



Biometric Recognition of the Hand in Unconstrained Images

Thesis submitted to the University of Nottingham for the degree of
Doctor of Philosophy

September, 2025

Gabrielle Hornshaw
14314359

Supervisors: Steven Furnell, Michael Pound, Oliver Butler

School of Computer Science
University of Nottingham

Abstract

The aim of this thesis is to progress the use of hand recognition in forensic contexts to aid in intelligence gathering and identification in perpetrator-recorded images depicting criminal activity (such as sexual abuse). Most developments in hand recognition to date have been evaluated on highly constrained hand image datasets with little variation in factors such as hand pose and illumination, which poorly represents the images that would be seen in a forensic context.

In this thesis, a new hand image dataset, named VarHands, with 6436 images of 40 hands from 20 subjects is collected. In it, each hand is featured under four illuminations with large variety in hand pose. A subset of hand recognition methods identified from the literature are evaluated against it to determine which current approaches are best suited to the unconstrained domain. Numerous pre-processing and model training techniques are proposed to improve cross-domain performance when moving from highly constrained training datasets to the unconstrained evaluation data: region-of-interest extraction and matching outperformed whole-hand matching by 3.8% rank-1 accuracy on 11k Hands, and luminance-based illumination-normalisation achieved a 39.06% rank-1 accuracy improvement on the new collected dataset. Further to this, uncertainty quantification is explored as a way to handle frequent mis-detections, blurring, and occlusion found in unconstrained images, with the proposal of a new uncertainty-aware method of multi-biometric fusion which achieves +0.6% rank-1 accuracy. The results show strong potential of application but highlight the need for more data from which to learn representative distributions of hand features.

The findings from this thesis provide avenues of research to develop hand recognition algorithms that are invariant to hand pose and illumination and thus more suitable for forensic application or other unconstrained domains. To conclude, a review of the regulatory landscape surrounding the application of biometric technologies in law enforcement is performed to identify the requirements for adoption of hand recognition, which further supports the development of larger and more representative hand image datasets.

Acknowledgements

I would like to first thank my supervision team: Steve Furnell, who has been a wonderful mentor throughout this process; Mike Pound, for recommendations on computer vision literature; and Oliver Butler, for his continuous support and guidance as I learn a new topic unfamiliar to me.

I am deeply grateful to the people I have worked with at Leicestershire police for the past two years, especially for the many conversations and suggestions about my thesis which have helped shape it into what it is today. Having the opportunity to do this placement and develop my personal and technical skills has been invaluable.

I am appreciative of the support provided by the staff at Horizon CDT, for the training sessions, help in navigating dataset license agreements and expenses, and for the organisation of regular social events and writing retreats, all of which made the PhD process a lot less painful than it could have been. I also extend my deepest thanks to the everyone who contributed their hand images to the dataset, without whom this project would not have been possible.

For all of their support, distractions, and reassurances, I thank my friends: Iyola, Megan, Robin, Leigh, Fiona, North, Greg, Callum, and Torran. Thank you all for making these last few years so enjoyable. I also thank my family for their love and support: Jessi, Nan, and especially Mum, as it is thanks to your lifelong encouragement that I am in this position today.

Finally, my deepest thanks go to my partner, Emma, who has eagerly offered her help over the years as my emotional support, proofreader, hand model, dataset recruiter, and cheerleader. Thank you for being so wonderful.

Contents

1	Introduction	9
1.1	Motivation	11
1.2	Related Work	11
1.2.1	Biometrics	12
1.2.2	Forensic Biometrics	13
1.2.3	Deep Learning	14
1.3	Aims and Objectives	15
1.3.1	Research Questions	16
1.3.2	Research Objectives	17
1.4	Structure of the Thesis	19
2	Biometric Modalities of the Hand	20
2.1	Palm	22
2.1.1	Datasets	23
2.1.2	Region-of-Interest Extraction	25
2.1.3	Feature Extraction and Matching	27
2.1.4	Performance	28
2.2	Knuckles	29
2.2.1	Datasets	31
2.2.2	Region-of-Interest Extraction	31
2.2.3	Image Pre-Processing	33
2.2.4	Feature Extraction and Matching	35
2.2.5	Performance	37
2.3	Dorsum	37
2.4	Whole-Hand	39
2.5	Skin Markings and Others	41
2.6	Multi-Biometric Fusion Schemes	43
2.7	Conclusion	45

3	Data and Labelling Strategy	46
3.1	VarHands: A New Hand Image Dataset	48
3.1.1	Collection Protocol	50
3.1.2	From Video to Image Dataset	52
3.1.3	Analysis of Collected Dataset	52
3.1.4	VarHands-Easy and Dataset Splitting	54
3.2	Data Labelling	56
3.2.1	Research Questions	58
3.2.2	Methodology	58
3.2.3	Evaluation Metrics	61
3.2.4	Results	62
3.3	Discussion	66
3.3.1	Dataset	66
3.3.2	Labelling	67
3.3.3	Conclusion	69
4	Region Extraction and Illumination Invariance of Hand Images	70
4.1	Research Questions	71
4.2	Background	72
4.3	Image Pre-Processing	73
4.3.1	Hand Landmark Detection	73
4.3.2	Whole-Hand	74
4.3.3	Palm Extraction	75
4.3.4	Knuckle Extraction	76
4.3.5	Illumination Normalisation and Texture Enhancement	80
4.4	Methodology	82
4.4.1	Model Training	83
4.5	Results	84
4.5.1	Ranked Retrieval	84
4.5.2	Impact of Illumination and Hand Pose	88
4.6	Discussion	93
4.7	Conclusion	95
5	Uncertainty in Hand Recognition	97
5.1	Scenario	100
5.2	Research Questions	102
5.3	Background	103
5.3.1	Aleatoric and Epistemic Uncertainties	103

5.3.2	Uncertainty Quantification in Machine Learning	104
5.3.3	Fusion of Uncertain Information	108
5.4	Methodology	109
5.4.1	Grey-Level Co-Occurrence Matrix Statistics	110
5.4.2	Auto-Encoder Loss	111
5.4.3	Monte-Carlo Dropout	112
5.4.4	Deep Ensemble	112
5.4.5	Bayesian Triplet Loss	113
5.4.6	Multi-Model Metric Fusion	113
5.4.7	Expected Calibration Error	114
5.5	Results	115
5.5.1	Qualitative Samples	115
5.5.2	Uncertainty Fusion	123
5.5.3	Ranked Retrieval Performance	125
5.6	Discussion	127
6	Hand Recognition in Practice	130
6.1	Research Questions and Contributions	131
6.2	Definitions	132
6.3	Regulatory Scope	134
6.4	High-Risk AI	138
6.5	Requirements for the Proposed System	142
6.5.1	Legal Basis	142
6.5.2	Databases	145
6.5.3	Competency and Training	147
6.6	Conclusion	148
7	Conclusion	150
7.1	Connection of Studies	151
7.2	Research Questions	152
7.3	Contributions and Limitations	157
7.4	Future Work	158
7.5	More Data as a Priority	160
A	VarHands Recruitment Materials	178

List of Figures

2.1	Biometric Modalities of the Hand.	21
2.2	Examples of Palm Images.	23
2.3	Examples of Knuckle Images.	30
2.4	Examples of Other Hand Biometrics.	42
3.1	Twenty-five images each from the 11k Hands and VarHands datasets.	47
3.2	Distribution of factors in the VarHands dataset.	53
3.3	Varhands-Easy selection.	55
3.4	Test dataset split into probes and gallery.	56
3.5	Training behaviour using the four labelling strategies.	64
3.6	Rank accuracy and precision-recall of four labelling strategies on 11k Hands.	65
4.1	Hand landmarks detected using the MediaPipe library and extracted regions.	74
4.2	Palm projection process.	76
4.3	Knuckle sizing process.	78
4.4	Knuckle centre localisation process.	79
4.5	Image processing methods applied to two images of the same knuckle.	81
4.6	Rank-accuracy and precision-recall graphs of regional methods.	85
4.7	Visualisation of embedding space highlighting the effect of illumination.	90
4.8	Ranked accuracy of colour and greyscale illumination-normalised knuckle models.	91
4.8	Precision-recall curves of colour and greyscale illumination-normalised knuckle models.	92
5.1	Sources of uncertainty in hand images.	98
5.2	Examples of training dataset knuckle images.	100
5.3	Examples of test dataset knuckle images.	101
5.4	Multi-biometric fusion with uncertain samples.	102

5.5	Change in measured uncertainty with increasing quality degradation.	116
5.6	Lowest and highest five uncertainties produced by each method in the 11k Hands dataset.	119
5.7	Lowest and highest give uncertainties produced by each method in the VarHands dataset.	120
5.8	Anomaly-annotated TSNE plot of dorsal knuckle feature embeddings.	123
5.9	An example of multi-biometric fusion with five uncertainty measures.	124

List of Tables

2.1	Palm Datasets.	24
2.2	Knuckle Datasets.	32
3.1	Eight video categories provided by each participant.	50
3.2	Label categories.	59
3.3	Ranked-retrieval results for each subset-trained model on 11k Hands subsets.	66
4.1	Accuracy and mean-average precision of regional methods on 11kHands.	86
4.2	Accuracy and mean-average precision of regional methods on VarHands.	86
4.3	Accuracy and mean-average precision of regional methods on VarHands- Easy.	86
5.1	Ranked retrieval results for different multi-biometric fusion methods.	125
5.2	Ranked retrieval results for different uncertainty quantification methods.	127

Chapter 1

Introduction

This thesis covers the topic of hand recognition in unconstrained images for the purpose of forensic investigation of criminal image and video material. Criminal image material is inclusive of both explicitly criminalised images or images depicting criminal acts: physical abuse, sexual abuse, animal abuse, harassment, terrorist propaganda, 'trophy shot' images, or images in which a suspected criminal displays illegal items like guns, drugs, or stolen items. This thesis is concerned in particular with perpetrator-recorded abuse material (as opposed to third-party recorded, such as CCTV footage). Identifying the perpetrators in this material is often very difficult as they take steps to conceal their identity by hiding their faces, voices, location information, and any identifying tattoos or marks; what *can* be visible in some of this material is their hands [150]. While not universal, and eventually circumventable (perpetrators might start wearing gloves), the presence of hands presents an opportunity to gather information on the human subjects of the images, which could be used to link cases to other cases, link cases to suspects, or potentially identify victims.

The topic of this thesis was chosen following several calls-for-research from law enforcement agencies indicating the potential for hand-based recognition to afford valuable information in investigations of criminal imagery, primarily physical abuse or child sexual abuse material (CSAM) [150]. In the UK, (human analysis of) hand biometrics has led to the positive identification and conviction of offenders [11], and there have subsequently been research efforts to collect large volumes of hand image data, quantify the variance of hand features in the population, and develop automated methods of hand recognition [64]. Improved investigative efficiency via hand biometrics could help alleviate some burden on the human analysts and police officers who are dealing with an ever-increasing amount of harmful online material [42], from which it is becoming increasingly difficult to identify perpetrators [43].

This is an interdisciplinary thesis topic that touches on a lot of different research fields. The background and motivation are largely supported by contributions from criminology, forensic biometrics, and forensic anthropology (hand anatomy). The actual work done has a strong basis in both cyber security (biometrics) and computer vision, as well as covering more broad concepts of deep machine learning and information retrieval systems, and Chapter 5 delves into the combination of probability theory and machine learning through model uncertainty quantification. The final chapter then contains an analysis of the regulatory area around law enforcement use of biometrics. To support this, this thesis has had supervisory input from the schools of cyber security, computer vision, and law (data protection focus); as well as external support from an industry partnership and work placement with a local police force.

Following a literature review into hand biometrics with specific attention to those applied in forensic or similar unconstrained contexts, research gaps were identified around the use of unconstrained data in developing and testing feature extraction and matching models. Significantly, almost no papers at the time of the review evaluated their methods on data that varied in hand pose or illumination, even while claiming forensic application as their goal. As a result of this, there is no understanding of which approaches to hand recognition are best suited for the forensic context or any studies into methodology which may improve performance in unconstrained images.

Working within the research gap identified, the aim of this thesis is primarily to evaluate existing methods of hand recognition (especially those claiming to be useful for forensic or unconstrained applications) on new unconstrained hand image data, to determine which approaches are truly effective and transferable. This requires a preliminary objective to collect a dataset of hand images that replicates the conditions of unconstrained images; it should contain a suitable number of human participants while also containing as many images as possible of each participant in a non-controlled range of hand poses, illuminations, and image qualities. Following the establishment of which existing methods are most suitable for the unconstrained domain, this thesis will look to evaluate the instances of failure (where correct identification is not achieved) in the new data and then develop, from other related feature-extraction work, methods that can achieve higher performance. As an example, one expected area for improvement is illumination-invariance. To finish the thesis, a review of the regulatory landscape will be performed to identify the requisite steps to implementation and use of hand recognition for law enforcement in practice.

1.1 Motivation

The quantity of sexual abuse imagery reported online has been rising by approximately 5-20% year-on-year for over a decade [41], and the proportion of the more serious categories been growing [41]. The Internet Watch Foundation, which has been hashing (uniquely encoding data) child sexual abuse material (CSAM) since 2015, assessed 734,048 new unique criminal images and videos in 2024, of which 29% were of the most serious category [42]. In shifting proportions year to year, approximately 25% of these images are thought to be non-self-generated and feature the abuser [43]. The severity of the the creation of this content and the importance of identifying and bringing to justice the abusers cannot be understated. In the period of 2014-2018 there was a sharp increase in the number of CSAM-related cases in the UK closed because of an inability to identify the offender - from 145 (4%) to 1690 (12%); the Police Foundation attribute these low detection rates both to the increased strain on police forces (higher and more complex caseloads without larger workforces) and to the technical capabilities of offenders [43]. They recommend that more should be done to proactively investigate and collect intelligence on these hard-to-catch offenders (without conflating hard-to-catch with highest-risk).

Researchers for the FBI’s Forensic Audio, Video and Image Analysis Unit stated that “hands often appear prominently in [CSAM], as well as ‘trophy shots’ from homicides and kidnappings”, and that automated hand biometric systems would “greatly enhance the efficiency of forensic work” [150]. In the UK, comparison of hand features has resulted in the positive identification and charge of offenders in multiple child abuse cases [11], which has led to the development of a long-term programme to collect hand image data to analyse hand feature variation in the population to support hand biometric evidence in courts [64]. While statistical backing for evidential use is important, development of hand-based automated recognition technologies is also underway, and in the last 5 years there have been many research papers in the domain citing forensic application as motivation (covered in Chapter 2). All but three papers from one research group use constrained image data, much unlike what would be seen in a forensic context, to evaluate their methods’ performance, so this project will address the gap and focus specifically on unconstrained hand recognition.

1.2 Related Work

The topic of unconstrained hand recognition for forensic or law enforcement purposes intersects a large number of fields, most closely: forensic biometrics, cyber security

(computational biometrics), computer vision, information retrieval, criminology, and law. Brief backgrounds to these fields and their relevance to the thesis are given in the following sections. Where possible, general examples and scenarios are related back to hand recognition.

1.2.1 Biometrics

Biometrics are the measured physical and behavioural traits of human beings. Traits must meet some criteria in order to be considered suitable for recognition (authentication, verification, or identification), including *universality* (every person has this trait), *uniqueness* (no two people share the same measurement), and *permanence* (the trait does not change over time and cannot be easily intentionally changed) [67]. Biometric technologies apply automated processing and statistical analysis to biometric traits in order to perform tasks such as authentication (the verification that a person is who they are claiming to be), identification (the discovery of an identity from matching of traits), or clustering (the conflation of multiple samples as belonging to the same identity). In the development of biometric systems, several other criteria are often considered to determine if a trait will be a good choice: the efficiency of measuring and comparing that trait, the accuracy that can be achieved with that trait (while it may be physically unique, there is often a loss of granularity when capturing into a digital format), how spoofable that trait is, and whether people would find it acceptable to provide a sample of that trait in the circumstances [67]. Traits commonly used in biometric technologies include face, fingerprint, iris, voice, signature, keystroke behaviour, mouse movements or screen behaviour, hand geometry, and palmprint. Some less common (but still with a substantial amount of research behind them) include retina, ears, gait, body odour, knuckle print, fingernails, and dorsal hand veins [10]. Many of these traits are found on or within the hand, and are discussed in far greater detail in a deep literature review into hand biometric traits in Chapter 2.

The most prominent application of biometric technologies is in authentication, in which they are incorporated into wider digital systems for security and access control. Systems are often built with specialist sensors (e.g. infrared) that obtain high-quality samples, and because enrolment is voluntary, constraints can be imposed on the subject and samples can be re-collected as necessary if they are low-quality or invalid. This means that many biometric methods used are not very applicable to domains of unconstrained biometrics, which includes forensic applications, because they rely on the conditions in high-quality and controlled samples. In recent years however, research has developed towards more unconstrained biometrics as user-friendly and

seamless integrations into smart devices, such as facial or palm recognition used to unlock mobile phones [19].

1.2.2 Forensic Biometrics

Forensic biometrics is the intersections of the fields of forensic science and biometrics. Forensic science, per Meuwly and Veldhuis, is concerned with making inferences about *traces* (a trace could be, e.g., a photograph of a hand whereas the *source* is the actual hand) to answer questions related to a criminal offence or to individualise a perpetrator (among other things) [117]. Biometrics is the statistical analysis of distinct human traits in order to recognise human beings. Meuwly and Veldhuis describe specifically four ways biometrics can be used to aid forensic investigations: identification of missing persons; forensic investigation (searching a database for likely matches to a trace); forensic intelligence (associating cases based on traces likely coming from the same origin); and forensic individualisation (using biometrics to prove that the suspect source is the only possible source of a trace) [117]. Inferences made in forensic science must be logically sound and hold up to scrutiny in a court. To be admissible in a court of law then, individualisation based on biometrics must be based on solid statistical properties, and not simply the output of an algorithm. Meuwly proposed the likelihood ratio (based on Bayes theorem) method of evaluation of biometric individualisations that considers the within-source variability of the source (e.g. how much a single person’s hand can vary - such as from weight gain/loss, scarring, ageing or more) and the between-source variability of the *trace* (e.g. how much photographs of hands vary irrespective of how much the hands themselves vary) [115]. This is on the basis that uniqueness of trace cannot be assumed from the uniqueness of the source (loss of granularity through capture into a digital format could result in, e.g., two different and physically unique hands appearing identical in images).

To be able to produce the statistics necessary to perform evidence-based individualisation from biometrics, Muewly et al. suggest that datasets must both contain reference samples (e.g clear hand images) of the suspect and population (a large number of people), and trace samples taken in the same conditions as the trace (e.g. the same hand pose and illumination) [116]. Additionally, the choice of measured features is very important, since something such as finger length would be far less population-variable than something such as knuckle print. Jain and Ross suggest the use of multiple feature sets or encodings for biometric comparison including both hand-crafted features and learned features [69]. Using deep neural networks to learn complex features in images can greatly improve recognition performance, particularly

on more unconstrained or low-quality forms of data, however they are more ‘black box’ and less likely to hold up as evidence.

Such robust methodology (incorporating the likelihood evaluation into algorithms) may not be necessary for intermediary tasks in the course of an investigation, such as searching a database for likely matches or associating cases based on shared traces. For many of the commonly used biometrics today, such as face and fingerprint, this is the case: biometric algorithms are used as tools to search databases and the outputs are verified by human analysts, then evidence on biometric measurements and statistical likelihoods of individualisation are produced by a human expert without algorithmic input. The *investigative* application of biometrics and the *evidential* application of biometrics are distinct areas with distinct requirements. In this thesis, the focus is on the investigative application, for which speed and accuracy are priorities for algorithm building, with the understanding that some slight error is permissible (where it would not be for evidential application).

1.2.3 Deep Learning

The most widely researched domain of image recognition is facial recognition, and it has been dominated by deep learning since 2014 when DeepFace [156] achieved near-human level performance on the Labelled Faces in the Wild (LFW) dataset with a pair-wise matching accuracy of 97.34%, far surpassing the performance of hand-crafted local image descriptors (SIFT, LBP, etc.) that had been common prior. The DeepFace model was a 9-layer neural network with 120 million parameters, and was trained using the Social Faces Classification dataset from Facebook featuring 4.4 million images of 4030 subjects. Since that development, facial recognition research had been focused primarily on three areas: the development of better, often deeper backbone networks, the development of loss functions and distance measures to produce better feature embeddings, and the development of pre-processing techniques that reduce variation in the data, though the latter has mostly been made redundant by newer deep learning techniques that are robust to data variation [166].

Backbone networks map an input image to a feature embedding. A number of well-performing architectures have been used for image recognition tasks, including AlexNet [90], Inception/GoogLeNet [155], VGG-Net [147] and ResNet [56] (the most used of the lot, because it enables stable training of very deep networks with high accuracy). These common architectures are convolutional neural networks (CNNs); since the beginning of this project in 2021, vision transformers (ViT) [33] have become a popular alternative to CNNs for many image processing tasks and have achieved higher accuracies than CNNs on tasks like facial recognition [140], with the

downside of requiring substantially more training data than typical CNNs. Optimal training dataset sizes for CNN architectures are not small either, and vary depending on the exact architecture (number of parameters) and complexity of the task; for open-set recognition tasks such as facial recognition, training datasets are massive - DeepFace was trained with 4.4 million images, and Google’s FaceNet with 260 million [142], though they saw only a 1% drop in performance going down to 26 million and a more significant drop after that.

Loss functions measure the network output by some criteria during training and affect how much the network weights are then updated. In image recognition there are typically two categories of loss functions: classification losses, in which the model’s ability to map every class in the dataset is measured, and verification losses, in which the model’s ability to differentiate embeddings of different classes is measured. Verification losses, also called contrastive losses, involve comparing the embeddings of two inputs and penalising short distances between different classes and large distances between same classes. From the basic 2-input contrastive loss evolved triplet loss, which considers a positive and negative match at the same time, enforcing a margin between them [142]. Since, a number of loss functions have found marginal improvements on performances >99% accuracy on face recognition datasets, including quadruplet loss [23] which outperforms triplet loss and angular losses that consider hyperspherical representations [108].

Large-cohort open-set recognition problems, as forensic biometric systems would be, are ill-suited to classification losses which seek to indirectly learn feature embeddings through their transformation into class probability vectors (sub-optimal for training datasets of thousands-millions of identities) [48]. Contrastive losses can handle large cohorts better as they directly optimise the feature embeddings to have the best separation between classes. In image retrieval systems, as are considered the application scenario for this project, feature embeddings are used directly to rank or cluster a gallery (search database), so for this reason only contrastive loss is used in this project. However, it should be noted that some studies indicate better results when using a combination of the two, where classification losses allow learning of good inter-class representations and verification losses improve intra-class representation [152].

1.3 Aims and Objectives

The overall objective of this thesis is to contribute towards the development of unconstrained hand recognition technologies for their eventual application in forensic contexts such as criminal investigations of abuse material. This application context

is important to consider as it shapes the types of systems the biometric algorithms will be part of, and so informs how the algorithms should be built and evaluated. The type of system considered in scope for this project is *information retrieval*, wherein an image is used to search a database of images for likely matches. Both the image used as a search and the database are assumed to be unconstrained images of any quality, and the database could be of any size with any number of true matches, including zero. Different from *identification* systems which are often closed-set and non-anonymous, information retrieval is an open-set recognition problem and possibly anonymous; though the same biometric technologies (the extraction and comparison of biometric features) underlie both systems, their differences can inform modelling methods (the use of contrastive losses over classification losses) and the evaluation strategy. The application context also makes the accuracy and robustness of the technology of critical importance, which should shape the research directions.

Following a literature review a significant research gap has been identified in the development and evaluation of hand recognition algorithms on unconstrained data. In addition to the lack of unconstrained data, existing hand image datasets are very small by biometric standards, placing strong limitations on not just how accurate and generalisable trained models can be, but also in how reliable their measured performance on test data is, which is a barrier to use for high-risk AI systems. There has also not yet been an exploration of the legal framework around hand recognition technologies for law enforcement use in the UK. These gaps present several areas for development in this thesis: firstly, the collection of unconstrained hand image data that could be used to more reliably evaluate hand recognition algorithms; secondly, the investigation of the impacts of unconstrained image conditions and the development of methods to overcome those impacts; and thirdly, identification of the steps needed to be taken for hand recognition technology to comply with the law and be used in practice.

1.3.1 Research Questions

RQ1 How should hand images be categorised and labelled for optimal training and recognition performance?

RQ2 Which whole-hand or regional approaches to hand recognition perform best on unconstrained data?

RQ3 To what degree does change in hand pose impact recognition performance?

RQ4 To what degree does change in hand illumination impact recognition performance?

RQ5 What are the behaviours of uncertainty quantifying methods when applied to hand biometric images of increasing quality loss?

RQ6 What impact can these quantified uncertainties have on recognition performance when incorporated into multi-biometric fusion?

RQ7 What steps are needed for compliance with the relevant regulations before hand recognition technologies can be used for law enforcement purposes?

Seven research questions to be answered were developed iteratively, given in the list above. Research Question 1 followed a literature review of hand recognition methods for which different data labelling strategies were used, and was something that needed to be answered in the early stages to support the labelling strategy chosen for a new collected dataset and later studies. Research Question 2 is again based on the literature review of related methods, and aims to determine which approaches seen in the surrounding literature are best suited to the unconstrained domain, again directing future research. Questions 3 and 4 arose out the desire to understand how the conditions present in unconstrained images might impact recognition performance, with the aim to identify potential techniques to overcome the impact without needing unconstrained training data. Questions 5 aims to discover if uncertainty quantification techniques can give meaningful uncertainty estimates in hand recognition for the purpose of building more robust and interpretable critical AI systems; Question 6 then asks if the uncertainty estimates can be used to increase recognition accuracy in multi-biometric fusion scenarios. Finally, Question 7 is asked to clarify the regulatory landscape around law enforcement use of biometric technologies in relation to hand recognition and identify the steps needing to be taken to make progress on implementing it.

1.3.2 Research Objectives

Several overarching research objectives were defined in order to answer the above Research Questions, given in the list below.

- Review hand recognition literature with particular attention to the pre-processing and feature-extraction methods that may be most applicable to unconstrained images.
- Collect a new dataset of unconstrained hand images with large within-subject variation in hand pose and illumination.
- Experiment with data labelling strategies for hand images and determine the best one.
- Implement hand recognition methods from the literature and evaluate them on the unconstrained dataset, especially with regard to cross-pose and cross-illumination recognition rates.
- Use pose-invariant and illumination-invariant processing techniques to improve performance on the unconstrained dataset.
- Implement uncertainty quantification methods and evaluate them qualitatively on all available data.
- Devise an uncertainty-aware multi-biometric fusion method, then implement and compare it against other fusion methods.
- Read the relevant legislation, court cases, and police policy to identify the barriers to implementation of hand recognition, recommend future directions to address these.

The first objective was to perform an extensive review of the hand biometric literature in order to identify approaches and methods that have potential application in the unconstrained domain, and to understand where the state-of-the-art and baseline performances are. The second, and most important, research objective was to collect an unconstrained hand image dataset to enable the evaluation of these hand recognition methods in the unconstrained domain. Following this, the objectives were to implement the most relevant methods from the literature review and evaluated them on this dataset, to demonstrate empirically a) which labelling strategy is superior, what the effect of illumination variation is on recognition performance, and what the effect of pose variation is on recognition performance. Building off of this evaluative work, the next objectives were to develop methods of region extraction and image pre-processing to improve performance on the unconstrained images, and to implement and evaluate uncertainty quantification methods and explore the potential for a novel uncertainty-aware biometric fusion method. The final objective

of the research was then to review UK and EU legislation and the policies of police organisations, and to make recommendations for the future of unconstrained hand recognition research based on identified requirements. These research objectives are summarised in the list below.

1.4 Structure of the Thesis

This thesis is broken down into seven chapters including this introduction and a final conclusion chapter. Chapters 4 to 6 each cover different studies and so each have their own background, methodology, results, and discussion sections. The unifying areas of background for all chapters is presented here in the introduction and in Chapter 2.

Chapter 2 is an in-depth literature review into existing hand biometrics and recognition methods with a specific focus on methodology that would be most applicable to unconstrained images (either because they were developed for those images, or because there are transferable aspects). This review also serves to identify and assess all of the available hand biometric datasets.

Chapter 3 describes the collection of a new unconstrained hand image dataset named VarHands, its analysis, and a series of experiments to identify the best data labelling strategy for hand images (based on several used in the literature).

Chapter 4 presents a series of experiments that compare performance of whole-hand and sub-biometric hand recognition approaches, as well as some methodology to illumination-normalise the images and enhance skin textures, and the comparison of these in cross-pose and cross-illumination recognition tasks.

Chapter 5 covers the implementation of several anomaly detection and uncertainty quantification methods to attempt to quantify the divergence of an unconstrained dataset sample from the constrained dataset ideal samples, then qualitatively analyses the values produced for dorsal knuckle images. It then presents a novel probabilistic multi-biometric fusion method incorporating this value, and compares this with different uncertainty values and non-probabilistic fusion.

Chapter 6, finally, shows a review of the regulatory landscape around use of biometric data (special category) and biometric technologies (high risk) by law enforcement in the UK and EU, identifying and describing three requirements that need to be fulfilled before hand recognition can be applied legally in a forensic context.

Chapter 2

Biometric Modalities of the Hand

This chapter presents a literature review of biometric modalities related to the hand with a specific focus on unconstrained application, in order to understand the available data, state-of-the-arts in unconstrained recognition, and benchmark performance on unconstrained tests.

A biometric modality, according to the authors of “Biometric Recognition: Challenges and Opportunities” [119], is a “system built to recognise a particular biometric trait”, with biometric traits being a biological or behavioural expression of an individual. The examples of biometric traits given by the authors are face, fingerprint, hand geometry, palm print, iris, voice, signature, gait, and keystroke dynamics. Definitions of biometric trait and biometric modality are malleable, so throughout this project hands are both considered to be one biometric modality, similar to faces, and a collection of more specific biometric modalities such as hand geometry, palm print, knuckle print, and fingerprint, as are displayed in Figure 2.1.

Because there is a large corpus of research in hand biometrics for the purposes of security and authentication, this literature review has been limited to include only those which have potential applicability to the unconstrained domain or which contribute some unique and important development to the field. The ratio of these suitable papers varies by trait according to the available image data. The areas of particular interest that will be focused on are in region-of-interest extraction, image pre-processing or enhancement, and feature extraction, encoding, representation, and matching schemes. Multi-biometric fusion schemes are considered more general and will be discussed separately in Section 2.6. For feature extraction and matching, all methods are considered relevant for this thesis, as region-of-interest extraction and image pre-processing are assumed to handle all variation found in unconstrained images and bring them to the same state used in controlled biometrics. In practice

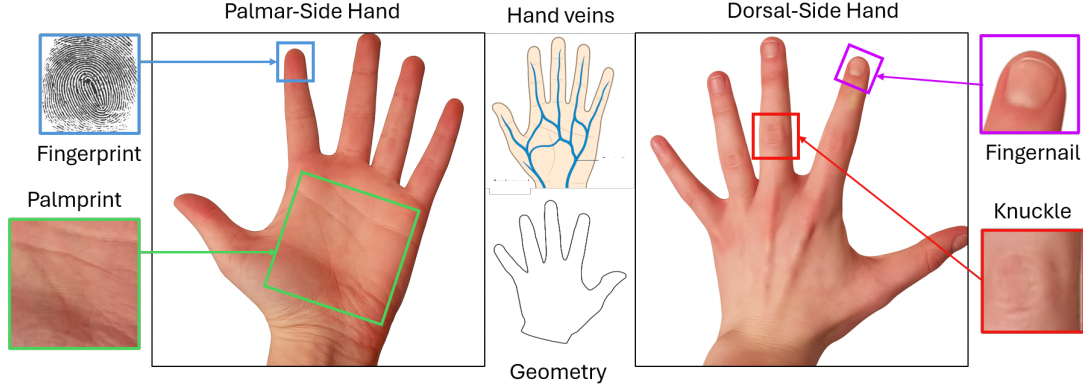


Figure 2.1: Some biometric modalities of the hand, including fingerprint, palmprint, hand veins, hand geometry, fingernail, and knuckles.

this is unlikely to be the case, and methods that are more robust to difference in image quality or small translations and transformations may be best suited to use in unconstrained images.

In this chapter Sections 1 and 2 discuss palm and knuckle biometrics, due to the large amount of research surrounding them. This is followed by two briefer Sections, 3 and 4, on hand dorsum recognition and whole-hand recognition, limited by their applicability and data availability, and Section 5 covering other hand biometric modalities such as fingerprints and soft biometrics. Section 6 then covers multi-biometric fusion schemes applied to combine results from multiple biometric modalities or algorithms.

An additional consideration which is not deeply covered in this review due to lack of supporting research is how the recognition methods handle ageing of the subjects. It is important to evaluate how algorithms perform when presented with images of the same subject at different periods in time, however this is difficult to do without suitable data, which would ideally be collected from the same participants over decades. Uhl and Wild analysed the short-timespan vs long-timespan performance of some hand recognition algorithms with a dataset collected over 5 years [158]. For palmprint, minutiae, and finger-shape based features, ageing caused a statistically significant worsening of the recognition performance measured in AUROC. For Eigen-hand features (hand geometry derived) however they found no impact from ageing. Their findings strongly suggest that ageing of a subject can have a substantial negative impact on recognition performance for a variety of hand biometric modalities, including possibly those they did not test (e.g. knuckles), which is important to keep

in mind going forward.

2.1 Palm

Palmprints consist of three layers of textural information: the principal lines (the most prominently visible lines, of which most hands have three-five), wrinkles, and skin friction ridges (raises and recesses formed to increase contact friction and improve grip when holding objects, like the treads on shoes or tyres). The friction ridges of the finger tips are what form fingerprints, and just like with fingerprints, palm friction ridge skin is unique to each individual, though can be obscured over time by the development of wrinkles. Comparisons between palmprint samples can be used to confirm or for exclude a match [35].

Palm prints can be acquired in a number of ways, including using ink to make an impression on paper (‘latent’ print) or simply holding the palm in physical contact with a special camera sensor (as in the Thupalm lab dataset, for example [27]). This form of palm recognition is not particularly well suited to the unconstrained domain and would only be applicable in situations where there are very close up or high resolution images in ideal illumination. The type of palm images most often found ‘in the wild’ - in uncontrolled digital images - are ‘contactless’. Contactless palm images are taken in most cases with a normal photo camera at some distance to the hand. In these contactless palm images it is often impossible to extract the detail of the skin friction ridges, and so recognition methods rely on more highly visible information such as the palm lines and creases. There is large variation in the quality and suitability of contactless palm image datasets for this form of recognition. Most are collected for the purpose of developing controlled authentication systems and so have the subject and camera in fixed positions with uniform lighting and backgrounds; while this limits their relevance to study unconstrained recognition, they give meaningful insight into the upper performance boundaries of palm-based biometrics in the best possible conditions. Figure 2.2 shows an example of the difference in texture detail available in a latent palmprint and a texture enhanced contactless image.

It should be noted that while ‘palmprint’ should in true definition refer only to latent prints or contact images with clear distinction of the skin friction ridges, the terms ‘palmprint recognition’ and ‘palm recognition’ are used interchangeably in the literature and both frequently refer also to contactless recognition.

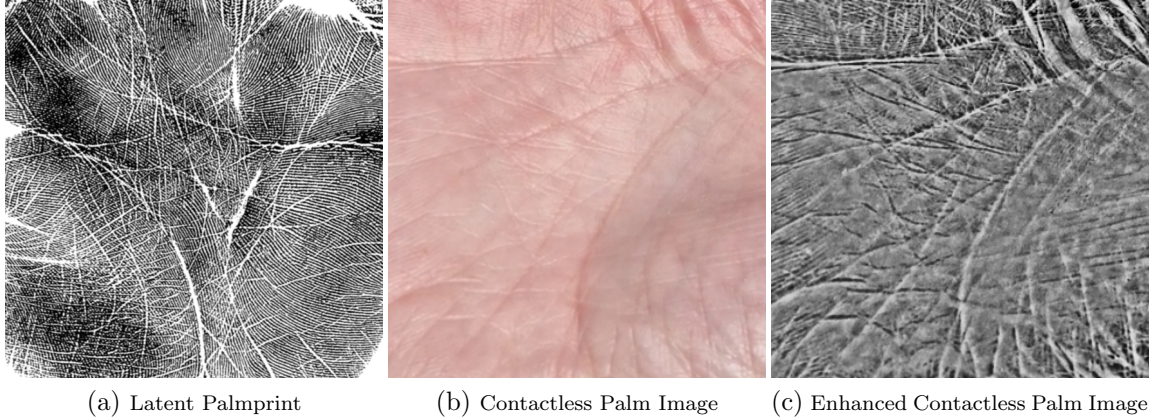


Figure 2.2: Examples of palm images: a) a latent palmprint, b) a palm region extraction from a contactless image, and c) that image with illumination-normalisation and texture enhancement applied. These images demonstrate the difference in textural information available in a latent print (high-density, precise friction ridges visible) and a contactless palm image (lower visibility of friction ridges, more interference from shadows and skin distortion).

2.1.1 Datasets

Palm datasets identified in this literature review are summarised in Table 2.1, consisting of 11 palm-specific datasets and two formed from datasets that contain both dorsal and palmar images of hands: Tecnocampus Hand Image Database [37], and 11k Hands [1]. Information about these datasets, including the number of palms (not participants - sometimes there are two palms per participant), number of total images, and a short description is given in the table columns. Where information was unobtainable the cell is left empty, and where the value is estimated from other available information, for example estimating the number of images per hand from the total number of images and the number of hands (though it may vary), it is prepended by $\sim$ (approximation).

No information on the image factors could be found for NTU Contactless Palms. Out of the remaining 12 datasets, six of them consist of images taken in one session under controlled and uniform lighting. Of the datasets where illumination is variable it varies very little, for some datasets (HIT-NIST) there is lack of clarity, without access, about whether each subject is captured in different illumination or if it just varies between subjects. In the latter case, illumination difference is likely to artificially inflate performance based on false patterns. In nine of the datasets the images

Dataset	Year	No. of Palms	No. of Ims	Ims per Hand	Description
CASIA [153]	2005	624	5502	~9	Greyscale images. Single illumination over simple background. No variation in hand pose and position-in-frame. Single-session acquisition.
IITD Touchless Palmprints [93]	2007	460	3220	7	Single illumination over simple background. Some minor variation in hand pose and position-in-frame.
PolyU Contact-Free 3D/2D Palms [82]	2011	228	570	~3	Single illumination over simple background. Controlled rotation of the hand, no variation in finger position. Single-session acquisition.
Tecnocampus - Palms [37]	2014	~220	~1000	~5	Multi-spectral, single illumination over simple background. No variation in hand pose or position-in-frame. Five-session acquisition over six months.
REST [22]	2016	~300			Single illumination over simple background. Some variation in hand pose (finger spread only) and some variation in position-in-frame. Single-session acquisition.
PolyU-IITD Contactless Palmprints [94]	2018	~1200	~12,000	~10	Some variation in illumination over simple backgrounds. Some variation in hand pose (finger spread only) and some variation in position-in-frame. Single-session acquisition. Diverse representation of disruptive features like scars, tattoos, and drawing on hands.
11kHands - Palms [1]	2019	380	5368	~14	Single illumination over simple background. Some variation in hand pose (finger spread only) and position-in-frame. Single-session acquisition.
BMPD [66]	2019	82	1640	20	Some variation in illumination over simple backgrounds. No variation in hand pose but some variation in position-in-frame. Two-session acquisition.
SMPD [66]	2019	110	4400	40	Close-up of the palm in single, clear illumination. Some variation in hand pose (perspective and rotation from the camera only). Single-session acquisition.
NTU Palmprints from the Internet [112]	2019	2035	7781	~4	Variation in illumination and complex backgrounds. Variation in hand pose and position-in-frame. Multi-session acquisition (varying for each palm).
NTU Contactless Palms [112]	2019	655	2478	~4	

Table 2.1: Palm Datasets.

Dataset	Year	No. of Palms	No. of Ims	Ims per Hand	Description
Tongji Mobile Palm-print [177]	2020	~400	~16,000	40	Some variation in illumination and complex backgrounds. Some variation in hand pose (mainly finger spread) and position-in-frame. Two-session acquisition.
HIT-NIST [20]	2023		3020		Variation in illumination between subjects and complex backgrounds. Some variation in hand pose (rotation and finger spread only) and position-in-frame. Two-session acquisition.

Table 2.1: Palm Datasets.

are captured against simple (flat, uniformly coloured) backgrounds, while four of them contain variable and complex backgrounds. Background affects the region-of-interest extraction process for palm and knuckle recognition; many papers that use the simple-background datasets use thresholding-based methodology that would not be applicable to images with complex backgrounds. Hand pose variations can be divided into translational (position-in-frame), rotational, and finger pose. Two datasets have no hand pose variation in any category, while most have free movement of the hand position-in-frame. Two datasets, PolyU Contact-Free 3D/2D Palms and SMPD, have controlled rotation of the hand while the rest are all direct views of the flat palm. Five datasets contain variation in the spread of the fingers (how separated or together they are) but no datasets contained any curling of the fingers.

2.1.2 Region-of-Interest Extraction

The majority of early palm region-of-interest extraction methods involve the positioning of a small palm-internal bounding box relative to the finger valley points that encapsulates the largest concentration of palm lines; this method was so pervasive that it is the only palm ‘region’ considered in Zhong et al.’s 2019 review on palm recognition methods [178]. Methods of positioning the box relative to these finger valley points include simple distance-based (the box is perpendicular to and a set distance from the finger valley line) [165], ratio-based (similar, but with the box size based on the valley line size) [19], and angle-based (the angle of the valley line is used to adjust and optimise the position of the box) [6].

There are many approaches to detecting the hand valley landmarks in the literature. Some of the most basic involve detecting the hand contour through thresholding, then locating the valley points as the minima of the contour relative to an as-

sumed middle (an example, but not original: [49]). In learned detection approaches, Bao et al. use a custom CNN to localise hand landmark points in the finger valleys [9], Shao et al. use a regression tree algorithm to locate the finger valleys and the outer edge of the palm [144], and Lin et al. use Yolo v5 to detect the general palm region and then a U-Net derived network to locate the finger valley landmarks [105].

Alternative to the valley-relative box placement is to use the maximum inscribed circle in the thresholded hand image, which is then either fit with the maximum inscribed square [106, 17], or projected/unwrapped into square space to retain all information within the circle [26].

Many of the basic hand contour localisation methods are not suitable for unconstrained images as they require clear distinction of the hand from the background, for which even skin-colour based thresholding fails when there is complex lighting and backgrounds. Learned landmark detection overcomes this problem, however still require that the fingers be unfurled and (mostly) spread apart, and placement of the bounding box is only effective in images with a direct (non angular) view of the palm from the camera.

Several papers have taken the approach of regressing the edge locations of the desired bounding box as done in object detection: Aykut and Ekinici use active appearance models to detect key landmarks on the sides of the palm from which they align a bounding box up to and perpendicular to the palm edges, resulting in a larger palm region than the valley-box method above that is still positioned within the palm (features no off-palm pixels) [5]; Jaswal et al propose a palm segmentation network based on R-CNN, sharing layers with a ResNet50 classification network that determines the type of image (spectral, contact, or contactless) to produce feature maps for anchor box coordinates, from which the final bounding box is regressed in the R-CNN layers [73]; Izadpanahkakhk et al. use an adapted faster R-CNN with a final four-neuron layer to output the edge values of the bounding box to be extracted [65].

In 2019 Matkowski et al. propose a region-of-interest localisation network (ROI-LAnet) to detect the positions of key hand landmarks used to extract a palm image. Their network is built from VGG-16 pruned after the 3rd pooling layer, a local response normalisation layer, and a two-layer fully connected network to output 18 values: x and y coordinates for 9 points forming a grid over the palm. An affine transform is applied to the detected grid, projecting palm landmarks to the same space in the final image regardless of hand perspective in the original image [112]. Building on this work in 2024, Matkowski et al. propose a multi-spatial transformer network that extracts perspective-normalised images for the palmar-side of the five fingers as well as the palm region [114]. This grid spatial transformation approach is

uniquely appropriate for unconstrained images from all methods considered, as it is able to work with (palmar) hand images in any orientation to the camera, although it would likely need to be adapted to handle images with curled fingers or other occlusion of the palm.

2.1.3 Feature Extraction and Matching

The majority of papers reviewed performed feature extraction, often with deep learning, on unprocessed palm images (after extraction). As the majority of palm datasets used include no or very little change in illumination, having no image pre-processing to normalise for illumination prior to model training or feature detection is likely to lead to illumination-dependent features. As a reference for the more texture-based work prior to 2018, deemed less relevant for this review, see Fei et al.’s review of palmprint feature extraction methods: for enhanced greyscale contactless images they identify three prominent methods that appear most often: SIFT, local binary patterns (LBP) and local line directional patterns (LLDP) [38]. There are two examples included in this review which focus on palm texture in feature processing: Ordered Palm Texture Code Matrices (OPTCM) proposed by Jaswal et al. in 2018 and Gabor-PCA convolutional networks proposed by Genovese et al. in 2019. OPTCM uses contrast-limited adaptive histogram equalisation for image enhancement, followed by vector quantisation that efficiently describes the pixel arrangements in patches of the image [73]; they then use hierarchical deep matching on the corresponding patches between two images to compare. Gabor-CNN is based on a three layer network architecture with two convolutional layers and one binarisation layer [49]. Genovese et al. provide options for Gabor tuning both layer 1 and 2 filters, and for PCA tuning layer 2 filters (Gabor-PCA); the benefit of this is primarily encoding the texture information through convolutions.

Eight papers reviewed used novel or unique applications of deep learning to palm recognition. Izadpanahkakhk et al. use the same faster R-CNN as their region detection model minus the final layer to extract the palm features, then use state vector machines (SVMs) and K nearest neighbours (K-NN) to perform matching of feature vectors [65]. Liu and Kumar use a custom residual feature network (RFN) which contains three convolutional layers followed by four residual layers and two more convolutional layers to extract palm features [109]. Both studies compare with and outperform non-deep learned features.

Shao et al. use VGG-16 and knowledge-distillation based deep hashing networks to convert palm images into 128-bit binary vectors (Softmax output), using Hamming distance to compare and match [144]. Matkowski et al. use a VGG-16 based

architecture to produce their ‘feature extraction and recognition network’ (FERnet); as with their region localisation network it is pruned after the pool3 layer with an added local response normalisation (LRN) layer and regression network ending with a Softmax layer [112]. Chai et al. also use a pretrained VGG-16 backbone to their model, pruned at the pool4 layer with added dropout, LRN, and convolutional layers; their palm recognition model is fused with a gender classification model at the feature level, which they show boosts the performance on both tasks [18].

Brown et al. compare their custom network with six convolutional and pooling layers followed by two densely connected layers against VGG-16 on palmprint recognition, and found their method outperformed VGG-16 [16]. Xu et al. also compared the performance of ResNet-18, ResNet-50 and VGG-16 architectures against texture-based methods competitive coding and binary patterns, finding ResNet-50 to perform best [169]. Matkowski et al. use separate pre-trained ResNet-50 models pruned after the 6th residual block to extract features from the six regions of interest in their paper (palm and five fingers), each followed by an embedding module, either convolutional or fully-connected variety, which produces a 512-D feature vector for the palm and 128-D for the fingers [114]. Finally, Chai et al. in their most recent work used a ResNet-152 pruned after the third convolutional layer, with added assistant connected and fully connected layers; they use ImageNet pre-trained weights for the ResNet portion which they fine-tune alongside training the assistant and FC layers from scratch [21] with combined centre and Softmax loss.

2.1.4 Performance

Despite the small region sizes and simple nature of palm images, results consistently favour larger and deeper neural networks for recognition performance [169, 21]. Chai et al.’s method with a large ResNet-152-based model achieved 100% accuracy on the Poly-U 2D/3D dataset, plus 99% accuracy on CASIA, IITD, and Tongji, greatly outperforming the simpler VGG-16 architecture [19]. Other papers, notably Matkowski et al.’s multiple spatial transformer method [112], have been able to achieve 100% rank-1 accuracy on the PolyU contactless palmprint and IITD datasets. Aside from maximising some down-rank performance or minimising small classification EER%, palm recognition on the illumination-consistent small databases is essentially a solved problem.

