing the full range of angles:

$$u_i = \begin{pmatrix} R\cos(\theta_u) + (4i+1) \\ R\sin(\theta_u) \end{pmatrix}, \theta_u \in [0, 2\pi), \tag{7.35}$$

$$v_i = \begin{pmatrix} \alpha R\cos(\theta_i) \\ \alpha R\sin(\theta_i) \end{pmatrix}, \theta_i \in [0, 2\pi), \alpha \in [0, 1], \tag{7.36}$$

$$v_{i+j} = \begin{pmatrix} -x_j + R\cos(\theta_{i+j}) \\ y_j + R\sin(\theta_{i+j}) \end{pmatrix}, \theta_{i+j} \in [0, 2\pi). \tag{7.37}$$

A straightforward calculation shows that

$$(u_i - v_i, v_{i+j} - v_i) = aR^2 + bR + c, \tag{7.38}$$

with

$$\begin{aligned} a &= A_1A_3 + A_2A_4 \\ b &= -x_jA_1 + (4i+1)A_3 + y_jA_2 \\ c &= -x_j(4i+1) \end{aligned} \tag{7.39}$$

and

$$\begin{aligned} A_1 &= \cos(\theta_u) - \alpha\cos(\theta_i) \\ A_2 &= \sin(\theta_u) - \alpha\sin(\theta_i) \\ A_3 &= \cos(\theta_{i+j}) - \alpha\cos(\theta_i) \\ A_4 &= \sin(\theta_{i+j}) - \alpha\sin(\theta_i). \end{aligned} \tag{7.40}$$

By the triangle inequality we have $|A_i| \leq 2$, so

$$|a|R^2 \leq 8R^2, \tag{7.41}$$

and since $|x_j| \leq 2$ and $|y_j| \leq 1$, we have

$$|b|R \leq 2(|x_j| + |y_j| + (4i+1))R \leq 8nR. \tag{7.42}$$

According to Lemma 7.3 we have

$$R \leq \min\{\sqrt{\frac{1}{32}(1-\cos\left(\frac{2\pi}{n}\right))}, \frac{1-\cos(\frac{2\pi}{n})}{32n}\}, \tag{7.43}$$

so

$$\begin{aligned} aR^2 + bR &\leq |aR^2 + bR| \\ &\leq |a|R^2 + |b|R \\ &\leq 8R^2 + 8nR \\ &\leq \frac{1-\cos(\frac{2\pi}{n})}{4} + \frac{1-\cos(\frac{2\pi}{n})}{4} \\ &\leq \frac{1-\cos(\frac{2\pi}{n})}{2} \\ &< x_j(4i+1) = -c \end{aligned} \tag{7.44}$$

This implies that $(u_i - v_i, v_{i+j} - v_i) = aR^2 + bR + c < 0$.

Proof of (7.34) We have, according to the triangle inequality $|t_i - r_{i+j}| \leq |t_i - r_i| + |r_i - r_{i+j}|$, so $|t_i - r_{i+j}| \leq 4i + 1 + 2 = 4i + 3$ as the maximum distance between two points on the unit circle is 2. This implies that $M(T_i, R_{i+j}) \leq 4i + 3 + 2R$ for $i = 0, \ldots, n-1$ and $j = 1, \ldots, n-1$, according to Lemma 7.2. By construction, $|t_{i+1} - r_{i+1}| = 4(i+1) + 1 = 4i + 5$, so by Lemma 7.2 we have $m(T_{i+1}, R_{i+1}) = 4i + 5 - 2R$. According to Lemma 7.3, $R < \frac{1}{4}$, showing that $M(T_i, R_{i+j}) < m(T_{i+1}, R_{i+1})$. □

We are now able to prove Theorem 7.1.

Proof. Given any $\delta > 0$, choose $n \geq \lceil \frac{1}{2\delta} \rceil$. According to Lemma 3, there exist disks T_i and R_i such that for every $i = 0, \ldots, n-1$ and $j = 1, \ldots, n-1$ we have

$$M(T_i, R_i) < m(T_i, R_{i+j}) \tag{7.45}$$

and for every $i = 0, \ldots, n-2$ and $j = 1, \ldots, n-1$ we have

$$M(T_i, R_{i+j}) < m(T_{i+1}, R_{i+1}). \tag{7.46}$$

For label i, use disks T_i and R_i to draw N_t test samples and N_r reference templates, respectively. We assume the use of a Euclidean score distance function $s(x, y) = -|x - y|$. Suppose that label i has a collection G_i of $G = N_t N_r$ genuine scores and a collection I_i of $I = (n-1)N_t N_r$ imposter scores. It follows from (7.45) and the reverse ordering property of the score function that for every i we have

$$I_i < G_i. \tag{7.47}$$

Similarly, (7.46) implies that

$$G_{i+1} < I_i. \tag{7.48}$$

We conclude that

$$I_{n-1} < G_{n-1} < \cdots < I_0 < G_0, \tag{7.49}$$

According to Lemma 7.1, $\mathrm{AUC}_i = 1$ while $\mathrm{AUC}_{gen} \leq \frac{1}{2} + \frac{1}{2n} \leq \mathrm{AUC}_{gen} + \delta$. □

7.3.4 Conclusion

This paper has presented a theoretical example that shows the largest possible discrepancy between a perfect label specific performance and random general performance. It complements other, empirical, work that already showed the variation of label specific performance around general performance; in particular, in which some labels exhibit a good label specific performance while the general performance is poor.

7.4 Chapter conclusion

This chapter presented two studies that addressed research question 2a: *To which extent do we observe or can we construct differences in general and subject based performance?* and

research question 2b: *How well can facial marks be used for forensic evaluation, also taking subject based data and subject based evaluation into account?*

The first included study presented three conclusions. The first conclusion referred to demographic and spatial properties of facial marks. Some differences between age and ethnic groups were found. Facial mark spatial patterns seemed to be clustered. The second conclusion mostly dealt with the influence of feature representation and classifier type on discriminating power, both seen at a general and a subject based level. It was found, amongst other things, that the grid cell size influences the discriminating power. If it became too small, all considered classifiers performed worse. Also, including subject based facial mark location data yielded the best classifier. A subject based evaluation showed that even for a relatively poor classifier, there are still subjects that can be discriminated. The third conclusion was drawn with respect to the influence of feature representation and classifier type on calibration. It was found that classifiers that were trained on general data were relatively well calibrated, even for small values of the grid cell size, whereas the classifiers that used subject based data exhibited a large calibration loss for grid cell sizes roughly smaller than half the interpupillary distance.

The second study presented a theoretical construction that showed that classifiers can exhibit perfect subject based performance, while the general performance is essentially random.

When we consider the first study and its conclusions in light of research question 2a, we observe that there are several occasions in which the differences between a general and a subject based evaluation are quite apparent and significant. The Tukey plots depicting discriminating power and calibration give an impression of their spread but also their limits. Moreover, there is a percentage of subjects that can be discriminated while the general performance of the used classifier might have moderate performance. The results of the second study show that it is possible to attain the largest possible discrepancy between general and subject based performance. This is a theoretical construction and is of relevance in light of the results of the first study.

Regarding research question 2b, we think that the first study partially shows that facial marks can be used for forensic evidence evaluation, in particular when classifiers have been trained on subject based facial mark spatial patterns, although the calibration results restrict the operational range of grid cell sizes. Also, we reiterate that the subject based evaluation shows that other, less well performing, classifiers might also be used to discriminate certain subjects. However, it should be noted that these results were obtained using good quality images. Therefore, it is worthwhile to explore the effect of using lower quality trace images as a better representative of forensic use cases. This is one of the topics explored in the next chapter.

Upon inspection, the presented six aspects of the proto-framework can be generalised into a framework that addresses the more general problem of designing and evaluating biometric classifiers for forensic evidence evaluation. This, and related topics, will be addressed in the next chapter.

Chapter 8

Subject Based: framework and random versus non-random performance

8.1 Introduction

As indicated in the chapter conclusion of Chapter 7, the six aspects in the proto-framework can be generalised into a framework with the justification of the use of subject based data and evaluation in mind. This is one of two studies included in this chapter which addresses in total four research questions. They are research question 2a: *To which extent do we observe or can we construct differences in general and subject based performance?*, research question 2b: *How well can facial marks be used for forensic evaluation, also taking subject based data and subject based evaluation into account?*, research question 2c: *In which manner can a biometric approach to FISWG characteristic descriptors be generalised into a framework for forensic evidence evaluation that also incorporates a subject based approach?*, and research question 2d: *What is a theoretical boundary between random and non-random behaviour of classifiers in a subject based performance evaluation based on AUC?*

The first study presents the framework and includes two example applications. The first application is the use of nine simple characteristic descriptors, applicable in the case when a perpetrator wears a balaclava. The second application extends the facial mark study of Chapter 7 by considering another forensically relevant dataset. Both applications show the large variation in performance seen from a subject based evaluation. The second study included in this chapter presents an exact expression and an approximation to the probability of the Area Under the Curve (AUC) values produced by a random classifier. The AUC measured on a finite set of scores is a random variable itself, and it is possible that the empirically measured AUC signals a low to moderate performance, while the underlying biometric classifier is random. This is of relevance as the subject based evaluation introduced in Chapter 7 and the first part of this chapter typically use a low number of same source and different source scores in which this effect is observable.

Section 8.2 has been submitted as “Mind the Gap: A Practical Framework regarding

Classifiers for Forensic Evidence Evaluation" [26].

Section 8.3 has been accepted for publication as "How Random is a Classifier given its Area under Curve?" [27].

Reading Guide
Section 8.2. This section should at least be browsed, in particular, the reader should get acquainted with the six aspects of the framework. The two applications can be browsed.
Section 8.3. The contents of the two theorems included in this section should at least be read.

8.2 Mind the gap: a practical framework regarding classifiers for forensic evidence evaluation

8.2.1 Abstract

In this paper, we present a practical framework that addresses six, mostly forensic, aspects that can be considered during the design and evaluation of biometric classifiers for forensic evidence evaluation. Although we use the term classifier, we are interested in the comparison score it produces, rather than a decision. Forensic evidence evaluation is a central activity in forensic case work; it deals with the assessment of strength of evidence to be used in a court of law. We envision the use of biometric classifiers that are capable of producing such strength of evidence. The addressed aspects consider the modality and features, the biometric score and its forensic use, and choice and evaluation of several performance characteristics and metrics. Advantages of this framework are that it attempts to bridge a "gap" between biometric and forensic research and that it makes choices during the design process more transparent. We also present two applications of this framework relevant to the domain of forensic face recognition. In the first application, a robber wears a balaclava making only the periocular and mouth region visible. We explore different classifiers using a collection of nine one-dimensional forensic facial features. We find large and explainable variations in discriminating power between subjects. In the second application, we use footage taken from surveillance cameras and explore how well facial marks can be used for forensic evidence evaluation. It is an extension of earlier work that considered facial marks on better quality images.

8.2.2 Introduction

Given trace specimens from a crime scene (for example finger marks or face images extracted from surveillance camera footage) and reference specimens taken from a suspect (for example, finger prints or good quality frontal and profile facial images), one of the tasks of the forensic examiner is to determine the strength of evidence supporting the hypothesis that trace and reference specimens have a common donor versus the hypothesis that the trace originates from another donor. Strength of evidence is very different if it is somebody taken by chance or a look alike. We refer to this process as forensic evidence evaluation. Computer assisted methods can support the forensic examiner during this process, ranging from pre-processing tasks to the computation of the strength of evidence.

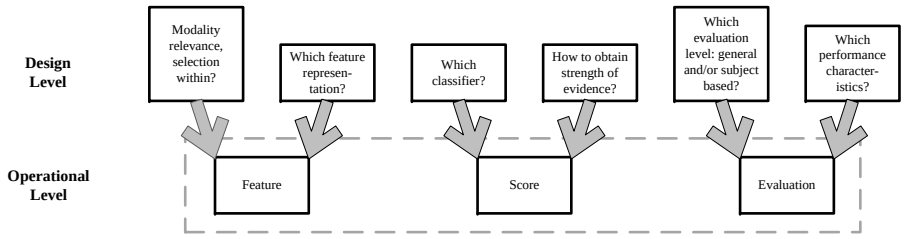


Figure 8.1: Design Framework that can be used during the design and evaluation of biometric classifiers for forensic evidence evaluation and relates to the forensic process published in [48]. Top row contains the aspects, bottom row shows essential components of an evaluated biometric system. Arrows are unidirectional "influences" relationships.

Not every modality has the same maturity level of automation. For example, the finger print modality has a straightforward three level representation (of which often only the first two are used) and mature extraction and comparison methods [162]. On the other hand, the face is a complex modality and especially forensic evidence evaluation using images of faces taken under realistic forensic conditions is largely a manual process [12].

Therefore, there is still a general need for research on biometric classifiers that are capable of producing strength of evidence, to be used in a forensic evidence evaluation process in which the human examiner remains to have a pivotal role. Biometric and forensic science have much in common, notably due to their strong interest in connecting individuals to traces (in the forensic nomenclature) or probes or tests (in the biometric nomenclature).

Due to this symbiotic relationship, a small but important "gap" between biometric and forensic science is easily overlooked outside the domain of forensic biometrics. We provide three differences. First, not every modality has the same potential in forensic science. Second, biometric scores produced by biometric classifiers in most cases cannot directly be used as strength of evidence in forensic case work. Third, there exist performance characteristics which are relevant from a forensic perspective [48], but which are not taken into consideration by standard biometric research.

Apart from these differences, even in both biometric and forensic science, the importance of subject based performance evaluation in relation to general performance evaluation seems to be somewhat underrated, since insight into this type of performance might be important from a forensic point of view. Here, a subject based performance evaluation uses traces from a single subject, whereas a general performance evaluation uses traces from multiple subjects. Having an extreme eye fissure angle opening might discriminate a specific subject well, while in general this angle has average to poor biometric performance.

These and other differences are further explored in this work. We believe that addressing this "gap" between biometric and forensic science in a practical framework, helps to bring biometric and forensic science closer together. In particular, this framework is designed to be applicable to research that considers modalities and feature representations, uses biometric classifiers to produce strength of evidence, and incorporates an evaluation of the suitability of features and score functions for forensic evidence evaluation purposes.

The contributions of this paper are:

- A systematic presentation of aspects to be considered during the design and evaluation of biometric classifiers for forensic evidence evaluation, including forensic performance characteristics and metrics;
- An emphasis on general versus subject based performance evaluation;
- A presentation of two relevant applications within the domain of forensic face recognition.

The paper is structured as follows. Section 8.2.3 presents the framework and explores the six aspects. In Section 8.2.4 we present the first application of the framework. It considers nine FISWG characteristic descriptors in the situation when trace images depict a perpetrator wearing a balaclava. FISWG characteristic descriptors [8] are facial features that can be taken into consideration during forensic evidence evaluation of trace and reference material. Section 8.2.5 contains the second, smaller, application of the framework. It extends the previously mentioned work on facial marks on good quality images [24]. In particular, it considers the case when trace images originate from surveillance cameras.

Two remarks are in order. First, both applications take the form of a small paper within a paper, that is, in principle they can be read independently from the framework description. Second, since related work is either confined to the framework or the two applications, we do not provide a separate related work section.

8.2.3 Framework

In this section we present our framework. As can be seen in Figure 8.1, it considers six aspects that reside at the design level of biometric classifiers for forensic evidence evaluation. These aspects influence the biometric system and its evaluation at an operational level. The aspects can be grouped into three groups of two aspects; they influence the feature, score calculation, and the evaluation, respectively. Choices addressed in or related to the first five aspects can serve as a template for their influence on Aspect 6, the chosen performance characteristic(s).

This framework is a generalisation of the six aspects discussed in [24] and contains the forensic performance characteristics presented in a recently published forensic guideline [48].

Note that the term framework might suggest that it is imperative to use all suggested choices discussed in each of the aspects. We rather see it as a *checklist* to at least consider those aspects while designing and evaluating these classifiers.

Modality relevance and selection within

The first aspect considers the relevance of the modality from a biometric and forensic point of view. Modalities can be composed of several smaller entities which on their own are modalities as well. Due to the forensic context, may be only some of these modalities can be used.

Jain et al. [30] identifies seven characteristics of biometric modalities: universality, distinctiveness, permanence, measurability, performance, acceptability, and circumvention. In particular, the distinctiveness property is used in forensic scenarios as identification (both closed and open), investigation, intelligence, and the evaluation of strength of evidence [1].

Most of these characteristics form a first test for suitability of a modality for forensic purposes. Acceptability is a notable exception; robustness to forensic scenarios and the availability of biometric information as a trace are additional forensically important characteristics. Distinctiveness does not need to be uniform amongst the subjects in order to have a forensic interest. We explore this in more detail in Section 8.2.3.

The second test for suitability is the forensic relevance of the modality, either at source or activity level inference. It involves the recognition in which forensic use case(s) the modality could be used for which forensic scenario(s) as described above. A forensic use case describes an act that produces a particular type of trace material captured at a crime scene. An example is a robbery in which the robber wears a balaclava and trace material only shows a few facial parts (eyes, eyebrows, mouth, possibly part of nose and part of the chin). This forensic use case is further explored in Section 8.2.4.

The forensic scenario(s) and use case(s) also determine whether within the modality a selection is made. The balaclava use case mentioned here above is an example of this selection. Another example is Zeinstra et al. [24] that only considers a selection of facial marks (prominent and permanent); this deviates from almost all other research regarding biometric suitability of facial marks. The rationale behind this choice is that the measurement of these facial mark types is assumed to be more robust to sensitivities as time lapse, image illumination, image resolution, and image blurring.

