Biometric quality and its impact on template ageing in a longitudinal fingerprint study

by

Henry John Harvey

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Electrical and Computer Engineering

Carleton University
Ottawa, Ontario, Canada, K1S 5B6

Carleton
UNIVERSITY

Abstract

Biometrics are increasingly deployed in domains ranging from social media authentication right up to border control. An important operational requirement for biometric systems is the supposed uniqueness and permanence of biometric records: physiological changes occurring between enrolment and verification are referred to as *template ageing*, and increase the likelihood of a misidentification. Its magnitude is hard to estimate, and the factors affecting it are relatively little studied.

This work proposes a measure of template ageing, called *biometric permanence*, and develops a methodology to estimate it in the presence of confounding factors. The measure is applied to a database of fingerprints obtained over a seven year period, using bootstrap resampling to obtain confidence intervals for the estimates of effect size. Fingerprint quality metrics are evaluated in terms of their ability to predict classification performance, and the subject-dependence of fingerprint quality is explored using the ideas of a "biometric menagerie". Statistically significant demographic factors underlying biometric quality and template ageing are highlighted and discussed.

The results of this work may have implications for the procurement and administration of biometric systems: for example, in ensuring consistent performance across a broad population demographic, and in the choice of credential lifetime and re-enrolment policy.

Dedication

For REE

Acknowledgements

I would like to thank David Dawson for coordinating and overseeing the data collection, and Bion Biometrics, Inc. for making available the ISBIT database and software.

The experimental phases of this work were supported in part by the Natural Sciences and Engineering Research Council (NSERC), grant number CRD 428240-11.

I would like to thank Dr. Andy Adler, my supervisor, for encouraging me to undertake this work and for his continued support; and Dr. John Campbell for his many helpful suggestions.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Biometrics refers to the use of certain characteristic attributes of an individual's person or behaviour in order to identify them or confirm their claimed identity – from the individual's face, fingerprint, or pattern of vasculature to their voice, gait or even heartbeat. They are deployed in applications ranging from unlocking a personal communication device such as a cellphone or tablet, authenticating to social media and banking apps, right up to government-issued identity documents such as biometric passports, visas and electronic travel authorizations.

1.1 Problem statement

The desireable properties of a biometric characteristic, enumerated by Jain in *Handbook of Biometrics* [42] are:

1 Universality

2 Uniqueness

3 Permanence

4 Measurability

5 Performance

6 Acceptability

7 [resistance to] Circumvention

Some of these properties are easy to comprehend. By *Universality* for example, we mean that the chosen trait should (for both operational and ethical reasons) exclude from participation as few individuals as possible: a gait-based biometric would exclude wheelchair users, while iris biometrics may disadvantage individuals with certain types of eye disease, such as cataracts [76]. *Measurability* addresses concerns of acquisition convenience, and leads us to favour fingerprints over toeprints for example, since access to the former is less likely to require removal of clothing. *Resistance to circumvention* encompasses anti-spoofing measures such as sensor fusion [62] and liveness detection [24].

Meanwhile *Performance* has a natural definition in terms of the binary classification problem [19], by which an individual's biometric presentation is classified as either a "match" or a "non-match" according to some decision rule, with well-established metrics – the Type I and Type II error rates, conventionally termed False Match Rate (FMR) and False Non-Match Rate (FNMR) in the biometric context – and comparison tools such as the Decision Error Tradeoff (DET) curve [83].

The properties of *Uniqueness* and *Permanence* are less well established. In the context of a biometric Identity Management System (IDMS), they concern the ability to distinguish unambiguously between any two individuals in a given cohort, and to continue to do so for as long as required by the particular application - such as the duration of a biometric electronic travel authorization (ETA). While several biometric modalities have established track-records of utility in such applications, this does not necessarily establish that they provide a permanent and unique record of an individual.

The concept of the uniqueness of a single object (such as a biometric record) is a difficult one, both philosophically and mathematically. In the mid twentieth century, Kolmogorov and others developed the notion of algorithmic complexity (see for example Li & Vitányi §2.8 [50]) to quantify the information in an object in terms of its shortest possible description in some universal programming language. To the author's knowledge, no attempt has been made to apply such methods in the field of biometrics: instead, biometric uniqueness is approached practically in terms of classification performance. That is, a biometric record is considered unique if it can reliably be distinguished from any other record.

In the particular case of fingerprints, Tabassi et al. in the US National Institute for Standards and Technology (NIST) have used machine learning techniques to identify sets of features that are likely to result in good classification accuracy (high genuine match score, low imposter match score) resulting in the public release of the NIST Fingerprint Quality (NFIQ) algorithms [74, 73].

From an information-theoretic point of view, a finite length string or data record can only convey a finite amount of information about an individual: in particular, there is an obvious lower bound of $\log_2 M$ bits required in order to uniquely index the members of a cohort of size M. The Data Processing Inequality (see for example Cover & Thomas §2.8 [9]) asserts that no subsequent processing can increase information content: so the sequence of steps from a physical attribute (such as a finger), to an image of that attribute, to a set of features (such as a fingerprint minutia record) describing that attribute each tends to reduce its information content – and hence its uniqueness. In the information-theoretic view, *permanence* becomes a question of the extent to which the mutual information between a biometric record and the physical attribute upon which it is based remains constant: that is, *biometric template ageing* expresses a decrease in mutual information over time. Some questions that arise naturally are:

- How should we define biometric permanence (or template ageing) in a way that is operationally meaningful?
- How can we estimate the magnitude of such ageing, in the presence of confounding factors such as environment and operator acclimation?
- Where different biometric capture technologies and feature extraction algorithms exist within a modality, is template ageing observed consistently between them?
- Do currently available measures of biometric quality adequately predict classification performance?
- To what extent can such biometric quality metrics be used to improve the overall classification performance of a biometric IDMS?
- What demographic factors affect biometric quality (and, by extension, biometric IDMS performance)?
- Are the same demographic factors significant in the observed template ageing?

It should be noted that while we may refer throughout this work to properties such as quality and permanence as characteristics of a biometric, they are (to the extent that we can evaluate them) in fact properties of a whole biometric system, in which everything from the underlying physiological trait, through the capture, processing and storage of a biometric record, to the match scoring and classification algorithm play their roles. In particular, when we consider (in Chapters 6 and 7) the demographics of fingerprint quality and classification performance, it is important to note that the design of the capture systems, the selection of features, and the training of scoring and classification algorithms may be as important as any intrinsic demographic factors of the underlying trait.

In the case of template ageing, one might suspect that physical degradation of the sensor (such as scratching or marring of the platen) could be a significant non-physiological factor. In principle, one might try to minimize this by procuring a number of identical devices from each manufacturer and using a new one for each data collection phase. Unfortunately such a procedure was beyond the scope of this study.

1.2 Goals

The goal of this work is firstly to develop an operationally meaningful estimate of biometric template ageing, and to apply it in a multi-year longitudinal fingerprint matching study. Secondly, to investigate the relationships between available measures of biometric quality, subject demographics, and classification performance in the same data. Finally to apply these findings on quality and demographics to the problem of template ageing.

1.3 Contributions

1.3.1 Methodology for estimating biometric permanence

In Chapter 4 we develop and define a measure of template ageing which we call biometric permanence P_B, based on the change in FNMR (at a given FMR) between the template ageing interval under test, and a short-time test. While intuitive, this definition of P_B is practically difficult to apply to estimate small changes in permanence in a longitudinal study subject to experimental error and visit-to-visit systematic biases. To address this issue, we introduce the "matched delta" method. Comparisons of these methods are performed using simulated data, and it is determined that the new method showed dramatically reduces sensitivity to systematic biases.

1.3.2 Observation and quantification of template ageing

In Chapter 5 we apply the preceding methodology to data collected in a multi-year, multi-vendor experimental fingerprint acquisition and matching study, involving over 350 participants, with a gallery size in excess of 12,000 ISO/IEC standards-compliant two-finger biometric enrolment templates obtained with a variety of commercially-available fingerprint sensor technologies. Confidence intervals for template ageing are estimated using non-parametric methods. The behaviours of different vendors' devices are compared and contrasted, and limitations of the methodology identified and discussed.

1.3.3 Investigation of the effect of biometric data quality on classification performance

In Chapter 6 we apply the NIST NFIQ fingerprint quality measures to the images collected in our study, comparing and contrasting the results for NFIQ-1 and NFIQ-2 across different device technologies. We investigate the relationship between reported quality and match score, for both genuine and imposter matches, and between quality score and classification performance. In particular, we confirm that the overall classification performance may be improved by rejecting a fraction of fingerprints based on their quality.

Interestingly, for our data, we find that NFIQ-1 and NFIQ-2 are equally effective at identifying any given fraction of low-quality presentations: the operational advantage of NFIQ-2 is that its more expressive quality scale allows the rejection fraction to be chosen much more precisely.

1.3.4 Observation and demographics of a biometric menagerie

In Chapter 7 we adapt a taxonomy first introduced by Doddington [17] in order to identify subsets of 'Goats', 'Lambs', and 'Wolves' in our data. We establish that these categorizations are to a large degree common across different fingerprint capture devices, and hence substantially reflect intrinsic properties of the underlying biometric. The demographics of the taxonomic subsets are explored: we find that, for our data, females and older individuals are overrepresented in the Goat subset (those individuals whose fingerprints contribute disproportionately to the FNMR), while younger individuals are overrepresented among the Lamb and Wolf subsets (the individuals whose fingerprints contribute disproportionately to the FMR). We discuss the extent to which these differences might be explained by training bias in the classification algorithms. Finally, we examine the impact of the Goat subset on template ageing, observing a significant improvement in biometric permanence when the subset is removed.

1.3.5 Outline for a biometric channel model

While the development of a comprehensive information-theoretic treatment of biometrics has remained an aspirational goal, it proved out of reach of the present work. However some steps towards a biometric channel model are outlined in the concluding chapter.

1.4 Publications

The following publications are based on the work presented in this thesis:

- In *Proceedings of the 2017 Annual IEEE International Systems Conference (SysCon)*: J. Harvey, J. Campbell, S. Elliott, M. Brockly and A. Adler *Biometric Permanence: Definition and Robust Calculation* [31]

- In *IEEE Transactions on Instrumentation and Measurement*: J. Harvey, J. Campbell and A. Adler *Characterization of Biometric Template Aging in a Multiyear, Multivendor Longitudinal Fingerprint Matching Study* [30]

Chapter 2

Background and literature review

2.1 History and early application of biometrics

While in its broadest sense, the term *biometrics* has historically been used to denote the measurement and statistical analysis of biological data in general [10] – which nowadays might more likely be referred to as *biostatistics* – it is now almost universally understood to mean the use of biological traits to establish or confirm the identity of an individual [42]. Among the biometric traits (or *modalities*) that have been studied and/or employed for this purpose include the features and morphology of the face [54], an iris image [46], a pattern of blood vessels [7], or an analysis of the individual's voice [69] or gait [51]. The focus of this work is the biometric modality of fingerprints [53].

Interest in the features of the human hand has a long cultural history from the point of view of their purported usefulness for divination or "cheiromancy" [6], but its modern development as a biometric modality really begins during the nineteenth century: Galton [25] gives a more-or-less contemporary (albeit subjective) account. Although his own primary interest seems to have been what the study of fingerprints might reveal about heredity, biological symmetry (homochirality) and specia-

tion, Galton devotes a whole chapter to their application to personal identification; in particular, the use of "signs-manual" in place of conventional written signatures in the attestation of contract documents, the pioneering of which he attributes to Sir William Herschel (and which was subsequently reported by him [33]). Galton also discusses the then-emerging forensic use of fingerprinting, first noting its distinction as a "much more difficult" one-to-many (identification) task rather than a one-to-one (verification) task, and going on to consider how even a relatively simple "A.L.W." (arch-loop-whorl) fingerprint classification scheme might greatly improve the utility of the French anthropomorhic system [3] of *Bertillonage*. Such a scheme was then already in use by police in Calcutta (now Kolkata), India under the direction of Henry [32] and latterly substantially attributed to Haque [70]. Roughly contemporaneous contributions in other jurisdictions, especially those of Vucetich in Argentina and Brazil, are also noted in a historical survey by Polson [59].

It should be noted that the Henry-Haque classification scheme was based on a coarse attribution of each finger's dominant feature, rather than the kind of minutia-based classification of single fingers provided by the devices used in the present work - although Galton (op. cit.) at least was familiar with the concept and terminology of fingerprint minutiae: one of the appealing features was its ability, after application of a coding scheme due to Bose [4], to be transmitted telegraphically – a valuable factor given the rise of mass transportation and the increased mobility of criminal suspects.

2.2 Modern biometrics

While much of the early history and development of biometric techniques focused on the identification and apprehension of criminal suspects, progress in automated capture, feature extraction, and algorithmic matching technologies has allowed biometrics to expand into the fields of large-scale identity management systems (IDMS). Such

schemes include those for machine readable travel documents [29], trusted traveller programs [78], and governmental personal identity verification (PIV) programs [28].

