Ph.D. thesis:

A Systematic Approach to Fingerprint Identification via Source Probabilities

---

Etienne Pillin

under the supervision of

Prof. Jeremy Levesley and Dr. Cheryl Hurkett

Department of Mathematics
University of Leicester

Project #4
INTREPID Forensics

Leicester, United Kingdom 21st June 2019

Abstract

This research project was carried out under the INTREPID Forensics programme, a doctoral program involving 10 Ph.D. students in various fields applied to Forensic Science, and funded by the European Commission. The purpose of this project was to produce innovative methods of pattern recognition for fingerprint ridge lines in order to improve the reliability and the amenability of automatic fingerprint identification to the court. This research provides a preliminary but systematic and necessary approach to achieve this.

First of all, the premise, software, and methodology for a data collection were developed for the purpose of a ground-truth database suitable for research and the training of identification algorithms. Two novel mathematical formulations of the fingerprint identification problem were made: source determination and source assessment. The latter provides a basis for the computation of source probabilities, namely the probability for two finger impressions to come from the same source, which is not considered sound. Despite current consensus, this thesis has established a new approach that proves that this can in fact be done in a mathematically justified manner.

Finally, this research culminated with the development of feature detection algorithms that proceed by fitting a section of a fingerprint image by a function which locally models the ridge line accurately, and which demonstrated promising results. The fitting methods used rely on optimisation algorithms known as Estimation of Distribution Algorithms (EDAs), which have been generalised to the context of mixed-discrete optimisation, and implemented and applied to fingerprint feature detection.

Acknowledgements

I first want to thank my supervisors, Jeremy and Cheryl, for their continuous support throughout my Ph.D., and for believing in my potential. Jeremy, thank you for encouraging me to pursue innovative ideas for these past years, for pushing me to find ways in which this research could be useful to forensic professionals, and for helping me to discover how to communicate my mathematical ideas in such a way that they can be made understandable to the intended audience. Cheryl, thank you for your moral support, especially during the Ethics approval process - your advice helped me stay grounded during these moments, including during the thesis write-up process.

I also can't thank Jessica enough for her help through these past years. You helped me find a way to make my research impactful and applicable to Forensic Science by making me understand the stringent requirements of the field. This helped me realise not only the societal impact that such research could have, but also the responsibilites that we have as researchers to produce transparent and honest research. You also challenged me and helped me improve my academic (and spoken!) English, when I was at a point where there wasn't much else I had to learn about that language. Above all else, you helped me go through these years at the times where they were the most challenging, all the way up to the submission. Your presence was invaluable to making this experience a successful one.

I also want to thank my parents for their continuous support during and before my Ph.D., and Lisa, the coordinator of the INTREPID Forensics Programme, who created this from the ground up, and without whom none of this would have been possible. Thanks for supporting not only myself but the entire group, and for putting such effort in ensuring that the entire group is successful in their research -

and also for putting up with my constant delays in dealing with paperwork.

Last but not least, I would like to thank Prof. Ivan Tyukin and Prof. Christophe Champod for their constructive criticism during the review process. Your feedback was valuable and challenged me to improve the quality of this document as well as the way I present my research in general.

Contents

Glossary

ACE-V Analysis, Comparison, Evaluation, and Verification. 15

ACID Atomicity, Consistency, Isolation, Durability. 34, 36

AFIS Automated Fingerprint Identification System. 15

AFSP Association of Forensic Science Providers. 54

API Application Program Interface. 33, 38–40

CNN Convolutional Neural Network. 72, 73

DBMS Database Management System. 7, 34, 36

D.F.O. 1,8-Diazafluoren-9-one. 48

EDA Estimation of Distribution Algorithm. 1, 89, 91, 92, 94–99

GMM Gaussian Mixture Model. 92, 94

GPU Graphical Processing Unit. 72, 95, 96

IAI the International Association for Identification. 15

LTS Long-term support. 39

OS Operating System. 39, 43

PSO Particle Swarm Optimisation. 90

RDBMS Relational DBMS. 34

SSD Solid-State Drive. 43

SWGFAST the Scientific Working Group on Friction Ridge Analysis, Study and Technology. 15

VAE Variational Auto-Encoder. 75

Notations

$f \circ g$	Composition of g and f.
$\mathcal{D}(f)$	Domain of function f.
$\mathbf{Im}(f)$	Image of the function f.
$\mathcal{F}(A, B)$	Set of all functions from A to B.
$A^{\complement}$	Absolute complement of A.
$\varnothing$	The empty set.
$A \cap B$	Intersection of A and B.
$A \cup B$	Union of A and B.

Chapter 1

Fingerprints: purpose, definition and significance

The purpose of this Ph.D. is to develop a systematic approach for developing a fully quantifiable and accountable fingerprint identification system in forensic science. It relies on Machine Learning, a field of Computer Science devoted to the design of algorithms which learn from data. This is motivated by the fact that despite the morphogenesis of fingerprints having been studied theoretically in the field of Biology, the resulting appearance of fingerprints is the subject of many confounding factors, and there is a significant empirical component to their study.

Machine Learning applied to fingerprint identification systems can address a significant limitation of current systems, which is their incapability to successfully identify crime-scene originated fingermarks [202]. This Ph.D. also sets out to provide introductory results that can address another existing issue in current fingerprint identification systems, which is the fact that they are not designed to determine whether two impressions have been produced by the same finger [202]. Finally, this Ph.D. will offer an unconventional approach to Machine Learning in an attempt to address this issue, and in the hope of addressing the fact that the probative value of AFIS systems and the reasons why they may give erroneous results is hard to assess [202].

These objectives are ambitious and this Ph.D. presents introductory work

towards that direction. Solving these objectives entirely requires that work in areas beyond the field of Mathematics be undertaken, which is why a holistic approach to solving the identification problem was undertaken as part of this Ph.D. As such, this manuscript follows the following plan:

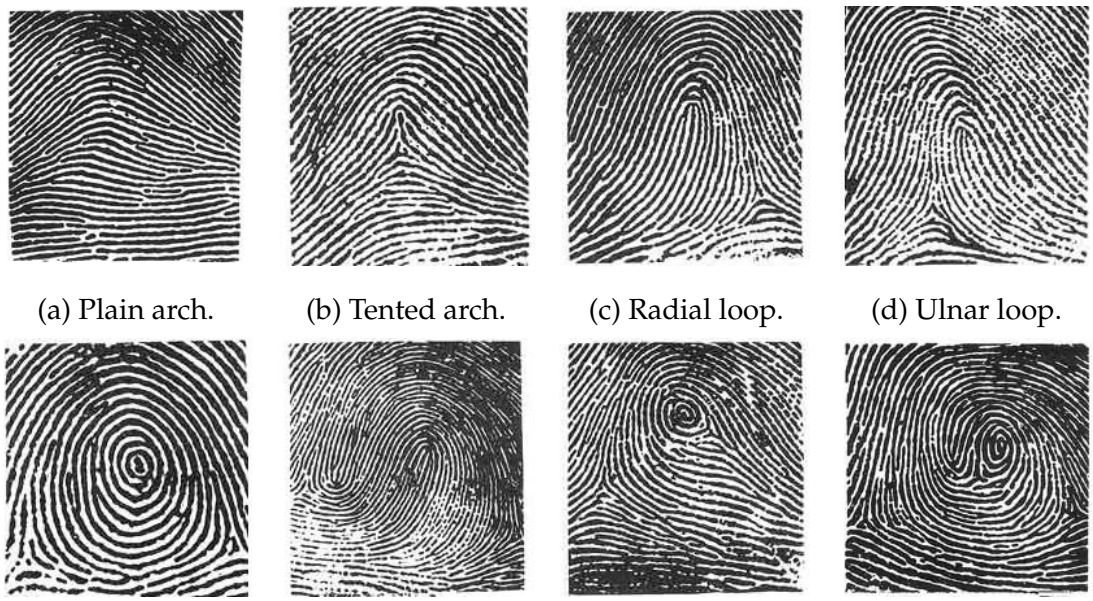
- fingerprint storage into a database,
- the analysis of fingerprints using image processing methods,
- the clustering of fingerprints,
- machine learning algorithms applied to clusters.

1.1 Fingerprints as Evidence

While fingerprints have been used as early as 18^{th} century BC for identification in Babylon and 3^{rd} century BC for legal documents in China, it was not until the 1860s that Sir William Herschel used them officially as a method of identifying criminals in India. It was then that Dr. Henry Faulds wrote the first scientific articles regarding fingerprints in 1880, and suggested their use in forensics to Scotland Yard in 1886. This suggestion was rejected at the time, most likely for lack of a solid scientific evidence of their uniqueness. Finally, Sir Francis Galton managed to classify fingerprints and estimate the probability of two humans having the same fingerprints. His findings were published in a series of papers and books dating from 1888 to 1895, and led to his recognition as founder of modern fingerprinting [12–14].

Sir Galton's classification of fingerprints relies on a number of parameters, the first of which is the overall shape of the fingerprints, which is commonly referred to as level 1 details. The different patterns he identified are: plain arches, tented arches, radial loops, ulnar loops, plain whorls, double loops, central pocket loops, and accidental patterns (see Figure 1.1). These are generally simplified into three categories: loops, whorls, and arches, which respectively account for 67.5%, 26%, and 6.5% of all fingerprints (estimation made on tenprints of 500 individuals) [12]. The formation of these ridge line patterns is a consequence of a buckling process

acting on the basal layer of the epidermis, which occurs between the 10^{th} and 16^{th} weeks of the pregnancy [15–19]. More precisely speaking, this phenomenon takes place on volar pads, which are "temporary eminences of the volar skin that form at about the 7^{th} week at the fingertips" [18]. Furthermore, the degree of asymmetry of the volar pad determines that of the ridge pattern: asymmetrical pads lead to loop patterns, while symmetric ones yield whorls. Finally, arches have been found to be a consequence of a late timing of the ridge pattern formation with respect to the volar pad regression process, both in symmetric and asymmetric volar pads [19].



(a) Plain arch. (b) Tented arch. (c) Radial loop. (d) Ulnar loop.

(e) Plain whorl. (f) Double loop. (g) Central pocket loop. (h) Accidental pattern.

Figure 1.1: Examples of fingerprints which are characteristic of each of the level 1 details [203].

Fingerprints are made of ridge lines which are not always continuous. They have imperfections, also called minutiae, which constitute the level 2 details of a fingerprint. The list of minutiae commonly include: ridge endings (end of a ridge), ridge bifurcations (a ridge splitting into two), short or independent ridges (ridge of "small" length), islands ("very small" ridges), ridge enclosures or lakes (ridge splitting into two, then merging back into one), spurs or hooks (short protrusion from a bifurcation), crossovers or bridges (short connection between two parallel ridges), deltas (triangular pattern in the ridges), and cores (center of the 1^{st}-level detail). Refer to Figure 1.2 for examples of minutiae.

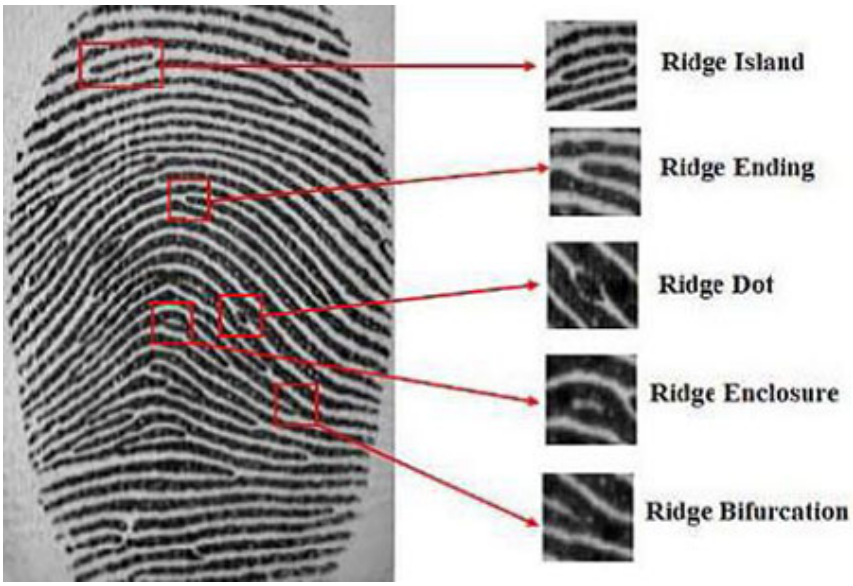


Figure 1.2: A fingerprint and some of its level 2 details [144].

Level 3 details refer to the specific shapes of the pores that form the ridges, and the outline of the ridges, see Figure 1.3. These are seen at a high magnification, but are less commonly accessible in crime scene fingerprints due to their susceptibility to distortion.

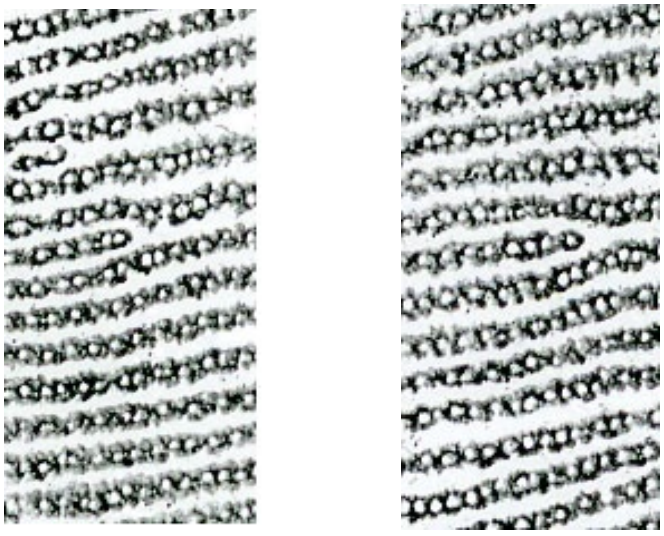


Figure 1.3: Level 3 details of two different fingerprints [20].

Nowadays, fingerprint evidence is commonly used worldwide to identify and convict criminals but its acceptance and usage vary across countries. Overall, it lies within the realm of forensic evidence, which is admissible in court via expert witness testimonies (see Table 1.1). Expert witness testimony is admissible in court

because fingerprint examination, along with many disciplines encompassed within forensic science, is considered an empirical science whose conclusions can help prove or disprove facts related to a legal procedure [20]. These facts rely on the assumptions that:

- fingerprints are unique to one given individual (uniqueness),
- an individual's fingerprints persist and remain identical throughout his/her life (persistence).

<table>
<tr><th>Topic</th><th>USA[1]</th><th>Canada</th><th>UK</th></tr>
<tr><td>Admissibility of evidence</td><td colspan="3">According to the Daubert standard [116–120], evidence must meet the following criteria:
relevance;
necessity in assisting the trier of fact;
should not be subject to any exclusionary rule;
must be given by a properly qualified expert.</td></tr>
<tr><td>Admissibility of fingerprint evidence</td><td>Treated as forensic evidence ([121], in particular rule 4).</td><td rowspan="2">Constitutes as evidence ([122], section 667).</td><td>Vaguely described, is used to identify individuals [123].</td></tr>
<tr><td>Admissibility of fingerprint examination</td><td>Treated as expert witness testimony ([121], FR702-706).</td><td>Treated as expert opinion, the jury should decide what value to give this opinion [124].</td></tr>
<tr><td>Fingerprint obtention by police forces</td><td rowspan="2">Identification records (which include fingerprints) may be acquired, collected, preserved and exchanged [125].</td><td>Can be taken without consent if in lawful custody or charged or convicted of an indictable offense [126].</td><td>Can be taken with written consent for an investigation of a criminal offense, without consent if arrested or charged for a recordable offense [123].</td></tr>
<tr><td>Fingerprint retention by police forces</td><td rowspan="2">Can request to have one's fingerprints erased [127–130].</td><td>Indefinite for a convicted adult [131].</td></tr>
<tr><td>Data rights</td><td>Wrt. police use, can be requested and challenged [132]. Wrt. commercial use, companies have the obligation to notify clients [133].</td><td>Can be requested [134].</td></tr>
</table>

[1] At the federal level.

Table 1.1: Simplified overview of the legislation regarding fingerprints across several countries.

American [116–118, 135] and Canadian case law [119] clarified the requirements

for expert testimony to be admissible in law: the witness has to be an expert in the field, whose results are peer-reviewed and published. As a result, well-established forensic science organisations such as the Scientific Working Group on Friction Ridge Analysis, Study and Technology (SWGFAST) and the International Association for Identification (IAI) define standards for fingerprint examinations. In SWGFAST's official documents [136, 137], the words *fingerprint identification* and *individualisation* are preferred to *fingerprint match*. They are defined as follows:

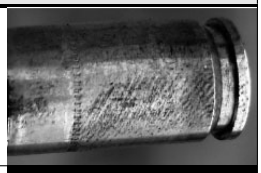
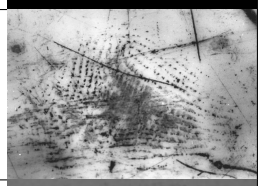
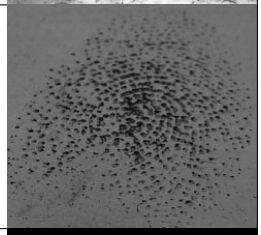
> "[A] decision by an examiner that there are sufficient discrimination friction ridge features in agreement to conclude that two areas of friction ridge impressions originated from the same source. Individualization of an impression to one source is the decision that the likelihood the impression was made by another (different) source is so remote that it is considered as a practical impossibility."

Such decisions are commonly made by first using an automated recognition software (also called Automated Fingerprint Identification System (AFIS), in reference to the FBI system), which returns a set of n likely matches within a given database (which can be local, national, or international) [138]. An expert will then perform a series of quantitative and qualitative assessments and comparisons according to the Analysis, Comparison, Evaluation, and Verification (ACE-V) methodology in order to reach a conclusion [137]. The extent to which the protocol is followed, however, may differ depending on the working environment, leading to numerous psychological biases that can affect the decisions. Even the use of algorithms can be manipulated to facilitate confirmation bias. For example, algorithms that use the annotation of the fingerprint - as opposed to the actual fingerprint image - can lead some examiners to input several different annotations in order for the algorithm to yield fingerprints that the examiner believes are more likely to come from the same source. The current system of decision making is therefore prone to bias and error [202].

Given that a fingerprint identification alone can be sufficient for conviction [139], the consequences of a wrong identification are dire. Since this system mostly relies on human-based decision making and that the very definition of a match is

subjective, it is conceivable that errors and abuses can happen, which experience has shown [140–143, 21]. As a result, there is growing concern about the possibilities of identification [22, 204, 23] and scepticism about the uniqueness of fingerprints [24, 25]. Galton's mathematical proof of uniqueness is rudimentary, and subsequent research has not lead to a definitive proof [26]. Instead, statistical research in this field is now aimed at estimating the probability distributions of fingerprint features [27] and quantifying the weight of fingerprint evidence in court [28].

In order to address these concerns, research is being done to analyse [29] and improve human fingerprint identification. In addition to this, the identification process could benefit from improved, or even completely autonomous, identification systems. Fingerprint identification is a difficult problem to solve due to the large amount of intra-class variability in the fingerprint patterns [30], which is exacerbated by the large amount of distortion and degradation present in crime scene originated fingermarks. See Table 1.2 for a list of the different artifacts that can be present in such fingermarks.

Category	Traits	Possible causes	Example
Deposition	Partial fingerprint	Incomplete contact with the substrate due to destruction, small subtrate, way the item is seized, obstruction	
	Slippage (smudges)	Movement during deposition	
	Double taps	Overlapping prints	
	Pressure distortion	Differences in pressure	
	Variying clarity	Varying levels of matrix	
	Pooling	Excess of liquid matrix on a horizontal surface	
	Feathering	Little amount of liquid matrix and smudging	

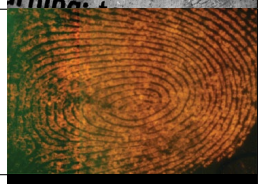
Category	Traits	Possible causes	Example
	Dripping	Excess of liquid matrix on an inclined surface	
	Other patterns (e.g. splattering, misting)	Confounding factors due to events that occur during bloodletting incidents	
Substrate	Ridge disruption	Textured surface (e.g. plastic wrapper, duct tape)	
	Complex background	Unclean surface, coloured surface, glass, plastic, thermal paper, paper money	
Development method	Air bubble & tape folds	Lifting tape	
	Glare	Photograph on glass, plastic or metal	
	Curvature	Photograph on curved item	
	Negative impression	Gun blueing, vacuum metal deposition	
	Fluorescence	Ninhydrin w. ALS, indanedione w. Zn/Cl and ALS, cyanoacrylate w. fluorescent dye, D.F.O. w. ALS	
	Speckles	Granular powders, indanedione w. Zn/Cl and ALS	
	Perspective distortion	Photograph not taken parallel to the fingerprint	

Table 1.2: Traits commonly found in fingerprints developed in practical scenarios [13, 31].

Chapter 2

Fingerprint Databases for Research

The purpose of this chapter is to provide a plan for the implementation and population of a database of fingerprints which is suitable for the purpose of this project, namely the training of Machine Learning algorithms for latent fingerprint identification; and to describe the efforts that have been undertaken in this direction as part of this doctoral work.

This chapter is broken down into sections which discuss the different aspects which need to be addressed for a database with the parameters mentioned above to be established. As such, Section 2.1 describes the fingerprint databases for law enforcement and for research and states the requirements of the one pursued by this project; Section 2.2 explores the possibility to populate the database with artificially-generated fingerprints ; Section 2.3 reviews the ethical ramifications of this project; Section 2.4 delves into the technical aspects of implementing the database; and Section 2.5 describes a scenario-based collection of simulated crime scene fingermarks for the initial population of the database .

2.1 Background and Requirements

Numerous private governmental fingerprint databases exist worldwide in order to enable the use of automated fingerprint identification systems for law enforcement purposes, see Table 2.1 for a list. Additionally, several datasets and

databases have been established worldwide for the purpose of enabling research in biometrics and forensics science, see Table 2.2 for a non-exhaustive list of such efforts. The difference in size between operational and research-oriented databases is in part due to the fact that developing fingerprint images suitable for forensic research is very time-consuming and labour-intensive, and that this collection is governed by the strict legislation and ethical standards regarding biometric data. See Table 1.1 for a succint comparison of the different legal aspects of fingerprints across several countries.

Name	Description	Size	Additional personal data stored
NGI[1] (formely IAFIS[2])[145, 146]	FBI database for investigation	75.3 million criminals, 60.5 million civilians, 2 million individuals of special concern	Criminal histories, mugshots, scars and tattoo photographs, physical characteristics (height, weight, eye, hair colour)
IDENT1 [147]	UK national database for investigation	6.5 million tenprints	
FAED[3] [148]	French national database for investigation	4.6 million individuals	Gender, ID if known
EURODAC [149, 205]	EU database for the identification of asylum seekers and irregular border-crossers	2.7 million sets of fingerprints	
Australian AFIS [150]	Australian national database for investigation	2.6 million tenprints	
INTERPOL Fingerprint database [206]	Database for international collaboration in criminal investigations, restricted access	$>$ 150'000 sets of fingerprints, $>$ 9'000 crime scene marks	

[1] Next Generation Identification
[2] Integrated Automated Fingerprint Identification System.
[3] Fichier Automatisé des Empreintes Digitales.

Table 2.1: Non-exhaustive list of notable fingerprint databases for law inforcement worldwide. Empty red cells signify that the information is not publicly available to the researcher's knowledge.

Efficient, detailed, large-scale databases are critical for developing advanced, robust systems for all applications [36], including the development of AFIS systems. In the field of Forensic Science, a growing number of practitioners believe in

the necessity to develop a reference research-oriented fingerprint database [34], including the UK Forensic Regulator [207], which could help meet the requirement for forensic laboratories in the UK to comply with the ISO 17025 standard by October 2018 in terms of their fingerprint comparison practices [208].

Tremendous potential in terms of resources and efforts are wasted in individual data collections devoted to the purpose of a single research project. This is caused by the traditional ethical submission and approval process which every research project must undergo at its host institution, and which customarily requires all data collected to be destroyed after the duration of the project has expired. At the time of writing, no storage efforts have been undertaken that are sufficient to meet the collaborative and accessible nature of the database outlined in this project, and no current standards exist. This Ph.D. therefore aims to set the standard for institutions involved in forensic science research.

The greatest impediments to the successful undertaking of one such collaborative database are the risks associated to it, and the ethical issues that therefore need to be addressed, as demonstrated by the feedback which has been given to this project (see Section 2.3). These challenges are caused by: the sensitive nature of the data; the context in the matter of personal data being stored and shared online at the time of writing, which is punctuated by personal data breaches and scandals of personal freedoms being impeded online, such as the Cambridge Analytica scandal [37]; and finally, the absence of a precedent which could provide a complete ready-made solution. For these reasons, there is value in discussing the premise and possibilities of one such database, in order to converge towards a satisfactory premise and design for such a project. This will also serve to educate Ethics Committees and other decision makers regarding the efforts undertaken to address the risks and ethical implications, as well as the missed potential that negative decisions represent.

As such, this research argues in favour of a worldwide, collaborative fingerprint database for research in forensic science. Such a project could: facilitate collaboration in the field by providing access to data sets used for research and publications [38]; improve transparency by facilitating the reproduction of research results [39]; offer

new possibilities for future research; and allow for long-term access to the data collected. Transparency in particular is especially needed for court purposes, and yet at the time of writing, commercial AFIS systems are not sufficiently transparent due to the fact that their methods remain proprietary [202].

In order for this database to truly be collaborative and to act as a suitable training set for Machine Learning identification algorithms, it must meet the following technical requirements:

(R1) allow for facilitated, software-assisted input and output by means of a website and an API, which has been previously explored by [33, 39];

(R2) be relevant to the forensic fingerprint identication problem, which means that it should encompass both fingerprints of good quality that can act as reference fingerprints, and fingermarks which are representative of those found in crime scene conditions (the difference between both is exemplified in Figure 2.1);

(R3) be ground-truth, meaning that all measures should be undertaken in order to increase the fidelity of the data contained, provided that the associated ethical risks are manageable. This is especially crucial regarding the source of each finger impression.

At the time of writing, neither law enforcement nor research databases meet all of these requirements simultaneously.

Figure 2.1: Livescan fingerprint to the left, and photograph of a developed latent fingerprint on a polyethylene bag which has clearly been degraded [40].

Name	Description	Size	File format	Additional personal data stored
NIST[1] Fingerprint databases [151]	NIST's public databases for the evaluation of fingerprint classification systems	$\sim$ 60′000 prints	JPG, 8-bit grayscale, lossless compression	None
CASIA-FingerprintV5 [152]	CASIA[2] database for research and education, accessible upon signup	20′000 fingerprints from 500 individuals	BMP, 8-bit grey-level	None
FVC2006[3] [153, 32]	Databases used for the FVC2006, accessible on request	7′200 fingerprints	BMP, 256 gray-levels, uncompressed	None
WVU[4] multimodal database [33]	Research dataset for Biometrics	7′136 fingerprints		Face, iris, hand, palmprint, and voice recording
ELFT public challenge #2 [34]	NIST data set for the evaluation of fingerprint identification algorithm	1′100 fingerprints	WSQ	
IIITD[5] Multi-surface [35]	Research data set of fingerprints developed from different surfaces, accessible on request	551 fingerprints, 51 individuals		None

[1] National Institute of Standards and Technology.
[2] Chinese Academy of Sciences' Institute of Automation.
[3] Fingerprint Verification Competition.
[4] West Virginia University.
[5] Indraprastha Institute of Information Technology Delhi.