Very few papers have evaluated their methods on unconstrained palm images. Chai et al. also assessed their above model on their newly collected HIT-NIST-v1 dataset, which contains smartphone images of hands in different scenes, and achieved 98.91% classification accuracy [21], though it is unclear if there is illumination/scene

change within each subject or only between subjects - if there is not change within-subject, then the illumination differences between-subjects could serve to falsely inflate the accuracy. Matkowski et al. test their multiple spatial transformer method on their NTU-P1-v1 dataset which was collected from the internet, and features a large range in hand pose, illumination, and quality within subject images. They achieve a best rank-1 accuracy of 52.42% and rank-30 accuracy of 75.59% from all compared models [112], significantly below the benchmarks set on all of the controlled datasets and indicating there is vast room for improvement in the unconstrained palm recognition domain.

2.2 Knuckles

Knuckles, in true anatomic sense, are the metacarpophalangeal (MCP) and interphalangeal (IP) joints of the thumb and four fingers. Colloquial terms for the knuckles vary; significant papers in the field refer to the proximal interphalangeal (PIP, closest to the hand) joints of the four fingers as "major" knuckles, the distal interphalangeal (DIP, furthest from the hand) joints as "upper minor" knuckles, and the MCP joints as "lower minor" knuckles [70]. Some other papers use ordinal terms (1st, 2nd 3rd) instead. This thesis from here on refers to the MCP joints as "base" knuckles, the PIP joints as "major" knuckles, and the DIP joints as "minor" knuckles to avoid ambiguity with the term "minor". Thumbs only have a base knuckle and one IP knuckle that is not designated major or minor.

In biometric literature, discussion of knuckles is not referring purely to the joints, but to the patterns of creases, ridges, and wrinkles that form on the skin over the location of the joints. As such, they are also often qualified as "outer/dorsal" or "inner/palmar", referring to which side is being viewed, with outer knuckles on the dorsal/back side of the hand, and inner knuckles on the palmar/inside of the hand. The vast majority of work in the area only considers dorsal knuckles, which contain a much higher density of high-definition creases compared to the palmar side. Little work has explored the impact of finger flexation on recognition potential, but Dan et al. show that there is a significant difference in observable knuckle crease features (lines, branches and intersections) at each end of the flexation range (which they measure in degrees from straightened) [29]. Kumar showed that match scores worsen as the difference in degree of flexation between two compared samples grows [91]. The difference in observable features is likely to strongly impact the outcome of any learned feature extraction, especially where training data does not contain sufficient within-subject flexation difference. The difference in the visibility of knuckle features through stages of flexation is shown in Figure 2.3, as well as the difference in density

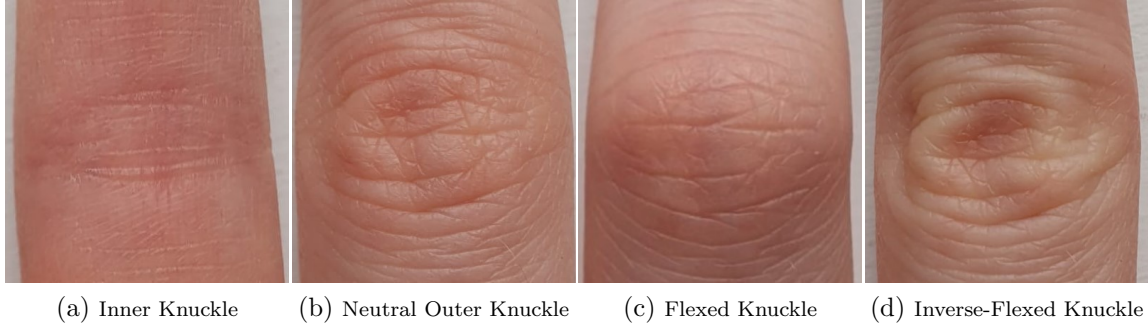


Figure 2.3: Examples of knuckle images, all taken of the same knuckle: a) the palmar side (inner), b) the dorsal side (outer) with the finger in a neutral position, c) the dorsal side with the finger flexed, and d) the dorsal side with the finger inversely flexed. These images demonstrate firstly the difference in texture and thus information density for the dorsal and palmar sides of the knuckles, and secondly how flexation can distort the knuckle creases and make the same knuckle appear radically different.

of textural information on the dorsal and palmar sides of the finger.

There are several significant reviews into knuckle biometrics to refer to here, and from which many of the specific sources below were found. Usha and Ezhilarasan’s 2015 review of simply “finger knuckle biometrics” is a comprehensive and analytical look at some of the early advances in dorsal-side major knuckle recognition, in particular which features can be extracted from the knuckle images and the methodology to do so. They identify in this review three key areas for further development: extraction of the knuckle region from contactless and unconstrained images, better multi-biometric fusion systems for knuckles, and methods which are invariant to the distortion of knuckle patterns which happens at different levels of flexation [160]. Jaswal et al. in 2016 performed a review of knuckle biometrics that considered more its use and benefits in the wider context of biometrics for authentication and security; in doing this, they compare it to other hand biometrics and also analyse the potential of the base and minor knuckles, and inner (palmar) sides. They conclude with eleven recommendations for future development, including (in addition to those mentioned by Usha and Ezhilarasan), the need for larger and more standardised knuckle image databases, and the potential for highly precise identification when knuckles are combined with other hand modalities, mentioning specifically a forensic context [70]. While not primarily about knuckle-based recognition, Al-Nima et al.’s 2020 review of inner (palmar side) finger texture (FT) should be mentioned as the primary source

of much of the textural information is the inner knuckles, particularly where they discuss horizontal lines across the finger as a source of identifying information [124].

2.2.1 Datasets

There are less datasets for knuckle recognition than palm recognition, and these dataset vary from images of single knuckles, images of fingers (containing 2-3 knuckles depending on the visibility of the base knuckle), and images of hands (containing 12-14 knuckles depending on the inclusion of the thumb). Since knuckles recognition is far more common and effective on the dorsal-side/outer knuckles only dorsal versions of whole-hand datasets are considered, however note that there are whole-hand palm datasets that could also be used for palmar-side/inner knuckle recognition. The knuckle datasets available include IITD Finger Knuckle Database Version 1.0 [100], Poly-U Contactless Finger Knuckle Images Database Version 1.0 [92], Poly-U Contactless Finger Knuckle Images Database Version 3.0 [95, 91], and Finger Major Knuckle (FMK) [53]. Knuckle regions can be extracted from Poly-U Contactless Hand Dorsal Images Database [99] and dorsal-side images from 11k Hands [1] and Tecnocampus Hand Image Database [37]. Information about these datasets is given in Table 2.2; the dataset ‘Poly-U Finger Knuckle Print Database’ is used in a number of papers referenced in this literature review but seems to be no longer available.

Of the eight datasets considered, two consist of close-up images of the knuckles, three are semi-close images of fingers, and three are images of the whole-hand dorsal side from which knuckle regions can be extracted. Only two of the datasets vary hand position-in-frame and finger spread, which will impact region extraction methods. Poly-U Contactless Finger Knuckle Images v3 is the only dataset to have variation in knuckle flexation.

2.2.2 Region-of-Interest Extraction

The vast majority of knuckle biometric work is performed on images that are already focused on the knuckle and so ROI-extraction techniques may have limited relevance to unconstrained work, though they have been included where potentially useful for positioning of a bounding box once rough localisation has occurred (through key-point detection perhaps). Some deep regression-based localisation is covered in the later papers.

In close-up knuckle images, Kong et al. first found the contour of the finger and rotated the image so that the lower finger edge was horizontal. They used the lighting in their specific images as a way to locate the knuckle centre (as the area of highest

Dataset	Year	No. of Hands	No. of Ims	Ims per Hand	Description
IITD Finger Knuckle Database [100]	2009	158	790	5	Close-up direct images of one knuckle from each hand in clear lighting. No variation in knuckle flexation.
Poly-U Contactless Finger Knuckle Images v1 [92]	2014	503	2515	5	Close-up images of the dorsal side of the middle finger of each hand in clear illumination against a simple background. No variation in knuckle flexation.
Tecnocampus - Dorsums [37]	2014	~220	~1000	~5	Multi-spectral, single illumination over simple background. No variation in hand pose or position-in-frame. Five-session acquisition over six months.
Poly-U Contactless Hand Dorsal Images Database [99]	2016	501	2505	5	Images of the dorsal side of each hand flat against a simple background, in varying illumination. Small variation in hand pose (finger spread only) and position-in-frame.
Poly-U Contactless Finger Knuckle Images v3 [95, 91]	2019	221	1950	~9	Images of the dorsal side of the index finger of each hand with direct illumination against a simple background. Variations in knuckle flexation cause deformations of the knuckle creases.
11kHands - Dorsums [1]	2019	380	5171	~13	Single illumination over simple background. Some variation in hand pose (finger spread only) and position-in-frame. Single-session acquisition.
Finger Major Knuckle [53]	2023	20	120	6	Close-up images of the dorsal side of the middle finger and thumb of each hand in clear illumination against a simple background. No variation in knuckle flexation.
Poly-U Finger-Knuckle-Print (FKP) Database		330	7920	6	Very close-up images of the major knuckle of the index finger and middle finger of each hand in bright, direct lighting. Knuckles in a controlled flexed position towards the camera.

Table 2.2: Knuckle Datasets.

pixel intensity) and grew a bounding box in rows and columns probabilistically [89]. Yu et al. built on this work to propose a two-step localisation process that also accounts for rotation of the finger (what they call 'vertical skew') by correlating it with the angle between the two halves of the upper finger edge, and off-setting their bounding box accordingly [172]. On the same data, Nigam et al. took the approach of binarising the image and using Canny edge detection to find the general knuckle region, then using Gabor filters to find the curved texture lines, which are generally symmetrical around the knuckle centre [123] (also used in [72]).

In images of fingers, Kumar and Wang initially naively positioned a bounding box at a fixed distance along the finger, found by thresholding, in all images [97]. Cheng and Kumar later detected finger edges (including knuckle creases) with thresholding and Canny edge detection, then slid a bounding box of fixed size to select the region with the highest edge-pixel density as the ROI, determining it to inherently contain the most biometrically-relevant information [24] (also used in Cheng and Kumar’s 3D knuckle reconstruction study [25]); they did not dynamically rotate the image to align the region, but stored multiple rotated templates of the region for matching. Attia et al. extracted the major and minor knuckles from finger images by breaking up the finger region to estimate the rough position of the knuckles, then finding the centroid of the convex hull of each region’s contour [4].

In whole-hand images, Kumar and Ravikanth initially extract whole finger ROIs by finding the contours of the hand through thresholding, then measuring the distances to the contour from an assumed palm centre position to find the finger valleys [96]. They used the finger valleys to place a base point at each finger and then divide the contour between valleys to find other landmarks, such as the fingertip as halfway along; the fingertip-base line is then used to rotate the image and extract a bounding box around the finger contour. Building on this work later, Kumar and Xu locate the base knuckle ROI as an extension of the fingertip-base line into the hand of $1/3$ the fingertip-base distance, then dynamically size the bounding box depending on the average distances between finger valleys for consistency with different hand sizes or camera distances [98, 99]. Once roughly localised, Kumar uses an orientation of gradients map to locate the centre of the knuckle for image alignment [95]

In terms of deep-learning approaches, Alghamdi et al. used the multiview bootstrapping method for hand keypoint localisation [146], then placed a bounding box of fixed size around their desired knuckle and fingertip points. [2]. Vyas et al. first trained a faster-RCNN [139] model to detect the bounding boxes for 12 knuckle points (the base, major and minor knuckles of the four fingers, no thumb) [163], then trained an ensemble of object detection networks including faster-RCNN (ResNet50 and ResNet101 based) and YOLO [138] (v2 and v3), and used the detection confidence scores from the networks to perform weighted box fusion in selecting the final box locations [162]. In none of these papers were the images rotated or transformed in any way to ensure alignment of the extracted knuckle regions.

2.2.3 Image Pre-Processing

Image pre-processing did not take place in every paper, with some opting to extract features from the raw or sometimes greyscale image. When it did take place, image

pre-processing on knuckle images was for the purpose of de-noising, enhancing, and sometimes extracting the knuckle skin texture into another representation like binary, 3D, or minutiae points. In the first automated knuckle recognition paper, Ferrer et al. focused on preserving the skin friction ridges and reducing noise in thresholding; They broke the images into 32x32 pixel blocks, each of which had its own thresholding filter calculated based on the direction of lines found by Fourier transform, then used Otsu’s algorithm to threshold [39].

The most common method to enhance the knuckle texture was to do histogram equalisation, used by Kumar et al. [92, 99], or contrast-limited adaptive histogram equalisation (CLAHE), used by Nigam et al. [123] and Jaswal et al. [72]. All four studies started by illumination-normalising the images using the same method of subtracting the mean pixel value in blocks, usually of 8-16 squared pixels. Nigam et al. reduced noise with a Wiener filter after histogram equalisation [123]. Jaswal et al. further encoded their histogram images with bubble ordinal pattern (BOP) and star ordinal pattern (SOP) using Sobel kernels, then used image ray tracing to emphasise the major line components in the BOP/SOP images [72].

In their a paper on knuckle crease minutiae points, Kumar and Wang experimented with several methods for texture enhancement similar to above (with median filtering for noise reduction) but opted to use Butterworth filtering plus self-quotient image calculation (an illumination-normalisation process proposed for facial recognition [164]) as pre-processing for their further work [97]. From the enhanced image, they then thresholded and used dilation and erosion to remove pixel noise, before applying a thinning algorithm to reduce the binary image to just the skeleton lines of the knuckle creases. Later, Cheng and Kumar performed 3D reconstruction of the knuckle skin surface texture using a stereo photometric method that requires multiple images of one knuckle (potentially usable for video-based identification) [25].

Effectiveness of the knuckle pre-processing methods is often done on the downstream recognition performance, which cannot be separated from the feature extraction and matching methods used, making direct comparison of pre-processing methods from different papers difficult. In practice, pre-processing methods may be deemed sufficient if they work for all (or most) images in a given knuckle dataset, which are at this time all constrained. It is likely that the existing methods will be insufficient for extracting knuckle texture through strong illumination effects in unconstrained images. While not for biometric purpose, Kim et al. proposed a method of illumination-normalising knuckle images using a conditional generational-adversarial network (CGAN) [87]. In this paper they train a generational network to remove the directional illumination effects on knuckle images with their own small private dataset of knuckles with strong directional lighting and ground-truth images.

Alongside the CGAN they use bilateral filtering to reduce noise, intrinsic image decomposition using second statistics patch descriptors, and histogram equalisation to enhance the texture information, from which they do 3D reconstruction.

2.2.4 Feature Extraction and Matching

The following papers have used a variety of features for knuckle images: thresholded binary images, texture filter responses, local and global feature descriptions, and deep learned features. Early on, Kumar and Ravikanth performed row pixel summation of a whole vertical finger image, unprocessed, to form the feature vector and then use PCA to reduce dimensionality [96]. Shortly after, Cheng and Kumar used 1-D log-Gabor filters on the greyscale image to encode the image textures [24].

Zhang et al. proposed competitive coding (CompCode) for knuckle texture encoding, which uses a bank of Gabor filters to calculate the dominant orientation at each pixel [176]. Gao et al. built on CompCode and created a version that keeps all filter bank responses - useful for when there may be intersections of lines - which they call MoriCode (multiple orientation code), and find the local binary patterns of the filter responses to form MtexCode (multiple texture code), then concatenating MoriCode and MtexCode together [45].

Kumar went on to evaluate local binary patterns, band-limited phase-only correlation (BLPOC), and 1-D log-Gabor filtering for feature encoding enhanced knuckle images, finding BLPOC to be the best [92]. He then proposed a new local feature descriptor that uses spatial filtering and quantization of the filter outputs, and compared with CompCode, BLPOC, local radon transform, and ordinal representations, of which the spatial filtering method performed second best to BLPOC [99].

Jaswal et al. extracted features from a knuckle image using SIFT and reduced dimensionality with PCA and LDA [71]. In a later paper they used various feature descriptors (such as SIFT and SURF) on encoded images to make correlation maps, pairwise comparing in a sliding-window style to account for potential spatial translations between the two images [72].

Nigam et al. encoded their enhanced knuckle regions using gradient ordinal relational patterns (GORP) and a derivative, Star-GORP (or SGORP), both of which they break into their vertical and horizontal components [123]. They found best results with SGORP-vertical when applied to their controlled data.

Attia et al. calculated binary statistical image features (BSIF) using PCA and independent component analysis (ICA) of the image; they then reduced dimensionality with PCA and LDA, and fused the two knuckle regions with feature concatenation [4].

These papers used a variety of methods to compare, rank, and make decisions on the equivalence of two vectors. Ferrer et al. in the first automated knuckle recognition study used Hidden Markov Models and Support Vector Machines to compare their thresholded images [39]. Several studies have used Hamming distance or a modified version of it [96, 24, 45] to compare images encoded as feature vectors; additionally, Zhang et al. [176] opted for angular distance while Jaswal et al. compared Euclidean, city block, and Spearman’s coefficient distance measures, finding Spearman’s to be best.

Kumar and Wang opted for a minutiae approach in 2015 by working with the traced skeleton lines of the knuckle creases (produced using the method described in the previous section) [97]. They detect the minutiae points of the lines in the region-of-interest - end-points, cross-points and bifurcations - using a pre-existing convolutional method for vein patterns from Olsen et al. [128]. They used local coherence of gradients to assess minutia point quality, with varying block sizes to account for the spread or contraction of knuckle crease lines with finger flexation. They then compare three methods for minutiae point comparison: triangulation, cylinder-code, and spectral, with and without their quality assessment included; they find best results when using spectral with quality assessment. The boost from quality assessment was only marginal in their results, but they assess with very controlled and high-quality data, so there is potential for it to be more beneficial when there is higher range in quality as in unconstrained images.

Most deep learning work in the area emerged post-2020, often with the application in unconstrained images for forensic purposes as a primary motivation. These papers move away from the close-up and highly controlled knuckle images used in most previous papers and use datasets like 11k Hands [1] and Poly-U Dorsal [99], which are still highly controlled in illumination and hand pose. As discussed in the previous ROI-extraction section, the non-dynamic bounding box localisation methods in these papers would lead to inconsistent and unaligned views of knuckles in unconstrained images, and so their results may be misleadingly effective as they evaluate only on similarly posed hands.

In 2021, Alghamdi et al. used DenseNet201 [59], pretrained on ImageNet and not further fine-tuned, to extract features from each knuckle (and fingernail) ROI and then performed feature fusion by concatenating the vectors [2]. They compare Euclidean and Bray-Curtis distance measures and conclude that Bray-Curtis is better. In the same year Vyas et al. opted to use DenseNet201 fine-tuned from layer 10 onwards to extract features from their twelve knuckle ROIs, then performed score-level fusion by averaging the match scores of the different regions [163]. Heidari et al. used AlexNet as a feature extractor for knuckle and fingernail images. They used

both score-fusion and rank-fusion to combine the contributions of each region [58].

2.2.5 Performance

Vyas et al. achieve 100% rank-1 accuracy on dorsal images of the 11k Hands dataset and 99.33% accuracy on the HKPU Hand Dorsal dataset, meaning the first returned subject was correct in nearly all instances [163]. This result demonstrates that in reasonably controlled images of the dorsal side of the hand (where the knuckles are visible), knuckle-based recognition can achieve near-perfect performance. Unlike with palms, there is no evaluation of knuckle-based recognition in truly unconstrained images.

Where results can be separated by finger, they show consistently higher performance for the middle finger over the index, ring, and pinky fingers [123, 99, 72, 58]. Similarly, where they can be separated by knuckle, performance is better with major knuckles over minor and base knuckles [92, 163, 162], and performance with base knuckles is notably worse (~ 27.41 - 29.41% lower rank-1 accuracy) compared to major knuckles [163]. Ablation analysis from the same study also show that the more knuckles that are combined, the better the performance, with the use of base, major, and minor knuckles on all four fingers achieving 100% rank-1 accuracy on 11k Hands.

Reasons for the difference in performance between knuckles and fingers could include the larger size of the middle finger and definition of the major knuckle creases providing more usable detail in the images, and the higher density of creases in the major knuckle creating more degrees of variation in the population. Cheng et al.’s combination of 2D and 3D pixel features in the image allow them to better separate genuine and imposter match scores, and they state their preference to use knuckle crease minutiae points over pixel features in future studies [25]. Interpreting from this, it is beneficial for models to utilise the texture (knuckle crease) information in feature extraction, and this should be considered when moving from the high-quality, single-illumination knuckle images used to date to unconstrained images in the future.

2.3 Dorsum

While there are no specific studies into the development of vein patterns in the hand, known mechanisms behind vein development in other areas of the body suggests that it is based on a combination of genetic and environmental factors, which would make the specific development unique to each individual [151] (see reference for quantification of variation of hand vein features in 178 participants). Despite the gap

in evidence for hand vein variability among the population, hand veins are a popular choice for secure and consistent biometric authentication due to the fact that they don't change over time with age, and they are very difficult (potentially impossible) to intentionally change or spoof. Per the survey below, minutiae and graph-based dorsal vein representation can consistently achieve 99% rank-1 recognition accuracy on datasets consisting of 1000 participants.

Jia et al.'s 2021 survey of dorsal hand vein biometrics covers the datasets, region-of-interest extraction methods, and feature extraction and comparison used for authentication applications [74]. Region-of-interest extraction is largely similar to the methods for controlled palmprint extraction: positioning of a square relative to the finger valleys, or maximum inscribed circle for the thresholded hand. The most common feature extraction method is minutiae (points representative of the ends, bifurcations, and intersections of lines), followed by graph representation (with nodes as the ridge ends and points of divergence or intersection, and edges as the lines between). This survey, and the majority of work in the area, covers application on infrared or near-infrared images, which enable good visibility of the veins under the skin, as shown in Figure 2.4 (c).

Restricted visibility of the veins in regular colour spectrum images is an obstacle to dorsal vein recognition in an unconstrained or forensic context, however there have been some occasions where there has been access to infrared spectrum images from some web cameras, and then dorsal hand vein matching has been used to confirm the perpetrator [11]. As demonstrated in Figure 2.4 (d), dorsal hand veins can be visible in some circumstances as they lay close to the surface of the skin, causing both a blue-ish discolouration and sometimes raised skin. This visibility however depends on several factors: the fat and hair coverage of the hand, the skin tone of the hand, and the rate of blood flow at the time the image was taken, which is itself affected by many factors such as time of day, temperature, and any recent exercise; even for the same hand, the veins could be highly visible in one image and not at all in another.

To overcome the limitation with non-infrared images there have been several studies into methods of extracting or enhancing the visibility of veins in colour images: Tang et al. use concepts of camera optics and skin biophysics to estimate the presence of veins under skin [175, 157]; Varastehpour et al. applied sparse auto-encoders to enhance vein definition in the RGB channels of colour images of skin [161]; and Keivanmarz et al. showed good performance with conditional generative adversarial networks (cGAN) in transforming small patches of colour skin images to their near-infrared counterparts, maintaining the underlying vein pattern [84]. Conversion of colour skin images to infrared-like counterparts for vein pattern extraction shows promise, but has yet to be demonstrated in unconstrained circumstances. While al-

lowing better visualisation for human analysis or simpler automated vein recognition methods such as skeleton graph comparison, it may be an unnecessary step if deep learning could be used to learn and embed the underlying vein patterns simultaneously. The use of generative networks to transform images in a forensic context also begs questions about the authenticity and reliability of the resulting data.

2.4 Whole-Hand

Matching of images containing a whole hand rather than a localised extracted region of the hand (e.g. palm, knuckle) is relatively sparse. For biometric security and authentication where special sensors can be deployed, the whole-hand offers a much less efficient basis for feature extraction than a regional extraction with higher density and higher quality features, such as fingerprints, knuckle prints, or palm prints. For a forensic context however, whole hand images are what will mostly be encountered, and so extracting features from the whole hand may be a more effective use of available information than focusing on a potentially low-quality extraction of a palm or knuckle.

The earliest and most prominent form of whole-hand recognition is based on the shape of the hand. In a very early paper on the topic Sanchez-Reillo et al. propose a system which captures geometric measurements including widths of the palm and fingers, lateral height from the surface, and deviations from determined points (essentially the 'volume' of the finger segments) [141]. With their method, capturing these measurements requires a controlled acquisition where pegs are used to consistently position the subjects hand, and a simple background is used for hand segmentation.

Moving on to the overall shape of the hand rather than measurements in specific regions, Yörük et al. use a modified Hausdorff distance and independent component analysis to compare two binary hand images [171]. These methods require complete alignment of the hands in each image to compare properly, however they propose a method of hand registration which removes the need for pegs to control the subject's hand placement: first, the hand contour (obtained from 2-means clustering segmentation) is rotated and aligned centrally, then the fingers are segmented and each rotated around a pivot point such that the fingertip-pivot point lines are consistent in all images. While this is a step towards unconstrained application, this method still requires simple backgrounds for segmentation and for the palm and outstretched fingers to be placed flat on a surface; the finger alignment method can fail if the fingers are not separated enough at first, merging in the binary image. They find that ICA performs better with 97.31% recognition accuracy on a dataset of 458 subjects

[171]. Dutağacı et al. continue this work using ICA and achieve better performance when incorporating texture information from the palm with the binary hand shape [36]. They also compare with some geometric measurements (palm width, finger widths, finger lengths, and overall area) and find recognition performance to be similar to ICA. Slot and Geradts compare Dutağacı et al.’s method with a checklist of human-assessed hand measurements, finding the non-automated checklist method to be more accurate.

Hand shape and geometry have not seen continued use as the field moves towards more unconstrained application. Low variability of hand shape and geometry among the population means that any error in measurement is likely to have a large impact on recognition performance, while variations in hand pose, distance to camera, and image quality increase the likelihood of errors occurring. Measuring the length of fingers, for instance, relies on accurately detecting the fingertip and base knuckle positions, which is exceedingly difficult when the fingers are bent or occluded. While regression-based hand landmark localisation could provide a solution, even the best performing models today, such as MediaPipe’s [174], have enough error and inconsistency that they would cause unreliable recognition results using this method.

In recent years, deep learning has been used for feature extraction and matching on whole-hand images. Yuan et al. use a convolutional network to produce feature vectors for registered pairs of colour and near-infrared whole-hand images, after which a support vector machine is used for matching [173]. Their custom convolutional network is comprised primarily of five modules from SqueezeNet following a single convolutional layer and followed by two fully-connected layers. The best performance of the proposed method (fusion of colour and NIR images with incorporation of local binary pattern for vein enhancement) underperforms VGG-16 on the Poly-U hand dorsal dataset.

In 2022 Baisa et al. proposed GPA-Net (global and part aware network), which is a ResNet-50 backbone with branching local feature extractors. They use unprocessed, whole hand images from 11k Hands and Poly-U Hand Dorsal to train the model with cross-entropy loss. GPA-Net achieves 95.83% rank-1 accuracy on 11k Hands and 94.64% on Poly-U Hand Dorsal [7]. Later, this architecture was adapted with channel and spatial attention modules to create MBA-Net (multi-branch with attention network) for the same task. MBA-Net achieved better rank-1 accuracies of 98.05% for 11k Hands and 95.12% for Poly-U Hand Dorsal [8].

The invariance to hand pose, image quality, camera distance, and illumination of deep-learned hand features will depend strongly on the training data used. An appropriately representative dataset to produce a truly unconstrained and condition-invariant model would have to include numerous images of the same hands in different

scenes, illuminations, and poses; a sufficiently large dataset like this does not exist to date. The existing work evaluates models trained on constrained data on more constrained data, which does not allow anything to be inferred about the feasibility of the methods in more unconstrained circumstances.

2.5 Skin Markings and Others

There are many more biometric traits of the hand not discussed above: palmar-side knuckles; fingerprints; fingernails; vein patterns in the fingers and wrist; androgenic hair patterns; and skin marks like scars, tattoos, freckles, birthmarks, and moles. These vary from highly identifying strong biometrics (e.g. fingerprints) to soft biometrics that can offer some valuable information when used in conjunction with others (e.g. androgenic hair). Their reasons for not being used wider in unconstrained hand recognition studies vary too; for some of them there is a lack of image data on which to train reliable algorithms (e.g. wrist veins); for many there is a poor trade-off between increased computational complexity and benefit to recognition performance (e.g. fingernails); systems to detect and match skin markings often already exist and serve better as independent, additional systems than incorporated into a multi-modal hand recognition system; and for some, contactless images are unlikely to be of a high enough quality to contain the necessary information (e.g. fingerprints).

Most contactless fingerprint research is done for authentication on consumer devices (e.g. mobile phones), solving the problem of deformations to the skin ridges that occur when pressing the skin against a surface for contact fingerprints and making the detected minutiae more consistent. A 2025 systematic review by Kaplesh et al. covers the methodology used for fingerprint extraction and matching in contactless images, including more recent deep learning approaches [83]. At present, all methods require very close-up images of the finger in good illumination (usually direct lighting from the mobile camera flash) for the skin ridges to be visible (see Figure 2.4 (b) for an example compared to a latent print in Figure 2.4 (a)), so has little relevance to unconstrained hand recognition. Finger vein recognition also has little relevance to the unconstrained domain because it requires specialist sensors (infrared/near-infrared) to capture the vein patterns, similar to dorsal veins but with even more limited visibility due to their depth. Finger veins are mostly used in tandem with other modalities such as fingerprints, and are most often proposed for authentication in high-security processes like ATM withdrawals and border crossings [101]. Just as discussed in Section 2.3, developments in the extraction of vein patterns from colour images of skin could lead to seeing this incorporated into unconstrained whole-hand

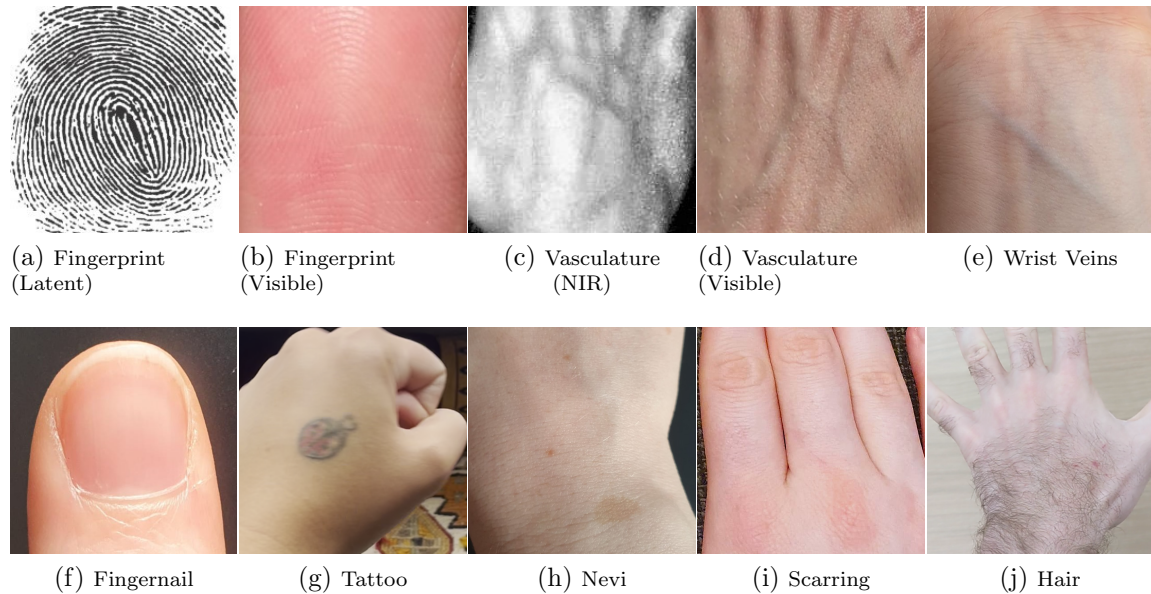


Figure 2.4: Examples of other hand biometrics: a) a latent fingerprint, b) a fingerprint visible with contactless flash photography, c) dorsal veins visible in near-infrared spectrum [37], d) dorsal veins partially visible in a non-infrared image, e) wrist veins and skin texture visible in a contactless non-infrared image, f) a fingernail, g) a tattoo on a hand dorsum, h) several nevi (birthmarks and moles) visible on a hand dorsum, i) a scar on a hand dorsum, and j) a hand dorsum with androgenic hair.

recognition, however it is likely to be superseded by better deep feature extraction when data allows it.

Two categories with more application in the unconstrained domain are wrist and fingernail biometrics. Matkowski et al. did a study on wrist identification in which they provide a new unconstrained wrist image dataset of 505 subjects for evaluation [113]; with deep learned features they achieve 24.88% - 38.64% rank-1 accuracy, though their dataset images have considerable variation in quality. The inner wrist offers two useful biometrics when visible: the skin crease pattern, and the vein pattern - see Figure 2.4 (e) for an example. Fingernails (Figure 2.4 (f)), similarly to finger veins, are most often used in conjunction with other biometrics in a multi-modal sense to increase recognition performances. Garg et al. have looked into solely-fingernail based recognition and achieved a genuine acceptance rate (GAR) of 80% [46]. It is most common for fingernails to be combined with the dorsal knuckles, as done by Kale et al. [81].

Soft biometric traits are traits often used subconsciously by humans to identify other humans, but which cannot be used for authentication due to being impermanent or non-unique [68]. For the hand, examples of soft biometrics include androgenic hair patterns (Figure 2.4 (j)) or skin markings like tattoos (Figure 2.4 (g)), scars (Figure 2.4 (i)), and birthmarks (Figure 2.4 (h)). While scars and birthmarks may technically be unique, they can appear identical to others to a very high degree. Tattoos and scars also lack in permanence, as they can be obtained and removed or healed over time. Distinct hair and skin markings are also not present on every hand: Black et al.’s study of scarring incidence on hand dorsums found that roughly 25% (61) of their 238 participants had no scars on either hand, and number of scars ranged from 1 to 26 with 50% more scars found on male hands than female hands [12]. In a related study on the same dataset, the authors found 9% of participants had melanocytic nevi (birthmarks or moles), with roughly double the incidence in female hands over male hands [13]. To incorporate these markings into recognition systems, there must also be some detection component, for which it is sensible to combine into one ‘skin abnormality’ detector and then classify, as done by Hefflin et al. for unconstrained recognition [57].

2.6 Multi-Biometric Fusion Schemes

Multi-biometrics refers to the combination of different biometric recognition responses to form one final outcome. It can be categorised as multi-modal, in which different biometric modalities are combined, multi-algorithm, in which different algorithms are applied to the same modality, or even multi-sample, which uses the same algorithm on multiple samples (images) of one modality from the same subject [32]. A multi-biometric system can also be any combination of these. For examples within hand recognition, multi-modal fusion would include the combination of different dorsal knuckles [4] or the palm with fingerprints [50]; multi-algorithm could be applying two different texture descriptors to obtain two feature vectors of the same modality (e.g. knuckle) [123], and multi-sample could be taking 2 images of the dorsal hand veins with different sensor types (such as colour image and near-infrared) [37] or several images of a finger from different angles for 3D-reconstruction [25].

The process of combining the different biometric responses into one is called fusion, and this can occur at many different levels of processing; these are commonly categorised as sensor-level, feature-level, score-level, rank-level, and decision-level, as done by Li et al. in their review of hand-based multi-biometric fusion [104]. Sensor-level fusion is the combination of the initial sensor output; for most biometrics this is an image, and the multiple images can be concatenated or combined pixel-

wise, such as through averaging the pixel values, or even used for 3D-reconstruction. Feature-level fusion is the combination of multiple feature vectors; again, the most common method for this is to concatenate them (so e.g. two 128-dimension vectors become one 256-dimension one), however there have been several other template-based methods devised as described in Li et al.'s review. Score-level fusion is applied to the metrics which measure the similarity between the input and the template (gallery) data; this could be a set of class probabilities, or the feature-distances in information retrieval. Li et al.'s review covers class score fusion based on min, max, or mean score, however when the scores behave as true probabilities there are many probabilistic frameworks that can be used, such as Bayesian inference or Dempster-Schafer theory [122]. Rank-level is suitable for ranked-retrieval based tasks, in which a collection of potential matches (the gallery) are ordered by likelihood of being a match to the probe sample; for rank-level fusion, a ranking is produced for each biometric sample, then combined using any of a number of methods which include averaging at simplest or probabilistic Monte-Carlo methods at the advanced end. However, it is also common to see a score-level fusion of the metric used for ranking (often feature-distance) in ranked retrieval tasks instead. Decision-level fusion is again best suited to classification-style tasks where there is a final identity 'decided' by the system, often the identity with the highest score or highest rank. The most well-known and simplest form of decision fusion is majority vote, however this is not the most optimal as it does not perform well with only a few samples (the number of samples should at the very least be more than the number of possible classes, which very quickly becomes unscalable).

Jaswal et al. in their 2016 review of knuckle biometrics discuss the potential for multi-biometric fusion of knuckles with other hand modalities (seemingly overlooking the combination of different knuckles being a multi-biometric system in itself), particularly for forensic contexts. In the review they argue that score fusion is the most popular the most suitable for biometric applications because it avoids the comparison of non-comparable sources of data that would occur when combining, for example, a knuckle and palm image at sensor-level or combining the knuckle and palm feature vector at feature-level. Sensor and feature level fusion may be suitable for a multi-sample system, where it is expected that the input represents the same subject and so is combinable, and that the feature vectors they produce should be identical, but it is not appropriate where each input sample should have a full and individual set of features extracted, and the position in feature space relative to other samples of the same type matters. Indeed, score fusion is the type of fusion most seen in biometric literature.

2.7 Conclusion

The hand is a region of the body with a lot of identifying information, and as such, many different potential biometric modalities. The vast majority of the hand recognition studies found and analysed were specifically focussed on one of those modalities; mostly the dorsal major knuckles, followed by the palm. Some other studies focussed on different modalities, such as the veinous systems in the dorsum and fingers, the fingernails, the wrist, overall hand geometry, or even the incidence of markings such as birthmarks, scars, and tattoos. A smaller number of studies performed multi-biometric fusion of various combinations of those modalities. Largely, the intent of these papers was to develop recognition methods for authentication and security purposes where inference samples can be collected under somewhat controlled circumstances and even re-captured when necessary; as such, a lot of the methods for region-of-interest extraction, pre-processing, and feature extraction rely on certain image conditions and would not be applicable to unconstrained images. To facilitate this, there are a number of datasets of hand images (some of specifically close-up knuckle images) taken under constrained conditions, where the illumination does not vary, nor does the pose and position of the hand. At the time of this review no public datasets exist consisting of a large number of unconstrained images with variation for each image subject. Despite this, there were still several studies identified which were motivated by application in forensic or other similar unconstrained contexts. They often aimed to employ region-of-interest and pre-processing techniques that were image condition independent. Only one research group in these studies evaluated their methods on a dataset of unconstrained images (one that they collected and is not widely available), and in their papers, there is no discussion of the impact various unconstrained image conditions may be having on recognition performance, nor a wider comparison of various different approaches to hand recognition (different modalities).

From this review several avenues for future development have been identified: firstly, the collection of an unconstrained dataset with specific attention to ensuring that there is a controlled range of illuminations and hand poses for each identity in it, to enable experiments to investigate the impact that illumination or hand pose may be having; secondly, a comparison of various different regional hand recognition approaches (knuckle, palm, whole-hand) on unconstrained images to determine which may actually achieve superior performance; and thirdly, the development of methods specifically designed to address the shortcomings identified in unconstrained recognition, e.g. illumination invariance. The next chapter presents the collection of an unconstrained hand image dataset designed to enable these experiments.

Chapter 3

Data and Labelling Strategy

This chapter describes the collection of a new hand image dataset intended to address some of the shortcomings of current datasets and provide a way to evaluate hand recognition methods on a proxy of the kind of data conditions that would be seen in a forensic (or other unconstrained) application scenario. Following this, an experimental approach is taken to empirically establish which labelling strategy (per individual, per hand, or per side-of-hand) is superior, and discussion is given to how the approach used intuitively in related studies may have flaws when applied to unconstrained data.

The literature review performed in Chapter 2 highlighted a lack of hand image data to work with when developing hand recognition systems that are intended to be applied to unconstrained data, such as in a forensic context. A total of 19 hand image datasets were discovered; of these, 6 were exclusively knuckle images or the dorsal side of the hand (most often used for knuckle recognition, but containing more than just the knuckles), 11 were palm or palmar-side images, and two were datasets that consisted of whole-hand images of both sides (dorsal and palmar) of the hand - 11k Hands [1] and Tecnocampus Hand Image Database [37]. One limiting factor common to the datasets is the small cohort sizes, having a mean of 557.83 with standard deviation 531.56 for palmar datasets, and mean of 291.63 with standard deviation 157.94 for dorsal/knuckle datasets. This cohort size is defined as the number of hands rather than the number of individuals as the textural features on each hand will differ, however, it should be noted that for most of these datasets the subject cohort is half the amount of the hand cohort, meaning there will be less diversity of high level features such as those affected by gender or ethnicity. Another limiting factor is that each of these datasets contains very few images per hand: a mean of 14.18 with standard deviation 13.07 for palmar datasets and a mean of 6.75 with



Figure 3.1: Twenty-five randomly selected images from a) the 11k Hands dataset and b) a subset of 8 subjects in the VarHands dataset. Image filenames were randomly sampled from the annotations files.

standard deviation 2.68 for dorsal/knuckle datasets. Having few images per hand means there is little variation in the representation of that hand, such as under different illuminations or angles, with different hand poses, or with time-dependent changes to blood flow, hair, nails, scarring, and so on. Indeed, since the datasets were largely produced for authentication and security contexts, they do not even intend to represent variation in illumination and hand pose to the degree that would be necessary for unconstrained applications, as in their context there can be some control exercised over those factors at inference time.

In an unconstrained scenario, the recognition system may be presented with images of hands in a variety of poses, lightings, sceneries, and image qualities. Images could also be horizontally flipped (and often are when taken with a smartphone), and it is not always possible to know if this has been done. For these reasons we would expect an unconstrained hand matching system to be:

- Pose invariant, such that it matches images of the same hand in different poses and doesn't match images of different hands in the same pose.
- Horizontal-flip invariant, similar to above.

- Illumination invariant, similar to above.
- Background invariant, such that the background does not impact feature extraction nor prevent region-of-interest extraction.
- Able to handle low quality images appropriately, such that undue confidence is not given to a match or non-match result from unreliable data.

Considering only a deep-learning approach with minimal pre-processing of the images, in order to learn the image features that are consistent under different illuminations, e.g., but not shared among images of different hands, the training dataset would have to contain many images of the same hands under different illuminations as well as having many different subjects. Having diversity of representation for every image subject is important, as under some circumstances it could be possible for the model to learn false patterns; for example, if the illumination varies between subjects but not within the images of individual subjects, then the model would falsely learn those illumination factors to be highly discriminative features, and so in any future application fail to match the same hand in different illuminations, and possibly falsely match two different hands in the same. Because of the risk of learning false patterns, applying deep learning with no pre-processing of images to a small-size non-representative dataset might not be the best approach. Many image pre-processing techniques or deterministic feature extractors are able to overcome differences in some factors like illumination, but would need to be incorporated with attention to whether they are having the expected impact. To ensure the chosen methods are actually learning the desired and effective patterns in the data, it is very important to have a large and representative evaluation dataset. Given that effective models can be learned from constrained data using some techniques (pre-processing, augmentation), it would be more prudent to assign any small amount of unconstrained data that exists to evaluation.

In this chapter Section 1 describes the collection protocol and analysis of the new dataset, and Section 2 describes the experimental approach to establishing the labelling strategy to be used for the data going forward into later experiments.

3.1 VarHands: A New Hand Image Dataset

As outlined in the introduction, a new dataset is required to enable the evaluation of hand recognition models with regard to their robustness or invariance to changes in image conditions. These conditions were categorised broadly as hand pose, image mirroring (horizontal flip), illumination, background, and image quality. Horizontal

flipping and loss of image quality by reduction of image resolution can both be applied to the data after the point of collection, and so are not considered important requirements. Variety in backgrounds (the scenery behind the subject’s hands in the images) is important to ensure there are no background features (distinct carpet patterns, perhaps) influencing feature extraction, however, it is also important that the backgrounds remain relatively simple to not prevent accurate region-of-interest (ROI) extraction. While robust ROI extraction is important for an unconstrained system, this project focuses on the consistent extraction of hand features and uses an external algorithm for ROI detection prior to fine-tuned extraction; improving the ROI detection process is out-of-scope for this project. This leaves hand pose and illumination as the two conditions for which variety within images of the same subject must be maximised in a new dataset. For both of these conditions there are an almost unlimited number of states. Illumination is affected by brightness, hue, and direction of the light source, which all exist on a continuous spectrum. Hand pose is a combination of the state of flexation of fifteen finger and thumb joints, the rotation and flexation of the wrist, and the angle of view from the camera, a total of seventeen ‘variables’ affecting the complete pose, each of which can individually have large variation.

To capture a large range of natural hand poses for each participant the dataset was collected as video recordings rather than photographs, from which frames were later extracted for an image dataset. For illumination, each participant provided videos in four categories: clear daylight, shaded daylight, clear artificial light, and shaded artificial light, with the artificial videos being taken at night so the only source of light is artificial. The collection protocol for this and the specific instructions to participants are given in the next section. This dataset has been named VarHands (variable hands) for reference throughout the subsequent studies. Because of conditions of the consent agreement with participants, this dataset cannot be shared publicly, however it is hoped that providing the details of the collection protocol can allow other groups to capture a similarly representative dataset for reproduction of results.

VarHands consists of 6,436 images of 40 hands from 20 participants, with approximately 100 to 250 images per hand. The intention of the dataset is not to evaluate the discriminate performance of the recognition methods, which would require far more subjects, but to evaluate consistency in feature extraction and invariance to various image conditions. For this reason it should not be the sole evaluation dataset for recognition performance, and should be used alongside a larger-cohort dataset. The dataset chosen for this is 11k Hands [1], which contains 10,539 images of 380 hands (190 subjects), with approximately 28 images per hand, roughly half on the

Video Name	Hand	Light 1	Light 2
Left_Day_Clear	Left	Day	Clear
Left_Day_Shaded			Shaded
Left_Artificial_Clear		Artificial	Clear
Left_Artificial_Shaded			Shaded
Right_Day_Clear	Right	Day	Clear
Right_Day_Shaded			Shaded
Right_Artificial_Clear		Artificial	Clear
Right_Artificial_Shaded			Shaded

Table 3.1: Eight video categories provided by each participant.

dorsal side and half on the palmar side. A dataset with lots of different hands is more suited to evaluating false positives and the overlap between hand features. The controlled illumination and hand pose also allow evaluation against a baseline of algorithm performance under ideal conditions (no quality loss).

3.1.1 Collection Protocol

Participants were recruited through email outreach, personal connections and word-of-mouth, with the offer of a £10 Amazon voucher in exchange for their participation. Physical posters, provided in Appendix A, were used in some University buildings but did not result in any recruitment. Only two restrictions were placed on recruitment: that participants be 18 or over, and that they have not previously contributed to any other hand image dataset. The second point was to ensure there would be no train-test data contamination when doing cross-dataset analysis.

Each participant was instructed to self-record 8 videos on their personal mobile device and submit them through an online portal. They were told or shown how to hold their mobile device in one hand while they filmed the other, and instructed to keep their hand in focus and in frame as much as possible. This resulted in each video containing one hand close-up from a first-person perspective. The videos were to be recorded in at least two sessions with a small period of separation, during the day time for daylight videos and night for artificial-light videos, to avoid the sets of videos appearing too similar. Participants were also asked to remove any jewellery that is not worn permanently prior to recording the videos, to limit the impact of accessories on falsely high recognition rates. Permanent accessories, such

as wedding bands, were allowed to remain as they are common, and removing them was likely to leave a visible impression on the hand that would have the same effect as the jewellery. The email text sent to participants with the specific wording of instructions given is provided in Appendix A.

