Face comparison in forensics

A deep dive into deep learning and likelihood rations

Macarulla Rodríguez, A.

Publication date
2024

Document Version
Final published version

Face Comparison in Forensics: A Deep Dive into Deep Learning and Likelihood Ratios

Andrea Macarulla Rodríguez

Face Comparison in Forensics

A Deep Dive into Deep Learning and Likelihood Ratios

Andrea Macarulla Rodríguez

Face Comparison in Forensics

A Deep Dive into Deep Learning and Likelihood Ratios

Andrea Macarulla Rodríguez

This dissertation was typeset by the author using LaTeX 2_ε, originally developed by Leslie Lamport and based on Donald Knuth's TeX. The body text is set in 12 point Egenolff-Berner Garamond, a revival of Claude Garamont's humanist typeface.

A template that can be used to format a PhD dissertation with this look & feel has been released under the permissive AGPL license and can be found online from the author's github repository, at github.com/lhoangan/template-uva-thesis, which originates from its lead author, Jordan Suchow, at github.com/suchow/Dissertate.

Face Comparison in Forensics

A Deep Dive into Deep Learning and Likelihood Ratios

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. ir. P.P.C.C. Verbeek
ten overstaan van een door het College voor Promoties ingestelde commissie,
in het openbaar te verdedigen in de Aula der Universiteit
op vrijdag 16 februari 2024, te 11.00 uur

door

Andrea Macarulla Rodríguez

geboren te Valladolid

Promotiecommissie

Promotor:	prof. dr. ing. Z.J.M.H. Geradts	Universiteit van Amsterdam
Promotor:	prof. dr. M. Worring	Universiteit van Amsterdam
Overige leden:	prof. dr. C. Champod	Université de Lausanne
	prof. dr. D. Meuwly	Universiteit Twente
	prof. dr. M.J. Sjerps	Universiteit van Amsterdam
	prof. dr. T. Gevers	Universiteit van Amsterdam
	prof. dr. S. Ghebreab	Universiteit van Amsterdam

Faculteit der Natuurwetenschappen, Wiskunde en Informatica

The work described in this thesis has been carried out within the Netherlands Forensic Institute, together with MultiX at UvA Informatics Institute, Amsterdam. The printing of this thesis was financially supported by the Co van Ledden Hulsebosch Center, Netherlands Center for Forensic Science and Medicine.

Face Comparison in Forensics
A Deep Dive into Deep Learning and Likelihood Ratios

Abstract

This thesis explores the transformative potential of deep learning techniques in the field of forensic face recognition. It aims to address the pivotal question of how deep learning can advance this traditionally manual field, focusing on three key areas: forensic face comparison, face image quality assessment, and likelihood ratio estimation. Using a comparative analysis of open-source automated systems and forensic experts, the study finds that automated systems excel in identifying non-matches in low-quality images, but lag behind experts in high-quality settings. The thesis also investigates the role of calibration methods in estimating likelihood ratios, revealing that quality score-based and feature-based calibrations are more effective than naive methods. To enhance face image quality assessment, a multi-task explainable quality network is proposed that not only gauges image quality, but also identifies contributing factors. Additionally, a novel images-to-video recognition method is introduced to improve the estimation of likelihood ratios in surveillance settings. The study employs multiple datasets and software systems for its evaluations, aiming for a comprehensive analysis that can serve as a cornerstone for future research in forensic face recognition.

Contents

Never tell me the odds!

—Han Solo

1 Introduction

Face recognition technology has evolved substantially, transitioning from manual methods to automated systems powered by machine learning and deep learning algorithms. This thesis explores the advancements and challenges in integrating face recognition into the domain of forensic science, specifically focusing on the application of deep learning techniques. The study addresses several key research questions aimed at advancing the field of forensic face recognition. These include the efficacy of deep learning in forensic face comparison, face image quality assessment, and the estimation of likelihood ratios, which are statistical measures used to evaluate the strength of evidence. Both manual and automated methods are examined, highlighting their respective strengths, limitations, and potential for synergy. The thesis also delves into specialized topics such as calibration techniques for likelihood ratio estimation and the challenges posed by surveillance video data. By leveraging state-of-the-art deep learning technologies and innovative methodologies, this work aims to contribute to the advancement of face recognition as a reliable and efficient tool in forensic investigations, thereby aiding the pursuit of justice.

1.1 Automated Face recognition in Forensics: Likelihood Ratios

The ability to recognize faces is deeply ingrained in our biological and social evolution [1]. It dates back to the earliest days of human history, when our ancestors could only rely on facial features to recognize each other. Over time, with the emergence of photography in

the 19th century, a paradigm shift occurred in the way we recognized and remembered people, and also how identities were stored and verified. This shift was dramatically illustrated in the work of Alphonse Bertillon, a French police officer who developed a system of identifying individuals based on a set of precise body measurements. His approach also involved capturing two types of photographs, one frontal and one profile view, which has influenced the practice of taking mug shots to this day [2]. From Bertillon's work onwards, face recognition no longer relied on human memory alone. This had a major impact on how investigations of crimes were conducted, as suspects could now be compared to photographs. Face recognition became a critical component of forensic science [3].

Face recognition was historically a manual process, relying on human perception and judgment to distinguish individuals based on their unique facial features. This could involve comparing two or more photographs or sketches and looking for commonalities or differences in features, such as the distance between the eyes, the shape of the nose, or the contour of the lips. Bertillon's system laid the groundwork for systematic identification and documentation, even though it didn't specifically involve facial recognition in the modern sense [2]. Still, his principles of precision and repeatability echo in the automated systems of today. Yet, the traditional face comparison process was time-consuming, labour-intensive, and prone to errors due to the subjective nature of human observation [4].

With the advent of computer technology in the late 20th century, face recognition began to evolve into an automated process. Computers were programmed to identify and compare various features of a face, thereby reducing the dependence on human experts and increasing the speed and consistency of the process [3]. These traditional methods focused on detecting facial landmarks and comparing distances and angles among them. Automatic landmark based methods marked the dawn of a new era in face recognition, setting the stage for the advanced technologies we see today [4].

The next major leap in face recognition came with the advent of machine learning and artificial intelligence in the 21st century. Today, powerful algorithms can analyze and compare millions of faces in a fraction of the time it would take a human expert. The level of accuracy often surpasses human capabilities [4] but these are aggregated results. In many difficult cases systems are not able to make accurate decisions. See figure 1.1 for an example of how difficult this can be. So, despite these technological advancements, it is essential to remember that face recognition, at its core, is about the intricate art of identifying individuals based on their unique facial features, a skill that humans have been refining since the dawn of our species. Solutions for face recognition in difficult cases should leverage the skills of both humans and machines.

(a) *(b)*

(c) *(d)*

Figure 1.1: Illustration of misidentification by an automated system. Images (a) and (b) represent the same individual, George Bush, erroneously identified as different individuals by the system. Conversely, images (c) and (d) depict two distinct individuals misclassified as the same individual, exemplifying an impostor scenario.

Transitioning from the generalized concept and applications of face recognition, we now delve into its specific integration within the domain of forensic science. Forensic science uses rigorous scientific methodologies to respond to inquiries within the legal system, often involving the analysis of various types of traces uncovered at a crime scene. These traces span from DNA, fingerprints, and footprints, to other forms of physical and biological evidence [5]. Particularly, visual evidence, such as images procured from surveillance cameras, often holds significant importance in the field of Forensic Science.

1

The application of face recognition technology within the field of forensic science introduces a unique set of challenges and opportunities. It enables investigators to identify individuals from surveillance footage by comparing their facial features with a repository of known faces, or pinpointing the individuals behind a terrorist threat announced on social media. This advanced use of technology has proven to be instrumental in modern forensic science, directly contributing to the pursuit of truth and justice. As we delve deeper into this technological era, the methods of implementing face recognition in forensic science are evolving, and the emphasis is now on enhancing accuracy and precision, setting the stage for the next phase which explores both manual and automatic methods [6; 7].

As the use of face recognition in criminal cases continues to progress, the accurate assessment and interpretation of face recognition results becomes increasingly critical to ensure credible conclusions and uphold the integrity of forensic science. This precision is essential for maintaining the reliability of forensic evidence, which forms the bedrock of many legal adjudications [8; 9]. To move forward in face recognition in forensic science we should consider how we can improve manual and automatic [6; 7] methods as well as their combination.

Manual face recognition depends on human experts visually analyzing facial features and their variations across different images, including pose, expression, illumination, occlusion, ageing, and disguise [7; 10]. This method is often carried out by forensic facial examiners, who are trained to compare faces using standardized procedures and guidelines [8; 9]. These examiners may also call upon individuals with exceptional face recognition abilities—often referred to as "super-recognizers"—to assist them in their tasks. A super-recognizer is a person who has an extraordinary ability to recognize faces, often remembering faces they've seen only once or in passing [9]. Additionally, dedicated software designed to interactively annotate and measure facial features can provide additional support to human examiners, thereby enhancing the consistency and efficiency of the manual face recognition process [7; 10]. Software to support manual face recognition is focused on the analysis of an individual image.

Automatic face recognition, on the other hand, is grounded in the extraction of numerical features from facial images and the comparison of these features through mathematical models and algorithms [11]. As indicated, initially, automatic methods focused on detecting facial landmarks such as the eyes, nose, mouth, and jawline, and measuring the distances and angles between these points. This approach essentially mimics the manual assessment conducted by human experts, but with increased speed and consistency. Various techniques have been developed to enhance this process. For instance, the Eigenfaces

method uses principal component analysis (PCA) to reduce the dimensionality of face images and represent them as linear combinations of eigenvectors, or 'eigenfaces'. The similarity between two faces is measured by the Euclidean distance between their eigenface coefficients [12]. The Fisherfaces method extends Eigenfaces by using linear discriminant analysis (LDA) to find the optimal projection that maximizes the between-class scatter and minimizes the within-class scatter of face images. The similarity between two faces is measured by the Mahalanobis distance between their Fisherface coefficients [13]. Another approach uses Local Binary Patterns (LBP), which are local texture features from face images which are extracted by dividing them into small regions and computing a binary code for each pixel based on its neighborhood. The histogram of these codes is used as a feature vector for each region. The similarity between two faces is measured by the chi-square distance between their LBP histograms [14]. The Scale-Invariant Feature Transform (SIFT) method detects and describes local interest points in face images that are invariant to scale and rotation. The SIFT descriptor is a 128-dimensional vector that captures the gradient information around each interest point. The similarity between two faces is measured by the sum of squared differences (SSD) or the ratio test between their SIFT descriptors [15]. Building upon these foundational techniques, the most recent advancements in automatic face recognition are predominantly powered by deep convolutional neural networks (DCNNs).These artificial neural networks have the ability to learn complex patterns from large volumes of data, leading to significant improvements in face identification tests over the past few years [16; 17]. This continuous progress in the realms of artificial intelligence and machine learning holds promise for further enhancements in the performance of automatic face recognition systems. Advancements in deep neural networks open the door for systems to manage an expanding range of forensic scenarios, while also adapting to a diverse array of conditions and constraints [17].

While both manual and automatic face recognition carry their respective strengths, they also possess limitations and uncertainties when applied in forensic processes [18]. Manual face recognition, being more subjective, is susceptible to human errors and biases [19]. On the other hand, automatic face recognition depends not only on the on the performance of the algorithms but also on the quality and volume of the training data [20]. It is worth noting that automatic systems may also carry inherent biases, often stemming from the data they are trained on. Thus, it becomes imperative to quantify the reliability and validity of face recognition results in forensic processes, employing statistical methods to do so. Due to their complementary abilities, the collaboration between human experts and automatic systems in the process of face recognition can greatly enhance forensic science. Human ex-

perts, with their ability to discern specific facial characteristics that may be difficult for algorithms, can provide invaluable insights. The combination with other biometric features from the body such as height, or shape of hands etc. will provide additional evidence. On the other hand, automatic systems excel in handling large-scale data processing tasks with efficiency [21]. A collaborative approach ensures that the strengths of both methods are capitalized on, leading to more accurate and reliable outcomes in face recognition [9; 22].

In the field of forensic science, being able to reliably define the value of evidence is crucial for appropriate decision making. A range of statistical methods are utilized to enhance the decision-making processes. Among these, likelihood ratios have found significant applications [18]. Likelihood ratios represent ratios of probabilities that measure the extent to which a piece of evidence supports or contradicts a given hypothesis [23]. In the context of face recognition, likelihood ratios provide a means to express the probability that two facial images belong to the same person or two different individuals, given their observable degree of similarity or dissimilarity [24]. These ratios can be derived from both human judgments and algorithm outputs. The concept and application of likelihood ratios have been extensively studied, not only for face recognition, but also for DNA, fingerprints, and other forms of physical and biological evidence. Integrating likelihood ratios into the forensic decision-making process can lead to more transparent and robust conclusions, thereby reducing the risk of wrongful convictions and enhancing public trust in the justice system [25].

While likelihood ratios can effectively be applied to still images in face recognition, the landscape of forensic science frequently requires the analysis of video evidence. Videos are prevalent sources of evidence in criminal investigations, with videos often providing a richer context by capturing facial movements, expressions, gestures, and situational context [8; 9]. However, video data presents additional challenges for face recognition, including factors such as low resolution, compression artifacts, motion blur, occlusion, and varying viewpoints. These factors require the development and application of more advanced techniques and tools for extracting and comparing facial features across different frames and modalities [10; 26]. Multimodal approaches integrate information from various sources, such as audio and contextual cues. They could potentially enhance the accuracy and reliability of face recognition in forensic applications [27]. This thesis will primarily focus on the analysis of images and videos for face recognition.

In addition, to address the limitations and uncertainties of both manual and automatic face recognition methods, enhancing accuracy and reliability in face recognition within forensic science can significantly increase its potential to contribute to the pursuit of justice.

This enhancement can be achieved through the development of new methods and standards. These may include advanced preprocessing techniques aimed at mitigating the impact of low-quality images, and innovative feature extraction methods tailored to capture facial details more effectively. Robust machine learning algorithms could be employed, capable of handling diverse and challenging scenarios. Additionally, the establishment of standardized protocols and best practices specifically for face recognition within forensic science is crucial. Proper preprocessing, robust machine learning, and standardized protocols can all increase the reliability and credibility of face recognition as a crucial tool in forensic investigations.

As face recognition technology continues to advance, its role within forensic science is expected to expand and evolve. Successful integration and application of face recognition in forensic science can only be achieved with careful implementation of statistical methods like likelihood ratios. This involves fostering a collaboration between human experts and machines, harnessing multiple sources of evidence, and encouraging the development of new methodologies and standards. The utilization of likelihood ratios in particular can provide a more nuanced understanding of the results, allowing for more precise and robust conclusions. By incorporating these elements, face recognition can become an increasingly valuable and powerful tool in the ongoing quest for truth and justice. Ultimately, the effective incorporation of face recognition, bolstered by the thoughtful use of likelihood ratios, can contribute to a more accurate, efficient, and fair legal system.

1.2 Research questions

With the rising tide of deep learning technologies, new opportunities are surfacing for advancements in various fields, one of which is forensic face recognition. Traditionally dependent on manual identification methods, forensic face recognition now stands at the precipice of a significant transformation led by deep learning algorithms. This thesis aims to traverse this chasm and determine how these innovative technologies can revolutionize the discipline. The pivotal question that spearheads this exploration is:

- **In what ways can Deep Learning techniques advance the field of Forensic Face Recognition?**

Delving into the potential applications of deep learning techniques in this field, our primary focus lies in three key areas: forensic face comparison, face image quality assessment, and likelihood ratio estimation. Each of these areas represents a crucial aspect of forensic

face recognition, holding unique challenges and opportunities for enhancement through deep learning. As an example of a difficult decision for the automated system to decide if it is the same person or different ones, see figure 1.1.

Forensic face recognition plays a pivotal role in legal proceedings, providing crucial insights to human investigators. A key part of this process involves the estimation of likelihood ratios, a technique commonly employed in fields such as DNA or glass source comparisons. However, it remains an open question whether such an approach can be effectively applied to the domain of forensic face recognition using deep learning. This leads to the following subquestion:

- **Can we estimate Likelihood Ratios when performing Forensic Face Comparison using Deep Learning?**

To tackle this question, in chapter 2 we compare the performance of three open-source automated systems—OpenFace, SeetaFace, and FaceNet—with that of forensic facial comparison experts. These systems, all based on convolutional neural networks, return either a distance (OpenFace, FaceNet) or similarity (SeetaFace), which is then converted to a likelihood ratio using three different distribution fits: a parametric fit with a Weibull distribution, a nonparametric fit based on kernel density estimation, and isotonic regression with the pool adjacent violators algorithm.

The results reveal that automated systems demonstrate superior performance in detecting non-matches with low-quality frontal images, achieving 100% precision and specificity in a confusion matrix compared to 89% and 86% respectively achieved by investigators. However, with good quality images, forensic experts deliver superior results. Notably, a rank correlation of around 80% was observed between investigators and software.

The estimation of likelihood ratios in forensic face recognition is influenced by the calibration process, where scores are transformed into these ratios. An important aspect that warrants investigation is the type of database used for calibration and its potential impact on the accuracy of the estimated likelihood ratios. Specifically, it is essential to ascertain if the calibration process is influenced by whether the calibrating pairs are random or share similar features with the test subjects. This leads us to the following subquestion:

- **How does the Calibration method affect Likelihood Ratio estimation?**

To address this question, in chapter 3 we explore the performance of three distinct calibration techniques - naive calibration, quality score-based calibration using typicality, and

feature-based calibration. These techniques have been successfully applied in other forensic disciplines but have not been thoroughly investigated within the context of facial image recognition. Maintaining transparency is critical in forensic procedures. Therefore, we compare the performance of state-of-the-art open software with a widely used commercial system. Using the European Network of Forensic Science Institutes (ENFSI) Proficiency tests as a benchmark, we evaluate the calibration results on three public databases: Labeled Faces in the Wild, SC Face, and ForenFace. Our findings suggest that quality score-based and feature-based calibrations outperform naive calibration. Yet, the commercial system outperforms the open software in estimating these likelihood ratios. Despite the commercial system's superior performance, the transparency offered by open software underscores the need for ongoing research to enhance the effectiveness and transparency of forensic facial image comparison methodologies.

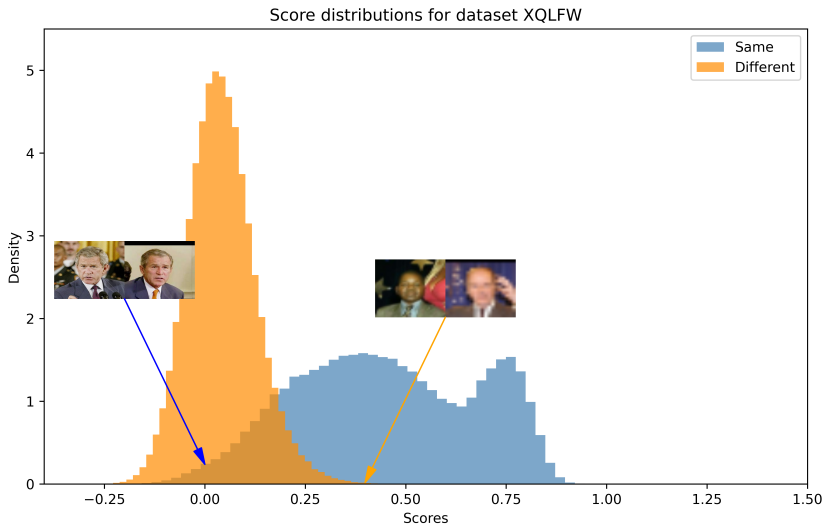


Figure 1.2: Calibration functions for same and different person. Images taken from XQLFW [28].

In forensic investigations, the extraction of suspects from surveillance footage is a crucial yet challenging task, often complicated by variable observation conditions and voluminous data. One critical element to consider in this process is the Face Image Quality (FIQ), a metric used to evaluate the utility of a face sample for facial recognition. Present automated FIQ assessment methods, while productive, carry two significant limitations: they yield only a scalar quality value without specifying the factors leading to low quality,

and they are computationally demanding, which inhibits their efficacy in managing large image volumes. An illustration of the calibration can be found in figure 1.2.

These limitations underscore the necessity for an FIQ assessment method that not only determines the quality of a face sample but also elucidates the factors influencing that quality assessment. Such a method would foster a more comprehensive understanding of face sample quality, rendering the assessment process more transparent and potentially more effective. This need motivates the following question:

- **Can we connect face image attributes to Face Image Quality?**

To address this question, in chapter 4 we introduce multi-task explainable quality networks (XQNets). Unlike traditional methods, XQNets not only provide the quality value but also identify the facial and environmental attributes contributing to that value, thereby enhancing our understanding of the factors influencing a sample's quality. During the training process, XQNets autonomously learn how each attribute contributes to the quality value. Moreover, this study proposes a dataset-agnostic quality pairing protocol (DAQP), ensuring that sample pairs are balanced across different datasets and evaluations are fair.

Our experimental results on the LFW, SCface, and ForenFace benchmarks indicate that the proposed approach can be generalized across different datasets and outperforms existing state-of-the-art methods. Consequently, the use of XQNets offers a more efficient and explainable approach to FIQ assessment, making it particularly suitable for large-scale forensic applications. Figure 1.3.illustrated face quality distribution.

Given the numerous challenges posed by surveillance videos for face recognition in forensic investigations, variations such as pose, illumination, and facial expressions can greatly compromise the effectiveness of recognition methodologies. The need for robust solutions that can accurately identify faces under these conditions is therefore critical. Prompted by this challenge, the following question becomes essential:

- **How can we improve the estimation of Likelihood Ratios in surveillance videos for more effective face recognition?**

In response to this question, in chapter 5 we propose a novel image-to-video face recognition method. This method pairs face images with multiple attributes (soft labels) and face image quality (FIQ), followed by the application of three distinct calibration methods to estimate likelihood ratios.

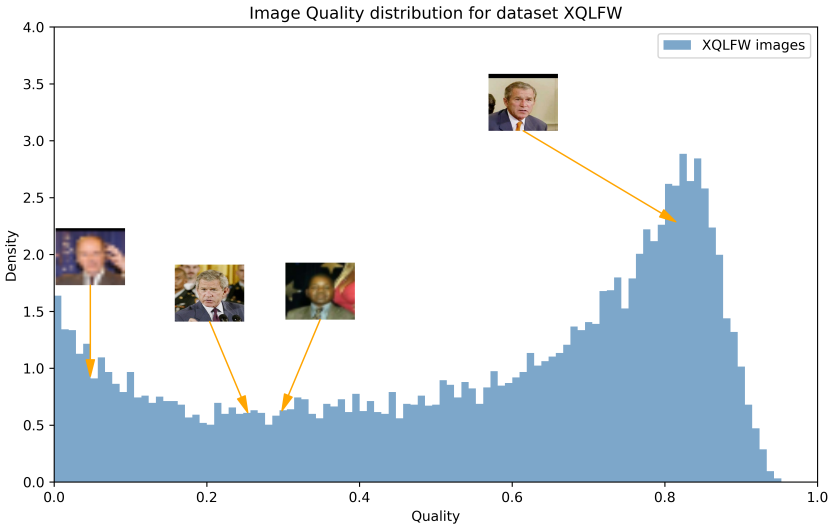


Figure 1.3: Quality distribution for XQLFW. Images taken from XQLFW [28].

The validation of this innovative approach is performed using the ENFSI proficiency test 2015 dataset, with SCFace and ForenFace serving as calibration datasets. Three different embedding models—ArcFace, FaceNet, and QMagFace—are utilized in the evaluation. The results suggest that focusing on high-quality frames significantly improves face recognition performance in forensic applications compared to using all frames. The most favourable outcomes are achieved when the highest number of common attributes between the reference image and selected frames is utilized, or when a single common embedding is created from the selected frames, each weighted according to its face image quality. This chapter introduces a new method for estimating likelihood ratios in surveillance videos, offering a significant contribution to the field of forensic face recognition and enhancing the practical applications and understanding of its implications.

An overview can be seen in figure 1.4.

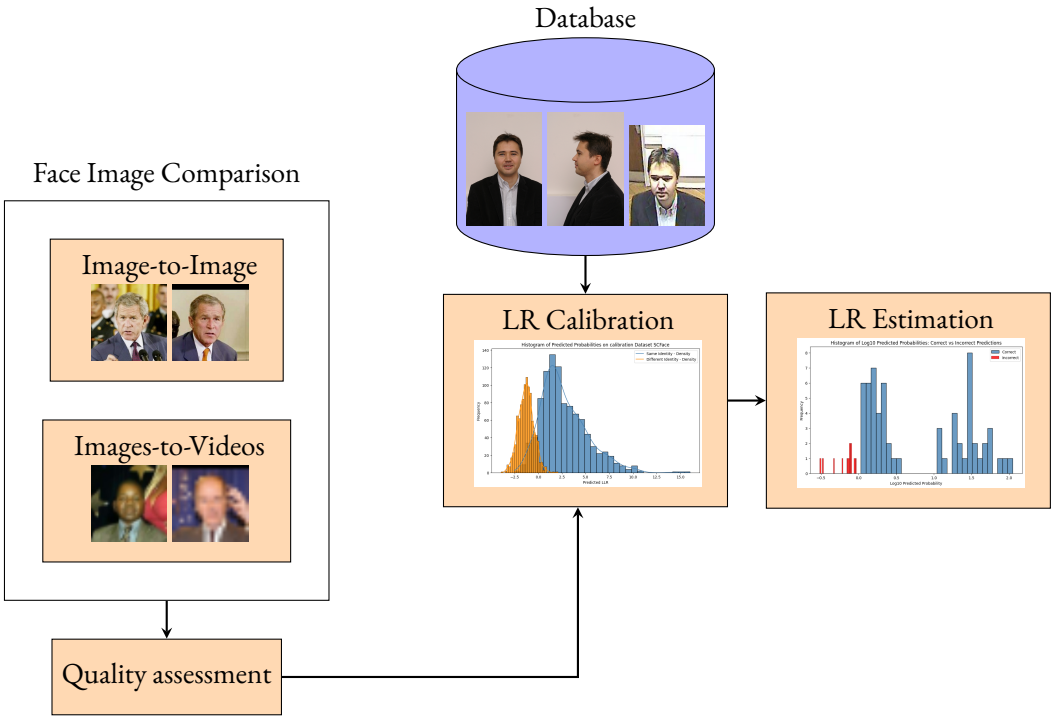


Figure 1.4: Likelihood ratio estimation outline. Reference images (usually ID or high quality pictures) are compared against questioned images (usually surveillance or low quality). Then, using an external database calibration is used to compute LR. Images taken from SCFace [29] and XQLFW [28].

1.3 Contributions

In this thesis, the contributions are organized by chapter, detailing the co-authors and their respective roles. The contributions are as follows:

- **Chapter 2:** Macarulla Rodriguez, A., Geradts, Z., & Worring, M. (2020). Likelihood ratios for deep neural networks in face comparison. Journal of Forensic Sciences, 65(4), 1169-1183.

  Andrea Macarulla Rodriguez: All aspects

  Zeno Geradts: Insight and supervision

  Marcel Worring: Insight and supervision

- **Chapter 3:** Rodriguez, A. M., Geradts, Z., & Worring, M. (2022). Calibration of score based likelihood ratio estimation in automated forensic facial image comparison. Forensic Science International, 334, 111239.

  Andrea Macarulla Rodriguez: All aspects

  Zeno Geradts: Insight and supervision

  Marcel Worring: Insight and supervision

- **Chapter 4:** Rodriguez, A. M., Unzueta, L., Geradts, Z., Worring, M., & Elordi, U. (2023). Multi-Task Explainable Quality Networks for Large-Scale Forensic Facial Recognition. IEEE Journal of Selected Topics in Signal Processing.

  Andrea Macarulla Rodriguez: All aspects

  Luis Unzueta: Insight and supervision

  Zeno Geradts: Insight and supervision

  Marcel Worring: Insight and supervision

  Unai Elordi: Technical implementation

- **Chapter 5:** "Improved Likelihood Ratios for Surveillance Video Face Recognition with Multimodal Feature Pairing". Under submission to Forensic Science International. Authors: Macarulla Rodriguez, A., Geradts, Z., & Worring, M., Unzueta, L.,

  Andrea Macarulla Rodriguez: All aspects

  Zeno Geradts: Insight and supervision

  Marcel Worring: Insight and supervision

  Luis Unzueta: Insight and supervision

1

2

Likelihood Ratios for Deep Neural Networks in Face Comparison

In this study, We aim to compare the performance of systems and forensic facial comparison experts in terms of likelihood ratio computation to assess the potential of the machine to support the human expert in the courtroom. In forensics, transparency in the methods is essential. Consequently, state-of-the-art free software was preferred over commercial software. Three different open-source automated systems chosen for their availability and clarity were as follows: OpenFace, SeetaFace, and FaceNet; all three based on convolutional neural networks that return a distance (OpenFace, FaceNet) or similarity (SeetaFace). The returned distance or similarity is converted to a likelihood ratio using three different distribution fits: parametric fit Weibull distribution, nonparametric fit kernel density estimation, and isotonic regression with pool adjacent violators algorithm. The results show that with low-quality frontal images, automated systems have better performance to detect non-matches than investigators. 100% of precision and specificity in confusion matrix against 89% and 86% obtained by investigators, but with good quality images, forensic experts have better results. The rank correlation between investigators and software is around 80%. We conclude the software can help reporting officers, as it can do faster and more reliable comparisons with full-frontal images, which can help the forensic expert in casework.

2.1 Introduction

Face recognition is a powerful biometric technique to recognize a person due to its non-intrusive characteristics [30]. Unlike other biometric recognition, such as fingerprints or DNA, face recognition does not require cooperation from the suspect, making it a useful source of evidence. Digital facial evidence can appear in CCTV footage, mugshots, mobile devices, or images from social media sites [8; 31], which are now commonly used in court [10]. An example use is a comparison between the ID image of a suspect and a face image retrieved from CCTV footage. This 1:1 comparison is known as verification or authentication. Organizations such as the European Network of Forensic Science Institutes (ENFSI) stimulate reporting the assertiveness of the statement match/nonmatch, that is, the verification stating whether it is the same person/different person or not, via a quantifiable amount [32]. To that end, ENFSI enforces the use of a likelihood ratio (LR) as the mensurable method to express the confidence in the match/nonmatch decision [32; 33] as also used in DNA or fingerprint comparison [34; 35].

LR is based on Bayes' rule. It is defined as the ratio of the probabilities of two hypotheses: the null hypothesis, here the hypothesis of the prosecution (H_p), and the alternative hypothesis of the defense (H_d) [33]. These terms are considered before certain findings, that is, the evidence E, are taken into account. Evidence in the case of face verification would come in the form of assessment if the face verification would be a match or a nonmatch. For face verification, we consider the null hypothesis a match, and the alternative hypothesis a nonmatch. The LR is defined as follows:

$$LR(H_p, H_d, E) = \frac{Pr(E|H_p)}{Pr(E|H_d)} \tag{2.1}$$

Would it be possible to obtain a valid LR in 1:1 face comparison suitable for forensics? For that end, we use the proceedings to attain an LR based on a biometric score [36; 37]. For face comparison, the biometric score is the value obtained from an automated system that can compute either the distance or dissimilarity between two given faces. Automated face recognition started with the eigenfaces in 1991 by M. Turk and A. Pentland [38]. Since then, automated face recognition has been an active subject of research in the computer vision community. In recent years, AI and Deep Learning have allowed progress and improvement in automated face recognition systems by leaps and bounds. In 2014, DeepFace [39] reached 97.35% accuracy identifying faces in the benchmark dataset Labeled Faces in the Wild (LFW) [40] versus a human performance of 97.53%. The current state-of-the-art

has boosted performance up to 99.80% [41]. Due to the improved performance, automated systems can become assistants of judgment in court [9; 42]. To assess this potential, the LR obtained through the process must be validated for suitability in the forensic field [23; 43].

The main contributions of this chapter consist of carrying out the process of a 1:1 verification end to end from an automated system to the final step of validation in the forensic field. We use three different open-source automated systems: OpenFace [44], SeetaFace [45], and FaceNet [46]. The reason to use these three automated systems is due to their availability and transparency to the user. We obtain either a distance (OpenFace, Seeta) or a similarity (SeetaFace) that is treated as a biometric score. We transform the score through three statistical methods: Weibull distribution [47], kernel density estimation (KDE) [48], and pool adjacent violators algorithm (PAVA) [49]. These methods use a set of scores to generate a probabilistic density function (Weibull, KDE) or a cumulative density function (PAVA). This process of obtaining such functions is commonly known as calibration. The set of scores is obtained from 1:1 comparison in the benchmark LFW, which is publicly available and contains a large set of unconstrained face images. After applying these steps, the LR is obtained. Once the LR is obtained, validation is performed through a comparison to the human expert. This conforms to our second contribution. The comparison with the human experts is based on the yearly ENFSI face recognition proficiency tests. These tests are performed by forensic experts giving a likelihood ratio to each pair of images analyzed, which may be of the same person or not. We will use these tests for both evaluating the performance of the automated system (match/nonmatch success through the Matthews correlation coefficient [50]) and the level of similarity to the forensic expert using rank correlation. The last contribution comes in the conclusion in the form of indications of how the automated tools can be of assistance to the expert based on the results found.

The chapter is organized as follows: First, we review the related work, subdivided on the use of likelihood ratio in forensics in general, automated face recognition advances, and likelihood ratio tied to face recognition. Second, we disentangle step by step the procedure of assessing the likelihood ratio from an automated system score in Methodology. We explain each of the open-source tools, the methods, and the dataset used. In the Results section, we present the accuracy of the automated system reached with the different statistic methods, that is, when it got better or worse combinations of match/nonmatch predictions and the rank correlation with the human investigators. Finally, in the Discussion and Conclusion section, the results are analyzed and the potential of the automated system to assess forensic

decisions is evaluated.

2.2 Related Work

2.2.1 Likelihood Ratio in Forensics

The idea of presenting evidence evaluation in court using a Bayesian probabilistic framework has been encouraged by institutions such as ENFSI in recent years as a suitable way to report evidence to justice [32; 51; 52] as it helps to standardize reasoning. In Europe, there have been initiatives to endorse this approach, for example, by the presentation of a guideline [43]. As a result, forensic laboratories around the world use the likelihood ratio as a means to summarize their findings [33].

The use of likelihood ratio to report results has been explored in several fields of forensic research. DNA trace comparison is probably the area with the largest known use of LR in Europe and has already frequently been used in court [34; 53]. There has been a study in forensic speaker recognition by Ref. [54] that evaluates the performance of different methods used for forensic automatic speaker recognition. In the reference, three methods of speaker recognition (VQ, GMM, and i-vectors) are evaluated in accordance with the methodical guidelines for best practice in forensic semi-automatic and automatic speaker recognition. They conclude that in the experimental conditions of the paper, the three methods compared produce similar results. In forensic fingerprint comparison, the performance of LR for comparisons of fingerprints with fingermarks is studied in Ref. [55]. They conclude that the results obtained could be used as a reference for score-based LR systems in other fields. In addition to applications in biometrics, LR computations have also been done for drug comparison [56], glass analysis [57], and gasoline analysis [58].

General guidelines for validation of the likelihood ratio approach can be found in Refs [23] and [43]. The proposed process of validation takes into account two ways of obtaining likelihood ratios from a biometric comparison: score-based and feature-based. In our chapter, we follow the majority of the work done in the biometric forensic field [31; 54] where validation is based on scores.

2.2.2 Automated Face Recognition

Many methods for automated face recognition are available, coming both from industry and academics [39; 44; 59; 60]. A survey carried out in Ref. [41] compares the current open-source best-performing face recognition algorithms and their accuracies in benchmarks [40]. The work concludes that, since 2014, all the best-performing algorithms are

based on convolutional neural networks [46; 61]. This state-of-the-art software outperforms human recognition in the benchmark dataset Labeled Faces in the Wild [40].

Face recognition algorithms in general consist of three steps: face detection, face normalization, and face identification or verification. Face detection aims to identify the presence of people's faces within an image [60]. It is very well developed and also commonly used, for instance in autofocus in cameras. In the next step, face normalization, faces are aligned by matching landmarks. Each picture is warped so that the eyes and lips are always in the same place in the image. This will make the comparison a lot easier [62]. Finally, identification tries to establish the identity of a person in an image by comparing it to a reference database. In face verification, the model has to determine whether two images of a person belong to the same individual [63].

2.2.3 Face Recognition and Likelihood Ratio

As indicated, face recognition has been widely researched in academia and industry, yet there has been little research in the field of forensic face recognition [31].

There have been attempts to compare automated systems to human performance. For instance, [9] researched groups of forensic experts (super-recognizers, i.e., people with significantly better-than-average face recognition ability, and trained facial reviewers) and untrained recognizers. In their study, they acknowledge that the best algorithms perform in the range of the best humans, that is, professional facial examiners.

The Carabinieri Forensic Investigation Department [42] in Italy carried out successful experiments on comparing commercial system performance in both the ENFSI test and 130 cases, focusing on the accuracy in recognition. The results show that two of the three automatic systems performed superior compared with the mean of the forensic experts. As a next step, the authors recommend computing likelihood ratios as recommended by the ENFSI guideline for evaluative reporting in forensic science. In their paper, they state a strongly optimistic view of the future use of support vector machines and convolutional neural networks.