2.2.1 Renewable biometric references

Password and public key infrastructure (PKI) based authentication systems provide the ability for an issuer to revoke and renew credentials simply by deleting passwords or keys and inserting new, randomly generated, ones. In contrast, raw biometric records provide limited opportunity for revocation or renewal - as noted by Shreier et al., "we have one face, two irises, 10 fingers" [66]. Considerable interest has been directed to addressing this deficiency in order to develop what have become known as renewable biometric references (RBRs) [39]. A key goal of such efforts is to protect the personally identifiable information (PII) of the individual [44, 43] while providing sufficient immunity to a variety of potential compromizes including attacks via record multiplicity (ARM), surreptitious key-inversion (SKI), and blended substitution attacks (BSI) [65].

2.2.2 Forensic applications

The matching of latent fingerprint (or partial fingerprint) images recovered from scenes-of-crime remains an important application of biometrics [52], with much current attention given to pre-processing of latent fingerprints [49], especially using chemical [21] and spectroscopic [11] techniques. Unlike many other biometric applications, that are dominated by one-to-one (or biometric verification) matches, forensic biometrics may include one-to-many (or biometric identification) tasks, such as identifying a list of suspects from an existing criminal database, as well as the association of criminal cases based on collection of latent fingerprints from an *unknown* common subject [52]. Biometric quality and template ageing are surely relevant to these applications, however they may have additional domain-specific aspects that are not

addressed in the present work - such as assessing the biometric quality of partial prints.

2.3 Biometric template ageing

An assumption underlying the deployment of biometric IDMS systems is the stability of the chosen biometric features – that is, that the biometric trait will remain, over the expected lifetime of the credential, sufficiently similar to that of the template to enable a positive comparison. In applications such as biometrically-enabled passports, stability over a period of five or ten years is desirable in order to align with current renewal policies for such credentials [36]. From a physiological point of view however, it is natural to expect some change in traits over time. For example, a subject's loss or gain in weight may affect measurements of hand geometry [68], while the onset of degenerative disease, injury, or occupational damage may affect fingerprints [18, 12]. As an instrumentation and measurement problem, biometric capture has in this respect something in common with many clinical monitoring and medical imaging systems: that is, the systems should be sensitive to clinically significant changes (in the case of biometrics, a change of identity) while remaining relatively insensitive to benign morphological changes arising from simple ageing or weight gain for example.

Slow changes in biometric features over time are typically referred to as *template ageing* [82, 79], and the performance of large-scale systems can be influenced by this effect. Unfortunately template ageing is hard to measure, because it is very sensitive to the visit-to-visit variability inherent in such a study (e.g. test personnel [5], test equipment and weather [20, 71]).

Attempts to quantify biometric permanence in fact go right back to Galton (op. cit.), who devoted an entire chapter to observations on the persistence of fingerprint minutiae in an (admittedly small) sample of 15 individuals. In one case, the inter-

val between observations was as large as 31 years, while in another he was able to record prints of a juvenile individual (aged approximately two-and-a-half years) and subsequently compare them to those obtained at age 15 years. Galton estimated that he could identify, on average, 35 "points of interest" (minutiae) from each digit, and that of the 700 such points provided by a full set of 10 fingerprints, 699 could be "inferred" to remain throughout an individual's life (and beyond - based on fingerprints apparently having been obtained from Egyptian mummies).

Modern attempts to quantify permanence (or template ageing) are generally based on statistical analyses of large numbers of biometric enrolments. The age progression of biometric traits has perhaps received most attention within the facial recognition and iris recognition modalities. Manjani et al. [54] evaluated both 2D and 3D facial recognition algorithms on a dataset of sixteen participants acquired over a period of ten years, comparing genuine acceptance rate (GAR) at 0.1% false acceptance rate (FAR) for short-term intervals (less than three months between enrolment and verification) versus long-term intervals (more than five years between enrolment and verification). Unlike the present work, the intervals were not blocked into absolute acquisition times i.e. all intervals greater than five years were taken together. They were able to reject at $\alpha = 0.05$ the null hypothesis that the short- and long-term genuine scores were drawn independently from normal distributions of equal mean and variance (t-test), or from the same continuous distribution (Kolmogorov-Smirnov test). In the case of the algorithm that performed best over the long-term intervals ("3D Region Ensemble: Product"), they found weak evidence against the corresponding hypotheses for the imposter scores: this is consistent with our model, in which the imposter distribution was assumed to be constant.

Lanitis & Tsapatsoulis [48] proposed a measure of biometric ageing that they called "Aging Impact" (AI), derived from the homogeneity and dispersion of a collection of templates. Although the primary focus of their work concerned facial images, finger-

and palm-print images were also considered; however they applied their method to individuals within different age classes, rather than to repeated measures of the same individuals over time as in the present work. The focus of much subsequent work has been the development and evaluation of artificial age progression algorithms for forensic applications [47, 58], rather than for biometric IDMSs.

Template ageing has also been reported in the iris modality [22, 26]. Hofbauer et al. [34] noted some controversy about its existence, and discussed the difficulty of controlling confounding factors independently – in particular, the cases of illumination and pupillary dilation. There was only a single long-term time interval – in this case of four years – while the study consisted of data from 47 subjects. The authors considered two schemes for re-normalization of pupil diameters: a "rubber sheet model" (RSM) and a "biomedical model" (BMM). They showed that while such re-normalizations were somewhat effective in improving long-term match accuracy, there was still a decrease in performance between intra-year and inter-year comparisons. This suggests that while systematic changes in pupillary diameter are a factor in iris template ageing, they are not the only such factor. Significant degradation over time in genuine iris match scores have been reported elsewhere [13].

Fingerprint ageing might be expected to share some of the same physiological factors as face ageing – in particular, skin textural changes and loss of tissue elasticity – and has been reported by Uludag et al. [77], who addressed the case of typicality and/or variability between presentations of the same biometric using novel template selection algorithms, based either on clustering or on mean distance. They then used this template selection to evaluate a number of template update schemes. They found that a scheme in which an original template was updated selectively using later presentations ("AUGMENT-UPDATE") outperformed one in which the original template data were discarded altogether ("BATCH-UPDATE"). From this, we might infer that the magnitude of the template ageing effect was not significantly greater

than that of the intraclass variance, at least over the relatively short interval of their study (approximately four months). Template ageing has recently been reported in two non minutia based fingerprint matching schemes [45]: FingerCode (FC), a Gabor-filter based technique similar to the widely adopted IrisCodes of Daugman [14]; and Phase Only Correlation (POC).

Template ageing has also been observed in speech biometrics [82].

Meanwhile, the influence of biometric sample quality on template ageing was highlighted by Ryu et al. [64], who found that lower sample quality (evaluated using the NIST NFIQ measure [74]) was associated with an increased number of matching errors.

The social and ethical implications of biometric ageing have also received recent attention [61]: most notably the potential role of biometrics in the "problematisation of ageing and of older people". The authors are careful to distinguish between a biometric subject's chronological age and biometric template ageing: their arguments for the exclusionary potential of the former (which is known to affect biometric system performance [67]) are stronger than those for the latter, which rely on a rather subtle semiotic analysis of the relationship between biometric features, subject, and biometric system as a whole.

2.4 Definition and evaluation of biometric data quality

Jain [42] identifies a number of properties that are desirable in a biometric characteristic, including *uniqueness* and *permanence*; *performance* and (resistance to) *circumvention*. *Performance* here may be interpreted as the system's ability to correctly identify the biometric presentation of a genuine subject, and to correctly reject the biometric presentation of an imposter subject. As with any such pattern classification problem, these abilities are inherently conflicting, and represent the Type I

and Type II error probabilities of a classical Neyman-Pearson hypothesis test [57] or signal detection problem. As such, biometric systems typically reduce the determination to a one-dimensional similarity score which may be thresholded in order to obtain a match/non-match decision; by appropriate choice of the decision threshold, the system integrator or operator may trade off security (lower false accept rate, FAR) against convenience (lower false reject rate, FRR) to suit the requirements of the particular IDMS application.

At any particular threshold, the fundamental performance (i.e. the minimum obtainable FRR at a chosen FAR, or vice-versa) will be intrinsically limited by the separability of the subjects' biometric characteristics over some biometric feature space. In the case of fingerprints (the focus of this proposed work), the feature space is usually a space of extracted fingerprint minutiae types and locations. In general, we would expect a feature space of higher dimensionality (more independent features) to permit higher classification accuracy. Features are, however, not always informative: in particular, the set of features that best describes a population may not coincide with that which best discriminates between its classes. Thus in the case of facial recognition for example, linear classifiers based on discriminant analysis (so-called "Fisherfaces") may outperform those based on principal component analysis (PCA) [2] (or "eigenfaces").

If biometric performance is defined in terms of classification accuracy in this way, then uniqueness is essentially a measure of intrinsic performance (i.e. the classification accuracy that might be obtained in the absence of any variability in the collection and/or processing of the biometric). Permanence becomes a measure of how well discriminability is maintained over time. These three characteristics each represent aspects of the informativeness of the biometric; in fact Adler et al. have sought to define *biometric information* formally in this sense as

> "the decrease in uncertainty about the identity of a person due to a set of

biometric measurements" [1]

In this view, one might expect a biometric's resistance to circumvention also to be related to its performance, since an attacker would need to expend more effort to spoof a more informative record - although in practice, external measures such as liveness detection and/or multi-factor authentication requirements are likely to be more significant.

Grother & Tabassi were among the first to formalize the evaluation of biometric quality as a predictor of genuine match score [27]. In particular, they addressed the fact that defining quality in this way necessarily involves the interaction of at least two biometric presentations[1] whereas, to posses utility, the quality measure so derived should be applicable to a single presentation. They discuss the appropriateness of various quality combination functions in order to explore the dependence of similarity score on match pairs of differing quality, as well as the useful number of levels of quantization of biometric quality. Distinctions were highlighted between positive identification (verification) applications, in which an enrolled subject is "motivated to submit high quality samples", and negative identification (blacklist) applications, where an individual is perhaps enrolled unwillingly and may be motivated to obscure or obfuscate their biometric: in the former case, they identified the key performance metric as false non-match rate (FNMR), while in the latter it is false match rate (FMR). They demonstrated the dependence of similarity score on quality in the positive identification case through the use of error versus reject curves. The notion of an ideal quality metric for the positive identification case was developed as follows:

Suppose a system is operating at a FMR determined by operational security requirements. There will be an associated FNMR x-%, meaning that x-% of genuine biometric classification scores fall below the decision threshold for that FMR, and are misclassified as imposters. An ideal quality metric for this case would be one that

[1] It may be more than two, since template generation may be based on multiple enrolment presentations.

identifies exactly these presentations, and removes them from consideration - thereby reducing the FNMR to zero.

This formalism was applied in the development of the NIST NFIQ [74] and NFIQ-2 [73] fingerprint quality algorithms used in the present work.

2.5 The biometric menagerie

Doddington introduced the idea of a "biometric zoo" [17] to describe the subject-dependence of biometric classification errors, originally applying it in the speaker verification modality. Four categories of individuals were posited: those whose biometric matched poorly against itself, and which therefore contributed disproportionately to the FNMR, were labelled "Goats"; those whose biometric was easily impersonated[2] as "Lambs", and those whose biometric is easily mis-attributed to such lambs as "Wolves", with Lambs and Wolves together contributing disproportionately to the FMR. Remaining individuals were classified, by elimination, as "Sheep". Similar categorizations have since been established in the face [81], iris [72] and fingerprint [80] modalities; the latter identifying Lambs and Wolves in a multi-finger dataset obtained from 510 individuals; Goats could not reliably be identified because of the relatively small number of genuine matches[3].

The biometric menagerie has subsequently been refined by considering the interactions of genuine and imposter scores [84]. The question of whether such categorizations generalize across biometric matching algorithms and data sets has been investigated by Teli et al. [75]. More recently, the concept of a biometric menagerie has been applied to biometric template update procedures [60] and to biometric fusion [63]. We believe that the present work is the first to extend the concept specifi-

[2]The term *impersonation* does not necessarily imply an actively malicious actor here: we are often interested in so-called *zero effort* imposters – individuals whose biometric naturally closely resembles that of another.

[3]Although the number of individuals was larger than that of the present study, only a single enrolment-verification pair was collected for each.

cally to the subject-dependence of biometric permanence.

2.6 Relationship between template ageing, quality, and demographics

A recent large-scale longitudinal study by Yoon & Jain examined both genuine and imposter match scores versus template age, NIST NFIQ fingerprint quality, and subject demographics [86]; the latter consisting of subject age, sex, and a binary race variate. The study size was large, consisting of 10-finger records of more than 15,000 subjects, with intervals between acquisitions ranging from five to twelve years. Bootstrapped estimates of mean genuine match score showed clear decreasing trends with time interval (i.e. template age), subject chronological age, and NFIQ score with only marginal dependence on the other factors.[4] Of the three significant predictors, NFIQ score was found to be the strongest. Although the effect size was large enough, for these factors, to be estimated with high confidence, the genuine and imposter score distributions remained sufficiently separable over the duration of the study that there was no observed change in either the true acceptance rate (i.e. $1 - \text{FRR}$) or false acceptance rate (FAR).