Each video is 10-20 seconds long and shows the participants moving their hands into a variety of poses and gestures (see Figure 3.1). The participants were not strictly guided on how to move their hands but were shown an example video posted online (<https://youtu.be/Z5OPjIjQwi4>) and given suggested movements as in the following list. This process was repeated 8 times, capturing each hand in four different lightning conditions, to achieve the set of videos described in Table 3.1.

1. Start with your fingers outstretched, palm facing the camera.
2. Slowly curl your fingers one at a time into a clenched fist.
3. Rotate your hand so the palm is facing away from the camera.
4. Slowly uncurl your fingers one at a time until they are all outstretched.
5. Rotate your hand so the palm is facing the camera again.

Three participants withdrew from the project either before or shortly after submitting their data, whilst it had not been anonymised and combined with the other participant data, for various reasons.

Initially 100 participants were budgeted for, and this was considered the target for recruitment within the first round of outreach occurring in the second year of study. Despite best efforts, and two more subsequent rounds of outreach, only 20 participants were recruited. From those who declined to participate in the data collection the most common reason given was concern over personal data privacy or leakage of their biometric information. Others had general objections to biometric technologies, or particular objections to the use of biometric technologies in law enforcement contexts. These individuals were provided with information on the data management plan, anonymisation process, and ethical reviews for the project, but refusals were respected. A small number of people declined due to self-consciousness about the appearance of their hands, and some simply did not respond to prompting about finishing the data collection. This experience suggests that biometric data collection for explicitly law-enforcement related projects may require more time and resources than expected, and result in low yields.

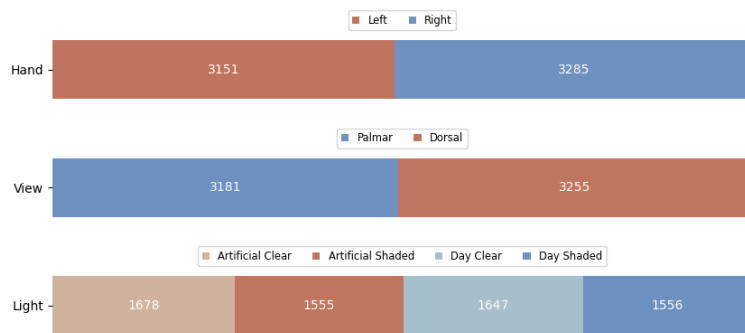
3.1.2 From Video to Image Dataset

In order to extract frames from the videos to form an image dataset, each video was loaded using OpenCV and hand landmark points were found using Mediapipe’s hand landmark detection library [174], which detects 3D coordinates (pixel coordinates x and y , and wrist-relative depth coordinate z) for 21 different hand landmarks, including the wrist, fingertips, and three knuckle joints on each finger. The frames and their landmarks were then divided into 10-frame sections, and in each section any frame without any successfully detected landmarks was discarded (this would happen due to blur or the hand moving out of frame). The variance of Laplacian was calculated for all remaining frames and any below a threshold of 0.5 were discarded, deemed to be too poor quality for recognition. The remaining frame with the highest variance of Laplacian (image sharpness) in each section was then selected to form part of the image dataset, stored along with its corresponding hand landmarks. Sectioning the video this way ensured a range of frames were taken from across the length of the video to reduce the similarity of images.

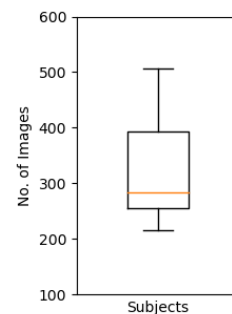
Classification of images as dorsal or palmar can be important for data labelling, as is discussed in more detail in Section 3.2, or for constructing sub-datasets for training of separate dorsal and palmar models and nuanced analysis. Due to the variation in poses in this dataset it is less clear whether some images are dorsal or palmar, as they can be somewhere in between and show no characteristics of either, though a label is assigned regardless using handedness and the relative positioning of the index and little finger MCP joints. For a left hand: if the index MCP is to the left of the pinky MCP the image is classed as palmar, and if it is to the right of the pinky MCP the image is dorsal; vice-versa for right hands. Handedness, while known in this circumstance, is also predicted by the Mediapipe hand landmark model; this can be useful for some tasks but not for forensic comparison as it may not reflect reality, images can be and often are flipped.

3.1.3 Analysis of Collected Dataset

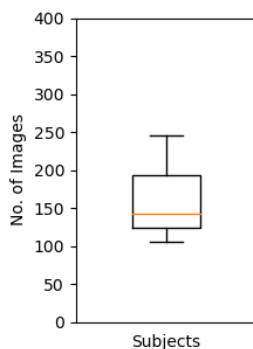
The number of images in the dataset for each subject varied and depended on both the length and quality of the videos provided by the subject, as an image was taken for every 10 frames, low-quality excluded. The box plot in Figure 3.2 (b) shows the range for this figure - the least amount of images for a subject is 216 and the most is 506, with a mean of 321.8. The majority of subjects lie in the 250-400 image range and there are no extreme outliers to skew dataset balance. As can be seen in Figure 3.2 (b) the images in the dataset are also well balanced with regards to the three factors of handedness, hand view, and lighting. Across all 20 subjects, there are



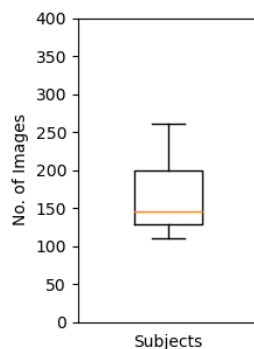
(a) Handedness, hand view, and lighting.



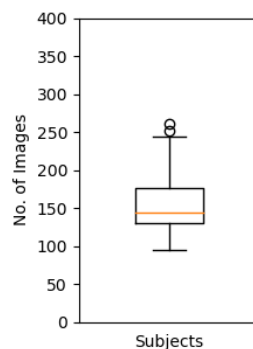
(b) Images per subject.



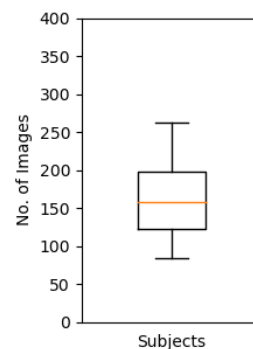
(c) Left Hands.



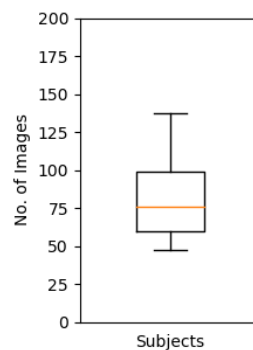
(d) Right Hands.



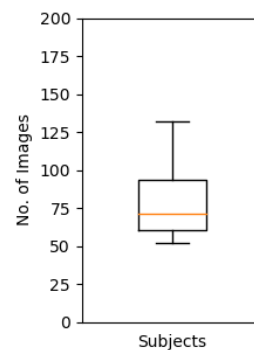
(e) Palmar Sides.



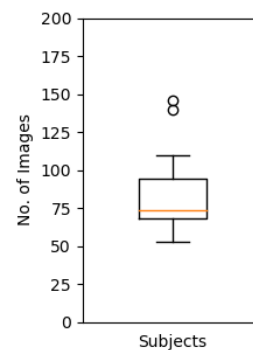
(f) Dorsal Sides.



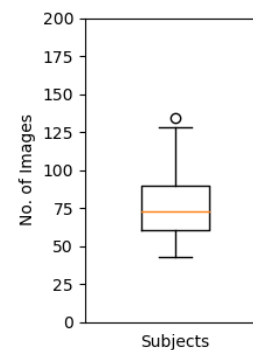
(g) Daylight Clear.



(h) Daylight Shaded.



(i) Artificial Clear.



(j) Artificial Shaded.

Figure 3.2: Distribution of factors in the VarHands dataset.

3151 left hand images to 3285 right hand images, 3181 palmar images to 3255 dorsal images, and 3203 images in daylight, 1647 of which are clear and 1556 are shaded, to 3233 images in artificial lighting, 1678 of which are clear and 1555 are shaded. Figures 3.2 (c)-(j) show the range of these numbers across subjects; while there is variation in the number of raw images provided by each subject (as in Figure 3.2 (b)) due to providing slightly longer or shorter videos, the proportion of each category was roughly the same for each subject, resulting in the balance seen in Figure 3.2 (a).

Twelve out of the 20 dataset subjects (24 out of 40 hands) had no noticeable accessories in any of their images. One dataset participant had two tattoos on their left hand: one on the dorsum and one on the wrist. Five hands from four subjects had permanent accessories, meaning they appeared in every image of that hand, including one instance of a watch and four simple rings (wedding bands). Nine hands from five subjects had temporary accessories, meaning that they only appeared in some of the images of that hand, including one bracelet, rings, and coloured nail polish. The temporary accessories would appear in the images from one or more video from the participant, meaning they could, as an example, appear in the daylight-clear and daylight-shaded images, but not the artificial light images.

Some demographic factors about the dataset are known, such as the gender distribution being 12 female to 8 male. The age range of the participants (at the time of collection) was heavily skewed young, with 14 participants falling into the 20-30 age range, 1 in 30-40, 2 in 40-50, and 3 in 50-60. Participant ethnicity was not recorded, however, the largest demographic would be white British, and the dataset is heavily skewed towards lighter skin tones; this presents a major limitation, as evaluation on this dataset can give no indication of performance on darker skin tones and potentially result in bias outcomes. Because the dataset was not collected over a long time period it does not represent any changes that may occur in a hand’s appearance over time such as skin sagging, growth or loss of hair, introduction of new tattoos or scars, or change to texture (e.g. skin rashes) or discolouration (e.g. age spots). With regards to disability, all dataset participants had two hands and all hands had five fingers and no noticeable abnormal development, which may also present a limitation of the resulting performance analyses.

3.1.4 VarHands-Easy and Dataset Splitting

For the purpose of being able to control for variation in hand pose for subsequent experiments and analysis, a subset of VarHands was created that only contains hands with a straight-on view (meaning all fingers are close to equidistant to the camera)

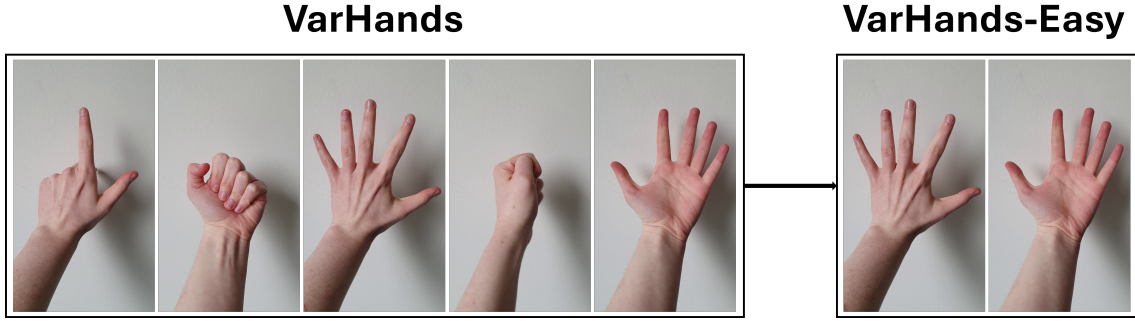


Figure 3.3: Visualisation of the manual selection of to form the VarHands-Easy sub-dataset. Images with excessive rotation, flexation, or curling of the fingers are removed.

and all fingers unfurled and straightened, as shown in Figure 3.3. These constraints mean that all of the features of the palm and palmar-side knuckles are visible in the palmar images, and all of the features of the dorsal-side knuckles are visible in the dorsal images (the features of the hand dorsum itself are always visible as they are not obscured when the fingers are curled). This subset has been named VarHands-Easy and can be used to evaluate performance with no pose variation, compared to performance *with* pose variation seen on the full VarHands dataset. The images that form VarHands-Easy were manually selected based on observation.

The 11k Hands dataset was the primary training dataset used throughout this project, however it is also still a valuable test dataset for insight into performance on controlled data and for direct comparison with other works. Half of the *subjects* in the dataset were assigned to the training set and the remaining half to the test set. Subjects for which all four hand and view subsets (right dorsum, left dorsum, right palm, left palm) had six or more images were prioritised for the training set so that they would be compatible with a balanced batch sampler that searches for sets of images (four to six) of the same class.

The images in both VarHands and the test set of 11k Hands were then split into probe and gallery subsets. For each label (ID4, defined in the next section), half of the images rounded up were put into the gallery and the remaining images acted as probes, as illustrated in Figure 3.4. Rounding up ensures that if there is only a single image belonging to a label then it goes to the gallery, preventing calculation errors (such as divide by zero when computing recall) and falsely lowered accuracy scores caused by having a probe image with no corresponding gallery image.

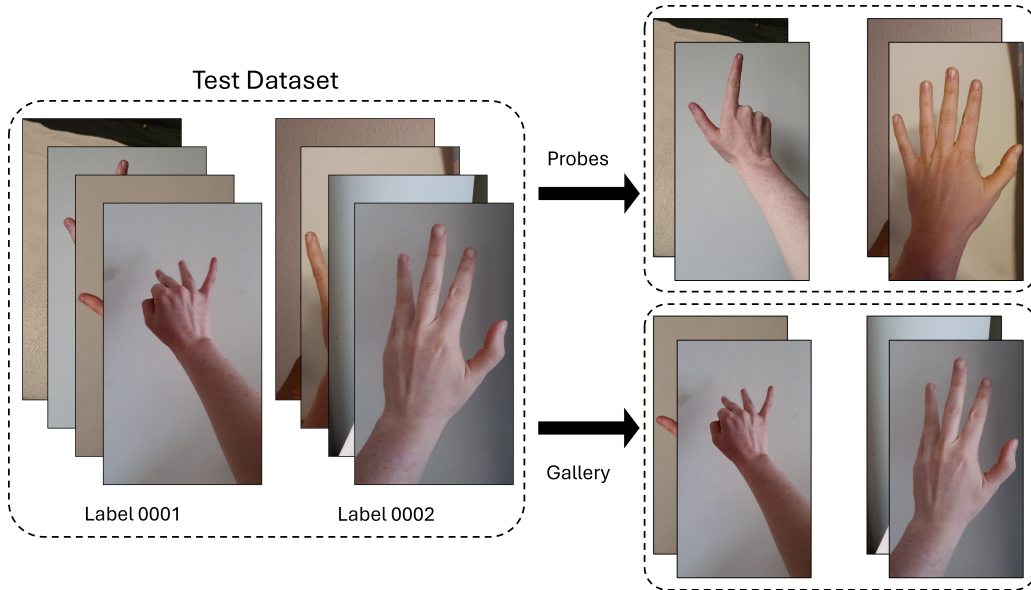


Figure 3.4: Splitting of images belonging to each label in the test dataset half into probe and half into gallery sets. Probe images are presented to the system individually, and for each probe the gallery is ranked by match likelihood.

3.2 Data Labelling

There are many possible ways to label a hand image dataset for model training, so in this section an empirical approach is taken to establish the best strategy to be applied to the new collected dataset. The 11k Hands dataset [1] is used for this purpose, as the larger number of hands and lack of influence from other image conditions allows the best evaluation of the embedding of pure biometric features. Comparatively, the VarHands dataset is not suitable, as there could be many potential biases introduced by the lack of biometric variety (only 40 hands) and large variation in illumination and hand pose, making the results non-generalisable. The aim of these experiments is to determine which labelling approach leads to the most separable biometric feature embeddings.

Biometrics for which there is a single source, such as faces, voice, or gait, are able to be labelled by subject only. Biometrics for which there are multiple non-comparable sources on one individual, such as iris or fingerprint, may (in circumstances that are not constrained to only one source) be labelled by subject and another denominator: left or right iris, and e.g. left index fingerprint, to prevent the false comparison of incomparable samples.

The left and right hand of a subject will be very similar in appearance (much closer than most other pairs of hands) yet not perfectly identical. The choice to label by subject or by hand is a trade off: at test-time, with by-hand labelling many of the close negative matches are likely to be the of the same subject, and so using subject labels may increase true acceptance rates; at train-time however the choice will have a large impact on the patterns and features learned by the model. The information shared between left and right hands that makes them appear similar is the same information that is also less variable among the population - skin tone and hand geometry. The information that differs between left and right hands on the same person are the highly identifying features such as palm lines, knuckle creases, dorsal veins, and marks like scars, freckles, and tattoos. If data were to be labelled by subject for training, models would learn to disregard the latter category of information to match left and right hands, and depend more on the first category, which is disadvantageous. Conversely, labelling by hand for training allows the model to learn the distinctive features of the hand while providing hard negative samples (close appearing imposters) that may not appear organically in a small-size dataset.

Unlike with faces, iris, or fingerprint however, hands have different features visible on each side. The dorsal side of the hand has the fingernails, outer knuckles, dorsal hand veins, and different skin marks and texture that would not exist on the palmar side. The palmar side of the hand has the fingerprints, inner knuckles, and palm lines, as well as unique skin markings. Largely, the information on each side of the hand is mutually exclusive - if the palm is visible, the dorsum is not, and vice-versa. As such, the same assumptions can be made about how a feature extraction model would behave as above. If using a by-hand only approach to labelling, the model would likely disregard the fine details to be able to match dorsal and palmar images, relying instead on the non-variable information like geometry and skin tone. Using a by-view labelling strategy will be important to ensure the model is able to properly extract the desired information in the dorsum, palm, and knuckles.

Most reviewed hand biometric studies use a by-hand labelling strategy (not needing to use by-view labelling as they only look at one side of the hand), and one study used a by-view but not by-hand strategy (meaning they treat left and right hands as the same, but dorsal and palmar sides as different) [1]. All work that does whole-hand deep recognition with datasets that feature both dorsal and palmar images use a by-hand and by-view labelling strategy [7, 123]. Additionally, there is divergence among studies where some train a single model on all of the data [1] (left and right hand images, dorsal and palmar views) while some train separate models on subsets of the data (e.g. dorsal, or left-dorsal and right-dorsal) [8]. Intuitively, there may not be much benefit to training separate models for the left and right hands, but there

might be for training separate dorsal and palmar models since the visual features are mutually exclusive and differ significantly. A combined dorsal-palmar model could result in less effective embeddings, where the model dedicates some of the embedding dimensions to features only found in palmar images or only found in dorsal images; if a model is trained exclusively on dorsal or palmar images, it could capture more fine-grained information in its representations. To date there have been no comparative studies to empirically show which labelling and training strategy results in the best recognition performance, so this will be addressed.

3.2.1 Research Questions

Based on this theory of data labelling, the research question addressed in this section is:

RQ1 How should hand images be categorised and labelled for optimal training and recognition performance?

To answer this question an experiment was performed following the methodology in Section 3.2.2.

3.2.2 Methodology

Biometric images are labelled with a numeric subject identifier so that matching and non-matching images can be identified during training and testing. In an image-retrieval scenario, models are not trained to predict this label, but to produce similar feature embeddings for images with the same label and distant embeddings for images of different labels, using contrastive losses. Using this scenario two tests were performed, with a total of ten models trained.

The first test evaluates which labelling strategy results in the best performance when training on *all* of the data. For this, the data was labelled according to four different criteria, named ‘ID1’, ‘ID2’, ‘ID3’, and ‘ID4’. For ID1 there is a unique identifier only for every *person* (subject) in the dataset, regardless of whether the image depicts their left or right hands in dorsal or palmar view. For ID2 there is a unique identifier for every hand, so the left and right hands of a subject are considered different identities but the dorsal and palmar view the same. For ID3 this is reversed, and the dorsal and palmar views are identified separately but the left and right hands of each subject are considered the same identity. Finally, for ID4, each combination of subject, hand, and view are given a unique identifier, such that each subject is broken down into four ‘identities’: right-dorsal, left-dorsal, right-

Subset	Description
Sub (ID1)	All images labelled by subject only
All (ID4)	All images labelled by subject, hand and hand view
Dorsal	Images of dorsums labelled by subject and hand
Palmar	Images of palms labelled by subject and hand
RD	Images of right-dorsums labelled by subject
LD	Images of left-dorsums labelled by subject
RP	Images of right-palms labelled by subject
LP	Images of left palms labelled by subject

Table 3.2: Label categories.

palmar, and left-palmar. Four ResNet-50 models are trained using triplet loss, one with each of the label sets. This test was evaluated in training performance (loss) and ranked-retrieval performance on all images in the 11kHands test dataset.

The second test evaluates whether models trained on only a subset of the data, e.g. dorsal-side images, outperform models trained on combined data. The ID1 (‘Sub’) and ID4 (‘All’) models from the previous tests are re-used for comparison, and six more models are trained using only subsets of the training data. These subset, outlined in Table 3.2, are dorsal, palmar, right-dorsal, left-dorsal, right-palmar, and left-palmar. These models are evaluated on each of the latter four subsets in the 11k Hands test set separately, so it can be assessed, for instance, whether training on exclusively left-dorsal images results in better performance on left dorsums than training on all dorsal images, or training on dorsal and palmar images.

Training and Evaluation

The backbone model used for feature extraction is ResNet-50 [56] with the final linear classification layer removed, to output the 2048-D feature vector. ResNet-50 was chosen because of its frequent use as a backbone in comparative papers throughout the subsequent studies. ResNet-18, ResNet-34, ResNet-101, and DenseNet-121 were tested and found to achieve worse or equivalent performance to ResNet-50 in these tasks.

These models are trained with triplet loss [142], given in Equation 3.2.2 on the 11k Hands training dataset, with a 90-10 training-validation split. The triplet loss uses a margin of $\alpha = 0.2$ following the original paper. Balanced batch sampling is performed such that every batch has 4 samples each of 8 classes (32 total images), allowing 48 combinations of matching samples, for which the negatives are chosen

using ‘triplet mining’ to produce triplets of anchors x_a , positives x_p , and negatives x_n . Following the original paper, negatives are selected using a semi-hard mining strategy for the first 100 epochs and then a hard mining strategy for another 100 epochs. The semi-hard condition is given in Equation 3.2.2 and means the negative sample is further from the anchor than the positive but within the margin. The hard condition is given in Equation 3.2.2 and means the negative sample is closer to the anchor than the positive sample. These batches are sampled ‘online’ meaning they are re-calculated at every epoch using the updated model to measure embedding distances, ensuring they get continually harder and adapt with the model.

$$\sum_i^N [||x_{ia} - x_{ip}||_2^2 - ||x_{ia} - x_{in}||_2^2 + \alpha] \quad (3.1)$$

$$||x_a - x_p||_2^2 < ||x_a - x_n||_2^2 < ||x_a - x_p||_2^2 + \alpha \quad (3.2)$$

$$||x_a - x_n||_2^2 < ||x_a - x_p||_2^2 \quad (3.3)$$

The learning rate is set to 1e-4 decaying by a third every 30 epochs, and Adam optimisation. Learning rate and optimisers are re-initialised between the first 100 semi-hard epochs and second 100 hard epochs. All images fed to the model are 256x256px resized square crops around the hand. Images with accessories such as rings were removed from both training and testing datasets to ensure the model only learns to extract features of the body, and to prevent influence on the results caused by rare accessories (easier to recognise) or common accessories (more likely to confuse). Each channel in the image was normalised according to the means and standard deviations in the training dataset. No other augmentations were applied at this stage.

The models are evaluated using a ranked-retrieval scenario wherein the test dataset is divided into probe and gallery using the process described in Section 3.1.4. Each probe image is used to rank the gallery images by the feature distance between them, then the metrics described in the next section are calculated based on the relative positioning of matching images (to the probe) and non-matching images in the ranked gallery. This evaluation protocol is used for all subsequent recognition tests in this thesis.

3.2.3 Evaluation Metrics

The metrics used to evaluate ranked-retrieval performance are accuracy, mean average-precision, mean precision, and mean recall. Each metric is measured at every rank k , denoted by $@k$, from 1 to the number of gallery samples n . For example, $metric@1$ measures the metric for only the first-ranked item for each probe, $metric@10$ measures the metric for the top 10 ranked items for each probe, and $metric@n$ measures the metric for the entire ranked list, for each probe.

Many real-world biometric systems use only the top ranked result for identification, thus rank-1 metrics are the most meaningfully interpretable. In a database search system as has been proposed, where no automated decisions will be made and instead the full rank is returned to a human for analysis, the significance of rank-1 metrics in particular is reduced, however it is still important to understand how often the system returns correct matches *first*, without falling back on human support. Another reason for choosing rank-1 metrics for analysis primarily, is that they complement well the rank- n metrics (whole ranking quality); without being defined by system requirements, other ranks (e.g. rank-5 or rank-10) metrics can be arbitrarily chosen and misleading about system performance, particularly where there are variable numbers of true matches in the database for each probe.

These metrics were measured for both the label (combination of the subject, left or right hand, and dorsal or palmar view) and the subject alone, to identify if images of the same subject in a different label are being returned at high ranks.

Accuracy@ k is the the percentage of instances (all probes) for which there is a positive match in the top k ranked images. This is most useful to know how often the top ranked image is correct (@1). It should always reach 1.0 by rank $n - p$, where p is the number of positive matches for the probe.

Precision@ k is the proportion of the top k ranked images that are a positive match for the probe, averaged across all probes (making it mean precision). This is equivalent to accuracy at rank 1, but gives more nuanced information in the following ranks where accuracy would be 1.0 whether the number of positive matches is $1/k$ or k/k . This metric does not give useful information at ranks much higher than p , as once all positive matches in the gallery have been encountered then precision can only decrease: at rank n , this value will be p/n .

Recall@ k is the proportion of all positive matches for the probe in the gallery that are captured in the top k ranked images, averaged across all probes (making it mean recall). This is useful at higher ranks to understand where the worst performing positive matches for the probe are placing. This metric is not useful at ranks lower than p , as it would not be possible to achieve high values. Recall will always increase and reach 1.0 by rank n .

Mean Average-Precision@ k (mAP) is the mean of the average-precision (AP) of the top k ranked images across all probes. The AP of the top k images is the precision at each *relevant* (positive match) rank up to k , averaged. For example, the average-precision at rank 5 if there are positive matches at rank 1 and rank 4, is the average of the precision at rank 1 (1.0) and rank 4 (0.5): 0.75. This metric gives a measure of the overall ranking quality: if all positive matches are in the top p images, then the mAP will be 1.0, whereas if all positive matches are in the bottom p images, the mAP will be roughly p/n (with a high number of images in the gallery, close to 0). Because this metric only considers the ranks where positive matches occur, it does not decrease at higher ranks as precision does. This is alternatively the area under the precision-recall curve.

Interpolated precision-recall curves show how precision changes as recall increases. Since non-relevant images in the ranking will mean a decreased precision while recall stays the same, the precision is interpolated to smooth the graph: for each recall step, the precision is set as the highest precision found at that recall step or any higher recall (to the right of it on the graph, see Figure 4.6 (d)), giving a stepped appearance. The area under this curve is the mAP. If all positive matches are in the top p ranked images, then the precision will remain at 1.0 all the way from the initial recall value to recall 1.0, and if all positive matches are in the bottom p ranked images then the precision value will remain at p/n for all recall values, since this is the final precision value and the graph shows the highest of all precisions to the right.

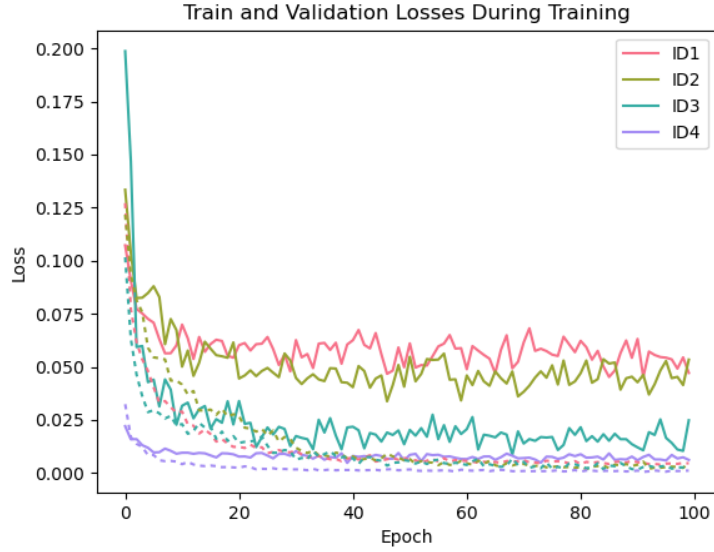
3.2.4 Results

Figure 3.5 shows the training and validation loss and validation accuracy over epochs for the four models trained with different label sets. All models displayed good training behaviour with a slow decline in loss to epoch 100 after initial large changes in loss in the first 10 epochs. The models trained with ID1 and ID2 displayed very similar loss values, with a large gap between the training and validation losses indicating a poor ability to generalise to unseen data; this gap for the model trained with ID3 is much smaller, so the performance on unseen validation data is almost equivalent to the performance on the training data. For the model trained on ID4 this gap is reduced even further, indicating a very good ability to generalise, and ID4 achieved the lowest loss values overall. The validation loss for the ID4 model is also shows significantly less fluctuation (noise) than for the other models, which shows that the model training is more stable and efficient. The validation accuracies, the percentage of all triplets in the validation batches for which the anchor is closer to

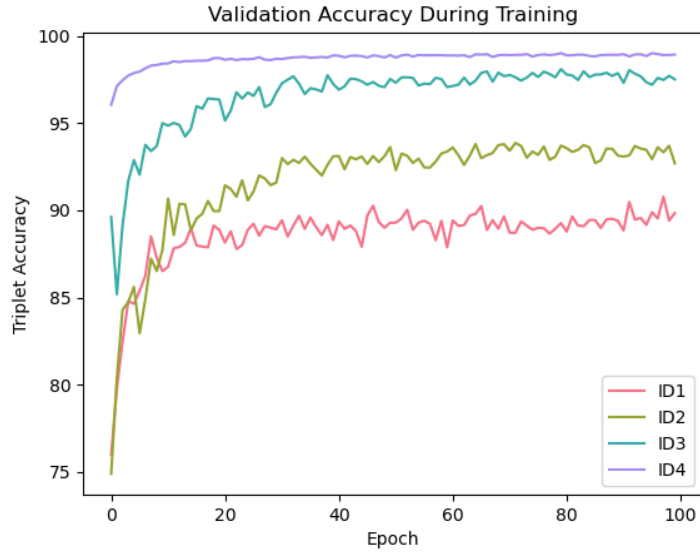
the positive than the negative, paint a similar picture. ID1 and ID2 start at a similar point but are separated by approximately 5% at epoch 100, ID3 is notably higher, and again ID4 achieves the best accuracy with a steady increase across epochs and none of the fluctuations seen in the other models. These graphs therefore suggest that training with the ID4 strategy can stabilise training and reduce over-fitting.

Ranked retrieval results for the four models are provided in Figure 3.6. Labelling solely by the subject of the image results in the lowest accuracies to rank 50 (57.1% Acc@1) and the lowest mean-average precision (0.2435) by quite a margin. Labelling with ID2, in which the left and right hands are considered different identities, produces a slight improvement to rank accuracy (64.18% Acc@1) and mAP (0.342) but largely follows the same trend in accuracy and precision-recall rates. ID3, however, sees a much larger improvement, with 77.08% Acc@1 (+19.98%) and 0.5671 mAP (+0.3236), with notable improvement to precision at midway recall rates (0.4-0.6). This could indicate that the information lost by the model when treating dorsal and palmar sides of the hand as the same identity is more significant than the information lost when treating left and right hands as the same. Since both result in significant improvements to performance it is unsurprising that ID4 achieves the best results, with 81.66% accuracy @1 and 0.6861 mAP, +24.56% Acc@1 and +0.4426 mAP higher than the subject-labelled baseline.

Table 3.3 gives the results for the subset-trained models (including the ID1 and ID4 models from above) in accuracy at rank-1 (Acc@1) and mean average-precision (mAP) on the right-dorsal (RD), left-dorsal(LD), right-palmar (RP), and left-palmar (LP) subsets of the test dataset. The dorsal model and palmar models are trained exclusively on dorsal and palmar images respectively, but left and right hands of the same subject are treated as different identities. As expected, dorsal-trained models performed better on dorsal subsets and palmar-trained models performed better on palmar subsets; it is also notable that the dorsal-trained models performed worse than the palmar subsets than the ID1 model trained on all data (a poor-performing baseline, as previously established), and vice-verse for the palmar-trained model. No benefit is seen to separately training models for left and right hands, as the left-palmar-trained model outperforms the right-palmar-trained model on right-palmar images, and the same for the dorsal models, and the right-dorsal-trained model performs better on the left-dorsal subset than on the right-dorsal subset. The ID4 model trained on all data achieved the highest Acc@1 on all subsets except left-palmar for which it achieved the second highest behind the palmar-trained model; as the palmar-trained model underperforms by -3.3% on the right-palmar subset then this could possibly be a fluke. These results could indicate that the models learn more effectively when presented with combined data (both dorsal and palmar images,

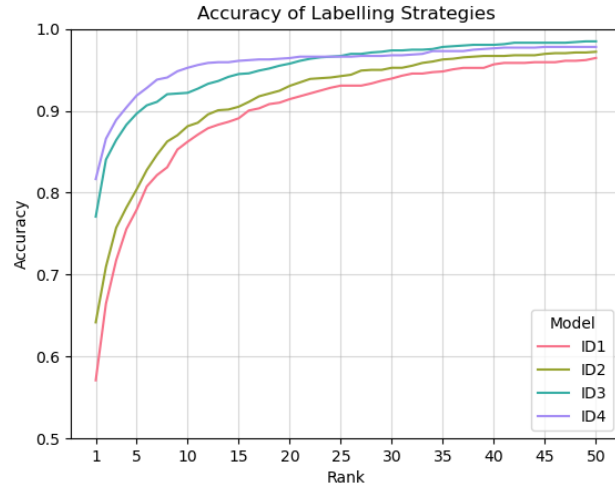


(a) Training Loss

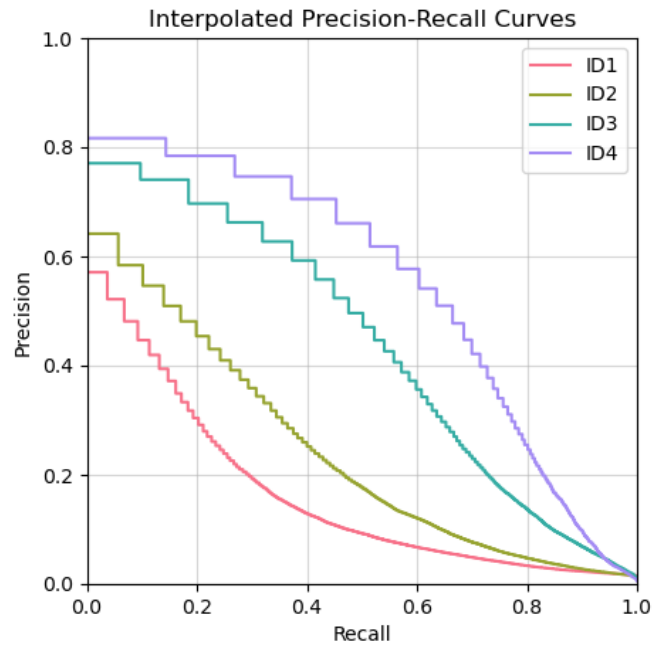


(b) Validation Accuracy

Figure 3.5: Training behaviour of the four labelling strategies in the first 100 (semi-hard) epochs: a) training (dashed line) and validation (solid line) losses, and b) validation accuracies measured as the percentage of all triplets for which the anchor was closer to the positive than the negative.



(a) Rank Accuracy



(b) Precision-Recall

Figure 3.6: Performance of the four labelling strategies on all 11k Hands test images: a) accuracy to rank 50, starting at 57.1% for ID1 and 81.7% for ID4 @1, and b) interpolated precision-recall curves, where more area under the line indicates better performance.

left and right hands) that are treated as different identities, compared to training individual models on subsets of the data (such as having a dorsal-specific or palmar-specific model). However, it is unknowable from these results whether this is simply caused by data sparsity, and if the amount of training data increased whether hand or view specific models would be able to learn more effective feature embeddings than combined models.

Model	RD		LD		RP		LP	
	Acc@1	mAP@n	Acc@1	mAP@n	Acc@1	mAP@n	Acc@1	mAP@n
All-ID1	0.6135	0.6898	0.7073	0.7769	0.6211	0.7013	0.6821	0.7635
All-ID4	0.6871	0.7447	0.7129	0.7754	0.685	0.7431	0.7123	0.7996
Dorsal	0.6421	0.7158	<u>0.7111</u>	0.7843	0.4692	0.5517	0.5684	0.6482
RD	0.5603	0.6432	<u>0.6116</u>	<u>0.6924</u>	0.4912	0.5603	0.5661	0.6481
LD	0.5971	0.6768	<u>0.621</u>	<u>0.6927</u>	0.4582	0.5391	0.5476	0.6265
Palmar	0.5665	0.6429	0.5497	0.643	0.652	0.7298	0.7471	0.8208
RP	0.5153	0.5893	0.4897	0.5757	0.5749	0.6566	<u>0.594</u>	<u>0.6884</u>
LP	0.5092	0.602	0.5422	0.6186	0.5837	0.6621	<u>0.652</u>	<u>0.7262</u>

Table 3.3: Rank retrieval results for each subset-trained model in Accuracy @ 1 and mAP @ n (final rank), on each of the right-dorsal, left-dorsal, right-palmar, and left-palmar subsets. Bold indicates the best result in the column, and underline indicates the best result in row.

3.3 Discussion

This chapter has introduced a new hand image dataset for evaluation of unconstrained hand biometric algorithms, addressing a gap in the available data. This chapter has also presented an area of confusion with the labelling of hand biometric images and empirically demonstrated which labelling strategies are the most suitable with the aim to align future hand biometric studies in this aspect.

3.3.1 Dataset

The dataset collected addresses the lack of hand pose and illumination variance within subject images found in hand images to date, which is necessary to truly understand how the hand recognition algorithms behave on ‘in the wild’ images that would be found in a forensic context. This data will finally enable the comparative evaluation of existing hand recognition methods which have been proposed for foren-

sic application in a more appropriate way (where they cannot rely on convenient image conditions for region-of-interest detection and feature extraction).

The dataset, with 100-250 images per hand (very high compared to some other datasets, e.g. 11k Hands with 28 images per hand), is one of the most within-subject diversely representative datasets to exist. It does however have limitations which must be considered. Firstly, there are only 20 subjects, or 40 hands. This is much lower than other datasets (e.g. 11k Hands with 380 hands) and limits the strength of conclusions that can be formed from evaluation on it. In biometrics, small datasets can often lead to higher-than-actual recognition accuracies as there are far fewer difficult negatives (identities that appear similar) and in a classification scenario, higher odds of a random selection being correct. A smaller dataset is also more likely to have a different statistical distribution to the true population, leading to potentially invalid conclusions. In addition to this, VarHands is limited in representation of several demographic categories, particularly age and ethnicity, and so analysis of performance on it would be unable to identify any problems with bias which could lead to discriminatory outcomes for groups of people.

Despite this, the dataset can provide insight into how various biometric methods handle real (non-augmented) changes to hand pose and illumination, and quality loss, providing direction for future research and enabling the exploration of methods to address these challenges.

3.3.2 Labelling

Understanding the impact on recognition performance of considering left and right hands, and dorsal and palmar views of the same hand, as different identities is an important first step to unconstrained hand recognition. The experiments performed in this chapter clearly demonstrate that this approach results in better modelling of biometric information, as models train inefficiently and perform poorly when they are encouraged to produce a unified feature embedding for these different hands or different sides of the hand.

However, while detection and categorisation of an image as left or right hand, or dorsal or palmar view, is simple in constrained datasets where factors about the data are known, it may not be so straightforward in unconstrained data.

Whether the hand is left or right would need to be predicted for unconstrained data. This is possible from the hand keypoints (detected locations for joints and fingertips) and angles between them, as is part of the functionality of the Mediapipe library [174], but may not always be correct and does not account for the possibility of horizontal image flipping, which is a common occurrence for images taken on

mobile devices. It may be possible to train models to extract the biometric features regardless of mirroring by applying horizontal flip augmentation during training, however it would likely take much longer for the models to converge; alternatively, and more efficiently, any image used as a search can be submitted twice, once being horizontally flipped. Doing can ensure the algorithm doesn't miss a match just because of horizontal flipping, without requiring extra training time for the models. In this sense horizontal-flip invariance is considered 'solved' for this project.

Whether the view of the hand in the image is dorsal or palmar is much more complicated for unconstrained images. Using the process outlined in Section 3.1.2 an image can be assigned 'dorsal' or 'palmar' based on the relative positions of the detected finger base knuckles, which is largely correct. In many of these images however the hand is rotated in such a way that the features typically expected of a dorsal or palmar image (e.g. full palm lines) are not visible, even to such an extent that no features of either the dorsal or palmar side are visible - the hand is entirely 'sideways'. Features can also be lost to occlusion, such as loss of the palm lines in a palmar image because of the fingers covering them. Dealing with only partial visibility of the expected features is one of the challenges of unconstrained biometric recognition, and one that will be addressed later in Chapter 5. The movement of fingers introduces more problems than simply occlusion of the palm however, as they can introduce typical dorsal-side features into palmar images (but not the other way around). When the fingers are curled, such as when someone is making a fist, a palmar view of the hand would show the dorsal knuckles (or fingernails), but using a by-view strategy of labelling as suggested above would result in a correct match between those features and a dorsal-side image being considered incorrect. This can be addressed at test-time by also considering the by-hand accuracy where dorsal and palmar views are not separated. For now, with the current hand image datasets not featuring changes to finger curl, this will not have an impact on model training, but any future attempts to train models on unconstrained data labelled by-view should pay careful attention to the impact this could have.

In addition to the findings on how data should be labelled, these experiments demonstrate that a single model trained on all subsets of the data (left and right hands, dorsal and palmar views) is at this time able to achieve better recognition performance on those subsets in tests than dedicated models trained only on that subset. This is in contrast with how many studies in the literature train hand recognition models (e.g. [1, 7, 8]), but may only be because of the limited size of training datasets currently and not hold as dataset sizes increase.

3.3.3 Conclusion

The findings in this chapter have determined that, moving forward, by-hand and by-view labelling should be used for the training and evaluation datasets, and unified models should be trained for each hand region (as opposed to training an individual model for the left and right hands or dorsal and palmar views), to maximise training efficiency and feature utility. Hand biometric systems used for unconstrained purposes should be able to handle inputs of varying illuminations and hand poses, which have not been represented in hand image datasets so far. The dataset collected in this thesis, VarHands, has enabled analysis across these attributes due to labelled illumination categories and tracked subsets of images that contain standard hand poses and hand poses with large variety in rotation, flexation, and finger position. The next chapter utilises this dataset to establish effective region detection-and-extraction methods across hand poses and evaluates existing hand recognition methods across illuminations, with proposal of illumination-normalising methods to boost performance.

Chapter 4

Region Extraction and Illumination Invariance of Hand Images

This chapter presents an analysis of whole-hand vs regional hand recognition approaches on unconstrained data for the first time, with additional insight into recognition performance cross-illumination category and with and without significant hand pose variation in the evaluation dataset. Novel methods of region-extraction that can be applied to unconstrained images are proposed. Additionally, a novel method of illumination-normalisation and skin texture enhancement for palm and knuckle images is proposed and evaluated against the use of the original colour images.

Hands could be viewed as a single biometric modality, like faces, or a collection of other modalities such as fingerprint, palm and knuckle lines, dorsal veins and more. This provides many options for hand recognition. Existing works have taken approaches that use both end-to-end deep learning on non-processed whole-hand images, or use a range of pre-processing techniques on sub-regions of the hand (most commonly palms and knuckles). Both whole-hand and regional hand recognition approaches have been able to achieve exceptionally high (100% rank-1 accuracy, 0.98 mAP [163]) on controlled and uniform hand image datasets such as 11k Hands and HKPU palm and dorsal datasets, but there is comparatively very little evaluation of these methods on unconstrained datasets due to their rarity. Performance on unconstrained datasets would be expected to be lower. As one example, Matkowski et al. use a private unconstrained hand image dataset collected from the internet to evaluate their palm recognition and achieve a best result from all compared models of 52.42% rank-1 accuracy [114], significantly lower than the standards on 11k

Hands and HKPU palm and dorsal datasets. Comparative evaluation of different hand recognition approaches on the same unconstrained dataset is lacking in the literature. Datasets would be expected to have unique biases that make comparisons of performance on *different* datasets meaningless. Being able to assess the relative performance of prominent approaches in the unconstrained domain could direct future research to the most suitable options.

In addition to this, it is desirable to understand some of the influencing factors in lower performance on unconstrained data, if any beyond the obvious loss of image quality. Three primary causes of visual differences in unconstrained images of hands are the pose of the hand, colouring of the hand due to illumination or skin tone changes (temperature, time), and long-term physical changes (wrinkling, weight change, scarring etc.) Two of these factors, pose and illumination, can be controlled for in the VarHands dataset.

The research questions to be addressed in this chapter are covered in the next section, followed by a section on the method. This chapter presents a series of research questions to be answered to address these gaps, describes the methodology used to answer them, and discusses the results from a series of qualitative and quantitative tests.

4.1 Research Questions

The subset of research questions addressed in this chapter are:

RQ2 Which whole-hand or regional approaches to hand recognition perform best on unconstrained data?

RQ3 To what degree does change in hand pose impact recognition performance?

RQ3 To what degree does change in hand pose impact recognition performance?

To answer these questions a review of the relevant literature was performed (hand recognition techniques intended to be applied to unconstrained data), from which the decision to consider and compare whole-hand recognition against contactless palm and knuckle recognition was made. Methods were devised to extract and pre-process

the relevant regions from unconstrained images, given in Section 4.3, and several models were trained and evaluated using the procedures described in Section 4.4.

4.2 Background

Both whole-hand matching, and matching based on sub-modalities of the hand such as the palm and knuckles, have achieved very good results on the constrained datasets.

The sole whole-hand recognition work at the time of this review is from Baisa et al., where they provide two iterations of a ResNet-50 derived network with branching attention mechanisms. The first iteration, GPA-Net (Global and Part Aware Network), combines feature vectors for the global image plus p (4) partitions of the image. GPA-Net achieved 94.84% rank-1 accuracy and 0.9583 mean average-precision (mAP, explained in Section 3.2.3 further on) on the dorsal images of 11kHands, and 95.78% rank-1 accuracy and 0.9626 mAP on the palmar images. The second iteration, MBA-Net (Multi Branch with Attention Network) adapts ResNet-50 with a channel attention module and a spatial attention module; it combines the feature vectors output by each module plus the non-attention global vector. MBA-Net achieves 97.08% rank-1 accuracy and 0.977 mAP on the dorsal images of 11kHands, and 97.74% rank-1 accuracy and 0.9813 mAP on the palmar images, as well as 95.12% rank-1 accuracy on Poly-U hand dorsal dataset.

In contrast to the few works on end-to-end whole-hand recognition, there are many more studies into using sub-regions of the hand in various combinations for recognition. Most of all of these select areas of interest on the hand and disregard other areas; Matkowski et al. however use a combination of palm and fingers for palmar-side hand recognition which encompasses the whole-hand. They separately region-detect and embed vectors for the five fingers and palm, and combine the results for ranked retrieval. They do not evaluate on 11kHands, but achieve 100% rank-1 accuracy on Poly-U and IITD palmar-side hand image datasets, and 52.42% rank-1 accuracy on their own unconstrained palmar-side hand dataset. In contrast, their own prior work which used only the palm, achieved 98.69%, 99.77%, and 41.92% rank-1 accuracies on IITD, Poly-U, and their unconstrained dataset, respectively, outperforming most other palm recognition methods. These results indicate that using more information (the fingers in addition to the palm) can improve recognition performance.