2.3 Methodology

The objective of this work was to compare the operation of automated facial recognition systems with the way forensic experts assess their findings, and to determine whether automated systems can be helpful tools to the investigator.

When comparing two images, the automated system returns a score or a score plus an empirically calculated threshold. The score does not directly give information whether it is the same person or not, rather the score of the system indicates the confidence of the system in the similarity of the two images under consideration. Therefore, the software output must be converted to LR values that facilitate the reporting of evidential value. To determine the usefulness of the automated systems, the results provided by the researchers must be compared with the LR values obtained from the automatic systems and the true relationship between the images. To compute a LR starting from a score, first calibration of the automated system is required. For that, we need an automatic system that provides a score, and then a statistical method to convert the score into a LR. This statistical method needs a database to perform the calibration. This calibration is done using the public database Labeled Faces in the Wild. Once the LR is obtained, the performance of the automated system is evaluated through the Matthews coefficient. The Matthews coefficient condenses in a single number the quality of the classification based on the confusion matrix. The next step is to compare the LR obtained from the automated system to the LR provided by the forensic experts. This comparison between the automated system and human experts was performed with rank correlation. The overall process, and with that the structure of the chapter, is illustrated in Fig. 2.1. In the process, the automated system and forensic experts act as actuators that receive input (both a pair of images to compare) and expel an output, scores, or distances in the case of the automated systems, and likelihood ratios in the case of forensic experts. For the automated system to output likelihood ratios, it needs to be calibrated through a reference database (in this case Labeled Faces in the Wild). The final goal and main contribution of this chapter is the comparison between the LR obtained by the automated system and the forensic experts, both in accuracy and similitude.

2.3.1 Likelihood Ratio Obtained from ENFSI Tests

ENFSI prepares every year a facial comparison test where forensic experts assess the likeliness of a match for face image pairs. Through the years, the subjects appearing in the comparisons change in nationality, quality of the picture, pose (frontal or different angle), different distances in 2011, or other challenges for face recognition such as compression of the image (2011) different ages (2012) or objects partially covering the face (2013). The characteristics of the tests evaluated can be found in Table 2.1.

In Table 2.2, the ENFSI criteria to determine the likelihood ratio associated with a certain pair of images are shown. Even though the true values of the likelihood ratio cover a

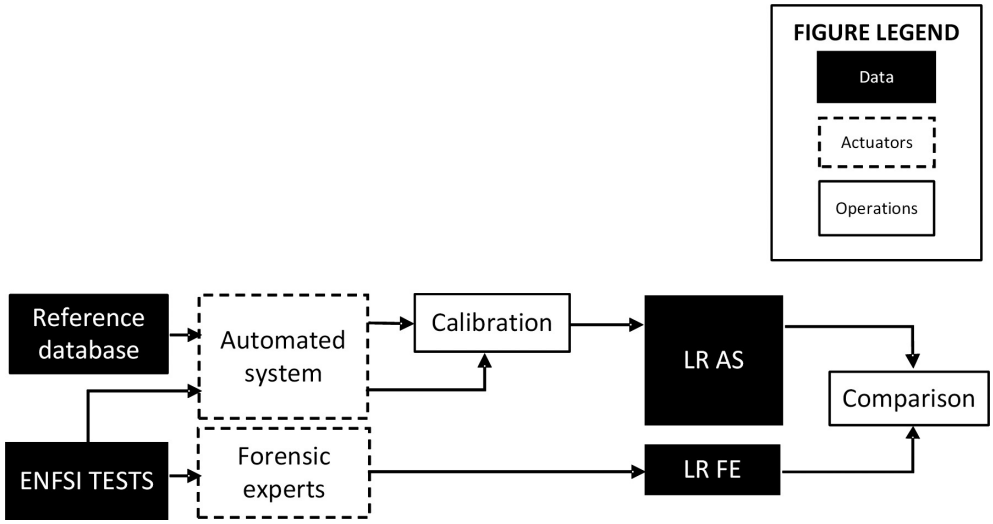


Figure 2.1: Overview of the chapter. Black boxes symbolize data. LFW and ENFSI tests are image datasets, and LR AS and LR FE are the two sets of likelihood ratios obtained from the ENFSI tests from the automated system and the forensic experts, respectively. Dash line indicates actuators, such as automated systems and forensic experts, that receive input (both of them a pair of images to compare) and expel an output, scores, or distances in the case of automated systems, and likelihood ratios in the case of forensic experts. A white box with solid black contour signifies an operation. For the automated system to output likelihood ratios, it needs to be calibrated through a reference database (in this case Labeled Faces in the Wild and the proficiency tests). The final goal and main contribution of this chapter is the comparison between the LR obtained by the automated system and the forensic experts, both in accuracy and similitude.

Table 2.1: Proficiency test characteristics for years 2011, 2012, 2013, and 2017.

	2011	2012	2013	2017
Country organizing the test	Sweden	Sweden	Sweden	Netherlands
Quality	Decent	Decent	Low (CCTV)	Good
Poses	Frontal	3 angles	Frontal	Frontal
Conditions	Distances	Similar	Similar	Similar
Other comments	Compression/ resolution	Up to 5 years in between	With glasses/scarves...	–

larger range, the experts in the ENFSI tests report them on a logarithmic scale for convenience. In Table 2.2, the original LR value is reflected as LR, the reporter logarithmic LR as LLR, and the verbal forensic report as verbal equivalence. For LR > 1, a logarithmic scale from 0 to +5 is used (LLR). When LR < 1, the LLR will be equivalent but with a negative value (from −5 to 0).

Samples from different years are shown in Fig. 2.2. They are referred to as match, that is, both of the images belong to the same person, or nonmatch, which means the pictures belong to different persons. Both the investigator and the automated system must report if the comparison corresponds to match/nonmatch and the degree of certainty about it through the likelihood ratio.

2.3.2 Likelihood Ratio Obtained from Automated Systems

Biometric Score Obtained from Open-source Automated Systems

In forensic science, transparency and explainability are important. Three methods are chosen due to their availability to the users since no license required and the source code is available. This transparency makes OpenFace, SeetaFace, and FaceNet open-source systems suitable for forensic study, in contrast to commercial software that is not open for examination. FaceNet is used due to its high performance in the dataset used to create the LR from the scores (99.65% accuracy). OpenFace and FaceNet are both based on Ref. [46], but OpenFace has faster running time than FaceNet because of its lower number of dimensions. In principle, a higher value of dimensions provides higher accuracy, but also more computational power. Finally, SeetaFace is based on VIPLFaceNet [45], which works with a different backbone network (the convolutional neural network that was trained to make

Table 2.2: Likelihood ratio scale that forensic experts use to assess their comparisons. Table based on Ref. (5) and ENFSI tests.

Values of likelihood ratio	LLR value	Verbal equivalent
10,000–1,000,000	5	*...provide very strong support for the first proposition rather than the alternative* *...are far more probable given...proposition...than proposition...*
1000–10,000	4	*...provide strong support for the first proposition rather than the alternative* *...are much more probable given...proposition...than proposition...*
100–1000	3	*...provide moderately strong support for the first proposition rather than the alternative* *...are appreciably more probable given...proposition...than proposition...*
10–100	2	*...provide moderate support for the first proposition rather than the alternative* *...are more probable given...proposition...than proposition...*
2–10	1	The forensic findings provide weak support for the first proposition relative to the alternative. The forensic findings are slightly more probable given one proposition relative to the other.
0.5–2	0	The forensic findings do not support one proposition over the other. The forensic findings provide no assistance in addressing the issue.

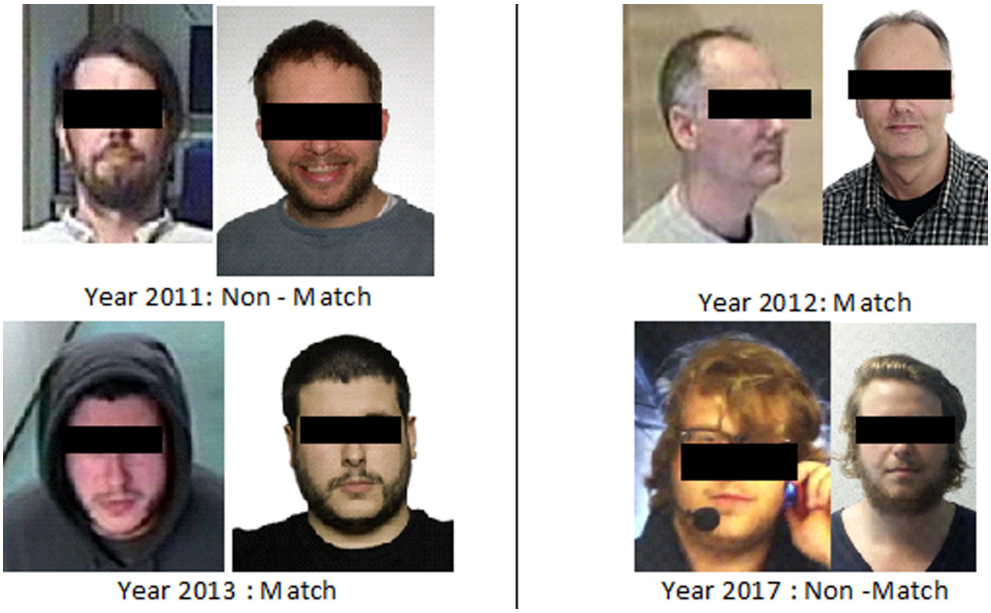


Figure 2.2: Samples of compared images. Match refers to a combination of two images that belong to the same person, and nonmatch refers to different persons.

the faces classification), and thus, the performance may be different from the other two software systems. All of them outperform human performance in the public database Labeled Faces in the Wild [41].

The three systems execute a 1:1 verification. In these automated systems (all based on a convolutional neural network), each detected face is represented as an N-dimensional vector in the space resulting from embedding the high dimensional image space to an N-dimensional feature space. Figure 2.3 shows a sketch of this procedure.

OpenFace is a Python and Torch implementation of face recognition and is based on Ref. [46]. The models are trained with a combination of the two publicly available face recognition datasets: FaceScrub and CASIA-WebFace. The software used for this chapter is a script that predicts a similarity score of two faces by computing the squared L2 distance between their representations, based on a normalized 128-dimensional embedding. A lower score indicates two faces are more likely of the same person. The lower the distance, the more similar the two faces are. It has accuracy on LFW of 92.92% [41]. The methods in Ref. [44] also form the basis for *FaceNet* which is a TensorFlow implementation. It has been trained on VGGFace2 [59], and face alignment has been done using MTCNN [64]. It does its calculations with a 512-dimensional normalized embedding and has an accuracy of 99.63% on LFW. It returns an L1 distance between 0 (same picture) and 2. Finally,

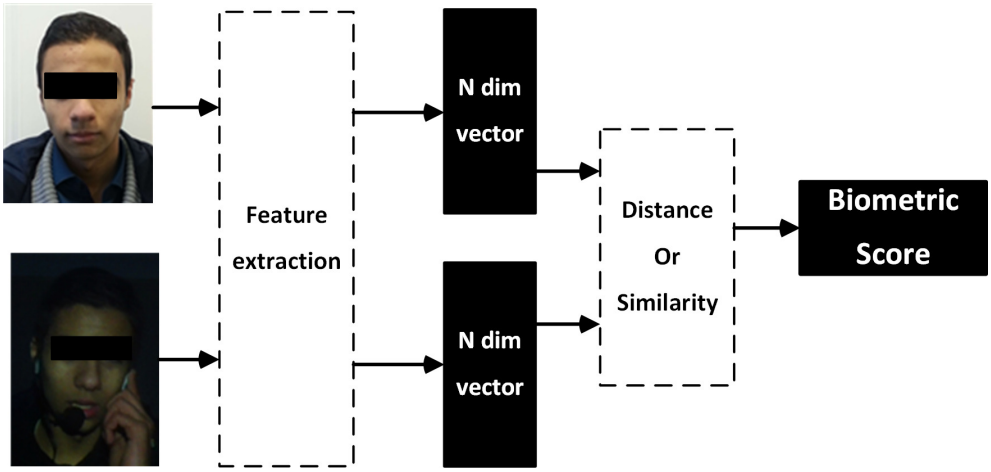


Figure 2.3: How automated systems generate a score. N is the number of dimensions of the embeddings for each representation. The distance measures how different the two embedded feature vectors are.

SeetaFace is a C++ face recognition engine, which can run on a CPU with no third-party dependence. It contains three key parts, namely SeetaFace detection [65], SeetaFace alignment [66], and SeetaFace identification [45].

The image representation is a 2048-dimensional embedding, and the score provided for the comparison between two images is calculated with the cosine similarity resulting in a value between 0 (completely different) and 1 (same image). It reaches 97.1% accuracy on LFW.

From Biometric Score to Likelihood Ratio

As indicated in the introduction, the LR is obtained from two conditional probabilities namely the probability of the evidence conditional to the hypothesis of the prosecution (the two faces belong to the same person) divided by the probability of the evidence conditional to the hypothesis of the defense (the two faces belong to a different person). When we use an automatic system to calculate the similarity between the two faces to be compared, it returns a score. This score in itself has no forensic relevance and that is why we aim to convert it to an LR.

In this chapter, we have chosen three methods commonly used in forensic literature [36; 67] to convert biometric scores into an LR. Methods used are the Weibull model approach [47], a parametric method that approximates two probability distribution functions (PDFs), kernel density estimation (KDE) [48], a parametric method that also generates two PDFs, and the nonparametric isotonic regression that computes a cumulative

distribution function (CDF) [49].

The Weibull distribution was chosen in the first place because it can assume the characteristics of many different types of distributions. It is flexible enough to model a variety of datasets. It can adapt to both skewed data and symmetric data. Weibull is a parametric distribution, which assumes parameters (defining properties) of the population distribution from which the calibration data are drawn. Because of that, the second choice is a kernel density estimation (KDE), which is a nonparametric test that does not make such assumptions. The third method chosen is isotonic regression commonly used machine learning model for statistical inference.

In Weibull distribution approach, if we use a sufficiently large set of scores obtained from comparisons between photographs that belong to the same person (within-source variability, WSV) and comparisons that belong to different ones (between-source variability, BSV), we can infer from these two sets two probability density functions (PDFs). Once we have these two functions, if a new comparison were made (which would be what we would consider evidence in a case), it would be enough to use the score obtained from the automated system as input and plug it in into the PDFs. Thus, we obtain two values, one for the prosecution hypothesis and another for the defense hypothesis. By dividing these two values, we obtain the likelihood ratio. A summary of this concept can be seen in Fig. 2.4.

The Weibull distribution is a continuous probability distribution that we fit the discrete set of scores obtained from the calibration set (LFW [40]). To approximate our set of data, we use the two-parameter Weibull, defined in Eq. 2.2.

$$f_w(x,\beta,\eta) = \frac{\beta}{\eta}\left(\frac{x}{\eta}\right)^{\beta-1} e^{-\left(\frac{x}{\eta}\right)^{\beta}} \tag{2.2}$$

The two-parameter Weibull distribution is commonly used in failure statistic studies and fits well with the histograms obtained with scores provided by automated systems, as seen in Fig. 2.5 The shape parameter (β) of the distribution changes the slope of the function, and the scale parameter (η) regulates the spread of the distribution. Their effects are illustrated in Fig. 2.6

Once the calibrated data are grouped into bins on a histogram, probabilistic functions have to be fitted to the data in order to calibrate. Using both Weibull functions (prosecution generated with BSV and defense generated with WSV), LR is calculated with the following:

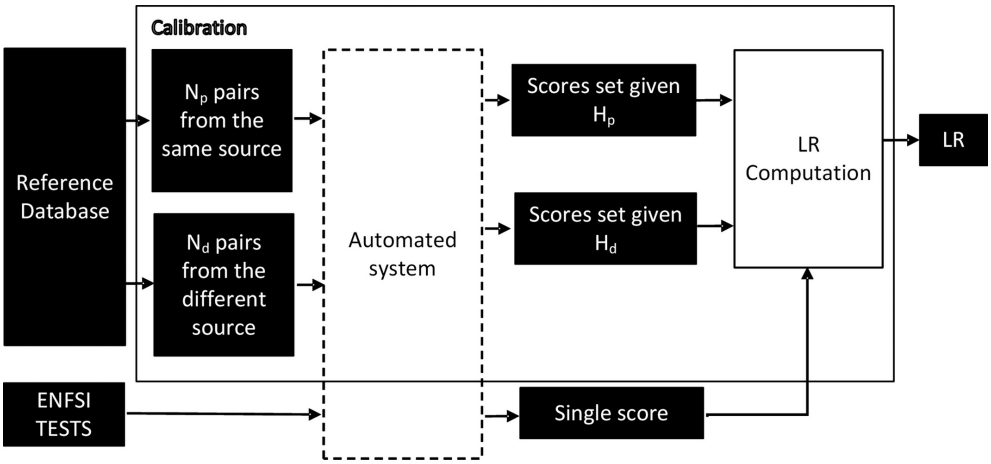


Figure 2.4: Computation of an LR for a pair of biometric specimens consisting of the suspect's biometric specimen and the trace biometric specimen. Figure based on Ref. [36]. The reference database is used to calibrate the automated system. From the calibration, two sets of scores are obtained, one for the same source pair of faces (H_p) and another one for different source pairs of faces (H_d). For each pair of question and reference image in the ENFSI test, the automated system will provide a score. The score is transformed to an LR through the calibration methods Weibull, KDE, and isotonic regression.

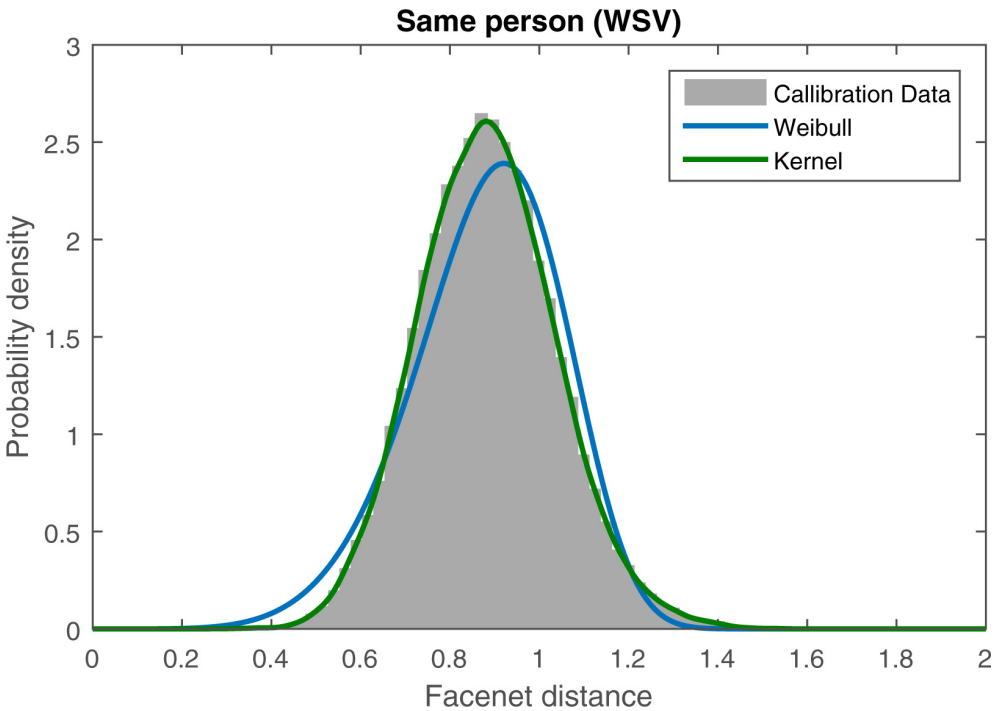


Figure 2.5: Weibull and KDE approximations to histograms generated with calibration data.

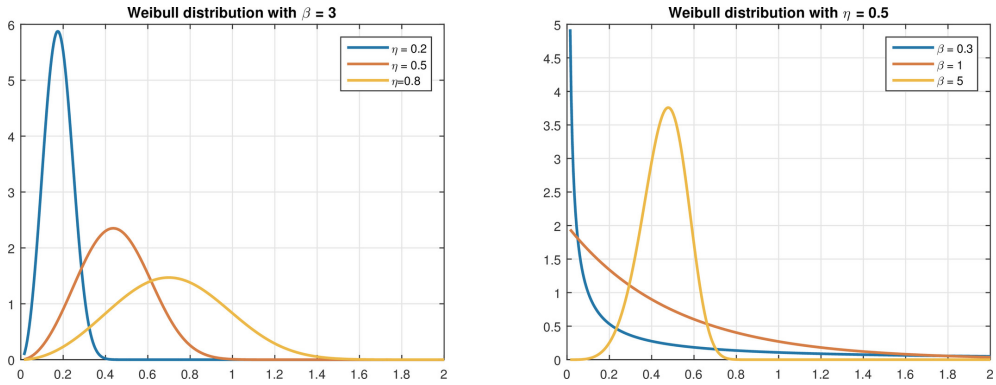


Figure 2.6: Different shape parameters (left figure) and scale parameters (right figure) in the Weibull distribution. The shape parameter (β) of the distribution changes the slope of the function, and the scale parameter (η) regulates the spread of the distribution.

$$LR_w(s) = \frac{Pr_w(s|H_p)}{Pr_w(s|H_d)} = \frac{f_w^p(s, \beta_p, \eta_p)}{f_w^d(s, \beta_d, \eta_d)} \tag{2.3}$$

In kernel density estimation, A kernel distribution iss a nonparametric representation of the probability density function (PDF) of a random variable. It is used when a parametric distribution cannot properly describe the data, or when avoiding making assumptions about the distribution of the data is desired. A kernel distribution is defined by a smoothing function and a bandwidth value h, which controls the smoothness of the resulting density curve. In other words, it is a technique that lets you create a smooth curve given a set of data [48]. It is given by the following equation:

$$f_k(x, h, K) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) = \frac{1}{nh}\sum_{i=1}^{n} K\frac{(x - x_i)}{h}. \tag{2.4}$$

where K is the kernel and h is the bandwidth. The kernel smoothing function defines the shape of the curve used to generate the probability distribution function. Similar to a histogram, the kernel distribution builds a function to represent the probability distribution using the sample data. Unlike a histogram, which places the values into discrete bins, a kernel distribution sums the component smoothing functions for each data value to produce a smooth, continuous probability curve. For this chapter, we will use a Gaussian kernel for the calibrations. The bandwidth steers the smoothness of the resulting approximation. The effect of this parameter is illustrated in Fig. 2.7. It can be observed that small bandwidth values (0.1) can generate overfitting.

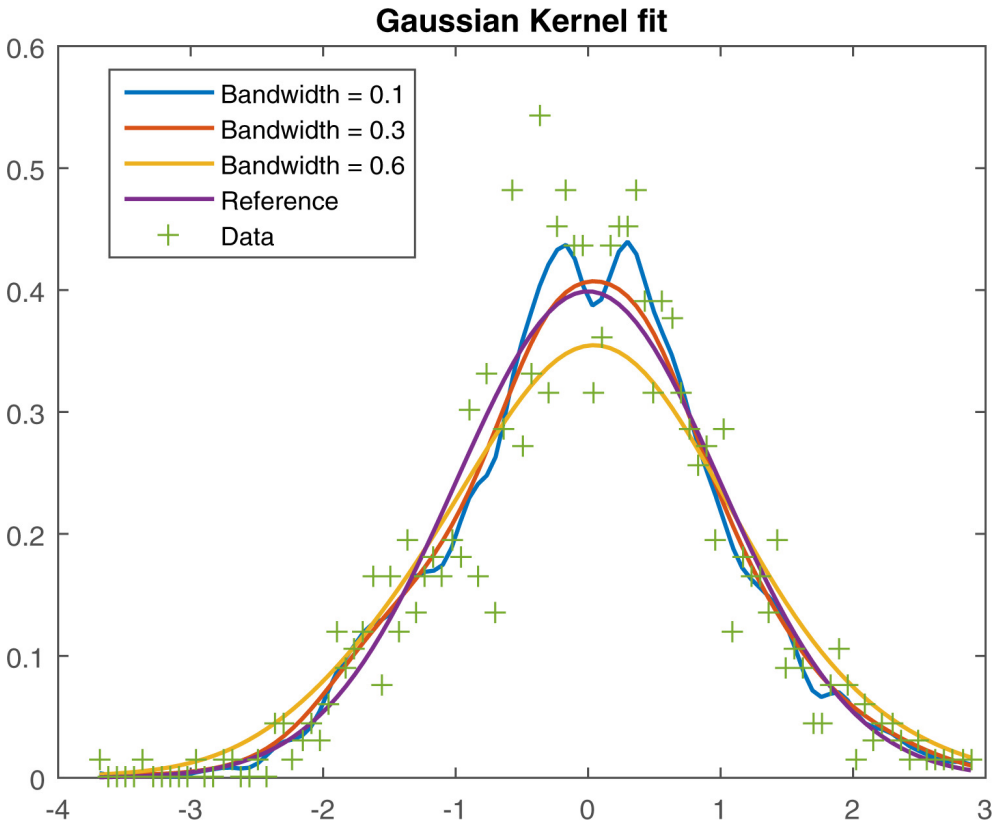


Figure 2.7: KDE with different bandwidth values (h). The bandwidth steers the smoothness of the resulting approximation. Higher values of h smooth the curve, whereas the low values make the curve fit the samples better. However, this can cause overfitting.

Using both kernel functions (prosecution generated with BSV and defense generated with WSV), LR is calculated with the following:

$$LR_k(s) = \frac{Pr_k(s|H_p)}{Pr_k(s|H_d)} = \frac{f_k^p(s, \beta_p, \eta_p)}{f_k^d(s, \beta_d, \eta_d)}. \tag{2.5}$$

Isotonic regression (pool adjacent violators algorithm) can be understood as approximating given series of 1-dimensional observations with a nondecreasing function which has to lie as close to the observations as possible. Isotonic regression is given by the following formula [68]:

$$\min_{g \in A} \sum_{i=1}^{n} w_i (g(x_i) - f(x_i))^2 \tag{2.6}$$

.

where A is the set of all piecewise linear, nondecreasing, continuous functions and f is a known function.

To apply the linear isotonic regression method, we use the pool adjacent violators algorithm (PAVA). Applying PAVA, an increasing function from the scores of a distance (OpenFace and FaceNet) or similarity (SeetaFace) is built. The input to feed the function is calibration scores from both WSV and BSV. In OpenFace and FaceNet, WSV corresponds to low score values (WSV corresponds to a comparison of the same person) and BSV corresponds to high values (comparisons of different persons). The larger the distance value, the higher the probability of the input being different persons. The relationships are completely the opposite of SeetaFace.

Each score obtained from the automated system is assigned a point in the xy plane. In this plane, x is the value of the obtained distance (in OpenFace and FaceNet) or similarity (in SeetaFace). The variable y will be assigned a value of 0 if it belongs to WSV and a value of 1 if it belongs to BSV (OpenFace, FaceNet), and the opposite for SeetaFace. Figure 2.8 left shows a scatter of this value allocation. To achieve isotonic regression, the requirements $y_i + 1 \geq y_i$ for every $x_i + 1 > x_i$ must be satisfied. As seen in Fig. 2.8, the distance values obtained are discrete, they do not satisfy $y_i + 1 \geq y_i$. To satisfy this term, PAVA is applied. The outcome of PAVA is a nondecreasing function with $y_i + 1 \geq y_i$.

There are points with x values that are equal (i.e., $x_i + 1 > x_i$ is not satisfied). All the points with the same x value are substituted by one that has the y value of the average. Also, that point is assigned a weight equal to the number of original points for that x value. With this step, a point cloud with different weights is obtained, but this time $x_i + 1 > x_i$ is satis-

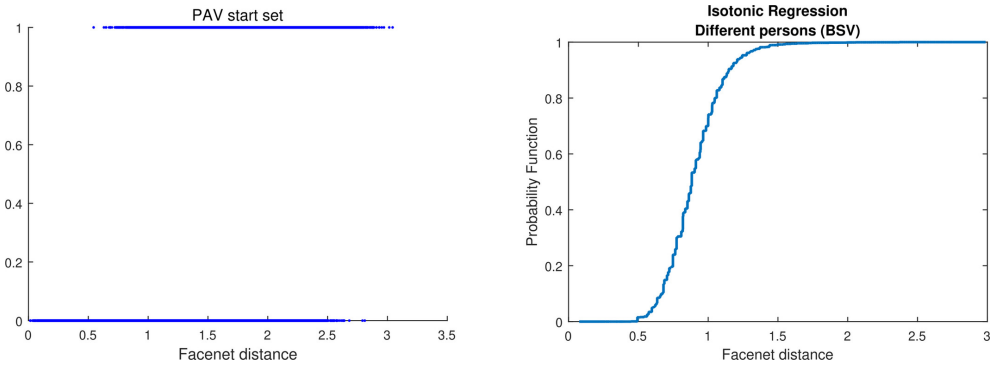


Figure 2.8: Left figure: points (x_i, a_i), where $a_i = 0$ or 1, depending on the scores obtained when the person is the same (0) or different (1). Right figure: outcome of PAVA. This curve is the nondecreasing curve which best fits the set of scores in the left figure.

fied for every i, as shown in Fig. 2.9. The next step is applying the pool adjacent violators algorithm (PAVA) making sure the requirement $y_i + 1 \geq y_i$ is satisfied. Going from the smallest x value in increasing order, if a violation of this requirement is encountered, the value of the point $y_i + 1$ (the violator) and the left adjacent points with the same y value are changed to the average of all of them, considering the assigned weights. With that, the decrease in the function is avoided at this point, augmenting the value of the violator and decreasing the value of the adjacent left points. However, after this step, it is possible that a new violator to the left of x_i has been created. It is for that reason that after a change in the value it is required to start from the smallest value of x again. The algorithm ends when all the violators are eliminated, that is, the obtained points define a nondecreasing function as shown in Fig. 2.8 right.

The resulting function can be considered an estimation of the probability of the comparison being two different persons, conditioned on a distance value or evidence. Also defined as following:

$$y(x) = P(BSV|x). \tag{2.7}$$

Hence, its complementary value to 1 corresponds to the probability that the two people in the comparison are the same person, conditioned to a distance value (or evidence):

$$1 - y(x) = P(WSV|x) \tag{2.8}$$

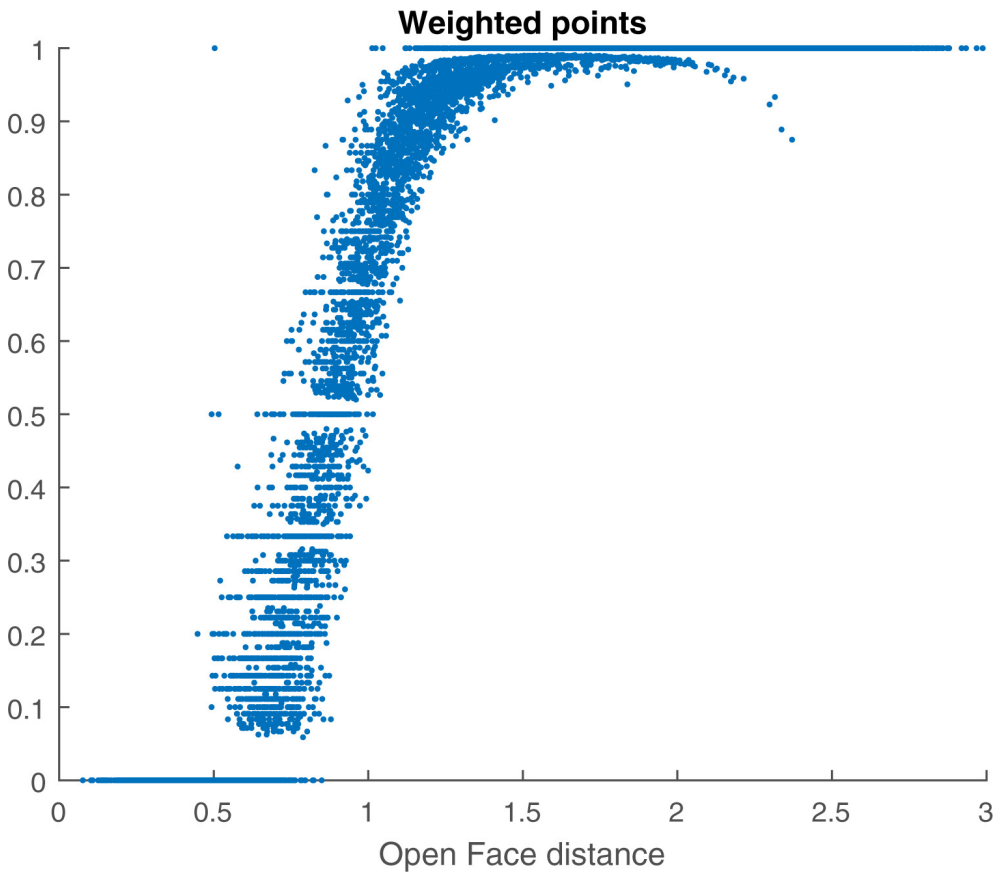


Figure 2.9: Point cloud with $x_i + 1 > x_i$ satisfied. Data points indicated in Fig. 2.8 left are not suitable for the PAVA. Points with the same value in the x-axis are substituted by a single weighted point. The result is a cloud of points, all of them with different x values.

The division of the two returns the LR *:

$$LR_{ISO}(s) * \frac{P(H_p)}{P(H_d)} = \frac{1 - y(s)}{y(s)} \tag{2.9}$$

Dataset for Calibration Data

To perform the actual calibration, a large dataset is needed from which we can learn the required probability functions. We do so by employing the Labeled Faces in the Wild database [40]. This is a database of face photographs designed for studying the problem of unconstrained face recognition. The dataset contains more than 13,000 images of faces collected from the Web. Each face has been labeled with the identity of the person pictured. 1680 of the people pictured have two or more distinct photographs in the dataset (13). It is widely used as a benchmark for face recognition performance. With this dataset, two sets of image pairs are generated: pairs of the same person (WSV) and a different person (BSV). Around 137,000 comparisons were performed in this dataset to achieve the calibration test.

2.3.3 Comparing ENFSI Investigators and Automated Systems

Correlation Between Automated Systems and Investigators

We now move to the comparison of the automated system and the human expert. This comparison is done with the Spearman correlation coefficient (referred to as rank correlation from now on). A graphical description of this comparison can be seen in Fig. 2.10.

The correlation between the n-dimensional vector LLR (logarithmic likelihood ratio) given by an investigator (x) and the vector LLR computed by the software (y) is as follows:

$$\rho_{xy} = 1 - \frac{6 \sum d^2}{\sqrt{n \sum (n^2 - 1)}} \tag{2.10}$$

where d is the difference between the ranks of the two vectors, and n is the length of each vector. The possible values of this coefficient go from -1 (opposing criteria between the investigator and the automated system) to $+1$, which expresses perfect concordance of criteria. A value of 0 means no relation between them or randomness. We use the LLR due to the nature of the ENFSI tests, in which the investigators provide LLR instead of LR. For automated systems, the LLR is computed using the values in Table 2.2.

*Given equal numbers of match/mismatch pairs, the prior probabilities ratio $\frac{P(H_p)}{P(H_d)}$ equals 1.

Confusion Matrix

To get insight into the performance of a set of results, being it from an investigator or an automatic system, we use a confusion matrix. The following terms play a role here:

TP: true positives —the number of cases where both images are considered belonging to the same person and it was a match.

FP: false positives —the number of cases where both images are considered belonging to the same person and it was not a match.

TN: true negatives —the number of cases where both images are considered belonging to different persons and it was not a match.

FN: false negatives — the number of cases where both images are considered belonging to different persons and it was a match.

		Actual	
		WSV	BSV
Prediction	Same person	TP	FP
	Different person	FN	TN

Table 2.3: Confusion Matrix

From these values, a set of other metrics can be calculated namely: **Precision** : TP/(TP + FP) **NPV** : negative predicted value = TN/(TN + FN) **Sensitivity** : TP/(TP + FN) **Specificity** : TN/(TN + FP) These values are expressed as percentages, and the classification is better when they are near to 100%.

Matthews Correlation Coefficient

Based on the confusion matrix, we can compute another measure of classification namely the Matthews correlation coefficient (MCC) given by

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}} \tag{2.11}$$

This coefficient condenses in only one value the quality of the binary classification. The absolute value of this coefficient is less or equal to 1. The higher the value, the better the classification is. A value of zero means that the classification is as good as a random one.