Table 2.2: Non-exhaustive list of notable fingerprint databases and data sets for research worldwide [34]. Empty red cells signify that the information is not publicly available to the researcher's knowledge.

2.2 Artificially Generated Fingerprints

Before even attempting to populate a database, the method of generating fingerprint data must first be considered. Two different approaches of generating fingerprint data exist - using artifically-generated fingerprints which is the most time-efficient approach, or undertaking a data collection of real fingerprints. This section explores the process of generating artifically-generated fingerprints and explains why they are not appropriate for a forensic research database.

The generation of artifical fingerprints has been accomplished by Kücken and al. in their computational study of fingerprint formation [18, 44, 45]. Their work relies on previous research in Biology that demonstrated that friction ridge pattern formation is a consequence of a buckling process in the epidermal layers of the skin during prenatal growth [15–17], as mentioned in Chapter 1. Kücken and Newell modeled this formation process computationally by determining the stress field given the boundary and normal forces imposed on that geometry, and then computing the resulting buckling pattern from the stress field obtained previously. The latter has been accomplished either by minimising the elastic energy of the system [44], or by solving the Föppl-von Karman equations [18, 45], which are derived from the former. See Figure 2.2 for some examples of the resulting generated fingerprints.

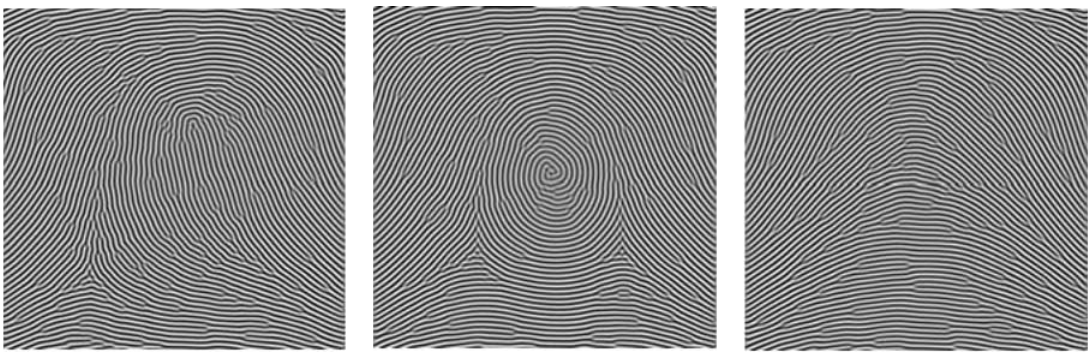


Figure 2.2: Fingerprint images representing the three main patterns: loop, whorl and arch (from left to right). These images were generated numerically by modeling the fingerprint formation process with different boundary and normal forces applied on the geometry [18, 44, 45].

Although this approach is scientifically accurate as it involves the modeling of the underlying physical phenomena, their mathematical formulations (either the elastic energy or the set of PDEs) requires the knowledge of the boundary and normal forces to which the volar pads are subjected. While these parameters were deduced based on the knowledge of the fingerprint formation process, they were not calibrated based on actual measurements. This is sufficient to study the theory of fingerprint formation, but not to generate artificial fingerprints that will be used for the purpose of automatic identification. Additionally, the results obtained are not realistic enough because this method does not take into account all the processes that occur after the formation of the fingerprints, such as events that can alter a fingerprint's appearance during the individual's life, or any circumstance surrounding the deposition and the development of the fingermark itself.

Research has more specifically been devoted to the generation of fingerprints for a large dataset, with the aim of producing fingermarks that are as realistic as possible [46–48]. This research culminated with the development of the software program SFinGe [154] (see Figure 2.3). They take a different approach by generating the fingerprint's geometry based on visual parameters such as the shape of the ridge lines, as opposed to modeling the physical phenomenon governing fingerprint formation. By focusing on visual parameters, the software manages to generate such convincing images that it has been used to generate one of the databases in the Fingerprint Verification Competition [154].

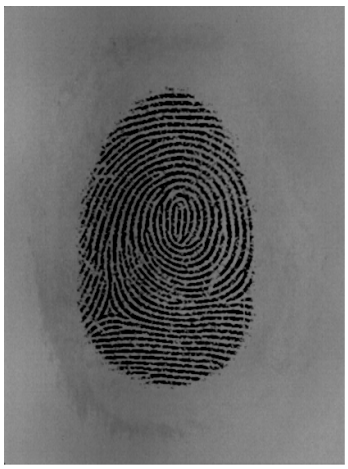


Figure 2.3: Fingerprint image generated with the SFinGe program [46–48].

While this avenue of generating artificial fingerprint images produces visually

more convincing results, substituting real data with generate data for the purpose of training humans and algorithms creates a risk that they become proficient with dealing with the latter, and not the former. While there is undeniable merit to the study of synthetic fingerprint generation from a scientific perspective, these methods should undergo strict validation in terms of: a) the fidelity of the images to real ones, based on the visual assessments of fingerprint examiners and quantitative assessments to be defined; b) their ability to precisely reproduce the parameters of the distribution of fingerprint images observed in the human population, including but not limited to the inter-class and intra-class variations, and the correlations between the presence and position of distinguishing features in the same fingerprint. The fact that the quantification of these parameters has not yet been established by current research [202], and that such an endeavour would require the collection of an extensive and appropriate data set, further justifies the purpose of the collaborative database described here.

Additionally, for the specific purpose of training Machine Algorithms, generating synthetic fingerprint images would not produce ground-truth information regarding their origin or the circumstance in which they have been found and developed. For these reasons , artificially generated fingerprints are not an appropriate means of generating fingerprint data that will act as a training set for fingerprint identification algorithms . Instead, real fingerprints emanating from a data collection will be used. Information pertaining to the donors, and the conditions in which each fingerprint has been collected will be documented so as to satisfy the requirement outlined in Requirement (R3). The specific scope and protocol the of data collection are described in Section 2.5.

2.3 Ethical implications

Forensic science aims at providing expertise and exactitude to the court in order to ensure that the legal system reaches accurate conclusions, and so that justice is served. As such, forensic science bears a crucial societal role. Therefore, ethical issues should not be tolerated any more than miscarriages of justice, and unethical

science is no more acceptable than inexact or sloppy science. Innovative science cannot be pursued at all costs, and it is capital for the ethical aspects of this project to be studied in order for it to have a positive societal impact.

This section sums up the main ethical aspects of the project which were discussed during the Ethics submission process at the University of Leicester, which took place from February 2016 to January 2017. They pertain to: the nature of the data stored, the individuals who have access to the database, and the lifespan of the database. Incidentally, these points are also the main aspects of most data protection legislation, including the UK Data Protection Act 1998 [134].

The nature of the data stored in the database is the most determining factor in terms of the level of risk that this project represents. The database aims at storing fingerprints, along with information pertaining (but not limited) to them such as their minutiae, and the quality according to the Bandey scale. There is also great research potential in storing data pertaining to the donors, such as, in order of increasing ethical concern:

a) their age and biological sex, their diet on a set period of time, or other factors that have already been scientifically identified as having an influence on the appearance of the fingerprints;

b) other factors that have not yet been linked to fingerprints, but that may or may not be worth investigating due to their potential value as evidence, intelligence, or for research;

c) their e-mail address, which can be stored in order to automatically ensure that there is no redundancy in the database (i.e. a situation where an individual gives his fingerprints on multiple occasions, but their fingerprints are not attributed to the same source).

While Data a) is acceptable provided that the donor agrees to provide each specific information, Data b) should be subject to a case-by-case review as donors may not understand the relevance of the information requested if a causality link with fingerprints has not been established by previous research. This review should therefore be done based on whether that information would be too intrusive, and

also if it can be used in addition with other information used to uniquely identify a donor. Regarding Data c), its purpose is to only be used internally by the database and to not be made public. However, that data may still be accessed by individuals maintaining the database, and those who breach its security. The inclusion of that data, as opposed to identifying each donor by a private ID of which they are in charge, therefore poses an additional risk as the database may be used for: cross-referencing with other databases and uncovering more information about an individual; framing a donor, provided that a physical fingermark can be deposited from an image file; and finally, intelligence and conviction purposes should the police access the data.

The second ethical aspect of this project concerns the range of individuals who have access to the database, which includes both the users of the data, and the contributors to the database. Providing access to the database to a wider audience increases the risk of the database being misused. Advertising the database also indirectly increases the probability of its security being breached. The less an individual is educated or involved in a project, the more likely they are to misuse it, and to be ignorant to its moral implications. Some of the suggestions which were brought up during the ethics process to mitigate these risks include:

(S1) requiring users to create an account and authenticate when accessing the database;

(S2) ensuring that all users of the database are properly informed of their rights and duties, and of the ethical implications of the database they access;

(S3) restricting the database for use within the University, or within a set list of countries;

(S4) or restricting the range of uses of the data (e.g. restricting the usage of the database to research applications and only allowing users associated to a verified academic institution).

The last topic of discussion during the ethics approval process pertained to the duration for which the data is to be stored. The purpose of this database is to serve as a reference database which future researchers may use and upon which they

can expand. Therefore it was suggested that the data be stored for an indefinite duration, which complies with the UK Data Protection Act as the data is meant for research use. This raises the issue of how the database will be managed over that duration. The solution offered was that this be made a commercialised research project managed by the researcher and his supervisors. Consequently, the support of the Research and Enterprise Division was sought.

In addition to the above, it was suggested that the ethical risks be mitigated by asking the donors what kind of data they are willing to provide, and to what kind of usage they consent. It is also more respectful when asking for consent in a physical or an online form to assume that they do not consent to anything by default. Ideally, the database should be periodically reviewed or audited by a third party in order for an up-to-date assessment of its risks to be made. All of these guidelines were compiled into terms and conditions as well as an FAQ that were to be used on the database's website. Unfortunately, these documents were never fully developed due to the lack of structure in the discussions with the Ethics Committee, as well as the lack of involvement of the Research and Enterprise Division, despite having secured the University of Leicester Prospects Fund (which amounts to $£10,000$).

On the 24th of November 2016, after almost 10 months of discussion about the ethical implications and the premise of the project and its research implications, the Ethics Committee suggested that this project be subject to a two-stage approval. According to these terms, the database would first be implemented and restricted for use within the University only. It would then have to be evaluated again by the Committee in order for it to be extended for use to other Universities or companies. However, the Committee did not specify what their requirements were, or if they agreed to any of the above premises of the database in terms of the type of data allowed, the range of users accepted for the second stage, or the duration for which the storage was allowed. They also did not specify according to what criteria or when that second evaluation would take place. In order to clarify how to move forward with the application, a meeting with the Chair of the Ethics Committee was organised. Unfortunately, the Chair came uninformed of the developments of the application over the last months, ultimately suggested that a new application be

made under a different name, and did not provide any additional insight into how the new application should be presented.

Following this, because of the lack of meaningful feedback received and the pressing need for data (as the active research part of the Ph.D. was due to end in October 2017), it became obvious that concessions needed to be made. Consequently, a very standard ethics application was filed on the 29th of November 2016 in order to provide a minimal data set that would act as a proof of concept for the scientific approach to fingerprint identification presented in this thesis. This new ethics application was made to be very similar to other applications related to fingerprint data collections. This meant that all the original and innovative aspects of the project such as its collaborative aspect, the storage of the data in a database, and an indefinite storage duration were removed in order to expedite the process of ethics approval. Despite the lack of innovative features, this new application was received as a very positive development by the reviewers:

> "This appears to be a much more focused version of an ethics application that the committee was asked to review about 6 months ago, and which the majority of committee members had fundamental ethical concerns about. This time, it is evident that the researcher has taken on board that constructive criticism, which is really good to see and this is a much better organised application from an ethical standpoint."

This application only required minor modifications, and was approved on the 31st of January 2017.

The reason why this process took so long and was ultimately unsuccessful are manifold. Firstly, the standard online application form, which is used across the entire University, was not well-suited for this application. It is suitable for some applications, such as small scale data collections, but is unfit to deal with complex or large-scale collaborative projects such as this one. This is due to the fact that there is no clear or appropriate part in which to discuss technical complexities or concerns. For example, in the "Project" part of the application, the questions focus on funding and participant recruitment, as well as a very general overview of the

project. The questions are not focused on, nor do they allow the explanation of technical details. Another example of this is in the "Permissions" section, where there is only a vague text box that asks about "legal, cultural, religious, or other" implications. The "Consent" and "Procedures" sections are almost exclusively formatted with yes-no questions, and the only text boxes in which the researcher has any chance of explaining anything are aimed at justifying these questions, rather than allowing the researcher to express complex technicalities. As a result, completing the online application took 2 months prior to its submission in February 2016, despite prior training in Ethics and a supervisor's assistance. Finally, the result was that the application itself was insufficient in addressing the ethical concerns surrounding this project in the absence of additional documents.

Secondly, there is no record of the level of risk that is deemed acceptable by the Ethics Committee, no history or record of real case research that has been accepted or refused, and no available example of failed and successful applications. Upon discussions with colleagues, it was discovered that the level of rigor and risk accepted by the same Committee can greatly vary, perhaps based on the specific reviewer involved.

Thirdly, there was no structure in the discussions that occurred. The discussions took place via notes that were sequentially added to the applications by the reviewers and the applicant. The researcher addressed the reviewers' concerns by regularly updating a Word document that was attached to the application. This lead to confusion, resulting in misinforming the reviewers about what solutions were suggested to address their concerns.

Finally, the Committee was not well-suited for this ethics review. The Committee is assigned to the University's College of Science, and none of the reviewers assigned are researchers in Computer Science, Mathematics, or Forensic Science. As a result, elementary Computer Science concepts and their implications were not understood, and even the premise of the database was unclear to them. This further made them feel like their concerns were not addressed.

Future advice for the ethics application process includes having very clear guidelines, which are currently lacking according to the researcher's experience.

These should be available before starting any application. The Ethics Committee should assess its ability to evaluate a given proposal, and either recuse themselves or appoint others with the appropriate background to assist them if needed. At the time of writing, this is the responsibility of the researcher as per the University of Leicester's Research Code of Conduct [213], section 3.2.5:

> "Researchers are responsible for ensuring that the correct sub-committee reviews the research."

Although it is important to include non-specialists during the assessment of an Ethical Application, it is equally important to include specialists that are so that a balanced consideration can be made. In addition, the Ethics Committee should be in charge of helping innovative research processes, which means providing clear advice and guidance as to how their concerns can be addressed. Conversely, if they do not believe a project should go through, this response must be clearly stated and given in a timely manner. In the case of complex applications, face-to-face meetings should be scheduled as soon as possible in order to collaboratively establish a document that lists the ethical concerns of the project. Acceptable solutions to these concerns should also be clearly established at this time. If a project aims to be a University or worldwide collaboration, or if it has commercial implications for the University, there should be a clearly defined procedure to start a discussion whereby all the different actors in the project can discuss what needs to be done in order for the project to move forward.

The suggestions mentioned above are meant to serve as an analysis of how the process could and should be improved in order to expedite Ethical approval processes without lowering ethical standards. The University itself acknowledged that there are severe delays in Ethical approval processes (see Appendix D), proving that there is a strong need for change to occur. If other researchers are not to be prevented from being successful in similar endeavours, the obstacles mentioned here must be duly addressed.

2.4 Implementation of the database, website, and API.

The purpose of this section is to provide details about the implementation of the database, as well as the website and Application Program Interface (API) which are interfaces to interact with the former in order to make the data more accessible to researchers..

A proof of concept was created for the front-end and back-end source code necessary for the definition of a website and API. The purpose was to provide a proof-of-concept which could convince the Ethics Committee of the feasibility of the project. This represents over 12,000 lines of code in Javascript, HTML, and CSS.

The following sections describe some of the aspects that have been considered during the implementation of this proof of concept. Those are technological decisions that have been made in order to develop the features required for this database, and the potential requirements of the Ethics Committee.

2.4.1 Web framework

The web framework refers to the program run on the server in order for it to host a website as well as make it accessible. These programs can also access locally hosted databases in order to make their content available via the website. This choice determines the ease of several things such as: establishing definitions; establishing access with other technology (like state-of-the-art cryptographic algorithms); maintaining the website in the future; and the performance of the website, among many other things. Given the number of web platforms and frameworks available (see [155] for an up-to-date listing and ranking) and the pace at which the environment of web development evolves, making a good long-term choice in this matter without prior practical experience is arduous. Choosing a solution which is fast, scalable, and popular enough to ensure that it will be well-maintained and updated in the foreseeable future is paramount. Given those concerns, the Javascript platform Node.js was chosen for this proof of concept [156].

Although Node.js is far from being the most popular solution at the time of

writing, its use is growing rapidly (see Figure 2.4). This is due to its accessibility, its scalability, its clear documentation, and very active community [157, 158]. It relies on V8, the engine developed by Google in 2008 that compiles Javascript into machine code, which addresses the performance drawback of interpreting Javascript in real time [159]. It also features a non-blocking I/O, event-driven paradigm, which allows it to be scalable and deal with many concurrent requests despite the fact that it is a single-threaded application, as opposed to most other platforms [160–162]. Last but not least, it comes with `npm`, a package manager that gives access to countless tools developed by the community [163].

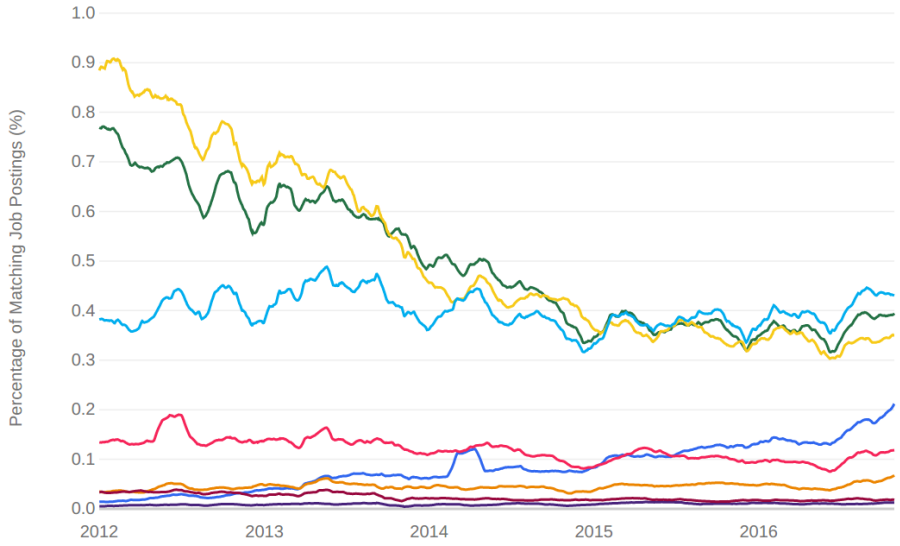
It remains that Node.js is an excellent solution for reaching proofs of concept quickly [166], and the researcher believes that it has great potential for dissemination and development in research.

2.4.2 DBMS

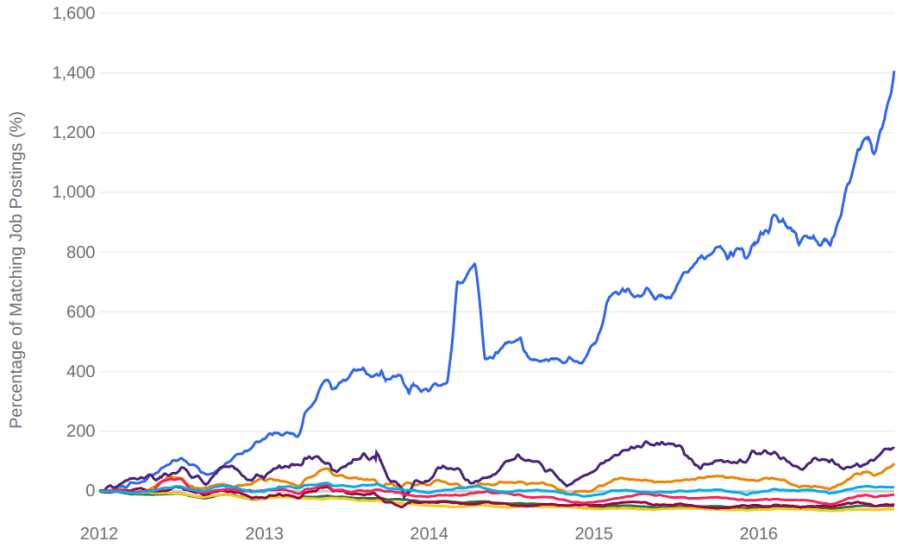
The Database Management System (DBMS) is the program which manages the database. It defines how the data is stored on disk and how it is accessed, and therefore has severe implications on the performance of the database. Additionally, choosing one which is well-implemented, well-documented, and well-maintained is important for the same reasons as the web framework.

Relational DBMS (RDBMS) have been around for decades and are still traditionally used. NoSQL, for Not Only SQL, DBMS have been increasingly used in the web community during the last decade, mainly because of the added flexibility and scalability of this new technology [167]. Despite that, there is no clear consensus in online communities concerning whether SQL or NoSQL is more effective, and SQL databases remain more popular overall [168]. This disagreement is most likely a consequence of the fact that the added flexibility of NoSQL solutions is achieved at the expense of ACID properties, which are properties that ensure that the database processes transactions reliably, and which are not complied by NoSQL DBMS.

Additionally, there are many different categories of NoSQL: key/value-based, column-based, document-based (or file-based), graph-based, each being more



(a) Graph in absolute scale.



(b) Graph in relative scale.

Figure 2.4: Graphs of the trends of various web frameworks in online job offers. Node.js, django, ASP.NET, j2ee, symfony, zend, ruby on rails and ASP.NET MVC were compared, and Node.js is represented in blue. The first graph is in absolute scale, while the second one is in relative scale, see [164, 165] for an up-to-date version. These graphs have been computed by Indeed.com.

efficient for different uses [49–51]. These terms refer to the general data structure used to organise the database and is very influential in determining how data is stored and accessed. Therefore, choosing a type of NoSQL database which is inappropriate for the required use may result in some data accesses being either very computationally expensive or even impossible in the worst case.

Based on the above arguments, relational databases are still useful in cases where the relationship between the different data entities is unlikely to change drastically over time. Their implementations also have the advantage of having been finetuned for decades, being very accessible to other programming languages, and being written in syntactically-similar languages.

Given the flexibility and versatility of NoSQL databases, they are very well-suited to allow a complex and / or quickly-evolving data structure. From a practical perspective, this means that it is possible to give researchers a set degree of freedom in annotating the data they collect in that they may add their own custom fields to the fingerprints. As a result, they are able to better describe and annotate data. SQL DBMS do not offer that flexibility. This makes NoSQL very valuable for research purposes, as mentioned in [39]. Both MySQL and the document-based NoSQL DBMS MongoDB [169] have been experimented with and compared in the early stages of the development of the database. On the basis that development with MongoDB was more straightforward; that the Mongoose package [170] to run MongoDB requests within Node.js is much better maintained and secure than those for MySQL; and that the structure of the database will evolve depending on the ethical and legal requirements of the partner countries, MongoDB was chosen for this project.

A final alternative to SQL and NoSQL is NewSQL. As the name suggests, it refers to a class of DBMS which offer NoSQL's flexibility and keep the relational mindset and desirable ACID properties of SQL [52]. As the landscape of such database implementations is fast-moving and heterogenous, those solutions have not been explored in the course of this Ph.D.

2.4.3 Donor anonymity

As mentioned in Section 2.3, donor anonymity can be implemented by very different means: storing the identity details of each donor but keeping that information private; storing a proxy such as the donors' email addresses and keeping them private; or not keeping any identifying detail and instead assigning a unique identifier to each donor.

The risks of storing any direct or indirect identifying information is that this information can be misused in the event that the database's security is compromised. This risk can be mitigated by storing identity details or a proxy in a different database which will be only accessed internally to perform redundancy checks; or even by hosting that database on a different server. This point was highly insisted upon by the Ethics Committee, although it may or may not result in increased security.

The second alternative, which consists in using unique identifiers to refer to each donor, can be achieved by having the database generate these identifiers upon the addition of a new donor. This is accomplished by default in MongoDB as any entity has an _id field, a 12-byte unique identifier, which is generated based on timestamp, machine ID, process ID, and a process-local incremental counter [171]. As a result, the uniqueness of each identifier is ensured throughout the entire database. However, this solution entails that each donor be in charge of their identifier, meaning that redundancy errors could occur as a result of them not remembering whether they participated, or not remember their identifier. Similarly, a donor who forgets their identifier may not have their fingerprints removed from the database.

Whichever alternative is used, it is important that the privacy and respect of the participants are kept. In addition to the security measures outlined here, no personally identiable information will be shared throughout the entire process of data collection and data storage, and any data that is shared such as fingerprint and participant-related information will be used for research, police training, and development purposes exclusively. Participants have the possibility to withdraw

their consent at any time, and have their fingerprints and data removed from the database. Additionally, users will be provided flexible means of accessing the database via a website and an API. As mentioned earlier, the latter allows secure programmational access to the database, in a way similar to social networks [172–175]. Finally, users will be engaged in the verification and the improvement of the database [53].

2.4.4 Authentication, access, and geolocation

In order to implement Suggestion (S1), it is possible to require users willing to access the database to create an account and authenticate systematically. Rate limits can also be implemented in order to control access to the website in terms of the number of requests made over a given period of time.

It is also possible to restrict access to the website and database to users within a given white list of countries, as per Suggestion (S3) made by the Ethics Committee. This can be accomplished by locating clients attemping to access the website either by HTML5 or IP geolocation, or a combination of both. HTML5 geolocation is achieved by requesting that the client provides their location via their browser. It is easy to implement and is precise, but it is also easy to spoof as a client may provide an incorrect location on purpose.

On the other hand, IP geolocation has been tested using different free services [176, 177]. The results cannot be falsified as easily as HTML5 geolocation, but they are also less precise: the city may be wrong but the country should be correct. However, it is still possible for clients to circumvent that issue by resorting to a proxy, in which case it is possible to couple the IP geolocation with a proxy check, which can be achieved by resorting to a specific service provider [178, 179].

2.4.5 Other security measures

Many other security measures have been researched and implemented for the proof of concept depite the fact that they have not been discussed with the Ethics

Committee.

It is possible to host the website and API onto a secure server, and enforcing secure communications with the client. This means that clients will be forced to navigate using the https prefix. Secure servers enable the bidirectional encryption of communications between the server and the clients, which aims at preventing man-in-the-middle attacks [180].

It is also possible to use secure software in order to host and manage the website, and access the database. This means using up-to-date LTS versions of the software programs used on the server including: the OS, node.js, and also any package used within node, including but not limited to cryptographic algorithms and mongoose [170], the package used to access MongoDB within node. These versions are more stable and better protected against known security faults.