The next most common hand region to use for recognition is the dorsal-side knuckles (and sometimes the palmar-side knuckles). Vyas et al. achieved 99.81% rank-1 accuracy on 11kHands dorsal-side images, and 99.33% on Poly-U hand dorsal, using

a combination of the base, major, and minor knuckles of the four non-thumb fingers. Their analysis showed that the more information used, the better the performance: the best rank-1 accuracy for a single finger was 98.83%.

4.3 Image Pre-Processing

This section describes the methodology used to detect hand landmarks, extract various image regions around the whole-hand, palm or dorsum, and knuckles, and to normalise image colour and illumination. The region extraction for palms and knuckles was performed once to produce separate on-file versions of the dataset in order to speed up data loading for training and testing; extraction of each region takes approximately 0.2s. The process for illumination normalisation was done to produce separate on-file versions of the 11kHands datasets, but performed live as part of the augmentation pipeline for the VarHands dataset. Implementing all steps as part of the data augmentation pipeline would be possible.

4.3.1 Hand Landmark Detection

The following work on locating and extracting hand regions is based primarily on the use of hand landmark coordinates detected using MediaPipe’s hand landmarker [174]. The landmarker includes a palm detection model with an average precision of 95.7%, prediction of whether the hand is left or right, and 3D keypoint localisation of 21 hand landmarks with tracking for video inputs. Subjectively, the performance of this method was deemed to be sufficient as it found all keypoints accurately in 11kHands and rarely failed on VarHands (as mentioned in Chapter 3, keypoint detection was used to select video frames for the VarHands image dataset, so occasional failure had no consequence). These keypoints detections are demonstrated on a VarHands image in Figure 4.1.

The purpose of these landmarks is to enable gesture analysis, so the coordinates they detect are often approximations and are inconsistent in specific pixel position over the region they denote. In biometrics, alignment of images around the same accurate pixel centre point or corner points can greatly improve performance. The following sections describe the methods to extract the regions-of-interest in the simplest way possible (“naïve”), and to further refine region localisation and sizing using the detected keypoints as a starting point.

Except where otherwise stated, all of the methodology used in this section is original. The majority of region localisation and extraction methods in the existing literature are applied to controlled images and so rely on certain image conditions,

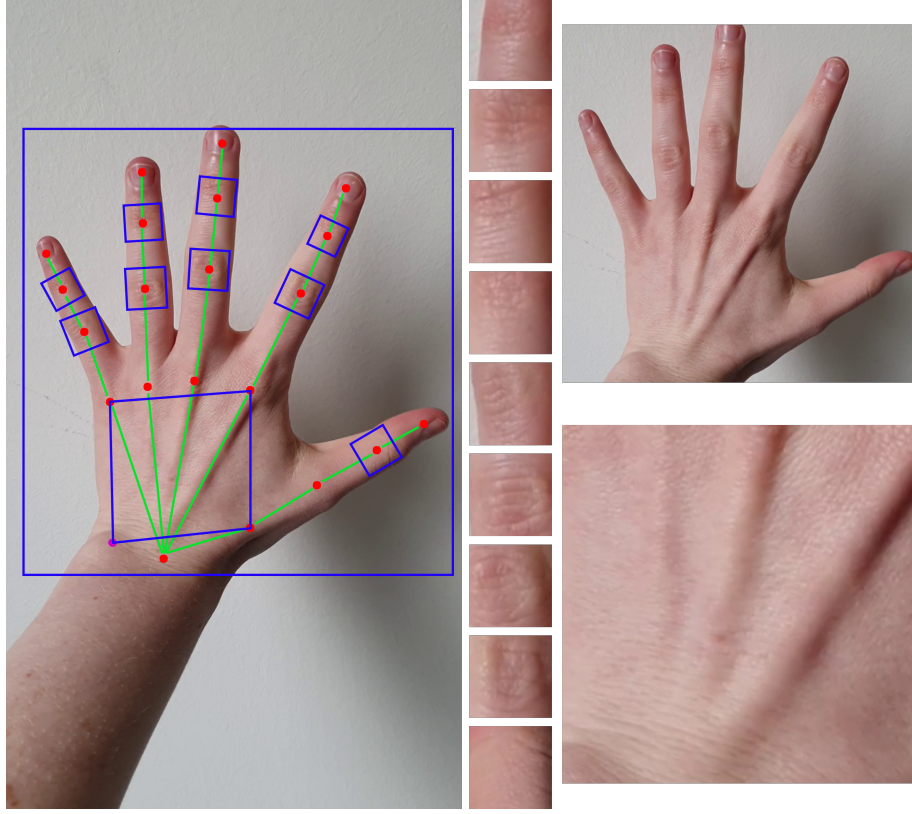


Figure 4.1: Hand landmarks and extracted regions.

like having a simple background to threshold against or spread fingers for finger valley detection. These methods were developed to use only the detected keypoints of the hand for generalisation to all kinds of image backgrounds and hand poses.

4.3.2 Whole-Hand

So that the image is not distorted when resized for model input, a square region around the hand is cropped for the whole-hand images. The centre of the region is calculated as the centre of the outermost hand keypoints: the midpoint between the minimum and maximum x and y coordinates of all hand keypoints. The dimension of the square region (width and height) is calculated as the distance between the minimum and maximum x and y coordinates plus a padding of 5% of that distance. If this exceeds one of the original image dimensions, it is set to the minimum of the image height and width instead. The cropped image is resized to 512x512 pixels

before being saved, which is smaller than the original size of the region in most images. This imposes an upper bound on the model input size of 512x512, however in practice input images this large result in memory errors and long training times - so the images are often further reduced in size to 324x324 or 256x256 (see Section 4.4).

4.3.3 Palm Extraction

The palm/dorsum region is extracted from each region using the wrist and metacarpophalangeal (MCP) joint (base knuckle) keypoints. The naïve method positions a square crop around the palm on the rotated raw image, while the advanced option uses spatial projection to better align the palm regions across images.

Naïve Method

The naïve method of palm extraction first rotates the image so that the line between the wrist a and middle finger MCP b is vertical with the wrist point below. An alternative method of making the line between the little finger and index MCP horizontal was tested, but resulted in more extractions with off-hand background sections included. The angle of rotation is calculated as $\text{atan2}(a_y - b_y, a_x - b_x)$ (arctan with respect to circle quadrants), converted to degrees and reduced by 90. The rotated image is padded to prevent information loss at the edges, with the new image dimensions as $w_{new} = w_i * \sin(\theta) + h_i * \cos(\theta)$ and $h_{new} = w_i * \cos(\theta) + h_i * \sin(\theta)$, the horizontal distance between the leftmost and rightmost corner, and the vertical distance between the topmost and bottommost corners. The resulting rotation matrix is applied to both the image and the relevant landmark coordinates. After rotation, the midpoint of the wrist - middle MCP line is used to centre a square crop with dimension equal to the distance between the index and little finger MCP joints plus 5%. The cropped image is resized to 512x512 pixels.

Spatial Projection

Since the palm is not a perfectly square region, and since rotation of the hand relative to the camera can cause the palm to appear skewed, perspective transformation is used to spatially project the palm to a square region from four corner points. Three of these corner points are the little finger MCP, index finger MCP, and thumb carpometacarpal (CMC) joint, which joins the thumb to the hand close to the wrist. The wrist and thumb CMC points are then used to calculate the fourth corner, which is not positioned near any detected landmarks in the lower thumb-opposing edge of

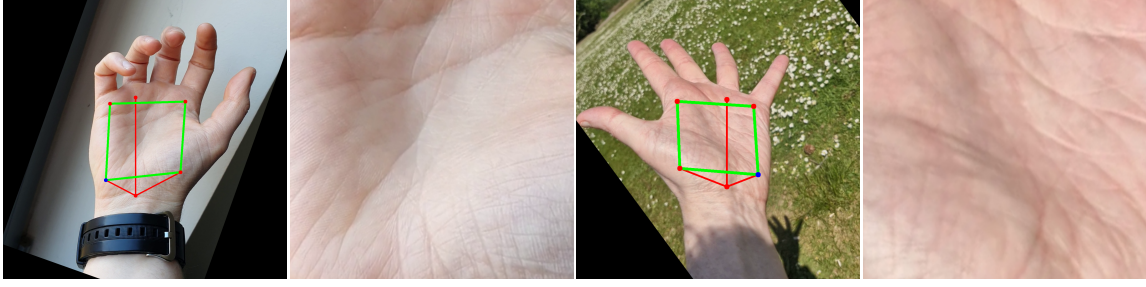


Figure 4.2: Two instances of palm projection. On the left the pre-existing landmarks (red circles), calculated lines (red), calculated fourth landmark (blue circle), and the boundary lines (green) are overlaid on the rotated image. On the right are the resulting palm extractions.

the palm. First, the angle of the fourth corner from the wrist point is calculated as the vertical inverse of the wrist a - thumb CMC b angle, using the formula given in Equation 4.1. Then the distance from the wrist to the fourth corner is taken as two thirds of the distance from the wrist to the thumb CMC, given in Equation 4.2. This process is shown in Figure 4.2; on inspection, it was found to be consistent in both the constrained and unconstrained datasets, which contained a range of hand poses and rotations, never projecting the fourth corner point off-hand.

$$\theta = \text{atan2}(a_y - b_y, a_x - b_x)$$

$$\theta = \begin{cases} \pi - \theta, & \text{if } \theta \geq 0 \\ -\pi - \theta, & \text{otherwise} \end{cases} \quad (4.1)$$

$$d = \text{int} \left(\frac{2}{3} \sqrt{(a_x - b_x)^2 + (a_y - b_y)^2} \right) \quad (4.2)$$

4.3.4 Knuckle Extraction

Nine knuckle points from each hand are extracted: the distal interphalangeal (DIP) and proximal interphalangeal (PIP) joints (minor and major knuckles) of the four fingers and the interphalangeal joint (also major knuckle) of the thumb; the base knuckles were excluded due to inconsistency in the detected landmarks that resulted in many off-target extractions, and a lack of strong visual information found in those regions compared to other knuckles.

Naïve Method

The naïve method uses the detected keypoint of each of the knuckles as the centre of a square region. Each knuckle is assigned a 'prior point': for minor knuckles it is the major knuckle on the same finger, for major knuckles the base knuckle on the same finger, and for base knuckles it is the wrist. For each knuckle independently, the raw image is rotated so that the knuckle point a and 'prior point' b line is vertical with the prior point below. This ensures the view of the knuckle remains aligned even if fingers are bent or the hand is rotated. The same method of image rotation and padding is used as for the palm.

The region dimension (height/width) is calculated as 10% of the original raw image's smaller dimension plus that value multiplied by the wrist-relative depth coordinate z (see equation 4.3). Multiplication by the wrist depth coordinate treats it as an approximation of distance from the camera, which is suitable in most instances but not entirely correct - sometimes it results in too-close images of the knuckle which excludes some information, and sometimes it results in too-far images of the knuckle which contain background or some of the neighbouring finger. Once the region has been cropped using the centre and box dimension, the resulting image is resized to 80x80 pixels, which is slightly smaller than most knuckles in the raw images (~ 100 px width).

$$d = 0.1(\min(h_i, w_i)) + 0.1z(\min(h_i, w_i)) \quad (4.3)$$

Dynamic Sizing

The naïve method calculates the knuckle region in the image to be a function of 10% of the image size and the depth coordinate z (see equation 4.3), however this results in variation in the scale of extracted knuckle from image to image, particularly where the distance of the hand to the camera differs, since z is wrist-relative and not absolute. In order to have a more consistent knuckle scale the box dimension is calculated from the measured finger size in the image.

To measure the finger size, the roughly localised knuckle image (using the naïve method) is converted from RGB (red, green, blue) colour space to HSV (hue, saturation, value) and its hue value is used to separate the finger from the background; assuming that while background pixel intensity may be similar to the hand pixel intensity, they will be of differing colours (which is lost in greyscale conversion). The hue image is binarised with Otsu's algorithm [129] to find the optimal global threshold value, and contour detection [154] is used to fill in the primary finger shape and remove binarisation defects from illumination. Then, the leftmost edge of the finger

is taken as the index of the first black pixel left of the centre point, and the right edge is the first black pixel right of the centre point. The new extract dimension is the distance between the right and left edge; if this distance is too small (defined as 1/4 of the dimension calculated using the naïve method) then it defaults to the naïve method.

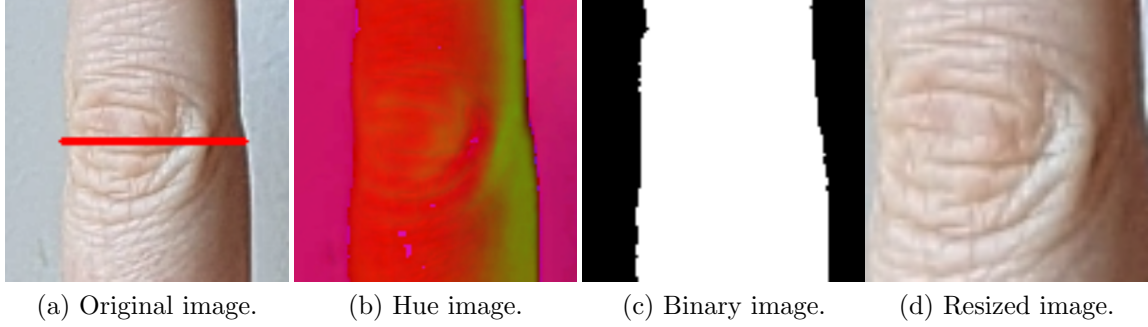


Figure 4.3: Knuckle sizing steps: (a) the original image with the measured width overlaid, (b) the HSV (hue, saturation, value) colour space, (c) the binarised image, and (d) the resulting resized image.

In practice this method still failed on some images, in which cases the naïve sizing was reverted to, resulting in some poorly sized knuckle extractions in the unconstrained test datasets, however applying this method greatly reduced the incidences of them.

Texture-Based Knuckle Centre Localisation

For more accurate alignment of the extracted knuckle regions, the methodology proposed by Kumar [95] was followed, which takes advantage of the circular nature of knuckle creases by using texture-based orientation fields to locate the 'true' knuckle centre.

$$\begin{aligned}
 Gy &= \sum_{u=8i}^{8(i+1)} \sum_{v=8j}^{8(j+1)} 2 \cdot Sx_{u,v} \cdot Sy_{u,v} \\
 Gx &= \sum_{u=8i}^{8(i+1)} \sum_{v=8j}^{8(j+1)} Sx_{u,v}^2 - Sy_{u,v}^2 \\
 O_{i,j} &= \frac{1}{2} \text{atan2}(Gy, Gx)
 \end{aligned} \tag{4.4}$$

$$F_{i,j} = \sum_{u=1}^n \sum_{v=1}^m \cos \left(\left| O_{u,v} - \frac{1}{2} \text{atan} \left(\frac{u-i}{v-j} \right) \right| \right) \quad (4.5)$$

First, the image is rotated and a roughly-localised knuckle image is extracted following the naïve method described above. The knuckle image is illumination normalised using luminance blur (described in the next section), converted to greyscale, and Gaussian blurred, and the Scharr gradients Sx and Sy are calculated (Figure 4.4 (b)). The image is then divided into 8x8 pixel blocks, setting block dimensions n and $m = 16$ with 256 blocks total; for each block $O_{0 < i \leq n, 0 < j \leq m}$, the gradient orientation is calculated using equation 4.4 (Figure 4.4 (c)). The 'fitness score' $F_{0 < i \leq n, 0 < j \leq m}$ of each block is then calculated according to equation 4.5; this results in higher values for blocks b in which the angles to the other blocks b' in coordinate space most align with the orientations of gradients in those blocks b' , or in simplistic terms, the block to which most other blocks are 'pointing' (Figure 4.4 (d)). The new knuckle centre point is taken as the central coordinate of that block in the extracted image, translated back to the original image by resizing to the dimensions produced in the naïve method and adding the offset.

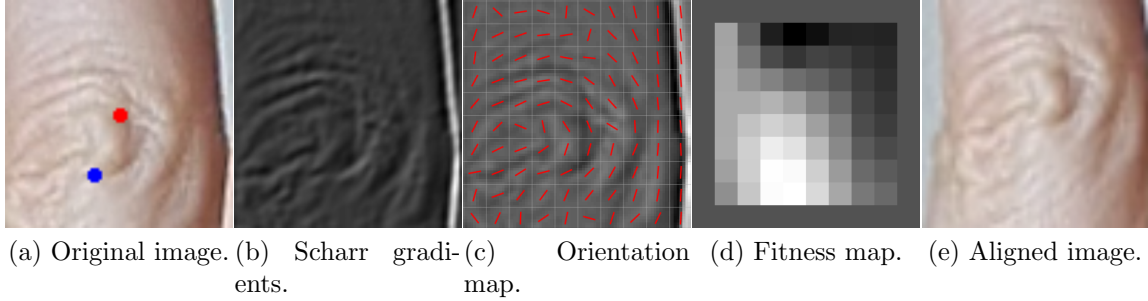


Figure 4.4: Knuckle centre localisation steps: (a) the original image with the original landmark (red) and new knuckle centre (blue) overlaid, (b) the Scharr gradient image, (c) the orientation of gradients grid map, with the red lines perpendicular to the gradient direction, (d) the fitness map, excluding the outermost squares in the grid, and (e) the resulting realigned image.

A square region sized according to the previous steps is extracted around this new centre. This process varies in effectiveness depending on the knuckle flexion and prominence of the knuckle creases. For some subjects in the VarHands dataset, the knuckle crease patterns were not circular enough to result in a correctly placed

centre point from this process. Visualisation of the steps described is given in Figure 4.4.

4.3.5 Illumination Normalisation and Texture Enhancement

In order to produce clear images of the skin texture in the images (palm lines, knuckle creases) the images must first be made colour and illumination invariant, and then the correct textural information must be enhanced over image artefacts such as noise. The reason for colour normalisation in addition to illumination normalisation (removal of shadows), is that the apparent skin colour of an individual may change even throughout a day due to temperature, blood flow, illumination, or other factors. Additionally, while skin colour may be a useful feature in a small-scale dataset (like the ones used for model training, thus the models may not learn to disregard it) it is not an appropriate feature for open-set recognition as it is non-unique and has limited variation among the population.

Three methods were investigated for illumination-invariance, two from prior biometric literature. Block-wise mean subtraction is a simple method used in several knuckle recognition papers [92, 99, 123, 72] that involves portioning a single-channel greyscale image into blocks (in this implementation, of 8x8px for an 80x80px image), subtracting the mean pixel value of each block from every pixel in the same block, then re-normalising the pixel range. The self-quotient image, proposed for facial recognition [164], and used by Kumar and Wang [97] for knuckle recognition, uses edge-preserving filters and the Retinex algorithm to eliminate lighting effects and normalise colour in the images. The third method, Gaussian blur of the luminance channel, was taken from research into vein enhancement in eye fundus photographs [136] and has not previously been used for texture enhancement in hand images. This involves first converting the image to the YCrCb colour space (luminance, chroma-red, and chroma-blue), then applying a Gaussian blur to the luminance channel (with a sigma of 3 for 80x80px knuckle images and 7 for 512x512px palm images) before conversion back to RGB colour space. This method produced the best and most consistent results and so was used going forward. In addition to these methods, deep-learned intrinsic image decomposition [110] - separating the object component of the image from the lighting component - was looked into, however the type of training image required for hands doesn't exist and would have been out-of-scope to produce. The conditional generational-adversarial network (cGAN) approach to knuckle image illumination-normalisation proposed by Kim et al. [87] was not done for the same reason.

A total of nine different unique steps, in various combinations, were tested for tex-

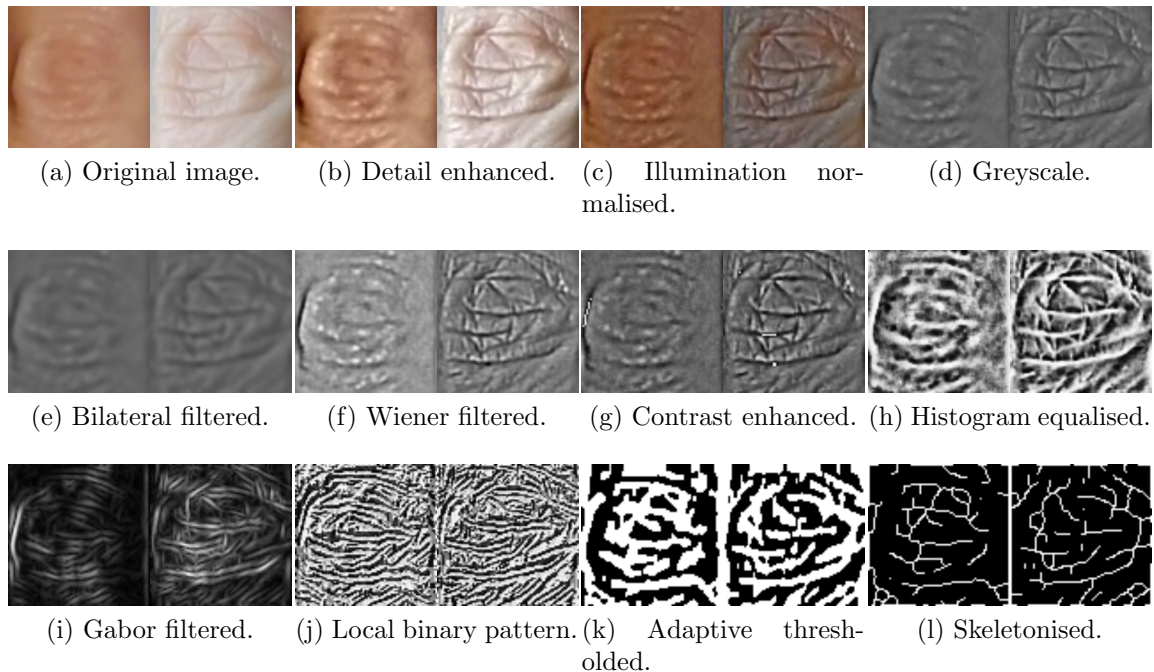


Figure 4.5: Image processing methods applied to two different images of the same knuckle. Methods (a) to (d) are applied sequentially to the result of the previous. Methods (e) to (k) are applied to the the output of (d), the illumination normalised greyscale image. Method (f) is applied to the output of (k).

ture enhancement. Bilateral filtering, which smooths images while preserving edges, was tested for noise reduction. After noise reduction, four methods were applied for enhancement of the remaining image texture: contrast enhancement using image morphology, wherein the original image has added to it the top-hat morphological transformation, then subtracted from it the bottom-hat transformation [34] (this is often used for medical x-ray enhancements); Wiener filtering [55]; image sharpening using edge-preserving detail enhancement [47]; and histogram equalisation, specifically contrast-limited adaptive histogram equalisation (CLAHE) [179]. In addition to texture enhancement, various image binarisations were tried: Gabor filtering at 6 different orientations (a method often used for fingerprints as it enhances ridge-and-valley structures [170], however is very sensitive to the scale of these structures so worked less well for multi-scale patterns like knuckle creases); local binary pattern applied on the Gabor filtered image [127], which is a texture descriptor for each pixel based on its relative intensity to its neighbours; and adaptive thresholding and

skeletonisation of the images. Skeletonisation, if successful, can be a precursor step to graph representations of the patterns, which have been used in dorsal hand vein recognition papers [107]. The result of all of these processes is given in Figure 4.5.

Figure 4.5 gives the results for two images of the same knuckle in different illuminations, to demonstrate which methods successfully result in similar representations and which are ineffective, having noticeable differences between the two. As can be seen, Gabor filtering, local binary pattern, thresholding, and skeletonisation produced variable results for images of the same knuckles, so were disregarded as options. Noise removal using bilateral filtering was found to have a negligible impact on the resulting image and in some cases removed the important textural information. Enhancement using Wiener filtering, contrast enhancement, and histogram equalisation was found to largely re-accentuate the illumination effects over the intrinsic image texture. Using these steps resulted in no improvement to performance and only added computational complexity. The sharpening using detail enhancement step however, when performed first on the colour image before illumination normalisation, did see an improvement in model performance. As a principal, if the resulting images cannot be used for manual feature extraction (e.g. graph representation from skeletonised images), then it is best to provide the model with less-processed images containing as much original information as possible.

The final pre-processing pipeline used was detail enhancement of the colour image, followed by illumination-normalisation by Gaussian blur of the luminance channel in YCrCb colour space, followed by conversion to greyscale colour space.

4.4 Methodology

To answer Research Question 2, outlined in the introduction of this chapter, five models were trained on the 11k Hands training dataset (obtained using the procedure described in Chapter 3 Section 3.1.4) and evaluated on the 11k Hands testing dataset as a baseline, VarHands, and VarHands-Easy. These models were (ResNet-50 [56]) RN50-Crop, RN50-Palm, RN50-Knuckle, (Multi-Branch with Attention Network [8]) MBA-Crop, and MBA-Palm. Additionally, the combination of the knuckles with the palm using both ResNet-50 networks is evaluated. The 'Crop' models refer to the whole-hand, around which the image is cropped. To answer Research Question 3 the performances of the models on both VarHands (lots of hand pose variation) and VarHands-Easy (no hand pose variation) was compared. To answer Research Question 4, models were evaluated cross-illumination, and performance was compared for the use of the illumination-normalisation model and without.

The models are evaluated on a ranked retrieval task. This uses feature-distance

(Euclidean) between the image embeddings produced by the model to rank a gallery of images relative to the probe image, such that the closest image in embedding space is at rank 1, and the furthest of n is at rank n . This is an appropriate evaluation task for open-set recognition problems (rather than closed-set, where the image subject is guaranteed to be part of a limited-size database). Assignment of the test dataset samples into the probe or gallery sets was random but ensured that all probe samples had at least one matching image in the gallery, though the exact number of true matches per probe varies, which is in contrast to real life where there could be no matching samples in the gallery. Having at least one sample per probe allows the evaluation metrics described in Section 3.2.3 to be used without issue to describe the general quality of the image rankings.

This task is most analogous to the described application scenario, in which the algorithm is used as a tool for searching, sorting, grouping, and retrieving information to be further analysed by a human, without any classification or other automated decision making.

4.4.1 Model Training

Two different model structures were implemented for these experiments: ResNet-50, which has seen effective use in biometric recognition tasks, and MBA-Net, which was proposed by Baisa et al. for whole-hand recognition and achieved exceptional results with minimal image pre-processing [8].

For both networks, the input images were normalised according to the mean and standard deviation of each image channel in the training dataset, and no further augmentation was applied except for random cropping to a minimum of 80% of the original image for knuckles, to improve robustness to scale changes which occur occasionally in the unconstrained datasets due to errors in sizing the knuckle region. The images were resized to 324x324px (the minimum input size for MBA-Net) for the whole-hand crop and palm regions, and 80x80px for the knuckle region. MBA-Net was not trained on the knuckle images as it would require increasing their size by a factor of 3, and because it takes significantly longer to train than ResNet-50 and there are nine times as many images for the knuckles. For ResNet-50, the final linear layer was removed so as to output the prior 2048-D feature vector. For MBA-Net, the output feature vector is a 6144-D concatenation of six 1024-D feature vectors, from each of the branches. These feature vectors were L2-normalised.

The models were trained with triplet loss in batches of size 32, made up of 4 samples each of 8 classes, to an effective batch size of 48 ($8 * 4C2$) triplets (anchor-positive pairs). The negative for each anchor-positive pair was selected from the

non-matching 28 images using the triplet mining strategies outlined in the FaceNet paper which proposed triplet loss [142]: semi-hard mining first (random selection of a negative for which the embedding distance is further from the anchor than the positive, but within the margin), followed by hard mining (random selection of a negative which is further than the anchor-positive distance plus the margin). The margin used was 0.2, from the original paper.

The models were trained for 200 epochs, 100 epochs each with semi-hard and hard triplet mining, with a weight decay rate of 0.0001, a learning rate of 0.0001 multiplied by 0.3 every 20 epochs, and Adam optimisation. Training was done using the Pytorch Lightning framework, with a 0.9-0.1 training-validation split and validation loss used to checkpoint the best performing model throughout training.

4.5 Results

Model performance in these tests was evaluated using the ranked-retrieval metrics described in Chapter 3 Section 3.2.3. Section 4.5.1 gives the results for the ranked-retrieval tests and compares the overarching hand recognition approaches (whole-hand, palm, knuckle, and palm + knuckle). Section 4.5.2 then gives results for the cross-illumination test and shows the impact that illumination and hand pose variation are having on recognition performances.

4.5.1 Ranked Retrieval

Table 4.1, Table 4.2, and Table 4.3 give the accuracy at rank-1 and mean average-precision at the final rank on the three evaluation datasets, each as their dorsal subset, palmar subset, and whole set. Figure 4.6 (a) shows the accuracy of the models from rank-1 through to rank-100, while Figure 4.6 (b) shows the precision-recall curves of the models. For both sets of figures the solid lines indicate the results where only the same label is considered a match, and the transparent dashed lines indicate the results if all four labels belonging to the same subject (left dorsal, right dorsal, left palmar, and right palmar) are considered to match; generally, subject-wise matching means higher accuracy (presence of at least one match) but lower mean average-precision (ranking quality).

Firstly, the MBA-Net results largely align with those obtained in the original paper. They achieve an average of 97.08% rank-1 accuracy across the left and right hand subsets of the dorsal images and 97.74% for the palmar images in 11kHands, compared to our 95.54% rank-1 accuracy on the 11kHands dorsal images and 96.76% rank-1 accuracy on the palmar images. Some discrepancy in the results would be

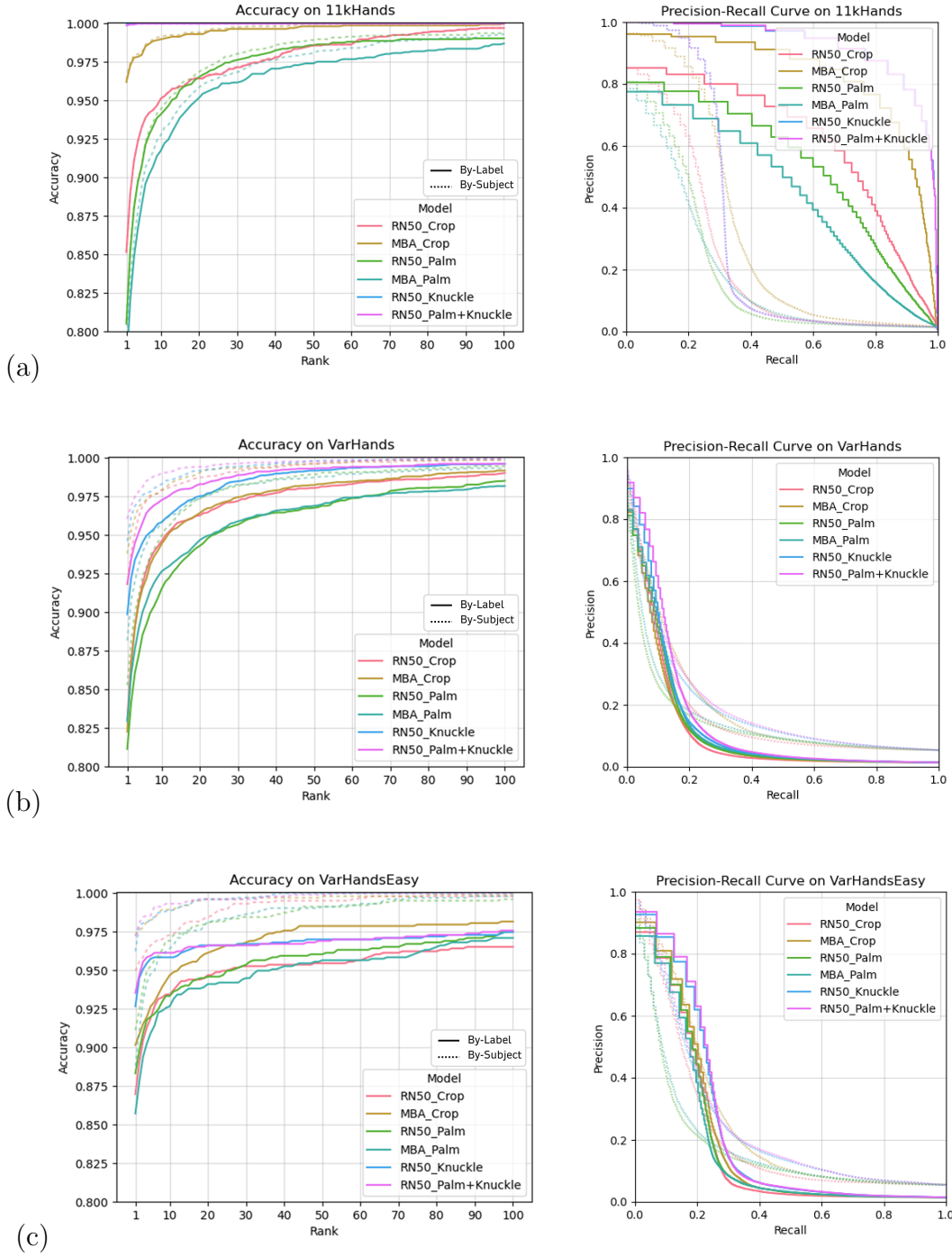


Figure 4.6: Rank-accuracy (a,b,c) and precision-recall (d,e,f) graphs for the 11kHands (a,d), VarHands (b,e), and VarHands-Easy (c,f) datasets.

Model	11kHands		11kHands-Dorsal		11kHands-Palmar	
	Acc@1	mAP@n	Acc@1	mAP@n	Acc@1	mAP@n
RN50-Crop	0.8516	0.7509	0.8463	0.7805	0.858	0.7393
MBA-Crop	0.962	0.9098	0.9554	0.9122	0.9676	0.9099
RN50-Palm	0.8049	0.678	0.785	0.6818	0.8238	0.6811
MBA-Palm	0.7752	0.595	0.7678	0.623	0.7938	0.5964
RN50-Knuckle	1.0	0.9896	1.0	0.9878	1.0	0.9917
RN50-Palm+Kn	0.9988	0.9887	0.9975	0.9857	1.0	0.9926
RN50-Knuckle-II	0.9983	0.987	.	.	.	.

Table 4.1: Accuracy at rank-1 and mean average-precision at the final rank of the dorsal, palmar, and combined sets of the 11kHands testing dataset.

Model	VarHands		VarHands-Dorsal		VarHands-Palmar	
	Acc@1	mAP@n	Acc@1	mAP@n	Acc@1	mAP@n
RN50-Crop	0.8302	0.1416	0.8888	0.1882	0.8722	0.1826
MBA-Crop	0.8227	0.1404	0.8665	0.1854	0.8614	0.1895
RN50-Palm	0.8114	0.1533	0.8894	0.2268	0.7823	0.1528
MBA-Palm	0.8293	0.1607	0.8881	0.2343	0.8006	0.16
RN50-Knuckle	0.8987	0.1641	0.9271	0.224	0.907	0.193
RN50-Palm+Kn	0.9184	0.1794	0.9506	0.2437	0.9209	0.2014
RN50-Knuckle-II	0.844	0.1952	.	.	.	.

Table 4.2: Accuracy at rank-1 and mean average-precision at the final rank of the dorsal, palmar, and combined sets of the VarHands dataset.

Model	VarHandsE		VarHandsE-Dorsal		VarHandsE-Palmar	
	Acc@1	mAP@n	Acc@1	mAP@n	Acc@1	mAP@n
RN50-Crop	0.8698	0.24	0.9349	0.2907	0.8797	0.2729
MBA-Crop	0.9016	0.2539	0.9372	0.3051	0.9044	0.2937
RN50-Palm	0.8833	0.245	0.9326	0.2904	0.8764	0.2673
MBA-Palm	0.8573	0.2356	0.9093	0.2956	0.8517	0.2545
RN50-Knuckle	0.9267	0.282	0.9488	0.3418	0.9259	0.3009
RN50-Palm+Kn	0.9354	0.2853	0.9605	0.3418	0.9341	0.3044
RN50-Knuckle-II	0.9701	0.5915	.	.	.	.

Table 4.3: Accuracy at rank-1 and mean average-precision at the final rank of the dorsal, palmar, and combined sets of the VarHands-Easy dataset.

expected for three reasons: one, even though the 11kHands dataset was split into

train and test in the same proportions, this was done randomly so would result in different images being assigned to each set; two, that this experiment uses a 50-50 split of the test set into probe and gallery sets, whereas they only use a single image of each subject in the test set as the gallery, which could result in lower average rank-1 accuracy but higher mean average-precision; and three, that they evaluate separately on the left and right hand subsets whereas in this experiment the left and right hand were mixed - since left and right hands of the same subject have a similar visual appearance they function as difficult imposter samples for each other and thus can drive down accuracy scores.

MBA-Net greatly outperformed ResNet-50 on whole-hand image recognition on the 11kHands test dataset, achieving +11.04% rank-1 accuracy and +0.1589 mAP. This was not repeated on the VarHands dataset, on which ResNet-50 achieved +0.75% rank-1 accuracy and +0.0012 mAP over MBA-Net; ResNet-50 also achieved higher rank-1 accuracy in both dorsal and palmar subsets but lower mAP on the palmar subset. On VarHands-Easy, MBA-Net marginally outperformed ResNet50, primarily on the palmar subset in which it achieves +2.47% rank-1 accuracy and 0.0208 mAP. In Figure 4.6, the MBA-Net and ResNet-50 models are only strongly separated in the 11kHands evaluation, and have almost identical precision-recall curves and accuracy through to rank-100 on VarHands. On palm images, ResNet-50 performed better than MBA-Net on 11kHands, with +2.97% rank-1 accuracy and +0.83 mAP, and similar relative performances on both dorsal and palmar subsets. Inversely to the whole-hand models, for the palm models MBA-Net does better than ResNet-50 on Varhands but worse on VarHands-Easy.

The palm models performed worse than the whole-hand models on 11kHands, with best (out of ResNet-50 and MBA-Net) mean average-precision of 0.678 compared to 0.9098. This could be explained by inconsistent hand positioning in the 11kHands dataset, which results in many images where the palm region is 'cut off' by the bottom edge of the image, resulting in large black-pixel areas with no information in the palm extractions. On VarHands, the palm models achieve lower rank-1 accuracies than the whole-hand models (82.93% compared to 83.02%) but higher mean average-precision (0.1607 compared to 0.1416), indicating more matching images are ranked highly when considering the palm/dorsum region only. On VarHands-Easy, where there is very minimal pose variation, the whole-hand models perform better than the palm models in both rank-1 accuracy and mean average-precision.

The knuckle model performs significantly better than both the whole-hand and palm on 11kHands, achieving a rank-1 accuracy of 100% and mean average-precision of 0.9896 compared to next-best 0.9098. It also performs better on VarHands, with

89.87% rank-1 accuracy compared to next-best 83.02%, and 0.1641 mAP over 0.1607. On VarHands-Easy, the Knuckle model has a larger margin of improvement in mAP over the next best, with 0.282 compared to 0.2539. The combined knuckle and palm model generally performs worse than the sole knuckle model on 11kHands, failing to achieve 100% rank-1 accuracy, except on the palmar image subset. On VarHands however the combined model achieves the best performance, with a rank-1 accuracy of 91.84% and mean average-precision of 0.1794.

All models performed roughly equivalently on 11kHands-Dorsal and 11kHands-Palmar, obtaining higher rank-1 accuracy on 11kHands-Palmar but lower mean average-precision. This could be because of the inconsistent placement of palms in the 11kHands-Palmar subset, which does not happen in 11kHands-Dorsal as much, meaning when the palm is visible there are high accuracy rates but a loss in mAP due to failing to recognise the misplaced palms. On VarHands, all models perform better on the dorsal subset than the palmar subset. This would be expected for the knuckle models as the dorsal-side knuckles have much more visible textural information in low-quality images, but the biggest performance difference seen is with the palm model. One possible cause for this is the pose variability, where curling the fingers in palmar images will obscure the palm region, while in dorsal images there will be a much more consistent view of the dorsum.

The difference in performance when treating only the same class label as a match, versus treating all class labels belonging to the same subject as a match, can be seen in Figure 4.6. On 11kHands, the rank-1 accuracies are generally the same for label and subject, whereas there is a large separation between label and subject precision-recall curve volumes (mAP) - as would be expected. This would be the case where the model is projecting, e.g., the right hand of a subject as far from the left hand as every other hand; it would be correctly separating the difficult examples of non-matching hands based on the unique textural information and not using the colour and illumination information. The opposite of this is seen in the VarHands and VarHands-Easy datasets: there is a similar mean average-precision for label and subject (noting that for subject, the minimum achievable precision is about four times higher), but a significant increase in accuracy through the first 100 ranks. This indicates that in a significant amount of instances the non-matching images from the same subject are being projected close to the probe image.

4.5.2 Impact of Illumination and Hand Pose

The impact of hand pose and the impact of image colour/illumination are evaluated together as they can cause compounding effects on the features extracted from the

image. Comparing ranked-retrieval results on the VarHands and VarHands-Easy datasets allows the evaluation of the influence of hand pose variation on recognition performance. Since the VarHands-Easy dataset is a manual selection of images from the VarHands dataset for which the hand is parallel to the camera and the fingers are all uncurled, it has very minimal variation in hand pose. If hand pose is having a significant impact, it would be expected that performance would be notably better on the VarHands-Easy dataset compared to the VarHands dataset. The results in Figure 4.6, which consists of models trained on the colour images with no illumination-normalisation, indicate a slight improvement to mean average-precision through the volume of the precision-recall curves, but notably lower accuracy on VarHands-Easy through to rank 100. Since accuracy only assesses whether at least *one* matching sample is present up to the rank in question (accuracy at rank-100 is the percentage of probes for which there is *one* matching image in the top 100 images), the combination of accuracy and mAP here could indicate that for VarHands-Easy, there are *more* matching images appearing at the higher ranks when at least one does. As Figure 4.8 shows, the relative performance on VarHands and VarHands-Easy changes when illumination is minimised as a factor. The illumination-invariant and texture-focused method has a far greater separation of results on VarHands-Easy and VarHands, particularly in the precision-recall curves, than the non-illumination-invariant method.

Figure 4.7 gives a visualisation of the embedding space for dorsal index finger knuckles with and without applying the process described in Section 4.3.5 to normalise illumination in the image, remove colour, and enhance the textural information. Both plots contain the same amount of knuckle images, however for the first, non-illumination-normalised plot the images are more densely grouped and overlapping. It can be seen clearly in sub-figure (a) that illumination and skin tone contribute massively to the features extracted from the image and thus where they are projected to in the embedding space: there are few distinct clusters of images that share a strong visual appearance based on colour. In sub-figure (b), the images are generally more distributed in the embedding space, and it can be seen that many of the images that were clustered together in sub-figure (a) based on illumination or skin tone are now well separated, as the model relies more on the textural information (creases) in feature extraction.

Prior ranked-retrieval results on VarHands and VarHands-Easy were obtained using a 50-50 probe-gallery split within each illumination category, that is, for each label, the daylight-clear images were split 50-50, as were the daylight-shaded images and so on. This means that for every probe image, there are several gallery images taken from the same video recording (same lighting conditions). As is seen in Figure

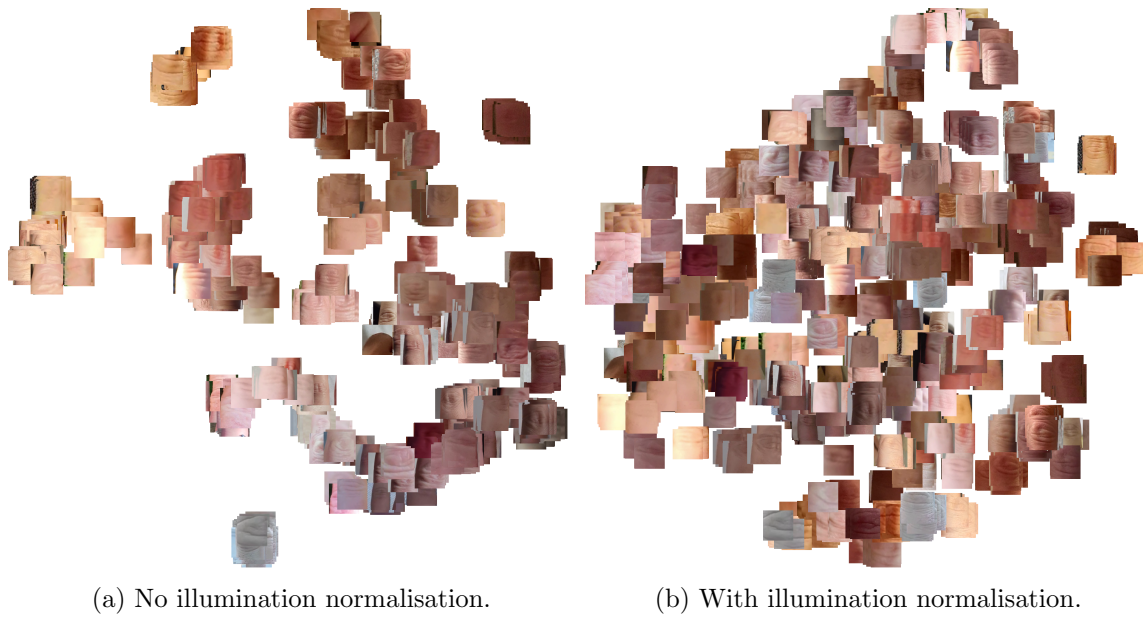


Figure 4.7: TSNE plot visualisation of the model embedding space of dorsal index knuckles in the VarHands-Easy dataset, a) trained without any illumination normalisation on knuckle images, and b) trained with greyscale illumination-normalised knuckle images.

4.6, this can lead to high ranked-accuracy (only reflects the presence or absence of *one* matching image at those ranks) with very low mean average-precision, as the matching images from other illuminations are falsely placed near the bottom of the rankings. To focus on the cross-illumination performance, the VarHands and VarHands-Easy datasets can instead be split such that one illumination forms the probe set and the other three form the gallery set; this way, there will never be a gallery image from the same video as the probe image.