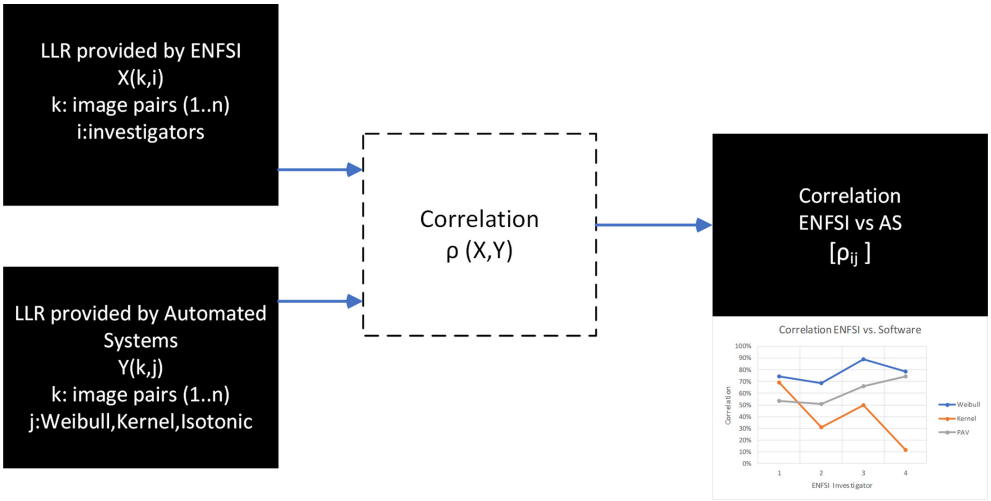


Figure 2.10: Correlation among ENFSI investigators and automated systems (AS). There are two logarithmic likelihood ratios (LLRs) obtained. First one from the forensic experts and the second one from automated systems. They are compared through a correlation, and a matrix is obtained and is represented in graphs.

Log-Likelihood Ratio Cost (C_{llr})

A final measure we consider is the log-likelihood ratio cost which is based on LR values directly [37]:

$$C_{llr} = \frac{1}{2 * N_p} \sum_{i_p} \log_2 \left(1 + \frac{1}{LR_{i_p}}\right) + \frac{1}{2 * N_d} \sum_{j_d} \log_2 \left(1 + LR_{j_d}\right), \qquad (2.12)$$

where N_p and N_d are the number of cases, H_p and H_d are true, respectively, and LR_p and LR_d are the likelihood ratios for these cases. This coefficient is always positive, and the lower the value, the better the performance of LR values is. In this chapter, C_{llr} is only used to compare calibration methods, not to compare them to forensic investigators.

2.4 Results

To present comparisons between the automated system and forensic investigators, correlation graphics and boxplots will be used. Although ROC and FAR/FRR are commonly used in literature, they do not apply to this chapter because they can only be obtained from calibration data. The data obtained from investigators are not enough for this kind of graph. We show for representation the correlation and results from ENFSI test 2011 in

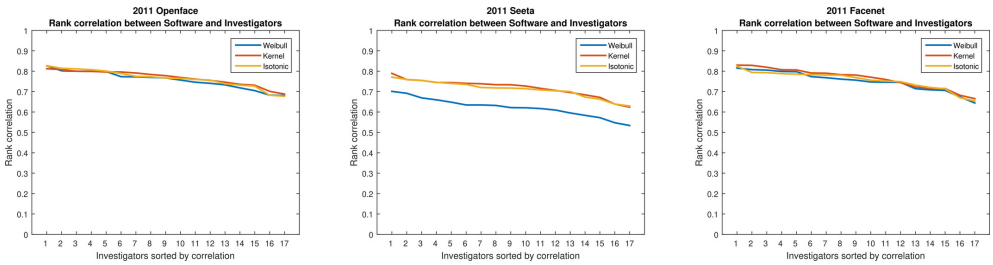


Figure 2.11: Correlation ENFSI vs automated systems, year 2011. These graphs show the correlation between each of the three scores to LR methods (Weibull, KDE, and IR) and every one of the investigators with the three types of automated system (OpenFace, SeetaFace, and FaceNet). Each figure represents one automated system: on the left, OpenFace; on the center, SeetaFace; and on the right, FaceNet. Higher values indicate higher concordance between the forensic expert and the automated software. The forensic experts are ordered from left to right according to the highest to the lowest correlation.

Fig. 2.11, and the rest of the years (2012, 2013, and 2017) are available in the annex.

2.4.1 ENFSI Test 2011

Figure 2.11, Figures S1, S3, and S5 (in the annex) show the rank correlation between each of the three scores to LR methods (Weibull, KDE, and IR) and every one of the investigators with the three types of automated system described before. They present the investigators ordered by their correlations concerning the three methods (Weibull, KDE, and IR).

Figure 2.12 (left figures, Figures S2, S4, S6) show the right (TP + TN) and wrong (FP + FN) answers of investigators (blue x) and automated systems (red triangles) and (right figures) the individual values of confusion matrix with investigators results (boxplot) and automated systems (red triangles).

For the experiments realized in the year 2011, one can see that out of the three software programs, the highest correlation is presented by FaceNet, closely followed by OpenFace. The three calibration methods have very similar results, except for Seeta, for which Weibull has less correlation than the other methods. Seeta has a higher number of wrong answers for an equivalent number of right answers to OpenFace. In OpenFace case, the most accurate method is the isotonic regression. In FaceNet, the number of correct answers significantly higher resembles the investigators. The best procedures are Weibull and KDE.

OpenFace has several right answers similar to the researchers, but more failures. The true positives of the three methods are equal to the researchers, and the true negatives are somewhat inferior. But OpenFace has more false negatives and false positives than researchers. Seeta hits all true negatives; however, it is below in the true positives. It has 0 false positives and high false negatives. FaceNet such as Seeta hits all the negatives but has

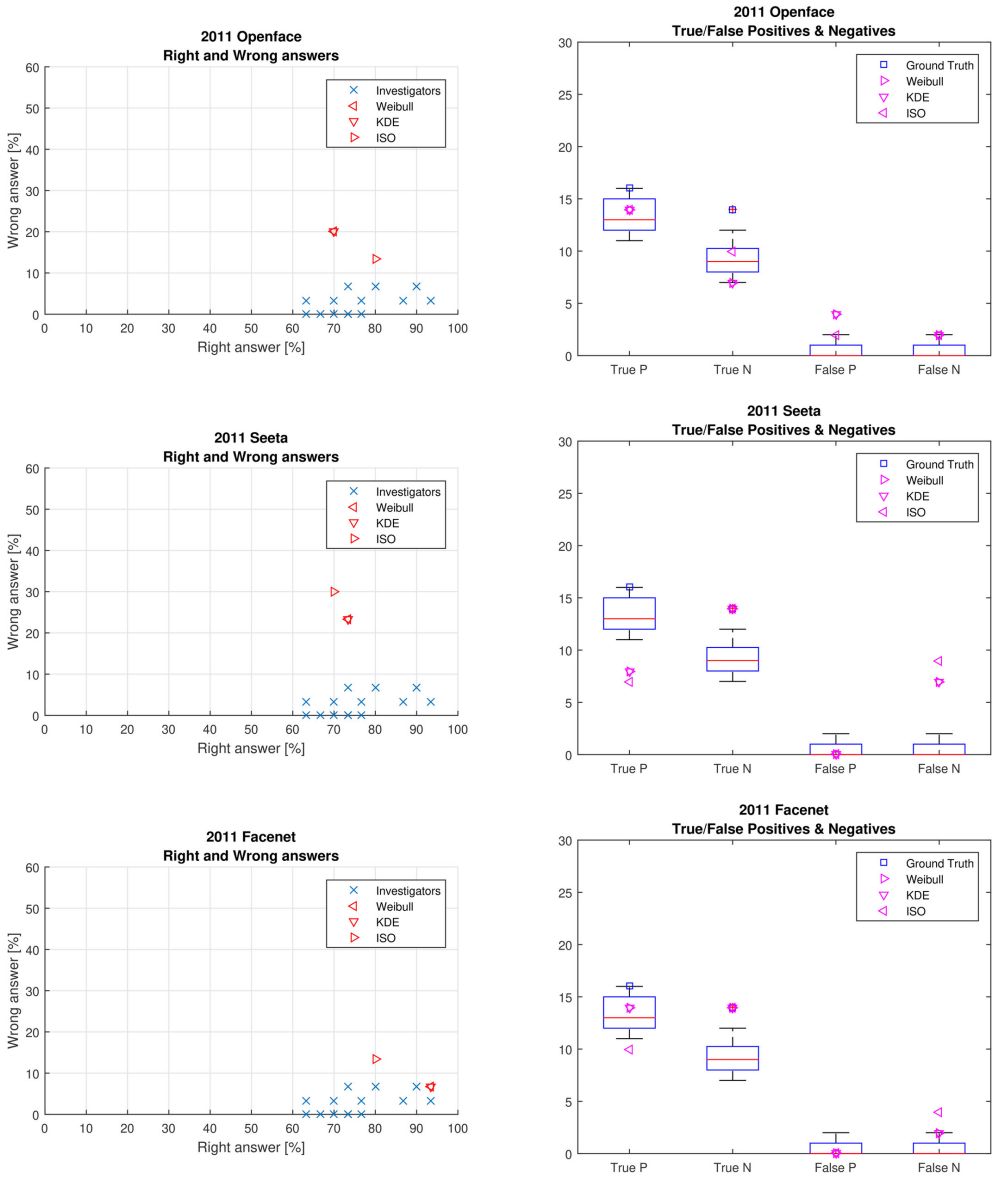


Figure 2.12: Right and wrong answers. Binary classification results. Year 2011. In the figure, the graphs are deployed as follows: Figures on the left correspond to right and wrong answers from the automated systems and the forensic experts. Crosses represent experts, and triangles, automated systems. On the right, a boxplot of the false positives, false negatives, true positives, and true negatives is shown. Boxplots are obtained from the forensic experts' data. The outcome from the three methods (Weibull, KDE, and isotonic regression) is superimposed in the same graph.

fewer false negatives. Weibull and KDE are as good as OpenFace to hit the true positives, and they also have the 3 methods 0 false positives.

2.4.2 ENFSI Test 2012

From Figure S1, one can see that the correlation between the methods and the investigators reaches negative numbers for OpenFace which indicates opposite criteria to forensic experts. With Seeta, correlation values stay positive but low. In two out of the three methods, Weibull performs better than the other two methods.

Looking at Figure S2 (left), it can be observed that OpenFace did not detect all of the faces and consequently returned few outputs (13 out of 30). The number of right answers is similar to the number of wrong answers. This software has poor quality results with images taken in different poses. Seeta performs a good number of true negatives, but it also has a high number of false negatives and low true positives. Nevertheless, investigators had a higher number of false positives. FaceNet behaves very similarly to Seeta.

In year 2012 experiments, the researchers have a great dispersion with the true negatives (Figure S2 right). Seeta and FaceNet have surpassed the researchers in the true negatives, and the three types of software have had a terrible rating in true positives, well below humans. Seeta and OpenFace have no false positives; however, they have many false negatives.

2.4.3 ENFSI Test 2013

With Figure S3, it can be noted that the correlations with FaceNet given by the three methods are very similar. However, with Seeta, Weibull calibration stands out among the other two. Correlations are higher in FaceNet than the others and in Seeta-Weibull higher than in OpenFace.

The number of right and wrong answers (Figure S4 left) with OpenFace is the same for the three density estimation methods, and similar to the ones Seeta has. For Seeta, the best density function model for calibration is Weibull. Seeta has less true positives and more false negatives compared with investigators. Nevertheless, its performance is better than investigators concerning true negatives and false positives. For FaceNet, isotonic regression results in a good number of true negatives, but a bad number of false negatives. Weibull and KDE behave similarly with a good number of false positives and negatives, and moderate numbers of true positives and negatives.

OpenFace has the highest rating in true positives, better than humans, and Seeta is the best with true negatives, also surpassing humans. OpenFace has many false positives; how-

ever, Seeta and FaceNet are at the same level as humans (Figure S4 right).

2.4.4 ENFSI Test 2017

For the year 2017 (Figure S5), Seeta calibration presents higher correlation values than other years, but FaceNet is the automated system with the best results in terms of correlation with investigators and KDE seems to be the best approximation. OpenFace has the worst results and isotonic performs better than Weibull and kernel.

The quality of results (right and wrong answers in Figure S6 left) is much better in Seeta than OpenFace with any of the three methods. The three density function estimation methods behave similarly in both Seeta and OpenFace. In FaceNet, the right answers and wrong answers are similar to Seeta with Weibull being the best option.

FaceNet using the Weibull and KDE methods is the one method with the highest number of true positives, equal to the majority of the researchers (median). However, the true negatives have been detected by Seeta very well and OpenFace very badly. While Seeta does not have any false positives, FaceNet and above all, OpenFace has many more than researchers as can be seen in Figure S6 right.

In conclusion, in all the tested years (2011, 2012, 2013, and 2017), the method that performs better is not always the same and it depends on the quality and poses of the images.

2.4.5 Confusion Matrix and MCC Results

A summary of the findings can be seen in the following Tables 3–5. From them, we can see that the quality of classification by the investigators is better than the one by the automated systems.

2.5 Discussion

When we compare images taken in frontal poses and lateral poses, the best results with all the automated systems are obtained when poses are frontal. The three automated system softwares give more incorrect answers when pose is lateral (45 Yaw, with a slight pitch ("from above") or with the time difference (age) between reference and questioned images). When the pose is 90° yaw, the software is unable to detect the face and returns an empty answer. To detect the face, the currently used software looks for two eyes, and this is not possible with a profile image.

With lateral poses, the correlation between software and human detection is random, it contains positive and negative values, and the software returns about 50% of wrong re-

Table 2.4: Confusion matrix values for OpenFace.

Openface	Metric	Weibull (%)	Kernel (%)	ISO (%)	ENFSI (%)
2017	Precision	68	71	79	96
	Negative predicted value	86	78	73	96
	Sensitivity	93	86	79	98
	Specificity	50	58	73	93
	Matthews correlation coefficient	48	46	51	91
	C_{llr}	99	97	82	-
2013	Precision	82	85	83	89
	Negative predicted value	91	83	82	100
	Sensitivity	95	89	83	100
	Specificity	71	77	82	86
	Matthews correlation coefficient	69	67	66	87
	C_{llr}	80	82	138	-
2012	Precision	67	100	100	84
	Negative predicted value	50	56	50	81
	Sensitivity	33	33	17	83
	Specificity	80	100	100	82
	Matthews correlation coefficient	15	43	29	65
	C_{llr}	138	123	175	-
2011	Precision	100	100	100	97
	Negative predicted value	86	87	70	95
	Sensitivity	83	83	57	96
	Specificity	100	100	100	96
	Matthews correlation coefficient	85	85	63	92
	C_{llr}	53	51	66	-

Table 2.5: Confusion matrix values for Seeta.

Seeta	Metric	Weibull (%)	Kernel (%)	ISO (%)	ENFSI (%)
2017	Precision	100	100	100	96
	Negative predicted value	75	71	70	96
	Sensitivity	79	75	70	98
	Specificity	100	100	100	93
	Matthews correlation coefficient	77	73	70	91
	C_{llr}	59	61	84	-
2013	Precision	93	93	100	89
	Negative predicted value	77	71	69	93
	Sensitivity	74	65	53	95
	Specificity	94	94	100	86
	Matthews correlation coefficient	69	62	61	82
	C_{llr}	65	75	109	-
2012	Precision	100	100	100	84
	Negative predicted value	61	59	56	81
	Sensitivity	25	18	15	83
	Specificity	100	100	100	82
	Matthews correlation coefficient	39	33	29	65
	C_{llr}	247	143	211	-
2011	Precision	100	100	100	97
	Negative predicted value	67	67	61	95
	Sensitivity	53	53	44	96
	Specificity	100	100	100	96
	Matthews correlation coefficient	60	60	52	92
	C_{llr}	167	98	154	-

Table 2.6: Confusion matrix values for FaceNet.

FaceNet	Metric	Weibull (%)	Kernel (%)	ISO (%)	ENFSI (%)
2017	Precision	83	83	100	96
	Negative predicted value	100	89	86	96
	Sensitivity	100	95	87	98
	Specificity	67	67	100	93
	Matthews correlation coefficient	75	67	86	91
	C_{llr}	59	58	50	-
2013	Precision	88	88	100	89
	Negative predicted value	92	86	72	93
	Sensitivity	94	88	53	95
	Specificity	85	86	100	86
	Matthews correlation coefficient	79	74	62	82
	C_{llr}	60	57	76	-
2012	Precision	83	83	100	84
	Negative predicted value	60	62	54	81
	Sensitivity	38	38	7	83
	Specificity	92	93	100	82
	Matthews correlation coefficient	37	38	20	65
	C_{llr}	164	138	163	-
2011	Precision	100	100	100	97
	Negative predicted value	88	88	78	95
	Sensitivity	88	88	71	96
	Specificity	100	100	100	96
	Matthews correlation coefficient	88	88	75	92
	C_{llr}	58	43	60	-

sponses, being isotonic regression the method with best results. Forensic experts provide better results in these cases but they also present low correlation among them, values about 0.4 which means that they present difficulties to take decisions and their criteria are different.

When the comparison is made only with frontal poses, the correlation between forensic experts and software is better. When the quality of questioned images is high, forensic experts have much better results (correct answers) and high values of correlation among them (greater than 90% in many cases). In this case, software methods give as many right answers as to when then image quality is low or decent. The method with best results and correlation is Weibull but with no significant difference with respect to the others. So, for frontal poses and low-quality images, the software systems are at the same level as forensic experts, but when the quality of images is good, the experts obtain better results. We conjecture that automatic systems are not able to take advantage of little details such as scars and freckles but, at the same time, are not sensitive to occlusions of the face by glasses, hats, or microphones.

To perform the calibration, the LFW database was used, which is unrelated to the ENFSI tests. LFW is large but may be biased due to most of the images being high quality and a lot of them frontal images. That gives room to better results in the LR obtained computed with scores in the case of fully frontal comparisons. Another public dataset, SC Faces was tested but offered similar results as LFW. To check that a large unrelated database provides better results than a small biased one, another experiment was made. The ENFSI tests were not used only as a test, but also as the mean to transform scores to LR. The number of comparisons was significantly reduced due to the number of pictures available (from 130,000 comparisons in LFW to 50 in ENFSI tests) resulting in score sets that are difficult to fit with a function. Hence, the LR computed using the ENFSI report as a data generator provides worse results than using a big, unrelated database. We could conclude that it is better to use a large unrelated dataset to the case material than to calibrate the system in data that are closer to the case material but biased. As proven by the results, the machine behaves more similarly to the forensic expert if the calibration dataset is large and unrelated to the test data than if it is of the same characteristics of the test data but a small number of images to calibrate. This can be seen in Fig. 2.13. The left graph corresponds to the results of a calibration computed with the ENFSI tests themselves for the year 2013 (few samples for both WSV and BSV), whereas in the right there are the results for calibration made with the LFW dataset. The difference is over 10 % of more right answers in the right graph than on the left.

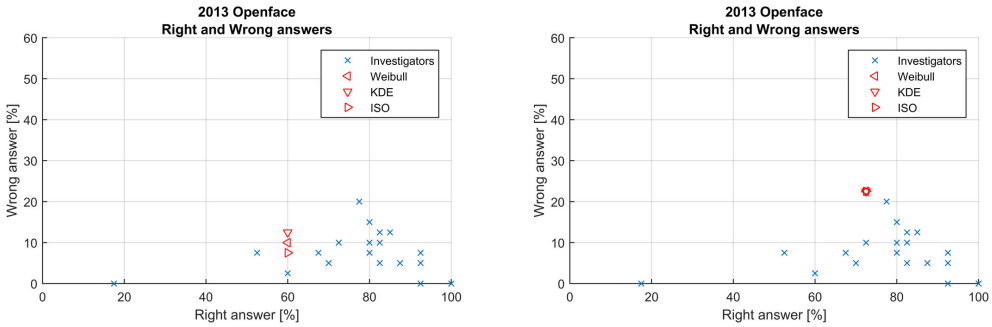


Figure 2.13: Right and wrong answers. Binary classification results. Year 2011. In the figure, the graphs are deployed as follows: Figures on the left correspond to right and wrong answers from the automated systems and the forensic experts. Crosses represent experts, and triangles, automated systems. On the right, a boxplot of the false positives, false negatives, true positives, and true negatives is shown. Boxplots are obtained from the forensic experts' data. The outcome from the three methods (Weibull, KDE, and isotonic regression) is superimposed in the same graph.

2.6 Conclusion

Observing the results obtained after comparing proficiency tests and likelihood ratios calculated from the scores provided by OpenFace, Seeta, and FaceNet, one can say the software can assist reporting officers, as it can do faster and more reliable comparisons with full-frontal images. Although the software presents limitations, these should not dictate what is feasible in terms of interpretation. It is expected that algorithms will evolve to adapt to all types of profiles and increase their performance. We have to think about it as a tool, never as a constraint to limit its usage. The expert cannot be replaced by this tool, but becomes more efficient, because the computer can help to reduce the amount of info to be managed doing appropriate filtering. If two independent experts conduct face comparison doing the comparison independently, the third might be an algorithm, and the experts can evaluate their findings and the findings of the algorithm to draw a conclusion. Due to the high accuracy of the automated systems in the full-frontal images, it makes this kind of open-source system especially adequate to full-frontal images comparison, such as an ID picture to a mugshot, which can be useful to forensic experts.

3

Calibration of Score based Likelihood Ratio estimation in automated forensic facial image comparison

Forensic facial image comparison lacks a methodological standardization and empirical validation. We aim to address this problem by assessing the potential of machine learning to support the human expert in the courtroom. To yield valid evidence in court, decision making systems for facial image comparison should not only be accurate, they should also provide a calibrated confidence measure. This confidence is best conveyed using a score-based likelihood ratio. In this study we compare the performance of different calibrations for such scores. The score, either a distance or a similarity, is converted to a likelihood ratio using three types of calibration following similar techniques as applied in forensic fields such as speaker comparison and DNA matching, but which have not yet been tested in facial image comparison. The calibration types tested are: naive, quality score based on typicality, and feature-based. As transparency is essential in forensics, we focus on state-of-the-art open software and study their power compared to a state-of-the-art commercial system. With the European Network of Forensic Science Institutes (ENFSI) Proficiency tests as benchmark, calibration results on three public databases namely Labeled Faces in the Wild, SC Face and ForenFace show that both quality score and feature based calibration outperform naive calibration. Overall, the commercial system outperforms open software when evaluating these Likelihood Ratios. In general, we

conclude that calibration implemented before likelihood ratio estimation is recommended. Furthermore, in terms of performance the commercial system is preferred over open software. As open software is more transparent, more research on open software is urged for.

3

3.1 Introduction

When face images are presented as evidence in court, the target most often is to interpret the result of the comparison between trace and suspect images. No standard method is, however, available for that task. The comparison technique, whether it is performed manually or using an automatic system, must meet legal requirements which vary per country [6; 37; 69]. Although the use of automatic systems is increasingly studied in the field of facial image comparison, for legal deployment it lacks standardization and validation. This is one of the reasons why cases of facial image comparison in court are currently still carried out manually by specialized facial image comparison experts [6; 37]. Having a unified and validated method for interpreting scores by experts and machine can provide the standardization needed in court.

The Likelihood Ratio (LR) comes as a possible solution [70; 71] for standardization, expressing the decision as the ratio of the probability given the evidence of a match against the probability of a non-match. Forensic experts endorse its use due to its compliance with the requirements of evidence-based forensic science: it is scientifically sound in particular it has transparent procedures, is testable, and it clearly separates the responsibilities of the forensic examiner and the court [72; 73]. For evidence in speaker recognition, fingerprints and DNA analysis, a distance or similarity based biometric Score likelihood ratio (SLR) is being studied and used [74–76]. Here, we aim to realize a similar approach for facial image comparison. As explored in chapter 2 and [36], automated systems for facial image comparison (especially when based on deep learning) combined with score-based likelihood ratio estimation have a great potential to help the forensic expert in the evaluative process [6].

In this chapter, we make a number of contributions. We develop a pipeline that given a score produces an LR estimation that can be compared to forensic experts and ENFSI participants. This serves as an SLR evaluation and validation for both open and commercial software. Thus we explore their differences and determine whether there is room for improvement in open software automated systems. Secondly, we estimate the influence of different LR calibrations in relation to resolution and image features, based mainly on

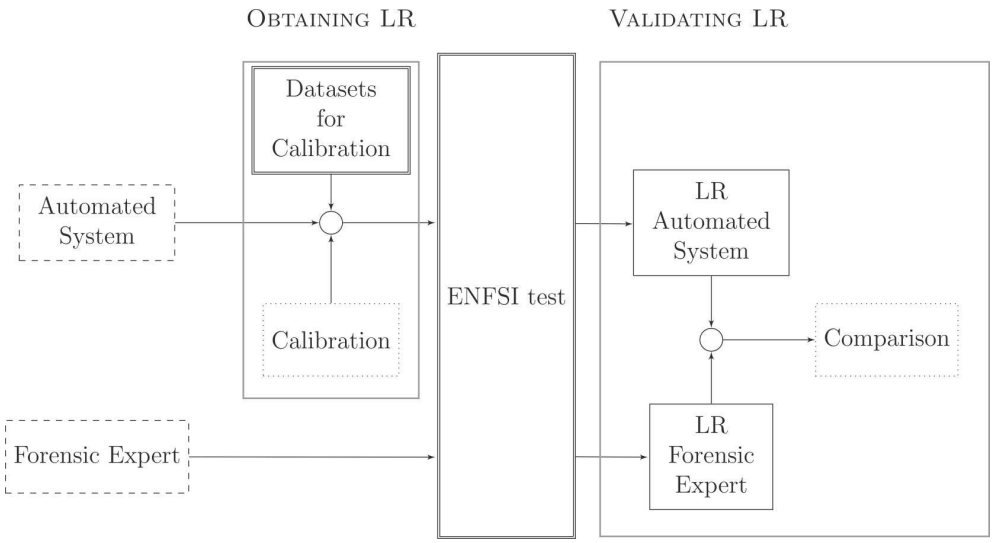


Figure 3.1: Overview of the main topics addressed in this chapter. Dashed boxes correspond to evaluating agents. Dotted boxes represent operations and double framed boxes correspond to data.

surveillance images which is a major source of evidence in forensic cases. Calibration has been researched and used in speaker comparison [75; 77] for similar types of voices. As identified in [6; 36] similar treatment in faces has yet to be researched. Thirdly, we compare the Likelihood Ratio estimation from both open software as well as commercial software to a set of forensic experts in the ENFSI Proficiency Face Comparison test (which include case work related images such as surveillance) using the statistic elements of Cost Log Likelihood Ratio (C_{llr}) [78; 79].

Figure 3.1 gives an overview of the main topics presented in this chapter.

3.2 Related work

We study related work by first considering how likelihood ratios are used in forensic fields other than facial image comparison. From there we consider how facial image comparison is currently being done. Finally, we look at the core step in standardization namely the calibration.

3.2.1 Likelihood Ratio in Forensics

Using a Bayesian probabilistic framework has been proposed in recent years as a logical and appropriate way to report evidence to a court of law [73; 80; 81]. The work of [82] states the requirements of evidence-based forensic science, which are: adoption of a basic-

research model, design of experiments that test said model and the ability of experts to inform court about the relative strengths and weaknesses and suggestion on how that knowledge applies to individual cases. They also recommend that for machine learning data should be collected based on the frequency with which markings and attribute variations occur in different populations. The Likelihood Ratio has been proposed in recent decades as a method which addresses these requirements by providing transparent procedures and being testable, as indicated in the introduction [73; 83]. When computed for a certain benchmark, different methods such as C_{llr} and ECE can be used to assess its predictions, see section 3.2.3 for more information about these methods. Score based procedures for the calculation of forensic likelihood ratios are popular across different branches of forensic science [84] especially in DNA [85], and speaker comparison [74; 75; 77]. They have two stages, first a function or model which takes measured features from known-source and questioned-source pairs as input and calculates scores as output, then a subsequent model which converts scores to likelihood ratios [84]. LR based on biometric similarity scores is referred to as Score based Likelihood Ratio (SLR) and defined as:

$$SLR = \frac{P(s|H_p, I)}{P(s|H_d, I)}, \tag{3.1}$$

where H_p is the null hypothesis or the prosecution hypothesis (evidence originates from the same source) and H_d is the alternative hypothesis or defense hypothesis (evidence originates from a different source). The value s is the score returned by the biometric system and I is the background information available in the case apart from the evidence. Although LR can be used for any type of forensic evidence (such as DNA or fingerprints), in our work it corresponds to face evidence.

According to [86], efforts to model or compensate the effects of adverse conditions in likelihood ratio computation should be improved. They evaluate the impact of these adverse conditions on glass samples. The analysis of [86] shows that integration of advanced machine-learning algorithms for the compensation of adverse conditions into forensic evaluation helps in this direction. They find this impact greatly affects calibration performance. There is a lack of a similar study in case of facial image comparison.

In [23] and [70], different LR validation methods are explored and analysed. The first question to consider is what and how to validate? In both papers, Cost Log Likelihood Ratio and ECE plot validation [73; 78] are proposed as promising characteristics. ECE is exposed in [87] as a method which measures both discrimination and calibration, and shows its potential. It also describes how other related measures such as Confusion Entropy

(CEN) or Matthews Correlation Coefficient (MCC) work with decision errors rather than probabilities. This implies the selection of a threshold and therefore they do not consider performance at different prior probabilities either. Other metrics are considered in [6], such as Tipett plots, Detection error trade-off (DET) and equal error rate (EER). They present an overview table summarizing the use and adequacy of these metrics for the assessment of model performance. In this overview, the graphical representation that scores the highest for both discrimination and calibration is again the ECE plot. In conclusion, for this work and according to the studied literature, the best indicators of both discrimination and calibration performances are C_{llr} and the ECE plots (explained in 3.3.3) [23; 71] which give a good view of both the calibration and discrimination power of the forensic experts and the automated systems.

3.2.2 Facial image comparisonin Forensics

Facial image comparison in Forensics has been largely studied from a manual point of view [83]. There have been tentative approaches on automated systems performing this task, whether for intelligence, investigation, or evaluative purposes Zeinstra et al. [8]; Ali [36]; Tistarelli and Champod [37]. And facial image comparison has proven to have potential to help the forensic expert if the likelihood ratio estimation method is properly standardized and validated (chapter 2). In manual comparison, four methods are typically used to analyse and compare faces: holistic, morphological and photo-anthropometric processes, along with direct superposition of the images [83].

These methods are not exclusive and can be combined in order to carry out the most exhaustive analysis with regard to the information available on the image. Recommendations in ENFSI practices are: out of these four methods, holistic comparison is only recommended when other more effective methods are not available, morphological (feature comparisons) is useful and recommended for facial image comparison. Both photo-anthropometric comparison and superposition are not recommended when using uncontrolled imagery.

Current face recognition systems [46; 88; 89], already reach very high levels of accuracy in public non-forensic benchmarks, and it is expected that in the coming years they will keep improving. If this improvement is accompanied with a standardization and proper validation in their decisions, they could become a powerful tool in Forensic Science [70]. An enforcement of this idea can be found in [6], where there is an extensive survey on the role of these automated system nowadays in the forensic field. They propose to improve the discussion between forensic expert, investigators and legal practitioners to best develop

this method with respect to the needs and constraints of each.

3.2.3 Calibration in Forensics

The calibration state of a model refers to the closeness of the computed value to the known value. Therefore, the calibration measures the extent to which the SLR points towards the correct proposition. It has been used in other fields of Forensic Science such as speaker comparison, DNA analysis or fingerprints [23; 73; 78]. In [71], the problem of incorrect selection of databases is put forward. This problem is tackled in [90] for the speaker comparison case. It discusses what should be the implications of a good calibration and proposes ECE as the preferred method of validation. For facial image comparison this implies that ECE methods for evaluation are adequate for detection if the performance of both the automated model and the forensic participant are affected in the same way by the chosen calibration population.

In literature, the term "calibration" is used to describe two different processes. It usually refers to SLR as described in the introduction, or it can more specifically point to the subsequent process to adapt models which have high discriminating power but are poorly calibrated [23]. As this second step is essential to enhance the overall performance of a model, [6] poses that the term "calibration" should not differentiate between the steps of score-to-SLR and SLR-to-calibrated SLR. Instead methods should cover every computation used from the initial score to the final reported SLR regardless of the number of treatment steps needed. In this work, we evaluate the effects of selecting the database to perform said calibration, for which the second step is not required. We evaluate the first interpretation of the term, so score to likelihood ratio with no subsequent computations, as they do in the work of [23; 73].

In [75], calibration on information extracted from speech is explored. It addresses the main issues in calibrating data: limited training data and dataset shift when score distributions change between calibration and test sets. Calibration in speaker recognition is based on features namely duration of audio, distance, language, and gender [75]. The work of [77] studies the impact in forensic voice comparison of lack of calibration and of mismatched conditions between the known-speaker recording and the relevant-population sample recordings.

A problem that could arise when calibrating [91; 92] is data scarcity. The references indicate that the use of simulated data gives a big improvement in data scarcity situations, but the testing of the validity of simulated databases for the operational use of systems in

a real setup is still controversial. For face images this would imply that a solution for data scarcity could be Generative Adversarial Networks (GAN) that generate realistic fake face images [93]. However, forensic implications of simulating face data should be evaluated.

3.3 Materials and Methods

We aim for the validation of automated facial image comparison systems computing an SLR. Referring again to figure 3.1, this validation has two parts. First is the SLR system itself, which consists of a scorer and a calibrator. In this case the scorer is the biometric system, i.e the facial image comparison automated system that will return either a distance or a similarity which will be treated as a score. The other element, the calibrator, will take a set of scores that either correspond to a group of facial images of comparisons within the same person (within source variability or WSV) or comparisons in a set of face images amongst different persons (between source variability or BSV). Having a set with different people, each of them with several images of themselves and using the two sets of comparisons defined, a likelihood ratio can be estimated. Once the SLR is obtained, it must be calibrated. A well calibrated LR will be accurate with its own predictions [71]. In the final step, LR estimation will be validated. This validation is done using three measures namely Cost Log Likelihood Ratio (C_{llr}) [78; 79]. Minimum Cost Log Likelihood Ratio (C_{llr} min) and Empirical Cross-Entropy (ECE) [23; 71] and compared to experts that have estimated a likelihood ratio for a series of tests issued each year [94–100].

3.3.1 Materials

Calibration of datasets: LFW, SC Face and ForenFace

The Likelihood Ratio is the ratio of two probabilities. As the probability functions of WSV and BSV are unknown, it is necessary to obtain them empirically. Using the scorer to generate multiple intermediate scores of both populations in which the ground truth is known, histograms can be computed. Subsequently, the histograms are approximated with probability functions through different methods, namely Isotonic Regression [101], Kernel Density Estimation [102] and Logistic Regression [103]. There has been some discussion on which type of datasets are optimal for calibration [73; 77], where there are some recommendations such as defining the WSV set with pairs that are highly similar (small distance between their embeddings) or choose a WSV set with the same features as the comparison at hand. The discrimination is robust independently from the dataset the system was

calibrated with, but calibration itself is highly dependent on the conditions [76]. In particular, in the work of [76] the effect of duration, distance, language, and gender in speaker comparison by using a variety of datasets makes a difference in the calibration results. Intuitively, the higher the number of comparisons and the more similar the dataset is to the tested data, the better the calibration will be. In our setting the datasets used, in which surveillance images predominate to be compliant with the forensic nature of the tests, are described in table 5.1[29; 104; 105].

3.3.2 Methods

Obtaining the SLR

Following similar procedures as in DNA and speaker comparison [74; 75] and other face recognition works in forensics such as [36; 106], the score obtained when comparing two faces is transformed to a Likelihood Ratio. Although the process of calibration has been studied and analysed in speaker comparison works such as [77] or [76], [36] and in facial image comparison in [106] those studies in facial image comparison did not take into account how different calibration characteristics such as features affect the results. It is for that reason that in this work we select different calibration types based on the work of speaker comparison and test them against ENFSI tests. The following section gives details on how this process is carried out.

The Scorer

The scorer is the system or person whose goal is to provide an estimation of a Likelihood Ratio, possibly through the intermediate determination of a distance or similarity. This scorer can e.g. be a pre-trained neural network which is calibrated so the intermediate score can be transformed to a Likelihood Ratio or a forensic expert who directly provides an estimated likelihood ratio based on the visual comparison of the face features [83]. The scorers used in this work are as follows:

Automated system The scorer compares two facial images and returns either a distance or a similarity as intermediate score. The scores group in two sets. As mentioned in 3.3.2, the first set is for estimating WSV in which two images corresponding to the same person are compared and the second set in which the comparisons correspond to different persons for estimating the BSV. Our open-source scorer uses Deepface state-of-the-art face recognition built with Deep Learning [107]. According to [107], the supported models FaceNet-512 got 99.65%; ArcFace got 99.41%;

Table 3.1: Dataset description

Dataset	Number of images	Characteristics						
		Age	Gender	Ethnicity	Occlusions	Pose	Illumination	Resolution
Labeled Faces in the Wild (LFW)	13233 images 3233 labeled 5749 subjects	Mostly adults Few children	Mostly men	Mostly Caucasian	Not many	Mild	Acceptable	Mostly high
SC Face	4160 images 130 subjects	From 20 to 75	115 males 15 females	All Caucasians	Beard Moustache Glasses	from -90 to +90 in equal steps of 22.5 degrees	Uncontrolled	Variable
ForenFace	Mostly videos 97 subjects	Contains both facial images and videos	Mixed	Mostly Caucasian	Caps Beards Glasses	Different	Acceptable	High Low

Dlib got 99.38% and VGG-Face got 98.78%; accuracy scores on Labeled Faces in the Wild benchmark whereas human beings could have just 97.53%. The commercial automated system we use is FaceVACS version 5.5.2 [108] from Cognitec. Ths commercial system only provides the final similarity score between two facial images. Open software exposes the architecture and weights that output the representation of each of the facial images in the n-dimensional space, which gives flexibility for tasks such as clustering or comparison. Also, open software allows to change the method to compute the similarity score between facial images. While the similarity score of Cognitec is a number between 0 and 1, but not disclosed how it is exactly computed, open software has different distance functions such as euclidean distance or cosine similarity which can be computed and compared.