Finally, Kirchgasser & Uhl attempted to relate observed biometric template ageing, over a four year interval, in the fingerprint modality to decreases in biometric quality [45], again using the NIST NFIQ metric of Tabassi (op. cit.) as well as BRISQUE - a non fingerprint specific measure of image spatial quality. However – perhaps due to the relatively small study size of only 49 participants – they were unable to do so, even observing some counter-intuitive negative correlation between NFIQ score and genuine match score among false non-matches.

[4] A *decreasing* trend because NFIQ scores from 1 (best) to 5 (worst).

2.7 Open questions

Yoon & Jain's study demonstrates an important feature of biometric template ageing: namely, that changes in match score do not necessarily result in observable changes in classification performance - at least, not over the time periods available for study. An open question therefore is how should we characterize template ageing in a way that is operationally meaningful?

Comparing the results of Yoon & Jain with those of Kirchgasser & Uhl highlights another important question: how can we estimate the magnitude of template ageing robustly in smaller cohorts, where we do not have the advantages of large sample size to reduce estimation variance?

Turning to issues of biometric quality, we would like to know how well the publicly-available NFIQ metrics perform as predictors of classification accuracy, in an independent dataset collected under different conditions and protocols than those on which their classifiers were trained. Demographic factors affecting biometric quality have been reported by Yang [85], but the analysis did not consider the demographics of classification accuracy (or of template ageing) directly.

The study described in this thesis is somewhat larger than that of Kirchgasser & Uhl and, although far smaller than that of Yoon & Jain, we believe that it supports their main conclusions concerning template ageing in the fingerprint modality, as well as providing valuable additional evidence for the role of biometric quality and subject demographics.

Chapter 3

The "Norwood" dataset

3.1 History and demographics

The data used for this study were collected in four phases between 2006 and 2013. The first two phases, in 2006 and 2008, were collected as part of a study on biometric system interoperability undertaken on behalf of the International Labour Organization (ILO) and known as the "Seafarers' Identity Documents Biometric Interoperability Test", or ISBIT. Data collection in these phases, known as ISBIT-3 and ISBIT-4, was undertaken by Bion Biometrics, Inc. with subject recruitment from a general population in and around Ottawa, Canada.

In 2012, Carleton University, in collaboration with Bion Biometrics, obtained funding from the National Research Council (NRC) in Canada for a project entitled "Effect of template ageing and sensor technologies in fingerprint recognition" [1]. The project was to leverage and extend the database and software from the ISBIT studies, with a re-focus towards template ageing. Funding was sufficient for a further two phases of data collection, which were undertaken in 2012 and 2013. Vendors who had provided fingerprint capture devices and API software for the ISBIT studies were approached to permit their use in the new study, and to renew software licences

[1] NSERC CRD 428240-11

Table 3.1: Pseudonymized devices and sensor technologies.

ID	Sensor technology	Image dimensions (pixels)
A.	Optical	420x480
B.	Optical	456x480
C.	Optical	524x524
D.	Optical	640x480
E.	Optical	416x416
F.	Optical	512x512
G.	Optical	524x524
H.	Multispectral optical	352x524
J.	Optical	524x524
K.	Optical	620x620
L.	Capacitive semiconductor	256x360

where required: one vendor declined to participate, and their device was eliminmated from the study. The remaining available devices and technologies are summarized in Table 3.1 using their pseudomomized identifiers from the ISBIT study reports [37, 38].

In order to minimize confounding factors due to API software and library differences, operating systems and vendor supplied software were frozen at their 2008 release points and steps were taken to isolate the experimental setup from further updates.

Recruitment for the new phases began in January 2012, under the oversight of Carleton University Research Ethics Board 'B' (CUREB-B). Approximately 40% of the original studies' participants were re-recruited into the new study (Figure 3.1); taken together with the ISBIT phases this provided intervals of approximately one, two, four, five, six, and seven years between biometric enrolment and verification events (Table 3.2).

Note that although the periods of Visit 7 and Visit 8 overlap, no individual subject's data was recorded in any later visit before an earlier one; the dates simply reflect difficulties in scheduling appointments for a small number of the subjects in

Table 3.2: Dates of data collection visits

	Visit start (YYYY-MM-DD)	Visit end (YYYY-MM-DD)
1	2006-02-07	2006-03-03
2	2006-02-22	2006-03-24
3	2008-09-26	2008-10-15
4	2008-10-13	2008-10-27
5	2012-02-12	2012-03-03
6	2012-03-12	2012-03-31
7	2013-03-06	2013-04-22
8	2013-04-05	2013-04-27

the cohort.

Figure 3.1: Overlap of participants between data collection phases (the 2013 collection is omitted for clarity; it overlaps almost completely with 2012).

3.2 Data collection protocol

In our study, data were collected in four phases, each consisting of a pair of subject visits separated by approximately two weeks in each of the years 2006, 2008, 2012 and 2013. Approximately 200 participants were recorded in each phase, with more than 100 taking part in at least two phases and over 70 being present in all four (Figure 3.1). The protocol for each subject visit consisted of a sequence of two-finger enrolments, followed by a sequence of single-finger verification presentations [37, 38]. Preferred fingers for enrolment were right and left index in the first instance; however if either of these was unavailable (or failed to enrol) alternate fingers were offered in the order right thumb, left thumb; right middle, left middle; right ring, left ring; and finally right and left "pinky" fingers. In subsequent enrolments, previously enrolled fingers were preferred in order to maximize the number of potential genuine matches. Three bitmapped images of each candidate finger were captured during each enrolment, and a further six images (in two distinct three-presentation verification attempts) per enrolled finger during each verification, such that a typical visit resulted in eighteen single-finger images per subject per device. In each subject visit, the order in which devices were presented for both enrolment and verification was randomized under software control in order to counterbalance for subject and operator acclimation.

In order to minimize labelling errors, the captured images were examined at intervals during or immediately after every visit by an experienced human operator[2]. While this procedure cannot guarantee that finger labelling is correct (i.e. that an image labelled as "Subject k, finger d" does in fact come from that subject-finger) it at least ensures, with high probability, that the labelling is consistent across all records for a particular subject-visit. Images that were corrupted (due to malfunctions of the device hardware or capture software for example) were also flagged during this examination, and removed from the dataset at this stage.

[2]Dr. John Campbell, Bion Biometrics, Inc.

Table 3.3: Numbers of genuine and imposter scores.

ID	Genuine	Imposter	ID	Genuine	Imposter
A	92243	24418495	G	62476	15301808
B	93630	25282974	H	61698	14901522
C	91326	24352257	J	57803	13646908
D	98725	27124531	K	98872	27125117
E	56047	14296890	L	99328	27350928
F	98874	27215472	Tot.	911022	241016902

A custom data acquisition software program *isbitDirector* was provided as an in-kind contribution to the project by Bion Biometrics, Inc.. In order to make the subject visits more interesting (for both subjects and test operators), the verification protocol programmatically generates a pre-determined fraction - by default, 20% - of imposter matches. These "online" matches are recorded in the database but are not used for the data analysis: instead, a separate "offline" process *isbitGrinder* was used to extract and crossmatch the desired sets of verification images and enrolment templates.

Twelve different commercially-available fingerprint sensor devices were initially present in the study, representing multiple vendors and technologies: single-spectral optical, multi-spectral optical and capacitive. One device became unreliable in the later phases, and was dropped from the capture protocol. One further device became unavailable due to software licensing restrictions and was removed from the study altogether (Table 3.1). To our knowledge, all of the optical sensors are based on frustrated total internal reflection. Ages of the participants at the time of the most recent collection ranged from 15 years to 70 years. In excess of 15,000 ISO/IEC standards-compliant two-finger biometric enrolment templates were generated, and nearly 200,000 bitmapped single-finger verification images were collected: together, these allowed us to synthesize nearly 250 million single-finger match transactions, with approximately 900,000 genuine (same subject, same finger) matches (Table 3.3).

3.2.1 Carleton modifications to the software and protocol

Because the ISBIT study was focused on vendor interoperability, the original *isbit-Grinder* software was written to perform cross-matches between verification images obtained on one vendor's equipment and enrolment templates recorded by another vendor's. For the purpose of this work, the software was modified to eliminate these inter-vendor cross-matches in favour of inter-visit matches.

3.3 Study terminology and notation

In this work, a *visit* comprises a sequence of biometric enrolments and verification attempts conducted on a set of subjects over a short contiguous period (typically 2-4 weeks). Match scores are generated between verification images collected in the n^{th} visit, V_n and biometric templates obtained during the m^{th} visit, E_m (Table 3.2).

Although the - inherited - protocol used in this study is based on a two-finger biometric template, match scores are evaluated separately for each enrolled finger. Hence we define a *match score*, s_{nm}^{ji} between an image of subject-finger j presented during verification visit V_n and the minutia record of subject-finger i recorded during enrolment visit E_m. Although in principle both i and j are composed of a subject identifier (k say) and a finger identifier ($d = 1, 2, \ldots, 10$), in practice the offline match generating software *isbitGrinder* only makes same-finger matches, i.e. barring mis-labelling, genuine scores always correspond to same subject to same subject matches, while imposter scores are between same fingers of different subjects.

It is important to note that whereas the verification protocol takes the form of a two-attempt transaction involving multiple individual presentations of a pair of fingers, all the performance metrics used in this work are based on single-finger match scores. So, for example, classification performance is quantified in terms of *false match rate*, FMR, and *false non-match rate*, FNMR, rather than transactional measures such

as false accept rate (FAR) or false reject rate (FRR).

3.4 Biometric record storage and retrieval

Along with data acquisition and match score generation software, Bion Biometrics, Inc. provided data from the ISBIT phases of the study as a Microsoft SQL Server (MSSQL) database. The database was re-used and updated to include the Norwood phases at Carleton, with schema extensions as required to support the scientific objectives of this work – for example, additional tables for the biometric quality metrics and attempted taxonomies of Chapter 6.

3.4.1 Database structure

The MSSQL database provides the primary interface to the biometric records, collection timestamps and subject demographics whose analysis forms the basis of the present work. Descriptions of some of the more frequently used database tables are provided in Table 3.4. In addition to these, an 'Algorithm' table is used when it is required to map internal (proprietary) device designators to either the pseudonymized identifiers used for the experimental arrangement ("B (1)", "H (2)" etc.) or those used for the purpose of reporting ("Device L", "Device A" and so on).

Although not used in the generation of match scores (which are based on comparisons between VerifyImage and EnrolTemplate records), the images captured during enrolment, from which the template minutiae are extracted, are recorded in table EnrolImage. These are used in the examination of biometric quality in Chapter 6. Quality assessment was performed externally on both enrolment and verification images, and additional tables were created as required to hold the re-imported quality scores.

Sample MSSQL queries are provided in Appendix A.

Table 3.4: More frequently used database tables

Table name	Description
Subject	Contains pseudonymized subject identifier and demographic information (birth year, sex, geographic region of origin, occupational exposure)
Subject_Visit_Map	Records start and end timestamps of each subject visit, plus visit-specific observations such as fingers that are unavailable (due to injury, for example)
EnrolTemplate	Encapsulates an ISO standard two-finger biometric information template and hash, along with numeric identifiers of the primary and secondary fingers enrolled
EnrolOnline	Primary enrolment record: maps an EnrolTemplate to a specific subject, enrolment visit, and algorithm (hardware product)
VerifyImage	Encapsulates a single uncompressed bitmap image presented during a verification attempt and its hash. May contain "online" match information (which is ignored in this work).
VerifyOnline	Primary verification record: maps a VerifyImage to a specific subject, verification visit, and algorithm (hardware product)
MatchPresentation	The (offline generated) biometric match scores, re-imported into the database along with identifying VerifyImageId and EnrolOnlineId keys

3.4.2 Binary Biometric Information Record (BIR) structure

The biometric templates obtained during this study are stored in the database as *varbinary* typed Binary Large OBjects (or BLOBs), in the cross-vendor format developed for the original ILO ISBIT study [37].

For certain parts of this work, it was necessary to unpack the binary Biometric Information Record (BIR) structures. The API is described fully in [38], Annex B: each record consists of a 16 byte, little-endian header, followed by a big-endian data segment consisting of up to 52 5-byte minutia records for each of two enrolled fingers, preceded by a 22-byte record header. A collection of Python routines was implemented to unpack and manipulate these BIR records, a selection of which are included in Appendix B.

The unpacked minutia records are also valuable as an aid to visualization of biometric features (Figure 3.2).