In order to increase the security of personally identifiable information, hashed passwords of the researchers and participants' accounts can be stored. Account passwords will be hashed, using salts, prior to storage [181]. Hashing, unlike encryption, is designed to be a non-invertible operation. It is quick to compute a hashed value for any input, but it is extremely hard to retrieve the initial value from the hashed value, unless one compares the hashed value to all possible inputs. As a result, the initial value is basically lost, and yet the website can still make comparisons. This is perfectly well-suited to store passwords, whose true value does not need to be known, but which need to be compared to the input given by a user upon logging in [182]. This is a standard precaution meant to prevent someone who has had access to the database to also have access to the users' passwords. The matter of determining which cryptographic hashing algorithm is best used is subject to debate [183–186]. However, the bcrypt algorithm is generally the most recommended due to the fact that: it is very slow in comparison with other choices which makes it harder to attempt to invert; it can be made slower by changing its work factor parameter, in order to account for Moore's law [187]. Additionally, there exists a well-implemented node.js package for that algorithm [188].

While the names of the participants are stored, they are not disclosed to users and are referred to by their unique identifier. Participants are given the option -

prior or during the donation - to provide an e-mail address. This allows them, at any time, to have access to their participant account, to see what information is stored about them, and to request their information to be erased completely from the database. Not providing any contact information implies that the participant forsakes these rights, which is mentioned in the Participant Information Sheet in Appendix A.

In order to ensure that participant, user, and contributor accounts remain in control of their rightful owners, the website and API will provide state-of-the-art features such as device and location recording (which can be achieved by HTML5 or IP geolocation) and security settings change recording. However, if someone donates their fingerprints and then moves to a country that is blacklisted, their request to have their data wiped may be denied. In order to increase additional security, a 2-step verification process can be considered [189]. This and phone-based recovered methods can be implemented using text-messaging APIs [190]. Additionally, Google offers defense systems against bots [191], which they also use for data labeling purposes. Ideally, generalising this to the labeling and the verification of the fingerprint database would accomplish the same goal and provide a data verification mechanism.

Finally, it is important to keep track of account logins, database accesses, and suspicious activity in order to make sure research and participant accounts alike stay in the possession of their rightful owners. This can be done automatically by the website by 1) ensuring the consistency of the locations from which the database is accessed; and 2) asking for validation by email, or even by SMS, in case of inconsistency.

2.4.6 Image formats

The choice of a file format - or a list of acceptable file formats - that will be used to store the images is capital in terms of storage space and performance. There are two main categories of 2D image file formats: raster and vector images. The former stores the values of the colour pixel for each or some pixels, while the latter only

stores vector data in a formatted way. See [214] for a complete taxonomy of image file formats.

From the perspective of law enforcement databases, the WSQ format, which is named after Wavelet Scalar Quantization, has been used historically as a standard for the exchange and storage of greyscale fingerprint images [54]. This standard has shifted to that of the JPEG2000 format in many institutions worldwide [215–218], especially for 1000 ppi fingermarks [219]. From a research perspective, these conventions are not adopted by research databases, as shown in Table 2.2.

An alternative to the above conventions are vector images. Formats in this category tend to be much smaller due to the fact that they only store a fraction of the information while maintaining all important details. In fact, images are sharper and less heavy due to the fact that they are vectorized rather than rasterized. Furthermore, their size and clarity do not depend on their resolution. File size is not so much a concern for storage than for bandwidth purposes, low file sizes ensuring faster transfers. For these reasons, vector formats such as SVG have become extremely popular on the web [192].

Converting files from a raster to a vector format, however, is no easy task. There are online paid software programs that accomplish this [193, 194]. Refer to Figure 2.5 for a sample JPG fingerprint file and the resulting SVG compression done by Vector magic.

At first glance, the result is stunning: the ridge lines of the fingerprint are smooth and well-defined and the picture looks much clearer altogether. In terms of storage space, the input was 28.8Kb, the output 72.9Kb, and only 28.2Kb after a simple compression with `tar -zcvf`. This suggests that the SVG compression is also effective in terms of information stored, and the initial difference in size is an artifact due to the comparison of a compressed and an uncompressed format. However, a close inspection reveals that some areas of the initial fingerprint are simply absent from the output, and the 3rd-level details are approximately represented. Additionally, the resulting SVG paths cannot be uniquely associated to a ridge line - see Figure 2.5c where different paths are coloured differently. In summary, the conversion is lossy and the vector paths have little signification.

(a) Input JPG file. (b) Resulting SVG file. (c) Coloured SVG file.

Figure 2.5: Example of SVG compression of a fingerprint file using Vector Magic at a high quality setting [193]. Figure 2.5c has been coloured manually by assigning different colours to each SVG path. The computation is assumed to have been done by the server and the overall computation time seen by the client is around 10 seconds.

Aside from this, research has also been done to apply SVG conversion to comic pictures for compression purposes [55, 56], where the methods suggested are tailored to the nature of the pictures. This suggests that optimal results are due to tailoring the method to one purpose. As far as fingerprints are concerned, a suitable result would be a file that is smaller in size, such that every ridge line is represented as a single path, and where imperfections are represented as SVG masks, which has not yet been accomplished to the researcher's knowledge. For these reasons, there is merit to the study of vector file formats, or even the creation of a new specific vector file format for the representation of a fingerprint after the detection of its features. However, their usage applied to the storage of the raw fingerprint data is more debatable as the compression process is lossy.

Given the above considerations , vector file conversion is not suitable for fingerprint files. Instead, the lossless PNG format was chosen for this project; PNG files are handled by the back-end of the website using the ImageMagick software [195].

2.4.7 Storage of the image files

The first and foremost decision in terms of database organisation is the storage of images: is it better to store them in the database itself, or in a filesystem while keeping the filepath in the database? The former is colloquially called "to BLOB", which stands for Binary Large Object and refers to the name of the MySQL type which allows their storage.

Intuitively speaking, large fields of unknown length are detrimental to a database's performance and, conversely, filesystems are very well-suited for that task. As a result, it would make sense to store files in a filesystem if they are above a given size threshold [53]. Comparison tests have been performed between SQL server and the NTFS filesystem, and they confirmed that the former is significantly more efficient than the latter for files below 256Kb, and vice versa for those above 1Mb [57]. That efficiency has been measured in terms of read and write throughputs and resulting file fragmentation over usage time. Note that there is a general consensus on the fact that the downsides of the latter are significantly attenuated in modern Solid-State Drives (SSDs) [196, 58, 220].

Despite this, filesystem storage has the disadvantage of being prone to inconsistencies: the filesystem and the database go out of sync after long usage and garbage collection has to be dealt with. Additionally, this setup makes many tasks harder such as an OS change or a migration of the database [197]. Knowing that a single researcher is involved in the development of the database, the performance advantages of not BLOBing are not worth the effort.

2.5 Data Collection of Simulated Crime-Scene Fingermarks

2.5.1 Scope

In order to provide the database with initial content, promote the addition of quality data, and to meet Requirement (R2) regarding the relevance of the data collected to the fingerprint identification problem, it is necessary to design a protocol for acquiring such data. From a mathematical perspective, the images collected must be representative of the entire set of fingerprint images which can be encountered in a forensic context. As such, the focus of this data collection is the acquisition of finger impressions with varying appearances, which are representative of crime scene conditions as well as those in which reference fingerprints are collected.

The reasoning followed in order to establish this protocol aims consists in first investigating the different processes and factors involved in the acquisition of fingerprint images , and understand how they impact the appearance of the result. Those factors are:

- The finger: its ridge lines, minutiæ, and 3rd-level details.
- The matrix: the substance which is deposited (e.g. eccrine, sebum, blood).
- The substrate: the substance onto which the fingerprint is deposited.
- The deposition: the conditions in which the fingerprint is deposited.
- Degradation: the events that occur between the deposition and the development of the fingerprint which may interfere with the fingerprint (e.g. fingerprint deposited outdoors and subject to environmental factors; substrate has been burnt after deposition).
- The development method(s): the method(s) used in order to make the fingerprint visible. There is a wide variety of methods, and while the use of one over another mostly depends on the substrate, this choice is admittedly

based on experience and/or personal preference. See [59] for an in-depth overview, or Appendices E and F - based on [203] - for a more concise one.

- The acquisition method: method for acquiring an image file from a visible fingerprint (e.g. lifting tape and/or photography).

In crime scene conditions, different combinations of those factors result in very specific traits in fingerprints, see Table 1.2 for a non-exhaustive list. In order to reproduce these traits systematically in the scope of a data collection, let us classify deposition conditions according to the following categories:

- realistic conditions, where objects are handled without instructions or preparation.
- simulated conditions, where objects are prepared and handled with specific directions (e.g. items are cleaned, participants are given instructions on how to manipulate the items).
- ideal conditions, where the goal is to achieve the highest-quality fingerprints.

It is possible to limit the collection time and the number of variable parameters such as the deposition conditions, substrates, development and acquisition methods which need to be explored in the context of the data collection while still reproducing most of the artefacts observed in crime scene prints. This can be accomplished by following a scenario-based approach, where a scenario refers to a situation which participants will be requested to enact, and which require specific substrates and development methods. As such, this protocol can be seen as a generalisation of a previously established multi-surface fingerprint collection protocol [35]. See Table 2.3 for the list of scenarios which will be covered in the scope of this data collection. Finally, degradation conditions are a peripheral parameter and will not be dealt with since it would require a classification of such conditions, and a significantly larger number of samples.

2.5.2 Protocol

Given the volume of samples and the need for several development methods, it is not practical to have the fingerprints developed immediately. Therefore sufficient and appropriate storage space is required to preserve the fingerprints on the item before they are developed. It is also necessary to find a suitable means of transportation from the collection to the development and acquisition location (even if those are both on University campus).

In order to generate eccrine prints, the participants will wear latex gloves and be asked to undergo physical exertion (e.g. going up and down stairs for 2 minutes) prior to the deposition. For sebaceous prints, the participants will rub their forehead, scalp, and back of the ear or neck for 5 seconds each. This will only be done for ideal and simulated conditions. Prior to generating either eccrine or sebaceous prints, the participants will be instructed to wash their hands with soap for 30 seconds and then thoroughly dry their hands. This will ensure that no other substances will be transferred.

The experiments will be ordered in the following fashion: realistic conditions for every scenario first, followed by the generation of the matrix for simulated and ideal for every scenario. Additionally, any item involved in ideal or simulated conditions will be manipulated separately and with gloves. The items used for realistic and simulated will be the same, whereas for ideal conditions a flat substrate of a material that is similar to the items used in the realistic and simulated will be used instead. This is so that depletion prints may be taken under ideal conditions.

The forensic relevance of the scenarios can be justified by quoting statistics of the circumstances in which fingerprint evidence is found in crime scenes. However, such statistics are not publicly available, and would require cooperation with the Police National Computer (PNC) to acquire that data in the UK.Address Champod's remarks here: even if it only concerns a minority of cases, it is important to represent some cases provided that they produce image that are encountered in this environment, plain and simple.

Takeaway prints are not covered by this data collection, but as it is a type of

negative print, no new traits are expected to arise. Finally, if this experiment is conclusive with sebaceous and eccrine prints, another round of data collection will take place using blood. In that case, sheep blood can be used as a proxy for human blood.

The list of scenarios considered for this experiment is as follows, and is summarised in Table 2.3.

Scenario	Realistic conditions	Simulated conditions	Ideal conditions	Substrate	Development method
Livescan				Livescan	
Writing on paper				Paper	
Grabbing a receptacle				Plastic / glass / metal / ceramic	
Handling a garbage bag				Plastic	
Collecting a receipt				Thermal paper	
Swinging a tool				Wood	
Using tape				Tape	
Handling a knife				Metal	Gun blueing
Shoot with a gun and loading a gun				Metal	Gun blueing
Strangulating a victim				Human skin	

Table 2.3: Conditions covered for each scenario considered. The green cells indicate the most relevant scenarios for this research project.

Livescan

Substrate type(s), development method(s) and matrix(ces). Livescan device.

Ideal conditions. 10 flat and 10 rolled fingerprints.

Comments. Realistic and simulation conditions are not applicable.

Writing on paper

Substrate type(s), development method(s) and matrix(ces). A4 blank printing paper. Indanedione ZnCl (over D.F.O. and Ninhydrin). Eccrine (because it reacts with amino-acids).

Realistic conditions. Collect a a sample of their writing on a support no larger than an A4 sheet of paper (sticky notes and such are allowed).

Simulated conditions. Place their non-dominant on the paper, write the name of a celebrity without moving the non-dominant hand. Once this is done, ask them to leave their hand where it is, number the fingers and take a picture.

Ideal conditions. 5 depletion series on paper with all fingers using tapped fingerprints.

Grabbing a receptacle

Substrate type(s), development method(s) and matrix(ces). 2 ceramic mugs, 2 metallic cans, 2 plastic bottles, and 2 drinking glasses per participant. One set is used for realistic, and one for simulated conditions. Ceramic, metallic, plastic and glass surfaces for ideal conditions. They will be labelled and gridded in order for the depositions to have sufficient space and to be clearly identifiable. Powder will be used to develop fingerprints on ceramic and glass surfaces, and cyanoacrylate fuming for the plastic and metallic ones.

Realistic conditions. Each participant will grasp and simulate drinking from each receptacle with the hand(s) of their choice. If the receptacle needs opening, they will have to do it.

Simulated conditions. Each participant will be handed each receptacle that they need to grasp and simulate drinking on. The hand which they use will be written on a piece of tape and attached on the bottom of the receptacle. Pictures will be taken in order to identify each fingerprint.

Ideal conditions. 5 depletion series on each surface with all fingers using tapped fingerprints.

Handling a garbage bag

Substrate type(s), development method(s) and matrix(ces). Different rolls of bags for realistic and simulated ideal. Cyanoacrylate fuming for the realistic prints (because the location of the prints is not completely known), silver granulate powder for the simulated and ideal conditions for varying contrast. Eccrine prints (although sebaceous would also work because they both contain amino-acids).

Realistic conditions. Ask them to unfold, and grab a bag in order to put objects in it.

Simulated conditions. Garbage bags will be in foot-long squares and each will have right or left hand outlines. Each participant will be instructed to grab both squares, one after another, while keeping their fingers within the outlines.

Ideal conditions. Strips of garbage bags stretched over a hard surface. 5 depletion series with each finger using tapped fingerprints.

Handling a receipt

Substrate type(s), development method(s) and matrix(ces). Rolls of blank thermal paper, which will be manipulated with gloves, and from which the first outer layer has been removed. HCl / muriatic acid. Eccrine prints.

Realistic conditions. Ask each participant to hand in one of their own receipts from the past week.

Simulated conditions. Print a receipt and hand it to them. They are instructed to take it with one hand. Take a photograph then label the fingers.

Ideal conditions. Either print a blank strip or one with annotation for each finger. Have those strips / long receipts stretched over a hard surface. 5 depletion series using tapped fingerprints.

Comments. Buy rolls of thermal paper and test if fingerprints can be developed on it right away, or if printing is required.

Swinging a tool (e.g. hammer, shovel, baseball bat)

Substrate type(s), development method(s) and matrix(ces). Two hammers, to be used to hit stuffed pillowcases or cushions. A wooden board will be used for ideal conditions. Magnetic powder. Either sebaceous or eccrine.

Realistic conditions. Hand them the hammer without gloves and ask them to hit the pillows/cushions repeatedly at full strength for as long as they want.

Simulated conditions. Hand them the hammer with gloves, ask them to grasp the hammer with one hand and hit the pillows/cushions three times at full strength. The hand used will be written down and the process will be filmed in order to record a possible change in the position of the hand.

Ideal conditions. A wooden board delimited with duct tape and annotated with permanent marker will be used. 5 depletion series for each finger, using tapped fingerprints.

Handling a knife (as an offender, and/or a victim)

Comments. The ideal conditions are already covered by a previous experiment (grabbing a metal receptacle), both simulated and realistic versions of this scenario would be complicated to implement for both security and ethical reasons while not adding many of the traits that are reproduced by other experiments. For these reasons, this scenario will not be implemented.

Using tape on a victim

Substrate type(s), development method(s) and matrix(ces). Rolls of duct tape, use different rolls for realistic and simulated conditions. Sticky side powder will be used to develop the fingerprints on the adhesive side of the tape.

Realistic conditions. Manipulate the roll without gloves. Ask them to tear and apply one or several pieces of tape onto an object, within reason.

Simulated conditions. Hand them a piece of tape with gloves. Ask them to tear it in half without touching other parts of the tape. Photograph the position of their fingers and label the hands and fingers used on the sticky side of the tape.

Ideal conditions. Have strips of tape spread onto a rigid surface sticky face up. 5 depletion series with all fingers, flat fingerprints.

Shoot with a gun and / or loading a gun.

Comments. This scenario is forensically relevant but is unlikely to produce fingerprints that are significantly different from that produced by previous experiments (namely the one on metallic cans and with hammers). Additionally, this experiment could be made possible with a partnership with a nearby shooting range but would be impractical because of the limited range of development methods (they have to be non-destructive).

Strangulating a victim

Comments. Interesting for the sake of completion, but will not be performed since the most effective development method on skin is iodine fuming, which is illegal and not practiced in the UK for health reasons.

2.5.3 Data Collection

A student project was organised for a 4^{th} year B.Sc. student in Mathematics. The purpose of her project was to perform a data collection, annotate the resulting fingerprints for the presence of traits mentioned in Table 1.2, then perform a statistical analysis in order to prove or disprove that the above protocol does produce different fingermarks with regards to the traits they show.

The student was given a short training in Forensic Science and fingerprint evidence, and was also trained in how to generate, develop, and collect fingerprints,

which was organised and provided by Jessica Lam, Ph.D. student at the University of Leicester under the Intrepid Forensics Programme at the time of writing.

The student booked 4 afternoons in the Criminology department meeting room to host the data collections on her own. She was supplied with all of the necessary equipment for the data collection. After she completed the data collection and analyses, she was asked to provide the results to me so that I could incorporate the data into this project. Unfortunately, she failed to make the data accessible and ultimately stopped responding to any e-mails after her module was completed. As such, it is unknown how many participants took part, how many fingerprints were collected, and what the results of her data collection were.

Chapter 3

A Formalisation of Fingerprint Identification using Source Probabilities

Historically, fingerprint conclusions provided by fingerprint examiners to the court were statements of absolute certitude. This certitude relied on the experience of the examiners, and on the assumptions that the finger impressions analysed presented unique patterns, and that this uniqueness made the impression distinguishable from any other impression. These assumptions have also been applied to other discplines in forensic science, such as firearm examination, footware examination, and handwriting examination to name a few, and are referred to as the theory of discernible uniqueness [60, 61].

The scientific community has shown, however, that there is insufficient basis for these claims [221], because: it cannot be empirically demonstrated that fingerprints are unique without conducting a comparison of every existing fingerprint; and the fact that the uniqueness of a pattern in a fingerprint does not entail the ability to distinguish between all of the impressions it creates, due to the large amount of confounding factors involved [61].

In order to address these shortcomings, the scientific community has encouraged the use of likelihood ratios, which represent how much more likely it is to observe a

set evidence E given a hypothesis C than it is to observe the same evidence given the negation of this hypothesis

$$\frac{\mathbb{P}(E|C)}{\mathbb{P}(E|\bar{C})}, \tag{3.1}$$

and which play a pivotal in the computation of posterior odds in the Bayesian framework [62].

As such, research has been devoted to establish probabilistic evaluations of the evidentiary value of fingerprint comparisons [63]. In addition to this, the Association of Forensic Science Providers (AFSP) has proposed a verbal scale which matches different ranges of likelihood values, such as $[1, 10]$ or $[10'000 - 1'000'000]$, to verbal expressions which can be understood by laypeople, such as "weak or limited" and "very strong" respectively [64]. This laid the foundation which makes it possible for forensic scientists to substitude previous statements such as

> "It is moderately probable, highly probable, or practically certain that two items have a common source."

by more appropriate statements, such as

> "It is far more probable that this degree of similarity would occur when comparing the latent print with the defendant's fingers than with someone else's fingers."

which reflects the evidentiary value of the expert's analysis more accurately [61].

It is worth noting, however, that the usage of the expression "someone else" in the suggested statement, which remains vague. Let us consider a population P from which all fingerprints are known, and are being compared. For any given pair of impressions from that population, it requires more detail or features in common, or more similarity between these impressions, in order to be able to uniquely identify them. Determining the amount of similitude required in order to make a comparison is referred to as the sufficiency problem, which is actively researched [65]. As such it become obvious that the evidentiary value of a fingerprint comparison depends on - and may therefore be tuned to - the population to which the unknown mark is implicitly compared.

The purpose of this chapter is not to implement upon the existing framework, but instead to lay the mathematical foundation required to discuss fingerprint identification algorithms in Chapter 4. However, in the process of doing so, it provides a mathematical basis for the consideration of the probability of two fingerprint impressions coming from the same source, as opposed to merely comparing likeness between impressions. These probabilities, which are referred to as *source probabilities* [61], have been criticised because current analyses do not support a conclusion that allow a source probability to be computed, and it was thus concluded by the forensic community that their use is fallacious. This thesis, however, demonstrates that calculating source probabilities can in fact be done with sound mathematical reasoning under appropriate circumstances.

By doing so, this section proposes a rigorous approach to fingerprint identification, formulates conditions which ensure that identifications are performed accurately, and also makes suggestions as to how hypothetical populations can be used in the context of the computation of source probabilities in order to provide more specific conclusions to the court. As such, Section 3.1 will introduce mathematical definitions of the concepts at hand; then, a mathematical formulation of the identification problem is suggested in Section 3.2, under the assumption that the database contains a fingermark that comes from the same source as the unknown fingermark; finally, the identification problem is formulated without that assumption in Section 3.3.

3.1 Defining fingermarks, sources, and depositions

The purpose of this section is to define mathematically the notions of fingermark, source, and deposition in order to understand the formal relations between them and ultimately provide a mathematical formulation of the identification problem. Throughout this chapter, a distinction will be made between: a fingerprint, which refers to the part of the finger which is formed of ridges and furrows; and a fingermark, which refers to the deposition left when a fingerprint is put in contact with a surface. Additionally, the term “source”, which is the abstract term used in a

legal context in order to refer to the fact that a fingermark has been left by a given individual under certain circumstances, will be used to refer to all the existing information pertaining to a given fingerprint.

A fingermark deposited on a surface is the result of a source having been in contact with that surface. This source leaves a trace that is deformed depending on a variety of factors, including properties of the surface, the fingerprints themselves, and the conditions in which contact was made. A deposition can be defined as a function which maps a source to a fingermark. As such, if the set of sources is denoted by S, and the set of fingermarks by F, the set of depositions Δ is a subset of the set of functions from S to F,

$$\Delta \subseteq \mathcal{F}(S, F), \tag{3.2}$$

with constraints that will be introduced as part of this section. For technical reasons, the sets S and F will respectively be endowed with σ-algebras $\mathcal{B}$ and $\mathcal{C}$, so that $(S, \mathcal{B})$ and $(F, \mathcal{C})$ are both measurable spaces. Additionally, the set of depositions Δ considered is a subset of the set of measurable functions from S to F.

Definition 3.1:
It is said that a fingermark $f \in F$ can be assessed as coming from a source $s \in S$, which will be denoted $f \rightsquigarrow s$, if there exists a deposition which maps s to f

$$\exists \delta \in \Delta \ \mid \ \delta(s) = f. \tag{3.3}$$

Definition 3.2:
Similarly, it will be said that two fingermarks $f, f' \in F$ can be assessed as coming from the same source, which will be denoted $f \approx f'$, if both fingermarks have been deposited by the same source

$$f \rightsquigarrow s \text{ and } f' \rightsquigarrow s, \tag{3.4}$$

or equivalently,

$$\exists s \in S, \ \exists (\delta, \delta') \in \Delta^2 \ \mid \ \delta(s) = f \text{ and } \delta'(s) = f'. \tag{3.5}$$

Assumption 3.1:
Given that all fingermarks originate from a source, it is legitimate to assume that

each fingermark can be assessed from coming from at least one source. This means that the sets F, S and Δ are such that

$$\forall f \in F,\ \exists (s, \delta) \in S \times \Delta \ \mid \ \delta(s) = f. \tag{3.6}$$

However, as mentioned in the introduction, there may be such distortion during the deposition of a fingermark in a practical scenario that it is indistinguishable from another one, which comes from a different source. There, the assumption that each fingermark can only be assessed as coming from a single source is **not** made. This emphasizes that there is a strict difference between the fact that two fingermarks are said to come from the same source as a result of an assessment, which has been defined above by the relation $\approx$, or from available ground-truth knowledge, which will be denoted by $f \sim f'$.

By definition, the relation $\approx$ is: binary (it takes two arguments); reflexive (each fingermark can be assessed as coming from the same source as itself); and symmetric (if f can be assessed as coming from the same source as f', then f' can be assessed as coming from the same source as f). If this relation is also transitive, then this relation is an equivalence relation. For a given equivalence relation $\sim$ on E, it is possible to consider its equivalence classes, which are defined as

$$[y] = \{x \in E \,|\, x \sim y\}\,. \tag{3.7}$$

Additionally, it has been demonstrated that the set of all such equivalence classes for $\sim$ defines a partition on E, which means that any element of E belongs to a unique equivalence class. As such, it is possible to gain a greater understanding of a set by identifying an equivalence relation and studying its equivalence classes. Unfortunately, the relation $\approx$ is not transitive on F as a given fingermark can be assessed as coming from two different sources.

Definition 3.3:
Let us now consider the subset F_u of F, the set of fingermarks that originate from a unique source, which is defined as such:

$$F_u := \{f \in F \ \mid \ \exists! s \in S,\ \exists \delta \in \Delta \text{ such that } \delta(s) = f\}\,. \tag{3.8}$$

The purpose of studying fingermarks from unique sources is not to make a statement or enquiry as to whether fingerprints are unique or not, but rather to study the implications of that definition on the relation $\approx$. Furthermore, let us note that the nature of F_u depends on that of the set of sources S and of depositions Δ considered, in that F_u can be interpreted as the set of fingermarks that can be distinguished from any other fingermarks given the set of sources and depositions considered.

Property 3.1:
The relation $\approx$ is transitive on F_u.

Proof. Given three fingermarks of unique source $(f, f', f'') \in F_u^3$, we have

$$\begin{aligned}
&\begin{cases} f \approx f' \\ f' \approx f'' \end{cases} \Leftrightarrow \exists!(s_1, s_2) \in S^2,\ \exists(\delta_1, \delta_1', \delta_2, \delta_2') \in \Delta^4 \ \mid \ \begin{cases} f = \delta_1(s_1) \\ f' = \delta_1'(s_1) = \delta_2(s_2) \\ f'' = \delta_2'(s_2) \end{cases} \\
&\Rightarrow \exists! s \in S,\ \exists(\delta_1, \delta_2') \in \Delta^2 \ \mid \ \begin{cases} f = \delta_1(s) \\ f'' = \delta_2'(s) \end{cases} \\
&\Rightarrow f \approx f''.
\end{aligned} \tag{3.9}$$

□

Despite the fact that this result shows that the relation $\approx$ is an equivalence relation on F_u, it does not give much insight into identification in a practical context. That is the case because there is no a priori knowledge as to whether a fingermark can be assessed as coming from a unique source, which amounts to stating that it is distinguishable from any other fingermark. That is why it is necessary to introduce the concept of a reference fingerprint, which refers to a fingerprint which has been collected in the best possible conditions, and which can be used as a reference for the identification of a lower-quality crime-scene print. In order to do so, let us first define the set of ideal depositions.