Forensic expert The forensic participants are members of the European Network of Forensic Science Institutes (ENFSI). Each year, a Proficiency Face Recognition test is distributed among laboratories within the organization and experts can assess which factors affect face recognition and their own assessments on Likelihood Ratio estimation. The manual forensic facial comparison process is a pair by pair comparison in which the experts estimate the likelihood ratio based on facial image features. The experts use a structured method to reach matching/non-matching conclusions for an image pair.

Calibration

As mentioned, calibration is the process of obtaining a Likelihood Ratio from a score. Likelihood Ratio is defined in section 3.2.1.

Now, there are two questions that need to be addressed according to similar studies where Score-based Likelihood Ratio is used for comparison assessment. First, which images to use for calibration? The whole dataset or just a subset having the most relevant features? Second, how to model the WSV and BSV distributions given the available data [6; 37; 73]?. Given that the performance of facial image comparison highly depends on the quality of the data that a model is built with, the author in [109] suggests to use images having similar conditions to the real life facial image comparisons. Regarding the BSV modelling, [36] uses what is known as "pseudo-traces", that is using several pictures of the reference individual in the comparison instead of generic pictures of the same person not related to the case at hand. In their results, on average 59,2% of the cases using this approach were more effective than the generic approach. In the case of BSV, no modeling

other than generic between-source comparisons has been done [36; 106]. However, taking this approach is paramount due to the importance of choosing the relevant population to obtain a suitable $P(E|H_d, I)$. According to [6], no study has yet shown the impact of variations in the choice of the relevant population for automatic face recognition. Moreover, in speaker comparison, in works such as [76], they calibrate according to divisions of the dataset with the same features, e.g. age or gender. It is for that reason, that in this chapter, three types of BSV calibration were carried out attending the methods practiced in other Forensic disciplines.

Naive calibration SC Face and ForenFace datasets image pairs were used indistinctly. In this dataset, no filters according to scores or features (as done in [86; 90]) were applied when choosing the pairs for both WSV and BSV distributions. This approach is considered the "generic" approach.

Quality Score calibration This type of calibration is an attempt of detecting how rare or frequent it is to find a face similar to the suspect's face in the relevant population, also known as "typicality". The calibration is performed in the following way: first, each image of the SC Face and ForenFace Dataset is compared against 1000 randomly chosen images from Labeled Faces in the Wild. As all the identities in ForenFace and SC Face with respect to Labeled faces in the Wild correspond to a different person, all the scores obtained will belong to the BSV distribution. What we will call a "Quality Score" is the average of the ten highest score mismatches from both SCFace and ForenFace with respect to Labeled faces in the wild. The higher that score, said face (from either SCFace or ForenFace) is more easily confused against a "standard" dataset (LFW) than another image with a lower score. This "Quality score" will be used to create different sets of calibration BSV corresponding to the Quality Score of the compared test faces. In other words, later in the validation part of the pipeline, faces will be compared in pairs. Each image of these pairs will be contrasted against LFW and a quality Score will be assigned to said test pair. Then this pair will only be calibrated with images having the same "Quality Score" For example, a test pair with "Quality Score" of 7 and 8 respectively, will generate a BSV in which the comparison scores have been obtained with calibration pairs that are also a 7 and 8 in "Quality Scores".

Feature calibration For this type of calibration, more intuitive than the former, all images in the test pairs were labelled according to if they contain headgear, beard, glasses,

yaw, pitch, resolution or other occlusions. The databases SCFace and ForenFace have already this type of labeling so the BSV population was generated with only the images that presented the same features as the test images.

3

Regarding the WSV population, [36] uses images from the same subject as the test pair to generate the test WSV population. but in our work, the usecase is that only one image of the suspect is available, as the suspect is not yet convicted. This is the case presented in the ENFSI tests used to evaluate. Because of that, a generic approach was taken by generating the same WSV for each calibration using pairs from the databases LFW, SCFace and ForenFace with the same identity.

To obtain the Likelihood Ratio from a score, in this chapter we follow three types of statistical methods to fit the WSV and BSV distributions. Three calibration methods were evaluated, Isotonic Regression, Kernel Density Estimation and Logistic Regression. They were chosen as two non-parametric (Isotonic regression and KDE), and one parametric (Logistic Regression) method. The Logistic Regression was chosen in the first place because it can assume the characteristics of many different types of distributions. It is flexible enough to model a variety of datasets. It can adapt to both skewed data and symmetric data. It is a parametric distribution, which assumes parameters (defining properties) of the population distribution from which the calibration data are drawn. Because of that, the second choice is a kernel density estimation (KDE), which is a non parametric test that does not make such assumptions. The third method chosen is Isotonic regression commonly used in machine learning models for statistical inference. The choice of one method or another doesn't seem to have a correlation with the performance of the different models of Likelihood Ratio estimation. The software used for calibration computations was from [110].

Isotonic regression a Free-form linear model that can be fit to sequences of observations [101] and then used for prediction. A common algorithm to obtain the isotonic regression is pool-adjacent-violators algorithm (PAVA). If we have the data

$$(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R},$$

isotonic regression looks for $\beta_1, \ldots, \beta_n \in \mathbb{R}$ such that the β_i approximate the y_i while being monotonically non-decreasing.

$$\underset{\beta_1,\ldots,\beta_n}{\text{minimize}} \quad \sum_{i=1}^{n}(y_i - \beta_i)^2 \tag{3.2}$$

For the Likelihood Ratio estimation, x_i represents the score and $y_i = 0$ if it is a mismatch or $y_i = 1$ if the pair comparison is a match. Applying the PAVA algorithm, proceeds as follows: going from low values of x_i to high values of x_i, we set $\beta_i = y_i$. If this causes a violation of monotonicity ($\beta_i = y_i < y_{i-1} = \beta_{i-1}$), replace both β_i and β_{i-1} with the mean $\frac{y_{i-1}+y_1}{2}$. This could result in earlier violations. If this happens, we average β_{i-1} and β_{i-2}.

KDE Kernel Density Estimation is a non-parametric density estimator. It is an algorithm which seeks to model the probability distribution that generated a dataset [102]. To fit this distribution, it makes use of two parameters, which are the kernel, which specifies the shape of the distribution placed at each point, and the kernel bandwidth, which controls the size of the kernel at each point.

Logistic regression models the probability of a certain class or event existing [103]. Logistic Regression is used when the dependent variable(target) is categorical. The dependent variable is a binary variable that contains data coded as 1 (match) or 0 (mismatch). In other words, in this chapter, the logistic regression model predicts the probability of match given a score $P(Y = 1)$ as a function of X.

3.3.3 Validating LR

The validation (see section 3.2.3) for Likelihood Ratio assessments has been discussed in [77; 79]. There three metrics are introduced that consider not only if the decision taken by the automated system was correct, but also penalizes if the system provides an inconclusive answer. The metrics are C_{llr}, C_{llr} Min and ECE plot [23; 111]. Compared to equal error rate or ROC curves, these metrics provide a better representation of both the discrimination power of the model and its calibration performance. These metrics can be used to evaluate any set of Likelihood Ratio estimations, both for the automated systems and the forensic experts. In this chapter we will use them to evaluate their performance on the ENFSI Proficiency tests.

Evaluation criteria

Forensic experts and automated system are compared with respect to their estimated Likelihood Ratios. As explained in 3.1, validation requires both assessment of discrimination and calibration. In [71], proposes both Log-Likelihood Ratio cost (C_{llr}) and Empirical Cross-Entropy (ECE) as adequate metrics for validating calibration on an incorrect selection of databases, a bad choice of statistical models, low quantity and bad quality of the evidence. There are several methods that evaluate the model performance on discrimination and calibration, such as EER, DET, Tipett plots (see section 3.2). However, according to [6] and [71], the ones that condense this information better are C_{llr}, C_{llr} min and ECE plots, which are described in section 3.2.1.

The Cost likelihood ratio is defined as:

$$C_{llr} = \frac{1}{2N_p} \sum_{i_p} \log_2(1 + \frac{1}{SLR_{i_p}}) + \frac{1}{2N_d} \sum_{j_d} \log_2(1 + SLR_{j_d}), \tag{3.3}$$

where the indices i_p and j_d respectively denote summing over the computed LR scores for each face pair comparison where each proposition (respectively prosecutor or defense) is true. Minimizing the value of C_{llr} implies an improvement of both discrimination and calibration performance of the automated system [73]. The value ranges from zero (perfect decision making), to infinity (completely wrong). A value of one indicates the system makes a random selection. A value larger than one indicates that the system is making a decision worse than random, i.e. supporting the prosecution hypothesis when it should be supporting the defence hypothesis or vice versa.

Empirical Cross-Entropy in terms of prior odds and the SLR is given by [73]:

$$\begin{aligned} ECE(O(H_p), SLR) = \frac{P(H_p|I)}{N_p} \sum_{i_p} \log_2 \left(1 + \frac{1}{SLR_i \times O(H_p)}\right) \\ + \frac{P(H_d|I)}{N_d} \sum_{j_d} \log_2 \left(1 + SLR_i \times O(H_p)\right), \end{aligned} \tag{3.4}$$

where s_{i_p} and s_{j_d} denote the scores from the same subject and different subject scores in each of the facial image comparisons, where H_p or H_d is respectively true. $O(\theta_p)$ is the value of the prior odds.

To be more precise, the meaning of the ECE plot is as follows [73]:

LRs This curve is the ECE of the LR values in the validation set, as a function of the prior

log-odds. The lower this curve, the more accurate the method. This curve shows the overall performance of the LR method.

PAV LRs This curve is the ECE of the validation set of LR values after the application of the PAV algorithm. This shows the best possible ECE in terms of calibration, and it is a measure of discriminating power.

Reference This curve represents the comparative performance of a so-called neutral LR method, defined as the one which always delivers $LR = 1$ for each forensic case in the set of LR values. This neutral method is taken as baseline performance: the accuracy should always be better than the neutral reference. Therefore, the solid curve in an ECE plot should be always be lower than the reference curve, for all represented values of the prior log-odds.

3.4 Results

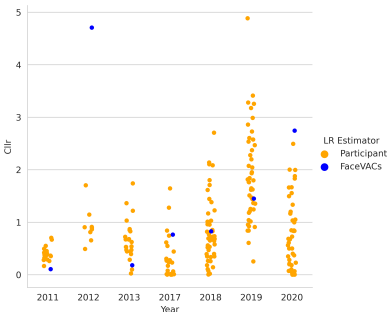
We used the three following three types of calibration: naive, quality score and same features calibration (see sections 3.4.1, 3.4.2 and 3.4.3).

3.4.1 Naive calibration

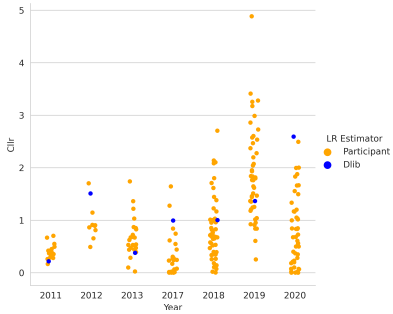
Calibration was performed with three datasets (LFW, SC face and ForenFace) with no filters related to the testing ENFSI tests. From the three datasets chosen, 10000 random pairs were selected as a representative sample. In figures 3.2a and 3.2b the C_{llr} from both face recognition and FaceVACs can be seen. In figures 3.3a and 3.3b ECE plots for the naive calibration can be seen.

We can appreciate that the year 2020 has a very poor C_{llr}, which approximates 4. This could be due to that year having identical twins in the ENFSI tests, which confused the algorithm and led it to classify as matches what should have been mismatches. For the year 2019, the C_{llr} is quite high, which indicates a poor performance, but it is in the same interval as the forensic experts. This year the comparison of faces was among children so both the algorithm and the experts had difficulties with the images.

For the years 2017 and 2018 the C_{llr} is approximately 1, which indicates the power decision of a random algorithm. On the other hand, human participants managed to have their C_{llr} below one in year 2011 (except 2 participants) and about two thirds of them had a C_{llr} below 1 in year 2018. For the rest of the years 2011, 2012 and 2013, both the commercial



(a) C_{llr} computed for calibration of naive calibration for FaceVACs.



(b) C_{llr} computed for calibration of naive calibration for Dlib face recognition.

Figure 3.2: C_{llr}s for naive calibration with Dlib and FaceVACs

software FaceVACs and the open software Face recognition present results comparable to the best performing experts.

Regarding the ECE plots, both FaceVACs and Face Recognition seem to make less errors in the prosecution priors than in the defence priors.

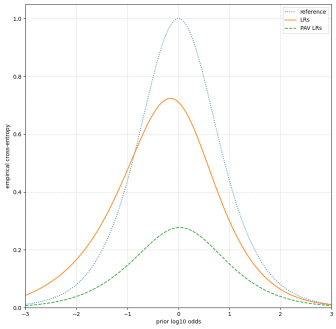
3.4.2 Quality score calibration

For each WSV pair of the calibration datasets (LFW, Sc and ForenFace) the corresponding BSV (i.e. pairs that correspond to a mismatch) is chosen according to a 'quality-score'. Through experiments, it can be seen that in higher resolutions there is a clearer threshold in which the system distinguishes which comparisons are a match and which ones are a mismatch. When the size of the image (measured in megapixels) is above 0.3, the similarity of matched pairs is close to 1, and close to 0 in the case of mismatches. As resolution of the images decreases, similarity for matched images also decreases and similarity for mismatches rises for some cases.

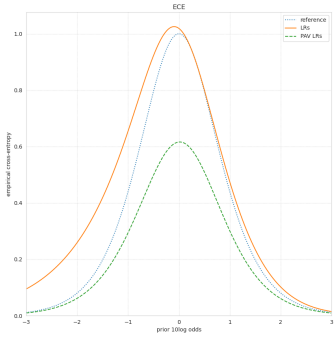
In figures 3.4a, 3.4b, 3.5a, and 3.5b, the validation of the automated systems against experts is checked. The results are shown for the years 2011, 2012, 2013 and 2017, 2018, 2019 and 2020 and both C_{llr} and C_{llr} min are plotted.

3.4.3 Feature calibration

The feature calibration was performed with pairs of the two datasets (SC face and ForenFace). For each test pair (from ENFSI tests), the set of features of image one and the set of features of image two are considered to calibrate only with the pairs of the calibration

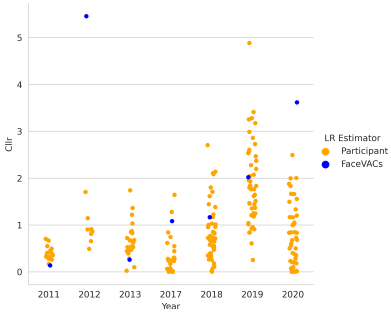


(a) ECE plot computed for calibration of naive calibration for FaceVACs.

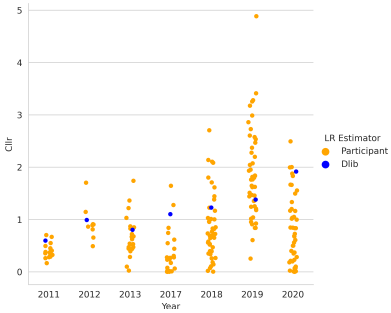


(b) ECE plot computed for calibration of naive calibration for face recognition.

Figure 3.3: ECE plots for naive calibration with Dlib and FaceVACs. LR curve is orange line. PAV LRs correspond to dashed green curve and reference is the dotted blue curve.

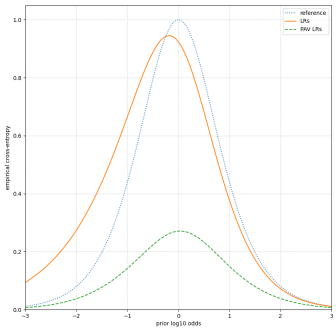


(a) C_{llr} computed for calibration with same quality score with FaceVACs system.

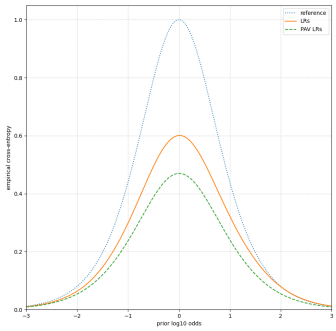


(b) C_{llr} computed for calibration with same quality score with Face Recognition system.

Figure 3.4: C_{llr}s for same quality score calibration with Dlib and FaceVACs



(a) ECE plot computed for calibration with same quality score with FaceVACs system.

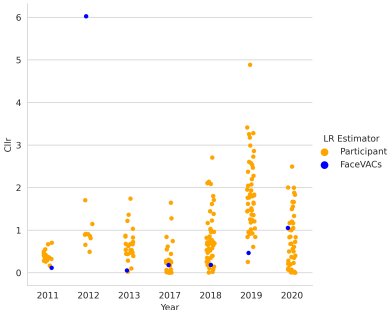


(b) C_{llr} computed for calibration with same quality score with Face Recognition system.

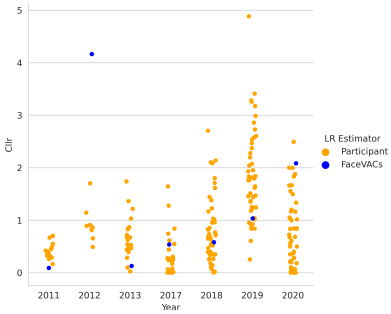
Figure 3.5: ECE plot for quality score calibration with Dlib and FaceVACs

dataset that have the same set of features as these two images. The features to be considered were: glasses, beard, headgear, other occlusions, and low quality. The datasets were manually annotated.

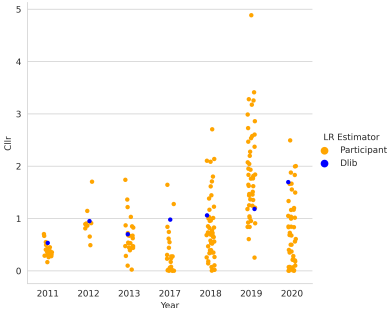
In figures 3.6a, 3.6b, 3.6c, and 3.6d, it can be seen that C_{llr} calibrating the system with comparisons that have the same features has improved results with respect to C_{llr} calibrated with comparisons of the same quality score.



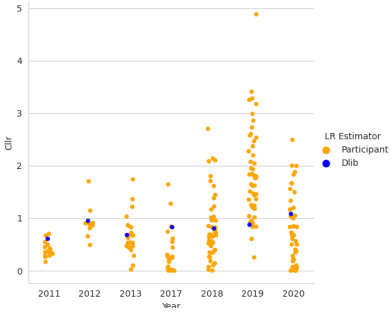
(a) C_{llr} computed for calibration with same with same yaw and pitch with with FaceVACs system.



(b) C_{llr} computed for calibration with same with same low quality with with FaceVACs system.



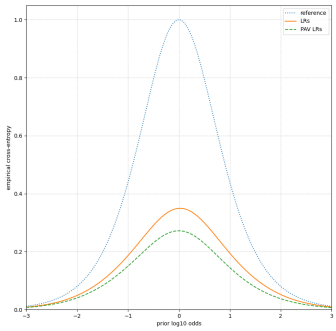
(c) C_{llr} computed for calibration with same yaw and pitch with face recognition system.



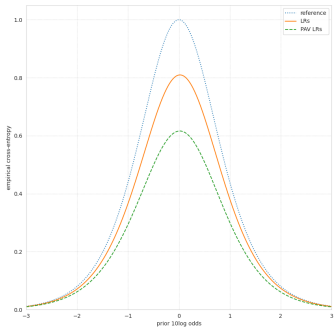
(d) C_{llr} computed for calibration with same low quality (manually annotated) with face recognition system.

Figure 3.6: C_{llr}s for features calibration with Dlib and FaceVACs

In figures 3.7a and 3.7b ECE plots for both automated systems are plotted.



(a) ECE plot computed for calibration with same features with FaceVACs system.



(b) ECE plot computed for calibration with same features with face recognition system.

Figure 3.7: ECE plot for features calibration with Dlib and FaceVACs

3.4.4 Overview

An overview of the results can be seen in table 3.2. For this results, the open source Dlib, is compared to the commercial software FaceVACS and to the ENFSI participants. The different results can be seen where the filters applied improve with respect to naive calibration. FaceVACS performs better than the open software system. The calibrator chosen for the results in the table was Isotonic Calibrator, although calibrating with any of the three would turn out to be similar to C_{llr}, the Isotonic seemed to outperform a bit with respect to Logistic Regression and Calibration. However, further work is necessary to make any recommendations on which cases each of the three calibration methods should be used.

Table 3.2: C_{llr} results summary for Dlib, FaceVACs (according to filters chosen) and participants.

Year	Filters	Dlib	FaceVACs	Average Participants
2011	No filters	0.22	0.11	0.40
	Confusion Score	0.60	0.14	
	Yaw, Pitch	0.53	0.11	
	Glasses, Beard	0.60	0.08	
	Low Quality	0.63	0.09	
	Head Gear	0.56	0.20	

2012	No filters	1.51	4.70	0.93
	Confusion Score	0.99	5.46	
	Yaw, Pitch	0.95	6.02	
	Glasses, Beard	0.94	3.67	
	Low Quality	0.95	4.16	
	Head Gear	0.96	5.10	
2013	No filters	0.38	0.18	0.67
	Confusion Score	0.80	0.26	
	Yaw, Pitch	0.70	0.05	
	Glasses, Beard	0.64	0.12	
	Low Quality	0.70	0.13	
	Head Gear	0.73	0.35	
2017	No filters	0.99	0.76	0.35
	Confusion Score	1.10	0.87	
	Yaw, Pitch	0.98	0.18	
	Glasses, Beard	0.81	0.53	
	Low Quality	0.83	0.54	
	Head Gear	1.00	1.24	
2018	No filters	1.00	0.83	0.84
	Confusion Score	1.23	0.93	
	Yaw, Pitch	1.06	0.18	
	Glasses, Beard	0.77	0.56	
	Low Quality	0.80	0.58	
	Head Gear	1.09	1.30	
2019	No filters	1.36	1.45	1.88
	Confusion Score	1.38	1.57	
	Yaw, Pitch	1.18	0.46	
	Glasses, Beard	0.82	1.07	
	Low Quality	0.87	1.03	
	Head Gear	1.23	2.21	

2020	No filters	2.59	2.74	0.78
	Confusion Score	1.92	2.78	
	Yaw, Pitch	1.70	1.05	
	Glasses, Beard	1.00	2.01	
	Low Quality	1.05	2.08	
	Head Gear	1.71	3.74	

3.5 Discussion

As we can see in the C_{llr} and ECE plots, the commercial software FaceVACs outperforms both the open software face recognition and the experts for full frontal images. However, the quality of images presented in the tests are easier for the automated system than the material that is normally handled in cases. Most of the images (especially in the year 2017) are frontal with little pose variation, which facilitates the task for the automated system. Most of the wrong assessments provided by the automated system were due to occlusions in the test images (caps, mics, scarfs) or to illumination. On the year 2011 dataset, where the illumination was constant but the images had different resolution and compression, there was not significant improvement neither in FaceVACs nor face recognition with respect to naive calibration or quality score and filtered base calibration, as the C_{llr} of the automated systems was already close to zero in the naive calibration. On the other hand, there is a significant improvement in the years 2018 and 2019 using the same features calibration instead of the naive calibration. These years had as peculiarity that year 2018 had a lot of variety in the test images (age variation, pose, quality...) and 2019 had pictures of children. The year 2020 has low performance in all the cases due to photos of twin siblings being present among the test images. The automated system had difficulty differentiating these faces and gives them a high similarity score, making the calibration prone to error.

This study takes a step further the usability of automated facial image comparison systems in the forensic field. In the literature, such calibrations are performed normally as suspect-anchored and trace-anchored [6; 73], however this type of calibration was not the use case in this study due to only having one sample of each identity in the comparison tests. This use case is given when the suspect is not yet convicted and only one image of the individual is available.

This has not impeded the automated system of reaching in most of the years the accuracy of the forensic experts. It may lead us to think that if on top of performing these cal-

ibrations with publicly available data-sets, data more relevant to the case was added (such as more images of the suspect, images of the suspect and other relevant population resembling the conditions in which the query image was taken) the results would only improve.

As future work, it would be convenient to indicate that the automated system performance (both FaceVACS and Deepface) is less reliable if there are occlusions. When the face was not detected by the automated system, it was not considered for the C_{llr} or ECE plot. A possible alternative to this is to add the lack of face detection as an inconclusive LR (i.e. LR = 1) which would drop the performance of the system in C_{llr} terms, as humans are habitually more efficient when finding faces in a picture than a automated system can be. As indicated, an important point made by [111] is that validation of Likelihood Ratio in the forensic field should take into account not only accuracy (if it is right or wrong assessment of match-mismatch) but also its calibration, i.e. the system capacity to make strong assessments. If a system provides an LR of around 1 for a comparison corresponding to a match, the assessment (i.e. discrimination power) is right, but the calibration and functionality to help to take a decision is not very useful. On the other hand, a second system that for the same match provides an LR of 1000 is both providing a high discriminating power and good calibration. The article [111] warns that validation of LR systems should check on both characteristics. For our work, measuring with C_{llr} and ECE plot has this warning covered, because looking at both equations 5.4 and 3.4, the cost will increase for those systems that provide wrong assessments or low discrimination power (LR close to one).

Regarding the three calibrator methods chosen (Logistic Regression, KDE and Isotonic Regression), none of them seemed to stand out from the others. Although Isotonic Regression seemed to achieve slightly better results than the other two, future work is required to assess in which use cases one calibration is better than the other. With respect to the three calibration methods chosen, although both the confusion score and labeled filters improved the C_{llr} with respect to applying generic calibration, also further research is needed to help the investigator to determine which method would suit best for each use case.

3.6 Conclusion

In conclusion, with this study it has been demonstrated that applying "filters" such as "Quality Score" and calibration with the same features as the test images improves the performance in the calibration, in terms of both C_{llr} and ECE. The results with open software are inferior, but they are more transparent so more research should be conducted to bring

open software at par with commercial vendors. On top of performing these calibrations with publicly available data-sets, more relevant data to the case, such as more images of the suspect, images of the suspect and other relevant population resembling the conditions in which the query image was taken could be added. The results would only improve. The expert cannot be replaced by this tool, but becomes more efficient because the computer can help to reduce the amount of information to be managed by doing appropriate filtering. If facial image comparison is conducted by two experts doing the comparison independent from each other, the third might be an algorithm, and the experts can evaluate their findings as well as the findings of the algorithm to draw a conclusion.

4

Multi-Task Explainable Quality Networks for Large-Scale Forensic Facial Recognition

Identifying suspects from surveillance footage is a crucial task in forensic investigations, but it is often hindered by the variable conditions of observation and the large amounts of data. Face image quality (FIQ) is a metric that measures the usefulness of a face sample for facial recognition. Existing methods for automated FIQ assessment only provide a scalar value for quality, and do not indicate which factors are causing low quality. Additionally, these methods are computationally expensive, which makes current FIQ assessment methods unsuitable for large numbers of images. To address these issues, we introduce multi-task explainable quality networks (XQNets). XQNets provide both the quality value and the associated facial and environmental attributes, and automatically learn how each attribute contributes to the quality value during the training process. We also propose a dataset-agnostic quality pairing protocol to ensure that sample pairs are balanced across datasets and evaluations are fair. Our experiments on the LFW and SCface benchmarks show that our approach generalizes well across different datasets and outperforms state-of-the-art methods. Our method offers a fast, explainable approach to FIQ assessment, making it suitable for large-scale forensic applications.

4.1 Introduction

Face recognition (FR) has improved significantly due to advancements in facial recognition algorithms and methodologies [107; 112]. Despite the improvements, recognition error rates remain high in real-world forensic applications, particularly in challenging conditions such as CCTV footage or ATM cameras [28; 113; 114]. As FR systems play an increasingly larger role in crucial decision-making processes, there is a growing need to explain the FR process to humans [115; 116]. As a solution, Face Image Quality (FIQ) assessment methods have been developed to output a quality score that can be represented as a single scalar value or a vector of values to explain how suitable a face image is for face recognition [59; 117; 118]. However, there have been few studies to explain these scores to users and provide an interpretable cause for a face image's low or high quality [117; 119]. An example of an explainable FIQ analysis pipeline is presented, where images with low quality scores are rejected and accompanied by attributes that help humans understand the system's decision [119]. This makes FR not only accurate but also explainable.

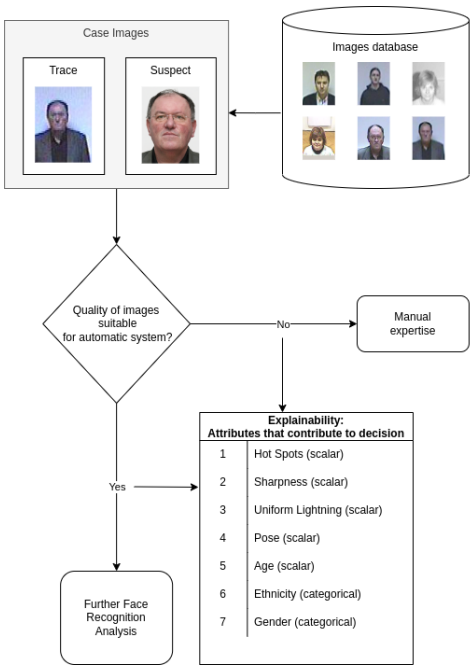


Figure 4.1: Face Image Quality pipeline, where the suitability of the image has to be assessed before face recognition analysis. Non suitable images can be manually compared or rejected. Images extracted from [29].

General image quality assessment algorithms such as BRISQUE [120], NIQE [121], and PIQE [122] do not achieve satisfactory performance when applied to face images because they aim to assess images in terms of subjective (human) perceptual quality [113]. On the other hand, FIQ assessment algorithms are concerned with the assessment of the biometric utility of facial images, which can be objectively defined in the context of specific FR systems. Hence, FIQ assessment methods obtain more accurate results for FR applications [115; 123; 124]. This occurs because FIQ algorithms for the purpose of biometric utility prediction can perform better than a general image quality assessment that has not been developed with facial biometrics in mind.

Predicting recognition utility in FR implies that the quality score has to indicate the "accuracy" or "certainty" of comparison scores generated for a sample pair that includes the assessed sample [115]. Thus, quality should be indicative of the face comparison performance. Note that this entails that the output of a specific FIQ assessment algorithm may be more accurate for a specific FR system. So the FIQ assessment utility prediction effectiveness ultimately depends on the combination of both the FIQ assessment algorithm and the FR system. According to [124], it is desirable to facilitate interoperability such that the FIQ assessment algorithm is predictive of recognition performance in general for a range of relevant systems instead of being dependent on only one.

FIQ measures enable various forensic applications. For example, in real-time recording sessions, photos can be accepted or rejected based on their scalar image quality values. If the image quality is too low, the system will reject it and collect a new image, which is particularly valuable during first enrolment when a reference photo is not available [113]. Scalar image quality values can also be used as a management indicator by summarizing the effectiveness of the collection process across different sites and conditions [125]. Additionally, they can be used to select the best image from a set of photos [117]. It would be useful to have an FIQ that can explain why an image cannot be used and which facial or environmental attributes the subject needs to improve to increase the quality.

Regarding FIQ assessment, literature often focuses on optimizing quality scores on benchmarks such as LFW or Adience [115; 116; 126]. However, there is a risk of saturation on datasets such as LFW [28], which led to the proposal of XQLFW, a benchmark derived from LFW with pairs of maximum quality difference [28]. The selection of images for XQLFW is based on BRISQUE and SER-FIQ quality scores, but this selection may introduce a bias towards SER-FIQ [28]. Additionally, LFW and XQLFW come with a predefined set of 6000 pairs for evaluation, requiring the generation of a new set of pairs for new datasets like SCFace or ForenFace. To overcome this issue, this study proposes a dataset-agnostic quality pairing (DAQP) protocol to ensure a balanced representation of the whole spectrum of qualities in pair generation. The study evaluates the widely used datasets LFW [40], XQLFW [28], and DAQP on forensic datasets such as SCFace [29] and ForenFace [105].

The overall FIQ value is related to descriptive facial attributes such as deviation from the frontal pose or hot spots; and environmental attributes such as sharpness or deviation from uniform illumination. One way to consider all these variables would be to process each separately and combine the scores afterwards but this would not allow to learn the common aspects and it would be inefficient. Learning paradigms such as multi-task learn-

ing (MTL) [127] could help leveraging the domain-specific information in the training signals of related tasks [128; 129], e.g., the FIQ value, and its related attributes to generalize better across different FR models. Moreover, MTL provides several outputs with a single forward inference, which allows accelerating the computation significantly. Thus, in this work, we study how MTL can be exploited to build efficient explainable FIQ assessment systems for large-scale forensic FR applications. More specifically, the contributions we make in this chapter are:

1) **Multi-task explainable quality networks (XQNets)** to efficiently assess FIQ value along with a set of facial and environmental image attributes that explain the calculated FIQ.

2) **A dataset-agnostic pairing (DAQP) protocol** to evaluate explainable FIQ assessment systems using the whole range of FIQ values in the test dataset ensuring that sample pairs are balanced.

3) **Experimental results** with the LFW and SCface FR benchmarks, following the DAQP protocol, demonstrating that XQNet has an accurate FIQ across different state-of-the-art FR assessment methods in complex surveillance scenarios and with **competitive inference times**.

The rest of the chapter is organized as follows: Section 2 describes prior related work; Section 3 explains our proposed XQNet and DAQP protocol; Section 4 presents experimental results with LFW and SCface FR benchmarks. Finally, Section 5 presents the conclusions and future lines of work.

4.2 Related Work

FIQ assessment algorithms can be classified as factor-specific and monolithic approaches. The former comprises methods for finding interpretable factors, such as blur and sharpness, which could help an operator to avoid inadequate face images when recapturing. The latter outputs an overall FIQ value leading to comparatively opaque assessments/quality scores.

Factor-specific approaches to FIQ assessment are based on either facial attributes (e.g. inter-eye distance, pose) or environmental attributes (e.g. illumination, blur) [117; 130–132]. For example, in [130], the authors trained and compared ten features of quality estimates of a single human to assess general image quality. In [131], the authors estimate only the pose angle without producing a normalized quality score, demonstrating that pose estimation can be used for FIQ assessment. In [132], 17 parameters based on ICAO Doc 9303 requirements are used to evaluate FIQ, resulting in an 88% correct classification rate.

In [117], FaceQvec is proposed as a method to estimate the conformity of facial images with ISO/IEC 19794-5, a quality standard for face images in official documents. The method consists of 25 individual tests related to the standard and other image characteristics, with accurate evaluation results. However, these methods tend to use high-quality images for evaluation and little attention is given to forensic applications.

The monolithic approaches are divided into: human ground truth training, FR-based ground truth training, FR-based inference and FR-integration. Hernandez-Ortega et al. proposes the FaceQnet model[124] with versions v0 and v1. For both versions, as part of the training data preparation, the BioLab-ICAO framework from [133] is employed to select suitable high-quality images per subject, which are used to compute the ground truth quality scores for the subjects' remaining training images. This ground truth quality score computation consists of the normalized Euclidean distances of embeddings produced by a number of FR feature extractors (three for v1; and only one, FaceNet [46], for v0). Both FaceQnet versions were based on a ResNet50 [134] model pretrained for FR using the VGGFace2 dataset [59], replacing the final output layer with two fully connected layers which are used for finetuning while the rest of the network weights were frozen.

SER-FIQ is a model proposed by Terhörst et al. with two variants: "same model" and "on-top model." Both variants estimate the quality of FR by comparing the embeddings of randomly chosen subnetworks without ground truth quality labels. The "same model" variant can be used on FR networks trained with dropout and the "on-top model" variant uses a small additional network trained with dropout on top of the FR model. The evaluations showed that the "on-top model" variant mostly outperformed the baseline approaches and the "same model" variant showed strong FNMR performance improvement for a fixed FMR of 0.001.

The MagFace model [116] integrates quality and FR, with quality directly indicated by the magnitude of the FR feature vector. The model extends the ArcFace [88] training loss with a magnitude-aware angular margin and magnitude regularization, resulting in larger magnitudes for higher quality images and smaller magnitudes for lower quality images. The magnitude is bounded during training, and a normalized quality score can be derived through linear scaling. The FR function after training remains unchanged from ArcFace.

The Pixel-Level FIQ approach [119] allows evaluating the pixel-level attributes of a face picture given an arbitrary FR network that does not require any training. A model-specific quality value of the input picture is computed and utilized to develop a sample-specific quality regression model to do this. Using this technique, quality-based gradients are back-

propagated and translated into pixel-level quality estimates. They evaluate the significance of their suggested pixel-level features subjectively and quantitatively using actual and fake disruptions and compare explanation maps on faces that do not meet ICAO rules. The findings show that the suggested method creates meaningful pixel-level characteristics that improve the interpretability of the full facial picture quality in all cases.