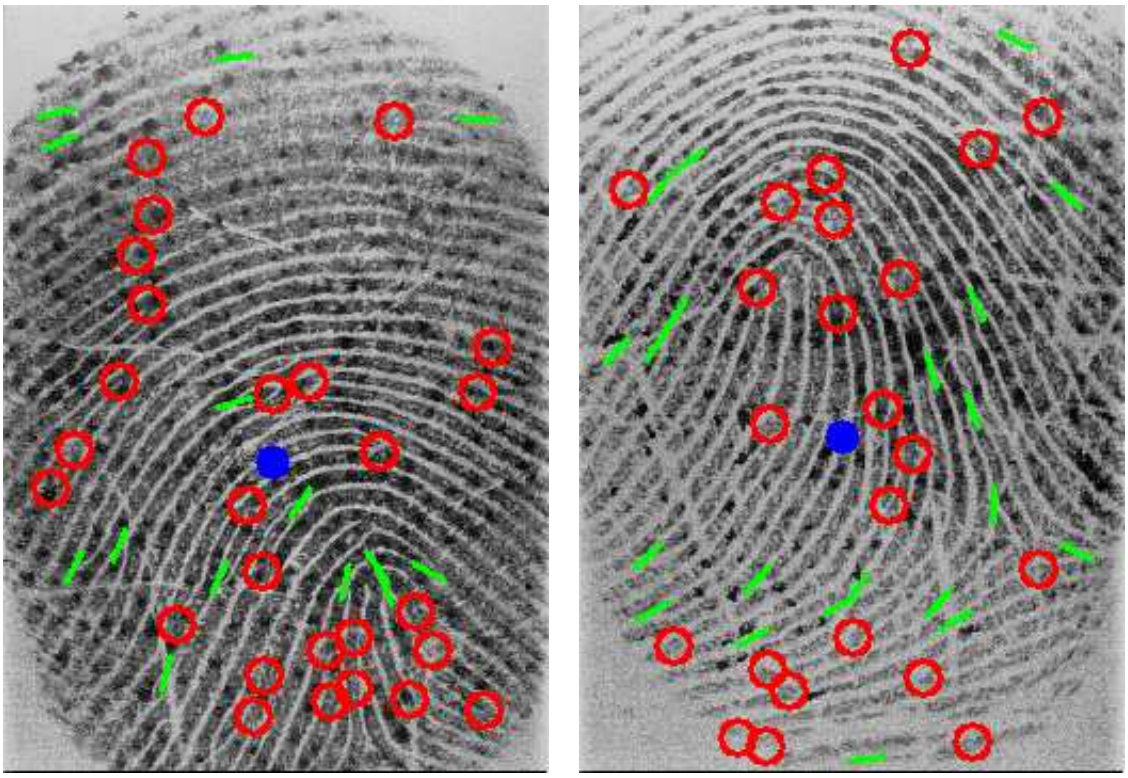


Figure 3.2: The author's own left and right index fingers, annotated with the minutiae obtained during enrolment on Device L

Chapter 4

Biometric permanence: definition and robust calculation

4.1 Introduction

Biometric systems allow identification of people based on analysis of images of their biometric features [41]. When a biometric is used for verification, a biometric sample image is tested against a previously captured sample from the person to be verified [53]. In verification, the performance of the system is measured in terms of its Type-I and Type-II error rates. One key criterion for a biometric modality is the stability of the underlying features. For example, for fingerprint recognition, the structure of the friction ridges is considered to be a unique and stable characteristic of each individual [53].

However, it is widely known [82, 22, 55] that, for any biometric modality, some degree of variation in the biometric features occurs over time. An example is the damage that can occur to fingerprints, which is more common in certain population groups and occupations [18]. Variation in biometric features over time, known as *template ageing*, results in a decrease in biometric recognition accuracy over time [40].

The importance of template ageing varies across different applications of biometrics. It is especially significant for many government programs, such as border security, in which stored templates are intended to be used for comparison over years or decades. There have been several studies done in assessing or describing the impact of template ageing [64, 77]. However, many of these studies have very small datasets (in terms of sample sizes and time periods). Several challenges associated with permanence were identified, including those associated with specific occupations and other environmental factors [20, 67]. Although biometric permanence has seen some investigation, to our knowledge no statistical measures have been defined to measure it or calculate it robustly. To address this deficiency, we define a new term, *biometric permanence*, P_B, and develop methods to calculate it. P_B has an inverse significance to template ageing: a biometric modality with high P_B shows little change over time.

We first propose a definition for *biometric permanence,* P_B, and a reference method to calculate it based on a traditional detection error trade-off (DET) analysis. Next, we consider how to measure P_B robustly for a given biometric modality. A sample population is recruited and biometric measures are performed at intervals over time (Δt), from which a complete set of cross comparisons is calculated [37]. When calculating P_B, the major difficulty in analysing these data arises in separating the visit-dependent factors from the Δt values, which are of course implicitly dependent on the absolute times of the visits. Since the effects of ageing can be small, the evaluation of changes is highly sensitive to estimation variance. To address this issue, we propose a strategy to improve the measure, which we call the *matched delta* method.

4.2 Methodology

An overview of the matched delta method is presented in Figure 4.1. We base the test protocol on that of [37]: in this Chapter however, we synthesize appropriate match

scores as described in Section 4.3. No actual human subject data is involved.

To calculate P_B under this protocol, data are required from a test crew of subjects who are biometrically tested over time at a series of visits. At each visit, k, enrolment E_k and verification V_k biometric samples are acquired. When biometric comparisons are made, match scores are calculated and assigned to a bin corresponding to the time difference (Δt) between visits. Thus, a comparison between E_m and V_n would be in bin Δt_{nm}. The highest match scores should be those from the same visit – during which no changes due to template ageing occur. Thus, comparisons of E_m to V_m have $\Delta t_{mm} = 0$. Given this set of biometric data, Figure 4.1 shows how DET curves from the match scores in each bin are calculated. At a selected value of false match ratio (FMR), the false non-match ratio (FNMR) is calculated for $\Delta t = 0$ and compared to that for a chosen value of Δt, from which P_B is calculated. We do not impose criteria on the selection of FMR; however, it should be chosen at some operationally meaningful level.

4.2.1 Definition

Given these data, we define *biometric permanence*, $P_B(\Delta t)$, for a given elapsed time Δt, as follows:

$$P_B(\Delta t, \mathrm{FMR}) = \frac{1 - \mathrm{FNMR}_{\Delta t}}{1 - \mathrm{FNMR}_0} \tag{4.1}$$

where $\mathrm{FNMR}_{\Delta t}$ is calculated from match scores in the Δt bin, and the "base" level, FNMR_0 is calculated from scores based on data captured during the same visit (i.e. $\Delta t = 0$). Some features of this formulation are:

- $P_B \to 1$ as $\mathrm{FNMR}_{\Delta t} \to \mathrm{FNMR}_0$ i.e. if there is no increase in FNMR over time, then the permanence is high;
- P_B decreases as $\mathrm{FNMR}_{\Delta t} \to 1$ i.e. as FNMR increases over time, permanence decreases.

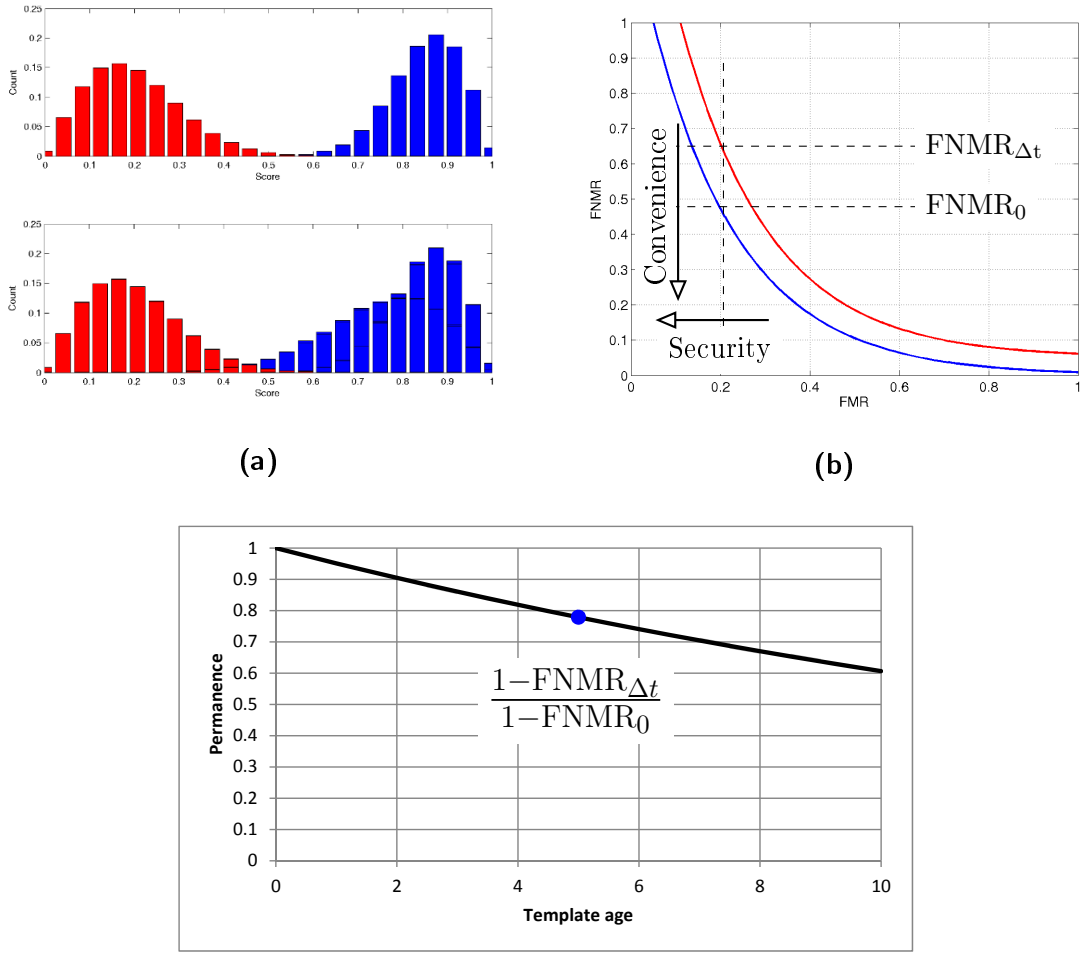


Figure 4.1: Overview of the method: (a) empirical match score distributions immediately after enrolment (top), and after some time interval (bottom); (b) change in classification accuracy represented on a decision error trade-off (DET) curve: arrows indicate the directions of increasing security and convenience; (c) permanence P_B derived from the change in FNMR at fixed FMR according to Eq. 4.1.

In the pathological case where $\text{FNMR}_{\Delta t} < \text{FNMR}_0$, P_B will be greater than 1.

4.2.2 Robust calculation

Given the above definition, it would appear relatively straightforward to calculate biometric permanence from a set of repeated biometric captures. Unfortunately, robust calculation of P_B is complicated. Primarily, the issue is that the effects of interest (small changes in the biometric features) occur in the context of many other changes which are difficult to control experimentally.

For example, in a longitudinal study over several years, there are changes in:

- *weather*: tests at different times of the year expose subjects to – not yet well understood – physiological changes which affect biometric performance (e.g. levels of skin dryness) [71].

- *test administrators*: over a period of several years there is inevitably some turnover in test staff. Not all staff are equally well trained. Some will be more attentive in ensuring proper positioning and placement during biometric tests than others [5].

- *test administrator training*: Even it it were possible to eliminate turn-over of staff, the training level of test staff will adapt over time as they become more familiar with the procedure.

- *ageing of the biometric sensors*: Biometric sensors are typically built of consumer grade electronics and not intended for many years of useful life. Degradation of some components in the sensors (e.g. lighting) can occur.

We address these issues with the matched delta methodology proposed in this section. In overview, match score data are used to estimate the visit-specific factors (which incorporate the variability above) and to separate them from the changes in match

scores caused by template ageing effects. These visit-specific factors may then be removed, leaving only the effect of template ageing.

Since we collect both enrolment templates and verification images during each visit, we can match all of the enrolment templates against all of the verification images and then visualize the available single-finger match transactions as an $N \times N$ matrix (N the number of visits in the study) in which the upper triangle elements are the 'forward time' matches and the lower triangle are the 'reverse time' matches. Along the diagonal are the *baseline* ($\Delta t = 0$) scores in which each finger image is matched against a template taken only a few minutes before, i.e. at $\Delta t = 0$. The essence of our proposed methodology is that we can substantially remove the per-visit score biases by looking at the difference in scores between a suitably chosen combination of visits and the corresponding baseline visits, and applying these to a composite distribution of the averaged baseline scores.

Our "matched delta" methodology is motivated by a simple additive model for the measurement errors in the similarity scores. In the following section, a *biometric presentation* refers to a single, fixed resolution, uncompressed bitmapped image of a fingerprint, while a *template* refers to a record of fingerprint minutiae types and locations extracted during subject enrolment, as described in [37, 38]. We assume there is some true score s_{nm}^{ji} between biometric presentation j in the n^{th} verification visit, and a template i from the m^{th} enrolment visit. In the context of fingerprints, i and j index a specific finger of a specific subject; $j = i$ therefore correspond to genuine matches, and $j \neq i$ to imposter matches. Then we postulate the following error terms:

- a pair of *visit biases* a_m, b_n representing systematic differences in the conditions of the data collections such as operator training, subject acclimation, humidity and so on;

- a stochastic term W^{ji} representing the natural variability between repeated

presentations of the same biometric.