Definition 3.4:
A deposition function $\delta \in \Delta$ is said to be ideal if it meets the following conditions:

- δ is injective, which means that depositing from different sources via δ produces different fingermarks;
- any fingermark which can be assessed as coming from a source via δ actually comes from that source

$$\forall s \in S,\ \delta(s) \sim s. \tag{3.10}$$

Assumption 3.2:

As a mean of formulating the fact that information regarding the source, such as features, may only be lost in the deposition process, the assumption is made that the set of deformations Δ is such that non-ideal depositions cannot produce the same fingermarks as those deposited by ideal depositions,

$$\forall (d, d') \in \Delta^* \times (\Delta \backslash \Delta^*),\ \text{Im}(\delta) \cap \text{Im}(\delta') = \varnothing. \tag{3.11}$$

Property 3.2:

Any fingermark which has been deposited through an ideal deposition can be assessed as coming from a unique source

$$\forall \delta \in \Delta^*,\ \text{Im}(\delta) \subset F_u. \tag{3.12}$$

Proof. Let $\delta \in \Delta^*$ be an ideal deposition, and $f \in \text{Im}(\delta)$ a fingermark which has been deposited through δ. δ is injective, and is therefore invertible as a function from F to $\text{Im}(\delta)$. Consequently,

$$\exists! s \in S,\ \delta(s) = f. \tag{3.13}$$

Given Equation (3.10), for any other ideal deposition $\delta' \in \Delta^*$ such that $f \in \text{Im}(\delta')$, we have $d'^{-1}(f) \sim f$ and $f \sim s$. Due to the uniqueness of the ground-truth source of any fingermark, we have $d'^{-1}(f) = s$. This means that any fingermark which has been deposited through an ideal deposition can only be assessed to a unique source through any ideal deposition. Finally, given Equation (3.11), there is no non-ideal deposition d' which is such that $d'(s) = f$. Consequently, $f \in F_u$. □

Definition 3.5:

The set of reference fingerprint F_r is the subset of fingermarks which has been deposited through ideal depositions

$$F_r := \underset{\delta \in \Delta^*}{\cup} \text{Im}(\delta). \tag{3.14}$$

Property 3.3:
Any reference fingermark can only be assessed to a unique source

$$F_r \subset F_u. \tag{3.15}$$

As a result, F_r defines a set of fingermark on which tangible assumptions have been made, such that all fingermarks it contains also come from a unique source. As a result, the relation $\sim$ is an equivalence relation on this set.

Definition 3.6:
For any source $s \in S$, the set of fingermarks which can be assessed as coming from source s is denoted by $[s]$, and defined as the following:

$$[s] := \{f \in F \mid f \rightsquigarrow s\}. \tag{3.16}$$

Theorem 3.1

Let f_u be a fingermark which can be assessed as coming from a unique source s. The equivalence class of f_u by the relation $\approx$ is given by $[s]$,

$$[s] = [f_u]. \tag{3.17}$$

Proof. For any fingermark f in $[f_u]$, there exists a source s' such that both f and f_u can be assessed as coming from that source. Given that f_u comes from the unique source s, then $s' = s$, which means that f can be assessed as coming from source s. Therefore $\forall f \in F,\ f \in [f_u] \Leftrightarrow f \in [s]$. □

Theorem 3.1 means that studying the fingermarks that come from a given source s can be achieved by studying the equivalence class of any fingermark of unique source s. Given Property 3.3, this result also holds for any reference fingermark. As a result, it has been demonstrated that: a) any reference fingermark can be used to uniquely identify other fingermarks to the same source; b) non-reference fingermarks may or may not be sufficient to perform a unique identification; and c) fingermarks that come from a non-unique source are insufficient to perform an identification.

In conclusion, a mathematical framework has been established in order to cover a variety of situations in terms of the choice of sets F, S, and Δ considered. It is now

possible to study varied mathematical formulations of the identification problem, and the properties of identification algorithms that learn from data. Regarding the Forensic Science field, the premise of this framework is to allow us to analyse different working environments, and eventually deduce which statements can and cannot be made with respect to fingerprint identification. It clearly outlines the hypotheses that are made on the fingermarks in order for some conclusions to be reached. These assumptions should be under consideration at all times in order for scientific statements to stay accurate for court purposes.

3.2 Source determination

Let us now consider the practical issue of identifying an unknown fingermark, denoted f_c (named after *crime scene*), with respect to an existing database of fingermarks $F_{\text{DB}} \subsetneq F$. This section tackles this problem under the following assumption.

Assumption 3.3:
There exists a reference fingermark f_m (named after *match*) in F_{DB} which comes from the same source as f_c,

$$\exists f_m \in F_{\text{DB}} \cap F_r \;\mid\; f_c \in [f_m]. \tag{3.18}$$

As a result, the problem of identifying the unknown fingermark f_c corresponds to determining which reference fingermark f_r in the database is such that $f_c \in [f_r]$. Practically speaking, there is limited knowledge about the set of sources S. The objective is to assess whether an unknown and a reference fingermark come from the same source solely based on the relation between both fingermarks, and the ground truth knowledge that the reference fingermark comes from a given source. Because of this, it is necessary to be able to map a reference fingermark to another fingermark, which is the purpose of the following definition.

Definition 3.7:
The set of modification functions $M \subset \mathcal{F}(F_r, F)$ is defined as

$$M := \left\{ m \in \mathcal{F}(F_r, F) \;\middle|\; \exists (\delta, \delta') \in \Delta^* \times \Delta \text{ such that } m = \delta' \circ \delta^{-1} \right\}. \tag{3.19}$$

Assumption 3.4:

The sets of depositions Δ and of modifications M are such that

$$\forall (m, \delta) \in M \times \Delta^*,\ m \circ \delta \in \Delta, \tag{3.20}$$

Theorem 3.2:

For any reference fingermark in the database $f_r \in F_{\text{DB}} \cap F_r$, f_r can be assessed as coming from the same source as the unknown print f_c only if there exists a modification $m \in M$ which maps f_c to f_r

$$f_c \in [f_r] \Leftrightarrow \exists m \in M \ \mid \ f_c = m(f_r). \tag{3.21}$$

Proof. Sufficiency.

$$\begin{aligned} f_c \in [f_r] &\Leftrightarrow \exists (s, \delta, \delta') \in S \times \Delta^* \times \Delta \ \mid \ \delta(s) = f_r \text{ and } \delta'(s) = f_c \\ &\Rightarrow \exists (\delta, \delta') \in \Delta^* \times \Delta \ \mid \ f_c = \left(\delta' \circ \delta^{-1}\right)(f_r) \\ &\Rightarrow \exists m \in M \ \mid \ f_c = m(f_r). \end{aligned} \tag{3.22}$$

Necessity.

$$\begin{aligned} \exists m \in M \ \mid \ f_c = m(f_r) \Rightarrow \exists (m, s, \delta) \in M \times S \times \Delta^* \ \mid \ &\begin{cases} f_c = m(f_r) \\ f_r = \delta(s) \end{cases} \\ \Rightarrow \exists (m, s, \delta) \in M \times S \times \Delta^* \ \mid \ &\begin{cases} f_c = m \circ \delta(s). \\ f_r = \delta(s) \end{cases} \end{aligned} \tag{3.23}$$

Assumption 3.4 grants that $m \circ \delta \in \Delta$, therefore $f_c \in [f_r]$. □

This result signifies that determining the reference fingermark f_m which comes from the same source as f_c can be achieved by establishing which reference fingermark f_r in the database is such that there exists a modification $m \in M$ that verifies $m(f_r) = f_c$. Next, the following concepts aim at formulating the identification problem in such a way that it can be solved numerically.

Assumption 3.5:

There exists a semi-metric d on the set of fingermarks F, namely a function from $F \times F$ to $\mathbb{R}^+$ such that, for any $(f, f') \in F^2$, $d(f, f') = 0 \Leftrightarrow f = f'$ and $d(f, f') = d(f', f')$.

Assumption 3.6:

The set of modifications M is a closed bounded set.

Definition 3.8:

The minimal distance after modification function, which will be denoted d_M, refers to the function which maps a pair (f_c, f_r) of an unknown fingermark, and a reference fingermark in F_{DB} to the minimal distance between f_c and a modification of f_r

$$\begin{aligned} d_M : \quad (f_c, f_r) &\mapsto \min_{m \in M} d\big(f_c, m(f_r)\big). \\ F \times F_r &\to \qquad\quad \mathbb{R}^+ \end{aligned} \tag{3.24}$$

The existence of this minimum on M is guaranteed by Assumption 3.6.

Theorem 3.3

For any unknown fingermark $f_c \in F$, the set $\widehat{[f_c]}$ defined as

$$\widehat{[f_c]} := \underset{f_r \in F_{\text{DB}} \cap F_r}{\operatorname{argmin}} \Big[\min_{m \in M} d\big(f_c, m(f_r)\big) \Big] \tag{3.25}$$

contains only and all reference fingermarks in the database which can be assessed as coming from the same source as the unknown fingermark f_c. In other words, for any reference fingermark in the database $f_r \in F_{\text{DB}} \cap F_r$,

$$f_c \in [f_r] \Leftrightarrow f_r \in \widehat{[f_c]}. \tag{3.26}$$

Proof. Sufficiency. As per Theorem 3.2,

$$\begin{aligned} f_c \in [f_r] &\Leftrightarrow \exists m \in M \;\mid\; d\big(f_c, m(f_r)\big) = 0 \\ &\Rightarrow d(f_c, m(f_r)) = \min_{\substack{f' \in F_{\text{DB}} \cap F_r \\ m' \in M}} d\big(f_c, m'(f')\big) \\ &\Rightarrow f_r \in \widehat{[f_c]}. \end{aligned} \tag{3.27}$$

Necessity. Given Assumption 3.3,

$$\begin{aligned} f_r \in \widehat{[f_c]} &\Rightarrow \forall m' \in M, \forall f' \in F_{\text{DB}} \cap F_r, \; \min_{m \in M} d\big(f_c, m(f_r)\big) \leqslant d\big(f_c, m'(f')\big) \\ &\Rightarrow \min_{m \in M} d\big(f_c, m(f_r)\big) = 0 \\ &\Rightarrow \exists m \in M \;\mid\; f_c = m(f_r). \end{aligned} \tag{3.28}$$

□

This theorem effectively proves that, under the assumption that there is a reference fingermark of the same source as the unknown fingermark in the database, it is possible to formulate the identification problem in F_{DB} as an optimisation problem on $(F_{\text{DB}} \cap F_r) \times M$. Specific details will be provided as to how this optimisation problem has been solved during this research: see Chapter 4 for the parameterisation of the set F and M, more specifically Section 4.4.2 for the computation of a distance d on F, and finally Chapter 5 for the specific optimisation algorithm used. The remainder of this section is an analysis of the performance of a theoretical identification system which solves the optimisation problem as specified by Equation (3.25).

In a computational setting, numerical error needs to be accounted for. Consequently, the minimum distance after modification between f_c and a fingermark f_r within $\widehat{[f_c]}$ would most likely not be exactly equal to zero, but would instead be in a close neighbourhood of it. As a result, in order for an identification system to work adequately, it is necessary to define criteria on the value $d_M(f_c, f_r)$ obtained which specify when an identification is said to be accepted, rejected, or inconclusive. Similarly, it is also necessary to define criteria which evaluate the performance of this identification system in order to ensure that it performs adequately on the dataset ground-truth dataset F_{DB}. In order to accomplish this, the following assumption and definitions are needed.

Assumption 3.7:
It is assumed that the database of fingermarks F_{DB} is such that there is ground-truth information regarding the source of every fingermark within F_{DB}. This means that, for any pair of fingermarks (f, f') within the database F_{DB}, it is known whether the statement $f \sim f'$ is true or false.

Definition 3.9 (Identification error)**:**
The identification error $e(f_c)$ committed by an identification system on a non-reference fingermark f_c corresponds to the maximum minimum distance after modification between f_c and any reference fingermark within F_{DB} which is known

to come from the same source as f_c. Therefore, the function e is defined as such

$$\begin{array}{rccc} e: & f_c & \mapsto & \max\limits_{\substack{f_r \in F_{\text{DB}} \cap F_r \\ f_r \sim f_c}} d_M(f_c, f_r). \\ & F_{\text{DB}} & \to & \mathbb{R}^+ \end{array} \tag{3.29}$$

Furthermore, the identification error committed by the identification system on the entire database, which will also be denoted by $e(F_{\text{DB}})$ corresponds to the maximum identification error associated to any non-reference fingermark f_c within F_{DB}

$$e(F_{\text{DB}}) := \max_{f_c \in F_{\text{DB}}} e(f_c). \tag{3.30}$$

Definition 3.10 (Free radius)**:**
Similarly, the free radius r associated with a non-reference fingermark f_c in F_{DB} corresponds to the minimum minimum distance after modification between f_c and any reference fingermark within F_{DB} which is known to come from a different source as f_c. The function r is defined as

$$\begin{array}{rccc} r: & f_c & \mapsto & \min\limits_{\substack{f_r \in F_{\text{DB}} \cap F_r \\ f_r \not\sim f_c}} d_M(f_c, f_r). \\ & F_{\text{DB}} & \to & \mathbb{R}^+ \end{array} \tag{3.31}$$

Additionally, the free radius associated to the identification system on the entire database F_{DB}, which will be denoted by $r(F_{\text{DB}})$, refers to the minimum free radius associated to any non-reference fingermark within F_{DB}

$$r(F_{\text{DB}}) := \min_{f_c \in F_{\text{DB}}} r(f_c). \tag{3.32}$$

With these notions defined, it becomes possible to define a decision-making process for a computational identification system. It is suggested that, given an unknown fingermark $f_c \in {F_{\text{DB}}}^{\complement}$, the set of potential candidates $\widehat{[f_c]}$ be computed as per Theorem 3.3. Then, for any fingermark f_r in $\widehat{[f_c]}$, the identification is said to be

$$\begin{cases} \text{accepted if } d_M(f_c, f_r) \leqslant e(F_{\text{DB}}), \\ \text{rejected if } d_M(f_c, f_r) \geqslant r(F_{\text{DB}}), \\ \text{inconclusive otherwise.} \end{cases} \tag{3.33}$$

This decision-making process is valid only if it correctly accepts all fingermarks within F_{DB} which are known to come from the same source, and correctly rejects all fingermarks within F_{DB} which are known to come from different sources.

Theorem 3.4: Validity of an identification system

If the identification system meets the following validity condition on F_{DB}

$$e(F_{\text{DB}}) < r(F_{\text{DB}}), \tag{3.34}$$

then all fingermarks in F_{DB} are properly identified. This means that there are no false positives and no false negatives in the identification of non-reference fingermarks in F_{DB} with respect to their ground-truth source, $\forall (f_c, f_r) \in F_{\text{DB}} \times (F_{\text{DB}} \cap F_r)$,

$$d_M(f_c, f_r) \leqslant e(F_{\text{DB}}) \Rightarrow f_c \approx f_r, \tag{3.35}$$

$$d_M(f_c, f_r) \geqslant r(F_{\text{DB}}) \Rightarrow f_c \not\approx f_r. \tag{3.36}$$

Proof. Let us first prove the contrapositive of (3.34)⇒(3.35). Assuming that the negation of Equation (3.35) is true, we have

$$\exists (f_c, f_r) \in F_{\text{DB}} \times (F_{\text{DB}} \cap F_r) \mid d_M(f_c, f_r) \leqslant e(F_{\text{DB}}) \text{ and } f_c \not\approx f_r. \tag{3.37}$$

This entails that $r(F_{\text{DB}}) \leqslant d_M(f_c, f_r) \leqslant e(F_{\text{DB}})$, which means that the negation of Equation (3.34) is true. The contrapositive of (3.34)⇒(3.36) can be also demonstrated in a similar fashion. Assuming that

$$\exists (f_c, f_r) \in F_{\text{DB}} \times (F_{\text{DB}} \cap F_r) \mid d_M(f_c, f_r) \geqslant r(F_{\text{DB}}) \text{ and } f_c \approx f_r, \tag{3.38}$$

we have $e(F_{\text{DB}}) \geqslant d_M(f_c, f_r) \geqslant r(F_{\text{DB}})$. □

This theorem provides a condition that ensures that the identification system is valid based on a database with ground-truth information about the sources of its fingermarks.

3.3 Source assessment

In this section, the problem of identifying an unknown fingermark with a reference database F_{DB} will be addressed in a similar fashion as in Section 3.2, but without Assumption 3.3 that there is indeed a reference fingermark F_{DB} which comes from the same source as the unknown fingermark. In other words, no prior

assumption is made and no information is given regarding the source of f_c, which therefore may or may not have a matching reference fingermark within F_{DB}.

Relaxing Assumption 3.3 invalidates the necessity part of Theorem 3.3, which means that the only result available from Section 3.2 is that, for any fingermark $f_r \in F_{\text{DB}} \cap F_r$,

$$f_c \in [f_r] \Rightarrow f_r \in \widehat{[f_c]}. \tag{3.39}$$

As a result, it is only possible to exclude reference fingermarks f_r in the database from coming from the same source as the unknown mark, by checking that $f_r \notin \widehat{[f_c]}$. Given that the converse of Equation (3.39) is untrue, computing $\widehat{[f_c]}$ may yield reference fingermarks f_r which do not come from the same source as f_c. This is due to the fact that the minimum distance after modification from one such reference fingermark f_r to a modification of f_c, $d_M(f_c, f_r)$ may or may not be equal to zero.

Despite this, alternative ways to provide an identification can be formulated. The first approach suggested here consists in applying the same method as the one suggested in Section 3.2 and specified by Equation (3.33). Whether the error is of numerical nature, or resides from the fact that there is no certainty as to whether F_{DB} holds a reference fingermark which comes from the same source as the unknown fingermark, it is possible to gauge whether that error is acceptable or not by comparing it to the identification error and the free radius of the system. These values are based on the limited knowledge of the entire set of fingermarks F which is provided by F_{DB}.

Alternatively, the second approach consists in considering a hypothetical population Π to which the source of the unknown fingermark is suspected or likely to belong. This is especially pertinent in investigations where intelligence information or evidence suggests that the perpetrator has certain characteristics that can narrow the list of possible sources. The point of this method is to effectively extend the previous one, whereby the F_{DB} is known to contain a match for the unknown fingermark, to a situation where the population Π is likely to belong. This can be achieved by formalising Π as a random variable on the set of sources S, and by determining a way to extend the premise of identification error and free radius to this situation.

Remark 3.1 (Random depositions)**:**
Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space. The set of depositions Δ can be endowed with the σ-algebra $\mathcal{D} := \sigma(\pi_s, s \in S)$ where

$$\begin{array}{rccc} \pi_s : & \delta & \mapsto & \delta(s). \\ & \Delta & \to & F \end{array} \tag{3.40}$$

The measurable space $(\Delta, \mathcal{D})$ is such that families of random variables $\delta = (\delta_s, s \in S)$ are $\mathcal{A} - \mathcal{D}$ measurable, which therefore define random variables on $(\Delta, \mathcal{D})$ [67]. Such random variables δ will be referred to as random depositions.

Remark 3.2:
Any pair (Π, δ) where Π is a random variable on S and δ a random deposition induces a random variable on F, which is denoted $F_{\Pi,\delta}$, and which is defined as

$$\begin{array}{rl} F_{\Pi,\delta} : & \omega \mapsto \delta(\omega)\big(\Pi(\omega)\big). \\ & \Omega \to F \end{array} \tag{3.41}$$

Given these remarks, three different decision-making possibilities are suggested. These choices involve different computations that may or may not be feasible depending on the computational cost incurred. The first consists in considering the identification error and free radius associated to $F_{\Pi,\delta}(\omega)$, for a given $\omega \in \Omega$. Instead of using a ground-truth database as a reference, the threshold used to perform the decision-making are based upon that of a simulated database associated to a hypothetical population, whose parameters are based upon the available knowledge of the entire set of fingermarks through F_{DB} and that of the set of depositions through the random variable δ on Δ.

The second one consists in considering not one, but several simulated datasets $F_{\Pi,\delta}(\omega)$ in order to compute the expected identification error and free radius associated to the random population Π. They can be respectively written as

$$\mathbb{E}\left[e(F_{\Pi,\delta})\right] := \mathbb{E}\left[\max_{\substack{(f_r,f_c)\in(F_{\Pi,\delta}\cap F_r)\times F_{\Pi,\delta} \\ f_c \sim f_r}} d_M(f_c, f_r)\right], \tag{3.42}$$

$$\mathbb{E}\left[r(F_{\Pi,\delta})\right] := \mathbb{E}\left[\min_{\substack{(f_r,f_c)\in(F_{\Pi,\delta}\cap F_r)\times F_{\Pi,\delta} \\ f_c \not\sim f_r}} d_M(f_c, f_r)\right]. \tag{3.43}$$

Finally, a the third approach consists in computing the probability of a fingermark $f_r \in \widehat{[f_c]}$ of coming from the same source as the unknown fingermark f_c in a population Π as such

$$\mathbb{P}\left[d_M(f_r, f_c) \leqslant e(F_{\Pi,\delta})\right], \tag{3.44}$$

and the probability of the fingermark f_r being rejected for identification within a population Π

$$\mathbb{P}\left[d_M(f_r, f_c) \geqslant r(F_{\Pi,\delta})\right] \tag{3.45}$$

Similarly to Theorem 3.4, this approach stays consistent provided that the random identification error is lesser than the random free radius almost surely

$$e(F_{\Pi,\delta}) < r(F_{\Pi,\delta}) \text{ a.s.} \tag{3.46}$$

The three above approaches to the production of a statement as to whether two fingermarks come from the same source differ in terms of the computational cost incurred, which depends on the means by which random fingermark datasets $F_{\Pi,\delta}$ are generated. Such methods are discussed in Section 4.4.3. Additionally, the first two approaches provide a quantification of the reasonable doubt associated to a fingerprint identification by means of the expected identification error, in a similar fashion as in the source determination setting; while the third one computes probabilities of two fingermarks coming from the same source, which is on par with current probabilistic approaches.

All in all, the setting described here addresses the fact that identification statements can be produced in the absence of Assumption 3.3. This was accomplished by considering a population Π, to which the source of the unknown mark is suspected to belong, and which is of forensic relevance. It was also shown that it is possible to produce different statements, either based on a quantification of reasonable doubt, or based on a probabilities, which which all pertain to the population Π considered.

Chapter 4

Machine Learning applied to Source Probability Computation or Ridge Line Modelling for Source Probability Computation

The purpose of the remainder of this thesis is to provide the basis for an identification system which is equipped with the tools required to solve the identification problem as it is formulated in Chapter 3, via the computation of source probabilities.

This chapter more specifically discusses the requirements needed to perform both the source determination and source assessment, which are respectively described in Sections 3.2 and 3.3. Although these concepts are both formulations of the identification problem, which refers to the identification of an unknown fingermark using a database, source determination assumes that there exists fingermarks coming from the same source in the database, whereas source assessment does not. As such, source determination requires the definition and implementation of a set of fingermarks F, a semi-metric d on F, and a set of modifications M.In addition to these requirements, it is also necessary for source assessments to have computational tools available to generate random variables on

the set of fingermarks F.

In order for the fingerprint identification system described here to fulfill the Daubert standard [116, 68], and so that it can be properly utilised by the court for forensic purposes, this system should meet the following requirements:

(R1) perform adequately on a training dataset of degraded fingermarks, and have a quantified error rate;

(R2) perform consistently as the database grows larger;

(R3) have a rationale that is understandable by a layperson;

(R4) be verifiable by a fingerprint examiner.

The ability of the algorithm to perform well with varied input, which is implied by Requirements (R1) and (R2), is particularly challenging to achieve in fingerprint identification due to: a) the large amount of intra-class variability in fingermarks [30], which is exacerbated by the large amount of deformation to which crime scene fingermarks are subjected; and b) the fact that a large population of individuals must be discriminated. Let us note that the second issue mentioned, which refers to the decrease of accuracy of fingerprint identification in growing databases, is a known pitfall in any identification technique in Forensic Science, and especially in DNA identification [69].

Achieving these objectives requires sophisticated algorithms which are tailored to the data that they are intended to process. This research does not claim to provide one such identification algorithm, but instead suggests the usage of a framework, which can be used to compare the methods employed by the identification algorithms that abide by it. This framework is modular, which allows its components to be implemented to any specification, or even to be optimised with data learning techniques in order to achieve better reliability - see Section 5.2.

Consequently, Section 4.1 discusses established Machine Learning algorithms and justifies why they are not applicable to the identification problem formulated via the computation of source probabilities. Section 4.2 introduces the premise and the technical aspects of the identification framework and its components. Section 4.3 describes the implementation that was developed, and demonstrates

how a more reliable detection of features can be accomplished. Finally, Section 4.4 lists the components of the framework which have not been implemented and provides the beginning of a reasoning that should act as a basis for their future development.

4.1 Background and Purpose

Machine Learning is a field of Computer Science devoted to the design and analysis of algorithms which learn from data. Given the breadth of algorithms encompassed in this field, this analysis will be focused on Convolutional Neural Networks (CNNs) applied to image classification, which have previously demonstrated super-human performance at handwriting recognition [70]. Their success at image classification is due to: their usage of matrix convolutions or filters at different scales for picking up various features; their reliance of a large number of parameters, which allow them to explore a large number of potential approaches; the availability of substantial training data; and the leveraging of the computational power of Graphical Processing Units (GPUs) which speeds up the training process [70]. These algorithms have been successfully applied to the detection of features in latent prints [71]. The purpose of this section is to justify the approach undertaken as part of this research by describing the limitations of the application of established image classification methods to the fingerprint identification problem, which are twofold: their inability to compute source probability; and the limitations associated to the features they rely on.

As justified in Chapter 3, source probabilities play a central role in this research as they allow the quantification of the evidentiary value of fingerprint comparison based on their likelihood of coming from the same source, rather than merely based on their similarity, which is more subjective. The difference between established Machine Learning classification algorithms such as logistic regression and the approach described in this chapter is as follows: the former aim to compute the probability of an element, in this case a fingermark, of belonging to a class, in which case it makes sense to consider classes which correspond to each known source in a

database. However, the probability computed by these algorithms remains the probability of a fingerprint belonging to the class associated to a source, rather than the probability of that fingerprint coming from the source. The difference between both lies on the fact that the latter requires a formulation of the meaning of a fingerprint coming from a source, which is the very purpose of Chapter 3, which includes both the quantification of the amount of detail in common between two impressions, but also that of the possible modifications that occur from the deposition to the creation of an image file.

Second of all, the features used for classification purposes by Convolutional Neural Networks (CNNs) are extracted by convolutional layers, whose parameters or weights are updated during the training of the network. The usage of optimal features with respect to a data set may be problematic as some of these features may be spurious, they may also be different from those used by fingerprint examiners - namely the main pattern, the 2^{nd} and 3^{rd} level details, or may even not be understandable or usable by examiners for verification purposes. On the other hand, ensuring the same features as fingerprint examiners are used would provide clarity to experts and laypeople alike, and may also allow both validation and learning by fingerprint examiners.

The two above arguments justify the need to focus on the implementation of algorithms which solve the identification as it is formulated in Chapter 3, and to provide feature extraction methods specifically designed to ensure that features used by human experts are detected.