Ou et al. proposed SDD-FIQA [123], a FIQ method that considers both the intrinsic properties and the recognizability of the face image. They argue that a high-quality face image should be similar to its intra-class samples and dissimilar to its inter-class samples. Thus, their method generates quality pseudo-labels by calculating the Wasserstein distance between the intra-class similarity distributions and inter-class similarity distributions. With these quality pseudo-labels, they are capable of training a regression network for quality prediction. Their method shows good generalization across different recognition systems. However, they do not provide the set of attributes that would affect the high or low quality of an image.

MTL has been applied to FR tasks such as landmark detection and anti-spoofing, but not to FIQ estimation. In [135], MTL is used for landmark detection and improves performance for faces with severe occlusion and pose variation. In [136], AENet uses rich semantic annotations as auxiliary tasks to boost the performance of face anti-spoofing. In [137], MHCNN is proposed for joint face detection, landmark detection, facial quality, and attribute analysis, but it is only used for face detection, not recognition. To date, there are no MTL methods applied to FIQ estimation.

4.3 Methodology

The proposed multi-task learning model consists of 3 steps. First, for a given facial image, the face must be pre-processed to obtain an input image of (3, 112, 112). Second, the input image is processed by the network, and third, the vector with facial and environmental attributes output will be obtained. The framework of the proposed multitask learning model XQNet is shown in figure 4.2. The model consists of a body with several heads. Each head processes one of the facial or environmental attributes. The loss function combines all of the head attribute outputs assigning a weight to each. Finally, the model outputs the target FIQ together with facial and environmental attributes that contribute to the decision for such quality.

To choose a backbone for the network, one has to take into account that better accuracies tend to be obtained by more complex DNN architectures that require significant

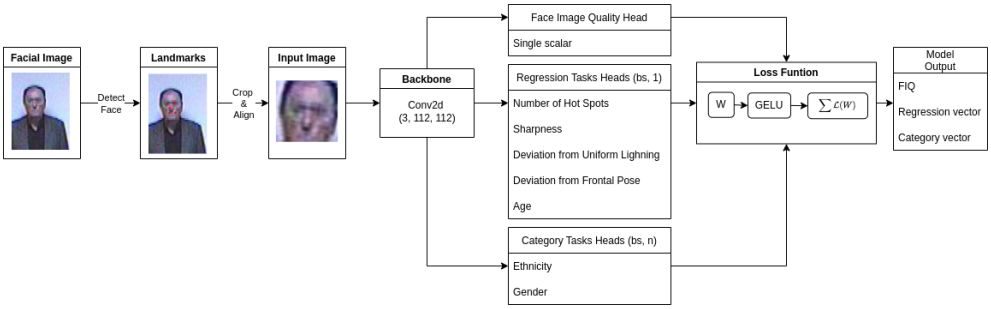


Figure 4.2: Multi-Task Learning Model architecture. First, the image must be prepossessed before feeding it to a backbone with several heads, 6 regression heads (including quality) and 2 categorical heads. Then all of their outputs are processed through a loss function that will output the vector for facial and environmental attributes.

computational resources. In our context, it is desirable that the network is able to process a large amount of images in a short time, and has a good trade-off between accuracy and performance. It is for this reason that the backbones chosen are EfficientNet [138] and ConvNext [139]. Vision transformers (ViTs) are another type of DNNs that are receiving the attention of the computer vision community recently, as they have demonstrated superiority in accuracy over CNNs [140]. But they currently have higher computational costs and therefore require further research on optimization techniques to efficiently deploy them in resource-constrained processors [141].

For MTL we use the hard parameter sharing approach [142]. It is generally applied by sharing the hidden layers between all tasks, while keeping several task-specific output layers. Hard parameter sharing greatly reduces the risk of overfitting [143]. The more tasks we are learning simultaneously, the more our model has to find a representation that captures all of the tasks and the less chance of overfitting on our original task, i.e. FIQ.

4.3.1 Training

The training pipeline is as follows: the landmarks of the facial image are first detected. In the case of the face not being detected, the image is not considered for training. Then the landmarks are used to crop and align the face in order to yield the appropriate shape (3, 112, 112) to be inserted into the network. In our method, we rely on detecting and aligning faces as the input to our model because the quality assessment should be performed on an aligned face in order to ensure accurate results. The model expects an aligned face as input in order to process the quality assessment in an adequate manner. The network itself is formed by a backbone and several heads. The outputs of the heads go to the loss

function. There are seven regression tasks to estimate the following variables: FIQ, number of hotspots, sharpness, deviation from uniform lightning, deviation from frontal pose, and age. Regression tasks are those which output a continuous variable. Moreover, there are two categorical variables to learn: ethnicity and gender. Category tasks are those which output a class value. Each of these tasks are represented in the network as a head that comes from the backbone.

The cost function is based on the work of [128] and has the weights as trainable parameters of the network. The loss function is defined as:

$$\mathcal{L}(\mathbf{W}, \sigma_1, \sigma_2) = \frac{1}{2\sigma_1^2}\mathcal{L}_1(\mathbf{W}) + \frac{1}{2\sigma_2^2}\mathcal{L}_2(\mathbf{W}) + \log(\sigma_1) + \log(\sigma_2), \tag{4.1}$$

where σ_i is the observation noise. The variable σ_1 represents the noise parameter for the model output y_1 (regression) and σ_2 represents the noise parameter for the model output y_2 (categorical). The losses $\mathcal{L}_1$ and $\mathcal{L}_2$ are defined by:

$$\mathcal{L}_1(\mathbf{W}) = \left\|\mathbf{y}_1 - \mathbf{f}^{\mathbf{W}}(\mathbf{x})\right\|^2, \tag{4.2}$$

and:

$$\mathcal{L}_2(\mathbf{W}) = -\log \text{Softmax}(\mathbf{y}_2, \mathbf{f}^{\mathbf{W}}(\mathbf{x})). \tag{4.3}$$

where $\mathbf{f}^{\mathbf{W}}(\mathbf{x})$ is the output of a neural network with weights $\mathbf{W}$ on input $\mathbf{x}$.

The equations 4.2 and 4.3 can be used for each of the regression and categorical tasks of our model, allowing us to learn the relative weights. This loss is smoothly differentiable, and it ensures that the task weights will not converge to zero. In addition to the model of [128], we apply a GELU [144] function before introducing the weights in the valid loss, both to ensure that the valid loss does not get in negative values and resulting in better training results.

The model is trained using [145], which demonstrates an improvement in the learning speed with regards to a cycle in which the learning rate (lr) and momentum are kept constant. It consists of the following steps: first, we progressively increase our lr from lr_{max}/f to lr_{max} and at the same time we progressively decrease our momentum from mom_{max} to mom_{min}. Second, we do the exact opposite: we progressively decrease our lr from lr_{max} to lr_{max}/f and at the same time we progressively increase our momentum from mom_{min} to mom_{max}. Thirdly, we further decrease our lr from lr_{max}/f to $lr_{max}/(fx100)$ and we keep momentum steady at mom_{max}.

4.3.2 Evaluation protocols

As done in [115; 116; 126], we use EVRC to evaluate the performance of FIQ assessment methods. The EVRC uses the partial area under the curve defined as:

$$pAUC = \int_0^a FNMR(\phi)d\phi, \tag{4.4}$$

where ϕ is defined as the percentage of images which are not considered and FNMR is the False Negative Match Rate at the given ϕ. For convenience and to be able to compare with The FNMR is defined as the number of false negatives (negative facial recognition claims which should have been accepted) divided by the total amount of real positives (false negatives or FN + true positives or TP) i.e.:

$$FNMR = \frac{FN}{FN + TP}. \tag{4.5}$$

FNMR is a useful metric for evaluating the performance of a face recognition system. In face recognition, false negative errors refer to the situation where the system fails to match a pair of face images that belong to the same person, i.e., it wrongly classifies them as different persons. A low FNMR indicates that the system is able to accurately match faces that belong to the same person. FNMR is particularly relevant in security and surveillance scenarios where failing to recognize a person can have serious consequences. For example, a false negative error could result in a person being incorrectly denied access, while in a criminal investigation, it could result in a suspect going undetected.

The use of the EVRC is beneficial because it demonstrates the effect of discarding low-quality face images on FR performance, as measured by FNMR. This curve shows the relationship between FNMR and reject rates, allowing us to understand how FNMR changes as an increasing amount of low-quality data is discarded. Using the EVRC curve is a fair method to compare the performance of different FIQ assessment algorithms, as it is independent of the absolute quality score values and their range. Additionally, the use of the ERVC provides a clear, visual representation of the relationship between FIQ and FR performance, making it an informative and effective evaluation tool.

When evaluating LFW, 6000 randomly generated pairs were used as a benchmark. We also used the 6000 pairs provided for XQLFW [28]. From each pair, the image with the minimum quality is taken as the 'pair quality'. Performing FR in all the pairs, FNMR is computed. Then 5% of the worse quality pairs are removed and FNMR recomputed. The process is iterated until no more pairs are left. However, this method does not contemplate

(a) LFW example pair *(b) XQLFW example pair* *(c) DAQP example pair*

Figure 4.3: Examples of representative pairs with different evaluation protocols, each with its SDD-FIQA [123] FIQ. Note that DAQP evaluation seeks pairs with similar (even equal) FIQ.

4

all the qualities in the database nor can it be extended to other databases that do not have a standard set pairs to evaluate such as LFW or XQLFW. If a new database arises such as SCFace [29], and it does not come with a preset evaluation of pairs. Should we generate one ourselves randomly? Even if the pair set is already available such as in LFW. Is it the most suitable for quality evaluation? XQLFW [28] proved that it is not, but they used both SER-FIQ and BRISQUE quality to generate the pairs, which can make the method biased. It is for this that we propose a new method which allows to extend the evaluation to other databases that do not have pre-established pairs.

First, we compute the quality for each image in the dataset. After that, we compute a histogram with $n = 20$ bins. For each bin, we compute the maximum number of pairs of the same identity that are available. In this way, we make sure that when performing face comparison pairs of similar quality are compared and not pairs that have very different qualities. Another way to form these pairs would be to perform cross-bin comparisons so that very high-qualities are compared against very low qualities, but as we are removing pairs with the lowest quality, and to be fair in the comparison (the quality of the pair is the minimum of the two image qualities), we decide to adopt this criterion of making pairs of similar quality and not maximizing the difference.

Once the quality pairs are computed, the same number of different identity pairs are obtained. Once this is done for all the bins in the histogram, we compute the FNMR 20 times each time removing the bin, the lowest quality. A summary of the DAQP algorithm is found in algorithm 1. A graph showing the pair-quality distribution for the datasets of LFW, XQLFW and DAQP is shown in figure 4.4. The FNMR performance of the FIQ model using DAQB (now renamed DAQP) provides a more comprehensive understanding of the model's performance across a diverse range of image quality, as compared to evaluating the model using randomly selected pairs. This is a crucial aspect in assessing the generalization capability of the FIQ method. Samples of different evaluation pairs are shown in figure 4.3.

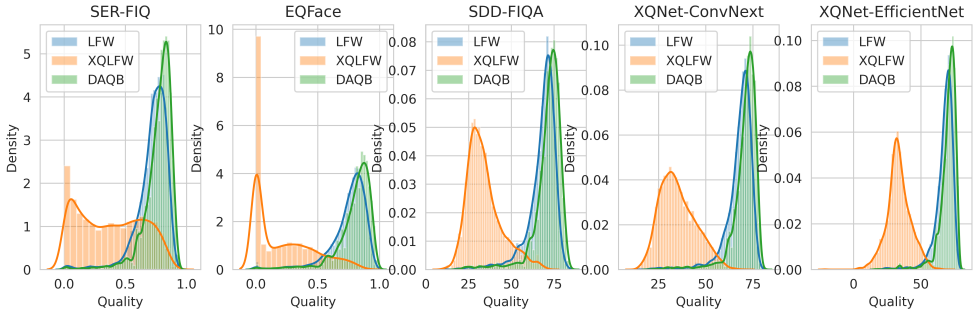


Figure 4.4: Set of qualities for all the state-of-the art methods used, including our own proposed method XQNet-ConvNext and XQnet-EfficientNet

Algorithm 1 Dataset-Agnostic Quality Pairing (DAQP)

Input: Set of images annotated by quality I

Output: Set of evaluation pairs P

DAQP (I, n = 20)

Distribute I in n quantiles based on quality

for $quantile = 1$ to n **do**

  $si \leftarrow$ all pair combinations of same identity

  $nsi = len(si)$

  $di \leftarrow nsi$ pair combinations of different identity

$P \leftarrow$ empty list

$minp \leftarrow$ find $min(nsi)$ in all the quantiles

$P = P.add(si, di)$ where quartile(si, di) has minp

iterate over the rest of quantiles

for $quantile = 1$ to $(n - 1)$ **do**

  $si = si.randomselect(minp)$

  $di = di.randomselect(minp)$

  $P = P.add(si, di)$

return P

4.4 Experiments

We use two types of databases: forensic-oriented, where the images have low-resolution, are taken at a distance, or the subjets have very different poses such as SCface [29] and Forenface [105]), and standard databases used commonly in literature to test FIQ algorithms, such as LFW [40], XQLFW [28] or UTKFace [146]. We use UTK Face for training and for testing we use LFW, XQLFW, SCFace and ForenFace. LFW and XQLFW are both public datasets, and although LFW almost reaches saturation in most FR systems [28], it is still widely used in literature for FIQ evaluation. On the other hand, SCFace and ForenFace are closed datasets but much more focused in forensics, proposing more challenging images for quality evaluation.

LFW [40] is a database of 13,000 images of faces collected from the web. 1680 of the people pictured have two or more distinct photos in the data set. SCFace [29] is a database in which images were taken in uncontrolled indoor environment using five video surveillance cameras of various qualities. The database contains 4160 static images (in the visible and infrared spectrum) of 130 subjects. ForenFace [105] contains video sequences and extracted images of 98 subjects recorded with six different surveillance camera of various types. Moreover, it also contains high resolution images and 3D scans for these subjects. A subset of 435 images (87 subjects, five images per subject) has been manually annotated, yielding a unique and very rich annotation containing almost 19.000 entries. It also contains a training/testing protocol. The UTK face dataset [146] has a long age span (range from 0 to 116 years old). It has of over 20,000 face images with annotations of age, gender, and ethnicity. The images cover large variation in pose, facial expression, illumination, occlusion, resolution, etc.

The detector used is MTCNN [64] and the face recognizers were ArcFace [88] and FaceNet512 [46]. These face recognizers where implemented using the library DeepFace [107]. The annotations used are as follows: the FIQ uses SDD-FIQA [123] for ground truth annotation, due to being to our knowledge the state of the art in quality estimation. Gender and Ethnicity are taken from the annotated UTK face database [146]. In case the ethinicity out-

Figure 4.5: Sample of dataset annotations

put in UTK dataset was "others", the software anotation used was [147]. The rest of the annotations, which are: number of hotspots, sharpness, deviation from uniform lightning and deviation from frontal pose are estimated using the commercial software library FaceVACs [108]. A sample of these annotations is shown in figure 4.5. A summary of the attribute definitions is in table 4.1. The purpose of XQNet is to be open software, so new training with other annotations and other databases is possible, making it less of the blackbox that FaceVACs commercial software is.

4.5 Results

The proposed explainable face quality estimation is analysed in four ways. It's important to note that determining what constitutes a "good enough" image is not an absolute term and depends on the desired FNMR and the specific dataset being evaluated. A higher FIQ score generally indicates a higher quality image and a higher likelihood of successful person identification, but the appropriate threshold will depend on the user's specific requirements and desired trade-off between false negatives and false positives. First, we used the pair-generation method DAQP algorithm 1 to compare the FNMR results in three datasets. Second, we compare our pair-generation method DAQP against 6000 randomly generated pairs from LFW and the algorithm of 6000 pairs used in XQLFW. Thirdly, as our method is intended for forensic large-scale usage, we compare both CPU and GPU performance times of different FIQ algorithms and fourthly, we show the attribute distributions produced by our explainability method in the different datasets.

Our pair-generation method DAQP algorithm 1 is used to compare the FNMR results in three datasets, (SCFace, ForenFace and LFW). Moreover, to assess the generalization against different Face Recognition systems, we perform the pair verification with ArcFace, FaceNet512 and SFace. The qualitative results are shown in figure 4.6, whereas the quantitative results are shown in table 4.2. the numbers in the table represent the evaluation in terms of partial area under the curve (pAUC) for reject fraction ranges from 5%, 15% and 35%. Lower values indicate lower FNMR, and thus, better performance of the model. For each rejection range, we have marked in bold the minimum FNMR, which indicates the best performance at that percentage of discarded images (see equations 4.5 and 4.4).

Our second analysis consists of comparing our pair-generation algorithm DAQP method against the 6000 randomly generated pairs in LFW [40] and the algorithm for generating 6000 pairs in XQLFW [28]. Equally to the first analysis, the quantitative results are shown in table 4.3.

Table 4.1: Attribute descriptions according to [148], [146] and [147].

Type	Attribute	Description	Range / Classes
Regression	Hot Spots	Bright areas of light reflected from the face	0-12000
	Sharpness	Focus and depth of field according to specification of ISO 19794-5:2005 section 7.3.3.	-
	Deviation from uniform lightning	-	-
	Deviation from frontal pose	-	-
Categorical	Age	-	0-116
	Ethnicity	-	White, Black, Asian, Indian, Hispanic, Middle Eastern
	Gender	-	Restricted to Male/Female

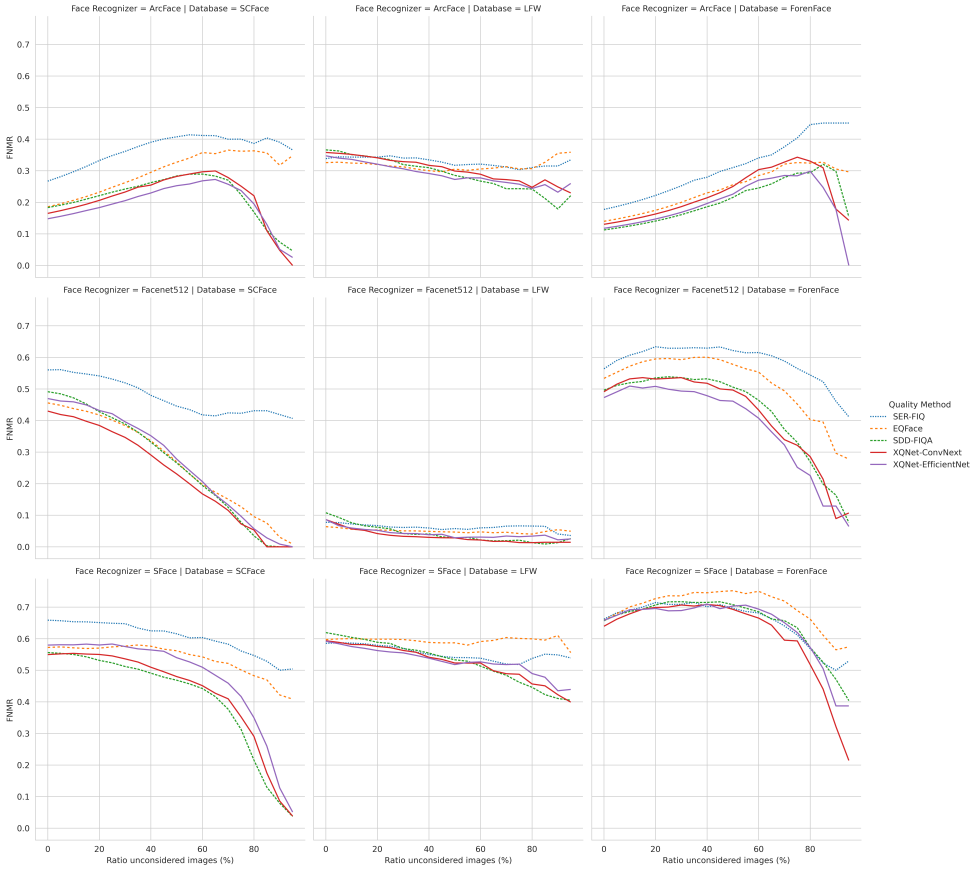


Figure 4.6: FNMR for quality bin pairs using DAQP. The lower the curve, the better the system.

Table 4.2: Table of results for DABP pair-generation method in 3 datasets and 3 Face Recognizers.

		pAUC (%)			
			Ratio of unconsidered images (%)		
Quality Method	**Face Recognizer**	**Database**	**5**	**15**	**35**
SER-FIQ [115]	Facenet512	SCFace	2,80	8,34	18,92
		LFW	0,39	1,11	2,40
		ForenFace	2,89	8,94	21,52
	SFace	SCFace	3,29	9,83	22,78
		LFW	**2,93**	8,79	20,27
		ForenFace	3,36	10,26	24,46
	ArcFace	SCFace	1,37	4,34	11,27
		LFW	1,71	5,14	12,00
		ForenFace	0,91	2,89	7,63
EQFace [149]	Facenet512	SCFace	2,26	6,64	14,63
		LFW	**0,31**	**0,87**	1,91
		ForenFace	2,71	8,42	20,31
	SFace	SCFace	2,87	8,58	20,05
		LFW	2,99	8,99	20,94
		ForenFace	3,34	10,33	24,96
	ArcFace	SCFace	0,95	3,02	7,97
		LFW	**1,63**	**4,88**	**11,18**
		ForenFace	0,72	2,27	6,03
SDD-FIQA [123]	Facenet512	SCFace	2,44	7,14	15,30
		LFW	0,50	1,28	2,35
		ForenFace	2,52	7,70	18,38
	SFace	SCFace	2,77	**8,26**	**18,71**
		LFW	3,08	9,12	20,74
		ForenFace	3,33	10,18	24,40
	ArcFace	SCFace	0,94	2,94	7,57
		LFW	1,82	5,35	11,98
		ForenFace	**0,58**	**1,83**	**4,85**
XQNet - ConvNext (Ours)	Facenet512	SCFace	**2,12**	**6,22**	**13,50**
		LFW	0,39	0,99	**1,75**
		ForenFace	2,52	7,80	18,45
	SFace	SCFace	**2,75**	8,28	19,13
		LFW	2,95	8,79	20,18
		ForenFace	**3,25**	**10,03**	24,05
	ArcFace	SCFace	0,85	2,68	7,08
		LFW	1,78	5,29	11,99
		ForenFace	0,67	2,12	5,62
XQNet - EfficientNet (Ours)	Facenet512	SCFace	2,33	6,90	15,21
		LFW	0,38	0,98	1,92
		ForenFace	**2,41**	**7,44**	**17,42**
	SFace	SCFace	2,90	8,71	20,28
		LFW	2,94	**8,70**	**19,86**
		ForenFace	3,33	10,21	**24,04**
	ArcFace	SCFace	**0,76**	**2,40**	**6,30**
		LFW	1,72	5,07	11,33
		ForenFace	0,60	1,91	5,08

Table 4.3: Table of results for DABP method in LFW agains random generated pairs and XQLFW generated pairs.

			pAUC (%)		Ratio of unconsidered images (%)
Quality Method	**Face Recognizer**	**Database**	**5**	**15**	**35**
SER-FIQ [115]	Facenet512	LFW	0,69	2,02	4,56
		XQLFW	3,01	9,25	22,23
		DAQP	0,39	1,11	2,40
	SFace	LFW	**2,87**	8,58	19,89
		XQLFW	3,15	9,70	23,49
		DAQP	**2,93**	8,79	20,27
	ArcFace	LFW	**1,77**	5,29	12,18
		XQLFW	2,54	7,84	19,09
		DAQP	1,71	5,14	12,00
EQFace [149]	Facenet512	LFW	0,69	1,98	4,20
		XQLFW	3,01	9,23	22,36
		DAQP	**0,31**	**0,87**	1,91
	SFace	LFW	2,88	**8,56**	19,73
		XQLFW	3,15	9,67	23,59
		DAQP	2,99	8,99	20,94
	ArcFace	LFW	**1,77**	5,25	11,96
		XQLFW	2,54	7,82	19,18
		DAQP	**1,63**	**4,88**	**11,18**
SDD-FIQA [123]	Facenet512	LFW	**0,68**	**1,89**	**3,76**
		XQLFW	**2,98**	**9,00**	**21,17**
		DAQP	0,50	1,28	2,35
	SFace	LFW	2,88	8,58	**19,68**
		XQLFW	**3,12**	**9,46**	**22,50**
		DAQP	3,08	9,12	20,74
	ArcFace	LFW	**1,77**	**5,22**	**11,82**
		XQLFW	**2,51**	**7,65**	**18,34**
		DAQP	1,82	5,35	11,98
XQNet - ConvNext (Ours)	Facenet512	LFW	**0,68**	1,91	3,83
		XQLFW	3,00	9,15	22,06
		DAQP	0,39	0,99	**1,75**
	SFace	LFW	2,88	8,60	19,79
		XQLFW	3,14	9,60	23,30
		DAQP	2,95	8,79	20,18
	ArcFace	LFW	**1,77**	5,27	11,92
		XQLFW	2,53	7,74	18,91
		DAQP	1,78	5,29	11,99
XQNet - EfficientNet (Ours)	Facenet512	LFW	**0,68**	1,93	4,05
		XQLFW	2,99	9,08	21,56
		DAQP	0,38	0,98	1,92
	SFace	LFW	2,89	8,64	19,91
		XQLFW	3,13	9,52	22,84
		DAQP	2,94	**8,70**	**19,86**
	ArcFace	LFW	1,78	5,29	12,07
		XQLFW	2,52	7,68	18,54
		DAQP	1,72	5,07	11,33

In both tables, the fraction of images that were deemed "unconsidered" refers to the percentage of images that were excluded (based on their quality scores) from the calculation of FNMR. If the quality model is accurately assigning quality scores, then discarding lower quality images for the computation of FNMR should result in improved performance and thus, lower FNMR. Conversely, if the FNMR deteriorates upon discarding low quality images, it suggests that these images were not in fact of low quality. In tables 4.2 and 4.3 best results for each database have been marked in bold. We observe that SDD-FIQA [123] and EQFace [149] tend to have better performance on LFW and XQLFW benchmarks, whereas XQNET (ours) has better performance with the DAQP evaluation protocol.

Thirdly, our method must be able to perform competitively in a large-scale dataset (such as CCTV footage). It is for that reason we make a table (see table 4.4) with performing times both with CPU and GPU. The CPU used was AMD EPYC 7B12, and the GPU used was Tesla T4. As seen in the table, our method performs competitively against other state-of-the-art algorithms such as SER-FIQ [115], EQFace [149] and SDD-FIQA [123].

Finally, we analyze the explainability component of the XQNets. The set of attributes is predicted and plotted in 4.7 and 4.8. In Figure 4.7, the correlation between pairs of variables is depicted for each database. The visual representation of the correlation between the variables can provide important insights into the relationship between the variables under study. The more concentrated and circular the curves are, the greater the correlation between the pair of variables. This can be seen in the quality/sharpness graph, where the dataset LFW lines occupy a very small area of the plane, indicating a strong correlation between these two variables. Conversely, in the hotspots/quality graph, the red curves (SCFace database) are widely dispersed, covering a fairly extensive area, indicating a weaker correlation between the two variables. This information can be used to make informed decisions about the variables that are most important to focus on for a given study or analysis. In addition to the correlation between variables, the diagonal of the figure also provides information about the distribution of each individual variable. Some variables have a pointed distribution, with a greater concentration around a dominant value, such as quality or hotspots in LFW database. Other variables have a flattened distribution, showing an almost uniform distribution within a range, such as age in LFW. Understanding the distribution of each variable can inform data preprocessing and modelling decisions, as well as provide insight into the underlying structure of the data.

Figure 4.8 displays the distribution of each categorical parameter value for each database. For example, in the LFW database, with regards to ethnicity and gender, it can be observed that there are no samples of women from the Middle East, while the quality of those for

Table 4.4: Computation times of different networks. N. Det. Imgs refers to Number of Detected images.

Database	N. Det. Imgs.	FIQ Algorithm	CPU Time	GPU Time
SCFace	3841	SER-FIQ	18h 22' 44"	15' 20"
		SDD-FIQA	8' 23"	48.2"
		EQ-Face	17' 51"	1' 18"
		xQNet-ConvNext	3' 56"	2' 4"
		xQNet-Efficientnet	3' 0"	1' 39"
LFW	13229	SER-FIQ	118h 5' 17"	37' 10"
		SDD-FIQA	2h 35' 42"	2' 0"
		EQ-Face	1h 14' 3"	3' 45"
		xQNet-ConvNext	5' 37"	42.7"
		xQNet-Efficientnet	2' 17"	37.8"
XQLFW	13140	SER-FIQ	112h 45' 55"	52' 0"
		SDD-FIQA	2h 35' 19"	1' 58"
		EQ-Face	1h 16' 55"	3' 42"
		xQNet-ConvNext	5' 31"	41.6"
		xQNet-Efficientnet	2' 21"	37.0"
ForenFace	2476	SER-FIQ	16h 20' 0"	9' 20"
		SDD-FIQA	24' 21"	31.8"
		EQ-Face	12' 12"	52.7"
		xQNet-ConvNext	1' 54"	49.2"
		xQNet-Efficientnet	1' 26"	48.3"

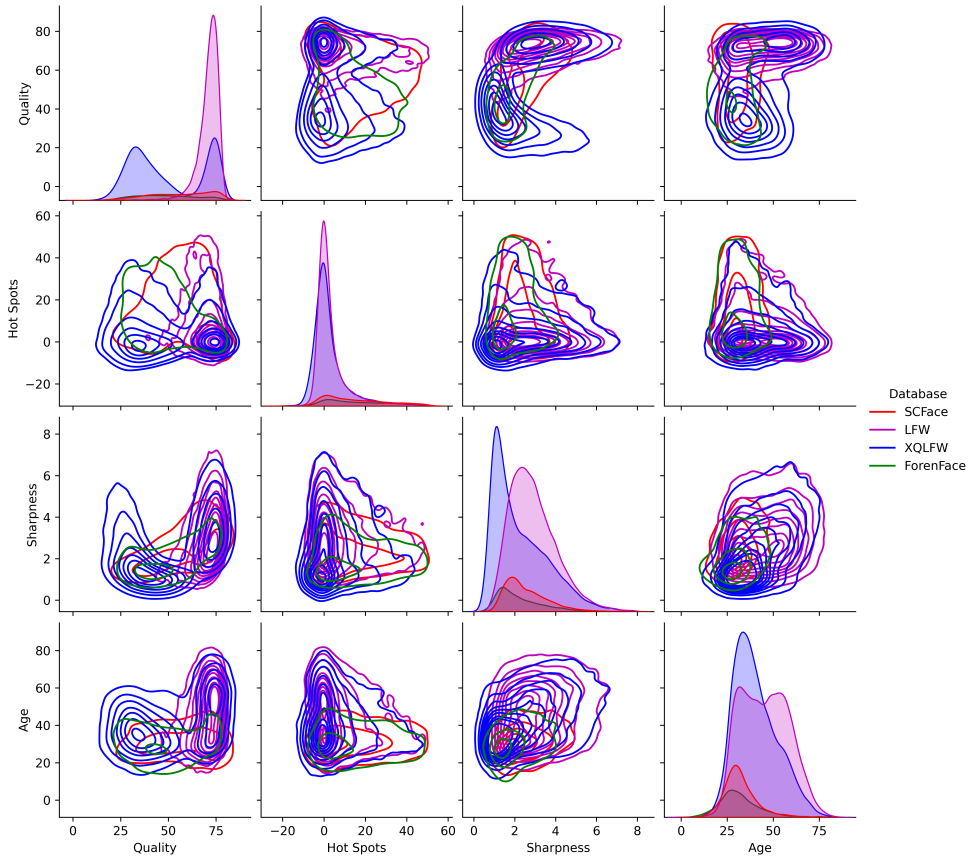


Figure 4.7: Attribute regression predictions by XQNet

the same ethnicity for men is highly concentrated. The elongated and thick graphs with little variation indicate a wide dispersion of the variable under study (quality), as can be seen with the Indian ethnicity in the SCFace database. In addition to the information about the distribution of each parameter value, figure 4.8 can also provide valuable insights into the underlying structure of the data. The distribution of values for each parameter can reveal the presence of biases or imbalances in the data, which can affect the results of further analysis and modelling. Understanding these patterns can inform the development of data preprocessing and balancing techniques, as well as provide guidance for future data collection efforts. Furthermore, the distribution of values for each parameter can also inform decision-making in terms of model selection and performance evaluation.

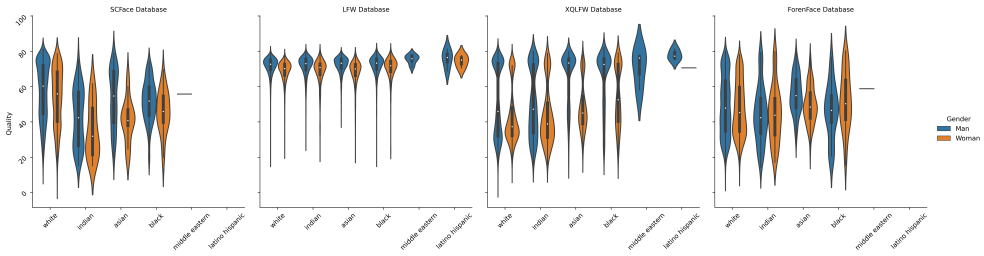


Figure 4.8: Attribute categorical predictions by XQNet

4.6 Discussion and Conclusion

Current definitions of face quality assessment are based on the suitability of a face image for the task of face comparison. However if these FIQ applications are meant to be used by humans, such as in forensics, this suitability score has to be accompanied with a sufficient degree of explainability. This explainability can be achieved through pixel values, such as the work of [119], or by measuring a set of standard attributes and weighting the contribution of each of them such as the work of [117]. In this work, we have chosen to develop a multi task learning model that jointly learns the suitability score with the facial and environmental attributes that contribute to it. The results show that FIQ highly correlates with sharpness, frontal pose and age. This can help the user to get real-time feedback on how to improve the quality of the image before further processing. Also, for forensic purposes in large databases, clusters of images with different qualities and different attributes can be produced to facilitate the investigation. As a caveat, it has to be mentioned that in cases where proper face detection and alignment are not possible, these images cannot be considered for computation. However, manual detection and alignment may be performed with the use of adequate software. This is important because XQNet relies on detecting and aligning faces as the input to our model because the quality assessment should be performed on an aligned face in order to ensure accurate results. The model expects an aligned face as input in order to process the quality assessment in an adequate manner. Another limitation of our current work is that it depends on the availability of labeled data for training and evaluating the model. This implies that the balance of the datasets used of training can have an impact on the results, and the results obtained from our work only apply to the datasets and embedding models used in the study and may not generalize to other datasets or models. Another aspect of training the score together with the attributes is that if the attributes are carefully chosen, unintended bias can be avoided. When the suitability estimation (i.e. facial image quality) is built on the deployed face recognition al-

gorithm, unintended bias can happen. Training several attributes can avoid this bias both in the training and the datasets chosen. Additionally, we could consider handling non-categorical variables that cannot be regressed directly. For example, continuous attributes could be discretized into ordinal bins.

As a conclusion, this chapter proposes a novel FIQ assessment approach, which adds explainability as FIQ annotation. The novelties of our algorithm are three-fold: First, we are the first to train a Multi-Task learning model considering several attributes that affect quality estimation of a face image. Second, we propose a new protocol to evaluate the traditional benchmarks such as LFW, but with a larger number of pairs and equal distribution of qualities. Third, an efficient implementation of multitask learning model shows that it speeds up the label generation and has competitive inference times. Our proposed method combines regression and classification, allowing it to be retrained for different labels (e.g., from quality to another type of float) or classes (e.g., from gender to another binary classification). This adaptability makes it suitable for tasks like person re-identification by adjusting labels and classes accordingly.

5

Improved Likelihood Ratios for Face Recognition in Surveillance Video by Multimodal Feature Pairing

The ability to accurately recognize faces in real-world surveillance videos based on a set of given images of a suspect and assess its value as evidence are critical aspects of forensic investigation and security monitoring systems. This task is affected by variations in pose, illumination, and facial expression that are commonly present in such videos. Currently, in cases where face comparison results are presented in court, manual facial comparison methods, such as holistic, morphological and photoanthropometric processes, are used. But these methods lack standardization and validation. These complexities highlight a critical gap: existing automatic methods may not be sufficient for robust face recognition in these challenging scenarios because of their susceptibility to the aforementioned variations. So, how can we enhance the reliability of facial recognition in forensic settings? We propose a method for image-to-video face recognition in challenging forensic scenarios by utilizing a new model that pairs a face image with multiple attributes, such as pose and facial expression, and face image quality. To statistically assess the strength of the evidence in a forensic investigation, we then apply three calibration methods to estimate the likelihood ratio. We validate the results of our proposed method, using the log-likelihood ratio cost (C_{llr}), on the ENFSI proficiency test 2015 dataset, using SCFace, XQLFW, ChokePoint and ForenFace as calibration datasets. We use three

face recognition models: ArcFace, FaceNet and QMagFace. Our results suggest that while using different viewpoints may improve recognition, focusing on higher quality frames alone can enhance face recognition performance for forensic purposes compared to using all frames. The best C_{llr} was achieved by employing the highest number of common attributes of the reference image and selected frames. Compared to using the top 25% best quality frames, this approach yields similar C_{llr} values. The second-best method involves creating a single common embedding from the selected frames and weighting it by the quality of each frame's face image. Upon preprocessing facial images with the super resolution CodeFormer, we observed an unexpected increase in the log-likelihood ratio cost, reducing the reliability of the evidence. Consequently, we discourage the use of CodeFormer in these forensic scenarios due to its detrimental impact on facial recognition performance.