Figure 4.8 shows the ranked-retrieval results for two knuckle models, one trained on and applied to the original colour knuckle images (RN50_Knuckle_Non), and one trained on and applied to the illumination-normalised greyscale images (RN50_Knuckle_IIGrey), in cross-illumination evaluation. Performance is slightly worse with the shaded-light images used as probes, as the image quality in those sets is generally lower. Averaged across all four probe sets, the illumination-normalised model achieved 50.69% rank-1 accuracy and 0.135 mAP on VarHands and 70.29% rank-1 accuracy and 0.4682 mAP on VarHands-Easy, while the non-illumination-normalised model achieved 11.63% rank-1 accuracy and 0.0505 mAP on VarHands

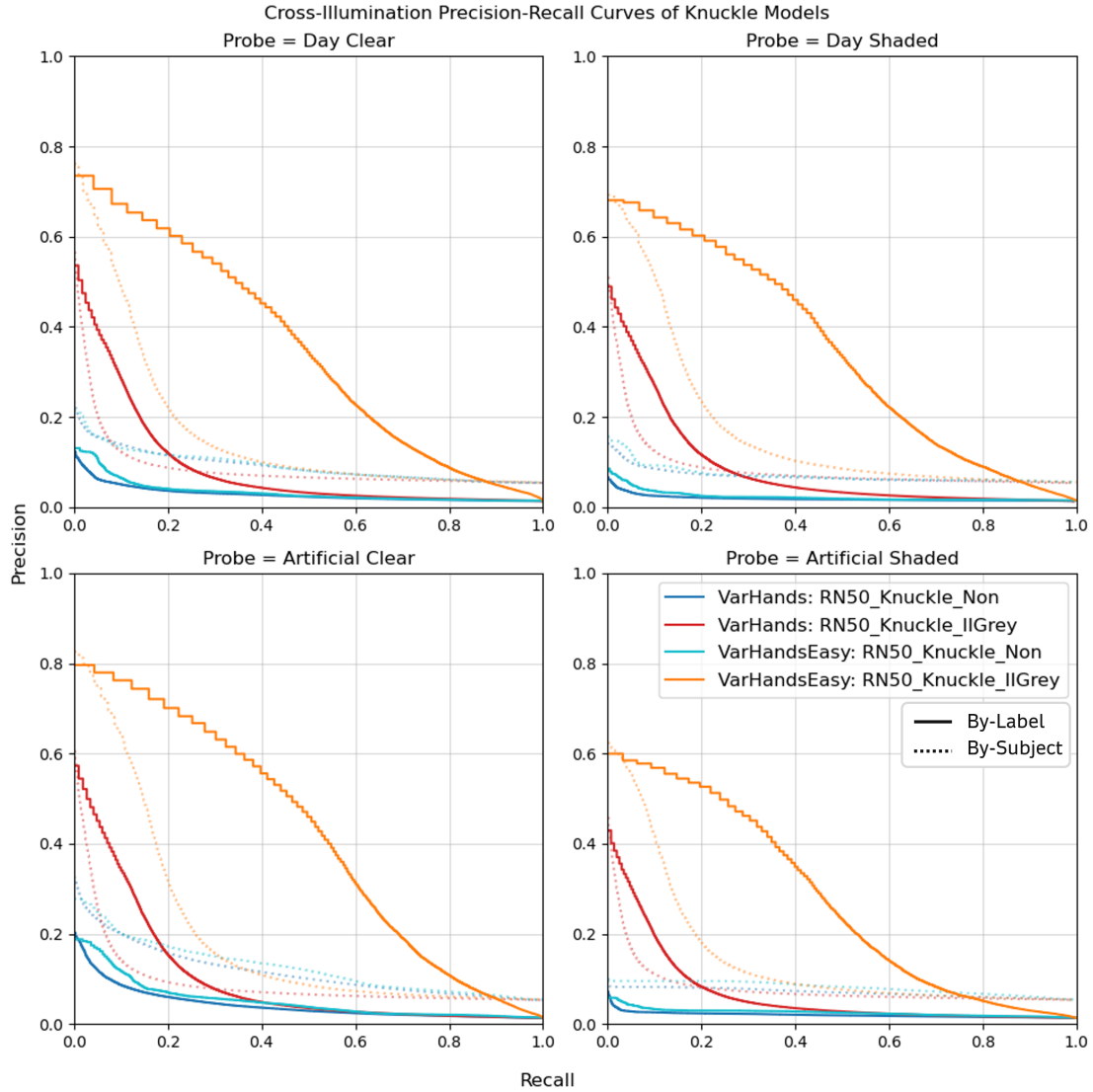


Figure 4.8: Ranked-retrieval accuracy to rank 100 of the knuckle model utilising colour images (Non) and the knuckle model utilising illumination-normalised greyscale images (IIGrey), on both the VarHands and VarHands-Easy datasets.

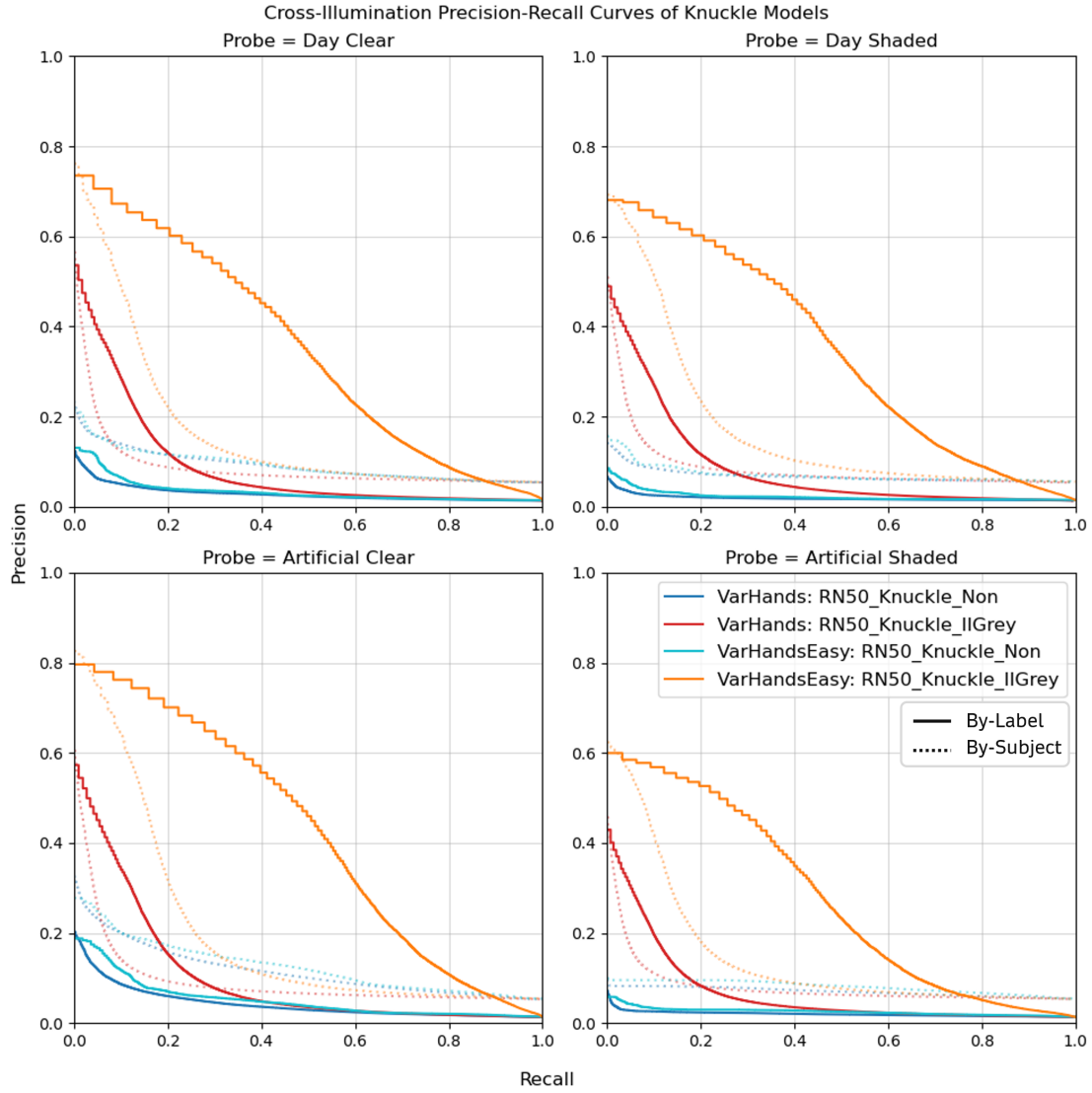


Figure 4.8: Precision-recall curves of the knuckle model utilising colour images (Non) and the knuckle model utilising illumination-normalised greyscale images (IIGrey), on both the VarHands and VarHands-Easy datasets.

and 10.87% rank-1 accuracy and 0.0769 mAP on VarHands-Easy. The illumination-normalised model performed significantly better than the model trained on the colour images, such that it achieved better results on VarHands than the colour-image model achieved on VarHands-Easy. In Table 4.1, Table 4.2, and Table 4.3, results are provided for the illumination-normalised model on the combined dorsal and palmar set for comparison. It achieves marginally worse rank-1 accuracy and mean average-precision than the colour model on 11kHands, where there is no variety in illumination in the images. On VarHands-Easy (combined illuminations, as opposed to the illumination-separated results in Figure 4.8), it achieves significantly better with approximately double the mean average-precision, 0.5915 over next-best 0.2853, and 97.01% rank-1 accuracy over next-best 93.54%. On VarHands, conversely, it suffers a loss in rank-1 accuracy with 84.4% compared to best 91.84%, but an increase to mAP, 0.1952 over next-best 0.1794.

Shown by the transparent, dashed line in Figure 4.8, the subject-matching scores for the illumination-normalised model follow a different pattern than those of the colour model. With illumination normalisation, there is less of an increase to subject-wise ranked accuracy over label-wise accuracy, compared to the colour model; this would indicate that the illumination-normalised model is less likely to highly-rank difficult non-matches (e.g. left and right hand of same individual) than the colour model, which may be using the shared colouring to match the images. Illumination normalisation also results in a lower subject-wise mean average-precision than label-wise, which is the opposite for the colour model, also indicating that the colour model is ranking many non-matching images from the same subject over matching images.

4.6 Discussion

This chapter presents the first comparison of different contactless hand recognition approaches (whole-hand, palm, knuckle, and combined palm and knuckle) on the same unconstrained dataset, as well as novel methods of extracting and palm and knuckle regions that rely exclusively on the keypoints detected by widely available hand landmarking models, and the first time application of some illumination normalisation and texture enhancement techniques for hand biometrics.

One of the bigger implications of these findings is that methods which show improved performance on controlled data, such as the use of MBA-Net over ResNet-50 for whole-hand recognition, do not necessarily improve performance on unconstrained data. While there was a large gap in performance for MBA-Net and ResNet-50 on 11kHands, this disappeared when evaluated on VarHands and VarHands-Easy. Changing from whole-hand recognition to knuckle or combined palm and knuckle,

and pre-processing the data to be colour and illumination invariant, are two actions which were much more impactful on unconstrained performance than using a more complex model.

In terms of the approaches considered for contactless hand recognition, combined palm and knuckle recognition (followed closely by sole knuckle recognition) achieved the best results on unconstrained data. Unlike on 11kHands, addition of the palm to knuckle recognition resulted in higher retrieval accuracy on VarHands and VarHands-Easy. It could be inferred that inclusion of more hand regions, like the base knuckles, wrist, or fingernails/tips, could improve performance even further. However, it appears to be beneficial, with the current quantity and quality of training data, for this to be done such that the features for each region or sub-modality are learned and compared independently. For example, having independent models process extracted knuckle and palm images results in better performance than simply including the knuckle and palm information in one image to be processed by one model, and so it may be appropriate to have a separate model for the wrist, e.g., instead of simply extending the extracted palm region to include the wrist and processing them together, though this should be tested.

The benefit of extracting the regions-of-interest as individual images is pose invariance. A single image containing all four major knuckles will have the knuckles at different scales (sizes) and image positions depending on the spread of the fingers, whereas extracting the knuckles individually allows them to be registered (spatially aligned) which can greatly improve feature extraction and matching. If this were having an impact, it would be expected to see a larger difference in performance between the palm-and-knuckle model and the whole-hand model on VarHands compared to VarHands-Easy, for which there is: a difference of 3.38% rank-1 accuracy and 0.0314 mAP between MBA-Crop and RN50-Palm+Kn on VarHands-Easy, and a difference of 9.57% rank-1 accuracy and 0.039 mAP on VarHands. The intuition of this is that as more pose variation is introduced (VarHands-Easy to VarHands) the whole-hand models would suffer a larger drop in performance than the palm-and-knuckle models.

These findings also demonstrate that normalisation of image colour and illumination and enhancement of the skin texture, even done with classical non-learning algorithms, can provide large improvements to performance on unconstrained data. All models achieved poor mean average-precision on the non-illumination normalised unconstrained data, with higher precision values only to about 20%-30% recall (Figure 4.6; meaning approximately 70%-80% of all matching images were being ranked lower than non-matching images. This would be consistent with failure to match images from different videos/illuminations ($\sim 75\%$ of all positive matches in the gallery), and is supported by the precision-recall curves in the cross-illumination tests (Figure

4.8) for the non-illumination normalised model having $<20\%$ precision for all recall values. The illumination-normalised model demonstrates a far more even trade-off between recall and precision, appearing to rank matches and non-matches in equal proportion.

There is still a long way to go in improving unconstrained recognition performance, much of which could be predicated on the collection of more diverse, varied, unconstrained data to learn from. The region extraction methods for palm and knuckles used in this study could be improved further, as they still demonstrate some error which leads to non-alignment in unconstrained images, but more consistent alignment (particularly with regards to sizing) in the constrained training data. More unconstrained data with better landmark labelling would enable learned region-of-interest extraction methods like the spatial project for palms and fingers done by Matkowski et al. [114]. For illumination-normalisation, there are several promising learning-based methods which would require specific datasets to enable them, such as Kim et al.’s conditional GAN method which requires images of hands under different direction lighting to train [87], intrinsic image decomposition [110] which would need to be fine-tuned on example decomposed images of hands. Being able to separate the underlying skin texture from illumination effects would be the ideal form of texture enhancement, bar noise, and open up 3D rendering as a further potential avenue for somewhat pose-invariant representation.

While not offering much benefit in the transfer from constrained and uniform training data to the unconstrained domain, more advanced networks and deep learning techniques (such as attention mechanisms) will be valuable when there is a significant amount of varied unconstrained hand image data to train on, from which the models can learn the discriminative hand information independent of hand pose, skin tone, and illumination. The primary directions of future research in the area of hand forensic biometrics should thus be to collect and make available substantial hand image datasets under varying conditions and if possible over a long period of time, to evaluate on unconstrained data where possible, and to make best use of the existing data by applying pre-processing techniques to generalise the models: focus on discriminative hand information through region-of-interest extractions, illumination-invariance, and robustness to changes in skin colour or cosmetic effects through normalisation or augmentation.

4.7 Conclusion

The findings of this chapter lead to the decision to use a sub-regional approach to hand recognition based on combination of the knuckles and palm going forward, and

the continued use of the proposed illumination-normalisation and texture enhancement, as they result in notably better recognition accuracy on the unconstrained images. Performance remained significantly lower on the pose-variable VarHands over VarHands-Easy, likely because there are a lot of low-quality (blurred) and invalid samples (mis-detections) resulting from hand pose variation. Regardless of the effectiveness of the models at extracting features when they are present, biometric systems used for unconstrained purposes should be able to adequately handle inputs with a *lack* of useable features without giving undue confidence to the output - particularly for critical systems which could lead to adverse effects on humans. The next chapter explores model uncertainty quantification applied to hand recognition for the first time as a theoretical way to capture model confidence and lessen the impact of low-quality inputs on recognition performance in multi-biometrics systems.

Chapter 5

Uncertainty in Hand Recognition

This chapter presents the first experimentation with the application of uncertainty quantification in hand recognition tasks and the first proposed use of uncertainty quantification for Bayesian multi-biometric fusion. A dual qualitative-quantitative evaluation approach is taken to assess the viability of uncertainty quantification in this domain with the current data.

Within texture images of palms and knuckles captured from unconstrained images there are many sources of uncertainty that can impact the performance of recognition algorithms. Figure 5.1 depicts some of these sources, loosely categorised as rotational, occlusive, and quality-related (noise or motion blur); depending on the region-of-interest extraction method, mis-detection can be another source of uncertainty when it results in extracted images that do not contain the desired region. Regardless of the source, the uncertainty stems from a loss of visual information required to accurately recognise the sample, on a spectrum of partial to total loss.

With traditional modelling, there is no understanding of the amount of (or lack of) relevant visual information to extract features from in the output. For feature-based ranked-retrieval of unconstrained images, uncertainty in the image has a strong and direct impact on the results - for example, a very blurry image of a knuckle will contain little of the visual information required to accurately match it to other knuckle images, but the model will project it to a point in feature space regardless, which will likely be a very short distance to other blurry or low quality samples from non-matching subjects. In classification-style systems, an uncertain (i.e. low quality) input sample would likely lead to low probability scores for all classes, but there is still a lack of information presented to the user about the data this is based on (is it a low-quality sample, or is it a high-quality sample belonging to a subject outside of the classification set?). This certainty of the model is often erroneously inferred from

the pseudo-probabilities output in classification tasks, but as Gal and Gharamani, the authors of MC-Dropout, state, “a model can be uncertain in its predictions even with a high Softmax output” [44]; probability outputs in classification often exhibit poor calibration (correlation with actual probabilities) [61]. Uncertainty doesn’t just pertain to image quality - a high-quality sample outside of the training dataset distribution can result in an uncertain output, as it contains information that the model is not equipped to handle (i.e. pixel values that the model is not optimised through training to process to the correct outputs).

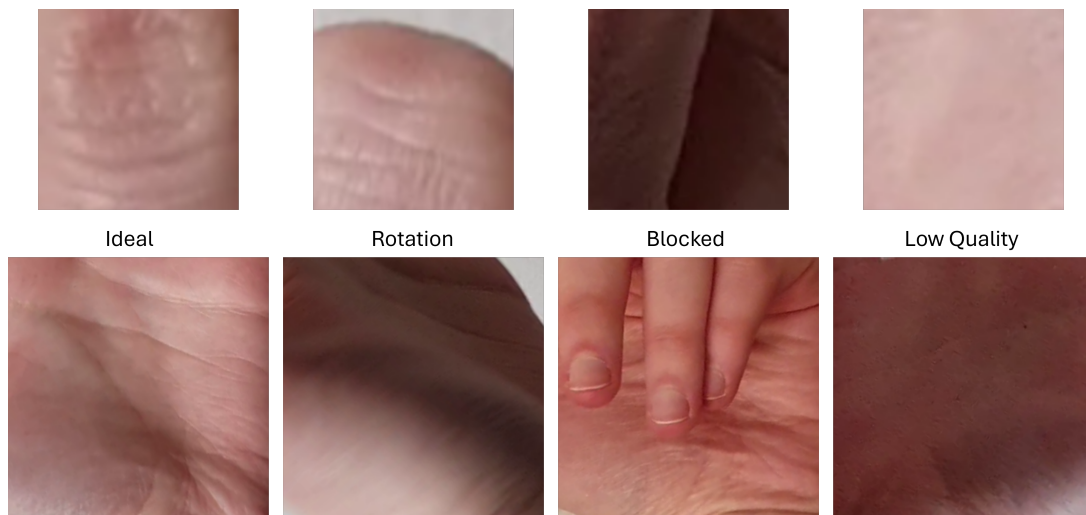


Figure 5.1: Sources of uncertainty in palm and knuckle extractions from unconstrained hand images.

Being able to quantify the uncertainty not only of the data (perhaps a proxy for data quality) but also in the model and model outputs could provide a number of benefits:

- Enabling the calibration of models to make their outputs, especially pseudo-probabilities like Softmax outputs, more robust and in line with actual likelihoods.
- Communicating to the system user that some action needs to be taken for this sample, such as re-capturing, enhancement, further scrutiny, limiting its impact on any outcomes, or in some cases, discarding it.
- Giving some value by which to weight samples in further processing, for example, in considering the contribution of different samples in the fusion stage of

multi-biometric systems.

Training models alongside uncertainty quantification with the proper inclusion of the uncertainty in the loss function has been demonstrated in several papers to result in better calibration of these uncertainty measures over psuedo-probabilities [86, 167]. This is largely only beneficial for better communication of the model output, however some papers have also established that uncertainty-aware models and loss functions also train more effectively and can lead to better performance on test datasets [168, 44, 145]. Communicating to the system user the uncertainty of a model output is an important aspect of critical or high-risk systems, to which forensic biometric systems used in law enforcement belong. As well as flagging a sample for re-capture (not possible when dealing with externally obtained images), this allows the system user to know when a sample may need further scrutiny (human analysis) or should not be relied upon for action. Providing this information when possible creates more robust and reliable biometric systems and is an ethical and legal obligation under the EU AI Act - this is discussed in more detail later in Chapter 6.

The third benefit, the use of quantified uncertainty measures as a weighting mechanism in multi-biometric or information fusion, is at the time of writing a novel proposition in this project. One study used known model accuracy in combination with estimated sample quality for multi-modal multi-biometric score fusion [149] using the probabilistic Dempster-Schafer theory (discussed in Section 5.3.3). No studies have been found to explore the use of quantified model uncertainty as the probability for probabilistic multi-biometric fusion. The benefit of using model uncertainty over proxies such as input sample quality is that many samples that are measurably high-quality (low noise or blur) might fall outside the distribution of the training datasets and thus result in unreliable model outputs (i.e. an image with a tattoo or other unique physical feature). This same mechanism is at play with demographic bias in biometrics: under-representation of certain physical features in the training dataset means that the model does not effectively learn to discriminate the fine details of those features. The implications for handling demographic bias are discussed further in Chapter 7.

The rest of this chapter presents research questions on the use of uncertainty quantification in unconstrained biometrics and multi-biometric fusion and a specific application scenario with an explanation of how uncertainty to other measures of divergence from the training dataset could be used, then covers the background of related work in uncertainty quantification, the methodology used to answer the research questions including implementation details, and a discussion of the results obtained.

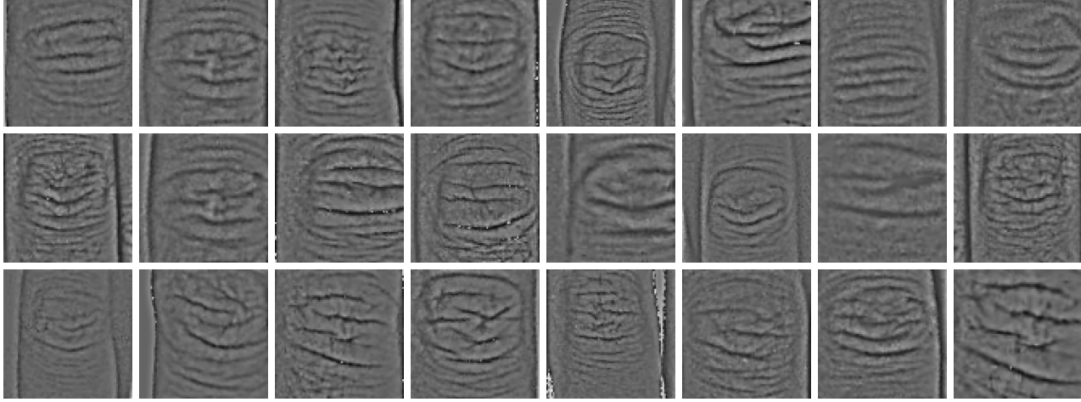


Figure 5.2: Examples of knuckle images extracted from the 11kHands training dataset, which consists of highly constrained hand poses under a single illumination.

5.1 Scenario

Consider a scenario where two dorsal-side unconstrained images of hands are being compared to determine if the hands match. Based on the results obtained in Chapter 4, a sub-regional multi-biometric approach is taken over a whole-hand approach, where the minor and major knuckles of the four fingers and the thumb knuckle are extracted and separately transformed into feature vectors by a single model. The Euclidean distance between each corresponding pair of knuckle feature vectors (e.g. index major knuckles) is then calculated, resulting in a total of nine distances. These distances are then fused (most simply: averaged) to give a single estimated feature distance between the two hand images, which can be used with a threshold to decide if they qualify as a match, or if more comparisons are performed for image A, to classify (closed set - image A is known to belong to one of the set of images containing image B) or rank results based on shortest distances (open set - image A may appear in the set containing image B any number of times, or not at all).

The single model is trained on images of all knuckles (minor and major on the four fingers plus the major thumb knuckle) extracted and pre-processed from whole-hand images in the same manner as they are at inference, however the training dataset consists of well-controlled images in one illumination, illustrated in Figure 5.2, while the inference data is unconstrained and illumination-variable, as illustrated (post-illumination-normalisation) in Figure 5.3. The lack of constraint in hand pose and variation in illumination seen in the inference images can lead to a spectrum of information loss from several sources, some demonstrated in Figure 5.1. When

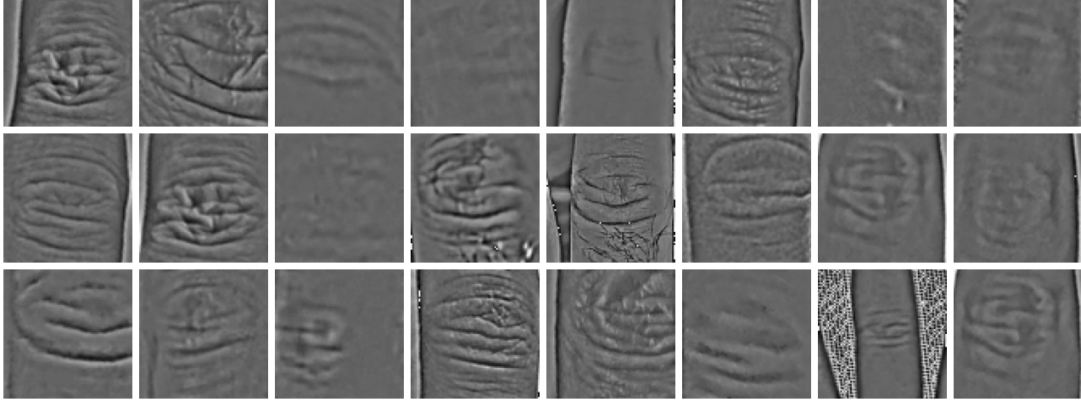


Figure 5.3: Examples of knuckle images extracted from the VarHandsEasy test dataset (in which all knuckles are visible), which consists of images in a variety of poses under a variety of illuminations.

presented with an image with information degradation or loss, the model will still project it to a point within feature space, but this is unlikely to be an accurate embedding; if it was presented with a totally uniform image with *no* information, which can happen when there is overexposure to bright lighting in the photos leaving them completely white, the model would project this image to the same point in feature space as every other uniform image, regardless of the subject they contain. By capturing the information loss of the input samples - on a scale, from complete loss to ideal image - it is possible to estimate the validity of the comparison between them.

In any given whole-hand image, the information loss of each knuckle will vary - Figure 5.4 shows this with the comparison of the major knuckles in two images of the same hand, one of which shows a clear view of the thumb, index, and middle fingers, and the other shows a clear view of the middle, ring, and pinky fingers. In this example the non-visible knuckle extracts may be completely invalid, however it is possible for there to be partially visible knuckles, or completely visible knuckles that lose some information through blurring or image noise.

It may not be the most optimal use of the data to do image quality-based exclusion of samples prior to comparing, as some other papers do [163], as this will disregard valuable information in the partially visible images and, as explained in the next section, the quality of the image does not necessarily correspond to the information useable by the model. Estimating the *model* uncertainty should capture the model output certainty based on the amount of information in the sample useable by the

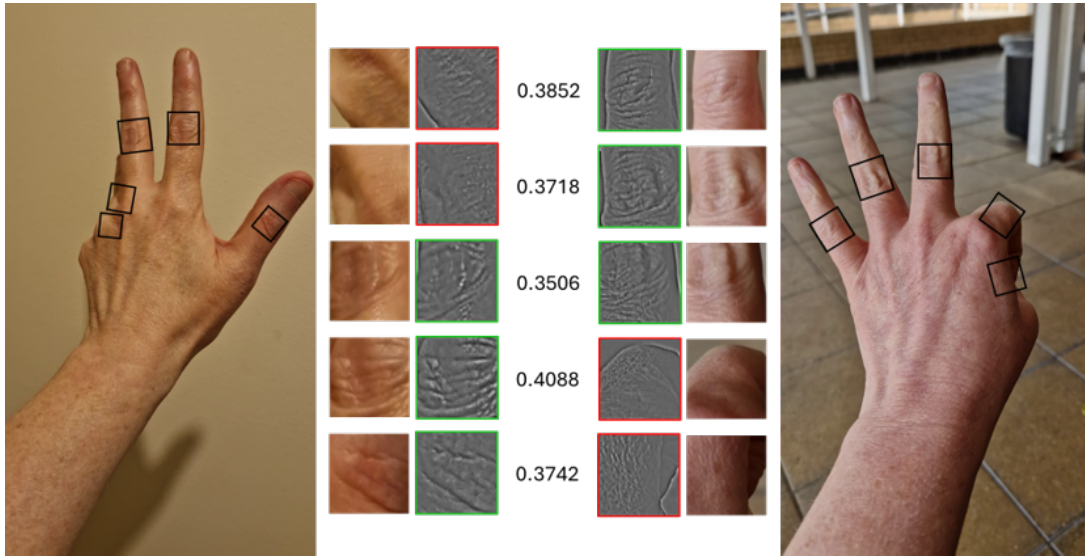


Figure 5.4: Example of two different images of the same hand in the VarHands dataset with major knuckles extracted. The knuckle images are all preprocessed and fed to the network to produce an embedding, and then the Euclidean distance between the corresponding knuckle embeddings calculated. The distance between the middle knuckles is lowest at 0.3506.

model (how well the model has been trained to handle that information), as well as, in a case such as this where the training dataset consists exclusively of ideal samples, the divergence of the sample from this ideal.

5.2 Research Questions

The subset of research questions addressed in this chapter are:

RQ5 What are the behaviours of uncertainty quantifying methods when applied to hand biometric images of increasing quality loss?

RQ6 What impact can these quantified uncertainties have on recognition performance when incorporated into multi-biometric fusion?

To answer these questions, a literature review was conducted to identify promising uncertainty quantification techniques, discussed in Section 5.3, and numerous exper-

imental set-ups were designed, discussed in Section 5.4 to evaluate the performance of the uncertainty quantification techniques against a baseline.

5.3 Background

Traditional model outputs are deterministic; they represent a single value, or a single feature vector, or a single set of Softmax classification ‘probabilities’. These outputs give us no information about the range of other potential output values that could have resulted from the input. We would expect that the range of potential feature embeddings for a low-quality, blurred image would be greater than the range of potential embeddings for a very clear image, but without uncertainty quantification this information is lost and we are presented with only a single embedding for each.

5.3.1 Aleatoric and Epistemic Uncertainties

Model uncertainty is often categorised into aleatory - randomness - and epistemy - systemic uncertainty [31]. Aleatoric uncertainty arises from inherent variation (noise) in real samples and their capture into a digital format. It can be further categorised as homoscedastic (the same for all inputs), such as from sensor noise in the camera, or heteroscedastic (different for all inputs), from things such as motion blur. Aleatoric uncertainty is unavoidable and cannot be improved with the introduction of more training data or better processing techniques. Epistemic uncertainty arises from a lack of knowledge, and is a measure of how well the model emulates the real system being represented. Theoretically, epistemic uncertainty can be reduced with more training data and better data processing and modelling techniques.

Uncertainty quantification in computer vision often seeks to separate these concepts and measure one of them in particular: aleatoric uncertainty can be to some degree comparable to image quality assessment, quantifying the blur or noise present in an image and giving a value for how uncertain a *sample* is, while epistemic uncertainty quantification is the task of measuring the uncertainty of the *model* in relation to a certain sample. For epistemic uncertainty, it is theorised that the model will be more confident (and thus more concordant with itself, always producing the same output for the input) with samples similar to those it was trained on and less confident with samples unlike the training dataset; this way, epistemic uncertainty can be a measure of divergence from the training dataset and be used for tasks like out-of-distribution detection [143].

In practice, aleatoric and epistemic uncertainties are very entangled, and capturing both can give a more complete view of the model’s confidence [86]. For this

scenario in particular, it would be prudent to capture a measure of both random input quality loss from factors such as noise or motion blur, and of divergence from the training dataset/loss of model confidence, caused perhaps by lack of sufficient representation in the training dataset of certain physical traits.

5.3.2 Uncertainty Quantification in Machine Learning

This section covers the background literature on computational methods of quantifying aleatoric (data) and epistemic (model) uncertainty in deep learning tasks.

Quantifying model or data uncertainty in practice means producing stochastic probability distributions as outputs rather than deterministic point values. For example, an expected (mean) feature embedding with variance for each dimension, rather than a single feature embedding. The variance of the distributions represents the certainty of the expectation: a small variance indicates the expectation may be highly accurate, whereas a large variance means there are many more possible values for the embedding. There are many ways to do this, and as will be discussed, most rely on approximate methods. It can be very difficult to ensure that the resulting probability distributions are well-calibrated (meaning the confidence intervals should be accurate - a 90% confidence interval should contain the correct value 90% of the time), and may require much more data than would typically be needed to train an equivalent deterministic model [75].

Bayesian Neural Networks (BNNs) are models in which all of the parameters exist as a stochastic probability distribution rather than a single value, and thus the model output is a propagated probability distribution with mean and variance [75]. The probability of the weights w given the training data D is the *posterior* distribution $p(w|D)$, and can be calculated from the *prior* $p(w)$ and the *likelihood* $p(D|w)$ through Bayesian inference [111] as shown in equation 5.1. The predictive distribution of a sample $p(\hat{y}(x)|D)$ can then be calculated as the expectation under the posterior, shown in equation 5.2, which is equivalent to sampling from an infinite ensemble of networks [14]. The problem with this approach is that it is very computationally expensive, particularly with large networks, as it requires integration over all possible values of w twice, to calculate $p(D)$ and to calculate the expectation of the posterior. Alternatively, there are Gaussian Processes, which model function outputs as joint Gaussian distributions with means and covariances then condition the posterior distribution of those functions on data observations; *Deep* Gaussian Processes (DGPs) stack multiple Gaussian Processes to capture complex hierarchical structures, like deep neural networks [28]. Integration over the latent variables at each layer of DGPs is similarly intractable, meaning approximations must be used.

$$p(w|D) = \frac{p(D|w)p(w)}{p(D)}t \quad (5.1)$$

$$p(\hat{y}(x)|D) = E_{p(w|D)}[p(\hat{y}(x)|w)] \quad (5.2)$$

Methods of approximation are generally categorised as either sampling-based or variational. Sampling-based methods involve sampling a large number of values from a distribution that is assumed to be approximate to the posterior, such as weights proximal to a trained network’s weights, and using them to estimate values for the true distribution and expectations [135]. Variational methods model the posterior distribution as a function of parameters, often denoted $q(w)$, utilising the Kullback-Liebler divergence (KL-divergence) to calculate similarity of the modelled and true posterior [135]. Generally, sampling methods only provide epistemic uncertainty - variance in the model outputs - and models need to be adapted to explicitly learn a value for aleatoric uncertainty. The naïve sampling method is to draw weights independently from the prior distribution, however as the number of network parameters grows, this requires exceptionally larger numbers of samples to ensure fitting with the true distribution. A less naïve approach is to use Markov-chain Monte-Carlo sampling, which generates dependent samples by selecting the next sample from within a distribution of the previous sample; this speeds up convergence at a high number of network weights [51].

Gal and Ghahramani derive that performing multiple forward passes through a network with dropout enabled is a Bayesian approximation of the Deep Gaussian Process, in a method they call Monte-Carlo dropout [44]. In effect, the mean and variance of outputs obtained from n forward passes of the network is a predictive distribution approximate to what would be obtained when sampling from a Bayesian network posterior distribution. The uncertainty (sample variance) obtained from this simple form is the epistemic uncertainty of the model, but the model can be modified to explicitly learn the aleatoric uncertainty of the data through an additional output value and a modified loss function. MC dropout is practical to implement as it doesn’t require any adaptation to network architecture, only the inclusion of dropout layers, and can be retroactively applied to trained models, so provides a computationally efficient way to obtain uncertainty estimates for non-Bayesian models. Compared to other methods, however, MC dropout may provide poor approximation of complex posteriors as it samples from a very limited distribution defined by the dropout layers and is very sensitive to weight magnitude and dropout rate, and some critics have refuted its Bayesian foundation [40].

Lakshminarayanan et al. demonstrate that averaging predictions across an ensemble of deep networks is able to provide competitive uncertainty estimates with Bayesian uncertainty quantification methods while remaining scaleable to large networks and datasets [102]. They modify the networks to output two values, the mean μ and variance σ^2 . They treat these values as a sample from a Gaussian distribution, and instead of minimising the MSE of the predicted and true value, they minimise the negative log-likelihood. They recommend adversarial training, where samples that are indistinguishable to humans but are misclassified by the network are used, as a way to smooth the predictive distributions. Experiments on ImageNet classification showed better calibration in uncertainty estimates and better performance than Monte-Carlo dropout.

Estimation of the weight posterior via sampling loses effectiveness as networks grow, and so variational approximation of the posterior becomes preferable for deep learning. Variational approximation methods aim to estimate distribution parameters θ that minimise the KL-divergence between the true posterior $p(w|D)$ and the estimated posterior $q(w|\theta)$ [88]; this cost function, the expected lower bound (ELBO) or ‘variational free energy’ as referred to in different papers that follow, is shown in equation 5.3.

Blundell et al., from whom equation 5.3 is obtained, use Monte-Carlo sampling of expectations in their backpropagation algorithm for variational Bayesian inference in neural networks, termed Bayes-by-Backprop, which allows probability distributions to be learned for the network weights instead of single values [14]. The authors describe the ELBO cost function as the sum of the data-dependent part (likelihood) and the prior-dependent part (complexity), thus the cost function aims to find a balance between the complexity of the data and the simplicity of the prior. Their experimental results demonstrate that using Bayes-by-Backprop performs better than stochastic gradient descent with and without dropout, and produces calibrated uncertainty estimates for unseen data.

$$F(\theta, D) = KL(q(w|\theta), p(w|D)) - E_{q(w|\theta)}(\log p(D|w)) \quad (5.3)$$

Shi and Jain presented Probabilistic Face Embeddings (PFE) as a method of stochastically embedding face images as a mean vector μ with variance for each dimension σ [145]. The variance vector is produced from two fully connected layers at the same level as the network bottleneck layer, and is trained to optimise the Mutual Likelihood Score for all genuine pairs (anchor-positives). The training of this variance head can also be applied to pre-trained and frozen networks, so a network can be trained deterministically to learn the mean vectors μ and then further trained to learn embedding variances σ . The mutual likelihood score is different from a

feature distance in the fact that it incorporates the uncertainty of each dimension, meaning low-variance dimensions contribute more; then, if all dimensions are high-variance, the mutual likelihood score would be low regardless of the mean distance. The authors demonstrate that applying PFE and using the mutual likelihood score boosted recognition performance on all tested models pre-trained with Softmax or Triplet losses.

Warburg et al. proposed a method for Bayesian triplet loss which allows networks to learn univariate probabilistic vectors for feature embedding through a likelihood function that adheres to the triplet constraint; they consider what the *probability* of the anchor being closer to the positive than the negative is, given the mean and variance of each embedding [168]. They additionally justify the continued use of L2-normalisation with both Gaussian and Uniform prior distributions over feature space, with a posterior distribution that is Gaussian for a Gaussian prior and von-Mises Fisher (vMF) for a Uniform prior. To obtain a mean and variance for each embedding, networks are adapted with extra layers to output two values: a mean embedding and a unit variance. Both network heads make use of generalised mean pooling (GeM) layers [137] with separate parameters. In place of the traditional triplet loss they opt to use the negative log-likelihood, which they demonstrate provides smoother and more bounded loss values. They show that Bayesian triplet loss can perform as well as traditional triplet loss while producing very well-calibrated uncertainty estimates, correctly identifying out-of-distribution samples at a higher rate than other uncertainty quantifying methods.

Daxberger et al. show Laplace approximation can be competitive with Bayesian variational methods to approximate the posterior of deep neural networks. Their method requires first training a network standardly to obtain the maximum a-posteriori (MAP) weights (a local minima), then treating some subset of those weights (or all) probabilistically through computationally-efficient approximations of the Hessian matrix [30]. They provide two options for tuning the prior and likelihood hyper-parameters: post-hoc, which uses cross-validation from the pre-trained network to select parameters, and online, which tunes the parameters alongside estimation of the MAP weights (regular model training). They demonstrate that the online and post-hoc methods can each be superior in different circumstances: post-hoc is more computationally efficient and better-performing when there is a strong pre-trained network and well-known hyper-parameters, whereas online can achieve superior performance otherwise. This method has been adapted for encoder networks to produce probabilistic embeddings by Miani et al. [118], and then for triplet-based metric (distance) learning of embeddings by Warburg et al. [167]. Miani et al. show that approximation of the Hessian diagonal is well-performing and scalable to high

dimensions in image encoding, despite Daxberger et al.’s findings that diagonal approximation was the worst performing option. They also demonstrate that post-hoc methods are sufficient for out-of-distribution detection but produce poorly calibrated uncertainties for in-distribution samples.

5.3.3 Fusion of Uncertain Information

This section covers the literature at the intersection of information fusion and uncertainty quantification.

Guo et al. performed a survey of belief theories and their application in machine learning and uncertainty quantification tasks [52], though they largely look at use in modelling rather than in information fusion. The two frameworks they define as applicable for information fusion, which is relevant for multi-biometric systems as well as others, are Dempster-Schafer theory (DST) and transferable belief models (TBM). Both of these methods use epistemic uncertainty arising from a lack of evidence (vacuity). They also cover the more widely-used method of Bayesian inference, as well as others which are less useful for the application domain considered in this thesis. The specific application considered here is sensor or data fusion, where the continuous and uncertain outputs (vectors and their distances) from multiple outputs must be combined and ranked. The belief theories will be discussed in relation to this scenario.

Dempster-Schafer theory applies probability masses to all options in a mutually-exclusive set of possible outcomes, then measures belief (evidence) in those outcomes (normalised so that all belief sums to 1), and calculates their plausibility based on levels of belief in other outcomes. Crucially for information fusion, DST includes a rule of combination, which evaluates plausibility from different evidence bases (different sets of beliefs) by discounting their disagreement. As DST requires a discrete and finite set of outcomes it is best suited to classification tasks, not regression or distance metric tasks. TBM, a modification of DST that treats empty set outcomes as unknown outcomes, can be adapted for use with continuous variables by discretizing them in intervals, but this is still not ideal for distance metrics. The benefit of DST and TBM is that they incorporate ignorance (lack of evidence to support a belief, vacuity) into final uncertainty where other methods do not. Bayesian fusion can be used to combine multiple (assumed Gaussian) distributed variables through the multiplication rule, where the mean of each variable is weighted by the inverse of its variance. Unlike DST methods, this can be used simply and effectively for continuous values, however as a downside it does not consider ignorance or unreliability in final uncertainty estimates.

There are numerous works applying DST and other belief theories to multi-biometric fusion. Nguyen et al. suggest that prior works rely on inappropriate measures of uncertainty, often considering Softmax classifier outputs to be equivalent to the true posterior probability of a class (as demonstrated in uncertainty-quantification works, this is not the case), and so propose the use of sample quality assessment as the uncertainty factor in DST fusion [122]. For face images they use a combination of different measures including the face detection score, illumination quality, and ‘frontalness’, and through incorporating these into DST multi-modal fusion they achieve better true positive - false positive ratios than prior DST fusion works. Quality assessment is however not equivalent to quantified model epistemic uncertainty, and no papers have been found that included the use of model epistemic uncertainty quantification for multi-biometric fusion.

5.4 Methodology

To answer Research Question 5, “what are the behaviours of uncertainty quantifying methods when applied to knuckle images of increasing quality loss?”, a well-performing appropriate backbone network was trained with a variety of deterministic and stochastic learning methods obtained from the literature review: standard triplet loss (Det.), Bayesian triplet loss (BTL), Monte-Carlo dropout (MCD), and deep ensembles (DE). In addition to these methods, non-probabilistic descriptors and metrics were tested for their correspondence with image ‘validity’ and for if they provide a performance boost in uncertainty-aware multi-biometric fusion, acting as a baseline to compare the uncertainty-quantification methods against; these were image texture description using grey-level co-occurrence matrices (GLCM), and out-of-distribution measurement using auto-encoder mean squared error (AESME). To answer Research Question 6, “what impact can these quantified uncertainties have on recognition performance when incorporated into multi-biometric fusion?”, several fusion methods described in Section 5.4.6 were devised and evaluated against their standard fusion baselines in a ranked-retrieval scenario on the VarHands dataset. Two other methods, probabilistic face embeddings (PFE) and Laplacian metric learning (LML) were implemented but did not result in successfully trained models at this time that could be fairly compared against the others.

Following from the previous chapters and for its frequent use in comparative papers, ResNet-50 was chosen as the backbone network and used for all models to ensure that only the method-specific output processing and learning protocols contribute to differences in performance. Other backbone networks were tested with traditional triplet loss, Monte-Carlo dropout, and Bayesian triplet loss: these were

ResNet-18, ResNet-34, DenseNet-121, and a custom smaller DenseNet structure. DenseNet was chosen in particular because of the incorporation of dropout making it suitable for the Monte-Carlo method. Due to a lack of difference in performance (neither the smaller ResNet models nor the larger DenseNet models outperformed ResNet-50) and similar training behaviour, and the fact that the DenseNet model took almost twice as long as the ResNet model to train, they were not implemented for the remaining methods in respect of time constraints.

Because Monte-Carlo dropout requires dropout layers in order to produce outputs with variance, the ResNet-50 model had dropout layers inserted after the first ReLU activation and then after every residual block. All models were trained with triplet loss (and Bayesian triplet loss as a variation of it), which were built on top of standard triplet loss. A margin of 0.2 was always used in accordance with the original triplet loss paper [142]. The batch size was always 32 images, made up of 8 classes with 4 samples each, for an effective batch size of 48 possible triplets. Every model was trained with semi-hard negative mining for 200 epochs and then hard negative mining for 200 epochs. Method-specific training configurations, such as learning rates, are given in the following sections.

Following the findings in Chapter 4 which demonstrate illumination normalisation techniques to optimise transfer performance from a single-illumination training dataset to illumination-variable test data, the training and testing data for these experiments were pre-processed in the same way. This includes detail enhancement, Gaussian blur of the image luminance (illumination normalisation), then conversion to greyscale and inversion. During training, the images were then randomly augmented with contrast shift in the range $[0.5, 2.0]$, where <1 reduces contrast and >1 increases it; after this, the images were sample-wise standardised to have 0 mean and unit variance. All output feature vectors were L2-normalised to exist on a unit sphere, as is a requirement for Bayesian triplet loss.

5.4.1 Grey-Level Co-Occurrence Matrix Statistics

In theory, knuckle images should have some textural properties that are not shared among many other types of images, particularly of images where there is more noise, blur, or occlusion, or images of a different scale where the knuckle crease patterns are either larger or smaller. Grey-level co-occurrence matrices (GLCMs) capture the probabilities of any two greyscale pixel values $[0, 255]$ being found adjacent to one another (or by some offset) [54]. From these matrices values for image contrast, dissimilarity, homogeneity, energy, correlation, and entropy can be calculated, as well as the co-occurrence means and variances [121]. To describe the textural distributions

of the training datasets, the statistics of mean, correlation, dissimilarity, and entropy were chosen for their small variation among the dataset samples but large divergence when noise or blur is artificially added.

To acquire these statistics, each training sample from the high-quality illumination-normalised greyscale image dataset (for knuckles and palms) is loaded, then four co-occurrence matrices are calculated at a distance of one pixel and angles of $[0, \frac{\pi}{4}, \frac{\pi}{2}, \frac{3\pi}{4}]$. The statistics are calculated for each of the four matrices and stored in an array, from which the means and variances across the whole dataset are calculated. In order to measure the divergence of a sample from these dataset averages, the sample statistics are calculated, subtracted from the training dataset mean (absolute value), and divided by the training dataset variance. The mean divergence across the four matrices is taken and then summed across all statistics.

The reason this is a baseline method and not considered to be uncertainty-quantification is that it provides no information about how the model processes the information. While it is reasonable to believe that this GLCM divergence will correlate with model performance on a given sample, it may not capture the same distribution of information as used by the model in its predictions.

5.4.2 Auto-Encoder Loss

This method, like GLCM, is a measure of the input samples divergence from the training dataset, and is not a measure of the model uncertainty.

Auto-encoder networks learn effective representations of images by using an encoder-decoder network architecture, in which the encoder reduces the image to a feature vector, and the decoder then reproduces the original image from that feature vector. The reconstruction loss of the decoder network - often the mean squared error (MSE), a pixel-wise difference between the two images - is used to encourage the network towards vector representations that well describe the input. In theory, the decoder (reconstruction) network will be far less accurate with image data that is unlike what it has seen during training, resulting in higher MSE losses for out-of-distribution samples than in-distribution samples; this assumption is why auto-encoder reconstruction loss has been used for anomaly or out-of-distribution detection [133]. In practice, however, auto-encoders often fail at anomaly detection as they can accurately reconstruct some inputs that would be out of the training distribution [15].

The encoder part of the auto-encoder network used in this study consisted of four 2D convolutions followed by ReLU activations which reduced the input image to a 6400-D vector, then a fully-connected linear layer to produce a final 2048-D embedding. The decoder network starts with another fully-connected linear layer to

reconstruct the 6400-D vector, then four transposed 2D convolutions (upsampling with trainable parameters) followed by ReLU activations to reconstruct the original input. This network was trained on the same high-quality training dataset as the ResNet-50 feature extractor models were trained on, with all of the same pre-processing and augmentation applied, in the same 90% train and 10% validation split. It was trained with a learning rate of $1e-3$ reducing to a third every 10 epochs, and trained for 50 epochs. At evaluation time, each input image is passed through both the ResNet-50 feature extraction network to produce the embedding for ranked retrieval, and the auto-encoder network to produce a MSE reconstruction loss as a validity measure.