Without loss of generality we can choose the W^{ji} to be zero-mean. In our protocol, we collect six images (in two contiguous verification attempts, each consisting of three presentations) and their averaged scores may then be modelled as

$$\overline{s}^{ji}_{nm} = s^{ji}_{nm} + a_m + b_n + \overline{W}^{ji} \tag{4.2}$$

This presentation averaging step is not essential to the methodology that follows; however it is expected to reduce the variance of the stochastic error term. We then observe that, in our experimental protocol, both enrolment templates and verification images are obtained from the same subject cohort at each visit. This allows us to evaluate the average difference, *forward and backward in time*, between the match score of biometric presentation j against template i with template age $|\Delta t_{nm}|$, relative to the average score at $\Delta t_{nn} = \Delta t_{mm} = 0$, as

$$\begin{aligned}
\Delta\overline{s}^{ji}_{nm}\left(a_m, b_n, W_{ij}; \Delta t_{ij}\right) &= \frac{1}{2}\Big(\overline{s}^{ji}_{nm} + \overline{s}^{ji}_{mn} - \overline{s}^{ji}_{mm} - \overline{s}^{ji}_{nn}\Big) \\
&= \frac{1}{2}\Big(s^{ji}_{mn} + a_m + b_n + \overline{W}^{ji}_0 \\
&\quad - s^{ji}_{mm} - a_m - b_m - \overline{W}^{ji}_1 \\
&\quad + s^{ji}_{nm} + a_n + b_m + \overline{W}^{ji}_2 \\
&\quad - s^{ji}_{nn} - a_n - b_n - \overline{W}^{ji}_3\Big)
\end{aligned}$$

where the $\overline{W}^{ji}_k$ are assumed i.i.d. variables with the distribution of $\overline{W}^{ji}$, i.e.

$$\Delta\overline{s}^{ji}_{nm}\left(W_{ij}; \Delta t_{ij}\right) = \frac{1}{2}\Big\{\left(s^{ji}_{nm} + s^{ji}_{mn}\right) - \left(s^{ji}_{mm} + s^{ji}_{nn}\right) + \sum_{k=0}^{3}(-1)^k \overline{W}^{ji}_k\Big\} \tag{4.3}$$

in which it is seen that the bias terms have been eliminated, leaving just the averages

of the forward and backward true scores and the baseline $\Delta t = 0$ scores for the corresponding visits. Meanwhile the stochastic terms, being uncorrelated, should add on an RMS basis such that the variance of presentation-averaged scores over fingers i, j adds as

$$\mathrm{var}\left(\frac{1}{2}\sum_{k=0}^{3}(-1)^k\overline{W}_k^{ji}\right) = \mathrm{var}\left(\overline{W}_k^{ji}\right) \tag{4.4}$$

leaving the signal-to-noise ratio of the measurement effectively unchanged[1]. The various noise averaging steps may be summarized as:

$$W_k^{ji} \xrightarrow{\text{over presentations}} \overline{W}_k^{ji} \xrightarrow{\text{over forward-backward visits}} \frac{1}{2}\sum_{k=0}^{3}(-1)^k\overline{W}_k^{ji}$$

4.2.3 Visit aggregation

A pervasive difficulty in the evaluation of biometric match performance is the inherent class imbalance. That is, for a set of K subject finger presentations, we have $K(K-1)$ imposter match scores but only K genuine match scores. This means that the genuine match score distributions tend to be less well defined than those of the imposter scores. In the context of the permanence measure Equation 4.1, this means that while it is relatively straightforward to interpolate the threshold corresponding to a chosen FMR, changes in the FNMR at that threshold may represent only a handful of individual genuine matches.

One way to improve the imbalance would be to aggregate scores across multiple devices in our study. Unfortunately however, while the API and biometric template were standardized across device, vendors were free to chose their own scoring schemes - which are hence largely incompatible. In order to improve the estimation, it was

[1] We have four i.i.d. random variables and we are subtracting two of them from the other two: the means subtract (to zero) but the variances add because the variance of $-W$ is the same as the variance of W.

therefore decided to aggregate groups of "short term" match scores into larger sample sets according to the scheme shown in Figure 4.2. Although this procedure implies some loss of temporal resolution, in practice the individual subject visits were typically spread over two or three weeks, so that the distinction between "1 year" and "1 year ± 2 weeks" is not really meaningful anyway.

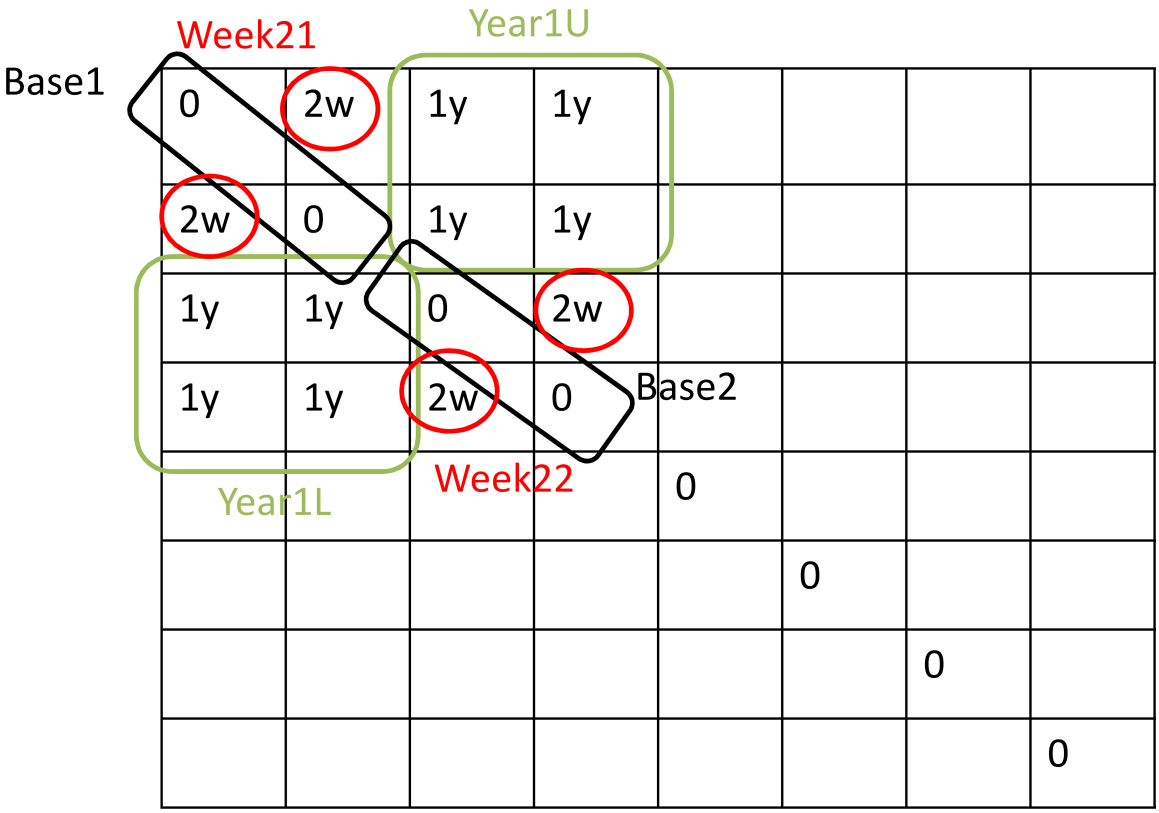


Figure 4.2: Matrix of visits for a single subject over the protocol. Each row and column represents a visit (with enrol, E_m and verification, V_n records). In our testing protocol, each round of testing has a pair of visits separated by two weeks. The upper triangle represents match scores "forward in time" (E_m vs. V_n, $m < n$), while the lower represents the corresponding match scores "backward in time" (E_m vs. V_n, $m > jn$). Match scores on the diagonal are from the same visit (i.e. $\Delta t = 0$).

In Chapter 4 we discard this aggregation approach in favour of a more formal bootstrap resampling procedure that allows estimation of confidence intervals for the permanence values.

4.3 Simulation

We have defined biometric permanence, P_B using, first, a model based on verification calculations alone ("reference methodology", the value calculated this way is $P_{B,R}$); and next, a "matched delta methodology" ($P_{B,M}$) in which visit variability is modelled and removed by calculating the shift in the genuine distribution. Our expectations are:

- In the absence of visit-to-visit variability, $P_{B,M}$ is an unbiased estimate of $P_{B,R}$.
- $P_{B,M}$ is a lower variance estimator of $P_{B,R}$ (since a single parameter is estimated from data before the DET calculation, which is known to be noisy).
- In the presence of visit-to-visit variability, $P_{B,M}$ will show more plausible results than $P_{B,R}$.

We take as our starting point a pair of canonical distributions for the underlying false and genuine match scores, where scores range between 0 and 1. In this work we have chosen to model the imposter scores via a simple Rayleigh distribution, and the genuine scores via a "flipped" Rayleigh distribution (Appendix D). In suitably normalized form the probability densities become

$$p_I(s) = \frac{s}{\beta_I^2} e^{-s^2/2\beta_I^2}; s \geq 0 \tag{4.5a}$$

$$p_G(s) = \frac{1-s}{\beta_G^2} e^{-(1-s)^2/2\beta_G^2}; s \leq 1 \tag{4.5b}$$

where the scale parameters $\beta_{I,G}$ are related to the mean scores by $\mu_I = \beta_I\sqrt{\frac{\pi}{2}}$ and $\mu_G = 1 - \beta_G\sqrt{\frac{\pi}{2}}$.

To model the sequence of enrolment-verification visits, we then make two assumptions, namely

- *The imposter distribution remains constant over time.* Since the imposter dis-

tribution represents non-matched samples, it will not be significantly sensitive to changes due to elapsed time. This is roughly equivalent to the assertion that for any pair of non-identical fingers whose match score is improved by some mechanism, there is another pair whose match score is correspondingly reduced.

- *With each visit k we can associate a pair of systematic biases a_k, b_k (for enrolment and verification, respectively) that affect all presentations equally.* We then model the effect of a match score between a verification presentation from visit n against a template recorded in visit m as a shift in the genuine distribution equal to $(a_m + b_n)$

We then model the template ageing as a further, time-dependent, shift of the genuine distribution. In this work we choose a simple time-symmetric assumption, namely $\delta(\Delta t) = -\alpha(1 - \exp(-\kappa|\Delta t|))$ i.e. a term that asymptotically approaches $\delta = -\alpha$ with time-constant $1/\kappa$.

To each variate generated according to Eq. 4.5 we apply an additional zero-mean Gaussian noise term W^{ji} to represent the natural variation in match score over repeated presentations of the same finger against the same template (effectively a kind of 'measurement noise').

4.3.1 Simulation of a single sequence of visits

To make a baseline qualitative evaluation of the efficacy of our method in removing the experimental biases, we constructed a single realization of a sequence of eight visits, with bias values a_k, b_k taken from a Gaussian distribution with standard deviation 0.025. We then generated canonical sample distributions according to Eq. 4.5 for each of two fingers for a sample size of 17500 subjects

with $\mu_I = 0.2$ and $\mu_G = 0.85$ for each finger. Next, for each of the $8 \times 8 = 64$

enrol-verify visit combinations, we modify the finger scores according to:

$$s_{n,m}^{j,i}(\Delta t) = s_0 + a_m + b_n - \alpha(1 - \exp(-\kappa|\Delta t|)) \tag{4.6}$$

Finally, we synthesize six independent presentations of each finger by adding a zero-mean Gaussian noise term with variance $\sigma^2 = w^2/6$: this scaling is a convenience, so that we can identify w^2 with the sample variance of the presentation-averaged experimental scores. We then processed the scores in two ways: (i) a simple direct calculation according to Section 4.2.1; and (ii) our matched delta method as described in Section 4.2.2.

We ran the simulation with three scenarios: first, with no visit biases ($b_n = a_m = 0$); second with bias in the creation of enrol templates only ($b_n = 0; a_m \neq 0$); and third with bias in both enrol template creation and verification presentation ($b_n \neq 0, a_m \neq 0$). Results of these simulations are illustrated in Figure 4.3: each sub-figure shows P_B (vertical axis) versus elapsed time (horizontal axis) in weeks between enrolment and verification. The {ON,OFF} status of parameters a (enrolment) and b (verification) represents whether the corresponding biases are simulated or not. Thus "a:ON b:ON" indicates simulation of both enrolment and verification biases. w is the standard deviation of simulated presentation noise. The value of the presentation-averaged sample standard deviation w was varied in the range 0 to 0.075 units and the ageing parameters were $\alpha = 0.1$ unit and $\kappa = 0.01$ week^{-1}. The reference FMR for the permanence calculation was 0.001 (0.1%).

Taken together, the visit bias terms and ageing correspond to a modification $(1 - s) \rightarrow (1 - a_m - b_n - \delta(t) - s)$ in the genuine score distribution of Eq. 4.5. The densities then become convolved with the Gaussian density — one can show (Appendix D) that the resulting tail integrals for the FMR and FNMR at some

threshold score θ become:

$$FMR(\theta) = \frac{\beta_I}{\beta'_I} e^{-\theta^2/2\beta'^2_I} \tag{4.7a}$$

$$FNMR(\theta; m, n; t) = \frac{\beta_G}{\beta'_G} e^{-(\chi(t;m,n)-\theta)^2/2\beta'^2_G} \tag{4.7b}$$

where $\beta'^2_{I,G} = \beta^2_{I,G} + \sigma^2_{I,G}$ (with $\sigma^2_{I,G}$ being the variances of the imposter and genuine presentation noise terms respectively) and

$$\chi(t; m, n) = 1 - a_m - b_n - \delta(t) \tag{4.8}$$

represents the mean degradation in genuine match score due to the per-visit biases and template ageing. The solid black curves in Figure 4.3 correspond to Eq. 4.7 , 4.8 with $a_m = b_n = 0$ i.e. they represent the 'ideal' (unbiased) ageing behaviour that would be observed due to presentation noise only: the deviation from this curve can be interpreted as the residual effect of bias that is not removed by the technique.