Feature Representation

Given the modality of interest, the possible feature representation(s) are to be considered. We believe that it is imperative to use the outcome of the previous step in the design of the feature representation. An example is again the use of facial marks. Since the measurement of the location of a facial mark is subject to for example within variation, one might also consider the use of a facial grid to represent the location in terms of the grid cell it belongs to, as a method to compensate for this within variation. The suitability of facial marks using such feature representation is given as an application in Section 8.2.5.

Classifier

The third aspect is whether we use any data to train a classifier and if so, whether that data is related to a general population or to a subject. As mentioned before, although we use the term classifier, we are interested in the comparison score it produces, rather than a decision. The rationale of this aspect can be illustrated by extending the facial mark example used in the previous aspect. One can argue that observing a facial mark at the same grid cell in two images is in general more rare than observing the absence of a facial mark. Moreover, the general probability of observing a facial mark at for example the cheek bone and cheek area is larger than on an eye lid, so the location of a facial mark influences the observation probability. This implies that using facial mark location data in a classifier could enhance its ability to discriminate. A statistical model based on subject based data could even be stronger than one based on general data, especially if the facial mark locations are very distinctive for that subject. We refer to Meuwly [163] for a general description of a framework that incorporates the use of subject based data for the calculation of strength of evidence. We further explore the use of general and subject based data in the next aspect.

Strength of evidence

A fundamental difference between biometric and forensic science with respect to the outcome of a comparison process, is what is being reported. In biometric science, we are interested in a comparison score; in forensic science we are interested in strength of evidence. We refer to Ramos and Gonzalez-Rodriguez [145] for a further discussion on this topic. One of the choices to be made is whether we convert a biometric score into strength of evidence or we adopt a feature based approach that produces strength of evidence directly based on the similarity and rarity of features. An example of the latter is the use of a statistical model in the classifier such that its outcome is interpretable as strength of evidence. Forensic work could study differences between these two classifier types. In the forensic community, this distinction is commonly referred to as score based likelihood ratio versus feature based likelihood ratio classifiers. Strength of evidence is commonly expressed as a likelihood ratio in modern forensic science[1]:

$$\mathrm{LR}(E) = \frac{\mathrm{p}(E|\mathscr{H}_s, I)}{\mathrm{p}(E|\mathscr{H}_d, I)}. \tag{8.1}$$

Here E denotes evidence, $\mathscr{H}_s$ is the same source hypothesis, $\mathscr{H}_d$ is the different source hypothesis, and I is background information. As described in Jackson et al. [44], the forensic examiner is responsible for the calculation of $\mathrm{LR}(E)$, whereas a court of law determines the prior odds $\frac{\mathrm{p}(\mathscr{H}_s|I)}{\mathrm{p}(\mathscr{H}_d|I)}$ and ultimately the posterior odds $\frac{\mathrm{p}(\mathscr{H}_s|E,I)}{\mathrm{p}(\mathscr{H}_d|E,I)}$:

$$\frac{\mathrm{p}(\mathscr{H}_s|E,I)}{\mathrm{p}(\mathscr{H}_d|E,I)} = \mathrm{LR}(E) \times \frac{\mathrm{p}(\mathscr{H}_s|I)}{\mathrm{p}(\mathscr{H}_d|I)}. \tag{8.2}$$

Finally, often (8.1) is used in its $\log_{10}$ form:

$$\mathrm{LLR}(E) = \log_{10}\left(\frac{\mathrm{p}(E|\mathscr{H}_s, I)}{\mathrm{p}(E|\mathscr{H}_d, I)}\right). \tag{8.3}$$

The advantage of using (8.3) over (8.1) is the emphasis on the magnitude of the likelihood ratio rather than its exact value.

The definition of the likelihood ratio (8.3) is not directly usable. The background information I is case dependent and typically involves auxiliary information like the model of the jacket worn by the perpetrator in the trace material. Therefore, we exclude the background information in the current work that describes a generic approach. Both the biometric comparison score and the features can be considered as evidence in (8.3).

If the evidence E is the simultaneous occurrence of trace x and reference y, we obtain the *feature based log-likelihood ratio*:

$$\mathrm{LLR}(\mathrm{x},\mathrm{y}) = \log_{10}\left(\frac{\mathrm{p}(x,y|\mathscr{H}_s)}{\mathrm{p}(x,y|\mathscr{H}_d)}\right). \tag{8.4}$$

One approach to calculate (8.4) is to assume parametric models for $\mathrm{p}(x,y|\mathscr{H}_s)$ and $\mathrm{p}(x,y|\mathscr{H}_d)$. Other approaches exist as well, including for example the use of copula models that relate joint probability distributions to their marginal distributions. We refer to the dissertation of Susyanto [45] for more information on copula approaches.

[1] Although Darboux, Appell, and Poincaré suggested its use already in 1906 for the appeal in the Dreyfus case [43], mostly during the last decade it has seen a mainstream acceptance.

If the evidence E is a biometric comparison score $s = s(x,y)$ computed on trace x and reference y, then (8.3) reverts to the *score based log-likelihood ratio*:

$$\mathrm{LLR}(s) = \log_{10}\left(\frac{\mathrm{p}(s|\mathscr{H}_s)}{\mathrm{p}(s|\mathscr{H}_d)}\right). \tag{8.5}$$

Several techniques can be used to estimate the numerator and denominator of (8.5). Examples are parametric model based (for example a normal distribution) or non-parametric (for example Parzen windows [46]). Another approach is the use of the Pool of Adjacent Violators (PAV) algorithm [47]. Given a training set of scores, this algorithm estimates $\mathrm{p}(\mathscr{H}_s|s)$ from which the likelihood $\mathrm{LLR}(s)$ can be derived:

$$\mathrm{LLR}(s) = \mathrm{logit}(\mathrm{p}(\mathscr{H}_s|s)) - \mathrm{logit}(\mathrm{p}(\mathscr{H}_p)), \tag{8.6}$$

with $\mathrm{logit}(x) = \log_{10}\left(\frac{x}{1-x}\right)$. Note that the prior $\mathrm{p}(\mathscr{H}_s)$ in (8.6) is the fraction of same source pairs in the training set and it is not the prior $\mathrm{p}(\mathscr{H}_s)$ set by a court of law. This process is an example of *score calibration* [91].

Both biometric comparison score functions and feature based log-likelihood ratio functions may include parameters that reflect general behaviour or even subject based behaviour. In particular, the same source $\mathscr{H}_s$ and different source $\mathscr{H}_d$ hypotheses can be formulated in two, distinct, manners. The general formulation is

- $\mathscr{H}_s = \mathscr{H}_s^g$: the trace x and reference y originate from a common donor.
- $\mathscr{H}_d = \mathscr{H}_d^g$: the trace x and reference y do not have a common donor.

The subject based formulation is

- $\mathscr{H}_s = \mathscr{H}_s^s$: the trace x and reference y originate from the same specific donor.
- $\mathscr{H}_d = \mathscr{H}_d^s$: the trace x and reference y do not have the same specific donor.

Since the subject based formulation is tailored towards a specific subject (the suspect), one could argue that the subject based formulation should be favoured over the general formulation. In the general based approach, a separate training set is used for the estimation of relevant statistical parameters; in the subject based approach for each subject a proportion of the measurements must be kept aside as a training set. This marks a drawback of the subject based formulation: it might lack data for a reliable estimate of the involved probability distribution.

Evaluation level

Another consideration is at which level performance characteristics are evaluated. From a biometric point of view, often only the general discriminating power is of interest. We refer to this as an example of *general evaluation*.

However, since some modalities are generally not very discriminating, they might be for certain subjects. This is illustrated in Figure 8.2. The left image in this figure represents the ideal situation in which features have a low within variation and a larger between variation, making discrimination between all subjects possible. The middle image represents the

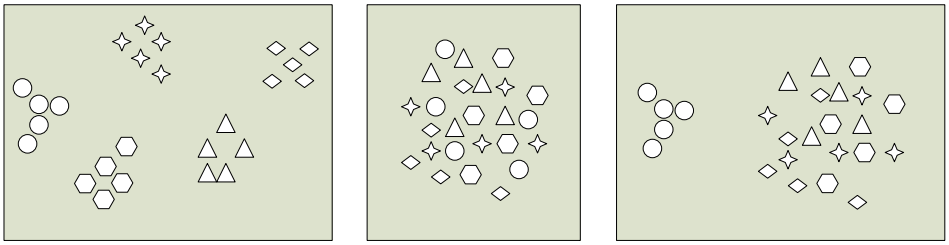


Figure 8.2: Five subjects with five observations in feature space. A generally strong biometric modality (left), a weak one (middle), and a weak one that can be used to discriminate one subject (right).

situation in which in general the within and between variation are similar and yields poor discrimination. The right image in this figure shows the situation in which one subject can be discriminated from the other subjects.

This example suggests that it makes sense to also report at a subject based level, at least to get an idea of the range of attained values. We refer to this as *subject based evaluation*. An additional advantage of this approach is that one can try to explain extreme performance metrics in terms of the phenotype of the feature. We suggest the use of Tukey or box plots [160] that show (a) the median, (b) whiskers with boundaries determined by a quartile proportion, and (c) outliers. Observe that the use of subject based data is independent of subject based evaluation: it is possible to perform a subject based evaluation on any classifier.

Performance Characteristics and Metrics

Biometric performance is often confined to ROC, AUC or EER. According to a recently proposed guideline by Meuwly et al. [48], used as a basis for an ISO standard, there are several other performance characteristics and corresponding metrics that are relevant in the context of our framework. In their work, they classify the performance characteristics into primary and secondary classes. The primary class encompasses

- Accuracy
- Discriminating power
- Calibration

Accuracy is the "closeness of agreement between computed likelihood ratio and the ground truth status" and is measured in Cllr. Given a set $\mathscr{S}$ of n_s and a set $\mathscr{D}$ of n_d scores under the same source hypothesis $\mathscr{H}_s$ and different source hypothesis $\mathscr{H}_d$) respectively, the cost of log-likelihood ratio [49] is defined by:

$$\text{Cllr} = \frac{1}{2}\left(\frac{1}{n_s}\sum_{s\in\mathscr{S}}\log_2(1+\mathrm{e}^{-s}) + \frac{1}{n_d}\sum_{s\in\mathscr{D}}\log_2(1+\mathrm{e}^{s})\right). \tag{8.7}$$

Discriminating power is a "property representing the capability of a given method to distinguish amongst forensic comparisons where different propositions are true", and is either

measured in EER or Cllr^{min}. With respect to the latter property, if we apply the PAV algorithm to the set of scores and reapply (8.7), we obtain the minimal achievable cost of likelihood ratio Cllr^{min}. This quantity measures the discriminating power and can be used as an alternative to EER.

Calibration is a "property of a set of LRs (...)". Perfect calibration means that LRs can be interpreted as strength of evidence. Its performance metric is calibration loss:

$$\text{Cllr}^{\text{cal}} = \text{Cllr} - \text{Cllr}^{\text{min}}. \tag{8.8}$$

Calibration loss essentially measures how well the computed likelihood ratio can be used as strength of evidence in a court of law.

The secondary performance characteristics are

- Robustness
- Coherence
- Generalisation

Robustness refers to "the ability of the method to maintain a performance metric when a measurable property in the data changes". Coherence is the "ability to yield likelihood ratio values with better performance with the increase of intrinsic quantity/quality (...)". Generalisation refers to the "ability to maintain performance under a dataset shift." The secondary performance characteristics are measured in Cllr or EER.

8.2.4 Application 1: Balaclava

Introduction

Suppose a perpetrator wears a three hole balaclava during a robbery. It covers the face, but might reveal the eyebrows, eyes, mouth and lower part of the nose[2] as shown in Figure 8.3a.

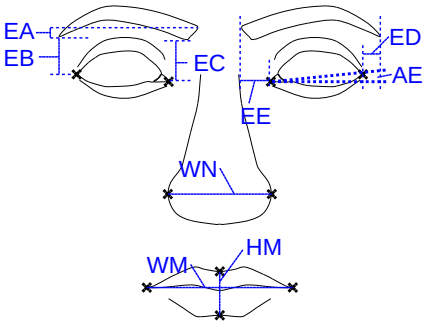
Although the shown example is of good quality, trace images are typically taken under challenging conditions that lower the image quality. Shape information can be difficult to discern in these low quality images, whereas it might be possible to extract angles, positions, and distances. Therefore, in this application we consider nine simple characteristic descriptors described in the FISWG Feature List [8], published by the Facial Identification Scientific Working Group [7]. This group, in which several forensic institutes are organised (including the FBI), has published several recommendations regarding facial comparison for forensic purposes.

The nine descriptors are shown in Figure 8.3b and they are the angle fissure, five distinctive eyebrow measures A-E, the height and width of the mouth, and the width of the nose. We expect that these descriptors have limited general discriminating power, but are also able to discriminate some subjects to a certain extent. Also, we are interested in differences in discriminating power between classifiers that produce a score and those that produce a likelihood ratio based on either general or subject based data.

This leads to the following two research questions:

[2] Balaclavas with one or two holes exist. Moreover, what is revealed depends on factors like the position of facial parts within the face and how the balaclava is worn.

a) Balaclava



b) Nine FISWG characteristic descriptors

Figure 8.3: From left to right: a) Subject wearing a balaclava with three holes, b) Face showing the nine considered FISWG characteristic descriptors: angle fissure (AF), five distinctive eyebrow measures A-E (EA-EE), the height (HM) and width (WM) of the mouth, and the width of the nose (WN).

RQ1 What is the discriminating power of an untrained classifier and those trained on general or subject based data viewed at a general and a subject based evaluation level?
RQ2 Which feature phenotypes correspond to good and poor subject based discriminating power?

Related Work

The periocular region (area around the eye) has been studied in its capacity as a soft biometric modality and as a biometric modality on its own. Soft biometric modalities can be used to augment other, "hard", biometric modalities when for example the image quality has been degraded. One of the papers that initiated research in the periocular region is Park et al [112]. They explored the use of a local (SIFT) and global (HOG, LBP) approach to describe the texture of the periocular region. Other studies further explored the use of LBP and its many variants, see for example the work by Miller et al. [127], Xu et al. [129], and Mahalingam and Ricanek [113]. Also, two studies by Hollingsworth et al. ([114, 115]) investigated what humans use during the recognition of the periocular region shown in near-infrared images. Finally, the periocular region has also been studied on a limited set of 40 images using all FISWG characteristic descriptors of the eye in Zeinstra et al [20]. It was found that some features like the shape of the iris (including to which extent the eye is opened) yields an AUC=0.95, whereas the angle fissure yields AUC=0.70.

Also the eyebrow modality itself has been the topic of several studies, for example Dong and Woodard [19] and Xu and Savvides [109]. In the latter study, it was shown that the eyebrow region accounts for $\frac{1}{6}$ of the facial region while it retains $\frac{5}{6}$ of the performance of the facial region. In Zeinstra et al. [18], a classifier using the Dong Woodard description of eyebrows was compared to classifiers that use the FISWG descriptors of the eyebrow. The

performances were found to be comparable.

The remaining modalities (nose and mouth) have almost never been studied. For example, the study of Moorhouse [164] considered the nose using photometric stereo images; lips as a biometric modality have been studied in Choraś [165].

In general, FISWG characteristic descriptors have been the subject of several related studies. For example, in Zeinstra et al. [22], the discriminating power in terms of EER of almost all descriptors was systematically investigated using four different trace image types. It was found that only in the case of the severest trace image quality, the hairline and a combination of binary descriptors performed somewhat better than a face recognition system that showed essentially random performance.

Framework applied

Modality relevance and selection within. The forensic relevance of the nine descriptors has already been explained. Moreover, these features are exemplary for a category of features with limited general discriminating power that can discriminate some subjects to a certain extent. Additionally, straightforward parametric statistical models can be trained and evaluated on these feature types.

Feature Representation. All measures are one-dimensional real numbers; all but one (angle fissure) are either a distance or a relative position. The angle fissure is measured in degrees.

Classifier. We employ classifiers that are untrained and ones are trained on either general or subject based data.

Strength of evidence. For each of the FISWG characteristic descriptors, we choose three different score comparison functions. The Euclidean distance score is

$$s(x,y) = -|x-y| \tag{8.9}$$

and is PAV calibrated, from which the likelihood ratio (8.6) can be computed.