5.4.3 Monte-Carlo Dropout

Monte-Carlo dropout for epistemic uncertainty quantification takes a standard deterministic backbone model and enables dropout at test time, forming probabilistic predictions by performing n forward passes to produce n feature embeddings, then calculating the mean and variance of each feature dimension. For this study $n = 50$ samples were used for evaluation, with a dropout rate of 0.3 (the same as in training). The model was not adapted to learn explicitly a value for aleatoric uncertainty. The same trained instance of the ResNet-50 model (same weight values) was used to obtain the deterministic, GLCM, AEMSE, and MCD results.

5.4.4 Deep Ensemble

Five instances of the ResNet-50 model without dropout enabled ($p=0$) were trained for the deep ensemble. In the original proposal of deep ensembles for uncertainty quantification the models are adapted to output a mean and variance, to explicitly learn the aleatoric uncertainty, and use a Gaussian negative log-likelihood loss function [102]. Because there is no simple way to learn a mean and variance output with triplet loss (this is what Bayesian triplet loss is solving), this modification is not made. Each model in the deep ensemble only outputs a single point embedding, then the final embedding used for ranking is the mean average of those five embeddings, and the value used for uncertainty is the sum of the variances for each feature dimension across the five embeddings. This measures only the epistemic uncertainty, or model disagreement, as with the implementation of Monte-Carlo dropout.

5.4.5 Bayesian Triplet Loss

Bayesian Triplet Loss is a modified version of triplet loss proposed by Warburg et al. [168] that uses a negative log-likelihood loss from a probabilistic interpretation of the triplet constraint over distributed embeddings. To implement this, the same ResNet-50 backbone network was used, but adapted with mean and variance network heads after the final feature layer. The mean head consists of a generalised mean pooling (GeM) layer [137] followed by a linear layer to output D-1 values. The variance head consists of a GeM layer followed by a rectified linear unit (ReLU), a linear layer to output 1 value, and a Softplus activation. For this project we assume a Gaussian prior and posterior distribution, thus the network mean output is normalised with a single trainable scale parameter. Code for the loss function, including calculation of T (the Bayesian triplet constraint equal to $ap_{dist} - an_{dist}$), the negative log-likelihood, and the Kullback-Liebler divergence, were taken directly from the supplementary material of [168] and adapted for the different data format.

The triplet loss margin was maintained at 0.2. The KL loss scale was set to 1e-4 instead of the 1e-6 used in the paper, as 1e-6 caused unstable training and led to divide by zero errors: as a fix, values less than zero were set to 1e-8, however this was not presented in the paper or original code. As in the paper, a learning rate of 1e-5 was used, a weight decay rate of 1e-3, step size of 1, and step gamma of 0.99 (meaning every epoch the learning rate is reduced by 1%).

5.4.6 Multi-Model Metric Fusion

A multi-model biometric system combines the outputs of several models into one result. In this scenario, feature embeddings are obtained for each of the nine knuckles, and then distance and covariance matrices between the probe samples and the gallery samples are calculated for each knuckle. The distance matrix is simply the L2-distance between each probe and gallery embedding, and the covariance matrix is the maximum uncertainty measure of each probe and gallery embedding pair. It is with these nine distances + covariances that distance-fusion is performed to produce final combined distances (one for every probe-gallery pair) by which to rank the gallery. The simplest way to combine multiple distances is to sum them (which in this instance is monotonically related to averaging while computationally more efficient). Four other methods are explored that take into account the covariances when combining the distances: selection of the minimum, Bayesian fusion, image-weighted Bayesian fusion, and region-weighted Bayesian fusion.

Selection of the minimum is another simple method in which only the distance of the region pair (i.e. index fingers) with the minimum covariance is used, and all other

distances are disregarded. Bayesian fusion, given in Equation 5.4.6, is performed by multiplying each distance value by its precision, which is the inverse of its variance, and then dividing that value by the sum of the variances.

For the image-weighted and region-weighted methods, the variance values are first normalised before applying Bayesian fusion using Equation 5.4.6, which is a combination of z-score standardisation of the variances followed by the logistic sigmoid function to project the values to the positive range $[0,1]$, compressing outliers to the boundaries. This was done over min-max normalisation to limit the impact outlier values (exceptionally high uncertainties) have on the resulting data distribution. All normalisation is separated by probe sample (using different means and standard deviations) so as to mimic the case of a single sample being submitted to the system, rather than using statistics calculated from all test probe samples at once. For image-weighting, different normalisation values (mean and standard deviation) are used for each probe-gallery pair, calculated only from the nine regions in that pair-wise comparison. For region-weighting, normalisation values are calculated separately for each region (i.e. the thumb) using all probe-gallery pairs.

Image-weighting gives more weight to the relative uncertainties of each region comparison within an image (the least certain region in each image will be discounted, as it will be normalised to near-0), however this does not account well for if all regions in an image have similar uncertainties, whether high or low. Region-weighting considers how different regions may naturally have higher uncertainties than others (the pinky finger compared to the index finger, for example), and limits the impact of this, since they are only ever compared within-region anyway.

$$D' = \frac{\sum_{n=0}^{N=9} D_n \frac{1}{V_n}}{\sum_{n=0}^{N=9}} \quad (5.4)$$

$$V' = \frac{1}{1 + e^{-\frac{V - \mu_V}{\sigma_V}}} \quad (5.5)$$

5.4.7 Expected Calibration Error

In addition to the other ranked-retrieval metrics described in Chapter 3 Section 3.2.3, the expected calibration error (ECE) is used to measure the correlation between model uncertainty outputs and their performance. In classification problems, this is the comparison of the predicted probability of a class and the actual likelihood of

that outcome; for instance, a predicted probability of 0.8 for class x would be well-calibrated if 80% of model predictions of class x were correct [120]. This is achieved by averaging the weighted differences between model confidence and model accuracy across confidence bins, as seen in equation 5.6. $|Bb|$ denotes the size of a given bin, which is divided by the overall number of samples for weighting.

$$ECE = \sum_{b=1}^B \frac{|Bin_b|}{n} |acc(Bin_b) - conf(Bin_b)| \quad (5.6)$$

For a ranked-retrieval task, the measures of accuracy and confidence are adapted: mean average precision is used for accuracy while probe sample variance is used for confidence. In order for the difference $|acc(Bin_b) - conf(Bin_b)|$ to be meaningful, the variance values are first standardised to have mean 0 and unit variance (to reduce the impact of different models producing variances in different ranges), then projected into the range $[0,1]$ using the logistic sigmoid function to compress outliers. This is then subtracted from one to produce the confidence, as variance and confidence are inversely related.

5.5 Results

This section is broken down into two sub-sections. The first gives qualitative results demonstrating the behaviour of the uncertainty-quantification techniques on different sample images, and when blur, noise, and occlusion are introduced to the image. The second gives quantitative results demonstrating the impact on model performance and calibration of the uncertainty-quantification techniques.

5.5.1 Qualitative Samples

Qualitative assessment is used to give demonstrable and visual examples of the relative behaviours of the five uncertainty quantification methods. These visualisations can be used to identify if any methods are failing to behave as expected on easy images, and provide a way to interpret and understand the later quantitative results. The limitations of these experiments are discussed in each section.

Artificial Uncertainty

The first qualitative experiment is to examine the change in estimated uncertainty produced by each method as Gaussian blur, Gaussian noise, and pixel erasure (to

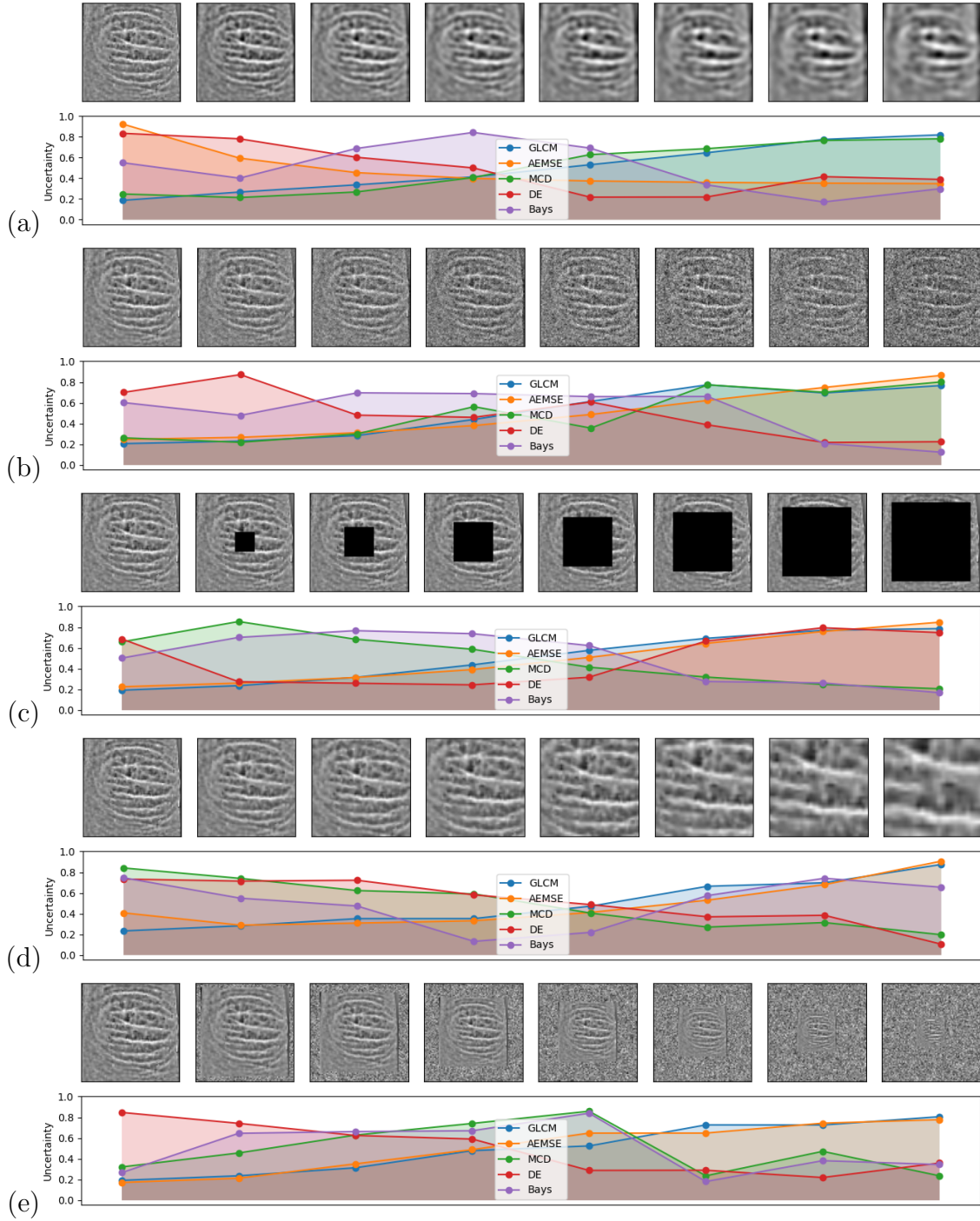


Figure 5.5: Change in measured uncertainty with increasing (a) blur, (b) noise, (c) occlusion, (d) zoom-in, and (e) zoom-out.

simulate occlusion) are increasingly added to an originally high-quality image. As well as this, scale change is simulated with zoom-in and zoom-out. The high-quality image used was manually selected from the knuckle images of 11kHands. For Gaussian blur, the image is convolved with Gaussian kernels of sizes 0 (no change), 3, 5, 7, 9, 11, 13, and 15. For Gaussian noise, the (accumulated) image is added to a matrix of random numbers multiplied by sigma values 0 (no change), 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, and 0.7. For occlusion, a central square is 'erased' (set to pixel value 0) with a height and width factor of the total image of 0 (no change), 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, and 0.8. For the zoomed-in images the image is centrally cropped with height and width factor of the original image of 1 (no change), 0.9, 0.8, 0.7, 0.6, 0.5, 0.4, and 0.3, then resized to the original image dimensions; for the zoomed-out images the original image is resized using the same factors as for zoom-in, then Gaussian noise is used to fill in the space to keep the image at the original dimensions (simulating image backgrounds).

The metric for uncertainty (MCD, BTL, DE, LML) or anomalousness (GLCM, AEMSE) is calculated for each of the modified images, and then normalised to have mean 0 and standard deviation 1 (separately for each method and modified image set) before being cast to the range $[0,1]$ with the sigmoid function to compress outlier values to 0 and 1. This normalisation allows the direct comparison of different methods, as they often produce values at different scales. Note that this process results in different normalised values for the first (unchanged) image in each set despite the raw values being the same, as the minimum and maximum are changed. The expectation for each of these tests is that the methods will produce their highest measures of uncertainty, variance, or anomalousness in the final instances, in which the image is most dissimilar to the training dataset and the visual information is corrupted so that any biometric comparison using that data should be unreliable.

The results of this experiment can be seen in Figure 5.5. No method showed a consistent trend in uncertainty for all of the types of introduced quality loss, and it is very clear that different methods can correctly quantify different types of uncertainty. Bayesian triplet loss (BTL) exhibited the same behaviour for almost all tests: an upward trend in uncertainty that peaks at the third or fourth instance, then a decrease in uncertainty through to the final instance. For the blur and the zoom-out tests there is a final slight up-tick in uncertainty at the final instance. This pattern is reversed for the zoom-in test, in which the uncertainty from Bayesian triplet loss drops to its lowest point at the fourth instance and then climbs again to the final instance. The deep ensemble (DE) showed the most undesirable behaviour, starting with the highest measured uncertainty in the base instance for all tests except occlusion, and then generally decreasing with each step. This pattern changes for occlusion, where

it begins to increase again from the fifth instance to the final. Monte-Carlo dropout (MCD) shows mixed behaviour, handling the uncertainty in some tests well and not in others. For the blur and noise tests, MCD appropriately starts with the lowest variance in the base instance and then increases with each step to the final instance. For the occlusion and zoom-in tests, it does the opposite, starting with the highest variances and then decreasing with each step. For the zoom-out test, the uncertainty increases to the fifth instance, then drops down again to its lowest value for the final instance. The two non-uncertainty quantifying methods, grey-level co-occurrence matrices (GLCM) and auto-encoder mean squared error (AEMSE), exhibit similar behaviour in all but the blur test and the most desirable behaviour for all methods. For each non-blur test they start with their lowest value in the first instance and steadily increase with step to have the highest value for the final instance. GLCM also does this in the blur test, whereas AEMSE does the opposite, starting with the highest value and decreasing to the lowest at the final instance. AEMSE’s behaviour in the blur test is predictable, as the auto-encoder reconstruction often loses fine details and appears as a blurred version of the original image, so would produce lower MSE scores for images that were already blurry.

The findings of this experiment are limited by the fact that only a single sample is used, which means it is unknown whether these behaviours are consistent across different samples. Similarly, while the sample was chosen from the training dataset, it is unknown whether it is an average sample or an outlier. If it is the case that it is an outlier and other training dataset samples have more blur or noise, then it would be expected to see higher uncertainty measures for this sample with methods which rely on the training dataset statistics (GLCM, and in theory, BTL, MCD, and DE) in the first unchanged instance than the next one or two stages of introduced blur and noise. What can be interpreted from these results is that the basic GLCM texture difference approach is most able to accurately describe the information loss across the five types; with the second-best being AEMSE. What cannot be interpreted is whether this information loss (e.g. a lot of blurring) correlates with the models ability to extract and match information from the image. In one test this is intuitively true - for the final stages of occlusion, there is no remaining information in which to extract features and form a valid match; however for other tests (blurring, noise, zoom-in, and zoom-out) it is unknown how many useable features the embedding model is able to extract.

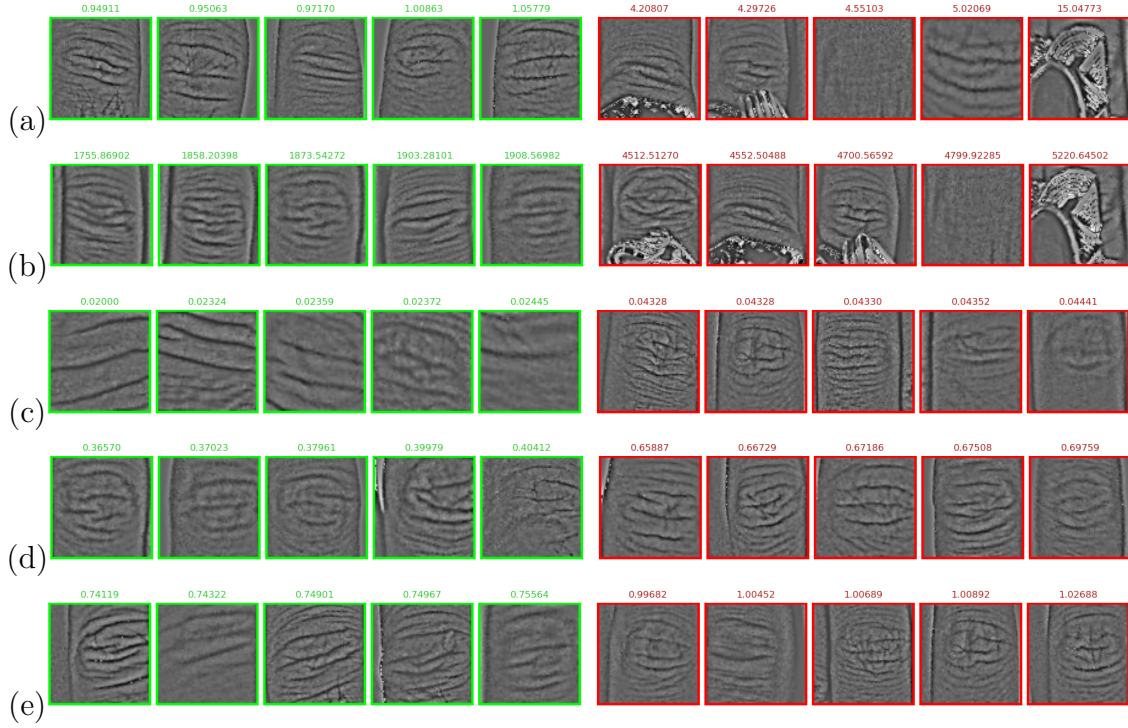


Figure 5.6: Lowest 5 uncertainties (green) and highest 5 uncertainties (red) in the 11k Hands test dataset knuckle images for (a) grey-level co-occurrence difference (GLCM), (b) auto-encoder mean squared error (AEMSE), (c) Monte-Carlo dropout (MCD), (d) deep ensemble (DE), and (e) Bayesian triplet loss (BTL).

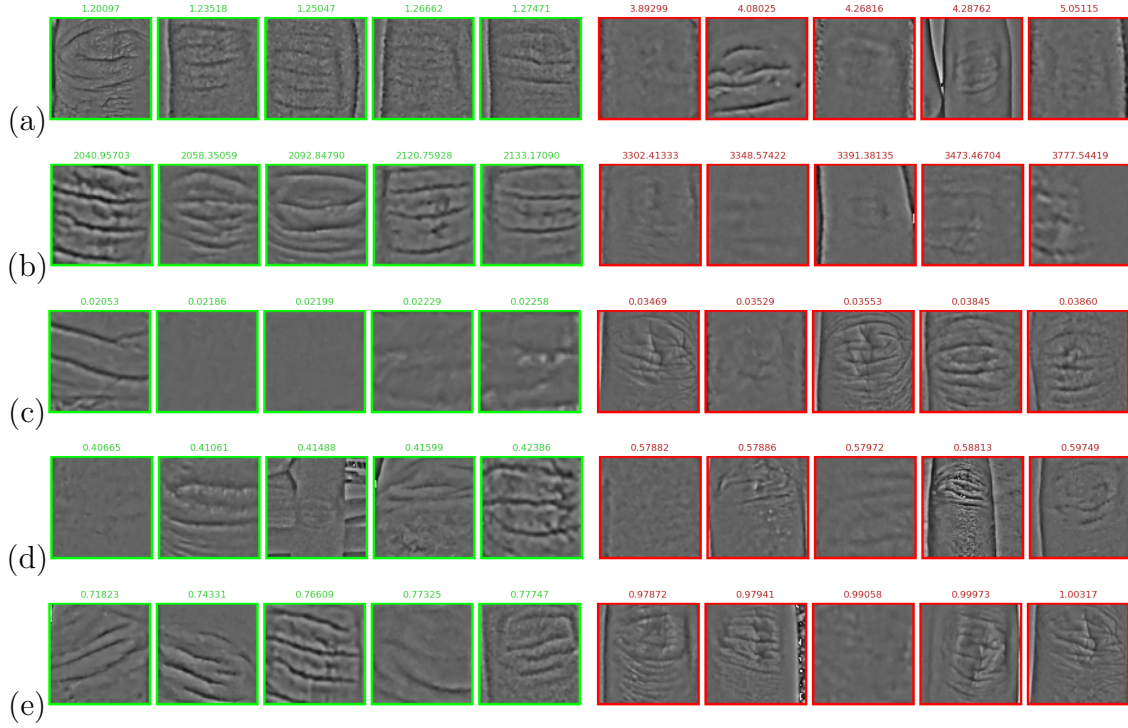


Figure 5.7: Lowest 5 uncertainties (green) and highest 5 uncertainties (red) in the VarHands dataset knuckle images for (a) grey-level co-occurrence difference (GLCM), (b) auto-encoder mean squared error (AEMSE), (c) Monte-Carlo dropout (MCD), (d) deep ensemble (DE), and (e) Bayesian triplet loss (BTL).

Whole-Dataset Evaluation

The second qualitative experiment was to process all images in both the 11k Hands test dataset and the VarHands dataset, probe and gallery sets combined, and observe which images are assigned the lowest and highest uncertainties by each method. Direct comparisons of the five images assigned the lowest and highest uncertainties or anomaly scores are presented in Figures 5.6 and 5.7; in order to diversify these results and prevent five images of the same class appearing, which would provide less useful insight into how the method is behaving, only the lowest and highest-uncertainty image of each class is shown. Figure 5.8 gives TSNE plots of VarHands knuckle image embeddings (limited to the index finger major knuckle for space) produced using the standard model and MCD model, with the image border indicating the relative AEMSE loss and MCD uncertainty. This TSNE visualisation shows the uncertainty to embedding correlation more broadly. Uncertainty values are not normalised for this analysis as only the order of the models’ values is needed.

Figure 5.6 gives the highest and lowest uncertainty-ranked knuckle images in 11k Hands. Images with accessories that were excluded during training are included in this evaluation. As accessories are non-unique (two separate people could wear the same ring, for instance), matching based on any features extracted from them (which will happen if they appear in the image) could result in false identifications. Even more so when they are excluded from the training dataset, the justification for which is given in Chapter 3, models are likely to conflate two similar looking accessories and project them to the same area of feature space. Given that images featuring accessories are not in the training dataset and so are out of distribution, it is expected that the uncertainty-quantifying methods should assign higher uncertainties to those images. On 11k Hands, AEMSE exhibits the most desirable behaviour, with all of the top 5 images (lowest MSE loss) being fairly clear, well-aligned images with strong knuckle crease patterns, and all of the bottom five images being ones that are unclear or contain accessories, making them significantly texturally different from the training dataset. This is also the case for GLCM. For MCD, all five images with the lowest variance were highly zoomed-in, which is consistent with the findings in the previous section, where for each stage in the zoom-in test the MCD variance decreased. The top-5 images for the deep ensemble and Bayesian triplet loss methods could subjectively be considered fairly high quality (except the second-highest BTL image), however for MCD, DE, and BTL, the bottom-5 images are also mostly of high quality and there is no appearance of the images with accessories or excessive blurring that are at the bottom of the GLCM and AEMSE rankings. One interesting note is that the top-5 images for BTL all appear to have the same slant to the knuckle creases, indicating some pattern to their assignment of a low uncertainty score.

Figure 5.7 gives the highest and lowest uncertainty-ranked images in the VarHands-Easy knuckle dataset. VarHands-Easy was chosen over VarHands as there was such a significant volume of occluded images in the VarHands dataset that visualisation of the bottom-5 did not offer any interesting insight into the relative behaviours of the methods. In VarHands-Easy, there are no visible accessories as there are in 11k Hands, but there are more instances of complete loss of visual information to blur and noise, as well as badly scaled images that give a very zoomed-out view of the knuckle. On Varhands-Easy, AEMSE again shows the most consistent and desirable behaviour, having the top-5 images as all well-scaled and centred images with clear knuckle crease patterns, and the bottom-5 as images that have too much blurring or noise for the crease patterns to be visible and extractable. GLCM similarly shows consistency in the images in the top and bottom-5, however the top-5 images are less clear and contain notable amounts of noise. The bottom-5 for GLCM, excluding the fourth-worst, are generally blurred with poorly visible crease patterns, the likely common factor between all 5 being a lack of fine-grained detail in significant portions of the image. For MCD, the top-5 images are again zoomed-in, but also include large amounts of blurring, including to where there is no visual textural information at all. The bottom-5 images for MCD are relatively high-quality compared to the bottom-5, with central positioning, good scaling, and visible crease patterns. The deep ensemble gives the most inconsistent uncertainty: of the top-5, there are images with no textural information, zoomed-in, zoomed-out, and with very strong textural information, and similarly for the bottom-5. BTL is consistent with which images are assigned low and high uncertainties: the top-5 images for BTL are all close-up with thick and highly contrasting lines for the knuckle creases, and the bottom-5 (with one exception where there is excessive blurring) are clear knuckle images where there are highlights on the skin creases from the illumination, giving them a more 3-dimensional effect compared to the top-5 images. While ideally BTL would have lower variance for more of the clear knuckle images (like those in the top-5 for AEMSE), the lighting effects the bottom-5 are out of the distribution of the training dataset and thus could be an indication that the BTL method is working (not producing collapsed uncertainty values) but in a less desirable way.

The TSNE plots in Figure 5.8 show the uncertainty or anomaly scores alongside the embedding distribution for each knuckle image. For both plots, similar invalid knuckle images such as images of bent knuckles or mis-detections showing fingernails are projected together in feature space (zoom-in may be needed to see the details). The AEMSE loss is well-correlated with these groupings, indicated by the green borders (lower loss) around the clear knuckle images and increasingly orange-red borders as information is lost to blur and mis-detection. The MCD uncertainty, by

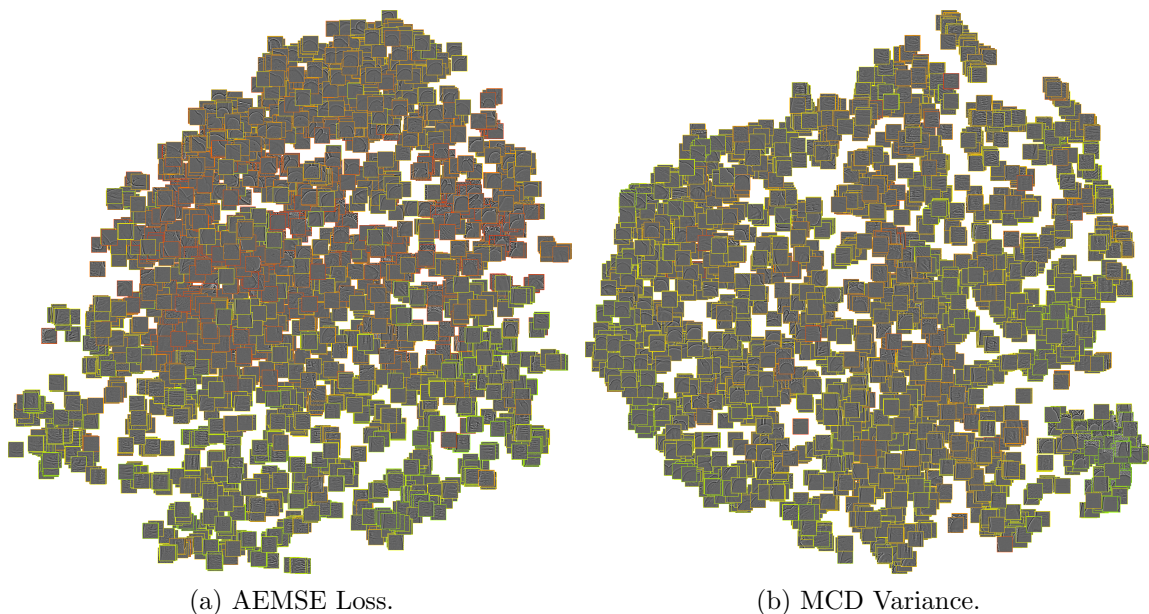


Figure 5.8: TSNE plot of embeddings of middle finger major dorsal knuckle images from VarHands, border colour indicating a) the AEMSE loss and b) the MCD uncertainty. Green indicates a lower loss/uncertainty and red indicates a higher loss/uncertainty.

contrast, is very ill-correlated with these groupings. In every area of feature spaces there are images of very low variance (green borders) and outlier images of very high variance (red border), independent of the validity or quality of those images for biometric purposes. In fact, for MCD, the images most consistently given low-variance are those that are significantly dissimilar from the training dataset - images of bent knuckles, fingernails, or containing strong background textures.

5.5.2 Uncertainty Fusion

The six fusion methods outlined in Section 5.4.6 were tested with the uncertainty values produced by each uncertainty quantification (or anomaly detection) method. Figure 5.9 demonstrates how three of these different fusion methods impact the distance calculations between a probe image (**A**, anchor), a genuine match (**P**, positive), and an imposter (**N**, negative) in the VarHands dataset, using only the index, middle, ring, and little fingers.

Using a non-uncertainty aware fusion method - the mean of the four distances -

	P	GLCM	AEMSE	MCD	DE	BTL	A	GLCM	AEMSE	MCD	DE	BTL	N
Index		0.7721 1.562	0.7721 2963.4	0.7379 1.58e-5	0.1817 6.8e-5	67.37 0.8741		1.1107 2.562	1.1107 3859.6	1.102 1.65e-5	0.1849 6.97e-5	92.98 0.9454	
Middle		0.5891 7.079	0.5891 3553.5	0.6227 1.79e-5	0.182 6.9e-5	17.59 0.9143		0.6994 6.959	0.6994 4263.8	0.7751 1.66e-5	0.1082 7.01e-5	69.04 0.8667	
Ring		0.9831 2.943	0.9831 6016.8	0.9227 1.84e-5	0.1845 8.1e-5	76.96 1.0847		0.4051 2.923	0.4051 6106.4	0.3992 2.11e-5	0.1728 8.03e-5	74.68 1.0778	
Little		1.1107 8.137	1.1107 4681.0	1.0943 2.19e-5	0.1803 8.05e-5	14.73 1.1568		1.0568 9.16	1.0568 6438.1	0.9095 1.92e-5	0.1867 8.04e-5	73.9 1.0872	
Fusion	Mean	0.8638	0.8638	0.8444	0.1821	44.16		0.818	0.818	0.7847	0.1812	77.65	
	Min Var	0.7221	0.7221	0.7379	0.1817	67.37		1.1107	1.1107	1.1052	0.1849	69.04	
	Bayesian	0.8601	0.8281	0.8289	0.1821	44.73		0.7998	0.8396	0.8032	0.1813	77.67	

Figure 5.9: An example of three fusion methods applied to five uncertainty measures to combine the index, middle, ring, and little finger distances for an genuine pair (**AP**) and a difficult imposter pair (**AN**). For each pair of images, the bold upper number represents the feature distance and the lower number represents the variance or uncertainty of that distance.

results in the distance to the negative sample being shorter than the distance to the positive sample (for GLCM, AEMSE, MCD, and DE, but not for BTL). Looking at each knuckle individually, the distances between the index and middle fingers to the positive sample are correctly shorter than the distances to the negative; as this comparison is for entirely visible and high-quality knuckle images, this assessment is valid. For the ring and pinky fingers the distances are lower for the anchor and negative sample which have similar mis-detections, as the model is not equipped to distinguish those images and sees more similarities between them than with a knuckle image (as in the positive sample) - this assessment is invalid.

The uncertainty values produced by the AEMSE, MCD, and DE methods demonstrate how uncertainty quantification could be used to smartly combine information based on its certainty, as opposed to giving all inputs equal weight. For these methods, the uncertainties produced for the index and middle comparisons were lower than the uncertainties for the ring and little finger comparisons, accurately reflecting the lack of information available in the anchor image for the latter two. Using Bayesian fusion with AEMSE loss then results in the anchor-positive distance being brought down from 0.8638 to 0.8281 and the anchor-negative distance being pushed

Fusion	Acc@1	mAP	ECE
Mean	0.7979	0.1976	0.3036
Minimum	0.5012	0.1095	0.3914
Bayesian	0.7985	0.2024	0.2988
Image-Weighted	0.7812	0.2001	0.3008
Region-Weighted	0.7392	0.1891	0.3119

Table 5.1: Ranked retrieval results of different fusion methods on dorsal knuckle images in VarHands, using the AEMSE uncertainty score.

up from 0.818 to 0.8396, correcting their position in the ranking. For MCD, the anchor-positive distance is decreased and the anchor-negative distance is increased, but not by enough to change their order. Most notably, for all methods except BTL, using only the distance with the minimum uncertainty (performing no fusion at all) results in the best separation of these distances - very low distance for the anchor-positive and very high distance for the anchor-negative.

The next section covers in detail the quantitative ranked-retrieval performance when using each uncertainty quantification method. Table 5.1 compares six fusion methods, including the three that rely on statistics calculated from mass comparisons (threshold, image-weighted, and region-weighted) when using AEMSE loss. Only non-normalised Bayesian fusion outperforms mean fusion, achieving +0.6% rank-1 accuracy and +0.0048 mAP, and -0.0048 ECE. The ‘minimum’ method, using only the distance with the smallest variance, results in a drastic drop in performance of -29.67% rank-1 accuracy, -0.0881 mAP and +0.0878 ECE, in contradiction with the example given in Figure 5.9. The two methods which performed normalisation of the uncertainty values, image-weighted and region-weight, performed worse than non-normalised Bayesian fusion and worse than not incorporating uncertainty at all (mean). Region-weighting in particular saw a large drop in rank-1 accuracy of -5.87%. Based on these results, only non-normalised Bayesian fusion is used to compare all of the uncertainty quantification methods in the subsequent experiments.

5.5.3 Ranked Retrieval Performance

The relative performances of each uncertainty-quantification method are evaluated quantitatively using a ranked-retrieval task, and the same evaluation metrics as are described in Chapter 3 Section 3.2.3: rank-1 accuracy (Acc@1), precision, recall, mean average-precision (mAP), and now also expected calibration error (ECE).

The results for ranked-retrieval with Bayesian distance fusion are given in Table

5.2. The results are given for the 11k Hands test set, which provides the baseline performance on ideal images of many subjects, and the VarHands and VarHands-Easy datasets. The evaluation on VarHands given in this table is for all illuminations combined, where rank-1 accuracy can be high due to matching images from the same video source, i.e. possibly adjacent frames, but mean average-precision is notably lower as cross-illumination images aren’t matched (demonstrated in particular in the VarHands-Easy results where pose difference is removed).

Rank-1 retrieval on 11k Hands is roughly equivalent for all methods, with the cap of 99.75% achieved by four: deterministic, GLCM, AEMSE, and MCD. Slight differences in performance are seen with mean average-precision and expected calibration error. The best mAP (overall ranking quality) seen on 11k Hands was 0.9861 by MCD however all methods achieved >0.98 . The expected calibration error however was highest for MCD at 0.4878, and lowest for GLCM at 0.4708. On VarHands, the best rank-1 accuracy, mAP, and ECE were achieved using the AEMSE method, at 79.85%, 0.2024, and 0.2988 respectively. AEMSE was the only uncertainty-based fusion method to outperform the basic deterministic model and mean combination of distances. Deep ensemble, despite achieving an average rank-1 accuracy, produced a much lower mAP and much higher ECE than the other methods on VarHands. On VarHands-Easy, where pose variation has been removed, the MCD method achieved the highest rank-1 accuracy and mAP, at 97.21% (joint best with Det.) and 0.7095. Interestingly, MCD had the worst calibration error at 0.2899, with the best of 0.1861 achieved by GLCM. This data suggests that AEMSE is more able to correctly handle variation in pose, as it achieves higher recognition accuracy for the whole VarHands dataset than any other method, but MCD and GLCM are better at handling uncertainties when pose variation is removed as in VarHands-Easy. This is consistent with the qualitative results, which demonstrated that GLCM and MCD produce high uncertainty for blurry knuckle images whereas AEMSE does not.

Bayesian Triplet Loss demonstrates lower accuracy and mean average-precision than the other methods, with the use of mean or Bayesian fusion. Using Bayesian fusion over mean slightly improves the performance of the BTL model - in mAP for all datasets, Acc@1 in VarHands, and ECE in VarHands and VarHands-Easy. Despite being a less effective embedding model, BTL achieves some of the lowest ECE scores for 11kHands and VarHands-Easy, showing its uncertainty estimates for those datasets are well-calibrated with its accuracy (i.e. where its embeddings are poor and causing low accuracy, it is producing a high uncertainty, and vice versa). This is not the case for the VarHands dataset where it produces ECE scores in line with the others (and an even more significant drop in performance), indicating it is failing to produce well-calibrated uncertainty estimates for the low quality images

Model	Fusion	11k Hands			VarHands			VarHands-Easy		
		Acc@1	mAP	ECE	Acc@1	mAP	ECE	Acc@1	mAP	ECE
Det.	Mean	0.9975	0.983	-	0.7979	0.1976	-	0.9721	0.7037	-
GLCM	Bayesian	0.9975	0.9829	0.4708	0.7713	0.1938	0.3302	0.9674	0.7041	0.1861
AEMSE	Bayesian	0.9975	0.9836	0.48	0.7985	0.2024	0.2988	0.9698	0.7067	0.201
MCD	Mean	0.9975	0.9854	0.4871	0.8004	0.1985	0.3397	0.9744	0.7135	0.2832
MCD	Bayesian	0.9975	0.9861	0.4878	0.7967	0.1966	0.3381	0.9721	0.7095	0.2899
DE	Mean	0.9963	0.9809	0.4848	0.7818	0.1286	0.3813	0.9674	0.5416	0.2359
DE	Bayesian	0.9963	0.9813	0.4852	0.7874	0.1292	0.3809	0.9651	0.5424	0.2385
BTL	Mean	0.9103	0.7852	0.2958	0.5865	0.119	0.3828	0.8419	0.3666	0.185
BTL	Bayesian	0.9079	0.7855	0.2959	0.5971	0.1196	0.3822	0.8349	0.3677	0.1847

Table 5.2: Ranked retrieval results for different uncertainty quantification methods on dorsal knuckle images in 11k Hands, VarHands, and VarHands-Easy.

that result from pose variation (such as invalid images due to mis-detection).

5.6 Discussion

Five methods of uncertainty quantification or anomaly detection were implemented with a ResNet-50 backbone network and trained with a contrastive triplet loss. This included auto-encoder reconstruction error and grey-level co-occurrence matrix difference for anomaly detection, and Monte-Carlo dropout, deep ensembles, and Bayesian triplet loss for uncertainty quantification. These methods were trained and evaluated on greyscale, illumination-normalised dorsal knuckle images. An ideal image was manually selected and then modified with increasing amounts of artificial quality loss in the form of blurring, noise, occlusion, zoom-in and zoom-out, in order to compare and understand how the uncertainty or anomaly scores produced by each method vary under different image conditions. In addition to this, each image in the test datasets was processed using each method and then ranked according the uncertainty or anomaly score, in order to see how the methods work on the real data as well as artificial data. Finally, the methods were compared in a ranked-retrieval scenario that allows the quantification of performance using the metrics of rank-1 accuracy, mean average-precision, and expected calibration error.

The qualitative results indicate a broad failure of all uncertainty quantification methods to capture well-calibrated uncertainty measures, or measures that aligned with what would naturally be considered ‘valid’ or ‘invalid’ knuckle images. The two anomaly detection methods, GLCM and AEMSE, were the closest to achieving the desired behaviour - assigning high values to blurry, occluded, noisy or otherwise low-quality knuckle images and non-knuckle images - in the controlled tests, but for

GLCM this did not translate into good performance in the quantitative or whole-dataset tests; Bayesian fusion with AEMSE on the other hand resulted in better performance over mean fusion on all test sets. When looking at the quantitative results, mean average-precision is the more important metric to evaluate the impact that the fusion method is having on the rank ordering, as accuracy at rank 1 only shows how a single rank position is changing and not whether matching items on the whole are being moved upwards in the ranking. Without being able to demonstrate that the methods achieve consistent and well-calibrated uncertainty estimates the insights from the ranked-retrieval test are limited. However, use of Bayesian fusion over mean fusion achieved an improvement in mAP for all of the methods (On all datasets for AEMSE, BTL, and DE, on 11k Hands for MCD, and on VarHands-Easy for GLCM).

The improvement seen with Bayesian fusion is only marginal, possibly because, looking back at Figure 5.9, there is not enough separation between the low and high uncertainty measures for any given model. Ideally the uncertainties for clear and in-distribution knuckle images would be close to zero, which would heavily skew the fused distances towards those images. As seen in these tests however the uncertainties given for those clear knuckle images are very close to the uncertainties for low quality and invalid knuckle images: for AEMSE, the lowest loss value is 2963.4 and the highest is 6016.8, for MCD it is $1.58\text{e-}5$ and $2.19\text{e-}5$, and for BTL it is 0.8667 and 1.1568. While the results show potential for uncertainty-aware fusion, to indisputably demonstrate that it can improve performance over standard fusion methods it would need to be applied with well-calibrated uncertainty measures. Having a dataset manually labelled by a human assessment of the image quality is one potential way to experiment with uncertainty-aware fusion methods before the development of suitable uncertainty-quantification methods in this domain, but would require substantial time input and, as discussed in the introduction to this chapter, a human assessment of image quality would not necessarily be consistent with the model’s certainty (for example, a human might consider an image high quality which contains physical features that are not well-represented in the training dataset, meaning the model could be very inaccurate with it).

Obtaining well-calibrated uncertainty measures is a difficult task and uncertainty learning often requires substantially more data than is necessary for learning deterministic embeddings. Learning aleatoric uncertainty, the per-sample noise, requires a large variation in the representation of each *class* in the data for the model to understand what information can signal noise, quality-loss, or variation within the class, and what information signals an entirely new class. Because, for example, images of bent knuckles are not present in the training dataset, the models cannot learn

that those images are low-certainty variations on the types of images it has seen, and instead sees them as different looking knuckles with high certainty. With 11k Hands as the training dataset there are approximately 14 images per class and they each have very little variation; for the extracted knuckle regions the only variation may be minor errors in localisation and registration. Even with augmentation, this provides little opportunity for the model to learn per-class distributions as would be seen in unconstrained data. To learn epistemic uncertainty, the model needs to be able to accurately approximate the true posterior distribution of the embeddings. To do this, it must be trained on a large dataset that is very representative of the types of knuckle images it will see at inference time. Without a well-approximated posterior it is possible for the model to predict even well in-distribution samples with high uncertainties; this is demonstrated with the BTL results in Figure 5.7, where it sees knuckle images with some illumination features not present in the training dataset (highlight of the skin ridges) as high uncertainty.

In order to address this problem of poorly-calibrated uncertainties the models would need to be trained on a much larger dataset that is more similar to the unconstrained evaluation dataset, containing low-quality images and mis-extractions. The models will not be able to learn to embed these low quality images so that they can be correctly matched with high-quality images of the same subject (the information to do this is simply not present) but it will allow them to learn that these are low-certainty images and treat them accordingly at inference time. To conclude, the results in this chapter show that there is promise in the use of uncertainty quantification and uncertainty-aware approaches to multi-biometric fusion (e.g. Bayesian combination), but before the work continues a substantial improvement in the size and representation of hand image datasets will need to happen. The next chapter delves into the legal regulations around biometric technologies used for forensic and law enforcement purposes, covering data requirements and discussing how sufficiently sized datasets should be defined in relation to proximal tasks (e.g. facial recognition) and expected performance.

Chapter 6

Hand Recognition in Practice

This chapter presents an analysis of the legal and regulatory landscape around the use of biometric technologies by law enforcement (in the UK and EU), related to the emerging area of hand recognition as is developed in this thesis. The systems considered for this, as defined in Chapter 1, are unconstrained hand recognition systems proposed for application in forensic information retrieval; specifically, an unconstrained image (perhaps a controlled and high quality suspect reference image, or perhaps any image obtained from the wild) being used as the search term (‘probe’) to rank a database (‘gallery’) by likelihood of containing the same hand. An important part of the development of this technology is to consider what would be required in terms of data, performance, and security in order for it to be legally and ethically used in practice.

The use of AI systems for law enforcement in the EU is governed by a combination of general data protection regulations (GDPR) [131], AI-specific laws (AI Act) [132], and law enforcement specific laws (Law Enforcement Directive) [130] all interacting with the rights to privacy afforded by Article 8 of the European Convention on Human Rights (ECHR) and Articles 7 and 8 of the Charter of Fundamental Rights of the European Union. Biometric systems are particularly contentious with the Article 8 right to privacy, and require the utmost care in data protection due to the inherently individualising nature of the processed data. For this reason, biometric data and systems are one of a few categories singled out as ‘sensitive’ and ‘high-risk’, requiring extra precautions in their deployment. Throughout this chapter these laws are additionally interpreted through the judgments of the Court of Justice of the European Union (CJEU) in several high-profile data protection cases. Interpretations from the CJEU in these instances can give good insight into how they would be likely to interpret similar questions in the future.

For several reasons, EU laws are considered in this analysis over local UK laws. Firstly, many of the UK’s implementations of data protection and AI laws are inherited from the EU, with only minor changes; the EU GDPR is implemented in the UK through the UK GDPR and supplements in the UK Data Protection Act (DPA) Part 2, and the EU Law Enforcement Directive is implemented as Part 3 of the UK DPA. Secondly, the EU constitutes the largest single market in the world and is in close proximity to the UK, with much cooperation in the areas of law enforcement and intelligence; compliance with EU law allows the technology to have the widest impact. Thirdly, EU standards are the highest in the world for data protection and human rights, so any EU-law compliant technology is likely to be acceptable anywhere and as future-proof as possible. The high standards of the EU laws promote high technical quality, putting security, accuracy, and safety above convenience; adherence to EU law would most limit the potential negative impacts of the technology while building public trust.

The rest of this chapter presents the research question to be answered, a discussion of the legal definitions and how they might be interpreted, the regulatory scope (which laws apply) under those interpretations, and then finally, a discussion of the requirements needed to be met to bring the technology forward.

6.1 Research Questions and Contributions

The subset of research questions addressed in this chapter are:

RQ7 What steps are needed for compliance with the relevant regulations before hand recognition technologies can be used for law enforcement purposes?

To answer this question, a review of the relevant regulation and court judgments was performed with the aim of determining a) the laws that apply, b) the consequences of those laws applying, and c) what objectives must be achieved to comply with those laws. The scope of laws and thus which apply are discussed in section 6.3. The consequences of categorisation as a high-risk AI system are discussed in section 6.4. Finally, the objectives that must be achieved are described in section 6.5.

6.2 Definitions

This section covers the relevant definitions in the law and discusses where they diverge from the common language used in technical biometric literature. Additionally, some instances in which the definitions may be more vague are considered, as they could have possible implications for the scope and consequences that are discussed in later sections.