4.3.2 Simulation of an ensemble of visit sequences

Over an ensemble of experiments (that is, sequences of enrol-verify visits) with experimental biases taken independently from some zero-mean distribution(s), one would like to show that the mean of the reference permanence measure (Eq. 4.1) does indeed converge to the value obtained by our new technique. Accordingly we ran the same procedure up to 180 times to simulate multiple independent realizations of our experiment. At intervals of 5 simulated experiments we calculated the mean deviation (across time differences Δt_{nm}) between the permanence calculated according to our method (Section 4.2.2) and reference method (Section 4.2.1) (Figure 4.4).

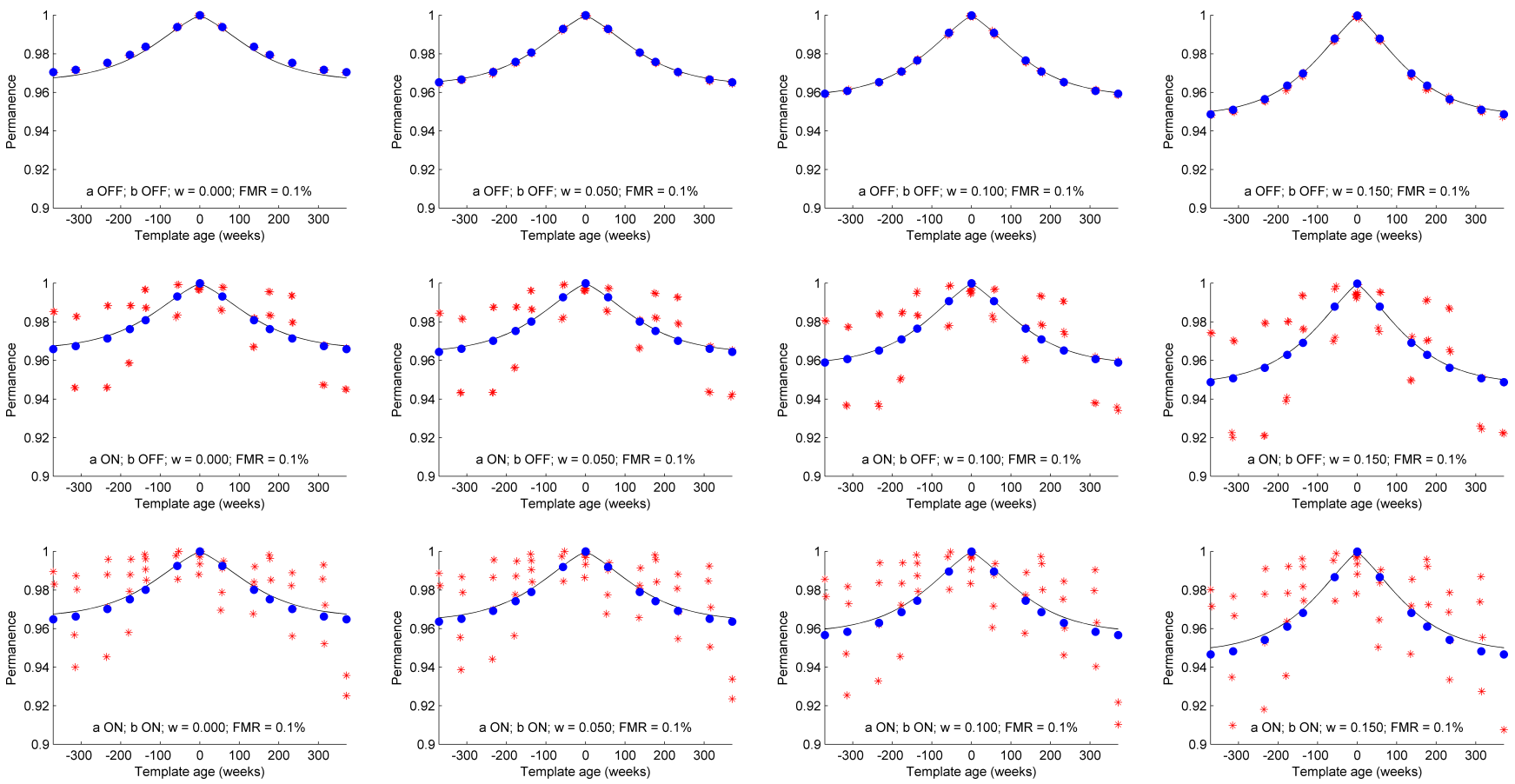


Figure 4.3: Simulation of the effect of visit biases. *Top row*: noise only; *middle row*: noise and enrol visit biases; *bottom row*: noise, enrol and verify bias. The case of noise plus verify bias only is omitted for brevity. Red stars are the reference method of Section 4.2.1 while blue circles are our matched delta method of Section 4.2.2. The black curve is derived from the analytical tail integrals of Eq. 4.7.

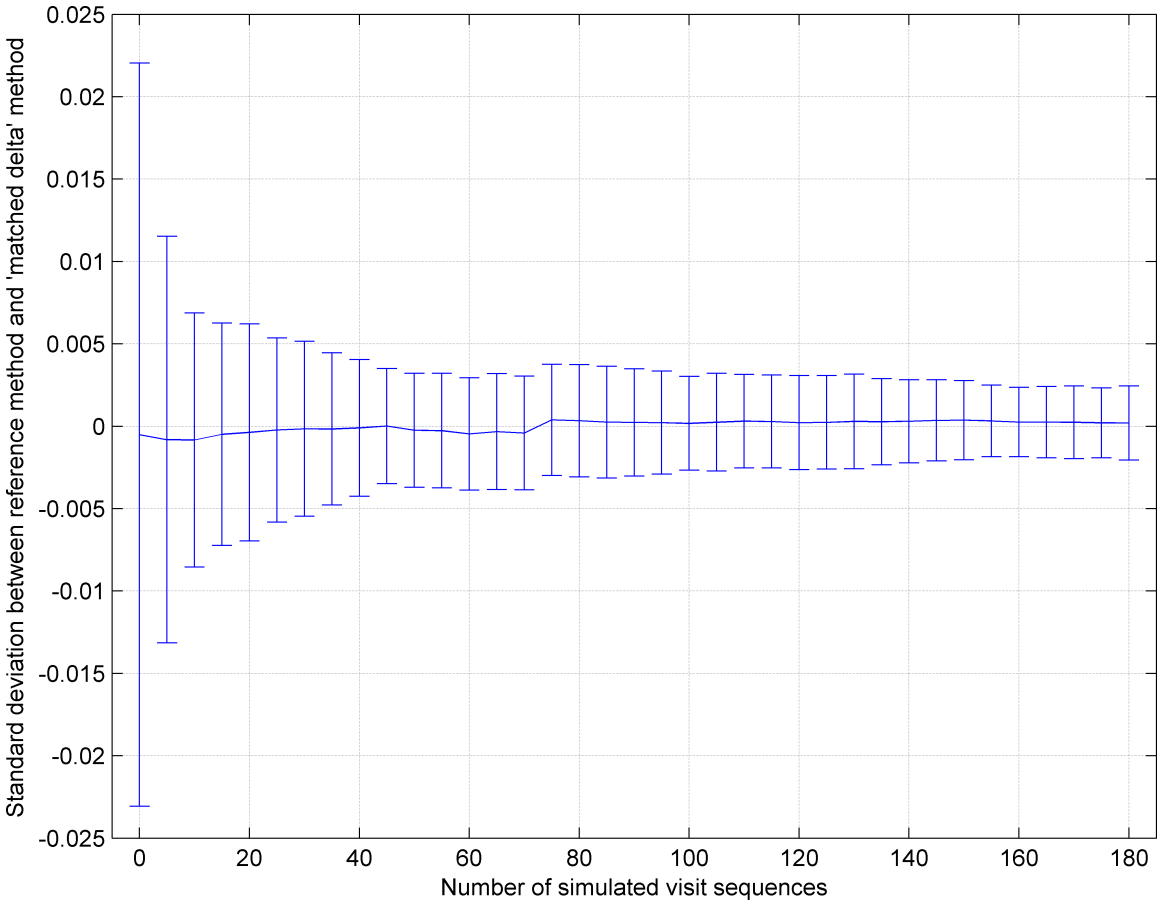


Figure 4.4: Difference (±SD) between the reference measure Section 4.2.1 and the matched delta method of Section 4.2.2 as the size of the experiment ensemble is increased.

4.4 Discussion

We have developed and defined a measure of template ageing which we call biometric permanence P_B, based on the change in FNMR (at a given FMR) between the template ageing interval under test, and a short-time test. While intuitive, this definition of P_B is practically difficult to apply to estimate small changes in permanence in a longitudinal study subject to experimental error and visit-to-visit systematic biases. To address this issue, we have introduced the matched delta method. Comparisons of these methods were performed using simulated data, and it was determined that the new method showed dramatically reduced sensitivity to systematic biases. Simulations were designed to evaluate two aspects of the proposed robust calculation method in comparison to the calculation of P_B from Eq. 4.1. First, simulations test the first and second order statistical properties (i.e. bias and variance); and, second, the sensitivity of the methods to visit-to-visit biases.

Figure 4.3 compares the two methods to the analytical values (Eq. 4.7). For two different values of presentation noise (columns), the presence or absence of visit biases is evaluated. A single sample of visit biases is evaluated in each case; multiple values at a single time interval indicate different ways in which the given time offset can be calculated. Without visit biases, the methods perform similarly, while their presence dramatically impacts the values calculated using Eq. 4.1. Meanwhile Figure 4.4 investigates the statistical properties of the methods. As sample number increases, the variance decreases and bias between methods decreases towards zero, as expected from our analytical model.

Chapter 5

Characterization of biometric template ageing in a multi-year, multi-vendor longitudinal fingerprint matching study

In Chapter 4 we outlined a method for evaluating the permanence of a set of biometric templates, based on a simple phenomenological model for confounding factors resulting from changes in physical environment and acclimation.

The goal of this chapter is to characterize template ageing in the fingerprint modality, for a number of commercially-available fingerprint sensor devices and technologies, and to understand its impact on the deployment and operation of fingerprint-based IDMSs.

In Section 5.2, results are presented for each of the devices. Finally, in Section 5.3 we attempt to justify, through further data analysis, the key assumptions underlying the methodology.

5.1 Methodology

The methodology follows that of the previous chapter, with the following exception.

In Chapter 4, we noted that aside from the visit-to-visit confounding factors such as variations in environmental conditions, operator training and subject acclimation, the chief difficulty in estimating template permanence according to our definition of Equation 4.1 is the relatively poor definition of the empirical genuine score distributions. This in turn results from the inherent class imbalance i.e. that whereas the number of imposter matches increases quadratically with the number of individuals in the study, that of the genuine scores only increases linearly. For the simulated study of that chapter, we implemented a "visit aggregation" scheme to boost the sizes of the genuine match score sets. Here we develop a more robust scheme based on bootstrap resampling [56].

As in our earlier procedure, the averaged "matched deltas" $\Delta\overline{s}^{ji}_{nm}$ from Equation 4.3 are averaged again across a particular pair of enrolment and verification visits m, n to give mean genuine and imposter score offsets $\Delta\overline{s}^{G}_{nm}$ and $\Delta\overline{s}^{I}_{nm}$ for the chosen visit pair n, m. We then aggregate the corresponding zero-time genuine and imposter scores $\{\overline{s}^{ii}_{kk}\}, \{\overline{s}^{j\neq i}_{kk}\}; k \in 1 \ldots N$ and use these aggregate distributions shifted by the respective mean offsets $\Delta\overline{s}^{G}_{nm}$, $\Delta\overline{s}^{I}_{nm}$ to evaluate P_B according to Equation 4.1 at time interval Δt_{mn}. In order to perform the bootstrap resampling, we arrange the aggregate genuine and imposter scores into a vector $\left(\overline{s}^{ii}_{kk}, \overline{s}^{j\neq i}_{kk}\right)$ along with a vector of class labels $(\mathbb{1}_{n_G}, \mathbb{0}_{n_I})$ where n_G, n_I are the genuine and imposter class sizes in the sample. The bootstrap procedure was then developed in a number of stages, as follows.

The initial implementation consisted of a simple N-fold resampling, with replacement, of the entire labelled datasets i.e.