We also use the feature based log-likelihood ratio (8.4) and assume that the feature values are normally distributed. Using the general model, we assume that under the same source hypothesis we have

$$\begin{pmatrix} x \\ y \end{pmatrix} | \mathscr{H}_s^g = \mathscr{N}\left(\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix} \right) \tag{8.10}$$

and under the different source hypothesis

$$\begin{pmatrix} x \\ y \end{pmatrix} | \mathscr{H}_d^g = \mathscr{N}\left(\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \begin{pmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_y^2 \end{pmatrix} \right). \tag{8.11}$$

Here all parameters ($\mu_x, \mu_y, \sigma_x^2, \sigma_y^2$, and ρ) are estimated on a separate training set. We also formulate a subject based model, for which under the same source hypothesis we have

$$\begin{pmatrix} x \\ y \end{pmatrix} | \mathscr{H}_s^s = \mathscr{N}\left(\begin{pmatrix} \mu_x^s \\ \mu_y^s \end{pmatrix}, \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix} \right) \tag{8.12}$$

and under the different source hypothesis

$$\begin{pmatrix} x \\ y \end{pmatrix} | \mathscr{H}_d^s = \mathscr{N}\left(\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \begin{pmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_y^2 \end{pmatrix} \right). \tag{8.13}$$

The only difference between the general model and this model are the subject specific means in the same source model. These means are estimated from a small subject based training set. Under these models we can derive two analytic expressions for the log-likelihood ratio (8.4). Under the general model we have

$$\begin{aligned}\mathrm{LLR}(x,y) &= \log_{10}\left(\frac{\mathrm{p}(x,y|\mathscr{H}_s^g)}{\mathrm{p}(x,y|\mathscr{H}_d^g)}\right)\\ &= -\frac{1}{2}\log_{10}(1-\rho^2) - \frac{1}{2}\log_{10}(\mathrm{e})g(x,y)\end{aligned} \tag{8.14}$$

with

$$\begin{aligned}g(x,y) &= A(x-\mu_x)^2 + B(x-\mu_x)(y-\mu_y) + C(y-\mu_y)^2\\ A = \frac{\rho^2}{\sigma_x^2(1-\rho^2)}, B &= -\frac{2\rho}{\sigma_x\sigma_y(1-\rho^2)}, C = \frac{\rho^2}{\sigma_y^2(1-\rho^2)}.\end{aligned} \tag{8.15}$$

Under the subject based model we have

$$\begin{aligned}\mathrm{LLR}(x,y) &= \log_{10}\left(\frac{\mathrm{p}(x,y|\mathscr{H}_s^s)}{\mathrm{p}(x,y|\mathscr{H}_d^s)}\right)\\ &= -\frac{1}{2}\log_{10}(1-\rho^2) - \frac{1}{2}\log_{10}(\mathrm{e})p(x,y)\end{aligned} \tag{8.16}$$

with

$$\begin{aligned}p(x,y) &= A(x-\mu_x^s)^2 + B(x-\mu_x^s)(y-\mu_y^s)\\ &\quad + C(y-\mu_y^s)^2 + D(x-\mu_x)^2 + E(y-\mu_y)^2,\\ A = \frac{1}{\sigma_x^2(1-\rho^2)}, B &= -\frac{2\rho}{\sigma_x\sigma_y(1-\rho^2)}\\ C = \frac{1}{\sigma_y^2(1-\rho^2)}, D &= -\frac{1}{\sigma_x^2}, E = -\frac{1}{\sigma_y^2}.\end{aligned} \tag{8.17}$$

Evaluation level. Given research question RQ1, we are interested in using both the general and subject based evaluation level.
Performance characteristics. We select discriminating power as the performance characteristic.

Experimental setup

We use the FRGCv2 dataset [33]. This dataset has been used in many face recognition studies and algorithm evaluations; it consists of 568 subjects for which 2D/2.5D frontal images are available. The 2D images are taken under controlled and uncontrolled conditions (hallway with non-frontal illumination, larger distance and blurring/movement) and without or with expression (smiling). All images are taken in four subsequent semesters.

For this experiment, we select all (12307) images taken under controlled conditions showing subjects with a neutral expression. We follow the same procedure on training and the

general and subject based evaluation of classifier discriminating power as described in [24]. We highlight the main points here.

In total 376 subjects have less than 25 recordings and are considered to be only train subjects. For the remaining 192 subjects with 25 or more recordings, the first ten recordings are reserved as subject based training data; the remainder of the recordings constitute the test set.

For each subject we construct (a) a set of same source comparisons: test recordings of this subject compared to other test recordings of this subject and (b) a set of different source comparisons: test recordings of this subject compared to test recordings of other subjects. By construction, each subject has a least 15 (25-10) test recordings, yielding at least 100 same source scores. The general evaluation is done on a set of same/different source scores sampled from the subject based sets; for details we refer to [24].

The nine FISWG characteristic descriptors are determined indirectly in an automatic setup. We use the One Millisecond Deformable Shape Tracking Library (DEST) [106] for landmark location. It uses an ensemble of randomised regression trees and gives accurate landmark locations. We do not employ the default landmark model of DEST as it is too coarse for our purposes. We train DEST using all available (2330) images in the HELEN database [137] and the available ground truth annotation provided by STASM [107] of a model containing 199 landmarks. HELEN contains "...high-resolution, accurately labeled face images and has a larger degree of out-of-plane orientation and rotation...". An affine transformation is then applied on the landmark positions such that the found pupil coordinates of each image are mapped to fixed locations. Finally, we extract the nine descriptors from the landmarks in this coordinate system.

Results and discussion

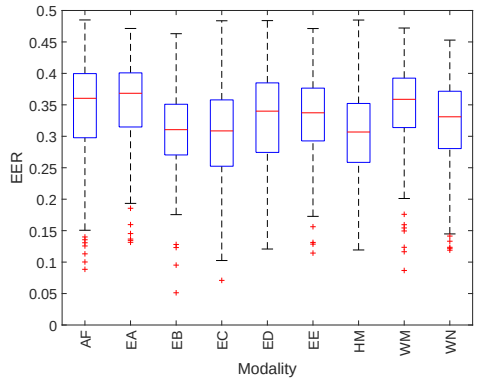
Regarding RQ1, the discriminating power of the three classifier types at a general and subject based level, Figure 8.4 we shows the box plots of the EER for comparison methods that do not require training, those trained on general data, and those trained on subject based data. We observe that all considered characteristic descriptors can be seen as soft biometric modalities as they have a very moderate median EER.

Although the box plots appear very similar, we can show, using a Wilcoxon signed rank test, that for each considered characteristic descriptor, the subject based method is better than the general method which in turn is better than the score based method ($p < 0.1\%$). This relationship is reinforced by their corresponding high correlation coefficients: $\rho \in [0.91, 0.99]$. Stated differently: although the Wilcoxon signed rank test shows the relative ordering of the models, there is a very strong relationship between the EER values of the three models. This seems to imply that even using an untrained classifier, one could draw stronger conclusions as if the strength of evidence had been calculated using subject based statistics.

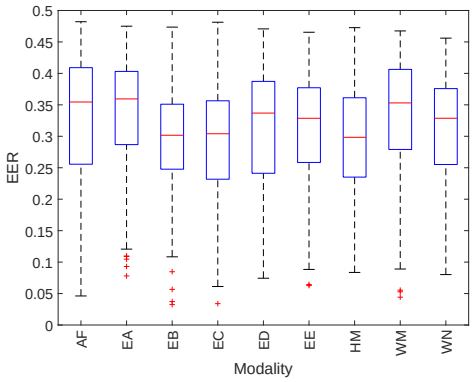
Also a number of correlations between the EER of different characteristic descriptors exist. In particular, we found that the EER of the eyebrow A/C, eyebrow B/C and width of mouth and nose are correlated with $p < 0.1\%$.

What makes the considered characteristics particularly interesting is the performance difference between some subjects. In Figure 8.5, for each of the descriptors, we show an outline[3] of the best (green) and worst (red) performing subjects with their performance, alongside with

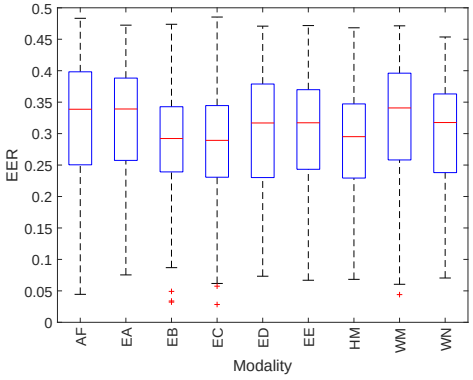
[3] Showing the outline is clearer than showing the actual facial patch.



a) Score LLR



b) Feature LLR, general model

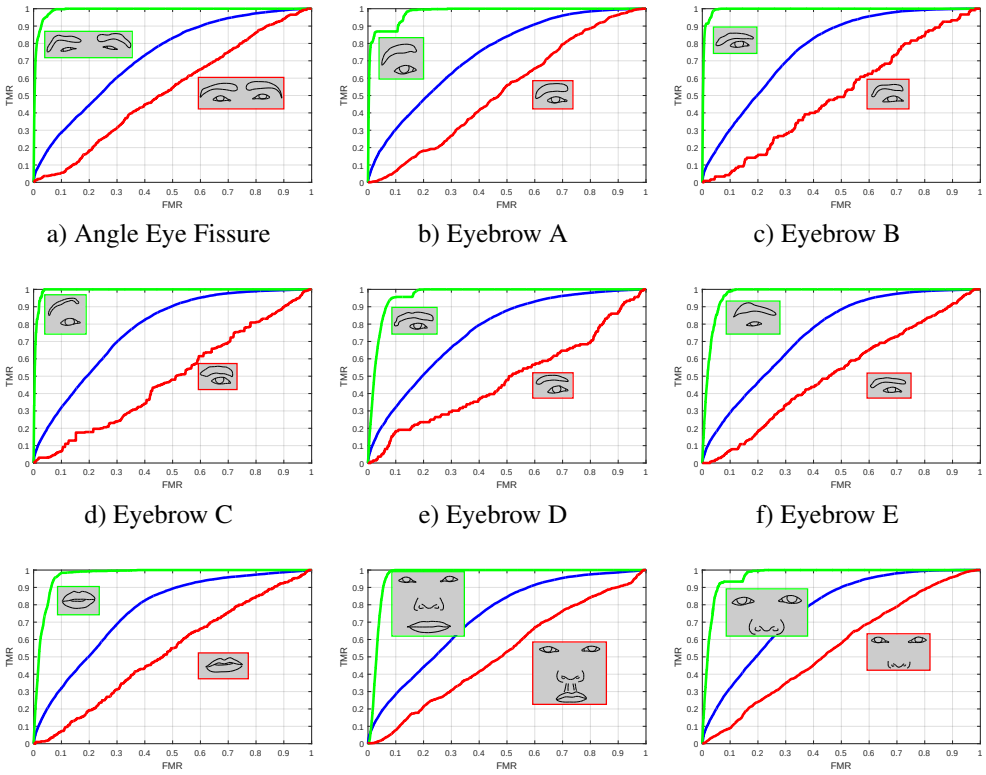


c) Feature LLR, subject based model

Figure 8.4: From top left to middle bottom: box plots of (a) EER of score based likelihood ratio, (b) EER of feature based likelihood ratios using a general model, and (c) EER of feature based likelihood ratios using a subject based model. The following nine FISWG characteristic descriptors are considered: angle eye fissure (AE), eyebrow A-E (EA-EE), height mouth (HM), width mouth (WM), and width nose (WN).

the general performance (blue) in a ROC curve. These examples show that the performance at a subject level can be explained in terms of the phenotype of the feature. As an example, Figure 8.5a shows that having low outer eye corners in relation to the inner eye corners is discriminative, whereas they are more leveled, they are essentially random. Also, the general performance is in line with the one reported in Zeinstra et al [20]. These results address RQ2 on the connection between phenotype and discriminating power.

This is one of the key observations in relation to forensic evidence evaluation and it reiterates what is depicted in Figure 8.2. In forensic evidence evaluation, in principle we are only interested in discriminating a particular subject from a group of subjects, rather than the much stronger property of discriminating everyone, including this particular subject. Although we do not claim this insight is new, but given the large variation between subjects, it seems warranted to emphasise this in the context of the validation of likelihood ratio methods



a) Angle Eye Fissure b) Eyebrow A c) Eyebrow B

d) Eyebrow C e) Eyebrow D f) Eyebrow E

g) Height Mouth h) Width Mouth i) Width Nose

Figure 8.5: The variation in performance of nine FISWG characteristic descriptors: (a) angle fissure, (b)-(f) eyebrow A-E, (g) height mouth, (h) width mouth, and (i) width nose. Green and red refers to the best and worst performing subject and performance respectively, blue is the general performance.

for forensic evidence evaluation as described by [48]. This guideline does not specify the level of evidence evaluation.

8.2.5 Application 2: Grid Based Facial Mark Likelihood Ratio Classifiers

Introduction

The second application is a small extension of previous work [24] on facial marks. In that work, primary performance characteristics (discriminating power and calibration) of six classifiers operating on a facial grid representing facial mark locations were compared. Moreover, the influence of grid cell sizes ranging from 0.05 IPD (interpupillary distance) to 1.0 IPD on these characteristics was studied as well. A facial grid example is shown in Figure 8.6c with $\Delta = 0.25$ IPD. A feature based likelihood ratio classifier trained on subject based data was

found to outperform every other considered approach. However, it was also discovered that the calibration of classifiers trained on subject based data was severely hampered when the grid cell size was lower than approximately 0.40 IPD, restricting its mode of operation for the computation of strength of evidence.

The study did not consider secondary performance characteristics like generalisation of discriminating power. In particular, only a subset of images of FRGCv2 [33] taken under controlled conditions and showing subjects with a neutral expression was used. Therefore, in this work we aim to extend the previous work by using another dataset, SCFace [92], in order to address the generalisation of discriminating power. This dataset contains stills taken from several surveillance cameras at three distances and high resolution images, and is used in numerous studies on low resolution face recognition.

Due to the inherent poor image quality and low(er) resolution of surveillance camera stills, we expect a large reduction in the number of detected facial marks in trace images relative to the corresponding reference images, introducing a systematic difference and rendering even classifiers trained on general data useless. Therefore, we only consider the Hamming classifier that (a) only uses the presence of facial marks in grid cells and (b) does not use any general or subject based facial mark location data.

This leads to the following two research questions:

RQ1 Can we generalise the discriminating power of Hamming based classifiers that use a facial mark grid?

Finally, we expect there exist large differences in EER at a subject level:

RQ2 How many subjects can still be discriminated, using their facial mark grid?

Related Work

Since this work is a small extension of [24], we briefly mention some of the literature discussed in that work.

Facial marks are an integral part of the Bertillonage system [53] that was used to describe individuals for law enforcement purposes. Facial marks are still relevant today from a forensic perspective and are mentioned in the FISWG Feature List [8]. Studies have shown that facial marks can indeed be discriminative, especially when comparing mono zygotic twins; see for example [148, 149]. A survey by Jain et al. [50] discussed several specific applications of forensic face recognition including mugshot retrieval using facial scars and marks.

Automatic detection of facial marks has also been studied. Notable examples include a study by Park and Jain [101] in which a facial mark detection system was presented that used an Active Appearance Model [152] for the location of facial features and a Laplacian of Gaussian blob detector, and other operators for post processing. Another, more recent, study is that of Srinivas et al. [102] in which the Fast Radial Symmetry Transform [153] for the detection of facial marks is employed. They compared the results of their automatically detected facial marks with those of a system in which the facial mark locations were post-processed by a human. They reported an EER on their High Resolution Face Database (HRFD) of 12%.

Framework applied

Modality relevance and selection within Several facial mark types exist, see for example [102] for an overview. As discussed in [24], not every facial mark type is suitable in a forensic

context. In particular, in [24] and the present study only the mole, pockmark, raised skin, and scars are taken into account.

Feature Representation. We assume that the facial mark locations are given in a coordinate system for which the pupil coordinates are $(-\frac{1}{2},0)$ and $(\frac{1}{2},0)$, making the IPD always equal to 1. We superimpose a grid with square cells having sizes Δ ranging from 0.05 IPD to 1.0 IPD in steps of 0.05 IPD. The feature is a binary vector in which each grid cell indicates whether it contains one or more facial marks.

Classifier. As discussed before, we only consider the Hamming comparison score

$$H((b_1^{ij}),(b_2^{ij})) = -\sum_{i,j} |b_1^{ij} - b_2^{ij}|. \tag{8.18}$$

Strength of Evidence. Since the Hamming comparison function does not produce likelihood ratio values, we use PAV calibration and (8.6) to create a score based likelihood ratio.

Evaluation Level We evaluate both at a general and a subject based level, as we expect that facial marks on low quality images have poor discriminating power, but might discriminate certain subjects.

Performance Characteristics Generalisation is chosen to augment previous work.

Experimental Setup

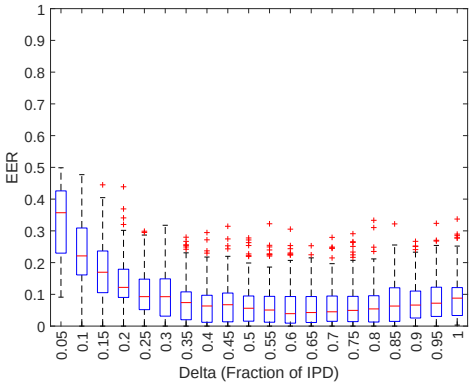
We use the FRGCv2 in the manner described in [24]. The SCFace dataset contains surveillance footage of six cameras in seven different configurations (visible and IR), depicting a subject at three different distances (4.20m, 2.60m, and 1.00m), IR mugshot and high resolution images for 130 subjects.

We define the experiment on the SCFace dataset as follows. We manually locate facial marks in reference images and then annotate all trace images, in random order. We use the same web application as described in [24]. The facial mark locations are subsequently mapped to the fixed coordinate system introduced before.

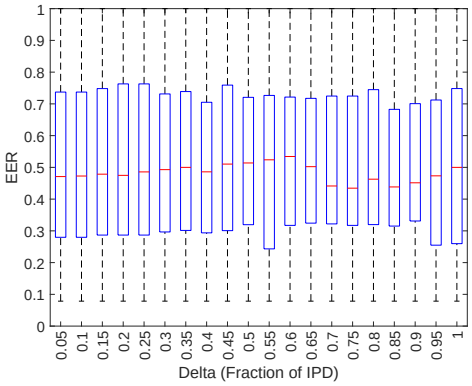
Results and Discussion

Regarding RQ1, we compare the EER on FRGCv2 (Figure 8.6a) with the EER on SCFace for Camera 1 and distance 3 (Figure 8.6b). Other cameras exhibit similar results and are therefore omitted. We observe that in the FRGCv2 case, the EER has some dependency on the grid cell size (notably with smaller grid cell sizes, its increase is explained by within variation) and some variation between subjects. On the other hand, in the SCFace case we observe a very poor EER that is even mostly independent of the grid cell size. With respect to RQ1, we conclude that the discriminating power of Hamming based classifiers using facial mark grids cannot be generalised.