4.2 A Framework For Fingermark Representation

The aim of this section is to define a representation framework for the set of fingerprints F, whose purpose is to define a means by which any fingerprint can be expressed or represented with fidelity by a limited set of meaningful parameters, so as to satisfy Requirements (R3) and (R4). With this framework, it becomes possible to tie the appearance of a fingerprint back to these parameters, thus allowing: the generation of a fingerprint from a set of parameters; and, conversely,

the parameterisation of a fingerprint image associated to a given fingerprint according to the model considered. This specific relation between a fingerprint and its associated parameterisation is formalised via a model, which is defined as a function from P, the set of parameters chosen, to F, the set of fingerprints introduced in Chapter 3,

$$\begin{aligned} m : \ p &\mapsto f, \\ P &\to F \end{aligned} \tag{4.1}$$

where (P, d_F) is a metric space of finite dimension, and p will be referred to as the parameterisation of f. The definition of a model plays a pivotal role within an identification system as it represents how a fingermark is apprehended and represented.

An example of a model consists in representing each fingermark by their list of minutiae, which is the standard in forensic and research practices, and in identification algorithms at the time of writing. This model will be referred to as the canonical model. However, more comprehensive definitions of a model could, for instance, include the presence of pores within the ridge lines and other 3rd-level details, which have proved to be useful to discriminate fingermarks in practical scenarios [13].

As mentioned in Theorem 3.3, the formulation of the source determination problem requires the definition of distance on the set of fingerprints F. This can be accomplished by noting that, provided that the model function m is bijective, it induces the following definition of a distance d_m on F:

$$d_m : (f, f') \mapsto d_P\big(m^{-1}(f), m^{-1}(f')\big). \tag{4.2}$$

This function is binary, symmetric, and it also satisfies the triangle inequality due to the fact that m is surjective, which proves that it does define a distance on F. From a practical perspective, this statement justifies the value in attempting to numerically invert the model function m. The function m^{-1}, which maps a fingerprint to its representation $p \in P$ will be referred to as the representation function. The purpose of the representation framework is to define a set of algorithms which are well-suited to compute the representation of any fingerprint according to a given model m.

Many representation algorithms can and have already been created. Taking the example of the canonical model, a representation algorithm is a minutiae detection algorithm, which has been the subject of considerable research. Similarly, Variational Auto-Encoders (VAEs) are unsupervised deep neural networks which aim to determine an efficient way to represent a data set in a latent space [72]. As opposed to these methods, the goal of the framework described here is to define a premise by which algorithms can abide in order for them to be understandable and presentable in court. It is designed to be modular in order for each of its components to be implemented in various ways.

The framework rests on the premise that the parameterisation of a fingermark can be established by making repeated observations of its associated image, and it includes the following components: observation tools, prediction tools, a saliency mapping, and integration tools. The structure of the framework is represented in Section 4.2, and a suggested pseudo-code is described in Algorithm 4.1. Observations can be interpreted as features or details, and can, in the context of fingerprint analysis, refer to minutiae or 3^{rd}-level details, for instance. Observations are made at a specific location within the image, according to a process which is defined by observation tools, the first component of the framework. Typical methods such as image filters can be used as observation tools within the framework. The locations at which subsequent observations are computed are determined by making predictions regarding the locations of future, which is accomplished by the prediction tools. The purpose of the prediction tools is to compute a distribution of the expected probability of presence of each type of observation across the entire image. These predictions are then used in order to determine where the attention of the algorithm is directed at by means of a saliency mapping, which proceeds by analogy with human vision [73], and returns the location or lists of locations that should be observed next. Finally, integration tools (which are unrelated to the mathematical meaning of the word) specify which observations are considered to be sufficiently reliable, and how they should be integrated within the parameterisation of the fingermark.

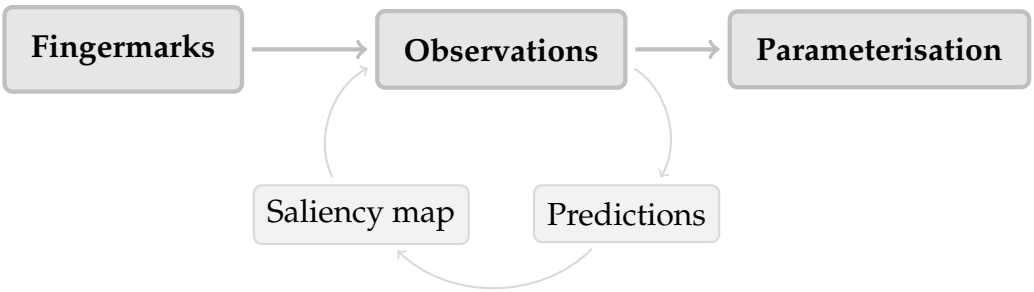


Figure 4.1: Diagram of the fingermark representation framework.

Algorithm 4.1 Representation framework

procedure REPRESENT(f)
    Initialise the saliency map, the observation set, the prediction set, and the representation.
    repeat
        Select a location on the image using the saliency map.
        Observe at this location.
        Add observation to the observation set.
        Integrate the current observation to the existing representation.
        Predict the location of other observations.
        Update saliency map.
    until fingermark is represented.
    return representation
end procedure

4.3 Implementation of the Observation Tools

Over the course this Ph.D. research, a simple implementation of the framework described in Section 4.2 was produced. At the time of writing, the program can open fingerprint image files and query them from an API, compute the representations of several images simultaneously using multithreading, and then post the resulting representations via an API for storage. It also implements the optimisation methods described in Section 5.1. An example of a representation computed and displayed by this program is provided in Figure 4.2. This implementation was made in C++, because it offers better performance over other computing languages, and is over

14′000 lines of code. The source code was written down from scratch and uses few dependencies: libpng to open image files [198], Qt for the graphical interface [199], nlohmann::json in order to manage JSON variables [200], and finally libcurl to make HTTP requests [201]. The source code is not provided as part of this thesis, nor was it made available on a public repository due to its lack of maturity. However, demonstrations can be made on request. The remainder of this section describes more specifically how the representation framework was implemented, and which observation tools were used.

Figure 4.2: Fingerprint image in black, and its ridge line annotations in red. The annotations have been automatically computed by the algorithm developed during this research.

The implementation is focused on the development of observation tools which can reliably detect the presence of ridge lines, identify its parameters, such as its orientation, width, and length, and also quantify the precision with which the observation was made. Furthermore, observations are made by following the ridge line, which was easily accomplished by using the ridge orientation and lengths computed previously. This effectively corresponds to a simple implementation of the representation framework where only one observation is considered, where the predictions are made in a very simplistic and heuristic fashion, and where the representation of a fingermark is restricted to its list of features.

The premise of the observation tool described here is to build upon filter-based approaches such as SIFT [74], a state-of-the-art algorithm in feature detection, which performs multiple convolutions of an image with filters of different dimensions and different parameters in order to deduce which feature is present, and which has already been successfully applied to fingerprint identification [75]. The approach described here aims to make this process more rigorous by defining a function which will be used to locally fit the ridge line. Given one such function $g_{c,o}$, centered in $c \in \mathbb{N}^2$, and with parameters o, its associated average squared error function is defined as

$$e : \ (I, f., c, o) \mapsto \frac{1}{\mu\big(\mathcal{D}(g_{c,o})\big)} \sum_{x \in \mathcal{D}(g_{c,o})} \Big(f_{c,o}(x) - I(x)\Big)^2, \tag{4.3}$$

where I is a greyscale fingerprint image represented as a function from $\mathbb{N}^2$ to $[0, 1]$, μ is the Lebesgue measure on $\mathbb{N}^2$, and $\mathcal{D}(f_o)$ is the domain of the function $g_{c,o}$, which corresponds to the window over which the modeling function $g_{c,o}$ is compared to the image I.

Given a point $c \in \mathbb{N}^2$ on the image, the function $g_{c,\cdot}$ can be fitted by minimising its associated average squared error function using an optimisation method. As a result, the resulting optimal error

$$\min_o e(I, f., c, o) \tag{4.4}$$

can be interpreted as the confidence of the algorithm in its observation at point c, and the associated optimal parameters o^* given by

$$o^* := \operatorname*{argmin}_o e(I, f., c, o) \tag{4.5}$$

provide the parameters of the ridge which have been computed.

In order for this reasoning to be applicable, it now suffices to define a fitting function which models the local appearance of a ridge line accurately. The fitting function used in this research is defined as follows:

$$\begin{array}{rccc} g : & p & \mapsto & s_{h,-a}\big(d_{c,\alpha}(p)\big) - s_{h,a}\big(d_{c,\alpha}(p)\big), \\ & \mathbb{N}^2 & \to & [0, 1] \end{array} \tag{4.6}$$

where

$$\begin{cases} d_{c,\alpha} : p \mapsto \cos(\alpha)(p_x - c_x) + \sin(\alpha)(p_y - c_y), \\ s_{h,a} : x \mapsto 1 - \dfrac{1}{1 + e^{-h(|x|-a)}}. \end{cases} \tag{4.7}$$

This choice is effectively a 2-dimensional generalisation of the sigmoid function $s_{h,\alpha}$ which incorporates parameters of the ridge line such as its orientation α, its half-width a, and the clarity with which it is defined h. The properties of this fitting function have been studied and are described in Appendix G. A graph of this function after it has been fitted to a section of a fingerprint image is represented in Figure 4.3.

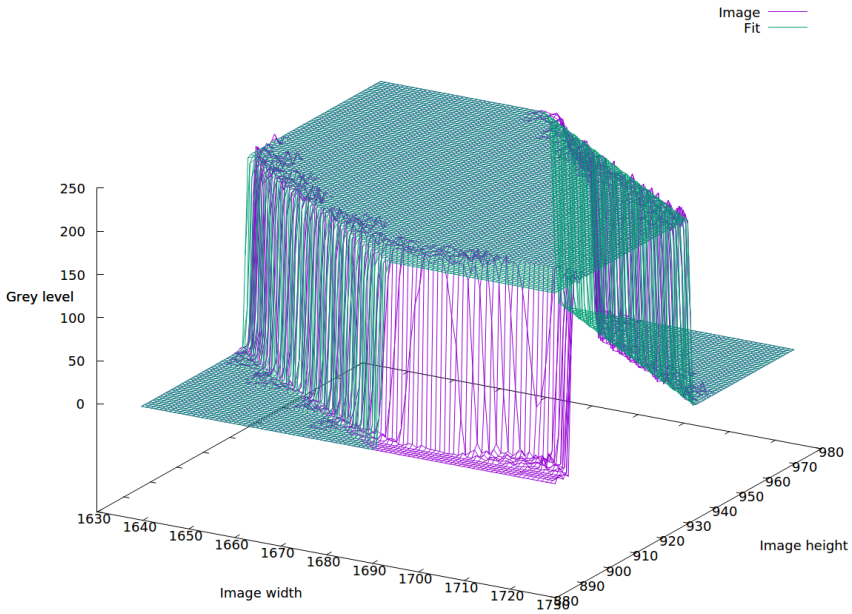


Figure 4.3: Graph of the fitting function defined in Equation (4.6) used for the detection of ridge lines, and of the fingerprint image in grey value on which it has been fitted.

Finally, prior to be added to the representation of the fingerprint, the algorithm confirms or rejects observations if their associated error is below a set threshold $t \in [0, 1]$,

$$e(I, g., c, o^*) < t; \tag{4.8}$$

otherwise, they are discarded. In that respect, the threshold t represents the acceptable average squared greyscale error between the function and the image which is deemed acceptable.

Based on a visual assessment, the results obtained with this algorithm, which are presented in Figure 4.2, are very promising. For more rigour, these performances

should be subjected to an evaluation based on a quality metric. That metric should be the result of a quantitative comparison of the positions of the ridge points found between the algorithm and an examiner. Further research should also be devoted to the simultaneous usage of several observation tools, to the inclusion of pores within the fitting functions used.

4.4 Completing the Identification System

After describing the identification framework and the simple implementation which was developed during this research project, the purpose of this final section is to describe how some of the components of the framework listed in Section 4.2 could be implemented in order to improve upon the implementation suggested.

4.4.1 Model

A limitation of the current implementation of the framework is that its associated model is limited to the list of observations which it makes. It is suggested that this limitation be addressed by representing a fingermark as a set of ridge lines, which in turn are defined as a set of Bézier curves. Let us first introduce the definition of a ridge line.

Definition 4.1 (Ridge line)**:**
A ridge line r is defined as a pair of an n_r-uplet of cubic Bézier curves $(B_i)_{i\in[\![1,n_r]\!]}$, where $n_r \in \mathbb{N}^*$ refers to the number of Bézier curves or sections within a single ridge line, and a real number $w \in \mathbb{R}^*$, the width of the ridge, which are such that:

a) the first control point of each path after the first one corresponds to the last control point of a previous path

$$\forall i \in]\!]1, n_r]\!], \exists j \in [\![1, i[\![\mid \ B_i(0) = B_j(1); \tag{4.9}$$

b) no Bézier curve intersects itself

$$\forall i \in [\![1, n_r]\!], \forall (s,t) \in [0,1]^2, \ s \neq t \Rightarrow B_i(s) \neq B_i(t); \tag{4.10}$$

c) and no Bézier curve intersects a previous Bézier curve on a point which is not the first control point of the former, or the first or last control point of the latter

$$\forall i \in [\![1, n_r]\!], \forall j \in [\![1, i[\![, \forall s \in]0, 1], \forall t \in]0, 1[,\ B_i(s) \neq B_j(t). \tag{4.11}$$

The points of bifurcation of a ridge line r (which are unrelated to bifurcation theory) are the set of points in $\mathbb{R}^2$ such that there exists two different sections B_i and B_j with $i \neq j$ such that $B_i(0) = B_j(0)$.

The purpose of imposing the above constraints on the set of Bézier curves which composes a ridge line ensures that they are not ill-formed, while still encompassing ridge lines that have points of bifurcation. Additionally, each ridge line is such that it does not have a point of bifurcation, then there is no permutation of its Bézier curves that also defines a ridge line. This definition could be refined to ensure that this property is still satisfied by ridge lines with points of bifurcation, for instance by adding a constraint on the ordering of such bifurcations which is based on their lengths. The next definitions give meaning to the width w of a ridge line, and introduce the suggested model for the representation of fingermarks.

Definition 4.2 (Graph and path of a ridge line)**:**
The path of a ridge line is defined as the union of graphs of its associated Bézier curves

$$\text{path}(r) := \bigcup_{i \in [\![1, n_r]\!]} \text{graph}(B_i). \tag{4.12}$$

Additionally, the graph of a ridge line is defined as the w-neighbourhood of its path, or, in other words, the subset of $\mathbb{R}^2$ of points which are at a distance less than w from $\text{path}(r)$.

Definition 4.3 (Parameterisation of a fingermark and modeling function)**:**
The parameterisation p of a fingermark is a n_p-uplet of ridge lines $(r_i)_{i \in [\![1, n_p]\!]}$ such that their graphs are not pairwise intersecting:

$$\forall (i, j) \in [\![1, n_p]\!]^2,\ i \neq j \Rightarrow \text{graph}(r_i) \cap \text{graph}(r_j) = \varnothing. \tag{4.13}$$

The modeling function for this model is defined as the function which maps p to the greyscale image which is black on the union of the graphs of its ridge lines, and

white everywhere else:

$$\begin{array}{rccc} m: & p & \mapsto & \mathbb{1}_{\bigcup\limits_{i\in[\![1,n_p]\!]}} \text{graph}(r_i). \\ & P & \to & \mathcal{F}(\mathbb{N},[0,1]) \end{array} \tag{4.14}$$

This model function is related to the notion of path in the .svg file format [223], but incorporates constraints on each curve in order to ensure that ridge lines are well-formed and carry meaning. This model can be used for the definition of a new vector image file format that addresses the remarks made previously about SVG compression in Section 2.4.6. However, unlike other compression methods, this model is restricted and tailored to the representation of fingerprint images.

While these definitions allow the modeling function m to reproduce the basic expected appearance of fingermark without its pores, they must be refined in order to ensure that the modeling function m is injective. This may be accomplished by imposing constraints on the parameterisation and the ordering of the ridge lines and their sections. Doing so would ensure that m is a bijective function from P to $\text{Im}(m)$, which induces the definition of a distance on $\text{Im}(m) \subset F$, as detailed in Section 4.2. It is assumed that the function m is injective in the remainder of this section. Future work should therefore include the mathematical analysis of the properties of this model, and the implementation of integration tools which specify how the observations made should be used to build such representations of a fingermark.

4.4.2 Fingermark distance

The fingermark distance is a quantitative measure of the difference between two fingermarks, which is necessary for the comparison of fingermarks required by the source determination problem defined in Theorem 3.3. As mentioned in Section 4.2, the definition of one such distance on the set of fingermarks F can be accomplished by considering the model function m defined in Section 4.4.1, which is assumed to be injective on its domain, the set of parameters P. The problem therefore consists in defining a meaningful distance between sets of continuous curves in $\mathbb{R}^2$. This section merely makes suggestions of functions which should be further studied

and implemented in order to ensure that they are applicable to the fingerprint identification problems.

The very first function which should be considered is the minimal distance between two curves. Given two ridge lines r and r', this function is defined as

$$d_1 : (r, r') \mapsto \inf_{(t,t')\in[0,1]^2} d_{\mathbb{R}^2}\big(r(t), r'(t')\big), \tag{4.15}$$

where $d_{\mathbb{R}^2}$ refers to the euclidian distance in $\mathbb{R}^2$. This first suggestion is limited as it does not take the entire geometry of both ridge lines into account, and does not satisfy the identity of indiscernibles $d_1(r, r') = 0 \Leftrightarrow r = r'$.

A first improvement can be made over this first definition by considering

$$d_2 : (r, r') \mapsto \int_{t\in[0,1]} \inf_{t'\in[0,1]} d_{\mathbb{R}^2}\big(r(t), r'(t')\big) = \int_{t\in[0,1]} d_{\mathbb{R}^2}\big(r(t), r'\big). \tag{4.16}$$

This function takes into account the shape of both ridge lines more comprehensively, but does not satisfy the identity of indiscernibles either, and also requires solving for t' for each t. The computational cost of this approach is significant as: a naïve approach has a complexity of $O(nm)$, n being the number of points used for the discretisation of r, and m for that of r'; or will require the computation of appropriate data structures with more appropriate nearest neighbour search methods such as quadtree and octree-based approaches [76], or the use of spatial hashing [77].

Another approach is to simply consider

$$d_3 : (r, r') \mapsto \int_{t\in[0,1]} d^2_{\mathbb{R}}\big(r(t), r'(t)\big), \tag{4.17}$$

which is straightforward to compute, satisfies the identity of indiscernibles and the triangle equality. This function therefore does define a distance and is an excellent potential choice.

However, a full implementation of a distance between two fingermarks f and f' requires not only that pairs of ridge lines be compared, but actually that there is a distance defined between sets of ridge lines. The suggested approach consists in: computing an overlap distance for each pair of ridge lines belonging to f and f'; then matching pairs which minimise that first distance and so that each ridge line

is matched to another one; and finally, computing the distance between f and f' as the sum of a second distance between all pairs of matched ridge lines. A distinction is purposely made between the first and second distance. While the second distance can be implemented by means of the function d_3 for instance, the first one - referred to as the overlap distance - aims at identifying which ridge line is most likely to match another one. Therefore, it should not penalise the fact that one is a perfect partial overlap for the other, and therefore does not define a distance in the mathematical sense of the word. The next paragraphs investigate candidates for one such function.

Let us first consider the convex hull delimited by the control points of the Bézier curves which compose two ridge lines r and r', which is denoted by $\text{Conv}(r, r')$. The area of this convex hull can be easily computed using the shoelace formula [78]

$$d_4 : (r, r') \mapsto \mu\big(\text{Conv}(r, r')\big) = \frac{1}{2}\sum_{i=0}^{n}\big(x_i y_{i+1} - x_{i+1} y_i\big), \tag{4.18}$$

where μ refers to the Lebesgue measure on $\mathbb{R}^2$, $(x_i)_{i\in[\![1,n]\!]}$ and $(y_i)_{i\in[\![1,n]\!]}$ refer to the Euclidian coordinates of the control points of the Bézier curves of the ridge lines r and r' ordered in such a way that the resulting polygon is non self-intersecting. This function satisfies the requirements of the overlap distance in that it does not penalise partial overlaps; however, there is still room for improvement.

It is possible to generalise the previous approach and potentially improve upon its precision by considering the area of a representative shape delimited two ridge lines r and r'

$$d_5 : (r, r') \mapsto \mu\Big(\text{significant shape}(r, r')\Big). \tag{4.19}$$

A mathematically natural choice for such a shape would be the convex hull delimited by these parametric curves; however they are not necessarily visually representative of the point cloud they are associated to [79]. The determination of a representative shape is a subjective choice, and several concepts have been explored in research in order to address this problem:

- α-shapes (also called α-hulls), which are generalisations of convex hulls and, which, for $\alpha \in \mathbb{R}^{-*}$, are defined as the intersection of all closed complements of discs that contain all the points of a point cloud [80];

- concave hulls, for which there is no strict mathematical definition, and which have instead been defined by the algorithms which compute them [81, 82];
- minimum area enclosure [83].

These approaches should be further analysed, but also tested empirically in order to assert their relevance for the computation of an overlap distance.

Another function which has been studied and applied to the field of partial curve mapping [84–89] is the Fréchet distance, which can be defined as

$$d_{\text{Fréchet}}(r, r') = \inf_{\alpha,\beta} \max_{t\in[0,1]} d\Big(r \circ \alpha(t), r' \circ \beta(t)\Big) \tag{4.20}$$

where α and β are two reparameterisations of $[0, 1]$, which are continuous non-decreasing surjections from $[0, 1]$ to $[0, 1]$. A more elaborate approach recommended here consists in computing the simplified integral Fréchet distance

$$d_6(r, r') = \inf_{\alpha,\beta} \int_{t\in[0,1]} d\Big(r \circ \alpha(t), r' \circ \beta(t)\Big) dt \tag{4.21}$$

where α is a reparameterisation of $[0, 1]$, and β is a partial reparameterisation of $[0, 1]$, which is a simplified version of the integral Fréchet distance introduced in [90] as

$$d_7(r, r') = \inf_{\alpha,\beta} \int_{t\in[0,1]} d\Big(r \circ \alpha(t), r' \circ \beta(t)\Big) \Big(\|(r \circ \alpha)'(t)\| + \|(r' \circ \beta)'(t)\|\Big) dt. \tag{4.22}$$

The purpose of these definitions is to ensure that they are unaffected by the parameterisation of two curves, and they take into account the entire shape of both curves into account. Computational methods for the integral Fréchet distance defined in Equation (4.22) have been investigated in [90, 91].

Future research should be dedicated to the definition and implementation of a rigorous algorithm for the computation of a distane between fingermarks, the comparison of different choices for both overlap distance and distance between two ridge lines, and their respective testing on a significant dataset.

4.4.3 Population modeling

Given that generating probability distributions $F_{\Pi,\delta}$ over the set of fingermarks F which are associated to a given population Π is a requirement to accomplish

source assessment as laid out in Section 3.3, this section aims at discussing means of generating such distributions based on an existing database of fingermarks F_{DB}.

In order to address this problem, let us first consider a list of discriminating factors $f_\cdot = (f_i)_{i\in[\![1,n]\!]}$ which:

(F1) are known or suspected to affect the formation or appearance of fingermarks;

(F2) are sufficiently invariable through time, and have a sufficently long-lasting impact;

(F3) can be recorded for research purposes;

(F4) are widely collected by governments, and are useable for intelligence and prosecution purposes.

Such factors include but are not limited to biological sex and age, and may arguably include ancestry depending on the jurisdiction considered. The population Π considered can be set to a population which can be based in a geographic area from which the perpetrator of a crime is suspected or known to come, or which bears characteristics which the perpetrator is believed to have.

Given Factors (F3) and (F4), the distribution of the target population Π and that of the donor population associated to F_{DB} according to each discriminating factor f_i is known. Therefore, it is possible to consider a subset of F_{DB} that bears the same demographics as Π, and which will be denoted $F_{\text{DB},\Pi}$. Given a model which specifies the modeling function m used and its domain, the set of parameters P, the subset of real fingermarks $F_{\text{DB},\Pi}$ can be used in order to infer the distribution of parameters on P associated to the population Π according to the available information provided by F_{DB}. In turn, this allows the generation of a random variable p on P, which aims to reproduce the parameters of fingermarks that are representative of individuals which belong to population Π.

Given that premise, the random variable p induces a random variable on F which is defined by

$$\begin{array}{rccc} F_P: & \omega & \to & m\circ p(\omega), \\ & \Omega & \to & F \end{array} \tag{4.23}$$

provided that the modeling function m is measurable on the set of parameters P. Therefore, future research should be devoted to determining and comparing adequate statistical methods by which a random variable p on P can be generated from the parameters of a subset of fingerprints $F_{\mathrm{DB},\Pi}$.

Chapter 5

Optimisation

The main problems pertaining to fingerprint identification which have been established throughout this thesis, such as Equations (3.25) and (4.5), rely on the computation of a minimum or argmin. The computation of minima and maxima such as

$$\min_{x \in E} f(x) \quad \text{or} \quad \max_{x \in E} f(x) \tag{5.1}$$

belong to the mathematical field of Optimisation. The nature and properties of both the function f which is optimised, and the space E in which f takes space, heavily determine which optimisation methods are applicable. As such, optimisation can be dissected into: combinatorial optimisation, which applies to cases where E is a finite set; and continuous optimisation, which addresses problems where E is a space with cardinality of the continuum, such as the set of real numbers $\mathbb{R}$.

These techniques are not applicable to the optimisation problems mentioned throughout this thesis, as the latter involve functions with both discrete and continuous parameters. This class of problems is commonly referred to in the literature as "mixed optimisation", "discrete-continuous optimisation", or "mixed discrete-continunous optimisation" [92, 93].

A good choice of optimisation method is crucial as it governs the quality of the result obtained, as well as the amount of time required to compute it. Considering the application to fingerprint identification, this decision is pivotal in ensuring that minutiae are detected accurately, that identifications are correct, and in making the

identification system sufficiently fast for it to be scalable - which is a requirement for it to be considered for large-scale applications such as nation-wide forensic identification.

This chapter describes why and how Estimation of Distribution Algorithms (EDAs) have been utilised for the purpose of feature detection in fingerprints as part of this research, and how they could be used to fully solve the fingerprint identification problem as it is formulated in Equations (3.25) and (4.5). As such, a review and an analysis of mixed optimisation methods and of EDAs in particular are provided in Section 5.1. Then the potential of these algorithms for Machine Learning applied to fingerprint identification is explained in Section 5.2.

5.1 Mixed Discrete-Continuous Optimisation

As mentioned previously, mixed discrete-continuous optimisation addresses maximisation or minimisation problems such as Equations (3.25), (4.5) and (5.1) where the objective function f has both continuous and discrete variables. While computational methods exist to address optimisation over continuous, and over discrete variables separately, neither class of methods is directly applicable to mixed optimisation. As a result, solving this problem poses a difficulty in that it involves comparing between adaptations of methods existing in other subfields of optimisation, and native methods that are custom-designed to the problem.

The algorithms discussed in this section have been studied in the literature from the 2000s onwards, and build upon older, more basic optimisation methods called metaheuristics such as simulated annealing, evolutionary algorithms, ant colony optimisation algorithms, and many more. See [94] for a review of these methods. These methods differ based on the following conditions:

- the nature of the finite spaces at hand. While many optimisation problems involve ordinal (numerical) values, some involve categorical values such as character strings. For these spaces, e.g.

$$E = \Big\{\text{"ridge line"}, \text{"ridge ending"}, \text{"bifurcation"}, \text{"island"}\Big\}, \tag{5.2}$$

there may not exist a meaningful mathematical definition of ordering or distance [224], which makes the task more arduous;

- whether they deal with constraints, and the kind of constraints they deal with, such as

$$\min_{\substack{x \in E \\ g(x) \geqslant a}} f(x); \tag{5.3}$$

- any other assumption made on the objective function in terms of derivability on the continuous component of E, for instance.