- select $n\mathrm{Tot} = n_G + n_I$ scores, randomly with replacement

- construct a DET curve for the sample
- interpolate the DET curve to find FNMR at the chosen reference FMR
- repeat N times to obtain a 95% confidence interval (CI) for the FNMR

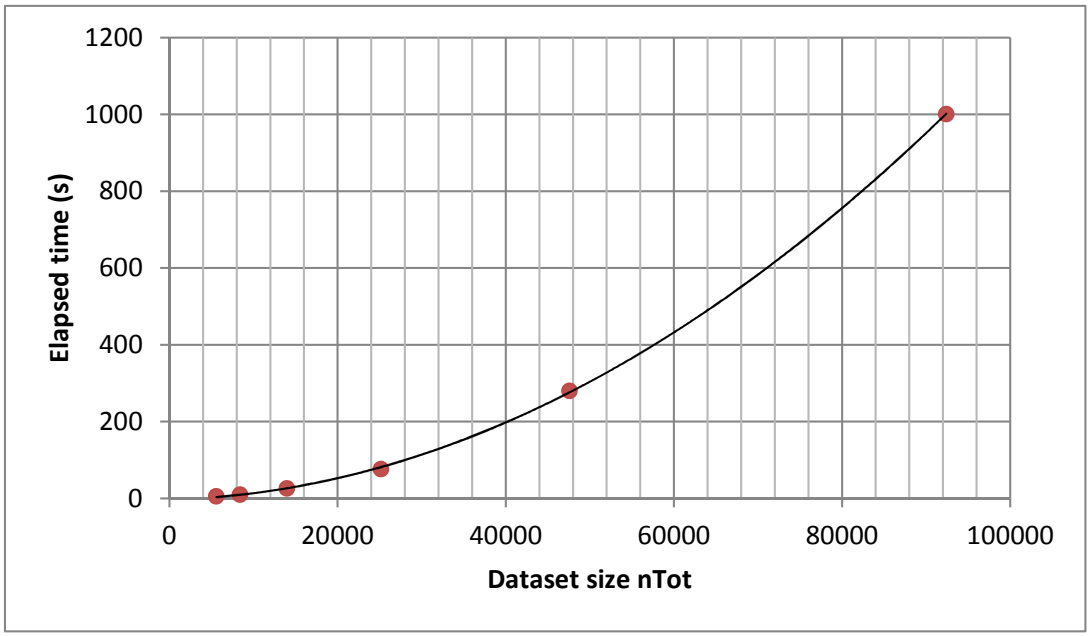


Figure 5.1: Computation time of the bootstrap CI versus dataset size, $n_{Tot} = n_G + n_I$ for the original procedure

This procedure is, however, computationally very inefficient, scaling poorly with dataset size (Figure 5.1) – there really is no benefit in resampling the imposter scores, since their distribution is already sufficiently well defined.

A refined procedure was then implemented in which a single DET curve was constructed using the full dataset, which was then interpolated to find a decision threshold θ_{Ref} corresponding to the chosen reference FMR. The imposter set was then sub-sampled to the same class size as the genuine set, $n_I = n_G$, and N-fold resampling implemented in the same manner as before in order to obtain a 95% CI for the FNMR at θ_{Ref}.

Although more efficient, a significant flaw remains in that the construction leads to a confidence interval for the FNMR, rather than for the permanence measure, P_B. To address this, an improved bootstrap procedure was developed. First, Equation 4.1 was re-cast in terms of true match rate (TMR) as

$$P_B(\Delta t, \text{FMR}) = \frac{\text{TMR}_{\Delta t}}{\text{TMR}_0} \tag{5.1}$$

Next we noted that, since an offset $\Delta\overline{S}^I$ to the imposter distribution is exactly equivalent to a shift in the threshold $\theta \to \theta + \Delta\overline{S}^I$ for the chosen FMR, while an offset to the genuine distribution is similarly equivalent to a shift $\theta \to \theta - \Delta\overline{S}^G$, we just need to evaluate $1 - \text{FNMR}$ (or, equivalently, the true match rate TMR) at a set of thresholds $\theta_{nm} = \hat{\theta}_0 + \Delta\overline{s}^I_{nm} - \Delta\overline{s}^G_{nm}$ (Fig. 5.2). In fact, since we defined P_B as a ratio, it suffices to work with the raw genuine score counts i.e. the permanence is estimated for each bootstrap sample as

$$\hat{P}_B(\Delta t_{nm}) = \frac{\left|\{\overline{s}^{ii}_{kk} : \overline{s}^{ii}_{kk} > \hat{\theta}_0 + \Delta\overline{s}^I_{nm} - \Delta\overline{s}^G_{nm}\}\right|}{\left|\{\overline{s}^{ii}_{kk} : \overline{s}^{ii}_{kk} > \hat{\theta}_0\}\right|} \tag{5.2}$$

evaluated for each enrolment-verification visit pair n, m, where $|C|$ denotes the cardinality of C.

The procedure was developed in MATLAB using the *perfcurve* function from the Statistics and Machine Learning Toolbox, with the permanence measure of Equation 5.1 as a parametrized *bootfun*. The resampling was class-weighted in inverse proportion to the original class sizes n_G and n_I in order to remove class imbalance.

The interested reader is referred to the MATLAB code in Appendix C for details of the implementation.

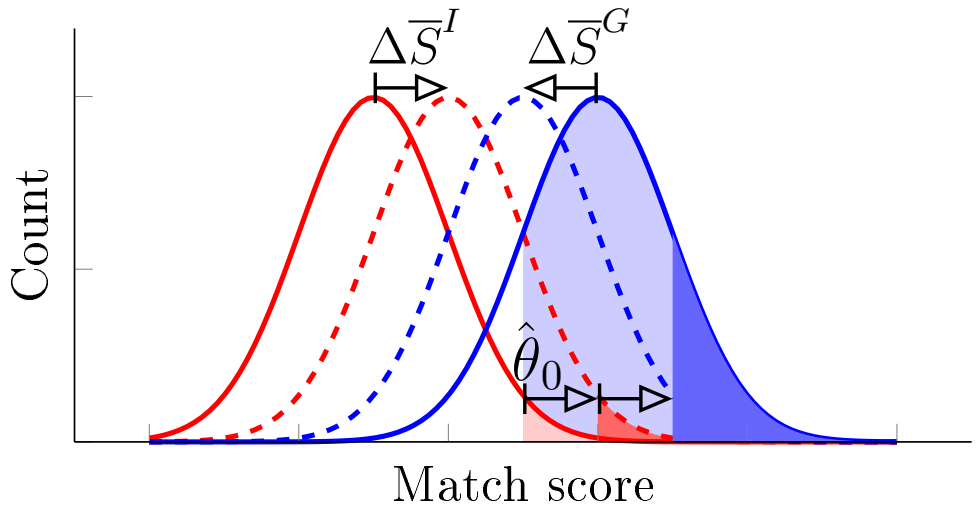


Figure 5.2: A shift in the mean imposter score results in a shift in the estimated decision threshold $\hat{\theta}_0$ for a specified FMR (red area) – and a corresponding change in the achievable TMR (blue area) for the shifted genuine scores.

5.2 Results

Results of this procedure are shown graphically in Figs. 5.3 – 5.11, with comparison to a "naïve" evaluation that does not attempt to account for visit bias.

The histograms are scaled to account for the large class imbalance between genuine and imposter scores. DET curves are generated using the "matched delta" methodology described in the text. The permanence results demonstrate the reduction in the confounding effect of visit biases due to our method; error bars correspond to the 95% bootstrap confidence intervals described in the text with bootstrap resampling factor $N = 1000$.

The solid lines in the permanence figures represent simple best fits to the data and are intended only as an aid to visualization. In particular, they may appear to suggest that P_B does not achieve a value of 1 at $\Delta t = 0$: in fact it does in all cases (as it must, given the definition of Equation 4.1) and what we actually observe is a steep drop in P_B between $\Delta t = 0$ and $\Delta t = \pm 2$ weeks, followed by a much more gradual

Table 5.1: Estimated 95% confidence intervals for permanence, P_B after 7 years, by device.

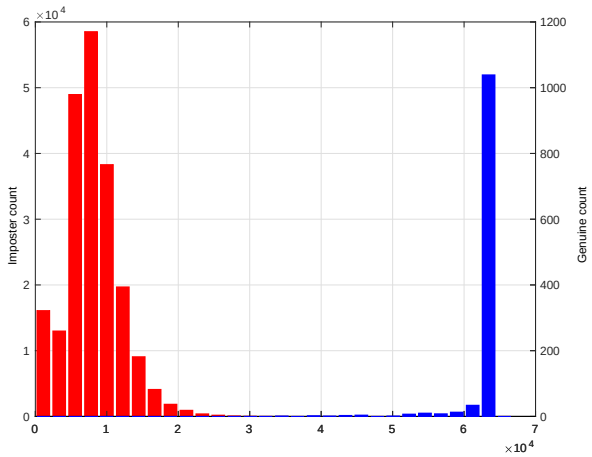
ID	Permanence, P_B (%)	ID	Permanence, P_B (%)
A.	92.4 ± 0.33	G.	95.9 ± 0.38
B.	100	H.	99.5 ± 0.12
C.	98.3 ± 0.24	J.	100
D.	96.1 ± 0.27	K.	97.2 ± 0.19
F.	98.6 ± 0.08	L.	95.5 ± 0.23

decline in permanence over the remaining duration of the study.

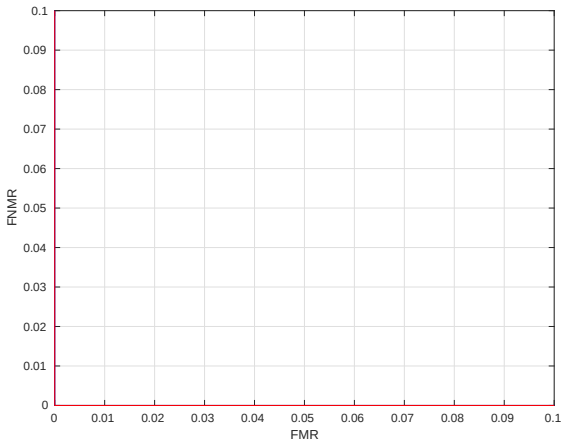
In Figs. 5.10 and 5.11 we see the evolution of the typical observed ageing behaviour of the devices in our study. First, we note that the baseline ($\Delta t = 0$) score distributions Fig. 5.10a, Fig. 5.11a are not separable; that is, there is no choice of binary threshold for which the probability of misclassification may be made arbitrarily small. Correspondingly, the decision error trade-off (DET) curves Figs. 5.10b, 5.11b are displaced from $(0, 0)$ at $\Delta t = 0$ (blue curve) and become further displaced as the template ages (red curve), indicating an increased misclassification probability. Finally in Figs. 5.10c, 5.11c we see the permanence P_B according to Equation 4.1 decrease monotonically away from template age $\Delta t = 0$.

Two of the available devices (B and J) did not show this typical behaviour. Instead, they showed well-separated genuine and imposter score distributions at $\Delta t = 0$ (Fig. 5.3a) which essentially remained separable over the whole duration of the study. Hence we see both $\Delta t = 0$ (blue) and $\Delta t = 7$years (red) DET curves achieving FNMR $= 0$ at FNMR $= 0$ (Fig. 5.3b) and correspondingly no discernable change in permanence P_B in Fig. 5.3c.

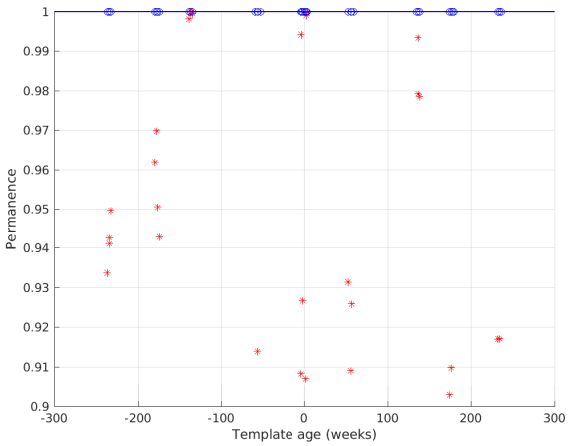
Results for all the available devices in our study are summarized in Table 5.1.