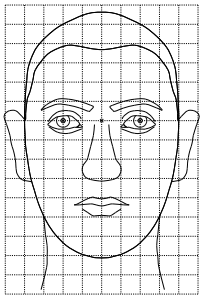
With respect to RQ2, we find that results between subjects vary to a large extent. There is a number of subjects that even have EER=1; this is caused by the large mismatch of observed facial marks in the trace and reference image belonging to that subject in relation to the differences in facial mark observations between its trace and reference images of other subjects. Mostly for distance 3 (1.00m) we found up to ten different subjects (depending on the grid cell size) that can be perfectly discriminated based on their facial mark grid. An example



a) FRGCv2

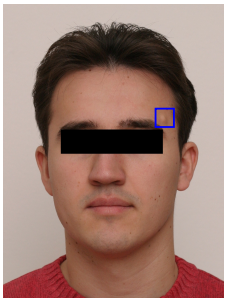


b) SCFace (Cam 1, Dist 3)



c) Grid

d) Trace



e) Reference

Figure 8.6: Top row: EER of Hamming comparison score function on a) FRGCv2, b) SCFace Camera 1, Distance 3. Bottom row: c) Example grid (with grid cell size 0.25 IPD), and a subject that can perfectly be discriminated based on the indicated facial mark location d) Trace, and e) Reference.

of such case is shown in Figures 8.6d and 8.6e. Subject 009 has a facial mark (raised skin) located at his left temple, clearly visible in both trace and reference images.[4]

8.2.6 Conclusion

In this paper, we have presented a framework that considers six aspects during the design and evaluation of biometric classifiers for forensic evidence evaluation. These factors influence the design of the feature, the biometric score and its forensic use, and the evaluation of performance characteristics. We also presented two applications of this framework.

The first application deals with the situation in which trace images depict a perpetrator wearing a balaclava. We explored the use of nine simple characteristic descriptors and found that the incorporation of either general or subject based data versus no training yields very similar classifiers. Moreover, we observed a large variation in discriminating power between

[4]These images have been anonymised due to the database release statement of SCFace, see Acknowledgement.

subjects. We believe that this behaviour can be attributed to, in some cases the extreme, phenotype of the considered features.

The second application is an extension of existing work on classifiers that use facial marks as features on a large subset of the FRGCv2 dataset. The extension considered a secondary performance characteristic and used the SCFace dataset. The EER of Hamming based classifiers is in the case of the SCFace dataset very poor and we concluded that its good results on the FRGCv2 dataset cannot be generalised. Despite the lack of facial mark observations in the SCFace case, we did find subjects that could be discriminated based on their facial mark grid.

We overall conclude that the framework has an added value for the forensic biometric community and can be used to design and evaluate biometric classifiers for forensic evidence evaluation.

8.2.7 Acknowledgement

We would like to thank Prof. Grgic for his kind permission to let us use an anonymised version of a subject of the SCFace dataset in Figures 8.6d and 8.6e.

8.3 How random is a classifier given its Area under Curve?

8.3.1 Abstract

When the performance of a classifier is empirically evaluated, the Area Under Curve (AUC) is commonly used as a one dimensional performance measure. In general, the focus is on good performance (AUC towards 1). In this paper, we study the other side of the performance spectrum (AUC towards 0.50) as we are interested to which extent a classifier is random given its AUC. We present the *exact* probability of AUC values of a truly random classifier, given a finite number of *distinct* genuine and imposter scores. It quantifies the "randomness" of the measured AUC. The probabilities involve the restricted partition function, a well studied function in number theory. Although other work exists that considers confidence bounds on the AUC, the novelty is that we do not assume any underlying parametric or non-parametric model or specify an error rate. Also, in cases in which a limited number of scores is available, for example forensic case work, the exact probability can deviate from these models. For completeness, we also present an approximation using a normal distribution and confidence bounds on the AUC value.

8.3.2 Introduction

The trade off between the False Match Rate (FMR) and True Match Rate (TMR) of a classifier while varying the decision threshold is commonly reported in a receiver operating characteristic (ROC) curve [166]. There exist several one dimensional classifier performance measures that can be derived from its ROC curve, for example, the Equal Error Rate and the Area under Curve [167]. In this study, we consider the Area Under Curve (AUC) measure. An ideal classifier has AUC=1, whereas a random classifier has AUC=0.50. The AUC is equal to the probability that a randomly chosen genuine score is larger than a randomly chosen imposter

score [42]. Also, the AUC can be interpreted as the Wilcoxon-Mann-Whitney statistic [168] when ordering the genuine and imposter scores produced by the classifier [167], [169].

In any empirical performance evaluation, only a finite number of genuine and imposter scores is available. Under the assumption that genuine and imposter scores are drawn from unknown probability densities, ultimately the AUC is also a random variable, having a probability distribution on its own. If we could replicate the experiment having the exact same number of genuine and imposter scores, we most likely would have obtained a different ROC curve and AUC value. In particular, this implies that the performance evaluation might yield an AUC value that is not identified as being produced by a random classifier. This could occur in the case of a subject anchored approach to evidence evaluation in which the available number of scores is limited, see [163] for a general framework.

The probability of AUC values produced by a random classifier are easily derived for trivial cases. More precisely, we assume that (a) this classifier draws genuine and imposter scores randomly from the *same* probability distribution and (b) the drawn scores are *distinct*. The last condition is a necessary technicality; if we for example assume that scores come from a continuous interval, this condition is typically met. Suppose we construct an ROC curve based on 1 genuine score g and n imposter scores i_k, $k = 1,\ldots,n$. We have $n+1$ possible orderings of the genuine score among the already ordered imposter scores:

$$g < i_1 < \ldots < i_{n-1} < i_n \text{ to } i_1 < i_2 < \ldots < i_n < g. \tag{8.19}$$

Since g and i_k come from the same distribution, each sequence in (8.19) has equal probability $\frac{1}{n+1}$. If l ($l = 1,\ldots,n+1$) is the position of g in any sequence in (8.19), then its AUC is equal to $\frac{l-1}{n}$. Hence, each of these possible values for the AUC is equally probable to occur. The one-to-one mapping in this trivial 1 genuine/n imposter case between sequences and the AUC does not hold in the general case. For example, both i_1, g_1, g_2, i_2 and g_1, i_1, i_2, g_2 yield AUC=0.50, and the situation becomes rapidly more complex when m and n attain values found in practice.

This paper presents an exact expression for the probability of AUC values produced by a random classifier for *any* finite number of genuine and imposter scores. We also present an approximation to the exact probability. This work can be used in the situation when we want to determine the probability that a random classifier produces the measured AUC; this is of interest when the measured AUC is low or the total number of scores is limited.

The remainder of this article is structured as follows. In Section 8.3.3, we present related work. Since the general approach involves the restricted partition function, we present its definition in Section 8.3.4. In Section 8.3.5, we present two theorems regarding an exact expression and an approximation to the probability of AUC values. Section 8.3.6 presents some examples of the exact probability and an application of the approximation. In Section 8.3.7, we discuss the two theorems. Finally, in Section 8.3.8 we present our conclusion.

8.3.3 Related Work

As indicated before, this work fits in a larger framework that studies whether two AUCs are significantly different by constructing confidence intervals. This is not only of importance in decision theory, but also for clinical medicine and psychology studies in which treatments are compared. We present some of these studies here.

For example, the work of [170] analytically derives exact and estimated confidence intervals based on a statistical and combinatorial analysis, using a fixed error rate and the number of genuine and imposter scores. Our work only uses the number of genuine and imposter scores, assuming that they are drawn from the same probability distribution. Another approach is the use of parametric models to construct confidence intervals. For example score distributions have been modeled as normal [171], binormal [172], exponential [173], and Gamma [174], from which expressions for the confidence intervals can be derived. Their main issue is the influence of the parametric assumption on the estimation of confidence intervals. To cater for that situation, several non-parametric methods have been explored, including Wilcoxon-Mann-Whitney and De-Long non-parametric interval [175]. The work of [176] compares nine non-parametric approaches in different simulation scenarios (moderate to good AUC and different combinations of genuine and imposter scores). They found that their own empirical likelihood approach [177] has a good coverage in different scenarios. Several studies have shown that methods can be negatively influenced by the number of considered scores. For example, [178] found that asymptotic methods are less accurate in this situation; the study of [179] shows how estimates for the AUC can differ significantly from the true value.

In summary, these studies emphasise on one hand the restriction of our work (random classifier) and on the other hand its uniqueness (exact probability, depending on the number of genuine and imposter scores).

8.3.4 Partition functions

The partition function is an essential function in number theory, a branch of mathematics that studies properties of integers [180]. A partition of a positive integer k is a decomposition of k as a sum of positive integers. The partition function p counts the number of different partitions of a positive integer, disregarding any permutations in the order of the terms. For example $p(5) = 7$, since

$$5 = 5 = 4+1 = 3+2 = 3+1+1 = 2+2+1 = 2+1+1+1 = 1+1+1+1+1. \quad (8.20)$$

It is customary to order the terms in a partition from the largest to the lowest value. This can be written more formally as $k_1 + \ldots + k_r = k$, and $k_1 \geq k_2 \geq \cdots k_r$. Also, by convention, the domain of p is extended by including $p(0) = 1$ and $p(k) = 0$ for $k < 0$.

There exist different "restricted" versions of the partition function. In particular, one can limit the number and value of the terms of a partition. Let $p(n,m;k)$ be the number of partitions of k which have at most m terms, each having maximum value n. In the sequel, we refer to this function as "the" restricted partition function. For example, $p(4,2;5) = 2$, since the maximum value is 4 and the maximum number of terms is 2:

$$5 = 4+1 = 3+2. \quad (8.21)$$

The restricted partition function has a generating function:

$$\sum_{k=0}^{nm} p(n,m;k)q^k = \binom{m+n}{m}_q, \quad (8.22)$$

in which

$$\binom{m+n}{m}_q = \frac{\prod_{j=1}^{m+n}(1-q^j)}{\prod_{j=1}^{m}(1-q^j)\prod_{j=1}^{n}(1-q^j)} \tag{8.23}$$

is the Gaussian binomial coefficient [181]. It generalises the binomial coefficient as for $\lim_{q\nearrow 1}$, (8.23) reverts to the standard binomial coefficient $\binom{k+l}{l}$. As an example, we expand $p(4,2;k)$ for $k = 0,\cdots,8$:

$$\sum_{k=0}^{8} p(n,m;k)q^k = \binom{6}{2}_q = \frac{\prod_{j=1}^{6}(1-q^j)}{\prod_{j=1}^{2}(1-q^j)\prod_{j=1}^{4}(1-q^j)} = \frac{(1-q^5)(1-q^6)}{(1-q)(1-q^2)}. \tag{8.24}$$

It is straightforward to verify that (8.24) is equal to $1+q+2q^2+2q^3+3q^4+2q^5+2q^6+q^7+q^8$. In particular, we observe that $p(4,2;5) = 2$ (the factor of q^5), a result that was also demonstrated by (8.21).

8.3.5 Exact Probabilities and an Approximation

We have the following theorem on the probabilities of the random variable AUC.

Theorem 8.1. *Given m genuine and n imposter scores, all distinct, the possible values for the random variable AUC are*

$$\{\frac{k}{mn} | k \in \{0,\ldots,mn\}\}. \tag{8.25}$$

Moreover, if the genuine and imposter scores are drawn from the same score distribution, then the probability of the random variable AUC is given by

$$\mathrm{P}\left(\mathrm{AUC} = \frac{k}{mn}\right) = \frac{p(n,m;k)}{\binom{n+m}{n}}, \tag{8.26}$$

where $p(n,m;k)$ is the restricted version of the partition function.

Proof. Having m genuine and n imposter scores, this divides the TMR and FMR space into $m+1$ and $n+1$ points with distance $\frac{1}{m}$ and $\frac{1}{n}$, respectively. Since we have distinct scores, whenever the threshold increases and passes a score, the corresponding operating point in ROC space will either move to the left with a step size $\frac{1}{n}$ or down with a step size $\frac{1}{m}$. Hence, the AUC can be seen as a sum of blocks of equal area of $\frac{1}{mn}$, showing that (8.25) holds.

Given the set of ROC curves for which the number of blocks under the curve is k, we can assign to each ROC curve a sequence $k_1,k_2,\ldots,k_r$ where k_1 is the number of blocks between $TMR = 0$ and $TMR = \frac{1}{m}$, until k_r, being the number of blocks between $TMR = \frac{r-1}{m}$ and $TMR = \frac{r}{m}$. By construction, (a) $k_1+\ldots+k_r = k$, (b) the size of k_i is restricted to n, (c) r is limited to m, and (d) $k_1 \geq k_2 \geq \cdots k_r$.

The reverse relation also holds: given a sequence $k_1,k_2,\ldots,k_r$ with properties (a)-(d), we can construct the corresponding ROC curve uniquely as follows. Place k_1 blocks to the right between $TMR = 0$ and $TMR = \frac{1}{m}$, until k_r blocks to the right between $TMR = \frac{r-1}{m}$ and $TMR = \frac{r}{m}$.

The properties (a)-(d) of a sequence $k_1, k_2, \ldots, k_r$ make it a restricted partition of k. Since there is a one-to-one correspondence between a ROC curve and a restricted partition, we conclude that the number of ROC curves with $AUC = \frac{k}{mn}$ is equal to $p(n,m;k)$.

Given that the total number of ROC curves is $\binom{n+m}{n}$, all being equiprobable due to the same score distribution assumption, we conclude that (8.26) holds. □

We can also approximate (8.26) with the normal distribution.

Theorem 8.2. *For large values of m genuine scores and n imposter scores, the random variable AUC behaves like a normal random variable with mean* $\frac{mn}{2}$ *and variance*

$$\sigma_{mn} = \sqrt{\frac{mn(m+n+1)}{12}}, \tag{8.27}$$

in particular

$$\lim_{\substack{m\to\infty \\ n\to\infty}} \mathrm{P}\left(\frac{(\mathrm{AUC} - \frac{1}{2})mn}{\sigma_{mn}} \leq t \right) = \Phi(t). \tag{8.28}$$

Proof. According to Theorem 4 of [182], we have, using our notation

$$\lim_{\substack{m\to\infty \\ n\to\infty}} \mathrm{P}\left(\frac{k - \frac{1}{2}mn}{\sigma_{mn}} \leq t \right) = \Phi(t), \tag{8.29}$$

with k related to AUC as $\mathrm{AUC} = \frac{k}{mn}$. Using this relation in (8.29), we conclude that

$$\lim_{\substack{m\to\infty \\ n\to\infty}} \mathrm{P}\left(\frac{(\mathrm{AUC} - \frac{1}{2})mn}{\sigma_{mn}} \leq t \right) = \Phi(t). \tag{8.30}$$

□

8.3.6 Examples

In this section, we provide three examples of the exact probability of the random variable AUC and one application of its approximation.

The 1 genuine/n imposter case

It is straightforward to show that $p(n,1;k) = \frac{(1-q)\cdots(1-q^{n+1})}{(1-q)\cdots(1-q^n)(1-q)} = \frac{1-q^{n+1}}{1-q} = \sum_{k=0}^{n} q^k$. Hence, $p(n,1;k) = 1$ for $k = 0, \ldots, n$. Moreover, $\mathrm{p}(\mathrm{AUC} = \frac{k}{mn}) = \frac{1}{\binom{n+1}{n}} = \frac{1}{n+1}$. This is in accordance with the example discussed in the Introduction.

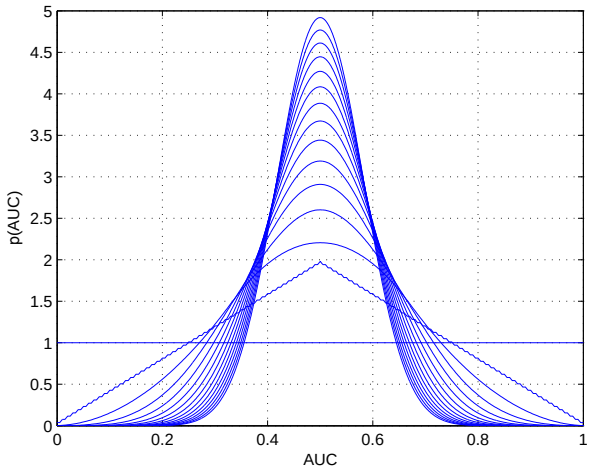


Figure 8.7: p(AUC) for $m = 1, \cdots, 15$ genuine and $n = 100$ imposter scores. Graphs are scaled such that they can be interpreted as continuous probability density functions.

The 2 genuine/Even n imposter case

Suppose n is even, then it can be shown that (8.23) can be written as

$$p(n,2;k) = (1+q+q^2+\cdots+q^n)(1+q^2+q^4+\cdots+q^n). \tag{8.31}$$

A straightforward calculation gives a staircase like shape:

$$\begin{array}{ll} p(2,n;2k) = p(2,n;2k+1) = k+1 & \text{if } 2k \leq n-1, \\ p(2,n,k) = \frac{n}{2}+1 & \text{if } k = n, \\ p(2,n;k) = p(2,n;2n-k) & \text{if } k \geq n+1. \end{array} \tag{8.32}$$

The 1-15 genuine/100 imposter case

In this example we plot P(AUC) for $m = 1, \cdots, 15$ genuine and $n = 100$ imposter scores in Figure 8.7. In particular, we see the uniform and staircase like shapes appearing for $m = 1$ and $m = 2$.

Confidence bounds

Theorem 8.2 can be used to construct a two sided $1-\alpha$ confidence interval $[\frac{1}{2} - x_\alpha, \frac{1}{2} + x_\alpha]$ around the AUC of a random classifier that depends on the number of genuine and imposter scores. Rewriting (8.28) shows that x_α is given by $x_\alpha = z_\alpha \sqrt{\frac{m+n+1}{12mn}}$, with z_α defined implicitly as $\Phi(z_\alpha) = 1 - \frac{\alpha}{2}$.

In Figure 8.8, we have chosen $m = n$, $\alpha = 5\%$ ($z_\alpha = 1.96$) and $\alpha = 1\%$ ($z_\alpha = 2.33$) and plotted the upper limit of confidence intervals as a function of the number of genuine and imposter scores. This illustrates the asymptotic behaviour of the approximation; for smaller numbers of scores, the AUC of a random system can still deviate much from AUC=0.50.