Categories of personal data are described in the GDPR Article 4 (Definitions). The other discussed regulations, the EU AI Act and the EU Law Enforcement Directive, inherit their definitions of these from the GDPR. ‘Personal data’, per Article 4 Paragraph 1, means “any information relating to an identified or identifiable natural person (‘data subject’)”. ‘Biometric data’, Article 4 Paragraph 14, is “personal data resulting from specific technical processing relating to the physical [] characteristics of a natural person, which allow or confirm the unique identification of that natural person”, with the examples given of fingerprints (‘dactyloscopic data’) and facial images. Biometric data can be used for a number of purposes, including ‘authentication’, ‘identification’, ‘categorisation’, and ‘recognition of emotions’ (AI Act Recital 14). Biometric data is one of the “special categories” of personal data (sometimes referred to as “sensitive” personal data) described in Article 9 (Processing of special categories of personal data) Paragraph 1, which refers to the processing of “biometric data for the purpose of uniquely identifying a natural person”. GDPR Recital 51, which covers the processing of special categories of personal data, states “processing of photographs should not systematically be considered to be processing of special categories of personal data as they are covered by the definition of biometric data only when processed through a specific technical means allowing the unique identification or authentication of a natural person”. Interpreting the recitals, it may be reasonable to treat the processing of biometric data that could *allow* the identification or authentication of an individual as the processing of special category data, irrespective of the specific purpose for that processing (be it for anonymous categorisation, emotion recognition, or image matching), despite the Article 9 stipulation that it be “for the purpose of uniquely identifying”. Whilst ‘anonymous’ is used here to denote that it is not necessarily associated with any other identifying information such as a name, biometric data is always personal data since it *allows* identification and relates to an identified or *identifiable* individual (and this is the interpretation agreed by the ICO [126]). There is also contradiction around whether the underlying images are biometric data or not (facial images are included as an example in Article 4 Paragraph 14), however as in the proposed system all images would be processed into biometric templates this question is out-of-scope for this chapter.

Actors involved in the production and use of AI or data processing are identically defined as “natural or legal person[s], public authorit[ies], agenc[ies] or other bod[ies]” in both of the EU GDPR and EU AI Act; as such, those terms will be substituted herein with the single term “actor” for avoidance of repetition. In the GDPR, data controllers and processors are defined in Article 4 Paragraphs 7 and 8 respectively. A controller is an “[actor] which [] determines the purposes and means of the processing of personal data”, while a processor is an “[actor] which processes personal data on behalf of the controller”. In addition to these terms which relate to the processing of personal data, the EU AI Act provides definitions for AI system providers and deployers in Article 3 (Definitions) Paragraphs 3 and 4 respectively. A provider is an “[actor] that develops an AI system or general-purpose AI model [] and places it on the market or [] into service under its own name”. A deployer is an “[actor] using an AI system under its authority”. It is possible that the same “natural or legal person, public authority, agency or other body” can be both the provider and deployer of an AI system, where the AI system is developed in-house to be used (deployed) by the same actor.

There are a variety of terms to describe the specific systems or actions that involve the use of biometric data: ‘recognition’, ‘identification’, ‘verification’, and ‘authentication’. In the technical literature many of these terms are used interchangeably. ‘Biometric recognition’ is often used as a general term to inclusively describe all other biometric systems, including identification, verification, and authentication; the Information Commissioner’s Office (ICO), which upholds the GDPR in the UK, however gives the following guidance: “Biometric recognition describes when you use biometric data to uniquely identify someone” [125]. ‘Biometric identification’ is described in the EU AI Act Article 3 Paragraph 35 as “the automated recognition of physical, physiological, behavioural, or psychological human features for the purpose of establishing the identity of a natural person by comparing biometric data of that individual to biometric data of individuals stored in a database”. ‘Biometric verification’ is defined in Paragraph 36 as “automated, one-to-one verification, including authentication, of the identity of natural persons by comparing their biometric data to previously provided biometric data”. So, in legal terms, biometric recognition is a colloquial umbrella term for the two primary biometric systems considered in law: one-to-N biometric identification, and one-to-one biometric verification/authentication.

What is not clearly defined or covered in the law are biometric systems which are not (directly) for the purpose of identification. The same underlying technology, which includes the extraction of biometric templates from images and comparison or ‘matching’ of two biometric templates, forms the basis of identification, verification,

and other systems such as image search and organisation. A biometric search system may involve the retrieval, ranking, or grouping of images based on a shared human subject, determined by the biometric data. For this process, no identity *needs* to be associated with either the search image (the ‘probe’) or the database images, and the direct purpose of the system is only to organise the data in some way, not to assign or confirm an identity. By the legal definitions, a biometric search system may not be considered a biometric technology covered by the EU AI Act (and thus the EU Law Enforcement Directive). Hand images used in a hand biometric search system (probe or database) may not even be considered a special category of personal data under the GDPR, assuming that it is not used “for the purpose of uniquely identifying a natural person”. Choosing to interpret the legislation this way could potentially lead to down-stream problems; technicalities of language should not be relied on and instead the intention behind the laws considered, as if an organisation’s interpretation is seen to be in contradiction with the intention of the law then that organisation is susceptible to unfavourable rulings. It is the safer option to go with the more human rights-protective interpretation, which in this case would be that processing biometric data for search and retrieval systems that *could* lead to identification is the processing of special category data, and that the potential for identification through these systems means they should be treated as biometric identification systems. This is discussed in more detail in the next two sections, with references to specific cases where judges have determined the wording of the law itself to be insufficiently protective.

6.3 Regulatory Scope

This section covers the scope of the GDPR, the EU AI Act, and the Law Enforcement Directive (LED) in regards to the proposed hand biometric system. Separate considerations are given for the training and development of models, and the deployment of models. Several possible interpretations of the applicable laws are given and explored in regards to what consequences they would have, before arriving at the likely most suitable interpretation in light of several court judgments.

The EU GDPR (UK GDPR and DPA in the UK) covers the processing of personal data “wholly or partly by automated means” (Article 2 Paragraph 1), with the exception of processing “by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences” (Article 2 Paragraph 2 Item D). The EU Law Enforcement Directive (implemented in the UK as Part 3 of the DPA) covers the processing of personal data “wholly or partly by automated means” (Article 2 Paragraph 2) for the purposes of “the prevention, investigation,

detection or prosecution of criminal offences or the execution of criminal penalties” (Article 2 Paragraph 1 and Article 1 Paragraph 1). The EU AI Act covers the provision, deployment, import and distribution of AI systems, plus the use of AI system outputs within the EU (Article 2 Paragraph 1). As such, the processing of data for the development of the AI systems and the data processing done in the course of using the AI system are covered variably by the GDPR and Law Enforcement Directive, while the AI system itself and its applications are covered by the AI Act. The permissions granted in each of these regulations are not absolute, and any action must be weighed against the rights afforded to individuals in the European Convention on Human Rights (ECHR) and the Charter of Fundamental Rights of the European Union. As a note, the ECHR is an instrument of the Council of Europe of which the UK is a part of, while the Charter is an instrument of the European Union and so does not directly apply to the UK. In general, the Charter encompasses the same rights as the ECHR but goes further on social issues; alongside the right to privacy (Article 8 ECHR, Article 7 Charter) the Charter contains the additional right to protection of personal data (Article 8). Interpretation of the Charter in specific cases is done by the Court of Justice of the European Union (CJEU), and their decisions are binding as EU law. The ECHR is interpreted by the European Court of Human Rights (ECtHR).

Since the jurisdiction of the GDPR and LED depend on whether the purpose of the processing falls under “prevention, investigation, detection or prosecution of criminal offences”, it is necessary to separately identify the purposes for processing data at each stage of the proposed system’s lifetime. In developing the system, i.e. the training, validation, and testing of the model, it is necessary to process personal data; the direct and immediate purpose of this is not related to tackling criminal offences, so such processing may fall under GDPR rather than the LED. In many instances, biometric recognition technologies (such as facial recognition technology used by the Met police [159]) are developed by private companies bound by GDPR, then sold as software to law enforcement agencies, for which data processed in their use is governed by the Law Enforcement Directive. It is becoming more common for law enforcement agencies to desire their technologies to be developed in-house, for better security and fine control over the functionalities of the systems. An argument could be made that if the technology was to be developed in-house by the law enforcement agencies, the processing would have an indirect (though far less indirect than if developed by a private company) purpose of tackling criminal offences, and as such could fall under the LED.

In terms of privacy rights for the data subjects, the GDPR enforces stricter rules and requirements on the processing of data than the Law Enforcement Directive.

For one, the GDPR requires that data is processed “in a transparent manner in relation to the data subject” (GDPR Article 5 Paragraph 1 Item A, ‘lawfulness, fairness and transparency’), meaning the data subject should be aware that their data is being processed and what for, while the LED loses this requirement, allowing data processing to be done without transparency to the subject (LED Article- 4 Paragraph 1 Item A). Processing under the LED, however, requires strict logging practices, making it far more transparent to regulators. Processing must meet one of seven points to be considered lawful under the GDPR, which include legal obligation of the controller, protecting vital interests, and explicit consent of the data subject (GDPR Article 6 Paragraph 1); whereas processing is considered lawful under the LED as long as it is *necessary* for a competent authority to carry out one of the LED purposes (LED Article 8 Paragraph 1). Data processed under the GDPR must also only be processed for the explicit purpose it was collected for and not re-used in any way that is incompatible with that purpose (GDPR Article 5 Paragraph 1 Item B, ‘purpose limitation’), while data processed under the LED can be re-used for another purpose under the scope of the LED, so far as that re-use is considered *necessary* and *proportional* (LED Article 4 Paragraph 2). It is worth noting that processing under the GDPR *can* be very permissive, depending on how broad the initial purposes consented to were or how far the re-use processing could be considered a legitimate interest.

So the question of whether the processing of data in the course of developing an AI model with the down-stream purpose of tackling criminal offences should be governed by the GDPR or the LED hinges on whether it is a *necessary* and *proportional* step to that aim. Several articles in the Law Enforcement Directive point to this not being an area it is intended to cover. One is that all data processed under the LED must be clearly labelled by whether it pertains to a suspect, a convicted criminal, a victim, or another associated with a crime (e.g. a witness) (LED Article 6), which would not comfortably apply to data collected for model training. Another is that all data must be designated as based on fact or personal assessment (assumption) (LED Article 7), which again, would not comfortably apply to training data (it is assumed that all training data is factual). The only other avenue for processing data for model development under the LED would be to dual-use data collected in the course of investigations for other purposes as training data, however, this would be inappropriate and ineffective (the volume of data collected in this manner is likely to be orders of magnitude lower than desired training datasets, see later Section 6.5.2).

As for its necessity or proportionality, it is unlikely that the courts would see processing under the LED for this aim as a reasonable interference with Article 7 and 8 rights to privacy and data protection. Judgments from the CJEU indicate

a tendency towards interpretations of regulations that are most protective of the human rights afforded in the Charter, whether that means taking a very narrow interpretation of the law as written or a more broad and generalised view of it. In *Digital Rights Ireland* (C-293/12) the court nullified a national directive (Data Retention Directive) which was designed in accordance with EU law as written, on the basis that it was not a proportional interference in Article 8 rights [80]. In *Schrems I* and *II* (C-362/14) and (C-311/18) the court determined that transfer of data to the US, in the face of their surveillance practices, was in contradiction of the intention of the law even if it was technically permitted [79, 76]. In *Google Spain v AEPD* (C-131/12), the *Right to be Forgotten* case, the court limited the scope of ‘legitimate aims’ on the tangible impact it was having on privacy rights, and notably took a wider-than-written interpretation of questions of territorial scope and responsibility [77]. In *La Quadrature du Net* (C-511/18 and others), the court ruled on the nature of proportionality and necessity of an action (blanket data collection and processing to detect terrorist threats) based on the impact it had on rights - choosing a much more narrow interpretation of the lawful purposes [78]. These cases are significant not only in their lawfully binding judgments, but because they demonstrate the courts attitudes to various issues. The court’s primary aim is to protect the fundamental rights afforded in the Charter, including Article 8 - Data Protection, and so will rule on interpretations of the law that are most compliant and even nullify laws that are in contradiction. Based on this, it could be assumed that the court would likely rule on the side of all model training data being collected under one of the GDPR purposes (e.g. consent), and then processed and maintained (not held indefinitely or used for another incompatible purpose) under GDPR rules. Data processed in the application of the model at later stages (investigation) however would fall under the LED.

Regardless, data processed through the model at application time, during investigations, would be processed under the Law Enforcement Directive rather than the GDPR. Article 13 Paragraph 2 (a) of the LED states that the “legal basis for the processing” must be made available to the data subjects on request, and per Recital 41 of the GDPR “a legal basis or legislative measure should be clear and precise and its application should be foreseeable to persons subject to it”. The legal basis must establish the permissions on which the processing is occurring and must be accessible and understandable to the data subjects.

The AI system itself is governed primarily by the EU AI Act, which categorises AI systems by several criteria - in addition to those categories covered in the prior section, also whether a system is ‘remote’, and then ‘real-time’ or ‘post-remote’. A ‘remote’ biometric system, per Article 2 Item 41 of the EU AI Act, is “an AI system

for the purpose of identifying natural persons, without their active involvement, typically at a distance”, which would include both pointing a camera at someone in a physical space and collecting their images through some other means such as over the internet. Remote biometric systems are then categorised as ‘real time’ “whereby the capturing of biometric data, the comparison and the identification all occur without a significant delay”, per Item 42, or ‘post-remote’ meaning a system other than a real-time system. The hand recognition system of this project would fall under the definition of a post-remote biometric system as it would not involve the active participation of the image subjects and would be applied to historic images rather than any real-time acquisition. How the system is defined according to the AI Act then pertains to whether it is classified as high-risk, is permissible for law enforcement under certain circumstances, or is totally prohibited. Prohibited AI systems are given in Chapter 2 (Article 5) including of special note the use of real-time remote biometric systems by law enforcement in a public space¹. Post-remote biometric systems are not prohibited, even when used for law enforcement purposes, and are likely to fall under ‘high-risk’ AI systems. The qualifying criteria for high-risk AI systems are given in Chapter 3 Section 1 (Articles 6-7) of the EU AI Act. Article 6 Paragraph 2 of the AI Act states that high risk technologies are those specifically referred to in Annex III, except in the cases given in Article 6 Paragraph 3, which describe “where [the AI system] does not pose a significant risk of harm to the health, safety or fundamental rights of natural persons”. If the system in question is classified as high-risk, then it is covered in scope by the additional regulations in Sections 2 through 5 in Chapter 3, including requirements for the AI system itself in Section 2 and obligations for the provider of the AI system in Section 3.

6.4 High-Risk AI

This section covers the regulation of high-risk AI under the EU AI Act. Annex III of the Act defines high-risk AI systems with reference to specific technologies, of which the two relevant sections are Paragraph 1, biometrics, and Paragraph 6, law enforcement.

The specific biometric systems designated high-risk in Annex III Paragraph 1 are emotion recognition, biometric categorisation by sensitive or protected attributes, and remote biometric identification except “to confirm that a specific natural person

¹With the exception of “the targeted search for [] missing persons”, “the prevention of a specific, substantial and imminent threat to [] life or physical safety”, and “the localisation or identification of a person suspected of having committed a criminal offence []”. One example of something not excepted would be deployment to prevent petty crimes such as theft

is the person he or she claims to be” (verification). Per Recital 54 of the Act, these critical use-cases of biometric technologies are considered high-risk because of their processing of biometric data, a special category of personal data; remote identification systems in particular are considered high-risk because of the risk of technical inaccuracies and bias results to “entail discriminatory effects” - additionally, Recital 17 which discusses remote biometric systems, states that they are more likely to have a non-minor impact on fundamental rights, compared to one-to-one verification, due to being applied to a “large number of persons without their active involvement”, or in other words that they are applied broadly and not targetted.

In addition to those instances listed, Chapter 3 Article 7 outlines the circumstances in which a technology not already included in Annex III may be added to it. Of particular importance to this analysis of the definitions and classifications of biometric systems are Paragraph 2 Item A, “the intended purpose of the AI system”, Item B, “the extent to which an AI system has been used or is likely to be used”, Item C, “the nature and amount of the data processed”, and Item D “the extent to which the AI system acts autonomously”. Item F also considers the AI system’s ability to “disproportionately affect a particular group of persons”, which is a common outcome of bias in biometrics. As the system proposed in this project is a post-remote biometric system, it would likely be considered high-risk. Two factors may change this: one, whether the system in question meets the definition of a biometric *identification* system (as discussed in Section 6.2), and two, whether it can be shown to be low-risk according to the criteria in Article 6 Paragraph 3.

Whether the the hand recognition system is likely to be classified as a biometric identification system is a question of purpose. Identification systems are those “for the purpose of establishing the identity of a natural person”; while this may not be the *direct* result of the proposed system as the data is (potentially, or potentially not) anonymous, identification could be seen as a down-stream goal of the system, as it is used to collect information and make connections that will eventually result in an identification. Based on the judgments discussed in Section 6.3, the courts demonstrate a willingness to go with whichever interpretations (broad or wide) are most protective of the fundamental rights afforded by the Charter of Fundamental Rights of the European Union (in this case, Article 7, respect for private life, and Article 8, protection of personal data), and so would likely consider the purpose of such a system suitable to classify it as ‘identification’.

If the system is *not* classified as identification and so does not automatically become high-risk, then it should be established whether it is likely to be added to Annex III according to the criteria in AI Act Chapter 3 Article 7. Item B of Paragraph 2, “the extent to which an AI system [] is likely to be used”, may be

relevant to a technology which is not necessarily identification, but can be in some circumstances (with images of a known origin) and is applied with the intent to eventually identify. Item C, “the nature and amount of the data processed”, would also apply with the system in question processing large amounts of biometric data, semi-indiscriminately (as is the case with remote systems) in regard to the subject of the images. Considering these points, it is likely that the proposed system, if not already considered high risk, would be added to the list in Annex III in any review.

An AI system is *not* high risk even if it is contained in Annex III however if it meets any of the four point given in Article 6 Paragraph 3, of which the relevant point to this analysis is Item D, “the AI system is intended to perform a preparatory task to an assessment relevant for the purposes of the use cases listed in Annex III”. As the proposed system is biometric search, retrieval, and organisation of images, it does not make any automated decisions and the output of it is expected to be further reviewed by human analysts, to confirm or reject any potential matches. This organisation of the data prior to a human analyst making the final decision (deciding if two images match based on hand biometric information) and controlling the outcome could constitute a “preparatory task”.

The “preparatory task” clause to define a system as low-risk has an interested interaction with one of the requirements of (specifically remote biometric) high-risk systems: that “no action or decision is taken by the deployer on the basis of the identification resulting from the system unless that identification has been separately verified and confirmed by at least two natural persons with the necessary competence, training and authority” (Article 14 Paragraph 5). If human verification of the output of a remote biometric system before action is taken is a consequential requirement of its classification as high-risk, and that complying with this requirement does not absolve it of that classification nor the need to comply with the other high-risk requirements, it is reasonable to assume that having the system be non-autonomous from the outset is not enough to prevent its classification as high-risk. In this context, “preparatory task” is likely referring to a processing step other than the system outputting predictions or likelihoods from which human analysts take action.

The AI Act has further categories for high-risk AI systems in law enforcement: systems to “assess the risk of a natural person becoming the victim of criminal offences”, “be used [] as polygraphs or similar tools”, “evaluate the reliability of evidence in the course of the investigation or prosecution”, “[assess] the risk of a natural person offending or re-offending [] or to assess personality traits and characteristics or past criminal behaviour of natural persons or groups”, and “be used [] for the profiling of natural persons”. None of these points relate directly to the proposed biometric system, and in fact no reference to biometrics is included at all,

likely because any biometric system thought to be employed for law enforcement would already be included in Paragraph 1. It could be asked whether the lack of inclusion of biometric systems that do not fit into the Paragraph 1 classifications in Paragraph 6 (law enforcement) is an oversight; one that could feasibly someday be rectified. In *Big Brother Watch* (Grand Chamber judgment) the European Court of Human Rights (ECHR) determined that indiscriminate data processing (bulk intercept), something akin to remote biometric technologies, for select purposes is not prohibited by EU law but must be done with adequate safeguards [62], similar to those provided for in the requirements for high-risk AI systems. Therefore it could be seen that a remote biometric system that doesn't necessarily perform identification would still be interpreted as high-risk and need to comply with the requirements of high-risk systems.

AI systems classified as high-risk must be compliant with the requirements set out in the AI Act Chapter 3 Section 2 (Articles 9-15), which cover the topics of risk management, data governance, technical documentation, record-keeping, transparency, human oversight, and technical quality (accuracy, robustness, and cybersecurity). Of particular relevance to the hand recognition system in question are Article 10, data and data governance, Article 13, transparency and provision of information to deployers, and Article 14, human oversight.

Per Article 10, the systems must be developed with training, validation, and testing datasets that meet certain criteria. This includes the requirement that the datasets be “sufficiently representative” and “complete in view of the intended purpose”. For biometric datasets this could be interpreted as the demographic diversity of the dataset should be accurately representative of the target population; for a remote biometric system this would be the whole (world) population. Rather than proportioning datasets strictly on world-population statistics, it may be sufficient to have an even distribution of demographic traits on a large enough scale, or to represent the demographics of the target domain based on historical or regional data. It is known that lower representation of certain demographic groups leads to higher false identification rates for that group, so this is a particularly contentious obstacle to adoption of hand recognition technologies: large and diverse datasets do not yet exist.

Per Article 13, high risk AI systems must be transparent in how they operate in a way that enables deployers to correctly operate them and interpret the output. Additionally, they should be provided with instructions for use and a record of accuracy on test datasets, including “its performance regarding specific persons or groups of persons on which the system is intended to be used”. This may be particularly relevant for biometric systems and constitute a requirement to specifically assess for

bias with certain demographic controls. The requirement for interpretable output, while intending that operators of the system should understand the mechanisms at play and what the outputs mean, could also connect to the research on uncertainty quantification in the previous chapter, which aims to provide users with a multifaceted output which represents not only the prediction from the model but also the strength of the evidence base (according to the model) for that prediction.

Finally, as previously mentioned, the systems must be designed so that they can be applied with appropriate human oversight, by “at least two [] persons with the necessary competence, training, and authority”. Per Paragraph 5 of Article 14 however, the requirement of “separate verification by at least two natural persons” is excepted for law enforcement applications where “Union or national law considers the application of this requirement to be disproportionate”. This relates to the “specificity” of some areas such as biometrics and the relative sparsity of training for those applications. For example, for a hand recognition system a qualifying person with the necessary competence, training, and authority could be someone with academic knowledge in hand anatomy - not a very common area of expertise. In these instances it may not be deemed proportional to expect two verifications when one may suffice.

As an additional note, as stated in Article 10, data and data governance, Paragraph 5 high-risk AI systems may process other special categories of personal data “to the extent that it is strictly necessary for the purpose of ensuring bias detection and correction”. This has interesting implications, as other studies have found combined biometric matching and gender prediction (biometric categorisation) to improve results and widen the search space for biometric systems [18, 1].

6.5 Requirements for the Proposed System

Analysis of the regulatory landscape around post-remote (retroactive) biometric systems reveals three key areas that need to be developed before hand recognition can be brought into use. These are a clear legal basis for application of the technology, sufficiently large and diverse datasets for training and evaluation of models, and clarity on the competency requirement for human oversight and potentially training schemes.

6.5.1 Legal Basis

Actions undertaken by public bodies such as law enforcement agencies must be expressly permitted by legislation. For law enforcement, an action must be based on

a statutory power or duty provided for in law: some sources include the Data Protection Act (DPA - Part 3 implements the Law Enforcement Directive) for data processing, the ‘Police and Criminal Evidence’ Act 1984 for acquisition of suspect samples, or the ‘Police, Crime, Sentencing and Courts’ Act 2022 for the collection of data during an investigation. The legal basis used must be clear and accessible to everybody. In addition to being supported by a legal basis, the action must also be demonstrably *necessary* to achieve the desired aim, and the action must be a *foreseeable* result of the basis.

Before continuing in the development of the technology it should be established that there is a clear legal basis for its use in UK law. Data processing for the training and development of the model and technologies would be covered by the GDPR, relying on one of the six bases provided within - the strongest option would be consent of the data subject, however an argument could perhaps be made for public interest. The remaining areas - the collection of images in the course of an investigation, the storage of images for a reference database for future investigations, (possibly) the collection of suspect reference images, and the application of a post-remote biometric technology - would be actions by a law enforcement body and need to rely on powers and duties given to law enforcement in legislation. None of these actions are unique to the proposed hand recognition technology, they are also required for the use of facial and fingerprint recognition, both of which have a long history of use. The legal basis given by agencies for their use of these technologies, and the legal challenges they have faced for them, can be used to understand what would be required in this area for hand recognition.

Retroactive (post-remote) facial recognition is relatively commonplace in law enforcement. Any time an acquired image (i.e. from CCTV footage) is used to search face image databases, that is retroactive recognition. The face image databases in question could be wanted criminals, missing persons, or others, and are assumed at this point to be lawfully acquired and held (this will be discussed later). As an example, the Metropolitan Police have provided a document outlining their use of retroactive facial recognition and the legal bases on which it stands [134]. Three points related to the acquisition and use of probe images (the image used to search) are touched on: the basis for the police to take images of suspects given by the Police and Criminal Evidence Act 1984, the basis for the police to take images into possession during an investigation given by the Police, Crime, Sentencing, and Courts Act 2022, and finally the basis for the police to process that image by extracting biometric templates for data search and retrieval is given in terms of related court judgments. They reference in particular *R. (on the application of Bridges) v Chief Constable of South Wales (2020)* and *Hellewell v Chief Constable of Derbyshire (1995)* in saying

that the courts have “recognised the right of the police to make use of a photograph of an individual”. Arguably, this is not a *clear* basis, however it seems implicitly true that the courts permit it given that the use of retrospective face recognition (in the ways it has been applied to date) has not been legally challenged. It could be seen that the use of an image this way (biometric template extraction and matching) is necessary for the pursuit of a legal law enforcement aim (as without this processing, the search would have to be undertaken by a human, which would require such time and resource as to not be feasible) and so permitted by the Data Protection Act Part 3 (Law Enforcement Directive) as a fair use of data processing to complete a law enforcement aim (identification). MPS justify the interference with the Article 8 right to private life in the Human Rights Act 1998 using this reasoning. They state that the interference with the right is negligible if no result is returned as the probe biometric template is simply deleted; when a positive result is obtained it is not negligible, but is justified as it is to pursue a legitimate aim (prevention of crime) which the public can understand, the action is necessary to pursue that aim, and it is proportional.

The use of *live* facial recognition by South Wales Police was challenged in court and subsequently prevented because, in the courts view, there was no clear legal basis for that kind of processing to take place [3]. In this situation, it was the application of the technology live and in a public place, resulting in the indiscriminate processing of the personal data of members of the public not connected with any crime, for which there was no basis. The particular judgment was that the legal basis used wasn’t clear, and therefore the the use of live facial recognition in public was not a foreseeable outcome, and that there were not sufficient safeguards to prevent arbitrary or unfair application. There are many differences between live and retroactive biometric recognition which eliminate some of the concerns (large-scale indiscriminate application) however this judgment reinforces that whatever statutory power is used to justify the processing should be thorough, clear, and have adequate limitations and safeguards in place.

Returning to the reference databases, for facial recognition, these ‘image reference libraries’ (as referred to by MPS) can be substantive or temporary. Substantive libraries consist of a large amount of images stored for a long period of time - for example, custody images [134], while temporary libraries are smaller and time-limited and have a specific purpose. These libraries in many cases store the biometric templates extracted from images, rather than the raw images themselves. For *each* library clear legal bases are given for the collection, storage, and use. There is a need to establish exactly what the reference library for the proposed hand recognition system would be - as discussed, it would contain hand biometric templates extracted

from perpetrator-recorded images depicting criminal acts (e.g. sexual abuse) collected from other investigations. The legal basis for the collection of these images is provided for, however a legal basis for the subsequent extraction of biometric templates and long-term storage would need to be determined. Based on the content of the images this would be applied to, and the fact that the biometric template would belong to someone either recording or participating in a crime (rather than someone simply in the vicinity, as is possible with CCTV footage), there could be little disagreement about proportionality. Similarly to the facial recognition justification, the information or identifications the database could lead to could meet requirements for necessity. However, despite the circumstances, the storage of biometric information about individuals who have not been *convicted* of a crime faces challenges; in *S and Marper v The United Kingdom (2008) (Grand Chamber)*, the judges decided that indefinite retention of samples from acquitted individuals was unlawful [63].

6.5.2 Databases

In determining the required size of datasets for training, the closest comparison for hand recognition is facial recognition. Fingerprints, another common biometric used in forensics and law enforcement, are matched using minutiae (ridge endings and bifurcations). Minutiae extraction is often done with rules-based algorithms (not learned functions) and so do not require large training datasets; even in the case of deep-learned minutiae extraction, this requires far less data than deep feature extraction, and the datasets that exist for this contain only a few hundred samples.

Biometric datasets are not measured purely by the total number of images in them, but by the number of identities (dataset depth) and the number of images per identity (dataset width), though the number of images per identity may have a large range. Deep biometric datasets with many different identities ensure that the models learn to better discriminate features that appear similar or reoccur across much of the population, such as, for example, certain nose shapes. Wide biometric datasets with lots of images per identity allow the models to learn to consistently extract the features under different image conditions, such as quality variation, different angles, occlusion, or ageing; this is most useful for unconstrained biometric systems where the conditions in the images can't be controlled. Ideally, biometric datasets are both deep and wide, but depth is more important for training as varying image conditions can be artificially introduced using image augmentation.

Popular face image datasets used to develop most commonly used facial recognition algorithms contain hundreds of thousands to millions of images: Facebook's DeepFace algorithm was trained on four million images of "more than 4000 Face-

book users” [156]; Google’s FaceNet was trained on 260 million images of an unknown number of identities [142]; the MegaFace dataset [85], which contains 4.75 million face images of 672,057 identities, has reportedly been used by law enforcement agencies in addition to large commercial companies [103]; most face recognition modules provided with common machine learning Python libraries are pre-trained on VGG-Face or VGG-Face2, the latter having 3.31 million images of 9131 identities. The datasets these models are evaluated on are considerably smaller: Labelled Faces in the Wild (LFW) has 12,233 images of 5749 identities, though only 1680 of those identities have two or more images [60]; Youtube Face (YTF) has frames from 3425 videos of 1595 identities.

The desire for such large datasets for facial recognition is because more data drastically increases recognition performance. In one example in which this was directly studied, Google’s FaceNet paper found an increase from 76.3% validation accuracy with a 2.6 million image training dataset to 86.2% with their final 260 million, though 85.1% was achieved with only 26 million. It is logical that similar amounts of data would be required for effective learning of hand features. The EU AI Act however does not place any requirements on dataset size, only that it be “sufficiently representative” and “complete in view of the intended purpose”. Arguably a biometric dataset could never be “complete” for its intended purpose as it would always benefit to include more identities, and where the line is drawn for “sufficiently” representative remains unclear; representation in biometric datasets is not just be along categorical demographic lines, but must allow the models to learn and discriminate the wide range of physical features that appear in various combinations across the whole population. There have not yet been any rulings about dataset sizes in other comparative biometrics (face, fingerprint, or other) to draw from when determining how large the hand recognition datasets should be, but a good approach may be to collect as much data as possible with the aim to match the sizes of face image datasets (hundreds of thousands of images, thousands of identities), understanding the practical barriers that make this more difficult (they cannot be as easily acquired from the internet as they exist in far lower quantities and with less connection to user identity, comparing to social media profile pictures).

Beyond simply collecting as much data as possible, further studies should be done to establish the link between facial recognition dataset size and hand recognition dataset size in terms of performance (it could potentially take more hand images to achieve a similar performance to face recognition due to the larger range of possible hand poses, but this depends on the approach taken). Once this factor is determined, then the studies which show how facial image dataset size affects performance and establish a point from which only marginal improvements are seen

(such as Google’s FaceNet only achieving +1.1% accuracy from 26 to 260 million images) can be used to infer that value for hand recognition datasets. This may not be completely accurate, but would provide a statistical basis on which regulators could reason about thresholds for ‘sufficiency’.

6.5.3 Competency and Training

Chapter 3, Article 14 of the EU AI Act lays out the requirements for human oversight of high-risk AI systems. Paragraph 5 of that chapter states that, for biometric identification systems, no decisions should be made based on the system’s identification without being “confirmed by at least two natural persons with the necessary competence, training and authority”. What constitutes the necessary competence or authority, or what the appropriate training should look like, will differ from (biometric) trait to trait, and so are left undefined in the legislation.

As examples, fingerprints and facial recognition are applied with very different levels of required authority. Fingerprints features are processed and fingerprint matches are verified by specially-trained fingerprint officers who have received training to analyse the minutiae of skin friction ridges (ridge endings, crosses, bifurcations etc.). A fingerprint officer should be able to assess if two fingerprint samples match without any technological input, and identify if automated technology has made a mistake. By contrast, officers involved in the deployment of facial recognition technology receive training in data protection, in how the technology functions, and how it should be used, but do not undergo extra training in how to match images of faces. Facial recognition is something innate in most people, such that the ability of any lay person to compare face images would be largely on par with someone trained to do it. People are not innately able to match and compare other biometric traits like fingerprints and hands because there are far fewer circumstances in life to learn to do so; thus, to verify the outcome of matching any non-facial biometric trait, a person must receive specific training in that biometric trait and be able to demonstrate a high accuracy.

It has been demonstrated that expert knowledge in another biometric field does not increase matching accuracy with hands compared to the scores of non-expert lay persons. Additionally, humans (using a checklist of physical features to compare images) and algorithms made different types of mistakes, with human mistakes being 66% false positives and algorithm mistakes being 87% false negatives (though this would depend on the sensitivity of the algorithm and thresholds used) [148]. This indicates that even an in-depth (checklist-based) examination by a human, never mind a cursory check, would not be sufficient oversight. The person verifying al-

gorithm responses before decisions are made should have a much more substantial understanding of hand biometrics and how to compare them.

For giving evidence of a biometric match, experts in anatomy report on the specific physical or behavioural qualities found in both samples that lead to a high likelihood, due to the distribution of the features in the population, that the two samples belong to the same person. Expert anatomic knowledge in a biometric trait as would be afforded by an anthropology qualification or similar may be excessive for verifying a routine database match during an investigation. For hands, perhaps a reasonable level of competence can be achieved through short training on the physical features that can confirm or exclude a match - the skin creases of the palm and knuckles, the dorsal veins, any birthmarks, scars, or tattoos, or digit proportions. In order to enable hand recognition in practice at a wider scale, this training would need to be finalised and approved by those with expert knowledge, then provided and certified in some way. However, it is likely that hand recognition is a niche enough scenario that the number of people needed to provide this service on a national scale would be low, and so the costs associated with this may also be low.

6.6 Conclusion

In order to progress hand recognition technology to a state where it can be used by law enforcement in investigations, it should first be understood how exactly the law applies and what would be required of it. To achieve this, the legislation for data processing and AI technology was looked at.

Several grey areas in the legal definitions of biometric data and systems were identified and worked through. Interpretations were chosen which were most protective of the Article 7 and Article 8 rights to privacy and data protection afforded in the Charter of Fundamental Rights of the European Union. This means interpretations that require more restricted, transparent, and secure use of the technology; aligning with these requirements now despite ambiguity is a form of risk-mitigation against future adverse court judgments or changes to legislation. In doing this, it was determined that the hand recognition system in question is a post-remote biometric identification system, making it a high-risk technology.

Because of this classification, it is bound by several additional requirements that would not apply to more trivial technologies. Most of these apply at the time of use and are relatively simple to implement, such as transparency and logging of every use and result. Two requirements were identified as important considerations for the development phase as they would likely need significant investment: the creation of a large hand image dataset for training the algorithms that meets minimum standards

of representation, and perhaps another (or subset of the same dataset) that could be used to properly evaluate the algorithms to demonstrate that they meet minimum performance standards; and the creation of training programmes and standards of competency to ensure that the human analysts verifying the matches made by the system have the appropriate expertise, i.e. that they themselves are capable of hand biometric matching. The specifics of these requirements are not yet established, but when implemented would need to be based on solid foundations. The recommendation of this chapter is that a dataset as large as possible should be constructed, and its sufficiency should be evaluated using statistical tests on face image datasets (as the most comparable biometric with the most abundant data). Additionally, in-depth training into hand recognition would be needed for a person to demonstrate recognition performance on hands that could meet competency standards, particularly to reduce the incidence of human false-positives. While this would require some up-front cost, the limited scope of the technology means that it is likely few people would be needed to fill such a role.

Finally, the legal basis for the processing of data and application of hand recognition models was considered. Hand recognition, as a pose-remote biometric technology, is very proximal to the retroactive facial recognition technology taking place in most police forces today, and so can utilise the same legal bases (for the acquisition of images, extraction and comparison of biometric templates, and use to search). A significant point of difference lies in the data that would be used as the reference database (or gallery or library), which would likely consist of biometric templates extracted from abuse imagery collected in prior investigations. Questions about the proportionality of the retention of this kind of data, or the necessity of it to tackle this type of crime, could be argued, but the legal basis would need to be firmly established with reference to legislation by any organisation intending to form the datasets and apply hand recognition.

Chapter 7

Conclusion

When it comes to perpetrator-recorded criminal imagery there is vast potential for uncommon biometric technologies to improve investigative efficiency, reduce human burden in viewing harmful material, and provide intelligence that would otherwise be missed through potential biometric links between cases and suspects. Alongside others like ear or vein analysis, hand recognition is considered particularly promising for perpetrator-recorded material because they appear commonly where other identifying information (face or voice) are hidden. In addition to this, sub-modalities of hands, primarily fingerprints, palm prints, and knuckle prints, have a substantial history of use for security and user authentication purposes in the cyber security field; there is even a solid base of research into the application of these recognition systems ‘unconstrained’ - placing little restriction on the system user in terms of hand pose or position-in-frame, or in illumination and background conditions - from which there are many transferable methods for building truly unconstrained (no restrictions, even on image quality) systems for forensic application. The literature review in Chapter 2 covered the most prominent, well-performing, and generalisable (transferable to the unconstrained domain) methods for hand-based recognition, which were primarily categorised as palm or knuckle recognition, with some using fingernails, vein patterns, geometric measurements, fingerprints (at a distance), or skin markings. Notably, some studies applied deep feature extraction to whole-hand images, achieving very high recognition performance on test datasets (98%+). The most pressing research gap identified in this literature review was a lack of unconstrained hand image datasets, which called into question the validity of the results (relative recognition performance of different hand recognition methods) of papers *intended* for application in an unconstrained forensic context, as it assumes that better performance on the constrained test datasets would equal better performance

on unconstrained images. To address this gap, a new hand image dataset called VarHands was collected, containing a large number of images of each hand in varying poses, illuminations, and backgrounds. The research in this thesis then focused on the development of robust and condition-invariant hand recognition methods and evaluation on the unconstrained dataset.

7.1 Connection of Studies

The studies in this thesis were formulated sequentially and strongly based on the outputs of the prior studies. The literature review of Chapter 2 was performed with the aim of collating and analysing current hand recognition research in the context of unconstrained or forensic application, something which had not yet been done by another review. Based on the findings from this, the objective to collect a new hand image dataset with varying pose and illumination was devised and implemented in Chapter 3, to be able to comparatively evaluate the identified methods. Additionally, a range of approaches to hand image labelling were observed in the literature, and it was determined that identifying which method was most optimal should take place alongside the collection of the dataset to inform how the collected dataset should be labelled, forming Research Question 1.

Following the collection of the dataset, Research Questions 2, 3, and 4, about the relative performance of different overarching hand recognition approaches on unconstrained data and the limiting factors of that data, were considered pressing research gaps that could now be addressed using the new dataset. Chapter 4 presents in-depth analysis of these methods on the new dataset simultaneously, allowing the evaluation of each factor (algorithm, hand pose, illumination) in the contexts of each other. While this study was able to provide some solution to illumination-invariance, quality loss or complete loss of the target trait due to pose variation remained a significant factor in poor performance; this resulted in a further literature review and identification of uncertainty quantification as a potential way to handle these low-quality samples in an unconstrained domain (where samples cannot be re-captured or simply excluded from analysis), forming Research Question 5. While uncertainty quantification has uses in and of itself for better output communication in critical systems (such as biometrics in law enforcement), since it was being applied in a scenario with multi-biometric fusion, it was conceived that an incorporation of uncertainty into multi-biometric fusion as a weighting mechanism could improve recognition accuracy, resulting in the formation of Research Question 6. Research Questions 5 and 6 were then answered sequentially in Chapter 5, the third study of the thesis.

Taking place in tandem with the other studies, the review in Chapter 6 addresses Research Question 7 which was in fact the first research question defined early in the project. The completion of this review was left until last as it was intended that the knowledge gained over the course of the project and the specific findings of the other studies would feed into the analysis and recommendations.

7.2 Research Questions

Here the research questions are re-visited and the findings of this thesis in relation to each one are discussed.

RQ1 How should hand images be categorised and labelled for optimal training and recognition performance?

This thesis presents the first empirical evaluation of different labelling strategies for hand biometrics, testing by-subject (person), by-hand (left or right), and by-view (dorsal or palmar) identity assignment. Research in the literature primarily used a by-hand and by-view labelling strategy (such that each subject has four hand ‘identities’), but some used only hand-based (not differentiating by view) or subject-based (not differentiating by view *or* hand) labelling. The results in study 1, Chapter 3 of this thesis show that, separately, by-hand and by-view labelling achieve quicker convergence in training and higher ranked-retrieval performance (measured in rank-1 accuracy and mean average-precision) than by-subject labelling, and ultimately the combination of both is best. The model trained on by-hand and by-view labelling converged in approximately one quarter of the epochs for the other models to converge, displayed high training stability (low validation noise), and minimal overfitting (had the smallest gap between train and test losses). It also achieved the best accuracy to rank-20 and best mAP on a ranked-retrieval task. This is strong evidence that the by-hand and by-view labelling strategy is optimal, however, discussion is given to how this labelling could be problematic on unconstrained images where some hands are not simply categorisable as either dorsal or palmar, true left or right handedness within the image may not be known (the image could be horizontally flipped), and some dorsal hand features can appear in otherwise palmar-view images (if the subject were to make a fist). These results and considerations are explored in more detail in Chapter 3.

RQ2 Which whole-hand or regional approaches to hand recognition

perform best on unconstrained data?

This thesis evaluated a baseline whole-hand recognition model (ResNet-50), a state-of-the-art whole-hand recognition model (MBA-net), a palm recognition model (in this case, the ‘palm’ region is referring to both the palm on palmar images and the dorsum on dorsal images), a knuckle recognition model, and the combination of the palm and knuckle models on the unconstrained VarHands dataset (using distance fusion for the knuckle and palm + knuckle models), which is the first evaluation of these methods on unconstrained data. For the palm and knuckle models, novel region-extraction and alignment methods were introduced that worked with the varying image conditions in the unconstrained images, relying only on the detected key points (joints) of the hand skeleton. Because there is a lack of recognition methods for palm and knuckles in the literature that could be directly applied to unconstrained images (without being heavily modified in their pre-processing, which is often what differentiates them from other works), this study serves as a comparison of the common overarching *approaches* to hand recognition rather than a comparison of specific algorithms. The results demonstrate that knuckle recognition performed best (measured in rank-1 accuracy, accuracy to rank-100, and mean average-precision) on the constrained 11k Hands dataset (with palm + knuckle close behind), aligning with expectations from the literature. On the unconstrained VarHands dataset the knuckle and palm + knuckle models are notably separated in rank-1 accuracy and mAP from the others, with palm + knuckle coming out on top. The lack of a boost when combining palms with the knuckle model on 11k Hands is theorised to be because of a quirk of that dataset: the hand images often start half-way through the palm, providing less opportunity for the learning of palm features and utilisation of them at test-time. This might also explain why palm models significantly underperformed, achieving lower scores than even the baseline whole-hand model. Interestingly, the results show that palm models perform better on dorsal images in both VarHands and VarHands-Easy; this cannot be solely explained by obstruction of the palm by fingers (this does not occur in VarHands-Easy, though the separation of dorsal and palmar scores is larger on VarHands). These results are explained in more detail in Chapter 4.

RQ3 To what degree does change in hand pose impact recognition performance?

This thesis has performed the first evaluation of the impact of hand pose in recognition accuracy by comparing the performance of recognition models on a

pose-variable dataset (VarHands) and a non-pose variable subset of that dataset (VarHands-Easy). The VarHands-Easy dataset was manually selected as a subset of the VarHands datasets, as images that do not contain curling of the fingers and are close to directly palmar or dorsal views (minimal hand rotation from the camera), as described in Chapter 3. By contrast, the whole VarHands dataset contains many images in which the hand is rotated (losing some visibility of the palm or dorsum) and the fingers are curled so that the palm may be obstructed and only some (or no) knuckles are visible. The results show a drop in performance of approximately 2-10% rank-1 accuracy and 0.8-0.12 mean average-precision from the pose-controlled dataset to the pose-variable dataset, across all models. When illumination is controlled for (as explained in the next section), this drop becomes approximately -20% rank-1 accuracy and -0.3 mAP (for the knuckle model). The drop in performance is lower for the sub-biometric models (palm and knuckle) than it is for the whole-hand models, suggesting that the use of sub-biometrics over whole-hand images can help mitigate some of the effects of pose variation (making the models more pose-invariant); this was expected as there would be less variability in the input images for sub-biometric models than for whole-hand models (e.g., the knuckles should appear the same whether the fingers are spread or together). These results and discussion of them are in Chapter 4.

RQ4 To what degree does change in hand illumination impact recognition performance?

This thesis presents the first evaluation of how illumination change impacts the recognition performance of hand biometric models, by comparing models in a combined-illumination ranked-retrieval task (each label in the dataset is split 50-50 into probe and gallery sets, balanced across illuminations) *and* in a cross-illumination task (four tests are run with each of the four illumination categories used as a probe set and the rest as gallery sets). A method of pre-processing for illumination-invariance (removal of shadows and normalisation of pixel colour and intensity) and enhancement of the skin textural information (palm and knuckle lines) was proposed and compared against the use of the original colour images. The results show rank-1 accuracy and mean average-precision were much lower for the non-illumination-normalised models when testing cross-illumination (e.g. 11% rank-1 accuracy for the knuckle model compared to 93% in combined-illumination), indicating that the majority of correctly matched high-rank images in the combined-illumination test are images from the same illumination category, and images from other illumination categories are being falsely placed lower in the ranking (having

a higher feature-distance), indicating an inability of the models to project them to the same embedding in feature space. Additionally, the results show a large improvement to mean-average precision in combined-illumination ranked-retrieval on both VarHands and VarHands-Easy when using the illumination-normalised images. For the cross-illumination test, use of illumination-normalisation achieved triple the average mAP on VarHands and six-times the mAP on VarHands-Easy. This study demonstrated that illumination is a significant limiting factor in unconstrained hand recognition performance (when training on single-illumination images), but that the proposed method of illumination-normalisation in pre-processing can mitigate this. However, even with the significant gains made by the illumination-normalisation method, the performance in the cross-illumination test remains lower than in the intra-illumination tests, indicating that many images of differing illuminations are still not being correctly matched. These results and a further discussion on this are presented in Chapter 4.

RQ5 What are the behaviours of uncertainty quantifying methods when applied to hand biometric images of increasing quality loss?

This thesis documents the first exploration of uncertainty quantification in any type of hand images (at the time of writing), specifically, knuckle images extracted from unconstrained hand images. Several popular uncertainty quantification methods (Monte-Carlo dropout, deep ensembles, and Bayesian triplet loss) and two anomaly detection methods (GLCM divergence and auto-encoder mean squared error) were implemented and trained on the constrained hand image data (which still features quality loss to blur and sensor noise) and evaluated qualitatively. A test that introduced artificial levels of quality loss (blur, noise, occlusion, zoom-in, and zoom-out) indicated that different quantification methods were able to capture different uncertainties, and all methods failed to produce appropriate estimates for at least one type. When visualising the estimated uncertainty within the embedding space (where low quality images such as completely blurry images were all projected together), the anomaly detection methods like AEMSE showed high correlation between embedding area and image type (high quality knuckles with low scores, and low quality knuckles with high scores); uncertainty-quantification methods like MCD on the other hand showed little correlation between estimated uncertainty and embedding space. Overall, the methods showed good ability to capture some uncertainties like blur and noise, but not the more complex ones like mis-detected knuckles (e.g. a sharp image showing something other than a knuckle, such as a fingernail). For the uncertainty quantification methods in particular (over the anomaly detection

methods) these uncertainties were poorly-spread (low range) and inconsistent. This is most likely explained by the fact that the training dataset did not have enough images or enough variation within the images for the models to accurately learn the posterior distributions they would see at inference on the evaluation dataset. This is discussed in more detail in Chapter 5.