These conditions naturally change with the application of the optimisation considered, which can vary greatly, from the optimal design of magnetic resonance devices [95, 96] to the definition of pacing strategies for team pursuit track cycling [97–99]. Not only the application, but also the exact formulation of the problem at hand heavily determines which choice of optimisation method is possible [99].

Given the nature of the mixed optimisation problems mentioned throughout this thesis, the class of problems of interest here are such that: the finite spaces involve categorical values; there are no constraints formulated on the optimisation problem; and no assumptions are made on the derivability of f. These choices are made in order to allow greater freedom in terms of the possible choices of observation tool for the fingerprint modeling framework, as described in Section 4.2. The mixed optimisation methods covered by this analysis fall within the following categories: continuous relaxation methods, two-partition approaches, pattern search algorithms - see [224, 225] for a detailed overview.

Continuous relaxation methods are designed to tackle cases where the finite spaces at hand involve ordinal variables. They are based off an existing continuous optimisation method, and are adapted to the mixed optimisation setting by fixing the resulting solutions. In that respect, they deal with ordinal variables as a constraint of a continuous optimisation problem. For example, [92] generalises an evolutionary algorithm to mixed optimisation by using truncation; Particle Swarm Optimisation (PSO) algorithms have also been generalised to mixed optimisation by rounding the velocity for integer variables [100], or by constraining the discrete variables to a grid [101].

Two-partition approaches consist in optimising a subset of the variables - usually the categorical ones - separately from the other ones. For instance, [95] describes a method that generalises a class of local continuous searches to mixed problems without making use of the derivatives of the objective function; [97] jointly use an evolutionary algorithm for the continuous aspect of the problem, and a random local search coupled with a simple evolutionary algorithm for the discrete part; [98] utilise two evolutionary algorithms that separately deal with the discrete and the continuous component of the problem, and such that one calls another; [102] implemented an algorithm whereby the continuous and discrete optimisation methods call one another via feedback mechanisms; finally, [99] designed a multi-objective optimisation framework that consists in applying different evolutionary operators on the discrete and on the continuous components of the problem.

Finally, pattern search algorithms are mesh-based approaches to mixed optimisation. These methods deal with both continuous and discrete variables simultaneously and rely on a search over a mesh that is updated at each iteration [103, 226, 104].

The suggestion made in this thesis is not simply based on the expected performance of the optimisation algorithm on the class of problems considered, but more importantly on the flexibility of the approach considered. The reason for that is that no identification system such as the one described in this thesis exists at the time of writing, and the primary aim of this research is to demonstrate the feasibility of the approach it describes. As such, the algorithm chosen should be flexible enough to be applicable to the identification and the observation subproblems described by Equations (3.25), (4.5) and (5.1), and to allow a varied choice of observation tools within the observation framework. EDAs have been chosen for these reasons. The remainder of this section describes: how these methods, which are traditionally used for continuous optimisation, can be used to address mixed optimisation problems; what their limitations are; and how they can be made sufficiently fast to be applicable to production environments.

The premise of EDAs is to randomly evaluate the objection function over

its domain, in a similar way as the Monte Carlo methods. However, instead of observing the function uniformly over its domain, EDAs use the information gathered about the function during previous iterations, and evaluate in priority areas that are more likely to yield optimised (either minimal or maximal) values. This is achieved by specifying the distribution according to which the algorithm should evaluate the objective function based on the previous optimised values found at the previous iterations. These algorithms can be classified both as probabilistic methods and evolutionary methods, as the algorithms generates a population of solutions at each iteration, see Algorithm 5.1 or [105, 106] for a generic pseudo-code of EDAs.

An EDA is defined by the method it uses to compute the random variable X used to evaluate the objective function from the previous generation - or all the previous generations - of solutions computed. This is made apparent by the call to the compute_parsing_rdv function in Algorithm 5.1, which is purposely left undefined. In practice, this is often accomplished by selecting a fraction of the best solutions found during the previous iterations of the algorithm, and building a distribution which favours evaluating neighbourhoods of these solutions. Some approaches involve building multimodal normal distributions, also called Gaussian Mixture Models (GMMs), which refer to distributions whose probability distribution functions can be written as

$$x \mapsto \sum_{i \in [\![1,n]\!]} \alpha_i \, f_{\mathcal{N}(m_i, \Sigma_i^2)}(x), \tag{5.4}$$

where $f_{\mathcal{N}(m_i,\Sigma_i^2)}$ is the probability distribution function of a multivariate normal distribution of mean m_i and covariance matrix Σ_i, $\forall i \in [\![1,n]\!], \alpha_i \in [0,1]$ and $\sum_{i \in [\![1,n]\!]} \alpha_i = 1$. See Figure 5.1 for the graph of a multimodal normal distribution on $\mathbb{R}^2$. For example, the GMM can be made such that each mode is centered around the top 10% of solutions found previously. The number of normal components present in the mixture, their associated weights and parameters (mean and covariance matrix) greatly impact how future values will be evaluated by the EDA, and therefore determine the performance of the algorithm.

Because of the fact that GMMs are a class of continuous probability distribution functions, the usage of EDA is usually limited to continuous optimisation. This

Algorithm 5.1 Generic pseudo-code for EDA optimisation methods.

```
1: procedure MINIMISE(f, E, n_gen, n_pop)
2:     ▷ Main EDA minimisation function.
3:
4:     ▷ Computing the first generation of solutions.
5:     X ← the uniform distribution over E.
6:     P[0] ← generate_population(f, X, n_pop)
7:
8:     ▷ Computing the subsequent generations.
9:     for i from 1 to n_gen do
10:        X ← compute_parsing_rv(P[i − 1])
11:        P[i] ← generate_population(f, X, n_pop)
12:    end for
13:
14:    ▷ Returning the minimum value achieved by the last population.
15:    return min(P[n_gen])
16: end procedure
17:
18:
19: procedure GENERATE_POPULATION(f, X, n_pop)
20:    ▷ Function that generates variables according to X and evalutes f for their
       values.
21:
22:    for ω from 1 to n_pop do
23:        Generate a random value x according to X
24:        A[ω] ← [x, f(x)]
25:    end for
26:
27:    return A
28: end procedure
```

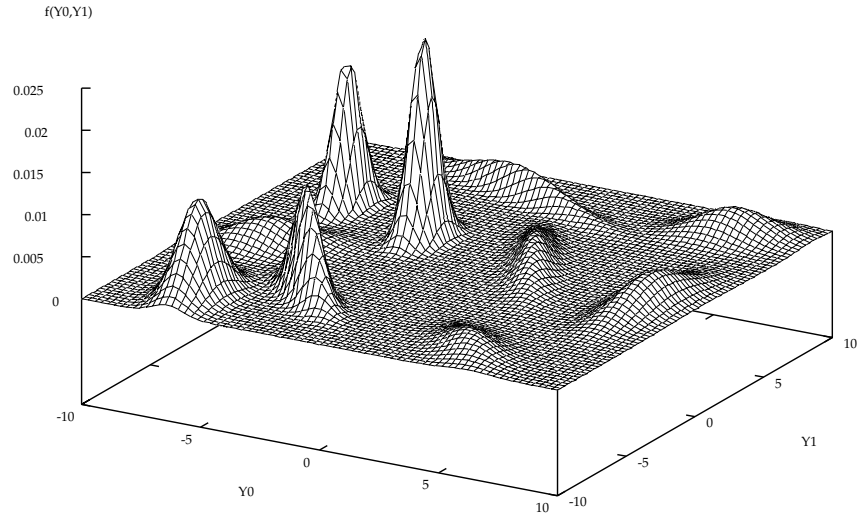


Figure 5.1: Graph of the pdf of a multimodal normal distribution on $\mathbb{R}^2$ [227].

restriction can be lifted by considering a probability distribution that extends GMMs to discrete spaces involving either ordinal or cardinal variables. The choice made in this research project is to consider a mixture model whose continuous component is modeled by a GMM, its ordinal component by a multivariate binomial distribution, and its categorical component by a categorical distribution.

A limitation of EDAs in comparison with continous approaches is that the dependencies between one or several variables are not natively taken into account, which is why research has been undertaken in introducing dependencies between variable via copulas [105–108].

Another limitation of EDAs is the computational cost associated with managing and generating random values according to a complex distribution such as a multinormal multivariate normal distribution [107, 109]. This issue has been previously addressed by clustering the best previous solutions into one or few, in order to limit the amount of modes involved [227], but this solution has been reported to result in the algorithm converging too quickly [110, 111]. The suggestion made here is to extend the classical GMM to a mixture of a GMM and a uniform distribution over the domain E as follows

$$x \mapsto \alpha \mathbb{1}_E(x) + \sum_{i \in [\![1,n]\!]} \beta_i \, f_{\mathcal{N}(m_i, \Sigma_i^2)}(x), \tag{5.5}$$

where $\alpha, \beta_i \in [0,1], \forall i \in [\![1,n]\!]$ and $a + \sum_{i \in [\![1,n]\!]} \beta_i = 1$. The rationale behind this definition is to ensure that that the EDA keeps exploring the entire domain E to an

extent controlled by the real parameter α. As a result, this parameter effectively controls how fast the algorithm converges, and can be reduced progressively with each iteration of the algorithm, as the certainty in the quality of the solutions obtained increases. This certainty can be quantified by the following quantity:

$$1 - \alpha = \sum_{i \in [\![1,n]\!]} \beta_i. \tag{5.6}$$

Let us note that this generalisation can be applied to any choice of multimodal distribution, and that the parameter α can be calibrated to set the speed of convergence of the algorithm.

Another way to address the computational cost of EDAs is to note that they are closely related to Monte Carlo methods, as they involve many repetitive random number generations as well as function evaluations, which are operations that are fairly adaptible to GPUs. The only component of EDAs which is not well-suited to GPU porting is the fact that previous solutions are partially sorted prior to the implementation of the parsing random variable. This can be accomplished using partial radix sort, whose adaptation to GPUs has already been investigated [112]. As a result, it is possible to leverage the computational power of these processing units, thus making the solution very amenable to large-scale applications.

All in all, there is an additional computational cost associated to EDAs in comparison with purely heuristic methods. This cost originates from the fact that the strategy followed by an EDA is justified by the use of a probabilistic model, which effectively represents our understanding and prediction of the behaviour of the objective function. Whether explicitly mentioned or not, each optimisation method makes an assumption on the objective function. That assumption can be a strong one such as the derivability of the function over its entire domain; or a weaker one such as stating that a low value of the objective function is somewhat indicative of the presence of other - potentially lower - values in the same neighbourhood. These assumptions are essential as they effectively allow the algorithm to beat the odds of a purely random strategy as the traditional Monte Carlo method. EDAs make possible the formulation of these weaker assumptions in the form of a probabilistic model, and the relevance of this model on the function or class of functions considered can be assessed by the proficiency of this model to

optimise these functions. In other words, the computational overhead of EDAs is comes from the added understanding of the objective function which they provide. Finally, it has been shown that this computational overhead can be managed, either by simplifying the multimodal distribution used by clustering the best solutions, and using the extended multimodal normal distribution described by Equation (5.5); and also by porting these algorithms to GPUs.

5.2 Optimisation-based Machine Learning and Application to Fingerprint Identification

The objective of this section is to describe how mixed optimisation with EDAs can be used to achieve robustness and precision in a fingerprint identification system.

Not only is it possible to use EDAs to perform the tasks required by the identification system, such as the detection of features with observation tools, or the identification of the most likely candidates for a match via the semi-metric d on the set of fingerprints, but these optimisation algorithms can also be utilised to optimise the parameters of the identification system itself. For example, these parameters can pertain to the fitting functions used to perform the detection of features (see Section 4.3), or to the generation of modification functions.

That being said, optimisation methods are only applicable to these contexts if they can be provided with an objective function which quantifies the quality of a given choice of a parameter for the task. In a similar way that the error between a fitting function and a fingerprint image was considered to accomplish feature detection in fingerprints in Section 4.3, it is possible to consider the annotation error of an annotation system, which is defined previously by the set of observation tools of an identification system, and the identification error committed by an identification system. It is assumed that reference annotation data provided by trusted experts is available to quantity the former, and that ground-truth information about the source associated to each fingerprint is provided. Both

of these requirements are met by the database described in Chapter 2, which is denoted by F_{DB} in Chapter 3.

Definition 5.1 (Annotation error)**:**
Let us also denote the annotation system with parameters p by $annotation_p$, which refers to a function that maps a fingermark to its computed annotation, and the reference annotation data associated to the fingermark f by a_f. The annotation error committed by the annotation system a_f can be defined as

$$\sum_{f \in F_{\text{DB}}} \Big| annotation_p(f) - a_f \Big|. \tag{5.7}$$

Definition 5.2 (Identification error)**:**
Similarly, let us denote by $identification_q$ the identification system with parameters q, which maps a pair of fingermarks (f, f') to either 1 or 0, which respectively correspond to whether these fingermarks come from the same source or not. Additionally, the ground-truth source associated to a fingermark f is denoted by s_f. The identification error committed by the identification system can then be defined as

$$\sum_{\substack{(f,f') \in F_{\text{DB}} \\ f \neq f'}} \Big| identification_p(f, f') - \delta_{s_f, s_{f'}} \Big|, \tag{5.8}$$

where δ refers to the Kronecker delta.

Consequently, Definitions 5.1 and 5.2 define objective functions that can be optimised by EDAs, regardless of the fact that the parameters involved have continuous, ordinal, or categorical components. This optimisation allows for the tuning of the parameters of the identification system or any of its components, such as the annotation system. Let us note that Definition 5.2 can be improved by taking different identification statements into account, such as "There is insufficient information to state whether both fingerprint come from the same source", and by weighing these results in a more comprehensive way.

The above reasoning of defining an error function associated with the identification system can be extended to any of its components whose performance must be monitored, provided that it is possible to define an objective function that quantifies that performance. This condition is often determined by the presence or

absence of associated ground-truth or trusted data, which further demonstrates the need to provide such data in the database described in Chapter 2.

The premise of optimising the identification system - or one of its components - with regards to its performance on an entire dataset can be interpreted as a calibration of the system on the data it is aimed to process, which also reinforces the need for the database F_{DB} to include fingerprints which are representative of those encountered in a forensic context. By noticing that F_{DB} effectively acts as a training set for the identification system, it can be said that this premise defines a mixed optimisation-based approach to Machine Learning, which can be utilised to ensure the robustness and precision of the entire fingerprint identification procedure.

All in all, the identification system described in this thesis has a somewhat set structure in that it abides to the identification framework described in Chapter 4, but remains modular in terms of the modeling functions and other parameters relied upon by its components. In this setting, EDAs not only perform the optimisation required to accomplish the observation, modeling, and identification tasks, but also aim to calibrate and improve the performance of the identification. As a result, this system is intended to meet Requirements (R1) and (R2), which were formulated in Chapter 4.

The remainder of this section briefly compares the optimisation-based approach to Machine Learning to other methods, and discusses the expected computational cost of the optimised identification system. The set structure of the identification framework, as well as the analogy of its rationale with human vision, is intended to make this solution approachable and understandable to laypeople and forensic examiners, so as to satisfy Requirements (R3) and (R4). In that respect, this approach is in contrast with most Machine Learning methods, which are perceived as black box algorithms and lacking in transparency. This is a crucial concern given the societal role played by identification algorithms. Furthermore, EDAs provide an understanding of the behaviour of the objective function by means of its optimisation strategy, as mentioned in Section 5.1. These optimisation methods can also return a quantification of the quality of the solution it provides, which can be interpreted, in the case of feature detection in fingerprints, as the confidence of the

system in the presence of a minutiae at a given point.

However, it is expected that a full-fledged optimised identification system has an expected computational cost, especially during its training phase on the dataset F_{DB}. This cost stems from the fact that the optimisation of the identification error function defined in Equation (5.8) involves many calls to the identification system, which also relies on EDAs applied to feature detection, for instance. This cost may however not be unlike that of large-scale Machine Learning systems, which are also considerable. Solutions have been proposed throughout this chapter in order to alleviate the computational burden, such as: leveraging the potential of Graphical Processing Units by porting EDAs onto them (Section 5.1); and optimising smaller components of the identification system, such as the annotation system, whose associated error function is less computationally expensive.

Conclusion

With the ever increasing world population, being able to discriminate between large quantities of fingermarks is crucial in ensuring the evidentiary value of fingerprint identification. Consequently, it is necessary to collect and recover fingermarks with more accuracy, and also to be able to make use of that information with more precision. The premise of this thesis was to improve the reliability of automatic fingerprint identification on latent fingermarks in order for it to play a more prevalent role within the legal system.

This research project culminated with the establishment of an alternate and novel approach to the identification problem, rather than iterating upon existing techniques. This new approach was necessary in order to address the current flaws in the identification process, which have been investigated in Chapter 1. Most notably, current methods of identification are not calibrated according to a ground-truth dataset that is representative of crime scene fingermarks.

The first outcome of this research project is the mathematical analysis of the identification problem which is made in Chapter 3. This study features two formulations of the identification problem based on optimisation, source determination and source assessment, which clearly state the hypothesis and requirements which need to be met in order these formulations to be valid, and so that a rigorous identification can be made. A major outcome of this analysis is that it identifies conditions under which it can be said that two fingermarks come from the same source beyond any reasonable doubt, which is in contrast with current techniques that provide a probability of a likelihood ratio as an output. This reasonable doubt is quantified and defined as the identification error of the system, which can prove to be crucial for the judge, jury, and court to properly evaluate the

weight and implications of forensic analyses. It is known that juries can struggle with understanding probabilities and statistical-based results [113], which is why there is value in attempting to formulate the output of research in different ways in order to ensure that forensic analyses are properly utilized by the court.

Additionally, this thesis created a computational framework described in Chapter 4 in order for identification algorithms within this framework to satisfy the Daubert standard and be amenable to the criminal justice system . A basic implementation of the fingerprint representation framework was produced in C++, which demonstrates that higher reliability can be achieved in the detection of minutiae, which is one of the main impediments to the robustness of current identification algorithms. This was accomplished by using optimisation methods in order to locally fit a fingerprint image with a function which represents the feature which need detecting, and which in this case were ridge lines. It is also advised that further research be lead into the full implementation of the identification framework, particularly in terms of the model, the fingermark distance, and the population modeling techniques to be used.

Given the prevalence of optimisation-based problems in the mathematical formulations that were made, this thesis also investigated adequate numerical methods to solve them in Chapter 5. The difficulty encountered lies in the fact that the objective functions to be optimised may involve continuous, but also finite spaces in the form of either ordinal and categorical variables. These problems fall within the field of mixed discrete-continous algorithms. As such, Estimation of Distribution Algorithms, methods which are traditionally known for continuous optimisation, were generalised to this setting, and were also numerically implemented. It was also shown how these algorithms could be applied at different stages of the fingerprint identification problem, including for the purpose of calibrating and adapting the entire system to a training dataset, thus suggesting an optimised-based approach to Machine Learning.

This thesis aimed to set the theoretical groundwork for more robust identification methods, but also to suggest practical avenues to implement a better identification system. That is why the premise, the software development, and the population of

a ground-truth database of fingerprints for research were undertaken, and were described in Chapter 2. However, despite the attempts at generating such a dataset within the context of this thesis, these endeavours were met with major limitations such as with the Ethics Committee (see Section 2.3), and with the data collection (see Section 2.5.3). These limitations represent in part the difficulties associated with novel approaches to research, since it is necessary to rely upon individuals from different backgrounds to understand how and why new approaches should be undertaken. This thesis therefore discussed possible ways to broach the topic of novel research with external bodies, such as Ethics Committees, which are applicable to researchers in any field who wish to pursue innovative ideas.

Academic sources

[1] International Association for Identification et al. *Fingerprint Sourcebook*. 1st ed. Washington DC, USA: National Institute of Justice, July 2011.

[12] F. Galton. *Finger Prints*. Macmillan and Co., 1892.

[13] D. R. Ashbaugh. *Quantitative-Qualitative Friction Ridge Analysis: An Introduction to Basic and Advanced Ridgeology*. 1st ed. CRC Press, Mar. 1999. ISBN: 1420048813.

[14] M. Bulmer. *Francis Galton: Pioneer of Heredity and Biometry*. Johns Hopkins University Press, Dec. 2003. ISBN: 9780801874031.

[15] A. Kollman. *Der Tastapparat der menschlichen Rassen und der Affen in seiner Entwickelung und Gliederung*. Voss Verlag, 1883.

[16] Kristine Bonnevie. "Studies on papillary patterns in human fingers". In: *Journal of Genetics* 15.1 (Nov. 1924), pp. 1–111. ISSN: 0022-1333. DOI: 10.1007/BF02983100.

[17] K. Bonnevie. "Was lehrt die Embryologie der Papillarmuster über ihre Bedeutung als Rassen- und Familiencharakter?" In: *Z. Indukt. Abstamm Ver.* 50 (Dec. 1929), pp. 219–248.

[18] M. Kücken and A. C. Newell. "Fingerprint formation". In: *Journal of Theoretical Biology* 235.1 (Feb. 2005), pp. 71–83. ISSN: 0022-5193. DOI: 10.1016/j.jtbi.2004.12.020.

[19] Kasey Wertheim. "Embryology and Morphology of Friction Ridge Skin". In: International Association for Identification et al. *Fingerprint Sourcebook*. 1st ed. Washington DC, USA: National Institute of Justice, July 2011.

[20] John R. Vanderkolk. "Examination Process". In: International Association for Identification et al. *Fingerprint Sourcebook*. 1st ed. Washington DC, USA: National Institute of Justice, July 2011. Chap. 9.

[21] S. Cole. "The Prevalence and Potential Causes of Wrongful Conviction by Fingerprint Evidence". In: *Golden Gate University Law Review* 28 (1 2006), pp. 39–105.

[22] National Research Council Committee on Identifying the Needs of the Forensic Sciences Community. *Strengthening Forensic Science in the United States: A Path Forward*. The National Academies Press, Aug. 2009.

[23] S. A. Cole. "Individualization is dead, long live individualization! Reforms of reporting practices for fingerprint analysis in the United States". In: *Law, Probability and Risk* 13 (Jan. 2014), pp. 117–150.

[24] S. A. Cole. "Forensics without uniqueness, conclusions without individualization: the new epistemology of forensic identification". In: *Law, Probability and Risk* 8 (July 2009), pp. 233–255.

[25] D. H. Kaye. "Probability, Individualization, and Uniqueness in Forensic Science Evidence: Listening to the Academies". In: *Brooklyn Law Review* 75.4 (2010), pp. 1163–1185.

[26] G. Langenburg. "Scientific Research Supporting the Foundations of Friction Ridge Examinations". In: International Association for Identification et al. *Fingerprint Sourcebook*. 1st ed. Washington DC, USA: National Institute of Justice, July 2011. Chap. 14.

[27] J. Abraham, C. Champod, C. Lennard, and C. Roux. "Modern statistical models for forensic fingerprint examinations: A critical review". In: *Forensic Science International* 232 (2013), pp. 131–150. DOI: 10.1016/j.forsciint.2013.07.005.

[28] C. Neumann, C. Champod, M. Yoo, T. Genessay, and G. Langenburg. "Quantifying the weight of fingerprint evidence through the spatial relationship, directions and types of minutiae observed on fingermarks". In: *Forensic Science International* 248 (2015), pp. 154–171. DOI: 10.1016/j.forsciint.2015.01.007.

[29] B. T. Ulery, R. A. Hicklin, J. Buscaglia, and M. A. Roberts. "Accuracy and reliability of forensic latent fingerprint decisions". In: *PNAS* 108.19 (May 2011), pp. 7733–7738.

[30] D. Maltoni, D. Maio, A. Jain, and S. Prabhakar. *Handbook of Fingerprint Recognition*. 2. Springer-Verlag London, 2009, pp. 1–494. ISBN: 978-1-84882-254-2. DOI: 10.1007/978-1-84882-254-2.

[31] J. W. Bond. “Visualization of Latent Fingerprint Corrosion of Metallic Surfaces”. In: *Journal of Forensic Sciences* 53 (4 July 2008).

[32] R. Cappelli, M. Ferrara, A. Franco, and D. Maltoni. “Fingerprint Verification Competition 2006”. In: *Biometric Technology Today* 15.7-8 (Aug. 2007).

[33] S. G. Crihalmeanu, Wv Morgantown, A. Ross, and L. Hornak. “A protocol for multibiometric data acquisition, storage and dissemination”. In: (May 2019).

[34] A. Sankaran, M. Vatsa, and R. Singh. “Latent Fingerprint Matching: A Survey”. In: *IEEE Access* 2 (Aug. 2014), pp. 982–1004. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2014.2349879.

[35] A. Sankaran, A. Agarwal, R. Keshari, S. Ghosh, A. Sharma, M. Vatsa, and R. Singh. “Latent fingerprint from multiple surfaces: Database and quality analysis”. In: *2015 IEEE 7th International Conference on Biometrics Theory, Applications and Systems (BTAS)*. Sept. 2015, pp. 1–6. DOI: 10.1109/BTAS.2015.7358773.

[36] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. “ImageNet: A Large-Scale Hierarchical Image Database”. In: *IEEE* (2009).

[37] “Facebook and Cambridge Analytica: What You Need to Know as Fallout Widens”. In: (Mar. 2018). Ed. by The New York Times.

[38] C. A. Goble, J. Bhagat, S. Aleksejevs, D. Cruickshank, D. Michaelides, D. Newman, M. Borkum, S. Bechhofer, M. Roos, P. Li, and D. De Roure. “myExperiment: a Repository and Social Network for the Sharing of Bioinformatics Workflows”. In: *Nucleic Acids Research* 38 (May 2010), pp. 677–682.

[39] M. Crosas, G. King, J. Honaker, and L. Sweeney. “Automating Open Science for Big Data”. In: *The Annals of the American Academy* 659 (May 2015). DOI: 10.1177/0002716215570847.

[40] B. Yamashita, M. French, S. Bleay, A. Cantu, V. Inlow, R. Ramotowski, V. Sears, and M. Wakefield. “Latent print development”. In: International Association for Identification et al. *Fingerprint Sourcebook*. 1st ed. Washington DC, USA: National Institute of Justice, July 2011, pp. 225–320.

[44] M. Kücken and A. C. Newell. “A model for fingerprint formation”. In: *Europhysics Letters* 68 (Oct. 2004), pp. 141–146.

[45] M. Kücken. “Models for fingerprint formation”. In: *ELSEVIER Forensic Science International* 171 (Apr. 2007), pp. 85–96.

[46] R. Cappelli. “Synthetic Fingerprint Generation”. In: *Handbook of Fingerprint Recognition*. Ed. by D. Maltoni, D. Maio, A. K. Jain, and S. Prabhakar. Springer, 2009. Chap. 6.

[47] R. Cappelli, D. Maio, and D. Maltoni. “An Improved Noise Model for the Generation of Synthetic Fingerprints”. In: *International Conference on Control, Automation, Robotics and Vision* (Dec. 2004).

[48] R. Cappelli. “SFinGe: an Approach to Synthetic Fingerprint Generation”. In: *International Workshop on Biometric Technologies* (June 2004), pp. 147–154.

[49] M. Indrawan-Santiago. “Database Research: Are We At A Crossroad?” In: *15th International Conference on Network-Based Information Systems* (2012), pp. 45–51.