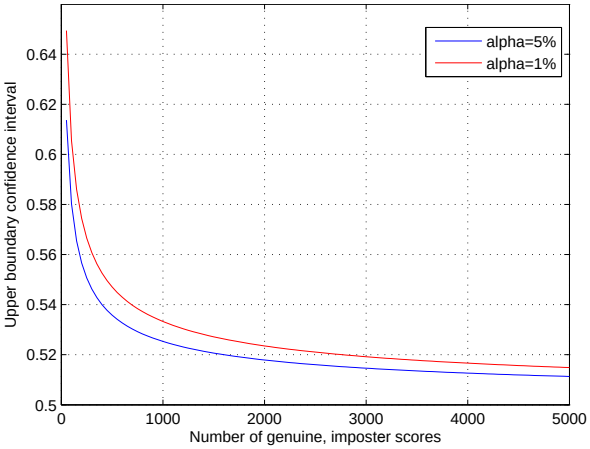


Figure 8.8: The upper limit of 95% and 99% confidence intervals as a function of equal number of genuine and imposter scores.

8.3.7 Discussion

Figure 8.7 visualises the dependency of $\mathrm{P}(AUC)$ on the number of scores. Especially, we observe that for a lower number of scores, the probability that a random system has an AUC that differs significantly from 0.50 is non-trivial. This is of relevance in, for example, the case of a subject anchored approach to evidence evaluation.

Although Theorem 8.1 provides an exact result, it can be challenging to calculate the value of the restricted partition function. One needs to resort to data structures to accommodate for values that are larger than those can fit into an IEEE-754 64 bit integer representation. This may result in an increased calculation time due to the lack of an efficient mapping from primitive operators to single machine instructions. Moreover, if we would be interested in the distribution $\mathrm{P}(\mathrm{AUC} \geq x)$, then a repeated calculation is not optimal as one could better use its generating function (8.22) for the simultaneous calculation of $p(n,m;k)$ over a range of k values.

The result of Theorem 8.2 is an approximative result, and it is instructive to see how well it approximates the true AUC probability for finite values of m and n. Figure 8.9 shows the exact probability and its approximation for three cases: $m = 5,\ 10,\ 15$, and $n = 100$. Even for moderate values of m and n the approximation seems satisfactory. Furthermore, if the number of genuine and imposter scores are equal (k) and $k \to \infty$, the probabilities become centered around AUC=0.50.

Although our work only considered approximative confidence bounds, we can also construct exact confidence bounds, especially when the number of scores is low.

8.3.8 Conclusion

In this paper, we have presented an exact expression for the probability of AUC values produced by a random classifier, given a finite number of distinct genuine and imposter scores. This work can be used in the situation when we want to determine the probability that a ran-

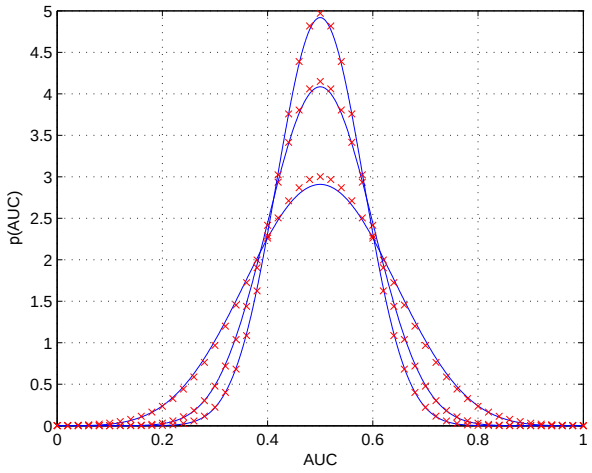


Figure 8.9: p(AUC) for $m = 5, 10, 15$ genuine and $n = 100$ imposter scores in blue, together with the approximation given by (8.28) in red. Exact graphs are scaled such that they can be interpreted as continuous probability density functions.

dom classifier produces the measured AUC; this is of interest when the measured AUC is low or the total number of scores is limited, masking the true nature of the classifier. The exact probability involves the restricted partition function and it can be approximated by a normal distribution. We used this approximation to derive confidence intervals for the AUC as a function of the number of genuine and imposter scores.

8.4 Chapter conclusion

In this chapter, two studies have been combined. They addressed four research questions. They were research question 2a: *To which extent do we observe or can we construct differences in general and subject based performance?*, research question 2b: *How well can facial marks be used for forensic evaluation, also taking subject based data and subject based evaluation into account?*, research question 2c: *In which manner can a biometric approach to FISWG characteristic descriptors be generalised into a framework for forensic evidence evaluation that also incorporates a subject based approach?*, and research question 2d: *What is a theoretical boundary between random and non-random behaviour of classifiers in a subject based performance evaluation based on AUC?*

The first included study presented the framework consisting of six aspects that influence the design of a forensically relevant feature, the biometric score and its forensic use, and the evaluation of performance characteristics. The first application involved the use of nine simple characteristic descriptors. It was shown in a subject based evaluation that there exists a large variation of discriminating power between subjects. Visual examples indicated that this variation could be explained by the phenotype of the considered characteristic descriptors. The second application was a small extension of the facial marks study presented in Chapter 7 and involved the generalisation of those results. The results indicated that, in general, facial

marks are not suitable for forensic evidence evaluation. However, it was also shown that in some cases subjects can still be discriminated.

The second study presented an exact expression for the probability of Area under the Curve (AUC) values produced by a random classifier. Since it involved the restricted partition whose calculation is very cumbersome, the study also gave an accurate approximation to them.

In light of the four research questions we conclude the following. Research question 2a is addressed by both examples of the framework and their results reiterate what was already concluded in Chapter 7: large variation in a performance measure is observable from a subject based evaluation. Research question 2b is addressed by the second application and shows that at least facial marks cannot be used for forensic evidence evaluation in general, although at a subject based evaluation level we observe that some subjects can be discriminated. Research question 2c is addressed by the presented framework that encompasses the relevant aspects to be considered during the design and evaluation of biometric classifiers for forensic evidence evaluation. The final research question 2d is addressed by providing an exact expression and an approximation to the probability of AUC values produced by a random classifier. It is possible to determine the boundary between random and non-random behaviour. Although research question 2d is placed in the context of subject based performance evaluation, the results of the second study can be applied in any performance evaluation context.

Chapter 9

Conclusion and recommendations

In this final chapter, we revisit the research questions and discuss in which manner they have been addressed by the papers bundled in this dissertation. We also present recommendations for extensions of the presented work. As indicated in Section 1.2, the two main research questions are addressed by their subordinate questions, and we will return to them in the final conclusion.

9.1 Conclusions

Research question 1a: *Under relatively well-conditioned settings, what is the performance of FFR-examiners in relation to non-examiners, both using FISWG characteristic descriptors and a best-effort approach in a verification task?*

This question has been addressed by the study contained in Chapter 3. In this study, FFR-examiners and non-examiners were asked to compare 100 eyebrow pairs using the FISWG characteristic descriptors and a "best-effort approach". The results of the study showed in essence that non-examiners should not use characteristic descriptors and that the FFR-examiners do not perform better or worse when using characteristic descriptors. The former conclusion does not have any real repercussion since non-examiners by definition do not compare, let alone use FISWG. The latter conclusion can lead to a questioning of the added value of characteristic descriptors; however, these results are confined to the eyebrow modality and have been obtained under relatively well-conditioned settings.

Research question 1b: *Under relatively well-conditioned settings, what is the general performance of biometric classifiers that use FISWG characteristic descriptors as their input and produce strength of evidence in relation to other non-forensic biometric classifiers?*

In Chapter 4, we presented two studies that addressed this research question confined to the periocular region. The results of these studies showed that the performance of the biometric classifiers that use characteristic descriptors of the eyebrow and the eye are comparable to those that used non-forensic features. Both studies also showed that the constituent parts of

the characteristic descriptors do not contribute equally well to the performance and it appears that the FISWG Feature List favours completeness over conciseness.

These results can be interpreted in several manners. In one interpretation, it shows there is merit in considering them as an alternative to other non-forensic features. However, in another interpretation, as mentioned in the previous discussion of research question 1a, it can question their added value, apart from their stated forensic relevance, over other non-forensic features. For example, the Dong Woodard feature set can in principle also be understood by a court of law and therefore possesses some forensic relevance. Also, although the used ULBP features are representative of more abstract and general features, the easiness of their automatic extraction does have its advantages over manual annotation. However, one can also argue that this comparison is not completely fair as some characteristic descriptors are difficult to extract automatically due to their definition and the reduced quality of trace images.

Our conclusion with respect to research question 1b is that classifiers using characteristic descriptors in terms of performance are mostly comparable to those that use non-forensic features under relatively well-conditioned settings, but their added value can be questioned. Also, we only considered the periocular region with respect to this research question, but we assume that this conclusion is representative of comparative studies on other characteristic descriptors. Moreover, these results have been obtained in relatively well-conditioned settings.

Research question 1c: *Under various forensic use cases, what is the general performance of biometric classifiers that use FISWG characteristic descriptors as their input and produce strength of evidence in relation to face recognition systems?*

This research question has been addressed by a study contained in Chapter 6, using the ForenFace dataset presented in Chapter 5. The results showed that in many cases, in terms of discriminating power, it was better to use a face recognition system. This goes against the desire to use features with forensic semantics instead of abstract and general features, but it is also in line with the results regarding ULBP in research question 1b. A notable exception to these results was encountered in the severest case involving trace images with 11px interpupillary distance. It was found that for example the shape of the hairline led to poor but still better discriminating power and strength of evidence than what was produced by a face recognition system. In line with the answer to research question 1b, the last result can be interpreted in several manners. A positive interpretation is that some combination of characteristic descriptors is shown to be better than the face recognition system. A negative interpretation is that these results still yield poor strength of evidence.

Research question 1d: *Under various forensic use cases, what is (a) the measurability of FISWG characteristic descriptors and (b) the influence of annotation variation on characteristic descriptors and strength of evidence produced by biometric classifiers that use these characteristic descriptors?*

As with research question 1c, this research question has been addressed by a study contained in Chapter 6 and used the ForenFace dataset presented in Chapter 5. The results of this study showed that the number of characteristic descriptors that can be measured mostly depended on the image quality and lowered in the severest cases, with some explainable ex-

ceptions. In general, these results question their usability and in particular the detailed nature of the characteristic descriptors as described by the FISWG Feature List [8] in relation to the image quality. However, one could also argue that these characteristic descriptors can be used on good quality trace images that show a partial face; the results of research question 1b indicate that other, non-forensic, features might be applicable in that situation as well.

The variability results of the second study showed the mostly negative influence of annotator variability on the variability of landmarks, shapes, and other, derived characteristic descriptors, especially when the trace image quality decreased. The influence of annotator variability on the strength of evidence generally decreased with decreasing image quality, but this cannot be seen as a positive result. In fact, it was shown that up to 70% of the evidential value intervals fully lie in the wrong region. These results reiterate the previously questioned detailed nature of the FISWG Feature List and their applicability on poor quality images.

Research question 2a: *To which extent do we observe or can we construct differences in general and subject based performance?*

This question has been addressed in Chapters 7 and 8. Although the original reason to consider this research question is the variation in discriminating power, in principle it can refer to any performance characteristic described in Meuwly et al [48].

In Chapter 7, the performance refers to discriminating power and calibration in the context of facial marks. The presented results showed that the differences between a general and a subject based evaluation are quite apparent and significant. The Tukey plots depicting discriminating power and calibration gave an impression of their variation. In the same study, it was shown that there is a proportion of subjects that can be discriminated while the general performance of the used classifier exhibited moderate performance. The same chapter presented a theoretical construction that showed that classifiers can exhibit perfect subject based performance, while the general performance is essentially random. Its aim was to complement the empirical facial mark study.

In Chapter 8, this research question has been addressed by the two included applications of the presented framework. Both applications considered discriminating power from a subject based perspective. In the first application that considers nine simple characteristic descriptors, a large variation in subject based performance was observed. This observation was reiterated by the second application that extended the facial mark study of Chapter 7. Even though the second application showed that the results regarding facial marks in Chapter 7 were not generalisable to a forensically relevant dataset, still some subjects could be discriminated perfectly. A similar, slightly, weaker result was true for the first application in which some subjects had characteristic descriptors that could be used to discriminate them to a much larger extent than the average discriminating power predicted.

Research question 2b: *How well can facial marks be used for forensic evaluation, also taking subject based data and subject based evaluation into account?*

This research question has been addressed by the study in Chapter 7 and one application included in Chapter 8. With respect to the study in Chapter 7, we found that that facial marks have a potential to be used in forensic evidence evaluation, in particular when classifiers had been trained on subject based facial mark spatial patterns. However, the calibration results

showed that the operational range of grid cell sizes should be restricted. The study in Chapter 7 was conducted on a set of good quality images; therefore this study was extended by one application of the framework of Chapter 7 that considered facial marks on traces originating from surveillance cameras. It was shown that the general discriminating power is poor, implying that, in general, facial marks do not possess the generalisation property. However, in conjunction with Chapter 7, we observed that there were still subjects that could be discriminated.

Research question 2c: *In which manner can a biometric approach to FISWG characteristic descriptors be generalised into a framework for forensic evidence evaluation that also incorporates a subject based approach?*

This research question has been addressed by another study included in Chapter 8. It did so by generalising the aspects presented in the proto-framework in the context of facial marks into a framework that is generally applicable to the design and evaluation of biometric classifiers for forensic evidence evaluation. The described aspects influenced the design of the feature, the biometric score and its forensic use, and the evaluation of performance characteristics. To emphasise its applicability, two applications were presented as well in the study.

Research question 2d: *What is a theoretical boundary between random and non-random behaviour of classifiers in a subject based performance evaluation based on AUC?*

This research question has been addressed by a study included in Chapter 8. This study provided an exact expression for the probability of Area Under the Curve (AUC) values produced by a random classifier. The probabilities used the restricted partition function, which is difficult to calculate. Therefore, also an accurate approximation in terms of the normal distribution was given. Based on this approximation, and given the number of genuine and imposter scores, it was shown how to construct a boundary between random and non-random behaviour.

9.2 Final conclusion

This dissertation has addressed two major research questions.

The first research question dealt with the suitability of FISWG characteristic descriptors as a means to discriminate, taking (a) human, (b) classifier, (c) feature, and (d) forensic aspects into account.

There does not exist a single simple answer to this research question. On one hand, one can argue that from a biometric point of view that is primarily interested in general performance, the results of (b) are both somewhat positive, but mostly negative when we consider the forensic circumstances. The former result was obtained in a limited and less representative setting, whereas the latter result was designed to be representative of various forensic use cases (d) and considered a large subset of FISWG characteristic descriptors. Also, the results of (c) showed that the characteristic descriptors are difficult to measure in representative forensic use cases (d), although almost any feature in general would be useless

under those circumstances. Results of (a) and (b) seem to indicate that there is little added value in using FISWG characteristic descriptors.

The second research question dealt with the suitability of a subject based approach in forensic evidence evaluation, taking (e) empirical results from specific applications, (f) theoretical results, and (g) a framework approach into account.

The framework (g) itself integrated the design of the feature, the biometric score and its forensic use, and the evaluation of performance characteristics, with a specific emphasis on a subject based approach. A theoretical construction (f), the results of the facial marks study (e), and the two applications (e) contained in (g) clearly showed that large differences between subject based and general performance existed. Also, the effect of using subject data in classifiers (e) was clearly seen. These results confirm the suitability of a subject based approach in forensic evidence evaluation; the presented framework can be used as a tool for such a subject based approach.

Overall, we conclude that from a general biometric perspective, FISWG characteristic descriptors are not suitable as a means to discriminate. However, if we also consider them from a biometric perspective that includes the use of subject based data and subject based performance, then in limited cases a FISWG characteristic descriptor can be used as a biometric feature to discriminate a particular subject from a group of subjects. More generally, subject based performance provides insight into the contribution of characteristic descriptors and their limits. The overall conclusion strongly points in the direction of possible future research: the creation of a large forensically relevant dataset with accompanying information that could shed more light on which, to which extent, and under which circumstances characteristic descriptors can be used in the forensic evidence evaluation process, both by FFR-examiners and classifiers.

As a final note, this dissertation has systematically considered FISWG characteristic descriptors, both directly and indirectly, starting from a human approach and zooming out to a general framework. This is a contribution to the scientific approach to forensic science encapsulated in the Daubert rule, but as mentioned in the recommendations, more research is needed to address the considered research questions to a full extent.

9.3 Recommendations for future research

Most empirical studies has some inherent limitation; the presented studies in this dissertation are no exception to this rule. The positive side of limitations is that they can serve as a recommendation for future work. Apart from these limitations, the results found in this dissertation also lead to additional recommendations. Somewhat in line with the main research questions, these recommendations have a human, a classifier and data, and a framework perspective.

9.3.1 Recommendation 1: Test the FFR-examiner performance in isolation

Studies on biometric classifiers using characteristic descriptors as their input and the characteristic descriptors themselves form the skeleton of this dissertation. The human, in its capacity of either the FFR-examiner or the non-examiner, has only been the focus in the study presented in Section 3.2. But as described in Chapter 2, the human is and will most

likely remain the main actor in FFR. In recent years, there have been some studies on how well the FFR-examiner performs in relation to non-examiners, and they seem to do a better job. However, these results pertain to the FFR-examiner in relation to non-examiners and we still do not have insight in which manner they perform their assessment. In particular, what remains unclear are (a) the influence of seeing all parts of the face at once on the outcome of the assessment and (b) which internal thresholds or strength of evidence values are used or assigned to characteristic descriptors in trace-reference image pairs. This issue is also closely tied to the central theme in the Daubert rule. Therefore, we advocate the study of human performance in trace-reference comparison tasks in which only a single facial part is shown. Such study would extend the approach presented in Section 3.2.