5.1 Introduction

Automated Face recognition (FR) is a method that has become increasingly important in recent years, particularly in the field of forensic investigation [114]. With the proliferation of surveillance cameras and the capture of images of criminal events, the comparison of faces has become a key tool for gathering intelligence, guiding investigations, and providing evidence in court [114][6]. While deep-learning based FR methods have demonstrated strong recognition performance for still images [150], such as those in the Labeled Faces in the Wild (LFW) dataset [40], video-based FR has not been as widely developed by the research community [151]. Video FR, however, offers additional information, such as temporal details and multiple views on the same person, which can be used in conjunction with frame based face recognition techniques to quickly identify subjects of interest in CCTV footage [152].

Despite the potential benefits of video-based FR, the process of analyzing such a large amount of data for each video is challenging due to the time needed to deal with all frames. Not all the frames in the video might be of equal importance though. Some frames can be useless for recognition due to low video quality, motion blur, occlusions, and frequent changes in the scene [153; 154] (see figure 5.1 for examples). An obvious method would then be to measure such characteristics and discard frames of low quality. Some works focusing on face image quality (FIQ) [115; 116; 123], however, have indicated that using human-based attributes for face image quality assessment might not be ideal. Aspects that humans perceive as affecting the quality of an image, such as illumination or pose, may not be the best characteristics for the face recognition system being used. The references above

Figure 5.1: Example of images-to-video scenario. Images taken from [105]. The reduced quality in this case is primarily attributed to the following factors: the challenging pose of the face, the subject wearing a cap, and increased subject distance.

use the SER-FIQ, MagFace, and SDD-FIQA face image quality assessment deep learning based methods to test on IJB-C [155] videos, and show that for 1:1 recognition on individual frames these assessment methods yield significant improvement. In current systems, image quality measures incorporating spatio-temporal information are not used.

So how to evaluate which method is best? In automated facial recognition systems, the similarity between two samples is usually reported in one or several score values intrinsic to each version of the facial recognition algorithm used [150]. To allow comparisons between facial scores from different face recognition systems, as well as for such an automated comparison to be useful in an evaluative forensic framework, there is a need to map the output scores to a Likelihood Ratio (LR) [73]. LR is defined as the probability of the evidence given hypothesis H_0 i.e., the probability of the reference being the same person as in the video, divided by the probability of the evidence given the alternative hypothesis H_1 i.e., the probability of the reference being a different person than the one appearing in the video. A possible approach to achieve this is use a score-to-LR mapping as a post-processing step in an existing score-producing facial recognition system [24]. Once a model for score-to-LR mapping has been set up, the forensic reporting can be presented using a level of conclusion, where each grade on the scale is connected to an interval of LR values [6; 83].

In this chapter, which is an extension of our conference paper [156], we propose a novel method for image-to-video face recognition in realistic forensic scenarios. We leverage a model that pairs face images based on multimodal face feature data, such as face attribute characteristics and FIQ. The aim is to accurately estimate likelihood ratios (LRs) for face recognition systems in practical settings. Our particular focus is on scenarios where multiple reference images of a suspect are available, and to verify if this person is the same individual appearing in a surveillance video. Previous studies, such as the work of Molder et al. [24], Rodriguez et al. chapter 3, and Jacquet et al. [6], have explored LRs in face recognition in still images, employing different techniques and considering various scenarios. Despite these contributions, there are still open questions, particularly regarding the accuracy of estimating LRs and their effective application in a forensic context. This chapter aims to address these gaps by improving the accuracy of LR estimation in automated face recognition using image-to-video comparisons, building on the work of researchers like Zheng et al. [154] and Huo et al. [157]. We apply three calibration methods to estimate LRs and validate the results using the log-likelihood ratio cost (C_{llr}). Our contributions include the following:

1) **MultiModal Feature Pairing** using FIQ to select frames with the highest quality and highest number of common attributes (soft labels), and combining them through a weighted average.

2) **Calibration** involving selection of random pairs with the same attributes and same FIQ as the test pairs.

3) **Validation** of the LR estimation system against a forensic test performed with hu-

man experts.

4) **Preprocessing with Super Resolution** method CodeFormer for preprocessing facial images and evaluate its effect on the C_{llr}.

The current study begins by providing an overview of the relevant literature pertaining to the estimation of likelihood ratios, face recognition in video, and the incorporation of FIQ in face recognition in images-to-video scenarios. Following this, the methodology for pairing and calibration is presented. The experiments and associated results are then discussed. Finally, the chapter concludes with a discussion of the findings and implications. The workflow for the computation of the Likelihood Ratio, giving a blueprint for the chapter, is depicted in Figure 5.2.

Compared to our earlier conference paper [156], this study introduces several significant advancements. Firstly, we implement a new computational model, CodeFormer [158], designed to optimize face recognition performance. Secondly, we expand our dataset collection to include XQLFW [**?**] and ChokePoint [159], enriching the empirical foundation of our research. We offer enhanced interpretability of our results by introducing a sunburst diagram as a novel visualization tool to better understand the relationships between various facial attributes and image quality. Finally, the text has been significantly extended to give more insight in the methodologies and results.

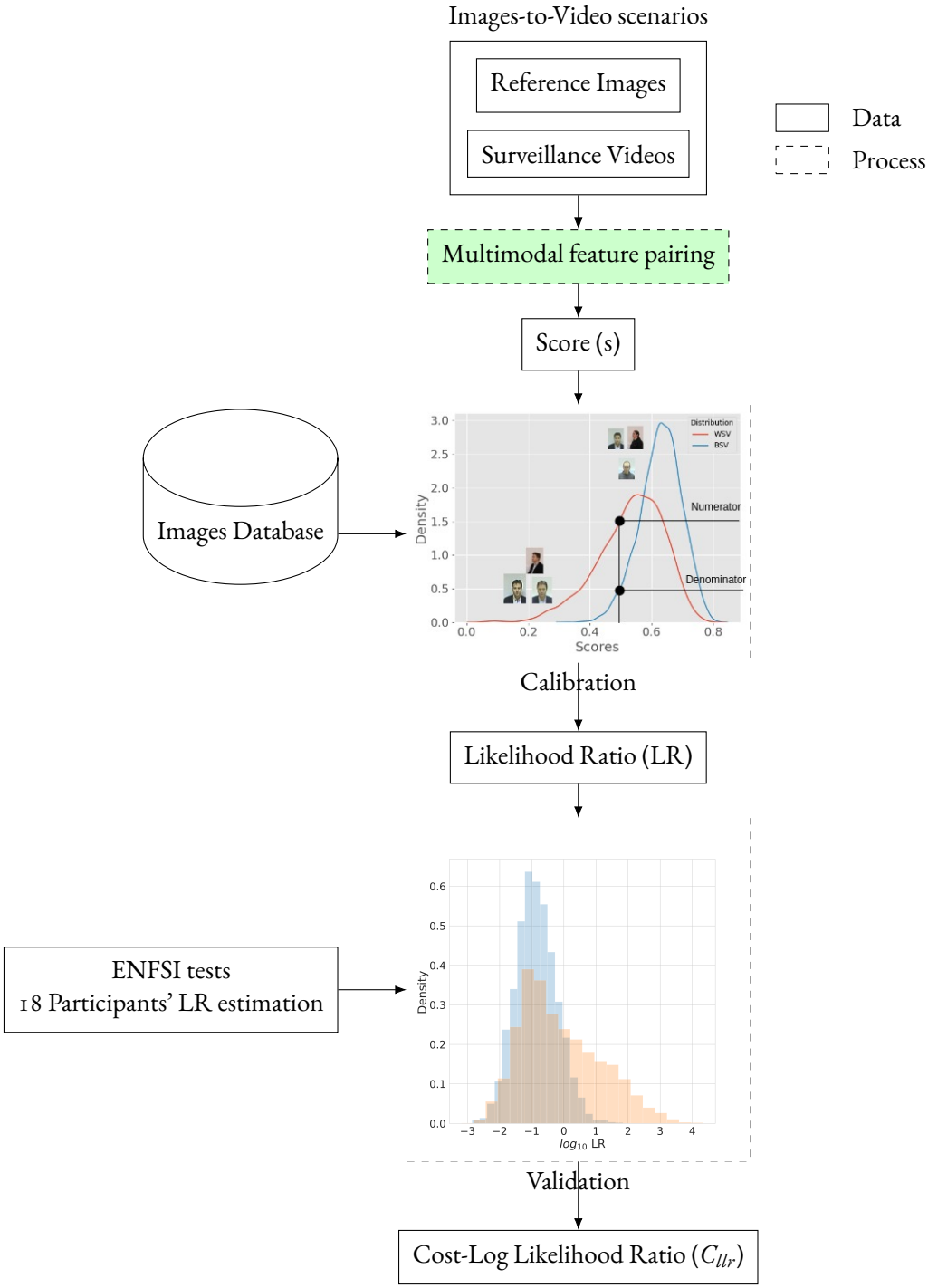


Figure 5.2: Workflow of the LR computation and validation process in ENFSI 2015 proficiency test [160].

5.2 Related Work

Likelihood ratios (LRs) have been applied in the field of face recognition. Molder et al. [24] test score-to-LR models in forensic data and find that the performance of the models is highly dependent on the available training data. Rodriguez et al. (chapter 3) and Jacquet et al. [6] also focus on this topic, with the former using facial attributes and quality scores to improve LR estimation, and finding that current commercial software outperforms open-source software. The latter reference explores the importance of LR in face recognition and assesses the performance of the model with respect to its discriminating power and calibration state. While these studies have made significant strides, the current state of LR research in video-based face recognition remains incomplete. An open question is how to accurately estimate LRs for face recognition systems in practice and how they can be effectively deployed in a forensic context, particularly when analyzing video sequences rather than isolated frames.

Spatio-temporal face recognition in videos has also been a topic of research. Zheng et al. [154] propose a system for image-to-video face recognition in unconstrained conditions, composed of modules for landmark detection, face association, and face recognition. They perform experiments on video datasets and demonstrate that their system can accurately detect and associate faces from unconstrained videos and effectively learn robust and discriminative features for recognition. Huo et al. [157] tackle n-shot face recognition in videos using metric learning methods and similarity ranking models, comparing a Siamese network with contrastive loss to a Triplet Network with triplet loss. They show that feature representations learned with triplet loss are significantly better in their setting, and that learning spatio-temporal features from video sequences is beneficial for face recognition in videos. Rivero et al. [161] propose an adaptive aggregation scheme based on ordered weighted average (OWA) operators, and develop two different implementations to validate its suitability for image-to-video face recognition. Nevertheless, the current state of spatio-temporal face recognition research is insufficient, as the problem of face recognition in forensic videos is still open, and the results are not generalizable to real-world scenarios.

To avoid processing a whole video, keyframe extraction methods for face recognition in videos have been developed. Abed et al. [151] propose a method based on face quality and deep learning. The first step is the face detection using the MTCNN detector, which detects five landmarks (the eyes, the two corners of the mouth, and the nose) and then limits face boundaries to a bounding box and from there provides a confidence score. This method involves two steps: the generation of face quality scores using three face feature

extractors (Gabor, LBP, and HoG), and the training of a deep Convolutional Neural Network to select frames with the best face quality. Bahroun et al. [153] propose a keyframe extraction method based on face image quality for video surveillance systems. Data is reduced by rejecting frames without faces, and then face images are clustered by identity. A set of candidate frames is then selected, and the face quality assessment is based on four metrics (pose, sharpness, brightness, and resolution). The frame with the best face quality is considered a keyframe. Experimental tests were conducted on several datasets to demonstrate the effectiveness of the proposed method compared to other state-of-the-art approaches. The issue with some existing methods of face image quality computation is their dependence on subjective or indirect measures of quality, which may not necessarily align with the needs of face recognition systems. In contrast, these newer methods, as exemplified by the works of Abed et al. [151] and Bahroun et al. [153], provide a more direct measure of face image quality, which is closely tied to the performance of the face recognition model itself.

Face image quality assessment for improving face recognition in videos has also been considered. Terhorst et al. [115] propose the SER-FIQ (Subjective and Objective Quality Factors of Images) method for assessing face image quality. They test the SER-FIQ method on the IJB-C [155] video dataset and show that it performs well in face recognition tasks. Meng et al. [116] propose the MagFace method, which uses a multi-attention guided face image quality assessment network to evaluate face image quality. They test MagFace on the IJB-C videos and show it outperforms other state-of-the-art methods. Ou et al. [123] propose the SDD-FIQA (Single Shot Detector based Face Image Quality Assessment) method, which uses a single shot detector to evaluate face image quality. They test SDD-FIQA on the IJB-C videos and show that it performs well in face recognition tasks. However, these works only evaluate face image quality in 1:1 (face verification) image-to-video scenarios, and do not consider the use of temporal information for face recognition as they use the frames as if they were isolated images.

Blind face restoration, which refers to the task of restoring faces in images without knowledge of the specific degradation processes they underwent, presents a complex challenge due to the inherent uncertainty stemming from its ill-posed nature and the potential loss of crucial details in degraded inputs. Together with super resolution techniques, they are emerging as important tools in improving image quality for various tasks, including face recognition. In a recent study, a novel approach has been proposed to handle the problem of blind face restoration, which is typically a highly ill-posed problem. Zhou et al. [158] introduces a learned discrete codebook prior in a small proxy space, reducing the uncer-

tainty and ambiguity of restoration mapping by casting the process as a code prediction task. The approach is called CodeFormer, a Transformer-based prediction network that models the global composition and context of low-quality faces for code prediction. This technique enables the discovery of natural faces closely approximating the target faces, even with severely degraded inputs. The study showed that CodeFormer outperforms state-of-the-art methods in both quality and fidelity, exhibiting superior robustness to degradation. The results were validated on both synthetic and real-world datasets, further underscoring the effectiveness of the method in addressing the challenges of face restoration and super resolution. Despite these advancements, the use of such advanced preprocessing techniques for image-to-video face recognition, particularly in the context of forensic investigations, is still an open research question.

5.3 Methodology

We propose a systematic workflow, illustrated in Figure 5.2, that is segmented into various interconnected stages. The process commences with the curation of 'Images-to-Video scenarios' incorporating both reference images and surveillance videos. These inputs undergo a 'Multimodal feature pairing' stage, further detailed in Figure 5.3. In this stage, all frames are compared to one another; frames of the highest quality are paired, as are frames with shared attributes between the reference images and the video. Additionally, a frame weighted by quality from the reference images is paired with a similarly weighted frame from the video. After generating a biometric score s, the workflow advances to the 'Calibration' phase. During this stage, scores are calibrated using distribution models derived from both within-source variability (WSV) and between-source variability (BSV). This calibration is done with data from an 'Images Database.' Subsequently, the calibrated scores are transformed into a 'Likelihood Ratio (LR),' which is then subjected to a 'Validation' phase. During validation, external data from 'ENFSI tests' involving the LR estimations of 18 participants, is incorporated. This offers a robust evaluation mechanism for the computed LRs, utilizing the 'Cost-Log Likelihood Ratio (C_{llr})' as a metric to evaluate the strength of our evidence [73]. We will proceed to further develop these concepts.

To estimate the LR as a measure of the strength of the evidence, the LR, expressed as the Score based Likelihood Ratio (SLR), is defined as:

$$SLR(s) = \frac{P(s|H_p, I)}{P(s|H_d, I)}, \tag{5.1}$$

where s is the biometric score, H_p is the null hypothesis that the evidence originates from the same source, and H_d is the alternative hypothesis that it comes from a different source. Logistic Regression is employed to fit the probability functions $P(s|H_p, I)$ and $P(s|H_d, I)$, considering the background information available in the case.

The workflow involves face detection, pairing of reference images and video frames, calibration of biometric scores using WSV and BSV, and validation against human performance. Following this methodology allows us to assess the likelihood of a person being present in a surveillance video, thus assisting in forensic investigations.

We propose an enhancement to the methodology by processing all frames in the video where a face is detected. For each of these frames, we compute the Face Image Quality (FIQ) and create an embedding vector e_i, which represents the compressed representation of facial features. The FIQ scores are then used to apply a weighting scheme when combining the embedding vectors to form:

$$\mathbf{e_{face}} = \sum_{i=1}^{n} q_i * \mathbf{e_i}, \tag{5.2}$$

This approach is applied to both the video frames and the reference images, allowing for a more comprehensive representation of the facial information. By incorporating FIQ-based weighting, we aim to improve the accuracy and reliability of face recognition in image-to-video comparisons.

The question we aim to answer is: How likely is this person the same as the one appearing in the surveillance video? To that end, we propose a workflow as seen in figure 5.2. Our focus is on the comparison of several reference images of the same person to a video in order to determine if the person appears in the video.

To estimate the likelihood ratio, the biometric score obtained from the comparison between the images and the video has to go through a process of calibration in which two distributions are computed: the WSV and the BSV. In this chapter, we focus on two specific aspects of this process as it pertains to images-to-video comparisons: (1) methods for pairing reference images with videos, and (2) the use of different types of images, such as different qualities or different attributes, to create the WSV and BSV distributions during the calibration step. The biometric score must be calibrated using these distributions to estimate the likelihood ratio.

5.3.1 MultiModal Feature Pairing

In this work, we aim to improve the accuracy of likelihood ratio (LR) estimation in automated face recognition using images-to-video comparisons. Examples are shown in figure 5.3.

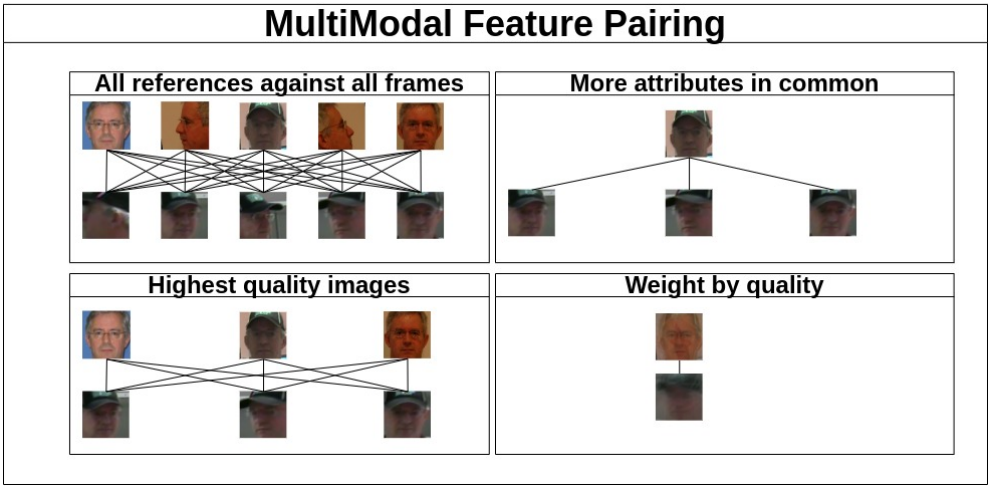


Figure 5.3: Examples of multimodal feature-pairing

One approach is to employ score pairs derived from the shared attributes of the reference image and the video frame. Let $S(i, v)$ denote the score for a given image i and video frame v. We define the score based on shared attributes:

$$S(i, v) = \sum_{a \in A} \delta(a_i, a_v) \tag{5.3}$$

where A is the set of all attributes, and δ is the Kronecker delta function. $\delta(a_i, a_v)$ is 1 if attribute a in image i matches attribute a in video v, and 0 otherwise.

Initially, we extract and calculate various attributes from all reference images and video frames, encompassing gender (for simplicity only comprising the categories man and woman), facial expression (including happy, angry, fear, and neutral), ethnicity (encompassing white, Asian, black, and Middle Eastern), yaw (representing frontal, slightly turned, and sideways orientations), pitch (Up, slightly up, frontal, slightly down, down), roll (frontal, slightly rolled, completely rolled), headgear, glasses, beard, and other occlusions (all of the latter booleans with value yes or no). Subsequently, we compare the attributes of each reference image with the attributes of every video frame, select pairs that exhibit the highest count of shared attributes, and we conduct likelihood ratio estimation defined in equation 5.1.

By considering the attributes and iterating through various numbers of shared attributes, the algorithm can make more informed decisions, potentially enhancing the accuracy of the face recognition system. A summarized depiction of the algorithm can be found in algorithm 2.

Algorithm 2 Pair frames with most attributes in common

Input: Surveillance video, reference images
Output: LR estimation
for each reference image **do**
  Extract and compute attributes
for each video frame **do**
  Extract and compute attributes
Initialize a list of all score pairs $S(i, v)$ using the defined formula in 5.3
$n \leftarrow$ maximum number of attributes in common
while $n > 0$ **do**
  Select all score pairs that have n attributes in common
  Compute *SLR* for each selected pair using the defined formula
  $n \leftarrow n - 1$
return SLR estimate for the highest score

An alternative approach for performing the pairing is to match all the reference images with all the video frames, and then order them according to their quality. Once sorted by quality, the LR is calculated using all pairs. Subsequently, a process of pruning is applied, starting with the removal of 10% of the pairs with the lowest quality, followed by the removal of an additional 10% of the pairs, etc. The objective of this method is to determine if the information lost by discarding pairs is valuable, i.e. the SLR improves, which would indicate that the discarded images were noisy and thus detrimental to the face recognition system. An algorithm for this method is presented in algorithm 3.

In addition, we propose to process all the frames in which a face is detected in the video, compute the FIQ of each frame, and create a combined embedding vector for the video using a weighting scheme based on the FIQ scores. Similarly, we process all the available reference images. This method is based on the equation 5.2. This process is applied to both the video frames and the reference images. A summary of this process can be found in algorithm 4.

An alternate approach involves integrating the techniques used in Experiments 1-4, while also adding an extra stage of preprocessing through the use of the super resolution CodeFormer. Our goal is to evaluate how such sophisticated image preprocessing might influ-

Algorithm 3 Pair only the highest quality frames

Input: Surveillance video, reference images
Output: LR estimation
Pair all reference images with all video frames
for each pair of frames **do**
  Assign the lower image quality value to the pair
Sort pairs by quality (lower values first)
Compute LR using all pairs
$p \leftarrow 10\%$
while $p \leq 100\%$ **do**
  Discard an additional $p\%$ of the remaining pairs with the lowest quality
  Compute LR using the remaining pairs
  $p \leftarrow p + 10\%$
return LR estimation

Algorithm 4 Weight all references and frames by quality

Input: Surveillance video, reference images
Output: LR estimation
for each frame in video **do**
  if face is detected **then**
    Compute FIQ q_i
    Compute embedding $\mathbf{e_i}$ of face image
Create combined embedding vector $\mathbf{e_{face}}$ of video using Equation 5.2
for each reference image **do**
  Compute FIQ q_i
  Compute embedding $\mathbf{e_i}$ of face image
  Create combined embedding vector $\mathbf{e_{face}}$ of reference image using Equation 5.2
Compare $\mathbf{e_{face}}$ of video and reference images to calculate LR estimation
return LR estimation

ence the accuracy and dependability of face recognition. See example of pre-processing in figure 5.4.

Figure 5.4: Example of super resolution image by Codeformer [158], on the left, the original, on the right, the processed image

5.3.2 Calibration

To improve the accuracy of the LR estimation for automated face recognition in video, we will consider three different approaches for selecting images from the calibration database to use in the estimation process. The baseline consists of using random images from the calibration database:

- **Same attributes:** Using images with the same attributes as the reference and video, such as pose or facial expression.

- **Quality pairs:** Using pairs that have the same FIQ group for the reference face and the combined face image qualities of the video frames. The FIQ group categorizes FIQ values into very low quality, low quality, medium quality, high quality, and very high quality.

By implementing these approaches, we aim to improve the accuracy of the LR estimation for automated face recognition in video.

5.4 Experiments

We will explore the workflow explained in section 5.3 doing experiments in the two parts of the method: pairing and calibration.

5.4.1 Datasets

Our study encompasses multiple datasets: ENFSI proficiency test [160], ForenFace [105], SCFace [29], and with respect to [156], we added the datasets XQLFW [**?**], and ChokePoint [159].

The ENFSI proficiency test 2015 focuses on matching mugshot images to CCTV video and includes 18 individual participants in 17 comparisons. ForenFace contains video sequences and extracted images of 97 subjects recorded with six different surveillance cameras. Its novelty lies in a subset of 435 images manually annotated, yielding forensically relevant annotation of almost 19,000 facial parts. SCFace has images taken in an uncontrolled indoor environment using five video surveillance cameras, consisting of 4160 static images and frames (in visible and infrared spectrum) of 130 subjects. The XQLFW dataset is a variant of the well-known Labeled Faces in the Wild (LFW) that focuses on cross-quality cases. It emphasizes the quality difference by containing only more realistically degraded images when necessary. It aids in assessing the robustness of face recognition models against various image quality challenges. ChokePoint was designed for real-world surveillance conditions. It was captured above several portals using an array of three cameras. The dataset features variations such as illumination, pose, sharpness, and misalignment. It comprises 48 video sequences and 64,204 face images of 54 subjects.

All these datasets consist of video sequences and face images with variations in illumination, pose, and sharpness. The study's objective is to train and test the performance of the proposed method on these datasets, seeking the most effective method to enhance the accuracy of the LR estimation for automated face recognition in video. A summary of these datasets can be found in table 5.1.

5.4.2 Face recognition models and face quality models

We chose to use three face recognition models, ArcFace, Facenet, and QMagFace [126], in our experiments, because they all have been proposed recently, have demonstrated state-of-the-art performance, and all have different characteristics. ArcFace has a clear geometric interpretation and significantly enhances the discriminative power. Facenet directly learns a mapping from face images to a compact Euclidean space where distances correspond to a measure of face similarity, which makes it highly generalizable. QMagFace combines a quality-aware comparison score with a recognition model based on a magnitude-aware angular margin loss, making it suitable to enhance the recognition performance under unconstrained circumstances. We implemented ArcFace and Facenet from [107].

Table 5.1: Summary of the five datasets.

Type	Dataset	Subjects/Images	Cameras	Description
Calibration	ForenFace	97/4000	6	Forensic annotations Video & images
	SCFace	130/4160	5	Indoor images Static imgs & frames
	XQLFW	3743/7263	N/A	LFW [40] variant Emphasizes FIQ Degraded images
	ChokePoint	54/64,204	3	Dif. Pose & illumination Video sequences
Test	ENFSI	18/NA	N/A	Mugshot and CCTV Individual comparisons

We use two quality models, SER-FIQ [115] and SDD-FIQA [123], as they both are unsupervised methods that have been shown to outperform state-of-the-art approaches in face image quality assessment and have good generalization across different recognition systems. SER-FIQ is based on the robustness against dropout variations as a quality indicator, and avoids the training phase completely. SDD-FIQA generates quality pseudo-labels by calculating the Wasserstein Distance (WD) between the intra-class and inter-class similarity distributions, which has been demonstrated to surpass state-of-the-art methods by an impressive margin.

5.4.3 Experimental cases

-**Experiment 1: Highest Number of Common Attributes**. Explained in algorithm 2. We aim to assess if using pairs that share attributes (multi-attribute, e.g., pairs with the same pose or facial expression) outperforms pairs that have nothing in common. We perform the LR estimation with 10,000 random images from the calibration set and 0 to 6 attributes in common (pitch, yaw, roll, facial expression, age, and gender).

-**Experiment 2: Quality-Based Drop**. Explained in algorithm 3. We aim to assess the influence of using the highest-quality frames on the LR estimation. We perform the experiment using 10,000 random images from the calibration dataset and compute the LR estimation by dropping 10% of the poorest quality face images in each iteration. We use the ENFSI test 2015 dataset for this experiment.

-**Experiment 3: Weighted Face Quality Images**. Explained in algorithm 4. We aim

to assess the effect of weighting frames by quality on the LR estimation. We use 10,000 random images from the calibration dataset for this experiment.

-**Experiment 4: Calibration**. Explained further in the text. Once the comparison of images-to-video is computed, we aim to assess the difference in calibration using random images or images with the same FIQ as the test pair.

-**Experiment 5: Super Resolution Preprocessing**. In this experiment, we incorporate all the methods from Experiments 1-4 but with an additional layer of preprocessing using super resolution CodeFormer [158]. We aim to assess the impact of advanced image preprocessing on the face recognition accuracy and reliability.

5.4.4 Validation

To assess the performance of our proposed methods, we use the log cost likelihood ratio as a measure due to its capacity to represent both discrimination and calibration [73]. C_{llr} is defined as:

$$C_{llr} = \frac{1}{2N_p} \sum_{i_p} \log_2(1 + \frac{1}{SLR_{i_p}}) + \frac{1}{2N_d} \sum_{j_d} \log_2(1 + SLR_{j_d}), \tag{5.4}$$

where the indices i_p and j_d respectively denote summing over the computed SLR scores using equation 5.1 for each face pair comparison. Specifically, i_p sums over cases where the proposition for the prosecutor is true, while j_d sums over cases where the proposition for the defense is true. The variable N refers to the number of samples for each proposition. Minimizing the value of C_{llr} implies an improvement of both discrimination and calibration performance of the automated system [73]. The value ranges from zero (perfect decision making), to infinity (completely wrong). A value of one indicates the system makes a random selection. A value larger than one indicates the system is making a decision worse than random, i.e. supporting the prosecution hypothesis when it should support the defense hypothesis or vice versa.

In addition, we also use boxplots to assess the impact of discarding pairs on the variability of our results, for both i_p (the summands corresponding to the prosecutor's proposition) and j_d (the summands corresponding to the defense's proposition). Specifically, we plot boxplots on the C_{llr} metric for each quality drop, indicating the percentiles of 25, median, and 75. The use of boxplots allows us to visualize the distribution of the C_{llr} metric and better understand how discarding pairs impacts the variability of the results, measure our approach for validating the performance of our proposed methods, and assess the impact

of discarding pairs on the variability of the results.

5.5 Results

This section presents the findings of our investigation. The first subsection focuses on the correlation between facial image quality and various attributes such as gender, pose, race, and facial expression. The second subsection delves into the results of several experiments aimed at understanding the effect of different pairing methods on SLR estimation.

5.5.1 Correlation Between Facial Image Quality and Attributes

To investigate the correlation between facial image quality and various facial attributes, Figure 5.5 presents the results of a study conducted on the test dataset ENFSI 2015. These facial attributes include gender, pose (specifically yaw), race, and facial expression. The results for the calibration datasets—SCFace, XQLFW, ForenFace, and ChokePoint—are presented in Figures 5.8, 5.9, 5.10, and 5.11, respectively, which can be found in the appendix section.

To visually depict the intricate relationship between these attributes and the resulting image quality, we chose to use a sunburst hierarchical graph. This type of graph offers a compact and intuitive representation of hierarchical data across multiple dimensions. The order of attributes in the sunburst graph was chosen strategically to reflect their relevance and potential interactions. Gender, as the first attribute, is an important factor in face recognition and analysis. Its inclusion allows us to examine if gender has a significant influence on the quality of facial images. Next, the attribute of yaw (pose) was selected to explore the impact of different face orientations on image quality. Pose plays a vital role as it can affect the visibility of facial features and details. Analyzing yaw within the sunburst graph enables us to discern how different pose angles relate to image quality. Ethnicity, as the third attribute, is crucial for understanding potential variations in image quality among different racial or ethnic groups. Its inclusion in the graph allows us to identify patterns or disparities that may exist in image quality based on ethnicity. Finally, facial expression was chosen as the last attribute in the sunburst graph. Facial expressions are essential for face recognition and emotional analysis. By including facial expression in the graph, we can assess whether different expressions impact the quality of facial images.

The sunburst diagrams depicted provide a clear visual representation of the relationship between these attributes and the resulting image quality. From the analysis, it is evident that pose plays a significant role in the quality of facial images. Particularly, in terms of

recognition, profile poses are associated with lower quality images, indicating a potential challenge in capturing sufficient detail and features in such poses.

Interestingly, gender does not appear to significantly influence the quality of facial images, suggesting that both male and female faces can be captured with comparable quality under the same conditions.

In terms of race, a noticeable pattern is seen in the XQLFW dataset, where images of individuals of Caucasian ethnicity seem to exhibit higher quality compared to other races. This could be due to many factors, such as lighting conditions, camera characteristics, or image processing techniques, and warrants further investigation.

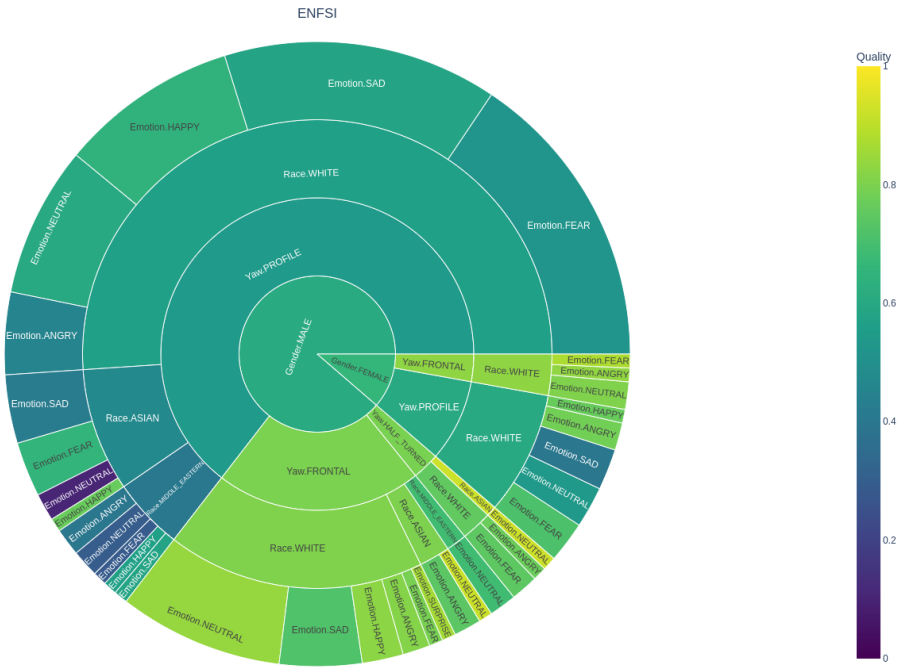


Figure 5.5: Facial attributes in the ENFSI database according to quality model SER-FIQ [115].

Results of four experiments on the effect of using different pairing methods on LR estimation in face recognition in videos are presented in figure 5.6. The rest of figures are included in the annex 5.8.

The outcome of applying super resolution is represented in figure 5.7. The former diagram indicates an enhancement in the quality of face images as per the Face Image Quality (FIQ) metric following the application of the super-resolution algorithm. Conversely, the

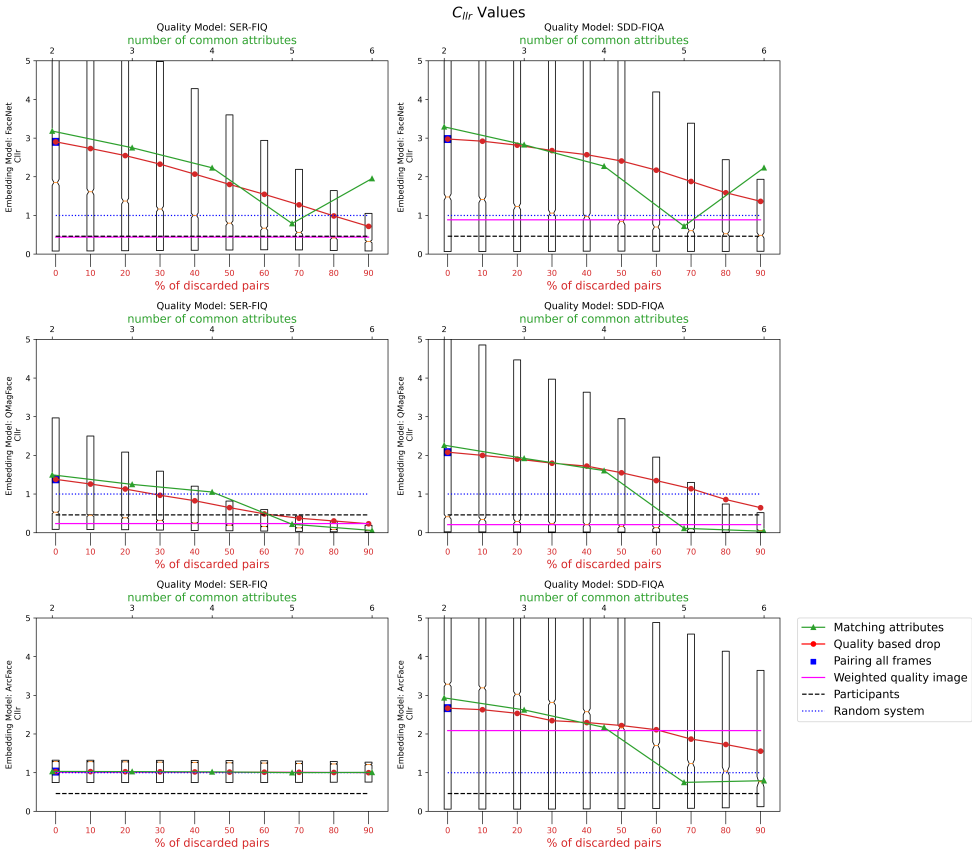


Figure 5.6: Graphical representation of the C_{llr} (log-likelihood ratio cost) values after calibration with attributes: yaw (top), pitch (middle), roll (bottom).

latter figure suggests the results deteriorate after the super-resolution processing.