[50] V. Abramova and J. Bernardino. “NoSQL Databases: MongoDB vs Cassandra”. In: *ACM* (2013), pp. 14–22.

[51] M. Qi. “Digital Forensics and NoSQL Databases”. In: *11th International Conference on Fuzzy Systems and Knowledge Discovery* (2014), pp. 734–739.

[52] K. Grolinger, W. A. Higashino, A. Tiwari, and M. AM. Capretz. “Data management in cloud environments: NoSQL and NewSQL data stores”. In: *Journal of Cloud Computing* 2.22 (2013).

[53] M. S. Brown, S. K. Shah, R. C. Pais, Y.-Z. Lee, M. F. McNitt-Gray, J. G. Goldin, A. F. Cardenas, and D. R. Aberle. “Database Design and Implementation for Quantitative Image Analysis Research”. In: *IEEE Transactions on Information Technology in Biomedicine* 9 (July 2004).

[54] A. Jain and S. Pankanti. “Automated Fingerprint Identification and Imaging Systems”. In: *Advances in Fingerprint Technology, 2nd Edition, Elsevier Science*. CRC Press, pp. 275–326.

[55] R.-I. Chang, Y. Yen, and Hsu T.-Y. "An XML-Based Comic Image Compression". In: *LNCS* 5353 (2008), pp. 563–572.

[56] C.-Y. Su, R.-I Chang, and J.-C. Liu. "Recognizing Text Elements for SVG Comic Compression and its Novel Applications". In: *International Conference on Document Analysis and Recognition* (2011).

[57] R. Sears, C. van Ingen, and J. Gray. "To BLOB or Not To BLOB: Large Object Storage in a Database or a Filesystem?" In: *Microsoft Research Technical Report* (June 2006).

[58] F. Chen, D. A. Koufaty, and X. Zhang. "Understanding Intrinsic Characteristics and System Implications of Flash Memory based Solid State Drives". In: *SIGMETRICS/Performance09* (June 2009).

[59] R. S. Ramotowski. *Advances in Fingerprint Technology*. 3rd ed. Vol. 1. CRC Press, Oct. 2012. ISBN: 9781420088342.

[60] M. J. Saks and J. J. Koehler. "The Coming Paradigm Shift in Forensic Identification Science". In: *Science* 309.5736 (2005), pp. 892–895. ISSN: 0036-8075. DOI: 10.1126/science.1111565. eprint: https://science.sciencemag.org/content/309/5736/892.full.pdf.

[61] W. C. Thompson, J. Vuille, F. Taroni, and A. Biedermann. "After uniqueness: the evolution of forensic science opinions". In: 102 (1 2018), pp. 18–27.

[62] I. W. Evett. "Towards a uniform framework for reporting opinions in forensic science casework". In: *Science & Justice* 38.3 (1998), pp. 198–202. ISSN: 1355-0306. DOI: https://doi.org/10.1016/S1355-0306(98)72105-7.

[63] C. Neumann, C. Champod, R. PuchSolis, N. Egli, A. Anthonioz, and A. BromageGriffiths. "Computation of Likelihood Ratios in Fingerprint Identification for Configurations of Any Number of Minutiæ". In: *Journal of Forensic Sciences* 52.1 (2007), pp. 54–64. ISSN: 0022-1198.

[64] Association of Forensic Science Providers. "Standards for the formulation of evaluative forensic science expert opinion". In: *Science & Justice* 49.3 (2009), pp. 161–164. ISSN: 1355-0306. DOI: https://doi.org/10.1016/j.scijus.2009.07.004.

[65] B. T. Ulery, R. Austin Hicklin, G. I. Kiebuzinski, M. A. Roberts, and J. Buscaglia. "Understanding the sufficiency of information for latent fingerprint value determinations". In: *Forensic Science International* 230.1 (2013), pp. 99–106. ISSN: 0379-0738. DOI: https://doi.org/10.1016/j.forsciint.2013.01.012.

[67] Denis Bosq. "Stochastic Processes and Random Variables in Function Spaces". In: *Linear Processes in Function Spaces: Theory and Applications*. New York, NY: Springer New York, 2000, pp. 15–42. ISBN: 978-1-4612-1154-9. DOI: 10.1007/978-1-4612-1154-9_2.

[68] G. D. Glancy and J. M. W. Bradford. "The Admissibility of Expert Evidence in Canada". In: *Journal of the American Academy of Psychiatry and the Law Online* 35.3 (2007), pp. 350–356. ISSN: 1093-6793. eprint: http://jaapl.org/content/35/3/350.full.pdf.

[69] W. C. Thompson. "How the probability of a false positive affects the value of DNA evidence." In: *Journal of Forensic Sciences* (Jan. 2003).

[70] S. Raschka and V. Mirjalili. *Python Machine Learning*. Packt Publishing, 2017. ISBN: 9781787126022.

[71] Y. Tang, F. Gao, and J. Feng. "Latent fingerprint minutia extraction using fully convolutional network". In: *2017 IEEE International Joint Conference on Biometrics (IJCB)*. Oct. 2017, pp. 117–123. DOI: 10.1109/BTAS.2017.8272689.

[72] M. Kingma D. P .and Welling. "Auto-Encoding Variational Bayes". In: (2013).

[73] L. Itti and C. Koch. "Computational modelling of visual attention". In: *Nature Reviews Neuroscience* 2 (Mar. 2001), pp. 194–203. DOI: 10.1038/35058500.

[74] D. G. Lowe. "Object recognition from local scale-invariant features". In: *7^{th} IEEE International Conference on Computer Vision*. Vol. 2. 1999, pp. 1150–1157. DOI: 10.1109/ICCV.1999.790410.

[75] R. Zhou, D. Zhong, and J. Han. "Fingerprint identification using SIFT-based minutia descriptors and improved all descriptor-pair matching". In: *Sensors (Basel, Switzerland)* 13.3 (2013). ISSN: 1424-8220.

[76] S. F. Frisken and R. N. Perry. "Simple and Efficient Traversal Methods for Quadtrees and Octrees". In: *Journal of Graphics Tools* 7.3 (2002), pp. 1–11. DOI: 10.1080/10867651.2002.10487560.

[77] J. Zhang, N. Mamoulis, D. Papadias, and Y. Tao. "All-nearest-neighbors Queries in Spatial Databases". In: *16th International Conference on Scientific and Statistical Database Management*. June 2004, pp. 297–306. DOI: 10.1109/SSDM.2004.1311221.

[78] Daniel Zwillinger. *CRC Standard Mathematical Tables and Formulae*. 31st ed. Chapman and Hall, 2002. ISBN: 1584882913.

[79] A. Galton and M. Duckham. "What Is the Region Occupied by a Set of Points?" In: *Geographic Information Science*. 2006, pp. 81–98. DOI: 10.1007/11863939_6.

[80] H. Edelsbrunner, D. Kirkpatrick, and R. Seidel. "On the shape of a set of points in the plane". In: *IEEE Transactions on Information Theory* 29.4 (July 1983), pp. 551–559. ISSN: 0018-9448. DOI: 10.1109/TIT.1983.1056714.

[81] A. Moreira and M. Y. Santos. "Concave hull: A k-nearest neighbours approach for the computation of the region occupied by a set of points". In: *GRAPP 2007 - 2nd International Conference on Computer Graphics Theory and Applications*. Jan. 2007, pp. 61–68.

[82] J.-S. Park and S.-J. Oh. "A New Concave Hull Algorithm and Concaveness Measure for n-dimensional Datasets". In: *Journal of Information Science & Engineering* 28.3 (May 2012), pp. 587–600.

[83] A. V. Vishwanath, R. Arun Srivatsan, and M. Ramanathan. "Minimum area enclosure and alpha hull of a set of freeform planar closed curves". In: *Computer-Aided Design* 45.3 (2013), pp. 751–763. ISSN: 0010-4485. DOI: 10.1016/j.cad.2012.12.001.

[84] H. Alt and M. Godau. "Computing the Fréchet distance between two polygonal curves". In: *International Journal of Computational Geometry & Applications* 5.01n02 (1995), pp. 75–91. DOI: 10.1142/S0218195995000064.

[85] A. Mosig and M. Clausen. “Approximately matching polygonal curves with respect to the Fréchet distance”. In: *Computational Geometry* 30.2 (2005). Special Issue on the 19th European Workshop on Computational Geometry, pp. 113–127. DOI: 10.1016/j.comgeo.2004.05.004.

[86] E. W. Chambers, É. C. de Verdière, J. Erickson, S. Lazard, F. Lazarus, and S. Thite. “Homotopic Fréchet Distance between Curves or, Walking your Dog in the Woods in Polynomial Time”. In: *Computational Geometry* 43.3 (2010), pp. 295–311. DOI: 10.1016/j.comgeo.2009.02.008.

[87] K. Witowski, M. Feucht, and N. Stander. “An Effective Curve Matching Metric for Parameter Identification using Partial Mapping”. In: *8th European LS-DYNA, Users Conference Strasbourg, pgs*. 2011, pp. 1–12.

[88] A. Driemel, S. Har-Peled, and C. Wenk. “Approximating the Fréchet Distance for Realistic Curves in Near Linear Time”. In: *Discrete & Computational Geometry* 48.1 (July 2012), pp. 94–127. ISSN: 1432-0444. DOI: 10.1007/s00454-012-9402-z.

[89] P. Accisano and A. Üngör. “Matching Curves to Imprecise Point Sets using Fréchet Distance”. In: *CoRR* abs/1404.4859 (2014).

[90] K. Buchin, M. Buchin, and Y. Wang. “Exact Algorithms for Partial Curve Matching via the Fréchet Distance”. In: *20th Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 645–654. DOI: 10.1137/1.9781611973068.71. eprint: https://epubs.siam.org/doi/pdf/10.1137/1.9781611973068.71.

[91] A. Maheshwari, J.-R. Sack, and C. Scheffer. “Approximating the Integral Fréchet Distance”. In: *Computational Geometry* 70-71 (2018), pp. 13–30. ISSN: 0925-7721. DOI: 10.1016/j.comgeo.2018.01.001.

[92] J. Lampinen and I. Zelinka. “Mixed Integer-Discrete-Continuous Optimization By Differential Evolution - Part 1: the optimization method”. In: *Czech Republic. Brno University of Technology*. 1999, pp. 77–81.

[93] S. S. Rao and X. Ying. “A hybrid genetic algorithm for mixed-discrete design optimization”. In: *Journal of Mechanical Design - Transactions of the ASME* 127.6 (Nov. 2005), pp. 1100–1112. ISSN: 1050-0472. DOI: 10.1115/1.1876436.

[94] El-Ghazali Talbi. *Metaheuristics: From Design To Implementation*. May 2009. ISBN: ISBN: 978-0-470-49690-9.

[95] S. Lucidi, V. Piccialli, and M. Sciandrone. "An Algorithm Model for Mixed Variable Programming". In: *SIAM Journal on Optimization* 15.4 (2005), pp. 1057–1084. DOI: 10.1137/S1052623403429573. eprint: http://dx.doi.org/10.1137/S1052623403429573.

[96] G. Liuzzi, S. Lucidi, V. Piccialli, and A. Sotgiu. "A magnetic resonance device designed via global optimization techniques". In: *Mathematical Programming* 101.2 (Nov. 2004), pp. 339–364. ISSN: 1436-4646. DOI: 10.1007/s10107-004-0528-5.

[97] M. Wagner, J. Day, D. Jordan, T. Kroeger, and F. Neumann. "Evolving Pacing Strategies for Team Pursuit Track Cycling". In: *MIC 2011: The IX Metaheuristics International Conference*. July 2011.

[98] C. D. Jordan and T. Kroeger. "An evolutionary algorithm for bilevel optimisation of Men's Team Pursuit Track Cycling". In: *2012 IEEE Congress on Evolutionary Computation*. June 2012, pp. 1–8. DOI: 10.1109/CEC.2012.6256558.

[99] M. Wagner. "Nested Multi- and Many-Objective Optimization of Team Track Pursuit Cycling". In: *Frontiers in Applied Mathematics and Statistics* 2 (2016), p. 17. ISSN: 2297-4687. DOI: 10.3389/fams.2016.00017.

[100] G. G. Dimopoulos. "Mixed-variable engineering optimization based on evolutionary and social metaphors". In: *Computer Methods in Applied Mechanics and Engineering* 196.46 (2007), pp. 803–817. ISSN: 0045-7825. DOI: 10.1016/j.cma.2006.06.010.

[101] C.-x. Guo, J.-s. Hu, B. Ye, and Y.-j. Cao. "Swarm intelligence for mixed-variable design optimization". In: *Journal of Zhejiang University-SCIENCE A* 5.7 (2004), pp. 851–860. ISSN: 1862-1775. DOI: 10.1631/jzus.2004.0851.

[102] F. Sambo, M. A. Montes de Oca, B. Di Camillo, G. Toffolo, and T. Stützle. "MORE: Mixed Optimization for Reverse Engineering - An Application to Modeling Biological Networks Response via Sparse Systems of Nonlinear Differential Equations". In: *IEEE/ACM Transactions on Computational Biology*

and Bioinformatics 9.5 (Sept. 2012), pp. 1459–1471. ISSN: 1545-5963. DOI: 10.1109/TCBB.2012.56.

[103] M. Kokkolaras, C. Audet, and J. E. Dennis. "Mixed Variable Optimization of the Number and Composition of Heat Intercepts in a Thermal Insulation System". In: *Optimization and Engineering* 2.1 (2001), pp. 5–29. ISSN: 1573-2924. DOI: 10.1023/A:1011860702585.

[104] M. A. Abramson, C. Audet, J. W. Chrissis, and J. G. Walston. "Mesh adaptive direct search algorithms for mixed variable optimization". In: *Optimization Letters* 3.1 (2008), p. 35. ISSN: 1862-4480. DOI: 10.1007/s11590-008-0089-2.

[105] R. Salinas-Gutiérrez, A. Hernández-Aguirre, and E. R. Villa-Diharce. "Using Copulas in Estimation of Distribution Algorithms". In: *MICAI 2009: Advances in Artificial Intelligence*. Springer Berlin Heidelberg, 2009, pp. 658–668. ISBN: 978-3-642-05258-3.

[106] R. Salinas-Gutiérrez, A. Hernández-Aguirre, and E. R. Villa-Diharce. "D-vine EDA: a new estimation of distribution algorithm based on regular vines". In: *GECCO*. 2010.

[107] L. F. Wang, Y. C. Wang, J. C. Zeng, and Y. Hong. "An Estimation of Distribution Algorithm Based on Clayton Copula and Empirical Margins". In: *Life System Modeling and Intelligent Computing*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 82–88. ISBN: 978-3-642-15859-9.

[108] R. Salinas-Gutiérrez, A. Hernández-Aguirre, and E. R. Villa-Diharce. "Estimation of distribution algorithms based on copula functions". In: *GECCO*. 2011.

[109] T. S. P. C. Duque, D. E. Goldberg, and K. Sastry. "Enhancing the Efficiency of the ECGA". In: *Parallel Problem Solving from Nature – PPSN X*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 165–174. ISBN: 978-3-540-87700-4.

[110] J. Grahl, P. A. N. Bosman, and F. Rothlauf. "The Correlation-triggered Adaptive Variance Scaling IDEA". In: *8th Annual Conference on Genetic and Evolutionary Computation*. GECCO '06. Seattle, Washington, USA: ACM, 2006, pp. 397–404. ISBN: 1-59593-186-4. DOI: 10.1145/1143997.1144071.

[111] S. Ivvan Valdez P., Arturo Hernández, and Salvador Botello. “Repairing normal {EDAs} with selective repopulation”. In: *Applied Mathematics and Computation* 230 (2014), pp. 65–77. ISSN: 0096-3003. DOI: 10.1016/j.amc.2013.12.081.

[112] Elias Stehle and Hans-Arno Jacobsen. “A Memory Bandwidth-Efficient Hybrid Radix Sort on GPUs”. In: *Proceedings of the 2017 ACM International Conference on Management of Data*. SIGMOD ’17. Chicago, Illinois, USA: ACM, 2017, pp. 417–432. ISBN: 978-1-4503-4197-4. DOI: 10.1145/3035918.3064043.

[113] A. Ligertwood and G. Edmond. “Expressing evaluative forensic science opinions in a court of law”. In: *Law, Probability and Risk* 11.4 (2012), pp. 289–302. DOI: 10.1093/lpr/mgs016.

[114] *Fingermark visualization manual*. eng. Stationary Office: London, 2014. ISBN: 9781782462347.

[115] M. Wang, M. Li, A. Yu, Y. Zhu, M. Yang, and C. Mao. “Fluorescent Nanomaterials for the Development of Latent Fingerprints in Forensic Sciences”. In: *Advanced Functional Materials* 27.14 (2017). ISSN: 1616-301X.

Legal sources

[116] *Daubert v. Merrell Dow Pharmaceuticals, Inc.* (1993 US Supreme Court). 509 U.S. 579 (1993).
URL: https://www.law.cornell.edu/supct/html/92-102.ZS.html.Last visited: 05/11/2016.

[117] *General Electric Co. v. Joiner* (1997 US Supreme Court). 522 U.S. 136 (1997).
URL: https://www.law.cornell.edu/supct/html/96-188.ZS.html.Last visited: 05/11/2016.

[118] *Kumho Tire Co. v. Carmichael* (1999 US Supreme Court). 526 U.S. 137 (1999).
URL: https://supreme.justia.com/cases/federal/us/526/137/.Last visited: 05/11/2016.

[119] *R. v. Mohan* (1994 Supreme Court of Canada). 1994 CanLII 80. [1994] 2 SCR 9. Case #23063.
URL: http://www.canlii.org/en/ca/scc/doc/1994/1994canlii80/1994canlii80.html.Last visited: 27/09/2015.

[120] *R. v. J.L.-J.* (2000 Supreme Court of Canada). 2000 SCC 51. [2000] 2 SCR 600. Case #26830.
URL: https://scc-csc.lexum.com/scc-csc/scc-csc/en/item/1815/index.do.Last visited: 04/11/2016.

[121] *Federal Rules of Evidence*. USA. 2015.
URL: https://www.law.cornell.edu/rules/fre. Last visited: 05/11/2016.

[122] *Criminal Code*. Canada. 1985.
URL: http://laws-lois.justice.gc.ca/eng/acts/C-46/. Last visited: 18/10/2015.

[123] *Police and Criminal Evidence Act*. UK. 1984.
URL: http://www.legislation.gov.uk/ukpga/1984/60/contents. Last visited: 28/09/2015.

[124] *Guidance on Expert Evidence*. UK. 2014.
URL: http://www.cps.gov.uk/legal/assets/uploads/files/expert_evidence_first_edition_2014.pdf. Last visited: 18/10/2015.

[125] *28 U.S. Code ğ534 - Acquisition, preservation, and exchange of identification records and information; appointment of officials*.
URL: https://www.law.cornell.edu/uscode/text/28/534. Last visited: 18/10/2015.

[126] *Identification of Criminals Act*. Canada. 1985.
URL: http://laws-lois.justice.gc.ca/eng/acts/I-1/FullText.html. Last visited: 28/09/2015.

[127] *Charter of Rights and Freedoms*. Canada. 1982.
URL: http://laws-lois.justice.gc.ca/eng/const/page-15.html. Last visited: 18/10/2015.

[128] *Privacy Act*. Canada. 1985.
URL: http://laws-lois.justice.gc.ca/eng/acts/p-21/. Last visited: 28/09/2015.

[129] *R. v. Beare* (1988 Supreme Court of Canada). [1988] 2 SCR 387. Case #20384.
URL: https://scc-csc.lexum.com/scc-csc/scc-csc/en/item/374/index.do.Last visited: 05/11/2016.

[130] *R. v. Doré* (2002 Supreme Court of Canada). 2002 SCC 12. [2002] 1 SCR 395. Case #33594.
URL: http://www.lexisnexis.ca/documents/2012scc12.pdf.Last visited: 18/10/2015.

[131] *Protection of Freedoms Act*. UK. 2012.
URL: https://www.gov.uk/government/publications/protection-of-freedoms-act-2012-dna-and-fingerprint-provisions/protection-of-freedoms-act-2012-how-dna-and-fingerprint-evidence-is-protected-in-law. Last visited: 27/09/2015.

[132] *28 CFR Part 16, Subpart C - Production of FBI Identification Records in Response to Written Requests by Subjects Thereof*.
URL: https://www.law.cornell.edu/cfr/text/28/part-16/subpart-C. Last visited: 18/10/2015.

[133] *The Consumer Privacy Protection Act*. USA. 2015.
URL: http://www.leahy.senate.gov/imo/media/doc/Consumer%20Privacy%20Protection%20Act%20of%202015_One%20Pager.pdf. Last visited: 18/10/2015.

[134] *Data Protection Act*. UK. 1998.
URL: http://www.legislation.gov.uk/ukpga/1998/29/contents. Last visited: 28/09/2015.

[135] *Frye v. US* (1923 US Court of Appeals). 293 F. 1013 (1923). Case #3968.
URL: http://www.law.ufl.edu/_pdf/faculty/little/topic8.pdf.Last visited: 27/09/2015.

[136] *Document #103 - Individualization/Identification Position Statement (Latent/Tenprint)*. 2013.
URL: http://www.swgfast.org/Comments-Positions/130106-Individualization-ID-Position-Statement.pdf. Last visited: 26/09/2015.

[137] *Document #10 - Standards for Examining Friction Ridge Impressions and Resulting Conclusions (Latent/Tenprint)*. 2013.
URL: http://www.swgfast.org/documents/examinations-conclusions/130427_Examinations-Conclusions_2.0.pdf. Last visited: 26/09/2015.

[138] *Document #12 - Standard Friction Ridge Automation Training (Latent/Tenprint)*. 2012.
URL: http://www.swgfast.org/documents/automation/121124_Automation_Training_2.0.pdf. Last visited: 26/09/2015.

[139] *People v. Jennings* ( ).
URL: http://law.justia.com/cases/california/court-of-appeal/2d/243/324.html.Last visited: 28/09/2015.

[140] *Mayfield v. US* (2007 US District Court, D. Oregon). 504 F.Supp.2d 1023 (2007).
URL: https://scholar.google.co.uk/scholar_case?case=4394360232307343544&q=mayfield+100%25&hl=en&as_sdt=2006.Last visited: 05/11/2016.

[141] *US v. Mitchell* (2004 US Court of Appeals, Third Circuit).
URL: https://scholar.google.co.uk/scholar_case?case=14529070956800992 98&q=fingerprint+false+match&hl=en&as_sdt=2006.Last visited: 27/09/2015.

[142] *Inspection of the Fingerprint Bureau of SCRO*. 2000.
URL: https://whereismydata.files.wordpress.com/2008/06/scro-fingerprint.pdf. Last visited: 28/09/2015.

[143] *3rd Year Reviews of Primary Inspections*. 2003.
URL: https://whereismydata.files.wordpress.com/2008/06/scro-fingerprint-2.pdf. Last visited: 28/09/2015.

Online sources

[144] D. Thakkar. *Minutiae Based Extraction in Fingerprint Recognition*. Oct. 2016.
URL: https://www.bayometric.com/minutiae-based-extraction-fingerprint-recognition/ (visited on 05/29/2019).

[145] *NGI Monthly Fact Sheet*. July 2018. URL: https://www.fbi.gov/file-repository/ngi-monthly-fact-sheet/view (visited on 09/04/2018).

[146] *NGI Officially Replaces IAFISYields More Search Options and Investigative Leads, and Increased Identification Accuracy*. URL: https://www.fbi.gov/services/cjis/cjis-link/ngi-officially-replaces-iafis-yields-more-search-options-and-investigative-leads-and-increased-identification-accuracy (visited on 09/04/2018).

[147] *Fingerprint Database - IDENT 1*. URL: https://www.gov.scot/Topics/Justice/law/dna-forensics/scottishdnadatabase/ident1 (visited on 09/04/2018).

[148] CNIL. *FAED : Fichier automatisé des empreintes digitales*. URL: https://www.cnil.fr/fr/faed-fichier-automatise-des-empreintes-digitales (visited on 12/06/2016).

[149] Directorate-General for Migration and Home Affairs (DG HOME). *Identification of applicants (EURODAC)*. URL: http://ec.europa.eu/dgs/home-affairs/what-we-do/policies/asylum/identification-of-applicants/index_en.htm (visited on 12/06/2016).

[150] Australian Police. *Automated Fingerprint Identification System (AFIS)*. URL: https://www.australianpolice.com.au/dactyloscopy/automated-fingerprint-identification-system-afis/ (visited on 12/07/2016).

[151] NIST. *Biometric Special Databases and Software*. URL: https://www.nist.gov/itl/iad/image-group/resources/biometric-special-databases-and-software#Fingerprint (visited on 12/06/2016).

[152] Biometrics Ideal Test (BIT). *CASIA-FingerprintV5*. URL: http://biometrics.idealtest.org/dbDetailForUser.do (visited on 12/06/2016).

[153] FVC2006. *FVC2006 Databases*. 2006. URL: http://bias.csr.unibo.it/fvc2006/databases.asp (visited on 09/28/2015).

[154] *Biometric System Laboratory*. URL: http://biolab.csr.unibo.it/sfinge.html (visited on 09/07/2018).

[155] *Web frameworks ranking*. URL: http://hotframeworks.com/ (visited on 12/03/2016).

[156] *Node.js*. URL: https://nodejs.org/en/ (visited on 12/03/2016).

[157] Quora. *Popularity of Node.js*. URL: https://www.quora.com/Why-is-Node-js-becoming-so-popular (visited on 10/04/2015).

[158] *Why Node.js is hitting the big time in Enterprise Markets*. URL: http://apmblog.dynatrace.com/2015/04/09/node-js-is-hitting-the-big-time-in-enterprise-markets (visited on 10/04/2015).

[159] Google. *Chrome V8*. URL: https://developers.google.com/v8/ (visited on 12/03/2016).

[160] A. R. Olakara. *Understanding the Node.js event loop*. URL: http://abdelraoof.com/blog/2015/10/28/understanding-nodejs-event-loop/ (visited on 12/03/2016).

[161] *Node.js Event Loop*. URL: https://www.tutorialspoint.com/nodejs/nodejs_event_loop.htm (visited on 12/03/2016).

[162] T. Norris. *Understanding the Node.js event loop*. URL: https://nodesource.com/blog/understanding-the-nodejs-event-loop/ (visited on 12/03/2016).

[163] *npm*. URL: https://www.npmjs.com/ (visited on 12/03/2016).

[164] URL: http://www.indeed.com/jobtrends/q-node.js-q-django-q-asp.net-q-j2ee-q-symfony-q-zend-q-ruby-on-rails-q-MVC.html (visited on 12/08/2016).

[165] URL: http://www.indeed.com/jobtrends/q-node.js-q-django-q-asp.net-q-j2ee-q-symfony-q-zend-q-ruby-on-rails-q-MVC.html?relative=1 (visited on 12/08/2016).

[166] Quora. *How good is Node.js?* URL: https://www.quora.com/How-good-is-Node-js (visited on 12/09/2016).

[167] Couchbase. *Couchbase NoSQL survey*. URL: http://www.couchbase.com/press-releases/couchbase-survey-shows-accelerated-adoption-nosql-2012 (visited on 10/02/2015).

[168] DB-Engines. *DB-Engines Ranking*. URL: http://www.project-voldemort.com/voldemort (visited on 10/02/2015).