9.3.2 Recommendation 2: Investigate how and to which extent FISWG characteristic descriptors are (actually) used by the FFR-examiner

This recommendation is closely related to the previous recommendation. Although the question itself has not been directly the object of any study included in this dissertation, one can question to which extent the FISWG characteristic descriptors are used by FFR-examiners. A related question is even more relevant in light of Recommendation (3) of the National Research Council of the National Academies report [15] that mentions validity: how and to which extent are FISWG characteristic descriptors *actually* used by the FFR-examiner? This is a serious problem: if there is no insight whether and to which extent they are actually used in relation to their *claimed* use, the validity of assigned strength of evidence and even their raison d'être can be challenged.

9.3.3 Recommendation 3: Explore other statistical models as well

In several studies contained in this dissertation, we have used straightforward, parametric, statistical models and approached the combination of strength of evidence in the simplest manner possible: independence. One major advantage of parametric models is that the strength of evidence can be computed by a closed form formula. We recommend the exploration of semi- or non-parametric statistical approaches as well. With respect to the combination of strength of evidence, for example copula models can be used. Copula models for score fusion has been the topic of the dissertation of Susyanto [45]. Also, Bayesian Belief Networks (BBN) are an alternative method to model[1] the dependency structure between characteristic descriptors.

9.3.4 Recommendation 4: More forensic data, more forensic information

As indicated in Chapters 2 and 5, the number of facial image datasets that can be used for forensic research is limited, let alone the ones that are tailored towards forensic evidence evaluation. The study presented in Chapter 6 uses a small dataset, and its size is even tiny when viewed from a modern big data perspective[2]. Another issue is the general lack of

[1] Actually, *to manage the complexity* might be a better description.

[2] This can be considered as a modern, reversed, redemption of Gulliver's Travels.

forensically important information that enriches the data, for example annotations. The last, related issue, is the lack of a clear set of forensic parameters under which the trace images have been recorded.

We have two recommendation that could progress the field of FFR. The first recommendation is to create a large scale forensically relevant dataset, with attached information and produced under a clear set of forensic parameters. Its disadvantage is that it requires a significant amount of effort to create such dataset and above all, it is static. Therefore, the second recommendation is to create a software package in which all forensic parameters can be described and which is capable to produce realistic trace images based on reference models. This would also require a major investment, but its significant advantage over the first recommendation is that it creates much more insight into the influence of all forensic parameters on the trace image. Also, datasets can be constructed "on-the-spot", which could be beneficial for real forensic case work.

9.3.5 Recommendation 5: Incorporate subject based evaluation in method validation

An insight that became apparent when the results on general discriminating power presented in Chapter 6 yielded overall moderate to poor results is that one does not need to discriminate everyone from everyone, as long as one can discriminate the suspect from the other subjects. This insight is certainly not new, but in our opinion it does have an implication for method validation as for example described by the recently published guideline for the validation of likelihood ratio methods used for forensic evidence evaluation [48]. This guideline describes performance characteristics that should be taken in account in order to validate a method that is capable of producing strength of evidence. One notable performance characteristic is discriminating power which can measured in for example the EER. This guideline does not describe at which level the evaluation should be performed. We believe that at least insight into the variation of a performance characteristic viewed from a subject based perspective leads to insight into the proportion of cases the method could be of value and more generally, its limits of usability. We recommend the explicit mentioning of such subject based evaluation in a validation process.

Bibliography

[1] D. Meuwly and R. Veldhuis. Forensic biometrics: From two communities to one discipline. In *2012 BIOSIG - Proceedings of the International Conference of Biometrics Special Interest Group (BIOSIG)*, pages 1–12, Sept 2012.

[2] Face Recognition Vendor Test (FRVT) - Performance of Face Identification Algorithms, NIST Interagency Report 8009. `http://biometrics.nist.gov/cs_links/face/frvt/frvt2013/NIST_8009.pdf`. Accessed: 2016-04-07.

[3] Kristin Norell, Klas Brorsson Läthén, Peter Bergström, Allyson Rice, Vaidehi Natu, and Alice O'Toole. The effect of image quality and forensic expertise in facial image comparisons. *Journal of Forensic Sciences*, 60(2):331–340, 2015.

[4] David G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *Int. J. Comput. Vision*, 60(2):91–110, November 2004.

[5] Timo Ahonen, Abdenour Hadid, and Matti Pietikainen. Face Description with Local Binary Patterns: Application to Face Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(12):2037–2041, December 2006.

[6] FISWG Guidelines for Facial Comparison Methods. `https://fiswg.org/FISWG_GuidelinesforFacialComparisonMethods_v1.0_2012_02_02.pdf`. Accessed: 2017-01-09.

[7] FISWG website. `https://fiswg.org`. Accessed: 2014-04-22.

[8] FISWG Facial Image Comparison Feature List for Morphological Analysis. `https://fiswg.org/FISWG_1to1_Checklist_v1.0_2013_11_22.pdf`. Accessed: 2017-01-09.

[9] Facial Features to look for based on experience in forensic case work. Private Communication with forensic facial examiner at NFI. Email: 2015-12-09.

[10] Facial Comparison List used at NFI Netherlands. Private Communication with forensic facial examiner at NFI. Email: 2015-12-09.

[11] Facial Comparison List used at NFC Sweden. Private Communication with forensic facial examiner at NFC. Email: 2015-09-06.

[12] Jason P. Prince. To examine emerging police use of facial recognition systems and facial image comparison procedures. `www.churchilltrust.com.au/media/fellows/2012_Prince_Jason.pdf`, 2012. Accessed: 2014-04-22.

[13] Glenn M. Langenburg. *A Critical Analysis and Study of ACE-V Process*. PhD thesis, Université de Lausanne, 2008.

[14] Gary Edmond, Katherine Biber, Richard I. Kemp, and Glenn Porter. Law's looking glass: expert identification evidence derived from photographic and video images. *Current Issues in Criminal Justice*, 20(3), 2009.

[15] National Research Council. *Strengthening Forensic Science in the United States: A Path Forward*.

[16] C. G. Zeinstra, D. Meuwly, A. C. C. Ruifrok, R. N. J. Veldhuis, and L. J. Spreeuwers. Forensic Face Recognition as a means to determine Strength of Evidence: a survey. *Submitted to Forensic Science Review*.

[17] C. G. Zeinstra, R. N. J. Veldhuis, and L. J. Spreeuwers. Examining the examiners: an online eyebrow verification experiment inspired by FISWG. In *International Workshop on Biometrics and Forensics, IWBF 2015 , Glövik, Norway*, pages 1–6, USA, March 2015. IEEE Computer Society.

[18] C. G. Zeinstra, R. N. J. Veldhuis, and L. J. Spreeuwers. Towards the automation of forensic facial individualisation: Comparing forensic to non-forensic eyebrow features. In *Proceedings of the 35th WIC Symposium on Information Theory in the Benelux, Eindhoven, Netherlands*, pages 73–80, Enschede, May 2014. Centre for Telematics and Information Technology, University of Twente.

[19] Yujie Dong and Damon L. Woodard. Eyebrow shape-based features for biometric recognition and gender classification: A feasibility study. In *IJCB'11*, pages 1–8, 2011.

[20] C. G. Zeinstra, R. N. J. Veldhuis, and L. J. Spreeuwers. Beyond the eye of the beholder: on a forensic descriptor of the eye region. In *23rd European Signal Processing Conference, EUSIPCO 2015, Nice*, pages 779–783. IEEE Signal Processing Society, September 2015.

[21] Chris G. Zeinstra, Raymond N.J. Veldhuis, Luuk J. Spreeuwers, Arnout C.C. Ruifrok, and Didier Meuwly. Forenface: a unique annotated forensic facial image dataset and toolset. *IET Biometrics*, May 2017. http://digital-library.theiet.org/content/journals/10.1049/iet-bmt.2016.0160.

[22] C. G. Zeinstra, R. N. J. Veldhuis, and L. J. Spreeuwers. Discriminating power of FISWG characteristic descriptors under different forensic use cases. In *BIOSIG 2016 - Proceedings of the 15th International Conference of the Biometrics Special Interest Group, 21.-23. September 2016, Darmstadt, Germany*, volume 260 of *LNI*, pages 171–182. GI, 2016.

[23] Chris Zeinstra, Raymond Veldhuis, Luuk Spreeuwers, and Arnout Ruifrok. Manually annotated characteristic descriptors: measurability and variability. In *International Workshop on Biometrics and Forensics, IWBF 2017, Conventry, United Kingdom*.

[24] Chris Zeinstra, Raymond Veldhuis, and Luuk Spreeuwers. Grid Based Likelihood Ratio Classifiers for the Comparison of Facial Marks. *Accepted for publication in IEEE Transactions on Information Forensics and Security*, 2017. http://dx.doi.org/10.1109/TIFS.2017.2746013.

[25] Chris Zeinstra, Raymond Veldhuis, and Luuk Spreeuwers. Label specific versus general classifier performance: an extreme example. *University of Twente Students Journal of Biometrics and Computer Vision*. http://dx.doi.org/10.3990/3.utsjbcv.i0.25.

[26] Chris Zeinstra, Raymond Veldhuis, Luuk Spreeuwers, and Didier Meuwly. Mind the Gap: A Practical Framework regarding Classifiers for Forensic Evidence Evaluation. *Submitted to Science & Justice*.

[27] Chris Zeinstra, Raymond Veldhuis and Luuk Spreeuwers. How Random is a Classifier given its Area under Curve? *Accepted for publication in BIOSIG 2017*.

[28] Aad Dijksma, Heinz Langer, Yuri Shondin, and Chris Zeinstra. Self-adjoint operators with inner singularities and Pontryagin spaces. In *Operator Theory and Related Topics*, pages 105–175. Springer, 2000.

[29] M. A. Kaashoek and C. G. Zeinstra. The band method and generalized Carathéodory-Toeplitz interpolation at operator points. *Integral Equations and Operator Theory*, 33(2):175–210, 1999.

[30] Anil K. Jain, Patrick Flynn, and Arun A. Ross. *Handbook of Biometrics*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2007.

[31] Fingerprint detail on male finger. `https://commons.wikimedia.org/w/index.php?curid=6158485`. Accessed: 2017-05-06.

[32] Human Iris. `https://commons.wikimedia.org/w/index.php?curid=3117810`. Accessed: 2017-05-06.

[33] FRGC website. `http://www.nist.gov/itl/iad/ig/frgc.cfm`. Accessed: 2014-04-22.

[34] DNA Overview. `https://commons.wikimedia.org/w/index.php?curid=694302`. Accessed: 2017-05-06.

[35] Ministry of Silly Walks. `https://en.wikipedia.org/w/index.php?curid=2757933`. Accessed: 2017-05-06.

[36] Dutch traffic sign C2. `https://commons.wikimedia.org/w/index.php?curid=2515101`. Accessed: 2017-05-06.

[37] Dutch traffic sign C3. `https://commons.wikimedia.org/w/index.php?curid=2515189`. Accessed: 2017-05-06.

[38] Minutiae in a fingerprint. http://biometrics.mainguet.org/types/fingerprint/fingerprint_algo.htm. Accessed: 2017-05-06.

[39] Pictorial Example of IrisCode. http://www.cl.cam.ac.uk/~jgd1000/examples.html. Accessed: 2017-05-06.

[40] Unique Identification Authority of India Aadhaar. https://uidai.gov.in/new/. Accessed: 2017-02-14.

[41] Arun A. Ross, Karthik Nandakumar, and Anil K. Jain. *Handbook of Multibiometrics (International Series on Biometrics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[42] David J. Hand and Robert J. Till. A simple generalisation of the area under the ROC curve for multiple class classification problems. *Machine Learning*, 45(2):171–186, 2001.

[43] Affaire Dreyfus, Rapport de Mr. les Experts Darboux, Appell, Poincaré. http://www.maths.ed.ac.uk/~aar/dreyfus/dreyfustyped.pdf. Accessed: 2016-12-12.

[44] G. Jackson, S. Jones, G. Booth, C. Champod, and I.W. Evett. The nature of forensic science opinion - a possible framework to guide thinking and practice in investigation and in court proceedings. *Science & Justice*, 46(1):33–44, 2006.

[45] Nanang Susyanto. *Semiparametric Copula Models for Biometric Score Level Fusion*. PhD thesis, University of Amsterdam, 2016.

[46] Emanuel Parzen. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076, 1962.

[47] Tom Fawcett and Alexandru Niculescu-Mizil. PAV and the ROC convex hull. *Machine Learning*, 68(1):97–106, 2007.

[48] Didier Meuwly, Daniel Ramos, and Rudolf Haraksim. A guideline for the validation of likelihood ratio methods used for forensic evidence evaluation. *Forensic Science International*, 2016. https://doi.org/10.1016/j.forsciint.2016.03.048.

[49] Niko Brümmer and Johan du Preez. Application-independent evaluation of speaker detection. *Computer Speech & Language*, 20(23):230–275, 2006.

[50] A. K. Jain, B. Klare, and U. Park. Face Matching and Retrieval in Forensics Applications. *IEEE MultiMedia*, 19(1):20–20, Jan 2012.

[51] Richard Russell, Brad Duchaine, and Ken Nakayama. Super-recognizers: People with extraordinary face recognition ability. *Psychon Bull Rev*, 16(2):252–257, Apr 2009.

[52] Are you a super recognizer? http://www.dailymail.co.uk/sciencetech/article-3125173. Accessed: 2017-01-03.

[53] A. Bertillon. *Identification anthropométrique: instructions signalétiques*. 1893.

[54] Jonathan Finn. *Capturing the Criminal Image: From Mug Shot to Surveillance Society*. University of Minnesota Press, New edition, 2009.

[55] Martin Paul Evison. *Forensic Facial Analysis*, pages 1713–1729. Springer New York, New York, NY, 2014.

[56] Meuwly, Didier. Le Mythe de l'Empreinte Vocale (I et II). *Revue internationale de criminologie et de police technique et scientifique*, 56(2):219–236, 2003.

[57] N. A. Spaun. Forensic Biometrics from Images and Video at the Federal Bureau of Investigation. In *Biometrics: Theory, Applications, and Systems, 2007. BTAS 2007. First IEEE International Conference on*, pages 1–3, Sept 2007.

[58] A. Sklar. *Fonctions de répartition à n dimensions et leurs marges*. 1959.

[59] D. Barber. *Bayesian Reasoning and Machine Learning*. Cambridge University Press, 2012.

[60] Guidance for evaluating levels of support. FIAG: Forensic Imagery Analysis Group. `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1550752`. Accessed: 2017-01-03.

[61] ENFSI Guideline for Evaluative Reporting in Forensic Science. `http://enfsi.eu/sites/default/files/documents/external_publications/m1_guideline.pdf`. Accessed: 2016-07-21.

[62] ENFSI website. `http://http://enfsi.eu/`. Accessed: 2017-01-05.

[63] Image Processing to Improve Automated Facial Recognition Search Performance. `https://fiswg.org/DRAFT_FISWG_ImageProcessingtoImproveFRSearchPerf_v1.0_2016_07_26.pdf`. Accessed: 2017-01-09.

[64] Physical Stability of Facial Features of Adults. `https://fiswg.org/DRAFT_FISWG_Physical_Stability_of_Facial_Components_v1.0_20160202.pdf`. Accessed: 2017-01-09.

[65] Nederlands Register Gerechtelijk Deskundigen. `https://www.nrgd.nl/`. Accessed: 2017-01-08.

[66] Guidelines and Recommendations for Facial Comparison Training to Competency. `https://fiswg.org/FISWG_Training_Guidelines_Recommendations_v1.1_2010_11_18.pdf`. Accessed: 2017-01-09.

[67] Standard for Facial Identification and Facial Recognition Proficiency Testing Programs. `https://fiswg.org/DRAFT_FISWG_Proficiency_Test_Program_Standard20140509.pdf`. Accessed: 2017-01-09.

[68] Xanthé Mallett and Martin P. Evison. Forensic facial comparison: issues of admissibility in the development of novel analytical technique. *Journal of Forensic Sciences*, 58(4):859–865, 2013.

[69] Saks, Michael J. and Koehler, Jonathan J. The coming paradigm shift in forensic identification science. *Science*, 309(5736):892–895, 2005.

[70] Alice J. O'Toole, Fang Jiang, Dana Roark, and Hervé Abdi. Predicting human performance for face recognition. *Face Processing: Advanced Methods and Models. Elsevier, Amsterdam*, 2006.

[71] Nicole A. Spaun. Facial comparisons by subject matter experts: Their role in biometrics and their training. In *International Conference on Biometrics*, pages 161–168. Springer, 2009.

[72] Megan H. Papesh and Stephen D. Goldinger. Infrequent identity mismatches are frequently undetected. *Attention, Perception, & Psychophysics*, 76(5):1335–1349, 2014.

[73] P. Sinha, B. Balas, Y. Ostrovsky, and R. Russell. Face Recognition by Humans: Nineteen Results All Computer Vision Researchers Should Know About. *Proceedings of the IEEE*, 94(11):1948–1962, Nov 2006.

[74] A. Mike Burton, Stephen Wilson, Michelle Cowan, and Vicki Bruce. Face recognition in poor-quality video: Evidence from security surveillance. *Psychological Science*, 10(3):243–248, 1999.

[75] Vicki Bruce, Zoë Henderson, Karen Greenwood, Peter J. B. Hancock, A. Mike Burton, and Paul Miller. Verification of face identities from images captured on video. *Journal of Experimental Psychology: Applied*, 5(4):339, 1999.

[76] Vicki Bruce, Zoë Henderson, Craig Newman, and A. Mike Burton. Matching identities of familiar and unfamiliar faces caught on CCTV images. *Journal of Experimental Psychology: Applied*, 7(3):207, 2001.