5.5.2 Effect of Pairing Methods on LR Estimation

-**Results for Experiment 1**. The results indicate that having a higher number of attributes in common between the image pairs significantly lowers the C_{llr} value. This suggests that multi-attribute pairing may be an effective strategy for improving the accuracy of likelihood ratio (LR) estimation in biometric systems.

-**Results for Experiment 2**. Our findings reveal that using higher-quality frames leads to lower C_{llr} values, thereby enhancing the performance of the LR estimation. However, an interesting observation was that adding more frames does not uniformly improve C_{llr}. In certain cases, incorporating lower-quality frames actually led to a worsened C_{llr}, high-

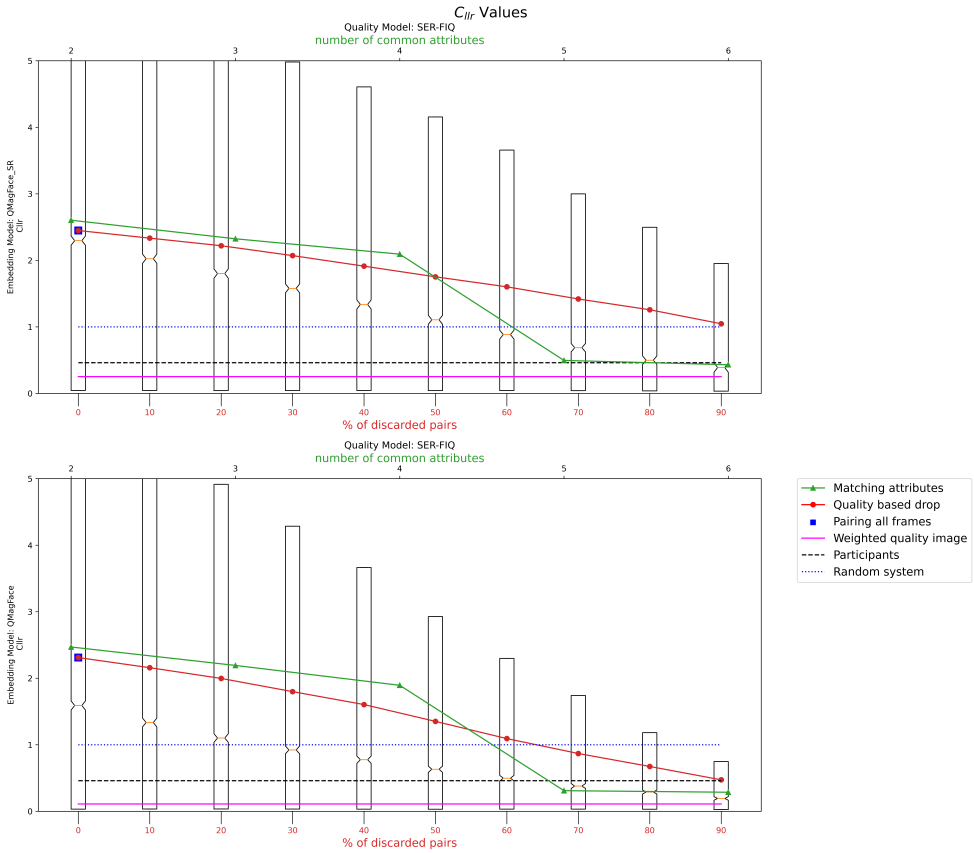


Figure 5.7: Graphical representation of the C_{llr} (log-likelihood ratio cost) values after the application of super-resolution processing (top). In the bottom graph, C_{llr} without pre-processing.

lighting the importance of frame quality in the estimation process.

-**Results for Experiment 3**. When all frames were used but weighted by their quality, the C_{llr} values decreased, suggesting an improvement in the LR estimation. This finding implies that taking quality into account in a weighted manner can improve the system's overall performance, even when low-quality frames are included in the mix.

-**Results for Experiment 4**. We found variability in the calibration based on the set of images used. Using 20,000 images with the same attributes yielded a lower C_{llr} value compared to using 20,000 random images or images of the same quality as the test pair. This suggests that the choice of calibration set can have a substantial impact on the resulting C_{llr} and, by extension, on the performance of the LR estimation.

-**Results for Experiment 5**. Applying an advanced image preprocessing step through super resolution CodeFormer actually had a negative impact on the face recognition sys-

tem's accuracy and reliability. The use of super resolution led to higher C_{llr} values, suggesting that it may not be beneficial for improving the performance of the likelihood ratio (LR) estimation in this specific context.

5.6 Discussion & Conclusion

In our experiments, we found that using higher quality frames improves the performance of face recognition in video compared to using all frames. We explored different methods for pairing reference images with video frames. We found that using images with the same attributes as the reference and video, or similar FIQ score for the reference face and the combined face image qualities of the video frames, can improve the likelihood ratio estimation. Furthermore, we found that using a weighted quality average of all available reference and video frames improved results even more. On the other hand, slightly poorer results were obtained when pairing facial images based on the maximum number of common attributes. Although SDD-FIQA[123] outperforms SERFIQ in the LFW [40] and IJB-C [155] benchmarks, SERFIQ [115] seems more robust in our experiments. The C_{llr} obtained in the best case is close to 1, which is worse than the 0.45 of the expert participants in the ENFSI proficiency test [160]. This could be due to the difficulty and low face quality of the video frames used. Even discarding those with the poorest quality, the remaining ones are not suitable for the face recognition system in this experiment (ArcFace). However, using Facenet as the face recognition system in our experiments, we achieved a C_{llr} of 0.8, which is a better result. With QMagFace, we achieved even better results, with a C_{llr} of 0.26 using the method of the weighted quality image, surpassing the human participants in the ENFSI 2015 test, who scored a C_{llr} of 0.46. The best result was obtained by QMagFace and SER-FIQ with the method of pairing the highest number of attributes in common, with a C_{llr} of 0.13. This demonstrates the effectiveness of using FIQ as a metric to improve the performance of automated face recognition in video surveillance. The boxplots suggest there is less variability when more pairs are excluded.

It is worth noting that in surveillance settings, errors in attribute estimation can occur, which may affect the accuracy of face recognition systems that rely on shared attributes to select reference images and video frames. It is therefore crucial to investigate how errors in attribute estimation impact the performance of the proposed method, which pairs the highest number of attributes in common. It is also important to explore alternative approaches for selecting reference images and video frames that do not solely rely on shared attributes, such as deep metric learning, which can learn discriminative features for face

recognition directly from the data. Future work should also consider examining the proposed methods on more diverse datasets, including those that present greater variability in facial attributes, to ensure the generalizability of the findings. Our results show the potential for using FIQ, spatio-temporal information and additional information, such as gait, clothes, or hair, to improve the performance of automated face recognition in video surveillance. Further research could explore the use of additional metrics for keyframe selection, and examine the performance of the proposed methods on a wider range of datasets and face recognition algorithms.

Intriguingly, our experiments also showed that preprocessing video frames with the super resolution Codeformer algorithm [158] did not lead to the anticipated improvement in face recognition performance. In fact, it seemingly deteriorated the outcome. One plausible explanation could be that the super-resolution process introduced some form of artifact or noise into the images that adversely affected the face recognition algorithms. Super-resolution algorithms like Codeformer generate high-frequency details that are not present in the original low-resolution image. If these details do not accurately represent the true high-resolution image, this could lead to mismatches compared with the reference face images, thereby deteriorating recognition performance.

In our study, we've delved into the intricacies of facial image quality metrics, attribute-based matching, and the impact of preprocessing techniques on face recognition. The importance of selecting high-quality frames has been reaffirmed, offering a tangible path forward for optimizing recognition performance in real-world scenarios. Our exploration into attribute-based pairings has illuminated both its potential benefits and areas requiring further study.

In conclusion, it is undeniable that facial image quality plays a pivotal role in face recognition, especially within video surveillance scenarios. Our results showcased the efficacy of using FIQ as a metric to enhance face recognition accuracy. Techniques such as utilizing weighted quality average and pairing based on shared attributes have proven to improve performance, underpinning the importance of considering these details in the recognition process. While not all explored methods yielded the expected enhancements—like the super resolution Codeformer algorithm, this exploration has provided valuable insights into the inherent complexities of automated face recognition, allowing us to better understand both its capabilities and limitations. These findings set a strong foundation for continuing advancements in the field, paving the way for further exploration of facial attributes, FIQ, and the potential integration of alternative super-resolution techniques.

5.7 Datasets quality Appendix section

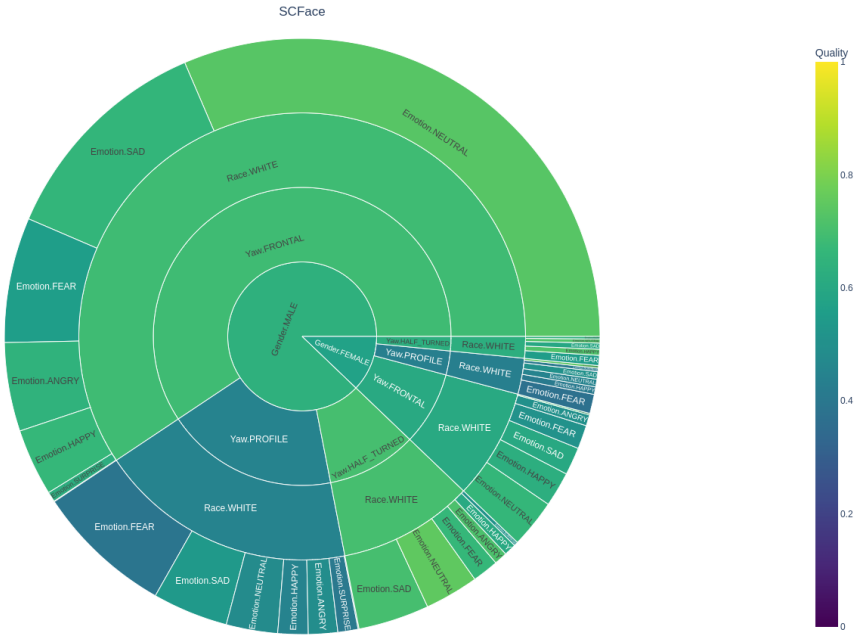


Figure 5.8: Facial attributes in the SCFace database according to quality model SER-FIQ [115].

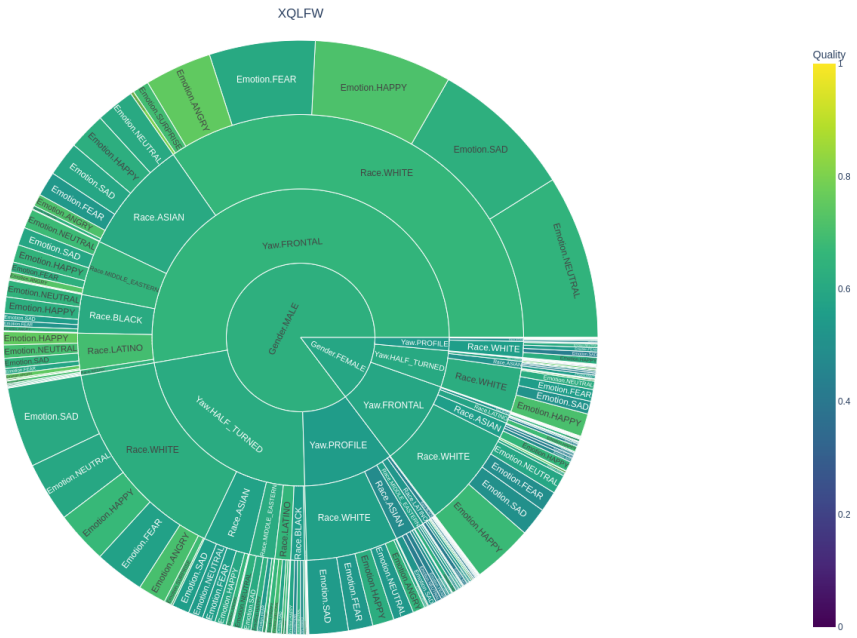


Figure 5.9: Facial attributes in the XQLFW database according to quality model SER-FIQ [115].

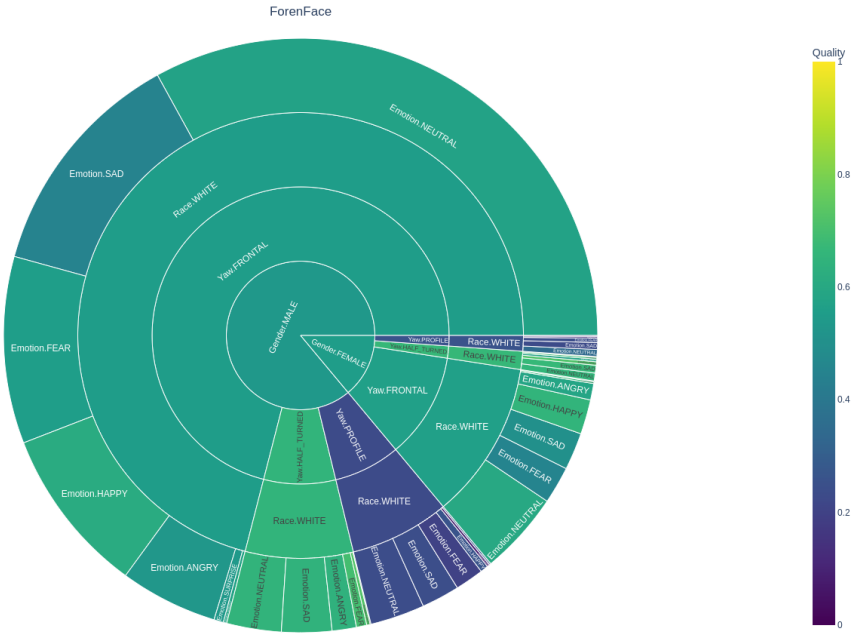


Figure 5.10: Facial attributes in the ForenFace database according to quality model SER-FIQ [115].

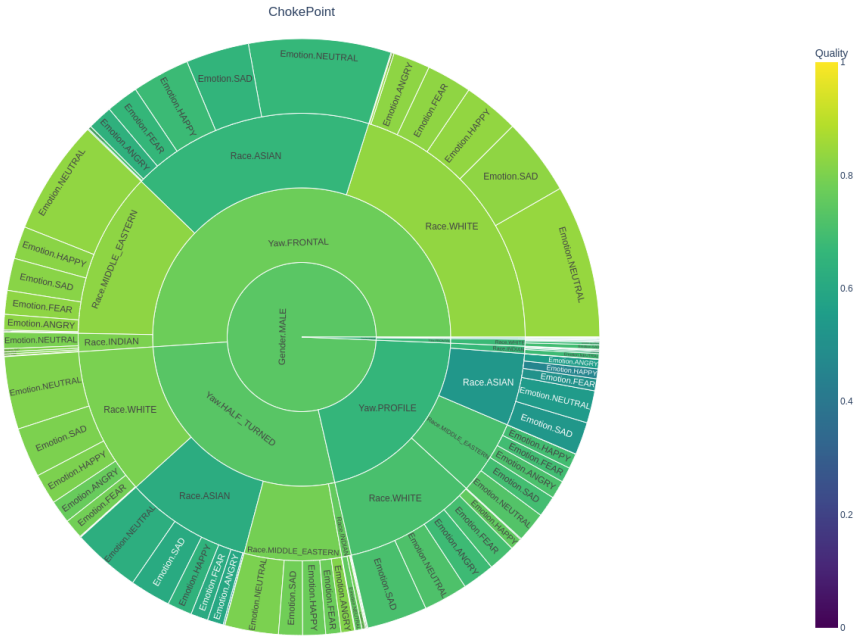


Figure 5.11: Facial attributes in the ChokePoint database according to quality model SER-FIQ [115].

5.8 Results Appendix Section

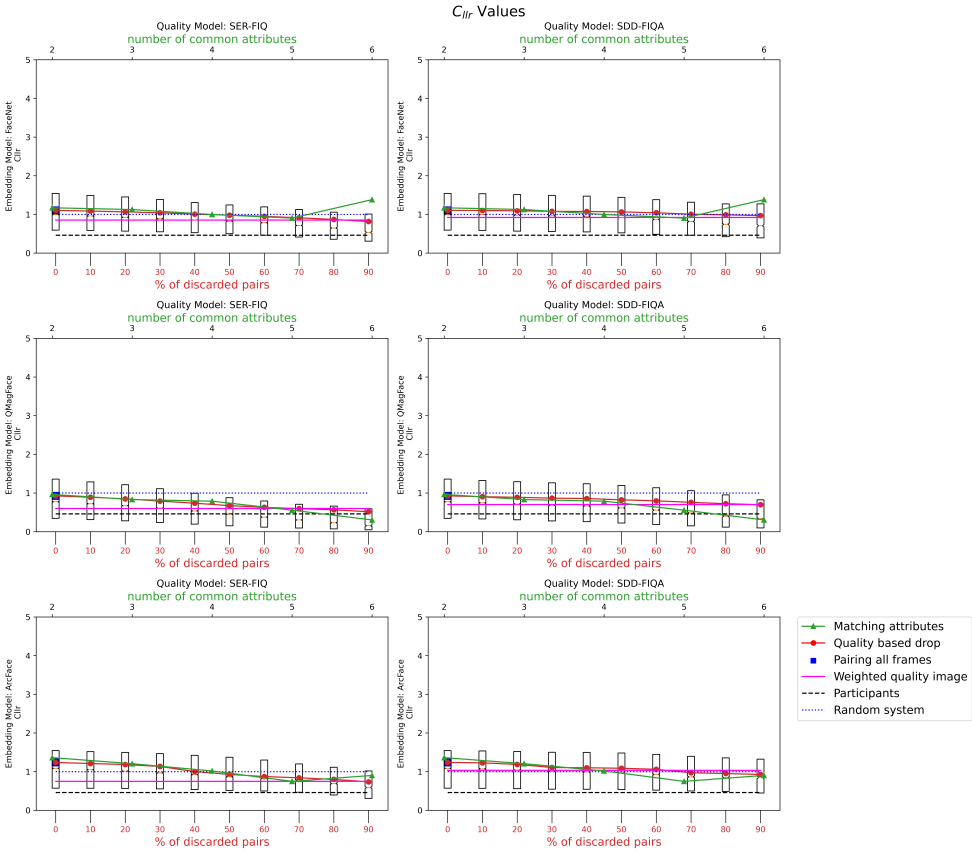


Figure 5.12: Graphical representation of the C_{llr} (log-likelihood ratio cost) values after calibration of random images

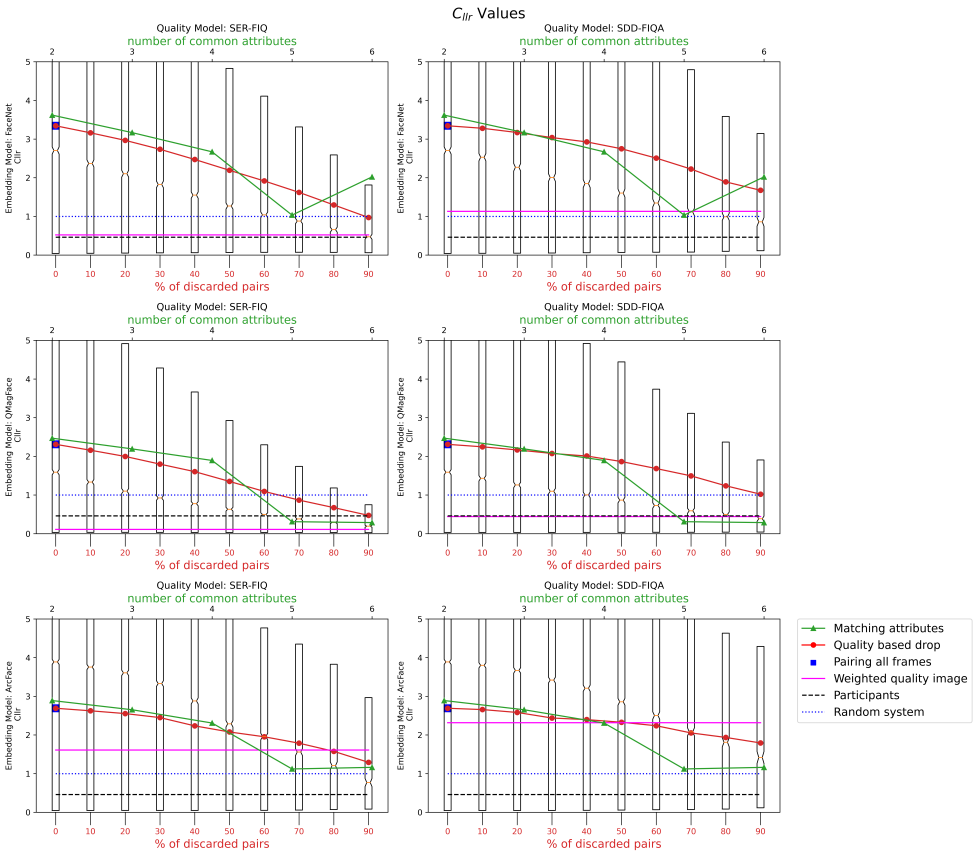


Figure 5.13: Graphical representation of the C_{llr} (log-likelihood ratio cost) values after calibration with quality filter.

5. IMPROVING LR FOR FR IN SV VIDEO BY MULTIMODAL FEATURE PAIRING

6

Summary and Conclusions

The transformative impact of facial recognition technologies is becoming increasingly apparent across various high-stakes fields, including forensics and security. This thesis offers a comprehensive exploration of key aspects that influence the effectiveness of these technologies. It evaluates the potential for automation to match or surpass expert human performance, highlights challenges related to the quality of visual data and criteria-based selection, and explores methods to optimize performance in dynamic settings like video surveillance. The research serves as a milestone, pointing to promising directions for future advancements while identifying limitations that warrant further inquiry.

6.1 Summary

In **Chapter 2**, we conducted a detailed comparative study to evaluate the performance of commercial and open-source face recognition systems against human experts. Our results show that commercial software generally outperforms both open-source alternatives and human experts, particularly for full-frontal images. However, the study also revealed that these automated systems have limitations, especially when handling poor-quality images or those with occlusions such as those caused by people wearing caps, mics, and scarfs. This study is a significant step forward in understanding the application of automated face recognition systems in the forensic field.

In **Chapter 3**, we investigated the impact of calibration techniques on the estimation

of likelihood ratios. Utilizing ENFSI Proficiency tests as benchmarks, we found that quality score-based and feature-based calibrations surpass naive calibration methods. While commercial software exhibits superior performance, the transparency of open-source systems underscores the importance of ongoing research to improve both effectiveness and accountability in forensic facial image comparisons.

Chapter 4, focused on the critical issue of explainability in facial image quality assessments. We developed a multi-task learning model that not only predicts suitability scores, but also identifies facial and environmental attributes affecting these scores. The model's dual capability offers both high accuracy and explainability, attributes that are crucial in forensic settings. This chapter emphasizes the importance of both suitability and explainability in face recognition systems. Our findings demonstrate a high correlation between facial image quality and attributes such as sharpness, frontal pose, and age. This dual capability of the model not only enhances the accuracy of the quality assessments but also provides actionable insights which are particularly useful in forensic settings. However, the model does have limitations; it requires well-labeled data for training and relies on proper face detection and alignment for accurate assessment.

In **Chapter 5**, we explored methods to enhance face recognition in video surveillance. The study found that the selective use of high-quality frames and their pairing with suitable reference images improves the likelihood ratio estimation. The chapter revealed that using a weighted quality average of all available frames produces even better results. Importantly, the chapter discussed the limitations and potential for improvement in different face recognition algorithms, including the counter-intuitive finding that super-resolution techniques might not always be beneficial, due to adding information and/or artifacts.

6.2 Conclusions

We structure our conclusions along a number of dimensions, reflecting the core research questions outlined in the introduction. Each of these dimensions highlights an aspect of forensic face recognition where deep learning technologies have a significant role to play.

6.2.1 Calibration Techniques

Responding directly to the research subquestion about the effectiveness of likelihood ratio estimation through calibration techniques, we found these methods to be highly impactful. Quality score-based and feature-based calibration techniques significantly elevate the accuracy and reliability of automated facial recognition systems. Furthermore, the type of

calibration method used has implications for the estimation of likelihood ratios. This advances our understanding of the question about the effect of calibration on likelihood ratio estimation. We discovered that these methods have a potential impact on the accuracy of these estimations, thereby making a case for their integration into standard forensic facial recognition processes.

6.2.2 Facial Image Quality (FIQ)

Our research directly addressed the question about the correlation between face image attributes and FIQ. We found that FIQ is an indispensable factor in the overall performance of face recognition systems. The multi-task learning model deployed in our study not only quantifies the quality of face images but also provides explainability. This layer of explainability is crucial for forensic applicability, enhancing the system's credibility and transparency. It resonates with our research question by contributing a new dimension to FIQ assessment. This is particularly important in high-stakes environments like forensics, where a detailed understanding of FIQ can mean the difference between accurate and misleading outcomes.

6.2.3 Video Surveillance

Turning to the subquestion about improving likelihood ratio estimation in video surveillance, our findings reveal two critical elements: quality frame selection and attribute-based pairing. These enhance the performance of facial recognition systems, especially in the challenging environments posed by surveillance video footage. However, our research also offers a cautionary note about the use of super-resolution algorithms. These do not consistently lead to anticipated performance gains, thus responding to our initial concerns in the research question about the effectiveness of different methodologies in video surveillance. The major takeaway is that attribute and quality based frame selection play pivotal roles in enhancing the performance and reliability of face recognition systems in video surveillance.

6.2.4 Limitations and Considerations

The limitations observed in current systems, especially their difficulty in handling images with occlusions and poor quality, offer additional directions for future research. These limitations should serve as cautionary notes for the deployment of such systems in high-stakes environments like forensics and surveillance.

6.2.5 Future Research Avenues

Our research has paved the way for several avenues for future research, shedding light on areas that need further investigation to maximize the capabilities of deep learning in forensic face recognition. These avenues include, but are not limited to, further studies into more effective calibration methods. Building on our discoveries, subsequent research could investigate the suitability of advanced statistical models for likelihood ratio estimation in forensic applications. This directly relates to the research subquestions about calibration methods and their impact on likelihood ratio estimation, affirming the need for ongoing, comprehensive studies.

Additionally, our work has revealed a critical need for a deeper understanding of attribute selection in video surveillance applications. By doing so, researchers may uncover methods that elevate the reliability and effectiveness of real-world surveillance systems, a finding that holds immense value in forensic settings. Lastly, there is an urgent need for expanded and more diverse datasets to test the generalizability of our findings and forensics face recognition in general. The broader the range of datasets, the more comprehensive and conclusive future studies could be, thereby providing answers that are not only scientifically compelling but also practically actionable. As we close this chapter on our contributions to the field, we open new doors for future scholars to walk through, carrying the torch forward in the quest for a more effective and accountable use of deep learning in forensic face recognition.

Bibliography

[1] R.S. Lacruz, C.B. Stringer, W.H. Kimbel, B. Wood, K. Harvati, P. O'Higgins, T.G. Bromage, and J.-L. Arsuaga. The evolutionary history of the human face. *Nature Ecology and Evolution*, 3(5):726–736, 2019. ISSN 2397-334X. doi: https://doi.org/10.1038/S41559-019-0865-7.

[2] Richard Farebrother and Julian Champkin. Alphonse bertillon and the measure of man: More expert than sherlock holmes. *Significance*, 11(2):36–39, April 2014. doi: https://doi.org/10.1111/j.1740-9713.2014.00739.x.

[3] Sarah Kember. Face recognition and the emergence of smart photography. *Journal of Visual Culture*, 13(2):182–199, 2014. ISSN 1470-4129. doi: https://doi.org/10.1177/1470412914541767.

[4] Lixiang Li, Xiaohui Mu, Siying Li, and Peng Haipeng. A review of face recognition technology. *IEEE Access*, 8:139110–139120, 2020. doi: 10.1109/ACCESS.2020.3011028.

[5] Ruth M. Morgan, Georgina E. Meakin, James C. French, and Sherry Nakhaeizadeh. Crime reconstruction and the role of trace materials from crime scene to court. *WIREs Forensic Science*, 2(1), nov 2019. doi: 10.1002/wfs2.1364.

[6] Maëlig Jacquet and Christophe Champod. Automated face recognition in forensic science: Review and perspectives. *Forensic Science International*, 307:110–124, feb 2020. doi: 10.1016/j.forsciint.2019.110124.

[7] Ipsita Pattnaik, Amita Dev, and A. K. Mohapatra. Forensic facial recognition: Review and challenges. In *Proceedings of International Conference on Data Science and Applications*, pages 351–367. Springer Nature Singapore, 2023. doi: 10.1007/978-981-19-6634-7_26.

[8] 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. *Forensic Science Review*, 30(1):21–32, jan 2018. ISSN 1042-7201.

[9] P. Jonathon Phillips, Amy N. Yates, Ying Hu, Carina A. Hahn, Eilidh Noyes, Kelsey Jackson, Jacqueline G. Cavazos, Géraldine Jeckeln, Rajeev Ranjan, Swami Sankaranarayanan, Jun-Cheng Chen, Carlos D. Castillo, Rama Chellappa, David White, and Alice J. O'Toole. Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. *Proceedings of the National Academy of Sciences*, 115(24):6171–6176, may 2018. doi: 10.1073/pnas.1721355115.

[10] Nicole A. Spaun. Face recognition in forensic science. In *Handbook of Face Recognition*, page 655–670. Springer London, 2011. doi: 10.1007/978-0-85729-932-1_26.

[11] Veena Mayya, Radhika M. Pai, and M.M. Manohara Pai. Automatic facial expression recognition using DCNN. *Procedia Computer Science*, 93:453–461, 2016. doi: 10.1016/j.procs.2016.07.233.

[12] Matthew Turk and Alex Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991. doi: https://doi.org/10.1162/jocn.1991.3.1.71.

[13] Peter N. Belhumeur, Joao P. Hespanha, and David J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):711–720, 1997. doi: 10.1109/34.598228.

[14] T. Ahonen, A. Hadid, and M. Pietikainen. Face description with local binary patterns: Application to face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(12):2037–2041, dec 2006. doi: 10.1109/tpami.2006.244.

[15] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, nov 2004. doi: 10.1023/b:visi.0000029664.99615.94.

[16] Rinku Datta Rakshit, Dakshina Ranjan Kisku, Phalguni Gupta, and Jamuna Kanta Sing. Cross-resolution face identification using deep-convolutional neural network. *Multimedia Tools and Applications*, 80(14):20733–20758, mar 2021. doi: 10.1007/s11042-021-10745-y.

[17] P. Jonathon Phillips. A cross benchmark assessment of a deep convolutional neural network for face recognition. In *2017 12th IEEE International Conference on Automatic Face &amp Gesture Recognition (FG 2017)*. IEEE, may 2017. doi: 10.1109/fg.2017.89.

[18] Philipp Terhörst, Jan Niklas Kolf, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Post-comparison mitigation of demographic bias in face recognition using fair score normalization. *Pattern Recognition Letters*, 140:332–338, dec 2020. doi: 10.1016/j.patrec.2020.11.007.

[19] Jacqueline G. Cavazos, P. Jonathon Phillips, Carlos D. Castillo, and Alice J. O'Toole. Accuracy comparison across face recognition algorithms: Where are we on measuring race bias? *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 3(1):101–111, jan 2021. doi: 10.1109/tbiom.2020.3027269.

[20] Hitoshi Imaoka, Hiroshi Hashimoto, Koichi Takahashi, Akinori F. Ebihara, Jianquan Liu, Akihiro Hayasaka, Yusuke Morishita, and Kazuyuki Sakurai. The future

of biometrics technology: from face recognition to related applications. *APSIPA Transactions on Signal and Information Processing*, 10(1), 2021. doi: 10.1017/atsip.2021.8.

[21] Saurabh Ravindranath, Rahul Baburaj, Vineeth N. Balasubramanian, NageswaraRao Namburu, Sujit Gujar, and C. V. Jawahar. Human-machine collaboration for face recognition. In *Proceedings of the 7th ACM IKDD CoDS and 25th COMAD*. ACM, jan 2020. doi: 10.1145/3371158.3371160.

[22] Steve Branson, Catherine Wah, Florian Schroff, Boris Babenko, Peter Welinder, Pietro Perona, and Serge Belongie. Visual recognition with humans in the loop. In Kostas Daniilidis, Petros Maragos, and Nikos Paragios, editors, *ECCV*, volume 6314 of *Lecture Notes in Computer Science*, pages 438–451. Springer, Berlin / Heidelberg, 2010. ISBN 978-3-642-15560-4. doi: 10.1007/978-3-642-15561-1_32.

[23] Daniel Ramos, Rudolf Haraksim, and Didier Meuwly. Likelihood ratio data to report the validation of a forensic fingerprint evaluation method. *Data in Brief*, 10: 75–92, feb 2017. doi: 10.1016/j.dib.2016.11.008.

[24] Anna Leida Molder, Isabelle Enlund Astrom, and Elisabet Leitet. Development of a score-to-likelihood ratio model for facial recognition using authentic criminalistic data. In *2020 8th International Workshop on Biometrics and Forensics (IWBF)*, pages 1–6. IEEE, apr 2020. doi: 10.1109/iwbf49977.2020.9107954.

[25] Rajesh Verma, Navdha Bhardwaj, Arnav Bhavsar, and Kewal Krishan. Towards facial recognition using likelihood ratio approach to facial landmark indices from images. *Forensic Science International: Reports*, 5:100254, jul 2022. doi: 10.1016/j.fsir.2021.100254.

[26] Interpol. Facial recognition. https://www.interpol.int/en/How-we-work/Forensics/Facial-Recognition, 2021.

[27] Jeremiah R. Barr, Kevin W. Bowyer, Patrick J. Flynn, and Soma Biswas. Face recognition from video: a review. *IJPRAI*, 26(5):1266002, 2012. doi: 10.1142/S0218001412660024.

[28] Martin Knoche, Stefan Hörmann, and Gerhard Rigoll. Cross-quality LFW: A database for analyzing cross- resolution image face recognition in unconstrained environments. In *16th IEEE FG 2021*, pages 1–5. IEEE, 2021. doi: 10.1109/FG52635.2021.9666960.

[29] Mislav Grgic, Kresimir Delac, and Sonja Grgic. SCface – surveillance cameras face database. *Multimedia Tools and Applications*, 51(3):863–879, oct 2009. doi: 10.1007/s11042-009-0417-2.

[30] Joseph C. Celentino. Face-to-face with facial recognition evidence: Admissibility under the post-crawford confrontation clause. *Michigan Law Review*, 114:1317–1353, 2016. URL https://repository.law.umich.edu/mlr/vol114/iss7/3.

[31] Adamo Quaglia and Calogera M Epifano. *Face Recognition: Methods, Applications and Technology*. Nova Science Publishers, Incorporated, New York, 2012. ISBN 9781619426634.

[32] Sheila Willis, Louise McKenna, Sean McDermott, Geraldine O'Donnell, et al. Enfsi guideline for evaluative reporting in forensic science. https://enfsi.eu/docfile/enfsi-guideline-for-evaluative-reporting-in-forensic-science/, 2015. Accessed: 2019-12-19.

[33] Steven P. Lund and Hari Iyer. Likelihood ratio as weight of forensic evidence: A closer look. *Journal of Research of the National Institute of Standards and Technology*, 122:1, oct 2017. doi: 10.6028/jres.122.027.

[34] A Collins and N E Morton. Likelihood ratios for DNA identification. *Proceedings of the National Academy of Sciences*, 91(13):6007–6011, jun 1994. doi: 10.1073/pnas.91.13.6007.

[35] Cédric Neumann, Christophe Champod, Roberto Puch-Solis, Nicole Egli, Alexandre Anthonioz, and Andie Bromage-Griffiths. Computation of likelihood ratios in fingerprint identification for configurations of any number of minutiae. *Journal of Forensic Sciences*, 52(1):54–64, dec 2006. doi: 10.1111/j.1556-4029.2006.00327.x.

[36] Tauseef Ali. *Biometric Score Calibration for Forensic Face Recognition*. PhD thesis, University of Twente, June 2014.

[37] Massimo Tistarelli and Christophe Champod, editors. *Handbook of Biometrics for Forensic Science*. Springer International Publishing, 2017. doi: 10.1007/978-3-319-50673-9.