RQ6 What impact can these quantified uncertainties have on recognition performance when incorporated into multi-biometric fusion?

This thesis involved the proposal of a novel method for probabilistic multi-biometric fusion incorporating quantified uncertainty. The method utilises Bayesian inference to weight and combine the distances between each pair of sub-images to produce an overall distance, treating the uncertainty value (maximum uncertainty of the two embeddings) as a variance. Each sub-image distance then has weight proportional to its contribution to the total image uncertainty, such that the fused distance will be closer to the distances with lower uncertainties. Despite being evaluated with the uncertainties produced in the answering of the previous research question, which were demonstrably not well-calibrated, the results showed a consistent small improvement to ranked-retrieval performance across nearly all models (except GLCM and MCD on the unconstrained datasets). Because the performance differences are small the significance of the findings are limited, however, it is interesting to see it consistently and suggests that potentially a bigger difference could be seen with better calibrated uncertainty estimates, meaning Bayesian fusion has promising potential to improve recognition accuracy in multi-biometric fusion. The implications of these findings are explored in more detail in Chapter 5.

RQ7 What steps are needed for compliance with the relevant regulations before hand recognition technologies can be used for law enforcement purposes?

This thesis provides the first review of the legal and regulatory landscape around hand recognition as an investigative technology to be used by law enforcement (not yet a reality). The legal categorisation of the hand recognition system in question (a retrospective biometric recognition system, as such, a high-risk AI technology) was interpreted in the face of several grey areas in the legal definitions and scopes. The scenarios of both in-house and external production of the algorithms were considered to determine that processing of the training and evaluation data is within the scope of the GDPR, and inference data under the scope of the Law Enforcement Directive.

As a high-risk biometric technology, three restrictions of consequence were identified that must be addressed on the path to implementation of hand recognition systems: establishment of an accessible and unambiguous legal basis for its use (proximal to the legal bases used for automatic retrospective facial recognition technologies); construction of large and varied hand image datasets based on statistically-backed ‘sufficiency’ (possibly derived from face image datasets); and the creation of training schemes on hand recognition for human analysts that will be verifying the outputs of the automated systems. These are discussed in more detail in Chapter 6.

7.3 Contributions and Limitations

This thesis makes several contributions towards the development of unconstrained hand biometrics for forensic application in both methodology and regulatory understanding.

The first is the establishment of the optimal labelling strategy. In the first systematic comparison of different hand recognition labelling strategies, this thesis demonstrates that a combined by-hand and by-view labelling strategy yields faster convergence, better training stability, and superior recognition performance. This finding is in contravention of the approach taken in some other hand recognition studies and provides solid evidence for the continued use of this labelling in future studies.

The second is in the quantification of the impact of hand pose variation and illumination on recognition performance in unconstrained images (when doing cross-domain application of models trained on constrained images). The results showed that illumination change had a significant impact on recognition performance while pose-variation had a slightly smaller impact. In this study, whole-hand and regional hand recognition approaches were compared for the first time on unconstrained images, demonstrating that regional approaches (palm and knuckle recognition) were more pose-invariant and so achieved higher recognition accuracy. These results show why research priorities should be shifted away from small performance gains on constrained data. In addition to this, a novel and effective illumination-normalisation and texture enhancement method was introduced for knuckle images, and the experiments strongly support that it can result in large performance gains on unconstrained images. These insights will shape the future of hand recognition studies as it is better understood which image conditions recognition methods must handle, and the methods proposed exist as baselines for invariant feature extraction.

Third, this thesis introduces uncertainty quantification into the domain of hand recognition, demonstrating how uncertainty estimates can be incorporated into critical biometric systems for better interpretability of outputs from low quality samples,

and incorporated into multi-biometric fusion to weight towards higher-quality samples and boost recognition performance. This contribution opens two new lines of research at the intersections of uncertainty quantification, trustworthy biometrics (ethical AI), and biometric fusion.

Finally, this thesis provides the first dedicated analysis of the legal and regulatory challenges facing hand recognition technologies in law enforcement. The identification of three essential areas for development - establishment of the legal basis, construction of large and ‘sufficient’ hand image datasets, and setting of competency standards for human oversight - can push future research in the directions that will be most beneficial to creating compliant hand recognition technologies.

While these contributions are valuable first steps and advances in the domain of unconstrained hand biometrics, they are limited by several factors. The first and most significant is the small size of the datasets, both the constrained training data and the unconstrained evaluation dataset. With only 40 hands from 20 participants in the evaluation dataset, it is likely that misidentification rates would be far lower in this dataset than in the population, and so evaluation is useful for *relative* comparison of methods but not to infer true accuracies. The small size and lack of variation in the training dataset also placed strong limitations on the development of effective feature extractors and uncertainty models. The comparative results then must be prefaced *with model training on current data levels*; it is not logical to infer, for instance, that just because whole-hand recognition underperformed in these studies that it wouldn’t be able to outperform regional recognition if there was significantly more data on which to train whole-hand models (though this would also mean more data for the sub-biometric models). For uncertainty quantification, the results were limited by the poor calibration of the uncertainty estimates. Poor calibration was expected from the amount of data available; at this current time, it is unknowable whether more data would result in better calibration of uncertainty estimates (in theory it would) and if using these estimates in multi-biometric fusion would have a more significant impact on performance than what was observed. Finally, the legal analysis would require re-examination in the future as it is based on current EU legal frameworks, and with shifting political attitudes it is unknown whether these will remain consistent or apply in the UK in the future.

7.4 Future Work

There are numerous avenues for future research identified throughout this thesis. Many are contingent on the creation of larger hand image datasets for development and testing, however, some could continue directly from the presented work and

were only not included due to time and scope constraints. Some of these areas that won't necessarily require more data are improvement of region detection and alignment, illumination-normalisation and texture-enhancement, and general modelling methods.

There is lots of room for improvement of boundary detection (sizing of the box around the knuckle centre) of knuckle images, where mistakes can result in either very zoomed-in (missing critical information) or very zoomed-out (containing background or information from neighbouring fingers) images. Using a depth estimate doesn't take into account differing hand sizes, and the colour-based method proposed in this study occasionally failed when fingers were in contact with each other. A regression-based localisation of boundary points would be able to achieve much better accuracy and consistency, but would require first labelling a substantial amount of hand images, or roughly-localised knuckle images, with these boundary points. Consistent sizing and alignment of the knuckle images should result in higher matching accuracies as the feature vectors, as learned from a training dataset of constrained and well-aligned images, are not scale-invariant. An alternative path would be to use augmentation of the knuckle images to train a model on knuckle images of various scales to produce scale-invariant vectors, but this would require longer training with more data (different background textures) and still wouldn't deal with cases where poor sizing results in knuckle information being missed.

It is also highly likely that the illumination-normalisation method proposed could be further developed, or another method used, to eliminate more deep shadows and highlights, and remove all effects of directional lighting. Conditional generational adversarial networks (CGANs) have shown promise in this area [87] but require a lot of data to train. More data (with set directional lighting) would also be needed to evaluate the effectiveness of any illumination-normalisation method.

Improvements could also be made in the modelling and feature extraction methods, making use of recent and future advances in network architectures or loss functions, such as perhaps using vision transformers [33]. While this project focused only on using contrastive losses which are generally better for large cohort open-set recognition, it has been noted that higher accuracies can be reached when using a combination of contrastive and classification losses [152], so this would be worth exploring in future studies.

Further exploration of uncertainty quantification in hand images would first require more data to be collected, as is described in the next section. Continuation of the development of probabilistic multi-biometric fusion however need not be constrained to hands and so may be worth exploring with other multi-biometric datasets, should they also contain varied and low-quality samples to properly learn posterior

distributions. It is possible that augmentation or synthetic data generation may allow the creation of more varied training datasets for this purpose, however it is very important that the training dataset is very similar and representative of the inference distribution or the learned distributions will not be accurate.

Finally, one line of work strongly considered in this project was the potential for well-quantified epistemic uncertainty to reflect demographic biases in biometric training datasets, offering a way to observe and handle these biases at test-time. This study was prevented by both a lack of measurable demographics and significant diversity in the datasets accessed, and by the lack of good calibration of uncertainty estimates. Moving away from hands, facial recognition may be a suitable domain to explore this proposal. Facial recognition datasets are large enough and contain enough sample diversity (in pose, expression, personal styling, etc.) to potentially learn effective uncertainties, and many are demographically-labelled for the purpose of studying algorithm biases. An experiment could be set up to train models on certain artificially-biased datasets (e.g. 80% and 20% female) then the uncertainties measured for each of those demographics at inference to determine if there is correlation between demographic representation and uncertainty. While this would only be *measurable* across demographic categories (e.g. race, sex) or ranges (e.g. age), algorithmic biases apply on a feature level, where certain sample features (e.g. a certain nose shape) being under-represented in a training dataset can result in poor distinction of those features and so false positive matching at inference time. Uncertainty estimated from modelling of the training dataset distributions should, in theory, be able to capture these granular biases.

7.5 More Data as a Priority

Many of the limitations in this study, and roadblocks to the future development of unconstrained hand recognition technologies, come down to the lack of unconstrained hand image data. Current hand image datasets are highly constrained, as they exist primarily for the development and evaluation of controlled biometric security systems using specific regions of the hand (e.g. palm, knuckle, fingerprints). These datasets are not sufficient for learning consistent hand features across a variety of hand representations (e.g. in different illuminations, in different poses, with occlusions, noise, and blurring), and they are certainly not sufficient to evaluate if proposed recognition methods and algorithms are invariant to these factors (with augmentations only being an approximation, unlikely to contain the full diversity of real unconstrained hand images). In addition to the lack of variety in the hand images (which must be represented *within* each subject, not just across all subjects), the

datasets are relatively small by number of participants compared to other domains such as face recognition. Having too few subjects in a biometric training dataset will lead to poor learned separation of classes, and in a testing dataset will result in falsely inflated recognition accuracies as there are too few close negative samples. Having too few subjects also prevents the detection of many possible biases. The lack of ability to identify true accuracy, conditions of failure, and biases means that unconstrained hand recognition systems cannot be compliant with the regulations on high-risk technologies, as discussed in Chapter 6.

It is important that future hand image datasets collected for the purpose of developing unconstrained recognition systems address the shortcomings present in the current datasets. While training datasets should contain variation equivalent to what is seen at inference (especially so when modelling uncertainties), it is more critical to focus on the expansion of evaluation datasets. Augmentation and data generation techniques can be used to build up training datasets in the interim.

An effective evaluation dataset should allow the identification of failure points. It should make it possible to analyse the relative performance on different demographic groups to identify biases, which would require a very large and diverse cohort of subjects with demographic information collected and stored alongside the data. It should also ensure total coverage of all possible hand poses, and a very diverse range of illuminations including direction lighting and different hues. In contradiction with the dataset collection protocol presented in Chapter 3, an efficient way to achieve this could be to separate the collection of illumination-variant and pose-variant images, and have them collected as photographs guided by a checklist rather than the completely unconstrained hand videos as done in this project (though perhaps it would be suitable to do both, as the videos would display a more ‘natural’ range of poses). For illumination, the images could be controlled to some degree (flat hand) and focus on showing lighting of different hues from set directions around the hand. For pose, the images could remain in clear illumination but show set intervals for the range of motion of each joint, much like how knuckle datasets with flexation are collected [29], but across all combinations of the finger knuckles and wrist joints. In addition to this, it would be highly valuable to be able to collect this data from the same participants over a long period of time (>5 years) so as to capture the effects of ageing.

Bibliography

- [1] Mahmoud Afifi. “11K Hands: Gender recognition and biometric identification using a large dataset of hand images”. In: *Multimedia Tools and Applications* 78 (2019), pp. 20835–20854.
- [2] Mona Alghamdi, Plamen Angelov, and Bryan Williams. “Automated person identification framework based on fingernails and dorsal knuckle patterns”. In: *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE. 2021, pp. 01–08.
- [3] The Court of Appeal of England and Wales. *R. (on the application of Bridges) v Chief Constable of South Wales*. EWCA Civ 1058. 2020.
- [4] Abdelouahab Attia, Zahid Akhtar, and Youssef Chahir. “Feature-level fusion of major and minor dorsal finger knuckle patterns for person authentication”. In: *Signal, Image and Video Processing* 15.4 (2021), pp. 851–859.
- [5] Murat Aykut and Murat Ekinici. “Developing a contactless palmprint authentication system by introducing a novel ROI extraction method”. In: *Image and Vision Computing* 40 (2015), pp. 65–74.
- [6] GS Badrinath and Phalguni Gupta. “Palmprint based recognition system using phase-difference information”. In: *Future Generation Computer Systems* 28.1 (2012), pp. 287–305.
- [7] Nathanael L. Baisa et al. “Hand-Based Person Identification using Global and Part-Aware Deep Feature Representation Learning”. In: *2022 Eleventh International Conference on Image Processing Theory, Tools and Applications (IPTA)*. 2022, pp. 1–6. DOI: 10.1109/IPTA54936.2022.9784133.
- [8] Nathanael L. Baisa et al. “Multi-Branch with Attention Network for Hand-Based Person Recognition”. In: *2022 26th International Conference on Pattern Recognition (ICPR)*. 2022, pp. 727–732. DOI: 10.1109/ICPR56361.2022.9956555.

- [9] Xianjie Bao and Zhenhua Guo. “Extracting region of interest for palmprint by convolutional neural networks”. In: *2016 Sixth International Conference on Image Processing Theory, Tools and Applications (IPTA)*. IEEE. 2016, pp. 1–6.
- [10] Debnath Bhattacharyya et al. “Biometric authentication: A review”. In: *International Journal of u-and e-Service, Science and Technology* 2.3 (2009), pp. 13–28.
- [11] SM Black et al. “Forensic hand image comparison as an aid for paedophile investigations”. In: *Police Professional* 184 (2009), pp. 21–24.
- [12] Sue Black, Briony MacDonald-McMillan, and Xanthe Mallett. “The incidence of scarring on the dorsum of the hand”. In: *International Journal of Legal Medicine* 128 (2014), pp. 545–553.
- [13] Sue Black et al. “The incidence and position of melanocytic nevi for the purposes of forensic image comparison”. In: *International Journal of Legal Medicine* 128 (2014), pp. 535–543.
- [14] Charles Blundell et al. “Weight uncertainty in neural network”. In: *International conference on machine learning*. PMLR. 2015, pp. 1613–1622.
- [15] Roel Bouman and Tom Heskes. “Autoencoders for Anomaly Detection are Unreliable”. In: *arXiv preprint arXiv:2501.13864* (2025).
- [16] Dane Brown and Karen Bradshaw. “Deep palmprint recognition with alignment and augmentation of limited training samples”. In: *SN Computer Science* 3.1 (2022), p. 11.
- [17] Dane Brown and Karen Bradshaw. “Improved palmprint segmentation for robust identification and verification”. In: *2019 15th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*. IEEE. 2019, pp. 1–7.
- [18] Tingting Chai, Shitala Prasad, and Shenghui Wang. “Boosting palmprint identification with gender information using DeepNet”. In: *Future Generation Computer Systems* 99 (2019), pp. 41–53.
- [19] Tingting Chai, Shenghui Wang, and Dongmei Sun. “Towards contactless palm region extraction in complex environment”. In: *International Journal of Biometrics* 10.3 (2018), pp. 232–254.
- [20] Tingting Chai et al. “Contactless palmprint biometrics using DeepNet with dedicated assistant layers”. In: *The Visual Computer* 39.9 (2023), pp. 4029–4047.

- [21] Tingting Chai et al. “Contactless palmprint biometrics using DeepNet with dedicated assistant layers”. In: *The Visual Computer* 39.9 (2023), pp. 4029–4047.
- [22] Nesrine Charfi et al. “Local invariant representation for multi-instance touchless palmprint identification”. In: *2016 IEEE international conference on systems, man, and cybernetics (SMC)*. IEEE. 2016, pp. 003522–003527.
- [23] Weihua Chen et al. “Beyond triplet loss: a deep quadruplet network for person re-identification”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 403–412.
- [24] KamYuen Cheng and Ajay Kumar. “Contactless finger knuckle identification using smartphones”. In: *2012 BIOSIG-proceedings of the international conference of biometrics special interest group (BIOSIG)*. IEEE. 2012, pp. 1–6.
- [25] Kevin HM Cheng and Ajay Kumar. “Contactless biometric identification using 3D finger knuckle patterns”. In: *IEEE transactions on pattern analysis and machine intelligence* 42.8 (2019), pp. 1868–1883.
- [26] H Kipsang Choge et al. “A circle-based region-of-interest segmentation method for palmprint recognition”. In: *2009 ICCAS-SICE*. IEEE. 2009, pp. 4993–4997.
- [27] Jifeng Dai, Jianjiang Feng, and Jie Zhou. “Robust and efficient ridge-based palmprint matching”. In: *IEEE transactions on pattern analysis and machine intelligence* 34.8 (2011), pp. 1618–1632.
- [28] Andreas Damianou and Neil D Lawrence. “Deep gaussian processes”. In: *Artificial intelligence and statistics*. PMLR. 2013, pp. 207–215.
- [29] Lilly Dan et al. “The persistence of knuckle creases during finger flexion for the identification of perpetrators from digital images of their hands”. In: *Forensic Science International* 348 (2023), p. 111603.
- [30] Erik Daxberger et al. “Laplace redux-effortless bayesian deep learning”. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 20089–20103.
- [31] Armen Der Kiureghian and Ove Ditlevsen. “Aleatory or epistemic? Does it matter?” In: *Structural safety* 31.2 (2009), pp. 105–112.
- [32] Lavinia Mihaela Dinca and Gerhard Petrus Hancke. “The fall of one, the rise of many: a survey on multi-biometric fusion methods”. In: *IEEE Access* 5 (2017), pp. 6247–6289.

- [33] Alexey Dosovitskiy et al. “An image is worth 16x16 words: Transformers for image recognition at scale”. In: *arXiv preprint arXiv:2010.11929* (2020).
- [34] Edward R Dougherty. “An introduction to morphological image processing”. In: *SPIE*. Optical Engineering Press. 1992.
- [35] Nicolae Duta, Anil K Jain, and Kanti V Mardia. “Matching of palmprints”. In: *Pattern Recognition Letters* 23.4 (2002), pp. 477–485.
- [36] Helin Dutagacı, Bülent Sankur, and Erdem Yörük. “Comparative analysis of global hand appearance-based person recognition”. In: *Journal of electronic imaging* 17.1 (2008), pp. 011018–011018.
- [37] Marcos Faundez-Zanuy, Jiri Mekyska, and Xavier Font-Aragones. “A New Hand Image Database Simultaneously Acquired in Visible, Near-Infrared and Thermal Spectrums”. In: *Cognitive Computation* 6 (Oct. 2013). DOI: 10.1007/s12559-013-9230-3.
- [38] Lunke Fei et al. “Feature extraction methods for palmprint recognition: A survey and evaluation”. In: *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 49.2 (2018), pp. 346–363.
- [39] Miguel A Ferrer, Carlos M Travieso, and Jesus B Alonso. “Using hand knuckle texture for biometric identification”. In: *Proceedings 39th annual 2005 international carnahan conference on security technology*. IEEE. 2005, pp. 74–78.
- [40] Loic Le Folgoc et al. “Is MC dropout bayesian?” In: *arXiv preprint arXiv:2110.04286* (2021).
- [41] Internet Watch Foundation. *IWF Annual Report 2021*. 2021. URL: <https://www.iwf.org.uk/about-us/who-we-are/annual-report-2021/> (visited on 09/26/2025).
- [42] Internet Watch Foundation. *IWF Annual Report 2024*. 2024. URL: <https://www.iwf.org.uk/annual-data-insights-report-2024/> (visited on 09/26/2025).
- [43] The Police Foundation. *Turning the Tide Against Online Child Sexual Abuse*. 2022. URL: https://www.police-foundation.org.uk/wp-content/uploads/2022/07/turning_the_tide_FINAL-.pdf (visited on 09/26/2025).
- [44] Yarin Gal and Zoubin Ghahramani. “Dropout as a bayesian approximation: Representing model uncertainty in deep learning”. In: *international conference on machine learning*. PMLR. 2016, pp. 1050–1059.

- [45] Guangwei Gao et al. “Integration of multiple orientation and texture information for finger-knuckle-print verification”. In: *Neurocomputing* 135 (2014), pp. 180–191.
- [46] Strati Garg, Amioy Kumar, and Madasu Hanmandlu. “Biometric authentication using finger nail surface”. In: *2012 12th International Conference on Intelligent Systems Design and Applications (ISDA)*. IEEE. 2012, pp. 497–502.
- [47] Eduardo SL Gastal and Manuel M Oliveira. “Domain transform for edge-aware image and video processing”. In: *ACM SIGGRAPH 2011 papers*. 2011, pp. 1–12.
- [48] Chuanxing Geng, Sheng-jun Huang, and Songcan Chen. “Recent advances in open set recognition: A survey”. In: *IEEE transactions on pattern analysis and machine intelligence* 43.10 (2020), pp. 3614–3631.
- [49] Angelo Genovese et al. “PalmNet: Gabor-PCA convolutional networks for touchless palmprint recognition”. In: *IEEE Transactions on Information Forensics and Security* 14.12 (2019), pp. 3160–3174.
- [50] Angelo Genovese et al. “Touchless palmprint and finger texture recognition: A deep learning fusion approach”. In: *2019 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA)*. IEEE. 2019, pp. 1–6.
- [51] Simon J Godsill. “On the relationship between Markov chain Monte Carlo methods for model uncertainty”. In: *Journal of computational and graphical statistics* 10.2 (2001), pp. 230–248.
- [52] Zhen Guo et al. “A survey on uncertainty reasoning and quantification in belief theory and its application to deep learning”. In: *Information Fusion* 101 (2024), p. 101987.
- [53] Siwar Hammami et al. *FMK (Finger Major Knuckle)*. 2023. DOI: 10.21227/n6wc-th47. URL: <https://dx.doi.org/10.21227/n6wc-th47>.
- [54] Robert M Haralick. “Statistical and structural approaches to texture”. In: *Proceedings of the IEEE* 67.5 (1979), pp. 786–804.
- [55] Russell Hardie. “A fast image super-resolution algorithm using an adaptive Wiener filter”. In: *IEEE Transactions on Image Processing* 16.12 (2007), pp. 2953–2964.

- [56] Kaiming He et al. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.
- [57] Brian Heflin, Walter Scheirer, and Terrance E Boulton. “Detecting and classifying scars, marks, and tattoos found in the wild”. In: *2012 IEEE Fifth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*. IEEE. 2012, pp. 31–38.
- [58] Hadis Heidari and Abdollah Chalechale. “Biometric authentication using a deep learning approach based on different level fusion of finger knuckle print and fingernail”. In: *Expert Systems with Applications* 191 (2022), p. 116278.
- [59] Gao Huang et al. “Densely connected convolutional networks”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 4700–4708.
- [60] Gary B Huang et al. “Labeled faces in the wild: A database for studying face recognition in unconstrained environments”. In: *Workshop on faces in Real-Life Images: detection, alignment, and recognition*. 2008.
- [61] Eyke Hüllermeier and Willem Waegeman. “Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods”. In: *Machine learning* 110.3 (2021), pp. 457–506.
- [62] European Court of Human Rights (Grand Chamber). *Case of Big Brother Watch and Others v. The United Kingdom*. Applications 58170/13, 62322/14 and 24960/15. 2021.
- [63] European Court of Human Rights (Grand Chamber). *Case of S. and Marper v. The United Kingdom*. (Applications 30562/04 and 30566/04. 2008.
- [64] Imaging and Computer Vision Lancaster. *H-Unique: In Search of Uniqueness*. 2019. URL: <https://wp.lancs.ac.uk/icvl/research/h-unique/> (visited on 09/26/2025).
- [65] Mahdiah Izadpanahkakhk et al. “Deep region of interest and feature extraction models for palmprint verification using convolutional neural networks transfer learning”. In: *Applied Sciences* 8.7 (2018), p. 1210.
- [66] Mahdiah Izadpanahkakhk et al. “Novel mobile palmprint databases for biometric authentication”. In: *International Journal of Grid and Utility Computing* 10.5 (2019), pp. 465–474.

- [67] Anil K Jain, Ruud Bolle, and Sharath Pankanti. *Biometrics: personal identification in networked society*. Vol. 479. Springer Science & Business Media, 2006.
- [68] Anil K Jain, Sarat C Dass, and Karthik Nandakumar. “Soft biometric traits for personal recognition systems”. In: *International conference on biometric authentication*. Springer. 2004, pp. 731–738.
- [69] Anil K Jain and Arun Ross. “Bridging the gap: from biometrics to forensics”. In: *Philosophical Transactions of the Royal society B: Biological sciences* 370.1674 (2015), p. 20140254.
- [70] Gaurav Jaswal, Amit Kaul, and Ravinder Nath. “Knuckle print biometrics and fusion schemes—overview, challenges, and solutions”. In: *ACM Computing Surveys (CSUR)* 49.2 (2016), pp. 1–46.
- [71] Gaurav Jaswal, Ravinder Nath, and Amit Kaul. “FKP based personal authentication using SIFT features extracted from PIP joint”. In: *2015 third international conference on image information processing (ICIIP)*. IEEE. 2015, pp. 214–219.
- [72] Gaurav Jaswal, Aditya Nigam, and Ravinder Nath. “DeepKnuckle: revealing the human identity”. In: *Multimedia Tools and Applications* 76.18 (2017), pp. 18955–18984.
- [73] Gaurav Jaswal et al. “DeepPalm-A Unified Framework for Personal Human Authentication”. In: *2018 International Conference on Signal Processing and Communications (SPCOM)*. IEEE. 2018, pp. 322–326.
- [74] Wei Jia et al. “A survey on dorsal hand vein biometrics”. In: *Pattern Recognition* 120 (2021), p. 108122. ISSN: 0031-3203. DOI: <https://doi.org/10.1016/j.patcog.2021.108122>. URL: <https://www.sciencedirect.com/science/article/pii/S0031320321003095>.
- [75] Laurent Valentin Jospin et al. “Hands-on Bayesian neural networks—A tutorial for deep learning users”. In: *IEEE Computational Intelligence Magazine* 17.2 (2022), pp. 29–48.
- [76] Court of Justice of the European Union (Grand Chamber). *Data Protection Commissioner v Facebook Ireland Limited and Maximillian Schrems*. Case C-311/18. 2020.
- [77] Court of Justice of the European Union (Grand Chamber). *Google Spain SL and Google Inc. v Agencia Española de Protección de Datos (AEPD) and Mario Costeja González*. Case C-131/12. 2014.

- [78] Court of Justice of the European Union (Grand Chamber). *La Quadrature du Net and Others v Premier ministre and Others*. Joined Cases C-511/18, C-512/18 and C-520/18. 2020.
- [79] Court of Justice of the European Union (Grand Chamber). *Maximillian Schrems v Data Protection Commissioner*. Case C-362/14. 2015.
- [80] Court of Justice of the European Union (Grand Chammbler). *Digital Rights Ireland Ltd v Minister for Communications, Marine and Natural Resources and Others*. Case C-293/12. 2014.
- [81] Karbhari V Kale et al. “Multimodal biometric system using fingernail and finger knuckle”. In: *2013 International Symposium on Computational and Business Intelligence*. IEEE. 2013, pp. 279–283.
- [82] Vivek Kanhangad, Ajay Kumar, and David Zhang. “Contactless and pose invariant biometric identification using hand surface”. In: *IEEE transactions on image processing* 20.5 (2010), pp. 1415–1424.
- [83] Pooja Kaplesh et al. “A systematic review of end-to-end framework for contactless fingerprint recognition: Techniques, challenges, and future directions”. In: *Engineering Applications of Artificial Intelligence* 150 (2025), p. 110493.
- [84] Ali Keivanmarz, Hamid Sharifzadeh, and Rachel Fleming. “Vein pattern visualisation using conditional generative adversarial networks”. In: *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE. 2020, pp. 1310–1316.
- [85] Ira Kemelmacher-Shlizerman et al. “The megaface benchmark: 1 million faces for recognition at scale”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 4873–4882.
- [86] Alex Kendall and Yarin Gal. “What uncertainties do we need in bayesian deep learning for computer vision?” In: *Advances in neural information processing systems* 30 (2017).
- [87] Donghyun Kim et al. “Illumination invariant skin texture generation using CGAN from a single image for haptic augmented palpation”. In: *2019 Third IEEE International Conference on Robotic Computing (IRC)*. IEEE. 2019, pp. 197–203.
- [88] Diederik P Kingma. “Auto-encoding variational bayes”. In: *arXiv preprint arXiv:1312.6114* (2013).
- [89] Tao Kong, Gongping Yang, and Lu Yang. “s”. In: *International Journal of Machine Learning and Cybernetics* 5 (2014), pp. 569–578.

- [90] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. “Imagenet classification with deep convolutional neural networks”. In: *Advances in neural information processing systems* 25 (2012).
- [91] Ajay Kumar. “Contactless finger knuckle authentication under severe pose deformations”. In: *2020 8th International Workshop on Biometrics and Forensics (IWBF)*. IEEE. 2020, pp. 1–6.
- [92] Ajay Kumar. “Importance of being unique from finger dorsal patterns: Exploring minor finger knuckle patterns in verifying human identities”. In: *IEEE transactions on information forensics and security* 9.8 (2014), pp. 1288–1298.
- [93] Ajay Kumar. “Incorporating cohort information for reliable palmprint authentication”. In: *2008 Sixth Indian conference on computer vision, graphics & image processing*. IEEE. 2008, pp. 583–590.
- [94] Ajay Kumar. “Toward more accurate matching of contactless palmprint images under less constrained environments”. In: *IEEE Transactions on Information Forensics and Security* 14.1 (2018), pp. 34–47.
- [95] Ajay Kumar. “Toward pose invariant and completely contactless finger knuckle recognition”. In: *IEEE Transactions on Biometrics, Behavior, and Identity Science* 1.3 (2019), pp. 201–209.
- [96] Ajay Kumar and Ch Ravikanth. “Personal authentication using finger knuckle surface”. In: *IEEE Transactions on Information Forensics and Security* 4.1 (2009), pp. 98–110.
- [97] Ajay Kumar and Bichai Wang. “Recovering and matching minutiae patterns from finger knuckle images”. In: *Pattern Recognition Letters* 68 (2015), pp. 361–367.
- [98] Ajay Kumar and Zhihuan Xu. “Can we use second minor finger knuckle patterns to identify humans?” In: *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 2014, pp. 106–112.
- [99] Ajay Kumar and Zhihuan Xu. “Personal identification using minor knuckle patterns from palm dorsal surface”. In: *IEEE Transactions on Information Forensics and Security* 11.10 (2016), pp. 2338–2348.
- [100] Ajay Kumar and Yingbo Zhou. “Human identification using knuckle codes”. In: *2009 IEEE 3rd International Conference on Biometrics: Theory, Applications, and Systems*. IEEE. 2009, pp. 1–6.
- [101] George Kumi Kyeremeh et al. “Fusion of Hand Biometrics for Border Control Involving Fingerprint and Finger Vein”. In: *IEEE Access* (2025).

- [102] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. “Simple and scalable predictive uncertainty estimation using deep ensembles”. In: *Advances in neural information processing systems* 30 (2017).
- [103] Jennifer Langston. *How well do facial recognition algorithms cope with a million strangers?* 2016. URL: <https://www.washington.edu/news/2016/06/23/how-well-do-facial-recognition-algorithms-cope-with-a-million-strangers/> (visited on 05/30/2025).
- [104] Shuyi Li et al. “Hand-based multimodal biometric fusion: A review”. In: *Information Fusion* (2024), p. 102418.
- [105] Chengrui Lin et al. “An unconstrained palmprint region of interest extraction method based on lightweight networks”. In: *Plos one* 19.8 (2024), e0307822.
- [106] Sen Lin, Tianyang Xu, and Xinyong Yin. “Region of interest extraction for palmprint and palm vein recognition”. In: *2016 9th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*. IEEE. 2016, pp. 538–542.
- [107] Fu Liu et al. “A recognition system for partially occluded dorsal hand vein using improved biometric graph matching”. In: *IEEE Access* 8 (2020), pp. 74525–74534.
- [108] Weiyang Liu et al. “Sphereface: Deep hypersphere embedding for face recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 212–220.
- [109] Yang Liu and Ajay Kumar. “Contactless palmprint identification using deeply learned residual features”. In: *IEEE Transactions on Biometrics, Behavior, and Identity Science* 2.2 (2020), pp. 172–181.
- [110] Yunfei Liu et al. “Unsupervised learning for intrinsic image decomposition from a single image”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 3248–3257.
- [111] David JC MacKay. “A practical Bayesian framework for backpropagation networks”. In: *Neural computation* 4.3 (1992), pp. 448–472.
- [112] Wojciech Michal Matkowski, Tingting Chai, and Adams Wai Kin Kong. “Palmprint Recognition in Uncontrolled and Uncooperative Environment”. In: *IEEE Transactions on Information Forensics and Security* 15 (2020), pp. 1601–1615. DOI: 10.1109/TIFS.2019.2945183.

- [113] Wojciech Michal Matkowski, Frodo Kin Sun Chan, and Adams Wai Kin Kong. “A study on wrist identification for forensic investigation”. In: *Image and Vision Computing* 88 (2019), pp. 96–112.
- [114] Wojciech Michal Matkowski, Xiaojie Li, and Adams Wai Kin Kong. “Improving Hand Recognition in Uncontrolled and Uncooperative Environments using Multiple Spatial Transformers and Loss Functions”. In: *arXiv preprint arXiv:2311.05383* (2023).
- [115] Didier Meuwly. “Forensic individualisation from biometric data”. In: *Science and Justice-Journal of the Forensic Science Society* 46.4 (2006), pp. 205–213.
- [116] Didier Meuwly, Daniel Ramos, and Rudolf Haraksim. “A guideline for the validation of likelihood ratio methods used for forensic evidence evaluation”. In: *Forensic science international* 276 (2017), pp. 142–153.
- [117] Didier Meuwly and Raymond Veldhuis. “Forensic biometrics: From two communities to one discipline”. In: *2012 BIOSIG-Proceedings of the International Conference of Biometrics Special Interest Group (BIOSIG)*. IEEE. 2012, pp. 1–12.
- [118] Marco Miani et al. “Laplacian autoencoders for learning stochastic representations”. In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 21059–21072.
- [119] Lynette I Millett and Joseph N Pato. *Biometric recognition: Challenges and opportunities*. National Academies Press, 2010.
- [120] Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. “Obtaining well calibrated probabilities using bayesian binning”. In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 29. 1. 2015.
- [121] Loris Nanni et al. “Different approaches for extracting information from the co-occurrence matrix”. In: *PloS one* 8.12 (2013), e83554.
- [122] Kien Nguyen et al. “Score-level multibiometric fusion based on Dempster–Shafer theory incorporating uncertainty factors”. In: *IEEE Transactions on Human-Machine Systems* 45.1 (2014), pp. 132–140.
- [123] Aditya Nigam, Kamlesh Tiwari, and Phalguni Gupta. “Multiple texture information fusion for finger-knuckle-print authentication system”. In: *Neuro-computing* 188 (2016), pp. 190–205.
- [124] Raid RO Al-Nima et al. “Finger texture biometric characteristic: a survey”. In: *arXiv preprint arXiv:2006.04193* (2020).

- [125] Information Commissioner’s Office. *Biometric recognition*. URL: <https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/lawful-basis/biometric-data-guidance-biometric-recognition/biometric-recognition/> (visited on 08/12/2025).
- [126] Information Commissioner’s Office. *What is special category data?* URL: <https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/lawful-basis/special-category-data/what-is-special-category-data/> (visited on 09/24/2025).
- [127] Timo Ojala, Matti Pietikainen, and David Harwood. “Performance evaluation of texture measures with classification based on Kullback discrimination of distributions”. In: *Proceedings of 12th international conference on pattern recognition*. Vol. 1. IEEE. 1994, pp. 582–585.
- [128] Martin Aastrup Olsen et al. “Convolution approach for feature detection in topological skeletons obtained from vascular patterns”. In: *2011 IEEE workshop on computational intelligence in biometrics and identity management (CIBIM)*. IEEE. 2011, pp. 163–167.
- [129] Nobuyuki Otsu et al. “A threshold selection method from gray-level histograms”. In: *Automatica* 11.285-296 (1975), pp. 23–27.
- [130] The European Parliament and The Council of The European Union. *Directive (EU) 2016/680 of the European Parliament and of the Council*. 2016. <https://eur-lex.europa.eu/eli/dir/2016/680/oj/eng>.
- [131] The European Parliament and The Council of The European Union. *Regulation (EU) 2016/679 of the European Parliament and of the Council*. 2016. <https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng>.
- [132] The European Parliament and The Council of The European Union. *Regulation (EU) 2024/1689 of the European Parliament and of the Council*. 2024. <https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng>.
- [133] Marco AF Pimentel et al. “A review of novelty detection”. In: *Signal processing* 99 (2014), pp. 215–249.
- [134] Metropolitan Police. *MPS Legal Mandate - Retrospective Facial Recognition System*. URL: <https://www.met.police.uk/SysSiteAssets/media/downloads/force-content/met/advice/lfr/new/rfr-legal-mandate-web.pdf%5D>. (visited on 07/04/2025).

- [135] Jerry Qinghui Yu et al. *Bayesian Neural Networks*. URL: https://www.cs.toronto.edu/~duvenaud/distill_bayes_net/public/ (visited on 09/27/2025).
- [136] Gwenolé Quellec et al. “Automatic detection of rare pathologies in fundus photographs using few-shot learning”. In: *Medical image analysis* 61 (2020), p. 101660.
- [137] Filip Radenović, Giorgos Tolias, and Ondřej Chum. “Fine-tuning CNN image retrieval with no human annotation”. In: *IEEE transactions on pattern analysis and machine intelligence* 41.7 (2018), pp. 1655–1668.
- [138] J Redmon. “You only look once: Unified, real-time object detection”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.
- [139] Shaoqing Ren et al. “Faster R-CNN: Towards real-time object detection with region proposal networks”. In: *IEEE transactions on pattern analysis and machine intelligence* 39.6 (2016), pp. 1137–1149.
- [140] Marcos Rodrigo, Carlos Cuevas, and Narciso García. “Comprehensive comparison between vision transformers and convolutional neural networks for face recognition tasks”. In: *Scientific reports* 14.1 (2024), p. 21392.
- [141] Raul Sanchez-Reillo, Carmen Sanchez-Avila, and Ana Gonzalez-Marcos. “Biometric identification through hand geometry measurements”. In: *IEEE Transactions on pattern analysis and machine intelligence* 22.10 (2000), pp. 1168–1171.
- [142] Florian Schroff, Dmitry Kalenichenko, and James Philbin. “FaceNet: A Unified Embedding for Face Recognition and Clustering”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2015.
- [143] Andreas Sedlmeier et al. “Uncertainty-based out-of-distribution detection in deep reinforcement learning”. In: *arXiv preprint arXiv:1901.02219* (2019).
- [144] Huikai Shao, Dexing Zhong, and Xuefeng Du. “Efficient deep palmprint recognition via distilled hashing coding”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*. 2019.
- [145] Yichun Shi and Anil K Jain. “Probabilistic face embeddings”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 6902–6911.

- [146] Tomas Simon et al. “Hand keypoint detection in single images using multiview bootstrapping”. In: *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 2017, pp. 1145–1153.
- [147] Karen Simonyan and Andrew Zisserman. “Very deep convolutional networks for large-scale image recognition”. In: *arXiv preprint arXiv:1409.1556* (2014).
- [148] Ana Slot and Zeno JMH Geradts. “The possibilities and limitations of forensic hand comparison”. In: *Journal of Forensic Sciences* 59.6 (2014), pp. 1559–1567.
- [149] Max Smith-Creasey and Muttukrishnan Rajarajan. “A novel scheme to address the fusion uncertainty in multi-modal continuous authentication schemes on mobile devices”. In: *2019 International Conference on Biometrics (ICB)*. IEEE. 2019, pp. 1–8.
- [150] Nicole A Spaun and Richard W Vorder Bruegge. “Forensic identification of people from images and video”. In: *2008 IEEE Second International Conference on Biometrics: Theory, Applications and Systems*. IEEE. 2008, pp. 1–4.
- [151] Harriet Stratton. “Dorsal hand feature analysis: an aid to forensic human identification”. PhD thesis. University of Dundee, 2015.
- [152] Yi Sun et al. “Deep learning face representation by joint identification-verification”. In: *Advances in neural information processing systems* 27 (2014).
- [153] Zhenan Sun et al. “Ordinal palmprint representation for personal identification [representation read representation]”. In: *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR’05)*. Vol. 1. IEEE. 2005, pp. 279–284.
- [154] Satoshi Suzuki et al. “Topological structural analysis of digitized binary images by border following”. In: *Computer vision, graphics, and image processing* 30.1 (1985), pp. 32–46.
- [155] Christian Szegedy et al. “Going deeper with convolutions”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 1–9.
- [156] Yaniv Taigman et al. “Deepface: Closing the gap to human-level performance in face verification”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014, pp. 1701–1708.

- [157] Chaoying Tang, Hengyi Zhang, and Adams Wai-Kin Kong. “Using multiple models to uncover blood vessel patterns in color images for forensic analysis”. In: *Information Fusion* 32 (2016), pp. 26–39.
- [158] Andreas Uhl and Peter Wild. “Experimental evidence of ageing in hand biometrics”. In: *2013 international conference of the BIOSIG Special Interest Group (BIOSIG)*. IEEE. 2013, pp. 1–6.
- [159] NEC Software Solutions UK. *Metropolitan police deploy NeoFace® Watch technology*. URL: <https://www.necsws.com/case-studies/public-safety/facial-recognition/metropolitan-police-deploy-neoface-watch> (visited on 08/12/2025).
- [160] K Usha and M Ezhilarasan. “Finger knuckle biometrics—A review”. In: *Computers & Electrical Engineering* 45 (2015), pp. 249–259.
- [161] Soheil Varastehpour et al. “An adaptive method for vein recognition enhancement using deep learning”. In: *2019 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*. IEEE. 2019, pp. 1–6.
- [162] Ritesh Vyas et al. “Ensemble-Based Bounding Box Regression for Enhanced Knuckle Localization”. In: *Sensors* 22.4 (2022), p. 1569.
- [163] Ritesh Vyas et al. “Robust end-to-end hand identification via holistic multi-unit knuckle recognition”. In: *2021 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE. 2021, pp. 1–8.
- [164] Haitao Wang et al. “Self quotient image for face recognition”. In: *2004 International Conference on Image Processing, 2004. ICIP’04*. Vol. 2. IEEE. 2004, pp. 1397–1400.
- [165] Jian-Gang Wang et al. “Person recognition by fusing palmprint and palm vein images based on “Laplacianpalm” representation”. In: *Pattern recognition* 41.5 (2008), pp. 1514–1527.
- [166] Mei Wang and Weihong Deng. “Deep face recognition: A survey”. In: *Neurocomputing* 429 (2021), pp. 215–244.
- [167] Frederik Warburg et al. “Bayesian metric learning for uncertainty quantification in image retrieval”. In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 69178–69190.
- [168] Frederik Warburg et al. “Bayesian Triplet Loss: Uncertainty Quantification in Image Retrieval”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Oct. 2021, pp. 12158–12168.

- [169] Nuoya Xu et al. “An effective recognition approach for contactless palmprint”. In: *The Visual Computer* 37 (2021), pp. 695–705.
- [170] Jianwei Yang et al. “A modified Gabor filter design method for fingerprint image enhancement”. In: *Pattern Recognition Letters* 24.12 (2003), pp. 1805–1817.
- [171] Erdem Yoruk et al. “Shape-based hand recognition”. In: *IEEE transactions on image processing* 15.7 (2006), pp. 1803–1815.
- [172] Hongyang Yu et al. “A new finger-knuckle-print ROI extraction method based on Two-stage center point detection”. In: *International Journal of Signal Processing, Image Processing and Pattern Recognition* 8.2 (2015), pp. 185–200.
- [173] Yimin Yuan et al. “HandNet: Identification based on hand images using deep learning methods”. In: *Proceedings of the 2020 4th International Conference on Vision, Image and Signal Processing*. 2020, pp. 1–6.
- [174] Fan Zhang et al. “Mediapipe hands: On-device real-time hand tracking”. In: *arXiv preprint arXiv:2006.10214* (2020).
- [175] Hengyi Zhang et al. “Matching vein patterns from color images for forensic investigation”. In: *2012 IEEE Fifth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*. IEEE. 2012, pp. 77–84.
- [176] Lin Zhang, Lei Zhang, and David Zhang. “Finger-knuckle-print: a new biometric identifier”. In: *2009 16th IEEE international conference on image processing (ICIP)*. IEEE. 2009, pp. 1981–1984.
- [177] Lin Zhang et al. “Towards contactless palmprint recognition: A novel device, a new benchmark, and a collaborative representation based identification approach”. In: *Pattern Recognition* 69 (2017), pp. 199–212.
- [178] Dexing Zhong, Xuefeng Du, and Kuncai Zhong. “Decade progress of palmprint recognition: A brief survey”. In: *Neurocomputing* 328 (2019), pp. 16–28.
- [179] Karel J Zuiderveld et al. “Contrast limited adaptive histogram equalization.” In: *Graphics gems* 4.1 (1994), pp. 474–485.

Appendix A

VarHands Recruitment Materials

 Information Sheet.pdf
140 KB

 Consent Form.pdf
72 KB

Thank you for your interest in participating in this data collection. This should take no more than 10 minutes in total but will need to be completed in at least two sessions (daytime and night-time), at your own pace. Before beginning, please can you read the attached information sheet and consent form, and respond to this email indicating your consent or lack of consent for sections **a** and **b** in the consent form.

You are asked to take eight 10-20 second video recordings of your hands with your personal mobile device. These eight videos will be: left hand in clear daylight, left hand in shaded daylight, left hand in clear artificial light (at night), left hand in shaded artificial light (at night), and the same again for your right hand.

Please take care to keep your hands in focus, in frame, and to not have any complex backgrounds or other identifying information in the videos. I also ask that you remove any jewellery beforehand, except items that are worn all of the time (such as wedding bands).

The audio will be stripped from the videos and they will not be stored alongside any identifying information such as your name or contact details. Frames from the videos will be extracted to form a hand image dataset. In order to obtain a wide range of views of your hands, please perform a variety of gestures and rotations, which could be:

1. Start with your fingers outstretched, palm facing the camera.
2. Slowly curl your fingers one at a time into a clenched fist.
3. Rotate your hand so the palm is facing away from the camera.
4. Slowly uncurl your fingers one at a time until they are all outstretched.
5. Rotate your hand so the palm is facing the camera again.

This video shows the movements and different lightings required: <https://youtu.be/Z5OPjIjQwi4>

Once you have recorded all eight videos, use the following private link to upload them all at the same time (you may need to open this email on your mobile device): []

Please feel free to contact me with any further questions,
Best regards,
Gabrielle Hornshaw

PARTICIPANTS NEEDED

Earn a £10 Amazon voucher for your hand images.

In images of abuse, hands can sometimes be the only available information to identify the perpetrator. Automated comparison of hands can provide insight about cases and suspects, improving the scale and efficiency of investigations and reducing the human burden of viewing harmful material. To make such automation possible we will need a new dataset featuring a wide variety of views of each participant's hands.



Want to contribute?

Participants are asked to provide 8 videos of their hands performing a series of gestures and movements, each no longer than 20 seconds and recorded under different lighting conditions.

To participate you must be over the age of 18 and have not contributed images to any other hand image dataset. The first 50 participants will receive a £10 Amazon voucher.

If you have a University of Nottingham email address, you can record the videos on your phone in your own time and submit them online at the link found below.

If you do not have a UoN email and would like to participate, or if you have any questions, please contact Gabrielle at the email address below.

gabrielle.hornshaw@nottingham.ac.uk

For further project information, instructions, and a video demonstration, please use this QR code or go to:

<https://forms.office.com/e/TnRStT97z>



Engineering and
Physical Sciences
Research Council

This research has been approved by the University
of Nottingham School of Computer Science ethics
committee.