[77] Ahmed M. Megreya and A. Mike Burton. Unfamiliar faces are not faces: Evidence from a matching task. *Memory & Cognition*, 34(4):865–876, 2006.

[78] Jason M. Gold, Jarrett D. Barker, Shawn Barr, Jennifer L. Bittner, Alexander Bratch, W. Drew Bromfield, Roy A. Goode, Mary Jones, Doori Lee, and Aparna Srinath. The perception of a familiar face is no more than the sum of its parts. *Psychonomic Bulletin & Review*, 21(6):1465–1472, 2014.

[79] Emily Pronin. How we see ourselves and how we see others. *Science*, 320(5880):1177–1180, 2008.

[80] Itiel E. Dror, David Charlton, and Ailsa E. Péron. Contextual information renders experts vulnerable to making erroneous identifications. *Forensic Science International*, 156(1):74–78, 2006.

[81] David White, P. Jonathon Phillips, Carina A. Hahn, Matthew Hill, and Alice J. O'Toole. Perceptual expertise in forensic facial image comparison. In *Proc. R. Soc. B*, volume 282, pages 2015–1292. The Royal Society, 2015.

[82] Krista F. Kleinberg. *Facial anthropometry as an evidential tool in forensic image comparison.* PhD thesis, University of Glasgow, 2008.

[83] Martin Evison and Richard Vorder Bruegge. *Computer-aided forensic facial comparison*. Taylor and Francis Group, Boca Raton, Florida, USA, March 2010.

[84] Josh P. Davis, Tim Valentine, and Robert E. Davis. Computer assisted photo-anthropometric analyses of full-face and profile facial images. *Forensic Science International*, 200(13):165–176, 2010.

[85] M. M. Roelofse, M. Steyn, and P. J. Becker. Photo identification: Facial metrical and morphological features in South African males. *Forensic Science International*, 177(23):168–175, 2008.

[86] S. Ritz-Timme, P. Gabriel, J. Tutkuviene, P. Poppa, Z. Obertova, D. Gibelli, D. De Angelis, M. Ratnayake, R. Rizgeliene, A. Barkus, and C. Cattaneo. Metric and morphological assessment of facial features: A study on three European populations . *Forensic Science International*, 207(13):239–239, 2011.

[87] Matthew Turk and Alex Pentland. Eigenfaces for Recognition. *J. Cognitive Neuroscience*, 3(1):71–86, January 1991.

[88] O. Déniz, G. Bueno, J. Salido, and F. De la Torre. Face Recognition Using Histograms of Oriented Gradients. *Pattern Recogn. Lett.*, 32(12):1598–1603, September 2011.

[89] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[90] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. DeepFace: Closing the Gap to Human-Level Performance in Face Verification. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[91] T. Ali. *Biometric Score Calibration for Forensic Face Recognition*. PhD thesis, University of Twente, Enschede, June 2014.

[92] Mislav Grgic, Kresimir Delac, and Sonja Grgic. SCFace - surveillance cameras face database. *Multimedia Tools and Applications*, 51(3):863–879, 2011.

[93] Yongkang Wong, Shaokang Chen, Sandra Mau, Conrad Sanderson, and Brian C. Lovell. Patch-based Probabilistic Image Quality Assessment for Face Selection and Improved Video-based Face Recognition. In *IEEE Biometrics Workshop, Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 81–88. IEEE, June 2011.

[94] NIST Mugshot Identification Database. `http://www.nist.gov/srd/nistsd18.cfm`. Accessed: 2016-04-25.

[95] Karl Ricanek Jr. and Tamirat Tesafaye. MORPH: A Longitudinal Image Database of Normal Adult Age-Progression. In *Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition*, FGR '06, pages 341–345, Washington, DC, USA, 2006. IEEE Computer Society.

[96] R. Vera-Rodriguez, P. Tome, J. Fierrez, N. Expsito, and F. J. Vega. Analysis of the variability of facial landmarks in a forensic scenario. In *Biometrics and Forensics (IWBF), 2013 International Workshop on*, pages 1–4, April 2013.

[97] João C. Neves, Gil Santos, Sílvio Filipe, Emanuel Grancho, Silvio Barra, Fabio Narducci, and Hugo Proença. *Quis-Campi: Extending in the Wild Biometric Recognition to Surveillance Environments*, pages 59–68. Springer International Publishing, Cham, 2015.

[98] ForenFace website. `http://scs.ewi.utwente.nl/downloads/show,ForenFace/`. Accessed: 2016-06-08.

[99] ICB-RW International Challenge on Biometric Recognition in the Wild. `http://icbrw.di.ubi.pt/`. Accessed: 2016-08-23.

[100] Pedro Tome, Julian Fierrez, Ruben Vera-Rodriguez, and Daniel Ramos. Identification using face regions: Application and assessment in forensic scenarios. *Forensic Science International*, 233(13):75–83, 2013.

[101] U. Park and A. K. Jain. Face Matching and Retrieval Using Soft Biometrics. *IEEE Transactions on Information Forensics and Security*, 5(3):406–415, Sept 2010.

[102] Nisha Srinivas, Patrick J. Flynn, and Richard W. Vorder Bruegge. Human Identification Using Automatic and Semi-Automatically Detected Facial Marks. *Journal of Forensic Sciences*, 61:117–130, 2016.

[103] J. E. Lee, A. K. Jain, and R. Jin. Scars, marks and tattoos (SMT): soft biometric for suspect and victim identification. In *BYSYM*, pages 1–8, 2008.

[104] Pedro Tome, Ruben Vera-Rodriguez, Julian Fierrez, and Javier Ortega-Garcia. Facial soft biometric features for forensic face recognition. *Forensic Science International*, 257:271–284, 2015.

[105] Rudolf Haraksim, Daniel Ramos, Didier Meuwly, and Charles E. H. Berger. Measuring coherence of computer-assisted likelihood ratio methods. *Forensic Science International*, 249:123–132, 2015.

[106] V. Kazemi and J. Sullivan. One millisecond face alignment with an ensemble of regression trees. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1867–1874, June 2014.

[107] S. Milborrow and F. Nicolls. Active Shape Models with SIFT Descriptors and MARS. *VISAPP*, 2014.

[108] J. Sadr, I. Jarudi, and P. Sinha. The role of eyebrows in face recognition. *Perception*, 32:285–293, 2003.

[109] F. Juefei-Xu and M. Savvides. Can your eyebrows tell me who you are? In *Signal Processing and Communication Systems (ICSPCS), 2011 5th International Conference on*, pages 1–8, Dec 2011.

[110] Qinran Chen, Wai-kuen Cham, and Kar-kin Lee. Extracting Eyebrow Contour and Chin Contour for Face Recognition. *Pattern Recogn.*, 40(8):2292–2300, August 2007.

[111] Yujian Li, Houjun Li, and Zhi Cai. Human eyebrow recognition in the matching-recognizing framework. *Computer Vision and Image Understanding*, 117(2):170–181, 2013.

[112] Unsang Park, R. Jillela, A. Ross, and A. K. Jain. Periocular Biometrics in the Visible Spectrum. *Information Forensics and Security, IEEE Transactions on*, 6(1):96–106, March 2011.

[113] Gayathri Mahalingam and Karl Ricanek. LBP-based periocular recognition on challenging face datasets. *EURASIP Journal on Image and Video Processing*, 2013(1):36, 2013.

[114] K. Hollingsworth, K. W. Bowyer, and P. J. Flynn. Identifying useful features for recognition in near-infrared periocular images. In *Biometrics: Theory Applications and Systems (BTAS), 2010 Fourth IEEE International Conference on*, pages 1–8, Sept 2010.

[115] K. P. Hollingsworth, S. S. Darnell, P. E. Miller, D. L. Woodard, K. W. Bowyer, and P. J. Flynn. Human and Machine Performance on Periocular Biometrics Under Near-Infrared Light and Visible Light. *Information Forensics and Security, IEEE Transactions on*, 7(2):588–601, April 2012.

[116] S. L. Phung, A. Bouzerdoum, and Sr. Chai, D. Skin segmentation using color pixel classification: analysis and comparison. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(1):148–154, Jan 2005.

[117] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[118] FISWG Facial Image Comparison Feature List for Morphological Analysis - Draft version. `https://www.fiswg.org/document/viewDocument?id=29`. Accessed: 2014-04-22.

[119] Yujian Li and Xingli Li. HMM Based Eyebrow Recognition. In *Proceedings of the Third International Conference on International Information Hiding and Multimedia Signal Processing (IIH-MSP 2007) - Volume 01*, IIH-MSP '07, pages 135–138, Washington, DC, USA, 2007. IEEE Computer Society.

[120] Xu Xiaojun, Yang Xinwu, Li Yujian, and Yang Yuewei. Eyebrow recognition using Radon transform and sparsity preserving projections. In *Automatic Control and Artificial Intelligence (ACAI 2012), International Conference on*, pages 1028–1033, March 2012.

[121] The BJUT Eyebrow Database. `http://mpccl.bjut.edu.cn/EyebrowRecognition/BJUTEyebrowDatabase/BJUTED.html`.

[122] S. Conseil, S. Bourennane, and L. Martin. Comparison of Fourier Descriptors and Hu Moments for Hand Posture Recognition. In *European Signal Processing Conference (EUSIPCO)*, 2007.

[123] Raymond N. Veldhuis, Asker M. Bazen, Wim D. Booij, and Anne J. Hendrikse. Hand-geometry recognition based on contour parameters. *Proc. SPIE*, 5779:344–353, 2005.

[124] Bianca Zadrozny and Charles Elkan. Transforming Classifier Scores into Accurate Multiclass Probability Estimates, 2002.

[125] PUT Face Database Description. `https://biometrics.cie.put.poznan.pl/index.php`. Accessed: 2014-04-22.

[126] T. Moriyama, T. Kanade, Jing Xiao, and J. F. Cohn. Meticulously detailed eye region model and its application to analysis of facial images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(5):738–752, May 2006.

[127] Philip E. Miller, Allen W. Rawls, Shrinivas J. Pundlik, and Damon L. Woodard. Personal Identification Using Periocular Skin Texture. In *Proceedings of the 2010 ACM Symposium on Applied Computing*, SAC '10, pages 1496–1500, New York, NY, USA, 2010. ACM.

[128] P. E. Miller, J. R. Lyle, S. J. Pundlik, and D. L. Woodard. Performance evaluation of local appearance based periocular recognition. In *Biometrics: Theory Applications and Systems (BTAS), 2010 Fourth IEEE International Conference on*, pages 1–6, Sept 2010.

[129] Juefei Xu, M. Cha, J. L. Heyman, S. Venugopalan, R. Abiantun, and M. Savvides. Robust local binary pattern feature sets for periocular biometric identification. In *Biometrics: Theory Applications and Systems (BTAS), 2010 Fourth IEEE International Conference on*, pages 1–8, Sept 2010.

[130] S. Bharadwaj, H. S. Bhatt, M. Vatsa, and R. Singh. Periocular biometrics: When iris recognition fails. In *Biometrics: Theory Applications and Systems (BTAS), 2010 Fourth IEEE International Conference on*, pages 1–6, Sept 2010.

[131] P. J. Phillips, P. J. Flynn, K. W. Bowyer, R. W. V. Bruegge, P. J. Grother, G. W. Quinn, and M. Pruitt. Distinguishing identical twins by face recognition. In *Automatic Face Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 185–192, March 2011.

[132] Walter Gander, Gene H. Golub, and Rolf Strebel. Least-squares fitting of circles and ellipses. *BIT Numerical Mathematics*, 34(4):558–578, 1994.

[133] Takeo Kanade. Picture Processing System by Computer Complex and Recognition of Human Faces. In *Doctoral Dissertation, Kyoto University*. November 1973.

[134] P. Jonathon Phillips, Hyeonjoon Moon, Syed A. Rizvi, and Patrick J. Rauss. The FERET Evaluation Methodology for Face-Recognition Algorithms. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(10):1090–1104, October 2000.

[135] P. J. Phillips, W. T. Scruggs, A. J. O'Toole, P. J. Flynn, K. W. Bowyer, C. L. Schott, and M. Sharpe. FRVT 2006 and ICE 2006 Large-Scale Experimental Results. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(5):831–846, May 2010.

[136] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[137] Vuong Le, Jonathan Brandt, Zhe Lin, Lubomir Bourdev, and Thomas S. Huang. Interactive Facial Feature Localization. In Andrew Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid, editors, *Computer Vision ECCV 2012*, volume 7574 of *Lecture Notes in Computer Science*, pages 679–692. Springer Berlin Heidelberg, 2012.

[138] Dallmeier website. `http://www.dallmeier.com`. Accessed: 2016-01-04.

[139] MeshLab website. `http://meshlab.sourceforge.net`. Accessed: 2016-01-08.

[140] Richard H. Bartels, John C. Beatty, and Brian A. Barsky. *An Introduction to Splines for Use in Computer Graphics and Geometric Modeling*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1987.

[141] Neurotechnology Verilook website. `http://www.neurotechnology.com/verilook.html`. Accessed: 2016-05-10.

[142] Cognitec FaceVACS website. `http://www.cognitec.com/products.html`. Accessed: 2015-12-29.

[143] A. K. Jain, B. Klare, and U. Park. Face recognition: Some challenges in forensics. In *Automatic Face Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 726–733, March 2011.

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

[145] Daniel Ramos and Joaquin Gonzalez-Rodriguez. Reliable support: Measuring calibration of likelihood ratios. *Forensic Science International*, 230(1):156–169, May 2013.

[146] Brendan Klare and Anil K. Jain. On a taxonomy of facial features. In *Biometrics: Theory Applications and Systems (BTAS), 2010 Fourth IEEE International Conference on*, pages 1–8. IEEE, 2010.

[147] Dahua Lin and Xiaoou Tang. Recognize High Resolution Faces: From Macrocosm to Microcosm. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1355–1362, 2006.

[148] N. Srinivas, G. Aggarwal, P. J. Flynn, and R. W. Vorder Bruegge. Analysis of Facial Marks to Distinguish Between Identical Twins. *IEEE Transactions on Information Forensics and Security*, 7(5):1536–1550, Oct 2012.

[149] N. Srinivas, G. Aggarwal, P. J. Flynn, and R. W. V. Bruegge. Facial marks as biometric signatures to distinguish between identical twins. In *CVPR 2011 WORKSHOPS*, pages 106–113, June 2011.

[150] Eliabeth S. Shalin, B. Thomas, J. J. Kizhakkethottam, and J. J. Kizhakkethottam. Analysis of effective biometric identification on monozygotic twins. In *2015 International Conference on Soft-Computing and Networks Security (ICSNS)*, pages 1–6, Feb 2015.

[151] S. Biswas, K. W. Bowyer, and P. J. Flynn. A study of face recognition of identical twins by humans. In *2011 IEEE International Workshop on Information Forensics and Security*, pages 1–6, Nov 2011.

[152] Timothy F. Cootes, Gareth J. Edwards, and Christopher J. Taylor. Active Appearance Models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(6):681–685, June 2001.

[153] G. Loy and A. Zelinsky. Fast radial symmetry for detecting points of interest. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(8):959–973, Aug 2003.

[154] J. S. Pierrard and T. Vetter. Skin Detail Analysis for Face Recognition. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, June 2007.

[155] Arfika Nurhudatiana, Adams Wai-Kin Kong, Keyan Matinpour, Deborah Chon, Lisa Altieri, Siu-Yeung Cho, and Noah Craft. The individuality of relatively permanent pigmented or vascular skin marks (RPPVSM) in independently and uniformly distributed patterns. *IEEE Transactions on Information Forensics and Security*, 8(6):998–1012, 2013.

[156] Arfika Nurhudatiana and Adams Wai-Kin Kong. On criminal identification in color skin images using skin marks (RPPVSM) and fusion with inferred vein patterns. *IEEE Transactions on Information Forensics and Security*, 10(5):916–931, 2015.

[157] Arfika Nurhudatiana, Adams Wai-Kin Kong, Noah Craft, and Hong Liang Tey. Relatively Permanent Pigmented or Vascular Skin Marks for Identification: A Pilot Reliability Study. *Journal of Forensic Sciences*, 61(1):52–58, 2016.

[158] Alvin F. Martin, George R. Doddington, Terri Kamm, Mark Ordowski, and Mark A. Przybocki. The DET curve in assessment of detection task performance. In George Kokkinakis, Nikos Fakotakis, and Evangelos Dermatas, editors, *EUROSPEECH*. ISCA, 1997.

[159] Daniel Ramos, Joaquin Gonzalez-Rodriguez, Grzegorz Zadora, and Colin Aitken. Information-Theoretical Assessment of the Performance of Likelihood Ratio Computation Methods. *Journal of Forensic Sciences*, 58(6):1503–1518, 2013.

[160] John W. Tukey. Exploratory Data Analysis. 1977.

[161] Validation Toolbox. `https://sites.google.com/site/validationtoolbox/`. Accessed: 2016-12-03.

[162] Davide Maltoni, Dario Maio, Anil Jain, and Salil Prabhakar. *Handbook of Fingerprint Recognition*. Springer Science & Business Media, 2009.

[163] D. Meuwly. Forensic individualisation from biometric data. *Science & Justice*, 46(4):205–213, 2006.

[164] A. Moorhouse, Adrian Evans, G. A. Atkinson, J. Sun, and M. L. Smith. The nose on your face may not be so plain: Using the nose as a biometric. In *3rd International Conference on Imaging for Crime Detection and Prevention, ICDP 2009*. Institution of Engineering and Technology, December 2009.

[165] Michał Choraś. The lip as a biometric. *Pattern Analysis and Applications*, 13(1):105–112, 2010.

[166] Tom Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006.

[167] James A. Hanley and Barbara J. McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36, 1982.

[168] Henry B. Mann and Donald R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, pages 50–60, 1947.