[38] M.A. Turk and A.P. Pentland. Face recognition using eigenfaces. In *Proceedings of the IEEE Conf. Computer Vision and Pattern Recognition*, pages 586–587. IEEE Comput. Sco. Press, 1991. doi: 10.1109/cvpr.1991.139758.

[39] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In Xianghua Xie, Mark W. Jones, , and Gary K. L. Tam, editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 41.1–41.12. BMVA Press, September 2015. ISBN 1-901725-53-7. doi: 10.5244/C.29.41.

[40] 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.

[41] Iacopo Masi, Yue Wu, Tal Hassner, and Prem Natarajan. Deep face recognition: A survey. In *2018 31st SIBGRAPI conference on graphics, patterns and images (SIBGRAPI)*, pages 471–478. IEEE, 2018. doi: 10.1109/SIBGRAPI.2018.00067.

[42] Alessandro Piva, Ilenia Tinnirello, and Simone Morosi. *Digital Communication. Towards a Smart and Secure Future Internet: 28th International Tyrrhenian Workshop, TIWDC 2017, Palermo, Italy, September 18-20, 2017, Proceedings*, volume 766. Springer, 2017. doi: 10.1007/978-3-319-67639-5.

[43] Rudolf Haraksim, Daniel Ramos, and Didier Meuwly. Validation of likelihood ratio methods for forensic evidence evaluation handling multimodal score distributions. *IET Biometrics*, 6(2):61–69, nov 2016. doi: 10.1049/iet-bmt.2015.0059.

[44] Brandon Amos, Bartosz Ludwiczuk, Mahadev Satyanarayanan, et al. Openface: A general-purpose face recognition library with mobile applications. *CMU School of Computer Science*, 6(2):20, 2016.

[45] Xin Liu, Meina Kan, Wanglong Wu, Shiguang Shan, and Xilin Chen. VIPLFaceNet: an open source deep face recognition SDK. *Frontiers of Computer Science*, 11(2):208–218, jan 2017. doi: 10.1007/s11704-016-6076-3.

[46] Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conf. Computer Vision and Pattern Recognition*, pages 815–823. IEEE, jun 2015. doi: 10.1109/cvpr.2015.7298682.

[47] Hoang Pham, editor. *Springer Handbook of Engineering Statistics*, volume 49. Springer London, 2006. doi: 10.1007/978-1-84628-288-1.

[48] Yen-Chi Chen. A tutorial on kernel density estimation and recent advances. *Biostatistics &amp Epidemiology*, 1(1):161–187, jan 2017. doi: 10.1080/24709360.2017.1396742.

[49] Jan de Leeuw, Kurt Hornik, and Patrick Mair. Isotone optimization inir/i: Pool-adjacent-violators algorithm (PAVA) and active set methods. *Journal of Statistical Software*, 32(5):1–24, 2009. doi: 10.18637/jss.v032.i05.

[50] B.W. Matthews. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure*, 405(2): 442–451, oct 1975. doi: 10.1016/0005-2795(75)90109-9.

[51] Colin G.G. Aitken and Franco Taroni. *Statistics and the Evaluation of Evidence for Forensic Scientists*. John Wiley & Sons, Ltd., 2004. doi: 10.1002/0470011238.

[52] Alex Biedermann, Christophe Champod, and Sheila Willis. Development of european standards for evaluative reporting in forensic science. *The International Journal of Evidence &amp Proof*, 21(1-2):14–29, dec 2016. doi: 10.1177/1365712716674796.

[53] Mark W. Perlin. Explaining the likelihood ratio in dna mixture interpretation. In *Proceedings of Promega's Twenty First International Symposium on Human Identification*, San Antonio, TX, 2010. URL https://www.cybgen.com/information/publication/2010/ISHI/Perlin-Explaining-the-likelihood-ratio-in-DNA-mixture-interpretation/page.shtml.

[54] Rudolf Haraksim and Andrzej Drygajlo. Measuring performance in forensic automatic speaker recognition: Vq, gmm-ubm, i-vectors. *Biosig 2016*, 2016.

[55] Anna Jeannette Leegwater, Didier Meuwly, Marjan Sjerps, Peter Vergeer, and Ivo Alberink. Performance study of a score-based likelihood ratio system for forensic fingermark comparison. *Journal of Forensic Sciences*, 62(3):626–640, feb 2017. doi: 10.1111/1556-4029.13339.

[56] Annabel Bolck, Haifang Ni, and Martin Lopatka. Evaluating score- and feature-based likelihood ratio models for multivariate continuous data: applied to forensic MDMA comparison. *Law, Probability and Risk*, 14(3):243–266, sep 2015. doi: 10.1093/lpr/mgv009.

[57] Andrew van Es, Wim Wiarda, Maarten Hordijk, Ivo Alberink, and Peter Vergeer. Implementation and assessment of a likelihood ratio approach for the evaluation of LA-ICP-MS evidence in forensic glass analysis. *Science &amp Justice*, 57(3):181–192, may 2017. doi: 10.1016/j.scijus.2017.03.002.

[58] P. Vergeer, A. Bolck, L.J.C. Peschier, C.E.H. Berger, and J.N. Hendrikse. Likelihood ratio methods for forensic comparison of evaporated gasoline residues. *Science &amp Justice*, 54(6):401–411, dec 2014. doi: 10.1016/j.scijus.2014.04.008.

[59] Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. VGGFace2: A dataset for recognising faces across pose and age. In *2018 13th IEEE International Conference on Automatic Face &amp Gesture Recognition (FG 2018)*. IEEE, may 2018. doi: 10.1109/fg.2018.00020.

[60] Bogdan Kwolek. Face detection using convolutional neural networks and gabor filters. In *Artificial Neural Networks: Biological Inspirations – ICANN 2005*, pages 551–556. Springer Berlin Heidelberg, 2005. doi: 10.1007/11550822_86.

[61] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. DeepFace: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE Conf. Computer Vision and Pattern Recognition*, pages 1701–1708. IEEE, jun 2014. doi: 10.1109/cvpr.2014.220.

[62] Shiguang Shan, Wen Gao, Bo Cao, and Debin Zhao. Illumination normalization for robust face recognition against varying lighting conditions. In *2003 IEEE International SOI Conference. Proceedings (Cat. No.03CH37443)*, pages 157–164. IEEE, 2003. doi: 10.1109/amfg.2003.1240838.

[63] Rajendra Kumar, Papendra Kumar, and Abhishek Gupta. Review paper on face recognition techniques. *International Journal of Computer Sciences and Engineering*, 7(5):223–229, may 2019. doi: 10.26438/ijcse/v7i5.223229.

[64] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, oct 2016. doi: 10.1109/lsp.2016.2603342.

[65] Shuzhe Wu, Meina Kan, Zhenliang He, Shiguang Shan, and Xilin Chen. Funnel-structured cascade for multi-view face detection with alignment-awareness. *Neurocomputing*, 221:138–145, jan 2017. doi: 10.1016/j.neucom.2016.09.072.

[66] Jie Zhang, Shiguang Shan, Meina Kan, and Xilin Chen. Coarse-to-fine auto-encoder networks (CFAN) for real-time face alignment. In *Computer Vision – ECCV 2014*, pages 1–16. Springer International Publishing, 2014. doi: 10.1007/978-3-319-10605-2_1.

[67] Geert HLM Heideman, FW Hoeksema, and HEP Tattje, editors. *Proceedings of the 13th Symposium on Information Theory in the Benelux*, Enschede, 1992. Werkgemeenschap voor Informatie- en Communicatietheorie (WIC). 204 pages.

[68] Thomas S. Shively, Thomas W. Sager, and Stephen G. Walker. A bayesian approach to non-parametric monotone function estimation. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 71(1):159–175, oct 2008. doi: 10.1111/j.1467-9868.2008.00677.x.

[69] Damien Dessimoz and Christophe Champod. A dedicated framework for weak biometrics in forensic science for investigation and intelligence purposes: The case of facial information. *Security Journal*, 29(4):603–617, sep 2016. doi: 10.1057/sj.2015.32.

[70] Didier Meuwly, Daniel Ramos, and Rudolf Haraksim. A guideline for the validation of likelihood ratio methods used for forensic evidence evaluation. *Forensic Science International*, 276:142–153, jul 2017. doi: 10.1016/j.forsciint.2016.03.048.

[71] Daniel Ramos and Joaquin Gonzalez-Rodriguez. Reliable support: Measuring calibration of likelihood ratios. *Forensic Science International*, 230(1-3):156–169, jul 2013. doi: 10.1016/j.forsciint.2013.04.014.

[72] Colin Aitken, Charles EH Berger, John S Buckleton, Christophe Champod, James Curran, AP Dawid, Ian W Evett, Peter Gill, Joaquin Gonzalez-Rodriguez, Graham Jackson, et al. Expressing evaluative opinions: A position statement. *Science & Justice*, 51(1):1–2, 2011. ISSN 1355-0306. doi: https://doi.org/10.1016/j.scijus.2011.01.002.

[73] Daniel Ramos, Ram P. Krish, Julian Fierrez, and Didier Meuwly. From biometric scores to forensic likelihood ratios. In *Handbook of Biometrics for Forensic Science*, pages 305–327. Springer International Publishing, 2017. doi: 10.1007/978-3-319-50673-9_14.

[74] David A. van Leeuwen and Niko Brümmer. The distribution of calibrated likelihood-ratios in speaker recognition. In *Interspeech 2013*, pages 24–29. ISCA, aug 2013. doi: 10.21437/interspeech.2013-406.

[75] Niko Brummer. *Measuring, refining and calibrating speaker and language information extracted from speech*. PhD thesis, Stellenbosch: University of Stellenbosch, 2010.

[76] Mahesh Kumar Nandwana, Luciana Ferrer, Mitchell McLaren, Diego Castan, and Aaron Lawson. Analysis of Critical Metadata Factors for the Calibration of Speaker Recognition Systems. In *Proc. Interspeech 2019*, pages 4325–4329, 2019. doi: 10.21437/Interspeech.2019-1808.

[77] Geoffrey Stewart Morrison. The impact in forensic voice comparison of lack of calibration and of mismatched conditions between the known-speaker recording and the relevant-population sample recordings. *Forensic Science International*, 283:e1–e7, 12 2017. doi: 10.1016/j.forsciint.2017.12.024.

[78] Niko Brümmer and Johan du Preez. Application-independent evaluation of speaker detection. *Computer Speech & Language*, 20(2-3):230–275, apr 2006. doi: 10.1016/j.csl.2005.08.001.

[79] Niko Brümmer and Albert Swart. Bayesian calibration for forensic evidence reporting. In *Interspeech 2014*. ISCA, sep 2014. doi: 10.21437/interspeech.2014-90.

[80] Colin Aitken and Franco Taroni. The evaluation of evidence. In *Statistics and the Evaluation of Evidence for Forensic Scientists*, pages 69–118. John Wiley & Sons, Ltd, jun 2005. doi: 10.1002/0470011238.ch3.

[81] 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, jul 2013. doi: 10.1111/1556-4029.12233.

[82] Michael J. Saks and Jonathan J. Koehler. The coming paradigm shift in forensic identification science. *Science*, 309(5736):892–895, aug 2005. doi: 10.1126/science.1111565.

[83] ENFSI. *Best Practice Manual for Facial Image Comparison*. European Network of Forensic Science Institutes (ENFSI), 2018.

[84] Geoffrey Stewart Morrison and Ewald Enzinger. Score based procedures for the calculation of forensic likelihood ratios – scores should take account of both similarity and typicality. *Science & Justice*, 58(1):47–58, jan 2018. doi: 10.1016/j.scijus.2017.06.005.

[85] Richard A Nichols. Interpreting DNA evidence: Statistical genetics for forensic scientists. *Heredity*, 82(5):585–586, may 1999. doi: 10.1038/sj.hdy.6885562.

[86] Grzegorz Zadora and Daniel Ramos. Evaluation of glass samples for forensic purposes — an application of likelihood ratios and an information–theoretical approach. *Chemometrics and Intelligent Laboratory Systems*, 102(2):63–83, jul 2010. doi: 10.1016/j.chemolab.2010.03.007.

[87] Daniel Ramos, Javier Franco-Pedroso, Alicia Lozano-Diez, and Joaquin Gonzalez-Rodriguez. Deconstructing cross-entropy for probabilistic binary classifiers. *Entropy*, 20(3):208, mar 2018. doi: 10.3390/e20030208.

[88] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conf. Computer Vision and Pattern Recognition*, pages 4685–4694, 2019. doi: 10.1109/CVPR.2019.00482.

[89] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. SphereFace: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE Conf. Computer Vision and Pattern Recognition*, pages 6738–6746. IEEE, jul 2017. doi: 10.1109/cvpr.2017.713.

[90] Geoffrey Stewart Morrison, Felipe Ochoa, and Tharmarajah Thiruvaran. Database selection for forensic voice comparison. In *Odyssey 2012: The Speaker and Language Recognition Workshop*, pages 62–77, Singapore, June 25-28 2012. International Speech Communication Association (ISCA). URL https://www.isca-speech.org/archive/Odyssey_2012/abstracts/Od12_P3-03_Morrison.pdf.

[91] Jesús Villalba and Niko Brümmer. Towards fully bayesian speaker recognition: integrating out the between-speaker covariance. In *Interspeech 2011*. ISCA, aug 2011. doi: 10.21437/interspeech.2011-142.

[92] Grzegorz Zadora, Agnieszka Martyna, Daniel Ramos, and Colin Aitken. *Statistical analysis in forensic science: evidential value of multivariate physicochemical data*. John Wiley & Sons, 2013. doi: 10.1002/9781118763155.

[93] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018.

[94] Fredrik Eklöf and Peter Bergström. The enfsi-diwg 2011 fic test, 2011.

[95] Fredrik Eklöf and Peter Bergström. The enfsi-diwg 2012 fic test, 2012.

[96] Fredrik Eklöf and Peter Bergström. Enfsi-diwg 2013 facial image comparisons proficiency test, 2013.

[97] Dr Arnout Ruifrok. Facial image comparison proficiency test 2017, 2017.

[98] Sergio Castro Martínez. Facial image comparison proficiency test 2018, 2018.

[99] Dr Dana Michalski and Gemma Snyder. Facial image comparison test 2019, 2019.

[100] Dr. Shelina Jilani. Facial image comparison proficiency test 2020, 2020.

[101] Jan de Leeuw, Kurt Hornik, and Patrick Mair. Isotone optimization inR: Pool-adjacent-violators algorithm (PAVA) and active set methods. *Journal of Statistical Software*, 32(5), 2009. doi: 10.18637/jss.v032.i05.

[102] JooSeuk Kim and Clayton Scott. Robust kernel density estimation. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, mar 2008. doi: 10.1109/icassp.2008.4518376.

[103] David G Kleinbaum and Mitchel Klein. *Logistic Regression: A Self-Learning Text*. Springer, 2002. doi: 10.1007/b97379.

[104] Gary B. Huang Erik Learned-Miller. Labeled faces in the wild: Updates and new reporting procedures. Technical Report UM-CS-2014-003, University of Massachusetts, Amherst, May 2014.

[105] 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*, 6(6):487–494, jul 2017. doi: 10.1049/iet-bmt.2016.0160.

[106] Andrea Macarulla Rodriguez, Zeno Geradts, and Marcel Worring. Likelihood ratios for deep neural networks in face comparison. *Journal of Forensic Sciences*, 65(4): 1169–1183, may 2020. doi: 10.1111/1556-4029.14324.

[107] Sefik Ilkin Serengil and Alper Ozpinar. Lightface: A hybrid deep face recognition framework. In *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*, pages 23–27. IEEE, 2020. doi: 10.1109/ASYU50717.2020.9259802.

[108] Cognitec. Facevacs, 2021. URL https://www.cognitec.com/facevacs-technology.html.

[109] Yuxi Peng. *Face recognition at a distance: low-resolution and alignment problems*. PhD thesis, UT, Netherlands, February 2019.

[110] NFI. Lir python likelihood ratio library. URL https://pypi.org/project/lir/.

[111] Itiel E. Dror and Nicholas Scurich. (mis)use of scientific measurements in forensic science. *Forensic Science International: Synergy*, sep 2020. doi: 10.1016/j.fsisyn.2020.08.006.

[112] Mei Wang and Weihong Deng. Deep face recognition: A survey. *Neurocomputing*, 429:215–244, 2021. doi: 10.1016/j.neucom.2020.10.081.

[113] Torsten Schlett, Christian Rathgeb, Olaf Henniger, Javier Galbally, Julian Fierrez, and Christoph Busch. Face image quality assessment: A literature survey. *ACM CSUR*, 2021. doi: 10.1145/3507901.

[114] A.C.C. Ruifrok, P. Vergeer, and Andrea Macarulla Rodrigues. From facial images of different quality to score based LR. *Forensic Science International*, 332:111201, mar 2022. doi: 10.1016/j.forsciint.2022.111201.

[115] Philipp Terhörst, Jan Niklas Kolf, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. SER-FIQ: unsupervised estimation of face image quality based on stochastic embedding robustness. In *2020 IEEE/CVF*, pages 5650–5659. IEEE, 2020. doi: 10.1109/CVPR42600.2020.00569.

[116] Qiang Meng, Shichao Zhao, Zhida Huang, and Feng Zhou. MagFace: A universal representation for face recognition and quality assessment. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14225–14234. IEEE, jun 2021. doi: 10.1109/cvpr46437.2021.01400.

[117] Javier Hernandez-Ortega, Julian Fierrez, Luis F. Gomez, Aythami Morales, Jose Luis Gonzalez de Suso, and Francisco Zamora-Martinez. FaceQvec: Vector quality assessment for face biometrics based on ISO compliance. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*. IEEE, jan 2022. doi: 10.1109/wacvw54805.2022.00014.

[118] Mei Ngan, Patrick Grother, and Kayee Hanaoka. Ongoing face recognition vendor test (FRVT) part 6b: Face recognition accuracy with face masks using post-COVID-19 algorithms. Technical report, nov 2020.

[119] Philipp Terhörst, Marco Huber, Naser Damer, Florian Kirchbuchner, Kiran Raja, and Arjan Kuijper. Pixel-level face image quality assessment for explainable face recognition. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 5(2): 288–297, apr 2023. doi: 10.1109/tbiom.2023.3263186.

[120] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12):4695–4708, 2012. doi: 10.1109/TIP.2012.2214050.

[121] A. Mittal, R. Soundararajan, and A. C. Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, mar 2013. doi: 10.1109/lsp.2012.2227726.

[122] Venkatanath N, Praneeth D, Maruthi Chandrasekhar Bh, Sumohana S. Channappayya, and Swarup S. Medasani. Blind image quality evaluation using perception based features. In *2015 Twenty First National Conference on Communications (NCC)*, pages 1–6. IEEE, feb 2015. doi: 10.1109/ncc.2015.7084843.

[123] Fu-Zhao Ou, Xingyu Chen, Ruixin Zhang, Yuge Huang, Shaoxin Li, Jilin Li, Yong Li, Liujuan Cao, and Yuan-Gen Wang. SDD-FIQA: Unsupervised face image quality assessment with similarity distribution distance. In *2021 IEEE/CVF CVPR)*, pages 7670–7679. IEEE, jun 2021. doi: 10.1109/cvpr46437.2021.00758.

[124] Javier Hernandez-Ortega, Javier Galbally, Julian Fierrez, Rudolf Haraksim, and Laurent Beslay. FaceQnet: Quality assessment for face recognition based on deep learning. In *2019 International Conference on Biometrics (ICB)*, pages 1–8. IEEE, jun 2019. doi: 10.1109/icb45273.2019.8987255.

[125] Philipp Terhörst, Jan Niklas Kolf, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Face quality estimation and its correlation to demographic and non-demographic bias in face recognition. In *2020 IEEE IJCB*, pages 1–11. IEEE, 2020. doi: 10.1109/IJCB48548.2020.9304865.

[126] Philipp Terhorst, Malte Ihlefeld, Marco Huber, Naser Damer, Florian Kirchbuchner, Kiran Raja, and Arjan Kuijper. QMagFace: Simple and accurate quality-aware face recognition. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, jan 2023. doi: 10.1109/wacv56688.2023.00348.

[127] Yu Zhang and Qiang Yang. A survey on multi-task learning. *IEEE Trans. on Knowledge and Data Engineering*, 2021. doi: 10.1109/TKDE.2021.3070203.

[128] Roberto Cipolla, Yarin Gal, and Alex Kendall. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *2018 IEEE/CVF*. IEEE, jun 2018. doi: 10.1109/cvpr.2018.00781.

[129] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Facial landmark detection by deep multi-task learning. In *European Conference on Computer Vision*, pages 94–108. Springer, 2014. doi: 10.1007/978-3-319-10599-4_7.

[130] Huitao Luo. A training-based no-reference image quality assessment algorithm. In *2004 International Conference on Image Processing, 2004. ICIP '04.*, volume 5, pages 2973–2976. IEEE, 2004. doi: 10.1109/icip.2004.1421737.

[131] Zhiguang Yang, Haizhou Ai, Bo Wu, Shihong Lao, and Lianhong Cai. Face pose estimation and its application in video shot selection. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, volume 1, pages 322–325. IEEE, 2004. doi: 10.1109/icpr.2004.1334117.

[132] M. Subasic, S. Loncaric, T. Petkovic, H. Bogunovic, and V. Krivec. Face image validation system. In *ISPA 2005. Proceedings of the 4th International Symposium on Image and Signal Processing and Analysis, 2005.*, pages 30–33. IEEE, 2005. doi: 10.1109/ispa.2005.195379.

[133] Davide Maltoni, Annalisa Franco, Matteo Ferrara, Dario Maio, and Antonio Nardelli. BioLab-ICAO: A new benchmark to evaluate applications assessing face image compliance to ISO/IEC 19794-5 standard. In *2009 16th IEEE International Conference on Image Processing (ICIP)*. IEEE, nov 2009. doi: 10.1109/icip.2009.5414000.

[134] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conf. Computer Vision and Pattern Recognition*. IEEE, jun 2016. doi: 10.1109/cvpr.2016.90.

[135] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Learning deep representation for face alignment with auxiliary attributes. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(5):918–930, 2016. doi: 10.1109/TPAMI.2015.2469286.

[136] Yuanhan Zhang, Zhenfei Yin, Yidong Li, Guojun Yin, Junjie Yan, Jing Shao, and Ziwei Liu. Celeba-spoof: Large-scale face anti-spoofing dataset with rich annotations. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XII*, volume 12357 of *Lecture Notes in Computer Science*, pages 70–85. Springer, 2020. doi: 10.1007/978-3-030-58610-2_5.

[137] Da GUO, Qingfang ZHENG, Xiaojiang PENG, and Ming LIU. Face detection, alignment, quality assessment and attribute analysis with multi-task hybrid

convolutional neural networks. *ZTE Communications*, 17(3):15–22, 2019. doi: 10.12142/ZTECOM.201903004.

[138] Mingxing Tan and Quoc Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/tan19a.html.

[139] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s. In *Proceedings of the IEEE Conf. Computer Vision and Pattern Recognition*. IEEE, jun 2022. doi: 10.1109/cvpr52688.2022.01167.

[140] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Representations (ICLR)*. OpenReview.net, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

[141] Xudong Wang, Li Lyna Zhang, Yang Wang, and Mao Yang. Towards efficient vision transformer inference. In *Proceedings of the 23rd Annual International Workshop on Mobile Computing Systems and Applications*. ACM, mar 2022. doi: 10.1145/3508396.3512869.

[142] Rich Caruna. Multitask learning: A knowledge-based source of inductive bias. In *Machine learning: Proceedings of the 10th international conference*, pages 41–48. Morgan Kaufmann, 1993. doi: 10.1016/b978-1-55860-307-3.50012-5.

[143] Jonathan Baxter. A bayesian/information theoretic model of learning to learn via multiple task sampling. *Machine learning*, 28(1):7–39, 1997. doi: 10.1023/A:1007327622663.

[144] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. doi: 10.48550/arXiv.1606.08415.

[145] Leslie N. Smith. A disciplined approach to neural network hyper-parameters: Part 1- learning rate, batch size, momentum, and weight decay. *CoRR*, abs/1803.09820, 2018. URL http://arxiv.org/abs/1803.09820.

[146] Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *Proceedings of the IEEE Conf. Computer Vision and Pattern Recognition*, pages 4352–4360. IEEE Computer Society, 2017. doi: 10.1109/CVPR.2017.463.

[147] Sefik Ilkin Serengil and Alper Ozpinar. Hyperextended lightface: A facial attribute analysis framework. In *2021 International Conference on Engineering and Emerging Technologies (ICEET)*, pages 1–4. IEEE, 2021. doi: 10.1109/ICEET53442.2021.9659697.

[148] Cognitec GmbH. *FaceVACS-DBScan ID Integrator Kit Reference Manual*. Cognitec GmbH.

[149] Rushuai Liu and Weijun Tan. Eqface: A simple explicit quality network for face recognition. In *IEEE CVPR Workshops 2021*, pages 1482–1490. Computer Vision Foundation / IEEE, 2021. doi: 10.1109/CVPRW53098.2021.00164.

[150] Samadhi P. K. Wickrama Arachchilage and Ebroul Izquierdo. Deep-learned faces: a survey. *EURASIP Journal on Image and Video Processing*, 2020(1), jun 2020. doi: 10.1186/s13640-020-00510-w.

[151] Rahma Abed, Sahbi Bahroun, and Ezzeddine Zagrouba. KeyFrame extraction based on face quality measurement and convolutional neural network for efficient face recognition in videos. *Multimedia Tools and Applications*, 80(15):23157–23179, aug 2020. doi: 10.1007/s11042-020-09385-5.

[152] Changxing Ding and Dacheng Tao. Trunk-branch ensemble convolutional neural networks for video-based face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):1002–1014, apr 2018. doi: 10.1109/tpami.2017.2700390.

[153] Sahbi Bahroun, Rahma Abed, and Ezzeddine Zagrouba. KS-FQA: Keyframe selection based on face quality assessment for efficient face recognition in video. *IET Image Processing*, 15(1):77–90, dec 2020. doi: 10.1049/ipr2.12008.

[154] Jingxiao Zheng, Rajeev Ranjan, Ching-Hui Chen, Jun-Cheng Chen, Carlos D Castillo, and Rama Chellappa. An automatic system for unconstrained video-based face recognition. *IEEE TBBIS*, 2(3):194–209, 2020.

[155] Brianna Maze, Jocelyn Adams, James A. Duncan, Nathan Kalka, Tim Miller, Charles Otto, Anil K. Jain, W. Tyler Niggel, Janet Anderson, Jordan Cheney, and Patrick Grother. IARPA janus benchmark - c: Face dataset and protocol. In *2018 International Conference on Biometrics (ICB)*. IEEE, feb 2018. doi: 10.1109/icb2018.2018.00033.

[156] Andrea Macarulla Rodríguez, Zeno Geradts, Marcel Worring, and Luis Unzueta. Improved likelihood ratios for surveillance video face recognition with multimodal feature pairing. In *2023 11th International Workshop on Biometrics and Forensics (IWBF)*. IEEE, apr 2023. doi: 10.1109/iwbf57495.2023.10157791.

[157] Jiahao Huo and Terence L van Zyl. Unique faces recognition in videos. In *2020 IEEE 23rd International Conference on Information Fusion (FUSION)*. IEEE, jul 2020. doi: 10.23919/fusion45008.2020.9190469.

[158] Shangchen Zhou, Kelvin C.K. Chan, Chongyi Li, and Chen Change Loy. Towards robust blind face restoration with codebook lookup transformer. In *NeurIPS*, 2022.

[159] 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 *CVPR 2011 WORKSHOPS*, pages 81–88. IEEE, jun 2011. doi: 10.1109/cvprw.2011.5981881.

[160] Reuben Moreton, Fredrik Eklöf, and Arnout Ruifrok. Facial image comparison test 2015, 2015.

[161] Jacinto Rivero-Hernández, Annette Morales-González, Lester Guerra Denis, and Heydi Méndez-Vázquez. Ordered weighted aggregation networks for video face recognition. *Pattern Recognition Letters*, 146:237–243, jun 2021. doi: 10.1016/j.patrec.2021.03.021.

Samenvatting

In **Hoofdstuk 2** hebben we een gedetailleerde vergelijkende studie uitgevoerd om de prestaties van commerciële en open-source gezichtsherkenningssystemen te evalueren tegenover menselijke experts. Onze resultaten tonen aan dat commerciële software over het algemeen beter presteert dan zowel open-source alternatieven als menselijke experts, vooral voor foto's van het volledige gezicht. De studie onthulde echter ook dat deze geautomatiseerde systemen beperkingen hebben, vooral bij het verwerken van afbeeldingen van slechte kwaliteit of met obstructies, zoals veroorzaakt door mensen die petten, microfoons of sjaals dragen. Deze studie is een belangrijke stap vooruit in het begrijpen van de toepassing van geautomatiseerde gezichtsherkenningssystemen in het forensische veld.

In **Hoofdstuk 3** onderzochten we de impact van kalibratietechnieken op de schatting van aannemelijkheidsquotiënten (*Likelihood Ratios*). Met behulp van ENFSI-proficientietests als benchmarks, vonden we dat kwaliteitsscore-gebaseerde en kenmerk-gebaseerde kalibraties de naïeve kalibratiemethoden overtreffen. Hoewel commerciële software superieure prestaties vertoont, benadrukt de transparantie van open-sourcesystemen het belang van voortdurend onderzoek om zowel de effectiviteit als de verantwoording in forensische gezichtsbeeldvergelijkingen te verbeteren.

Hoofdstuk 4 richtte zich op het cruciale probleem van uitlegbaarheid in beoordelingen van gezichtsbeeldkwaliteit. We ontwikkelden een multi-task leermodel dat niet alleen geschiktheidsscores voorspelt, maar ook gezichts- en omgevingskenmerken identificeert die deze scores beïnvloeden. De dubbele capaciteit van het model biedt zowel hoge nauwkeurigheid als uitlegbaarheid, eigenschappen die cruciaal zijn in forensische omgevingen. Dit hoofdstuk benadrukt het belang van zowel geschiktheid als uitlegbaarheid in gezichtsherkenningssystemen. Onze bevindingen tonen een hoge correlatie aan tussen gezichtsbeeldkwaliteit en kenmerken zoals scherpte, frontale pose en leeftijd. Deze dubbele capaciteit van het model verbetert niet alleen de nauwkeurigheid van de kwaliteitsbeoordelingen maar biedt ook bruikbare inzichten die met name nuttig zijn in forensische omgevingen. Het model heeft echter beperkingen; het vereist goed gelabelde gegevens voor training en is afhankelijk van juiste gezichtsdetectie en -uitlijning voor een nauwkeurige beoordeling.

In **Hoofdstuk 5** verkenden we methoden om gezichtsherkenning in cameratoezicht te verbeteren. De studie vond dat het selectieve gebruik van frames van hoge kwaliteit en

het koppelen ervan met geschikte referentiebeelden de schatting van het aannemelijkheidsquotiënt verbetert. Het hoofdstuk onthulde dat het gebruik van een gewogen kwaliteitsgemiddelde van alle beschikbare frames nog betere resultaten oplevert. Belangrijk is dat het hoofdstuk de beperkingen en mogelijkheden voor verbetering in verschillende gezichtsherkenningalgoritmen besprak, inclusief de tegenintuïtieve bevinding dat superresolutietechnieken niet altijd voordelig kunnen zijn, vanwege het toevoegen van informatie en/of artefacten.

Acknowledgments

Just like my fellow countryman Don Quixote in his epic battle with the windmills, during the journey of this doctorate, I often found myself facing what seemed to be formidable giants. These 'giants', whether they were complex research studies or challenging academic papers, initially presented themselves as insurmountable foes in my noble quest.

However, akin to Cervantes' tale, these daunting giants revealed themselves to be mere windmills, conquerable through calm, perseverance, and steadiness. It would not have been possible to see the true nature of these challenges without the help of my own Sancho Panzas - those mentors, colleagues, friends, and family members who walked alongside me, offering wise words, unwavering support, and clear perspective in the most critical moments.

To them, I owe a deep debt of gratitude for helping me understand that challenges, when approached with time and perseverance, are no more than tranquil windmills dotting the landscape of my academic and personal growth. Their presence and support have been as fundamental to my journey as Sancho was to Don Quixote, and for this, I extend my most sincere and heartfelt thanks.

First and foremost, I am grateful to my promotors: Zeno Geradts, for the pleasure of sharing an office and for his close guidance, even during the challenging times of the pandemic; and Marcel Worring, who offered invaluable support and patience from Science Park, even when things took longer than expected.

My gratitude extends to my office colleague at NFI and coauthor, Arnout Ruifrok. Sharing an office with you was a joy, filled with long and fruitful conversations about face recognition, image quality, and likelihood ratios.

In the ASGARD project, my heartfelt gratitude goes to Juan Arraiza for his excellent direction, and to Sean Gaines, who gave me the opportunity to work at Vicomtech as a researcher, and with whom I now have the pleasure of collaborating on a paper. I also thank my ASGARD colleagues from Ulster University: Sriram Varadarajan, Bryan Scotney, and

Wang Hui. Working with you was a pleasure, and our periodic meetings and the conference in Belfast were deeply enriching.

I also acknowledge the Dublin University team, especially Owen Corrigan, for the successful collaboration in ASGARD and clothes classification with my Master student Jette Korthal-Altes. I couldn't have asked for better teammates.

At Vicomtech, my supervisor Luis Unzueta deserves special thanks for his supervision and assistance with our papers on face image quality. Working with the V4 team, including Unai Elordi, in San Sebastian on the SHAPES project was an incredible experience.

In the STARLIGHT project, I am thankful to Jorge Garcia and Mikel Aramburu for their assistance, patience, and support during the project and tool fests. The time spent in San Sebastian was unforgettable.

I express my pride and appreciation to my NFI MSc students: Marissa Koopman, Illias Batskos, Aida Ploco. Supervising your theses was a pleasure, and I am impressed by your excellent work.

Thanks are also extended to my fellow NFI interns, especially Marianna Bedeli, for making the days at NFI much more enjoyable and for the memorable long lunches.

Thank you to my fellow NFI colleagues: Jeroen Waarnar, for enlightening me about another side of NFI, DNA; Carlos Alberca and Paula Cortes, for being part of our small Spanish-speaking group within NFI; Christos, Thijs, Loes, for the enjoyable long coffees at NFI, and for the pleasure of discovering the new NFI coffee place in the hall.

I want to extend my thanks to the MultiX Group. Sharing our MultiX meetings, presentations, and dinners was great: Nanne, Yen-Chia, Teng, Sarah, Jia-Hong, Jiayi, Ivona, Tom, Ujjwal, Thanos, Wangtuan, Tim, Shuai, Carlo, Sadaf, Inske, Gjorgi, Javier.

I also thank the AI4forensics group: Merel, Floris, Gonçalo, Eleni, Meike, Conor. Our cake gatherings on Wednesdays were invaluable. It was a pleasure being part of the "Activity Committee."

I would like to thank the FBDA team at NFI. Working with you in a sprint was extremely pleasant: Elina, Rolf, Judith, Jeannette, Nivea, Jeroen, Ankie, Rachid, Simone, Wauter. I learned a lot about teamwork, Trello, and sprints with you. Also, my thanks extend to the rest of the FBDA group: Bart, Huub, Mathijs, Hannah. I am very happy to have been in your team. The 'lir' Python library has been a pivotal contribution to my research. And thank you to Valentine Arendsen and Lisanne van Dijk, for being wonderful team managers.

Special thanks to Diego Luque for the fruitful discussions post-ENFSI meetings and for introducing me to the ICC, an extraordinary place to work.

Also, I'd like to thank my friends here in Delft who have supported me during these years in the Netherlands: Mario, Edu, Alvaro, Javier, Jordi, Jerry, Elizabeth, Chara. And I also thank my far-away supporters from Spain, from three different regions. From Cantabria, I thank Diana, Carmen Yolanda, Doni, Martin, Jorge. From Madrid, my two good old friends Miquel Larsson and David Lopez. And from my hometown in Valladolid, it was nice to reunite during Christmas, summer, and Valladolid fiestas and reminisce about the old times: Marta, Antonio, Jessi, Manuel, Cristina, Adrian, Dani, Isa, Javi, Pedro, Carlos, Marta, Luisfer, Leti, Marco.

Finally, I'd like to thank my family back in Spain: Amelia, Bea, Teresa. Although she can't read this, thanks to my sweet cat Misi, who has kept me company, warming to the heat of my laptop (sometimes awake, sometimes not) during long nights writing my manuscript drafts in Spain. Thank you to both of my parents, Jose Maria and Sonsoles, who have always supported and encouraged me in my studies. Without them, studying first in Madrid and then in the Netherlands would have been impossible. Their support and values made this thesis possible. They have been a great support during COVID times and have provided a home to which I can always return. Thank you to my sister Maria and to Luis, who, with their creativity, helped me distribute my thesis in a very original way with their QR keyring. I couldn't have asked for a better gift. And thank you to Edwin, for sharing my passion for AI, for all the great moments being nerds together, for helping me to create this thesis cover and for showing me that the Netherlands (including the windmills) can be both different and beautiful.

Thank you, each and every one, for being the companions, guides, and supporters in this grand adventure of mine.