[169] Inc MongoDB. *MongoDB for GIANT ideas*. URL: https://www.mongodb.com/ (visited on 12/07/2016).

[170] *The Mongoose package for MongoDB in Node.js*. URL: https://docs.mongodb.com/ecosystem/drivers/node-js/#node-js-driver (visited on 11/30/2016).

[171] *Glossary - MongoDB Manual*. URL: https://docs.mongodb.com/v3.0/reference/glossary/#term-objectid (visited on 11/30/2016).

[172] Twitter, Inc. *Twitter REST API*. URL: https://dev.twitter.com/rest (visited on 08/14/2015).

[173] Facebook, Inc. *Facebook Graph API*. URL: https://developers.facebook.com/docs/graph-api (visited on 08/14/2015).

[174] LinkedIn Corporation LAD. *LinkedIn API Console*. URL: https://apigee.com/console/linkedin (visited on 08/14/2015).

[175] Instagram. *Instagram API Console*. URL: https://instagram.com/developer/api-console (visited on 08/14/2015).

[176] *ipstack - Free IP Geolocation API*. URL: https://freegeoip.net (visited on 11/30/2016).

[177] *IP-API.com - Free Geolocation API*. URL: http://ip-api.com (visited on 11/30/2016).

[178] *Free Proxy / VPN / TOR / Bad IP Detection Service via API and Web Interface | IP Intelligence*. URL: https://getipintel.net (visited on 11/30/2016).

[179] *minFraud Overview | MaxMind*. URL: https://www.maxmind.com/en/minfraud-services (visited on 11/30/2016).

[180] *Wikipedia article about secure http*. URL: https://en.wikipedia.org/wiki/HTTPS (visited on 11/30/2016).

[181] Reuters. *LinkedIn suffers data breach*. URL: http://in.reuters.com/article/linkedin-breach-idINDEE8550EN20120606 (visited on 12/08/2016).

[182] *StackExchange discussion about password hashing*. URL: http://stackoverflow.com/questions/4494234/what-are-the-best-practices-to-encrypt-passwords-stored-in-mysql-using-php (visited on 12/05/2016).

[183] *StackExchange discussion about hash functions*. URL: http://security.stackexchange.com/a/4801 (visited on 11/30/2016).

[184] *StackExchange discussion about SHA1 and MD5*. (Visited on 12/05/2016).

[185] *StackExchange discussion about the bcrypt algorithm*. URL: http://security.stackexchange.com/a/6415 (visited on 11/30/2016).

[186] *Bcrypt*. URL: https://news.ycombinator.com/item?id=2004833 (visited on 12/05/2016).

[187] *Bcrypt*. URL: http://codahale.com/how-to-safely-store-a-password/ (visited on 12/05/2016).

[188] *bcrypt package for node*. URL: https://www.npmjs.com/package/bcrypt (visited on 09/06/2018).

[189] Google, Inc. *Google 2-step Verification*. URL: https://www.google.com/intl/en-GB/landing/2step/index.html#tab=how-it-works (visited on 10/04/2015).

[190] Twilio. *SMS Text Messaging API for Web Applications*. URL: https://www.twilio.com/sms (visited on 10/04/2015).

[191] Google, Inc. *reCAPTCHA: Easy on Humans, Hard on Bots*. URL: https://www.google.com/recaptcha/intro (visited on 10/04/2015).

[192] W3C. *SVG 1.1 (Second Edition)*. Aug. 2011. URL: http://www.w3.org/TR/SVG11 (visited on 10/02/2015).

[193] Cedar Lake Ventures, Inc. *Vector Magic*. 2007. URL: http://vectormagic.com (visited on 08/14/2015).

[194] Dice Holdings, Inc. *Autotrace*. URL: http://autotrace.sourceforge.net (visited on 08/14/2015).

[195] ImageMagick Studio LLC. *ImageMagick*. URL: http://www.imagemagick.org (visited on 10/01/2015).

[196] J. Kehayias. *Does Index Fragmentation Matter with SSD's?* Apr. 2013. URL: https://www.sqlskills.com/blogs/jonathan/does-index-fragmentation-matter-with-ssds (visited on 10/07/2015).

[197] StackOverflow. *BLOBing of image files*. URL: http://stackoverflow.com/a/4654590 (visited on 10/03/2015).

[198] Greg Roelofs. *libpng Home Page*. URL: http://www.libpng.org/pub/png/libpng.html (visited on 09/26/2018).

[199] The Qt Company. *Qt | Cross-platform software development for embedded & desktop*. URL: https://www.qt.io/ (visited on 09/26/2018).

[200] N. Lohmann. *GitHub - nlohmann/json: JSON for Modern C++*. URL: https://github.com/nlohmann/json (visited on 09/26/2018).

[201] D. Stenberg. *libcurl - the multiprotocol file transfer library*. URL: https://curl.haxx.se/libcurl/ (visited on 09/26/2018).

Other sources

[202] W. Thompson, J. P. Black, A. Jain, and J. B. Kadane. *Forensic Science Assessment - A Quality and Gap Analysis - New ways of reporting fingerprint evidence*. Tech. rep. American Association for the Advancement of Science, 2017.

[203] W. Knaap. *Scenes of Crime Officer Course Resource Manual*. Forensic Identification Services. Sept. 2011.

[204] Expert Working Group on Human Factors in Latent Print Analysis. *Latent Print Examination and Human Factors: Improving the Practice through a Systems Approach*. Tech. rep. NIST, Feb. 2012.

[205] eu-LISA. *Annual report on the 2014 activities of the Central System of Eurodac pursuant to Article 24(1) of Regulation (EC) No 2725/2000*. June 2015.

[206] INTERPOL. *Databases - Fact Sheet*. Mar. 2016.

[207] Forensic Science Regulator. *Fingerprint Quality Standards Specialist Group (FQSSG). Notes of the meeting held on 8 June 2016 at Room 4.035-36, Block B, 109 Lambeth Road, London, SE1 7LP*. June 2016. 8 pp.

[208] Forensic Science Regulator. *Codes of Practice and Conduct for forensic science providers and practitioners in the Criminal Justice System*. Version 2. Aug. 2014.

[213] University of Leicester. *Research Code of Conduct*.

[214] A. El-Haj, E. Abu-Taieh, and A. Abu-Tayeh. *Taxonomy of Image File Formats*. 2013.

[215] Government of India - Department of Information Technology. *Fingerprint Image and Minutiae Data Standard for e-Governance Applications in India*. Nov. 2010.

[216] UIDAI. *Biometric Design Standards for UID Applications*. Dec. 2009.

[217] NIST. *Data format for the Interchange of Fingerprint, Facial & Other Biometric Information*. Dec. 2013.

[218] F. L. Podio, J. S. Dunn, L. Reinert, C. J. Tilton, B. Struif, F. Herr, J. Russell, M. P. Collier, M. Jerde, L. O'Gorman, and B. Wirtz. *Common Biometric Exchange Formats Framework (CBEFF)*. Apr. 2004.

[219] S. Orandi, J. M. Libert, J. D. Grantham, K. Ko, S. S. Wood, F. R. Byers, B. Bandidi, S. G. Harvey, and M. D. Garris. *Compression Guidance for 1000 ppi Friction Ridge Imagery*. 500-289. Feb. 2014.

[220] O&O Software. *O&O Defrag and Solid State Drives*. 2014.

[221] President's Council of Advisors on Science and Technology. *Strengthening Forensic Science In the United States: A Path Forward*. Tech. rep. Aug. 2009.

[223] E. Dahlström, P. Dengler, A. Grasso, C. Lilley, C. McCormack, D. Schepers, and J. Watt. *Scalable Vector Graphics (SVG) 1.1 (Second Edition)*. Aug. 2011.

[224] T. Liao. "Population-based Heuristic Algorithms for Continuous and Mixed Discrete-Continuous Optimization Problems". PhD thesis. Université Libre de Bruxelles, 2013.

[225] K. Socha. "Ant Colony Optimization for Continuous and Mixed-Variable Domains". PhD thesis. Université Libre de Bruxelles, 2009.

[226] M. A. Abramson. "Pattern search algorithms for mixed variable general constrained optimization problems". PhD thesis. Rice University, Aug. 2002.

[227] P. A.N. Bosman and D. Thierens. "Mixed IDEAs". Dec. 2000.

Appendices

APPENDIX A: Participant Information Sheet

The FOREPRINT project

The principle

ForePrint is a collaborative forensic fingerprint database. Its purpose is to create a common repository for researchers worldwide to add the results of their independent data collections. These fingerprints are exclusively used for research, development and training purposes.

The names and identifying information of each donor are not made public, and are only known to the researchers in charge of each data collection and the administrators of the database.

Access to this database may be monetised depending on the needs of each user and the nature of their activity (commercial or not). However, the names and identifying information of each donor are neither sold nor communicated to any third party.

What it means for the donor

Your identifying and contact details are only requested in order to provide you with the possibility to have your data removed at a later date, and to ensure that there are no duplicates in the system. We appreciate your contribution to this database and respect your right to privacy, therefore we take every precaution to ensure that your data is kept securely.

You may request the removal of your data from the database by contacting either the primary researcher at ejap1@le.ac.uk, or the administrators of the database at consent@foreprint.org.

The guidelines of this project have been elaborated with the Ethics Committee and the Information Assurance Services of the University of Leicester, and in accordance with the Data Protection Act. Whilst this study will comply with thqt legislation at all times, it can only do so subject to the limitations of the law.

This data collection

Fingerprints deposited in crime scenes are frequently partial and of very poor quality. This makes it hard for fingerprint experts to identify the possible perpetrator. Additionally, those fingerprints may only be used for investigation and conviction purposes, and therefore not for research.

To this day, there is no large research database that contains fingerprints similar to crime scene prints. The purpose of this data collection is to address this and attempt to reproduce crime scene prints in research conditions, by using different substrates and development techniques, and inviting the donor to deposit fingerprints according to different scenarios and in different conditions.

The creation for this dataset could be invaluable for the development of automatic fingerprint identification algorithms, and for establishing conclusions about the circumstances of a fingerprint deposition. This will not only increase the ability to make identifications using automated systems, but it can also aid in crime scene event reconstruction.

APPENDIX B: Participant Consent Form

Background Information

Title : FOREPRINT - Data Collection #1.

Primary researcher: Etienne Pillin - ejap1@le.ac.uk.

Purpose of the experiment: Gathering fingerprints in different conditions for the purpose of: 1) creating a database that will serve as a training set for automatic fingerprint identification algorithms; 2) serving as a reference data set that can be accessed by other researchers for academic purposes. These fingerprints will not be used for criminal prosecution.

Scope of data collection: doctoral research supervised by Prof. Jeremy Levesley - jl1@le.ac.uk, and Dr. Cheryl Hurkett - cph9@le.ac.uk.

Details of participation: The participant will have to fill in a datasheet and will then be invited to leave his/her fingerprints on different items and in different conditions. These situations include, among others: drinking from a receptacle, writing on a sheet of paper, using tape, and making depositions in sheep blood. The participant is able to choose which scenario he/she wants to participate in, and how many. The entire experiment is expected to take up to three 2-hour sessions, and will take place on the University of Leicester campus (exact location TBC).

Consent Statement

- I have read the Participant Information Sheet, I am aware of what my participation will involve, and I may contact the primary researcher to clarify any questions that I have.
- My data will be stored indefinitely on a secure server. My anonymised data will be made accessible to other institutions for research and training purposes exclusively, and never for investigation or conviction. Access to that database may be monetised. My name and other identifying details will not be shared with anyone.
- My participation is voluntary and I may withdraw from the research at any time,

without giving any reason. I have the right to have my data deleted at any time, without giving any reason, provided I have provided identification and contact details at the time of the data collection. My contact details will not be shared with or sold to any third party.

- My anonymised data will be kept in countries with similar data protection standards than that of the European Economic Area (EEA).
- The overall findings from this research may be submitted for publication in a scientific journal, or presented at scientific conferences. Any data presented will be anonymised.
- I can request additional information a) about this project by contacting the primary researcher or the researchers supervising this project at the email addresses provided above; b) or about the ForePrint project by contacting contact@foreprint.org or visiting www.foreprint.org.

I understand the above statements and am giving my consent for my data to be used for the following purposes:

- for this research project Yes ☐ No ☐
- shared with academic institutions for research and development purposes Yes ☐ No ☐
- shared with police forces for training purposes Yes ☐ No ☐
- shared with companies for R&D purposes Yes ☐ No ☐

All questions that I have about this research have been satisfactorily answered. I agree to participate.

Date DD/MM/YYYY

Participant's name

Participant's signature

APPENDIX C: Data sheet

Disclaimer: Nothing is mandatory for you to fill in, although any information provided is helpful. Any field which is not filled in will be marked as *"Refuse to disclose"*. Your contact information will not be disclosed to anyone except the researchers in charge of this data collection. It will be privately stored on the database so that: 1) you may request to have your data deleted from it; and 2) you may be contacted for future data collections, should you choose to opt into this.

Identification details

First name	John	Surname	Doe
Birthdate	DD/MM/YYYY	Biological sex	M ◯ F ◯
Email address:			

Shared information

Height	180cm	Weight	75kg
Ethnicity			
Mother's country of bi		Father's country of bir	
Nationality		Current country of reside	

Additional information:

Non-intrusive, relevant information that impacts the appearance of your fingerprints, and that you are confortable to disclose, e.g. scar or skin disease on the hands.

To be filled out by the researcher

Date of deposition	DD/MM/YYYY	Reference number #	

Photo number range

Participated in all experiments?

Comments:

Comments about the fingerprint development process.

Submit

APPENDIX D: Ethics review problems for PGR research

Lam, Jessica F.

From: Sikanyiti, Mable N.
Sent: January 16, 2018 6:32 AM
To: Baddiley-Davidson, Kathy; ml-pgr-ClgSocSciArtsHum
Subject: Re: Ethics review problems for PGR research.

For me, the completeness percentage has never gone beyond 77% and therefore, the submission button has never appeared. In this case, i am unable to submit the application. However, I feel I have attached all the needed documents and have filled in all the sections. In addition, the completeness of all sections is 100% but the overall ends at 77%. Kindly help as I am now behind schedule because of this.
Mable
119045122

From: Baddiley-Davidson, Kathy
Sent: 09 January 2018 14:34
To: ml-pgr-ClgSocSciArtsHum
Subject: Ethics review problems for PGR research.

Dear all,

I am writing to you in my capacity as the PGR Rep in the Postgraduate Research Policy Committee ("PRPC").

The PRPC has had a number of reports of PGR researchers having difficulties with the ethics review system, whether because of delays or because of changes being requested that were surprising. The PGR Co-Directors and the Head of the Doctoral College are investigating the issue, with a view to determining what problems may exist and how they should best be resolved.

However, in order to report the issues and address the problems efficiently the Committee must have a clear picture of what the problems are (or aren't) with the ethics review system. The current goal for the ethics review process is that it should be completed within 3 weeks of the supervisor approving the online application. So an idea of how often that standard is or isn't met would be useful. In addition, it would be useful to know of any specific problems that you may have experienced.

Therefore I would kindly ask you to report to me any of the problems you might have in regards to the abovementioned issue in order for me to report these back to the PGRPC, so they can be incorporated into the review. The next PGRPC is on this coming Tuesday (16 January), so replies before that date would be appreciated.

Kind regards

Anna Liza Kyprianou Spiliakou
Graduate Teaching Assistant

APPENDIX E: Latent fingerprint development methods

Method	Glass	Metal	Painted surface	Plastic	Paper and cardboard	Wood	Leather	Human skin	Wet s.	Multi-colour s.	Tape
D.F.O.					Bonded/cheques, paper money						
Ninhydrin			Flat paints, latex[2]	Foamed	Bonded/cheques, paper money, kraft paper	Oiled, unfinished[5]				Yes[9]	
Indanedione (and zinc chloride)					Bonded/cheques, paper money, kraft paper)						
Physical developer				Soft	Bonded/cheques, wet, paper money, kraft paper				Yes		
Cyanoacrylate fuming		Yes		Hard surfaces, soft, vinyl, cellophane, foamed, arborite	Waxed cardboard	Waxed or polished	Yes			Yes[9]	
Powders[3]	Clean, greasy		Hard enamels, gloss and semi-gloss acrylic, flat paints & latex	Hard surfaces, soft, vinyl, cellophane, foamed, arborite	Waxed cardboard	Waxed or polished, oiled, varnished/painted, unfinished	Yes				
Powder suspension		Yes[7]		Soft, vinyl, foamed	Waxed cardboard	Waxed or polished, oiled[8]	If wet		Yes		
Iodine fuming[11]			Flat paints, latex		Waxed cardboard	Unfinished		Yes[10]			
Camphor smoke		Yes									
Gun blueing		Yes									
Muriatic acid[6]					Thermal paper						
Crystal violet or gentian violet[12]											Adhesive side
Sticky-side powder											Adhesive side
Titanium dioxide											Adhesive side
Vacuum metal deposition	Yes	Yes	Yes	Yes		Yes	Yes			Yes	Non-adhesive side

[2] though it causes damage by dying the area purple.
[3] Metallic (aluminium based), granular(black, white, grey, fluorescent) or magnetic.
[5] Use a spray.
[6] Hydrochloric acid.
[7] Thought it may remove or damage the impression.
[8] May not work.
[9] With fluorescent dyes.
[10] With silver plate method.
[12] Possibly carcinogenic.

Table E.1: Overview of latent fingerprint development methods and the associated substrates on which they are best used. [203, 114]

APPENDIX F: Patent fingerprint development methods

Method	Porous	Semi-porous	Non-porous
Amido black[1]			
Hungarian red[1,2]			
Leucomalachite green[1,3]			
Ninhydrin			
D.F.O.			
Indanedione (w. ZnCl)			
Acid Yellow 7 (AY7)			

[1] Is destructive.
[2] Fluoresces.
[3] Suspected to be carnicogenic.

Table F.1: Overview of patent fingerprint development methods and the type of surfaces for which they are adequate [203, 114, 115].

APPENDIX G: Function properties for ridge modeling

G.1 The rectangular sigmoid function

Let us consider the sigmoid function, also referred to as the logistic function, as the function s defined as such

$$s : x \mapsto \frac{1}{1+e^{-x}}, \tag{G.1}$$

for any $x \in \mathbb{R}$. The rectangular sigmoid function is defined as the function f

$$f : x \mapsto s\big(\lambda(x+h)\big) - s\big(\lambda(x-h)\big), \tag{G.2}$$

where $\lambda \in \mathbb{R}^{+*}$.

Property G.1:

$$s(x) - \frac{1}{2} = -s(-x) + \frac{1}{2}. \tag{G.3}$$

Proof.

$$\begin{aligned} s(x) - \frac{1}{2} &= \frac{1}{1+e^{-x}} - \frac{1}{2} \\ &= 1 - \frac{e^{-x}}{1+e^{-x}} - \frac{1}{2} \\ &= \frac{1}{2} - \frac{1}{1+e^{-x}} \\ s(x) - \frac{1}{2} &= -s(-x) + \frac{1}{2}. \end{aligned} \tag{G.4}$$

□

Property G.2:
The function f is even,

$$\forall x \in \mathbb{R},\ f(-x) = f(x). \tag{G.5}$$

Proof. This is a direct consequence of Property G.1 and the chain rule. □

In order to study the behaviour of f, let us study the derivatives of s and f.

$$s'(x) = \frac{e^{-x}}{(1+e^{-x})^2} \tag{G.6}$$

$$= \frac{1}{\left(e^{\frac{x}{2}} + e^{-\frac{x}{2}}\right)^2} \tag{G.7}$$

$$= \frac{1}{4\cosh^2\left(\frac{x}{2}\right)} \tag{G.8}$$

$$s'(x) = \frac{1}{2\big(1+\cosh(x)\big)}. \tag{G.9}$$

Therefore,

$$f'(x) = \frac{1}{2\Big(1+\cosh\big(\lambda(x+h)\big)\Big)} - \frac{1}{2\Big(1+\cosh\big(\lambda(x-h)\big)\Big)}. \tag{G.10}$$

It is now possible to study the roots of f by studying the numerator of the above expression,

$$\begin{aligned} f'(x) = 0 &\Leftrightarrow 1+\cosh\big(\lambda(x-h)\big) - 1 - \cosh\big(\lambda(x+h)\big) = 0 \\ &\Leftrightarrow \sinh(\lambda x)\sinh(-\lambda h) = 0, \end{aligned}$$

which occurs either if $h = 0$, a trivial case where $f = 0$, or when $x = 0$, which is the only extremum of f. Additionally, the above numerator of f' determines its sign (since its denominator is a product of strictly positive expressions). Thus, f' is positive on $\mathbb{R}^-$, and negative on $\mathbb{R}^+$, which proves that f reaches its maximum in 0, and it can be calculated as follows

$$\begin{aligned} \max_{x\in\mathbb{R}} f &= f(0) \\ &= s(\lambda h) - s(-\lambda h) \\ \max_{x\in\mathbb{R}} f &= 2s(\lambda h) - 1. \end{aligned} \tag{G.11}$$

G.2 The 2-rectangular sigmoid function

Let us now denote by f the 2-rectangular sigmoid function which is defined, for any $x \in \mathbb{R}$ as

$$f : x \mapsto s\big(\lambda(x+a+h)\big) - s\big(\lambda(x+a-h)\big) + s\big(\lambda(x-a+h)\big) - s\big(\lambda(x-a-h)\big), \tag{G.12}$$

where $\lambda \in \mathbb{R}^{+*}$.

Given Equation (G.6), we have

$$\begin{aligned} f'(x) = &\frac{1}{2\Big(1+\cosh\big(\lambda(x+a+h)\big)\Big)} - \frac{1}{2\Big(1+\cosh(\lambda\big(x+a-h)\big)\Big)} \\ &+ \frac{1}{2\Big(1+\cosh\big(\lambda(x-a+h)\big)\Big)} - \frac{1}{2\Big(1+\cosh(\lambda\big(x-a-h)\big)\Big)} \\ f'(x) = &\frac{1}{2}\left(\frac{1}{g(x+a+h)} - \frac{1}{g(x+a-h)} + \frac{1}{g(x-a+h)} - \frac{1}{g(x-a-h)}\right), \end{aligned} \tag{G.13}$$

where $g : x \mapsto 1+\cosh(\lambda x)$. Additionally, given that g is a positive function, the zeroes of f' are determined by the numerator of the above expression, namely

$$\begin{aligned} f'(x) = 0 \;\Leftrightarrow\; &\big(g(x-a-h) - g(x+a+h)\big)\, g(x+a-h)\, g(x-a+h) \\ &+ \big(g(x+a-h) - g(x-a+h)\big)\, g(x+a+h)\, g(x-a-h) = 0. \end{aligned} \tag{G.14}$$

Let us now remark that

$$\begin{aligned} g(x+a) - g(x-a) &= \cosh\big(\lambda(x+a)\big) - \cosh\big(\lambda(x-a)\big) \\ &= 2\sinh(\lambda x)\sinh(\lambda a), \end{aligned} \tag{G.15}$$

and

$$\begin{aligned} g(x+a)\,g(x-a) &= \big(1 + \cosh(\lambda x)\cosh(\lambda a) + \sinh(\lambda x)\sinh(\lambda a)\big) \\ &\quad \big(1 + \cosh(\lambda x)\cosh(\lambda a) - \sinh(\lambda x)\sinh(\lambda a)\big) \\ &= \big(1 + \cosh(\lambda x)\cosh(\lambda a)\big)^2 - \big(\cosh^2(\lambda x) - 1\big)\sinh^2(\lambda a) \\ &= \cosh^2(\lambda x)\big(\cosh^2(\lambda a) - \sinh^2(\lambda a)\big) + 2\cosh(\lambda x)\cosh(\lambda a) \\ &\quad + 1 + \sinh^2(\lambda a) \\ g(x+a)\,g(x-a) &= \big(\cosh(\lambda x) + \cosh(\lambda a)\big)^2. \end{aligned} \tag{G.16}$$

Equation (G.14) can now be rewritten as

$$\begin{aligned} f'(x) = 0 \Leftrightarrow \sinh(\lambda x)\Bigg[&\sinh\big(\lambda(a-h)\big)\Big(\cosh(\lambda x) + \cosh\big(\lambda(a+h)\big)\Big)^2 \\ &- \sinh\big(\lambda(a+h)\big)\Big(\cosh(\lambda x) + \cosh\big(\lambda(a-h)\big)\Big)^2\Bigg] = 0. \end{aligned} \tag{G.17}$$

The first solution to this corresponds to $\sinh(\lambda x) = 0$, which occurs when $x = 0$. Next, the second term in Equation (G.17) is either strictly positive or strictly negative unless $\sinh\big(\lambda(a+h)\big)$ and $\sinh\big(\lambda(a-h)\big)$ are of the same sign, which is equivalent to $a > |h|$ or $a < -|h|$. In this case, this second term can be rewritten as

$$\begin{aligned} &\sqrt{\Big|\sinh\big(\lambda(a-h)\big)\Big|}^2\Big(\cosh(\lambda x) + \cosh\big(\lambda(a+h)\big)\Big)^2 \\ &- \sqrt{\Big|\sinh\big(\lambda(a+h)\big)\Big|}^2\Big(\cosh(\lambda x) + \cosh\big(\lambda(a-h)\big)\Big)^2. \end{aligned} \tag{G.18}$$

Due to the strict positivity of the squared $\cosh$ terms, Equation (G.18) has the same roots as

$$\begin{aligned} &\sqrt{\Big|\sinh\big(\lambda(a-h)\big)\Big|}\Big(\cosh(\lambda x) + \cosh\big(\lambda(a+h)\big)\Big) \\ &- \sqrt{\Big|\sinh\big(\lambda(a+h)\big)\Big|}\Big(\cosh(\lambda x) + \cosh\big(\lambda(a-h)\big)\Big) \end{aligned} \tag{G.19}$$

Finally, the roots of Equation (G.19) are given by

$$\cosh(\lambda x) = c(\lambda, a, h) = \frac{\cosh\big(\lambda(a+h)\big)\sqrt{\Big|\sinh\big(\lambda(a-h)\big)\Big|} - \cosh\big(\lambda(a-h)\big)\sqrt{\Big|\sinh\big(\lambda(a+h)\big)\Big|}}{\sqrt{\Big|\sinh\big(\lambda(a+h)\big)\Big|} - \sqrt{\Big|\sinh\big(\lambda(a-h)\big)\Big|}} \tag{G.20}$$

Given that $\lambda > 0$, and assuming that $h > a > 0$, f' has a single root in 0. Moreover, Equation (G.17) gives us that f' is positive on $\mathbb{R}^-$ and positive on $\mathbb{R}^+$. Therefore, similarly to the rectangular sigmoid function, the 2-rectangular sigmoid function has a single

maximum in 0 of

$$f(0) = 2s\big(\lambda(a+h)\big) - 2s\big(\lambda(a-h)\big). \tag{G.21}$$

In the case where $a > h > 0$, g' has three zeroes in 0 and $\pm x^*$, where $x^* = \operatorname{arccosh}\big(c(\lambda, a, h)\big)$. Additionally, it follows from Equation (G.17) that g' is positive on $]-\infty, -x^*[$, negative on $]-x^*, 0[$, positive on $]0, x^*[$ and negative on $]x^*, \infty[$. Therefore, in this case, the 2-rectangular sigmoid function has its two maxima in $\pm x^*$ and a local minimum in 0.