[169] Simon J. Mason and Nicholas E. Graham. Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation. *Quarterly Journal of the Royal Meteorological Society*, 128(584):2145–2166, 2002.

[170] Corinna Cortes and Mehryar Mohri. Confidence intervals for the area under the ROC curve. In *Nips*, pages 305–312, 2004.

[171] Hsin-Neng Hsieh, Hsiu-Yuan Su, and Xiao-Hua Zhou. Interval estimation for the difference in paired areas under the ROC curves in the absence of a gold standard test. *Statistics in medicine*, 28(25):3108–3123, 2009.

[172] Charles E. Metz, Benjamin A. Herman, and Jong-Her Shen. Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data. *Statistics in medicine*, 17(9):1033–1053, 1998.

[173] Howell Tong. On the estimation of $\Pr\{Y < X\}$ for exponential families. *IEEE Transactions on Reliability*, 26(1):54–56, 1977.

[174] T. Pham and J. Almhana. The generalized gamma distribution: its hazard rate and stress-strength model. *IEEE Transactions on Reliability*, 44(3):392–397, 1995.

[175] Elizabeth R. DeLong, David M. DeLong, and Daniel L. Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, pages 837–845, 1988.

[176] Gengsheng Qin and Lejla Hotilovac. Comparison of non-parametric confidence intervals for the area under the ROC curve of a continuous-scale diagnostic test. *Statistical Methods in Medical Research*, 17(2):207–221, 2008.

[177] Gengsheng Qin and Xiao-Hua Zhou. Empirical likelihood inference for the area under the ROC curve. *Biometrics*, 62(2):613–622, 2006.

[178] Nancy A. Obuchowski and Michael L. Lieber. Confidence intervals for the receiver operating characteristic area in studies with small samples. *Academic Radiology*, 5(8):561–571, 1998.

[179] Blaise Hanczar, Jianping Hua, Chao Sima, John Weinstein, Michael Bittner, and Edward R. Dougherty. Small-sample precision of ROC-related estimates. *Bioinformatics*, 26(6):822–830, 2010.

[180] G. E. Andrews. *The Theory of Partitions*. Cambridge Mathematical Library. Cambridge University Press, 1998.

[181] George E. Andrews. Applications of basic hypergeometric functions. *SIAM review*, 16(4):441–484, 1974.

[182] Lajos Takács. Some asymptotic formulas for lattice paths. *Journal of Statistical Planning and Inference*, 14(1):123–142, 1986.

Summary

In this dissertation, FISWG characteristic descriptors are the main object of study. FISWG is an organisation in which several forensic institutes participate, notably the FBI and NFI. Characteristic descriptors are facial features that can be used by a forensic face examiner during forensic evidence evaluation of trace and reference images. The trace image often captures a crime scene and is most of the time taken under uncontrolled conditions. The reference image is a photograph of a suspect and is taken under controlled conditions. During this forensic evidence evaluation, the forensic face examiner pays attention to these characteristic descriptors, mostly shape like and potentially highly discriminating features, and computes the strength of evidence that can be used in a court of law.

The mere fact that the characteristic descriptors are documented does not automatically imply their suitability, in particular for their intended use under forensically relevant conditions. Actually, little research is done on this topic. Also, in light of in the recent adoption of the Daubert rule ("a trial judge must ensure that any and all scientific testimony or evidence admitted is not only relevant, but reliable"), there should be more insight into this matter.

This dissertation addresses two major research questions.

The first research question deals with the suitability of FISWG characteristic descriptors as a means to discriminate, taking (a) human, (b) classifier, (c) feature, and (d) forensic aspects into account. A classifier is a computational structure that uses characteristic descriptors, or more generally features, to compute a value that can be either interpreted as or converted to strength of evidence.

There does not exist a single simple answer to this research question. On one hand, one can argue that from a biometric point of view that is primarily interested in general performance, the results of (b) are both somewhat positive, but mostly negative when we consider the forensic circumstances. The former result is obtained in a limited and less representative setting, whereas the latter result is designed to be representative of various forensic use cases (d) and considers a large subset of FISWG characteristic descriptors. Also, the results of (c) show that the characteristic descriptors are difficult to measure in representative forensic use cases (d), although almost any feature in general would be useless under those circumstances. Results of (a) and (b) seem to indicate that there is little added value in using FISWG characteristic descriptors.

The second research question deals with the suitability of a subject based approach in forensic evidence evaluation, taking (e) empirical results from specific applications, (f) theoretical results, and (g) a framework approach into account.

The framework (g) itself integrates the design of the feature, the biometric score and its forensic use, and the evaluation of performance characteristics, with a specific emphasis

on a subject based approach. A theoretical construction (f), the results of the facial marks study (e) and the two applications (e) contained in (g) clearly show that large differences between subject based and general performance exists. Also, the effect of using subject data in classifiers (e) is clearly seen. These results confirm the suitability of a subject based approach in forensic evidence evaluation; the presented framework can be used as a tool for such a subject based approach.

Overall, we conclude that from a general biometric perspective, FISWG characteristic descriptors are not suitable as a means to discriminate. However, if we also consider them from a biometric perspective that includes the use of subject based data and subject based performance, then in limited cases a FISWG characteristic descriptor can be used as a biometric feature to discriminate a particular subject from a group of subjects. More generally, subject based performance evaluation provides insight into the contribution and limits of FISWG characteristic descriptors.

Recommendations of this dissertation include the conduction of more studies with respect to the use of FISWG characteristic descriptors by forensic face examiners, the collection or creation of more forensically relevant data and information, and the incorporation of a subject based evaluation in method validation.

As a closing remark, this dissertation has systematically considered FISWG characteristic descriptors, both directly and indirectly, starting from a human approach and zooming out to a general framework. It is a contribution that serves the scientific approach as meant by the Daubert rule.

Samenvatting

In deze dissertatie staan FISWG karakteristieke descriptoren centraal. FISWG is een organisatie waarin diverse forensische instituten deelnemen, in het bijzonder de FBI en het NFI. Karakteristieke descriptoren zijn gezichtskenmerken die door een forensisch gezichtsonderzoeker kunnen worden gebruikt tijdens de forensische evaluatie van spoor- en referentiebeelden. Sporen zijn vaak opnamen van een plaats delict en zijn meestal onder ongecontroleerde omstandigheden opgenomen. Het referentiebeeld is een foto van een verdachte en wordt onder gecontroleerde omstandigheden genomen. Tijdens deze forensische evaluatie besteedt de forensische gezichtsonderzoeker aandacht aan deze karakteristieke descriptorenvooral vormen en mogelijk sterk discriminerende eigenschappen-en berekent de bewijslast die vervolgens in een rechtbank kan worden gebruikt.

Het enkele feit dat de karakteristieke beschrijvers zijn gedocumenteerd, impliceert niet automatisch hun geschiktheid, in het bijzonder het beoogde gebruik onder forensische relevante omstandigheden. Eigenlijk is er weinig onderzoek gedaan naar dit onderwerp. Ook in het licht van de recente introductie van de Daubert-regel (“a trial judge must ensure that any and all scientific testimony or evidence admitted is not only relevant, but reliable”), zou er meer inzicht moeten komen in deze kwestie.

Deze dissertatie behandelt twee algemene onderzoeksvragen. De eerste onderzoeksvraag behandelt de geschiktheid van FISWG karakteristieke descriptoren om te onderscheiden, waarbij rekening wordt gehouden met (a) menselijke, (b) classifier, (c) kenmerk en (d) forensische aspecten. Een classifier is een rekenkundige structuur die karakteristieke descriptoren, of algemener kenmerken, gebruikt om een waarde uit te rekenen die óf kan worden geïnterpreteerd als óf omgezet kan worden in bewijslast.

Er bestaat geen eenduidig antwoord op deze onderzoeksvraag. Aan de ene kant zou men kunnen stellen dat vanuit een biometrisch standpunt dat voornamelijk geïnteresseerd is in algemene prestaties, de resultaten zowel een beetje positief zijn (b), maar vooral negatief als we de forensische omstandigheden in ogenschouw nemen. Echter, het eerste resultaat is verkregen in een beperktere en minder representatieve omgeving, terwijl het laatste resultaat is ontworpen om representatief te zijn voor verschillende forensische situaties (d) en een grotere deelverzameling van FISWG karakteristieke descriptoren beschouwt. Ook blijkt uit de resultaten van (c) dat de karakteristieke descriptoren moeilijk te meten zijn in representatieve forensische situaties (d), hoewel bijna elk kenmerk in het algemeen in deze omstandigheden nutteloos zou zijn. Resultaten van (a) en (b) lijken aan te geven dat het gebruik van FISWG karakteristieke descriptoren weinig toegevoegde waarde heeft.

De tweede onderzoeksvraag behandelt de geschiktheid van een persoonsgebonden aanpak tijdens een forensische evaluatie van bewijslast, waarbij rekening wordt gehouden met (e)

empirische resultaten, (f) theoretische resultaten en (g) een framework benadering.

Het framework (g) zelf integreert het ontwerp van het kenmerk, de biometrische score en zijn forensische gebruik en de evaluatie van prestatiekarakteristieken, met daarbij een specifieke nadruk op een persoonsgebonden benadering. Een theoretische constructie (f), de resultaten van de huidtypica studie (e) en de twee toepassingen (e) in (g) laten duidelijk zien dat er grote verschillen bestaan tussen algemene en persoonsgebonden prestaties. Ook het effect van het gebruik van persoonsgebonden data in classifiers (e) is duidelijk te zien. Deze resultaten bevestigen de geschiktheid van een persoonsgebonden aanpak in forensische evaluatie van bewijslast; het gepresenteerde framework kan worden gebruikt als een hulpmiddel bij zo'n persoonsgebonden aanpak.

We concluderen dat over het algemeen vanuit een algemeen biometrisch perspectief FISWG karakteristieke descriptoren niet geschikt zijn als een middel om te onderscheiden. Echter, als we ze ook vanuit een biometrisch perspectief dat persoonsgebonden prestaties omvat beschouwen, dan kan een FISWG karakteristieke descriptor in sommige gevallen als biometrisch kenmerk worden gebruikt om een specifiek persoon te onderscheiden van een groep personen. Meer algemeen, een persoonsgebonden prestatie evaluatie geeft inzicht in de bijdrage en de limieten van FISWG karakteristieke descriptoren.

Aanbevelingen in deze dissertatie zijn onder andere het verrichten van meer studies met betrekking tot het gebruik van FISWG karakteristieke descriptoren door forensische gezichtsonderzoekers, het verzamelen of creëren van meer forensisch relevante gegevens en informatie en het opnemen van een persoonsgebonden evaluatie in methodevalidatie.

Ter afsluiting, deze dissertatie heeft systematisch gekeken naar FISWG karakteristieke descriptoren, zowel direct als indirect en zoomt vanaf een menselijke aanpak uit naar een generiek framework. Het is een bijdrage die recht doet aan de wetenschappelijke benadering zoals bedoeld door de Daubert rule.

Dankwoord

Elke zeiltocht begint met een eerste briesje. Ergens begin 2009 stak dat briesje op, vier jaar later was er voldoende wind om óf op de boot te stappen óf deze kans definitief te laten varen. Een tocht zoals dit maak je niet alleen en ik weet zeker dat ik zonder de steun, humor en "de andere kijk op dingen" van de vele reisgenoten het in ieder geval een stuk minder leuk zou hebben gevonden. Promoveren is soms dobberen zonder een duidelijk reisdoel, maar het is uiteindelijk vooral een louterende ervaring. Ten minste, achteraf.

Ik wil een aantal van deze reisgenoten bedanken.

Ten eerste gaat mijn dank natuurlijk uit naar Raymond Veldhuis en Luuk Spreeuwers, de dagelijkse begeleiders van dit project. Ik dank jullie voor de geboden vrijheid, jullie combinatie van het grote plaatje en het detail en vooral de discussies, met regelmatig een goede grap. Mijn waardering gaat uit naar jullie manier van wetenschap bedrijven. Begin bij het probleem in plaats van "als je een hamer hebt is elk probleem een spijker". Overigens nam ik de "dagelijkse begeleiding" misschien iets te letterlijk als ik weer eens langsging bij Raymond voor een "kort vraagje" dat vaak toch uitmondde in een discussie van een half uur voor het whiteboard. Oh ja, dat lesgeven Luuk, stiekem was dat wel erg leuk, ook al mopperde ik wel over al die "tijd die ik wel ergens anders aan kon besteden".

Ook dank aan Didier Meuwly, die later in het promotieproject wat meer in beeld kwam. Ik heb diep respect voor je enorme kennis van en passie voor je vak; bij elke discussie leek je me weer op zaken te wijzen waar ik weer eens te onduidelijk was of gewoon niet aan had gedacht. Iets verder weg, maar zeker niet minder belangrijk, Arnout Ruifrok, "onze man bij het NFI". Je bent vanaf het begin betrokken bij dit project als een soort praktische vraagbaak en je wees ons al snel op FISWG. Ik dank je voor de behulpzaamheid als ik weer iets wilde weten, het meedenken over de opzet en uitvoering van allerlei experimenten en je interesse in de voortgang. Ook wil ik de overige commissieleden bedanken voor hun bereidheid om zitting te willen nemen in de promotiecommissie. In het bijzonder dank ik Chris Klaassen voor zijn vele gedetailleerde opmerkingen.

Tijdens een promotietraject breng je naast de begeleiding natuurlijk ook veel tijd door met je kamergenoten. Ik schoof als groentje aan bij Yuxi, Tauseef, Chris (1), Ahbishek, Pinar, Jen-Hsuan en Chanjuan. Ook andere promovendi zoals Meiru en Rita liepen toen nog rond in Carré. Na het ontstaan van de SCS leerstoel, verhuisden we naar Zi4070 en begon een periode met onder andere Nanang, Wasim, Johannes, Roeland (hofleverancier van drop, (programmeer)taal- en inburgeringsexpert en verder algemeen partner in crime) en Fieke. Alireza mag niet onvermeld blijven met zijn onvermoeibare textuurstudies.

Ik ben vrij lang de laatst binnengekomen promovendus van Raymond en Luuk geweest. In mijn vierde jaar kwam Erwin als extern promovendus binnen en zijn Diah, Pesi en Nova

daar nog later bijgekomen. Ook hebben Tolga, Soumik en Dan hun plekje in Zi4070 gehad. Gelukkig gaan we regelmatig naar de sportkantine voor goedkoop bier en “meatballs” met de verkeerde saus.

Buiten de “Zi4070” kosmos kom je uiteraard andere promovendi en postdocs tegen (waaronder Jan-Willem, Elmer, Tim, Riccardo, Marco, Bence, Dan en Alexandr) tijdens een lunch, in de social corner of een enkele keer zelfs bij een walk-in fridge naast een kegelbaan. Wat hebben we vaak pittig gediscussieerd en lol gehad. Uiteindelijk is dat, naast de wetenschappelijke verrijking, misschien wel het leukste van het werken op een universiteit: de waaier van culturen, gewoonten en het uiteindelijk beter begrijpen en anders waarderen van je eigen achtergrond.

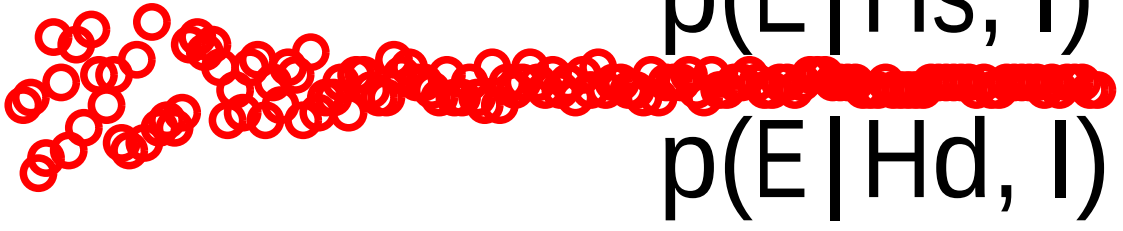
Waar zouden we zijn zonder secretaresses (Sandra, Suse en Bertine) of Geert-Jan (bringer of smiles en around problem solver): dank dat jullie altijd maar weer klaarstaan voor ons allemaal.

De afgelopen jaren heeft een aantal mensen niet zo goed nieuws gekregen en gelukkig zijn die er allemaal in redelijke of goede gezondheid nog bij, in het bijzonder mijn vader. Lieve Dad, fijn dat je nog gewoon veel dingen op jouw manier kan en blijft doen. Niets is leuker dan zittend aan jouw keukentafel een ideetje voor een artikel te schetsen terwijl jij ongehinderd commentaar geeft op een snookerwedstrijd.

Ook dank aan familie, vrienden en collega’s voor jullie voortdurende interesse, in het bijzonder Rob en Rashida, en allen die ook hebben geparticipeerd in experimenten: Yuxi, Chris (1), Roeland, Vincent, Marjolein, Peter, Hanneke en die vele vrijwilligers uit het netwerk van Arnout. Een bijzondere vermelding is op z’n plaats voor Liz en Loretta die zelfs aan meerdere experimenten hebben meegedaan.

Er is één persoon die deze hele tocht mogelijk heeft gemaakt, hoewel ze dat altijd systematisch ontkent. Terwijl ik de afgelopen vier jaar in het ruim zat en mijn tijd besteedde aan wetenschap en aanverwante zaken, stond jij, mijn liefste Hanneke, gewoon de hele tijd aan dek, achter het roer met een blik op de horizon en zorgde je voor een behouden vaart door weer en wind.

Elke zeiltocht heeft een laatste haven. Hanneke en ik zijn op de plaats van bestemming aangekomen, ik draai mij om en ik geniet in verwondering van het uitzicht. Maar ik weet ook: “When at last the work is done, don’t sit down, it is time to dig another one”.


p(E|Hs, I)
p(E|Hd, I)