(a) Genuine (blue) and Imposter (red) score counts at $\Delta t = 0$

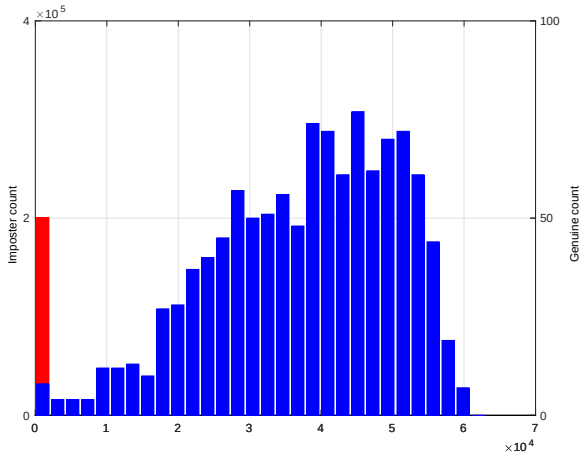


(b) DET curve at $\Delta t = 0$ (blue - hidden) and at $\Delta t = 373$ weeks (red)

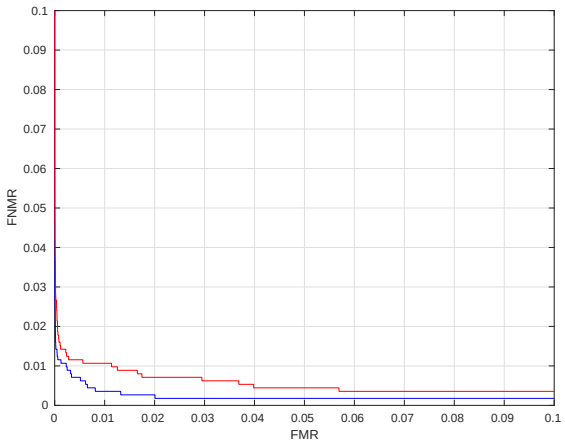


(c) Permanence vs. template age: naïve calculation (red); present method (blue)

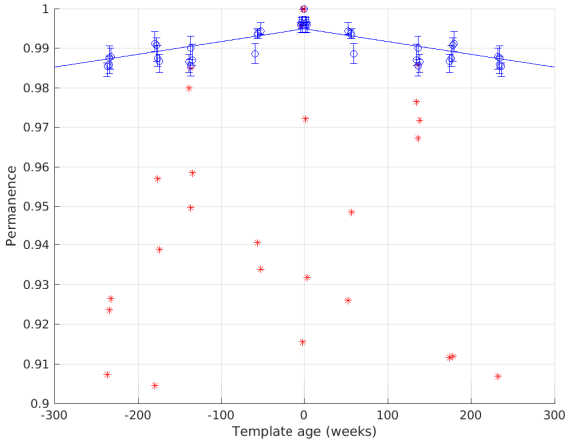
Figure 5.3: Match score distributions, DET, and P_B: Device B (optical)



(a) Genuine (blue) and Imposter (red) score counts at $\Delta t = 0$

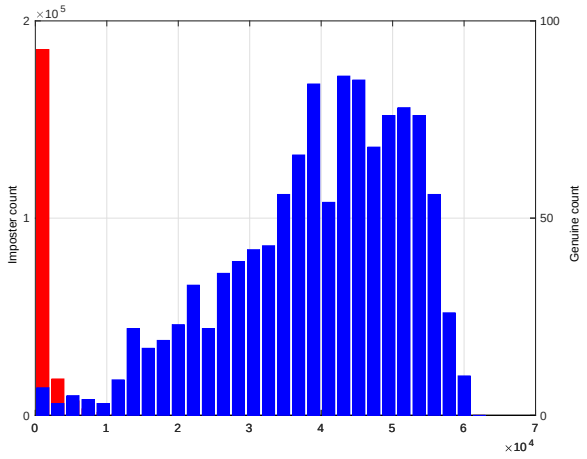


(b) DET curve at $\Delta t = 0$ (blue) and at $\Delta t = 373$ weeks (red)

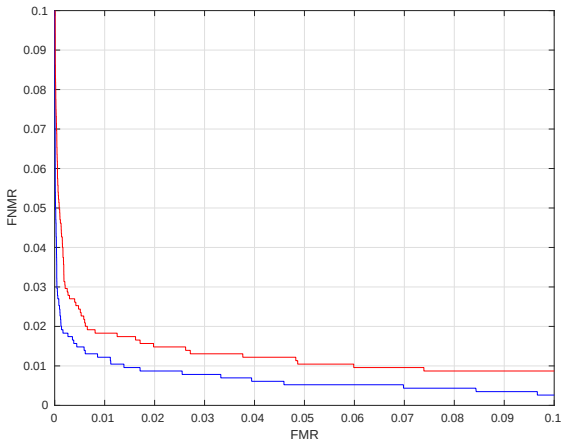


(c) Permanence vs. template age: naïve calculation (red); present method (blue)

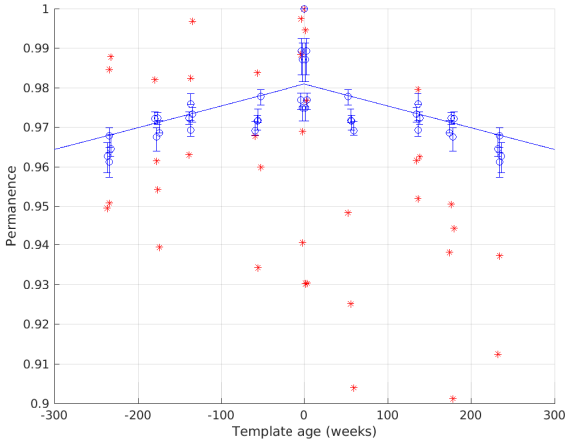
Figure 5.4: Match score distributions, DET, and P_B: Device C (optical)



(a) Genuine (blue) and Imposter (red) score counts at $\Delta t = 0$

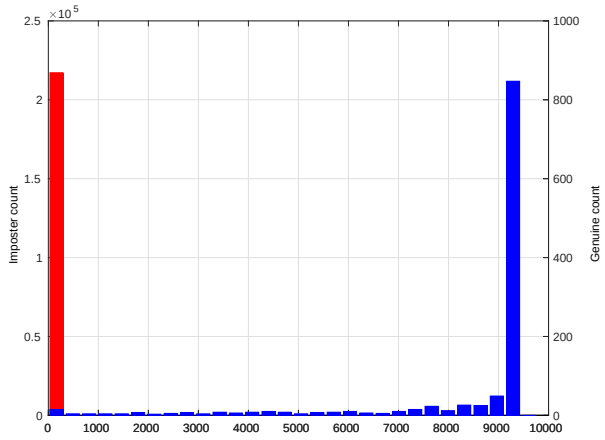


(b) DET curve at $\Delta t = 0$ (blue) and at $\Delta t = 373$ weeks (red)

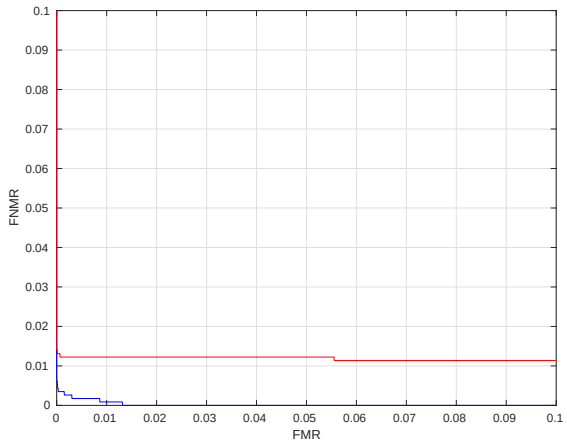


(c) Permanence vs. template age: naïve calculation (red); present method (blue)

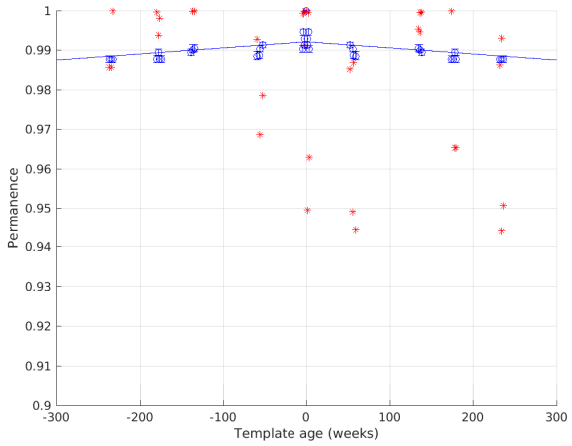
Figure 5.5: Match score distributions, DET, and P_B: Device D (optical)



(a) Genuine (blue) and Imposter (red) score counts at $\Delta t = 0$

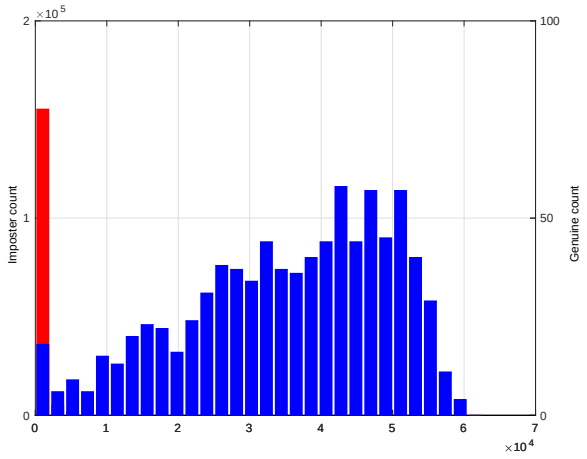


(b) DET curve at $\Delta t = 0$ (blue) and at $\Delta t = 373$ weeks (red)

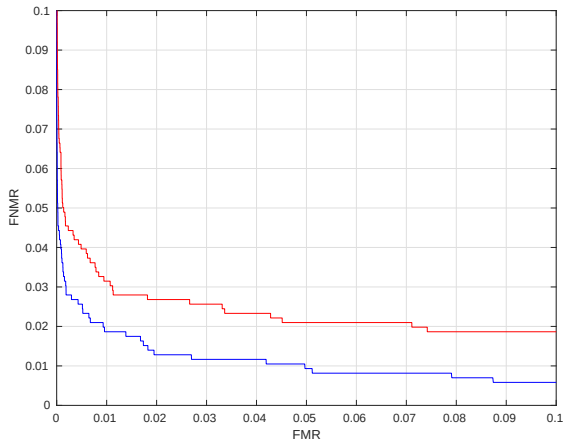


(c) Permanence vs. template age: naïve calculation (red); present method (blue)

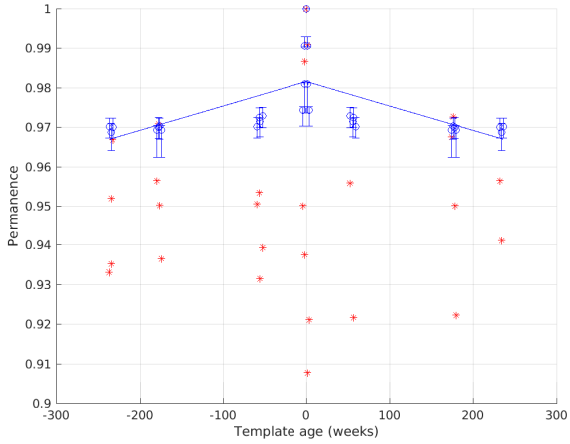
Figure 5.6: Match score distributions, DET, and P_B: Device F (optical)



(a) Genuine (blue) and Imposter (red) score counts at $\Delta t = 0$

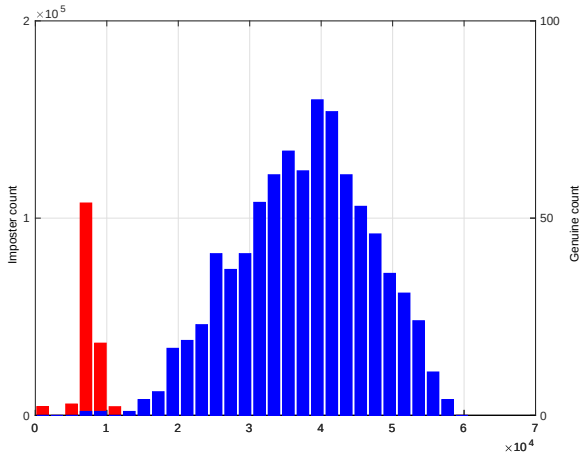


(b) DET curve at $\Delta t = 0$ (blue) and at $\Delta t = 373$ weeks (red)

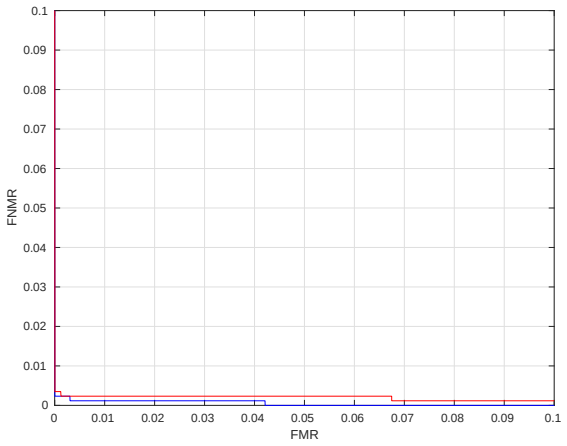


(c) Permanence vs. template age: naïve calculation (red); present method (blue)

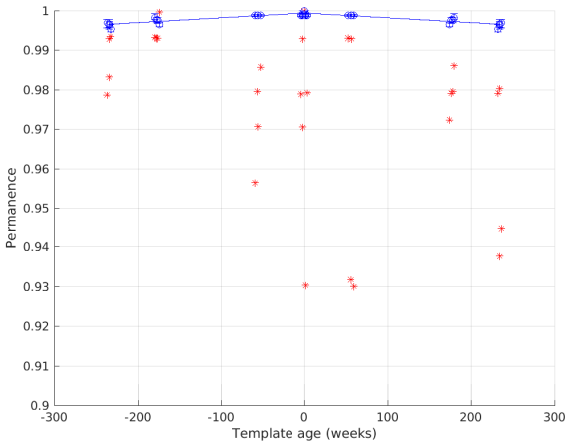
Figure 5.7: Match score distributions, DET, and P_B: Device G (optical)



(a) Genuine (blue) and Imposter (red) score counts at $\Delta t = 0$

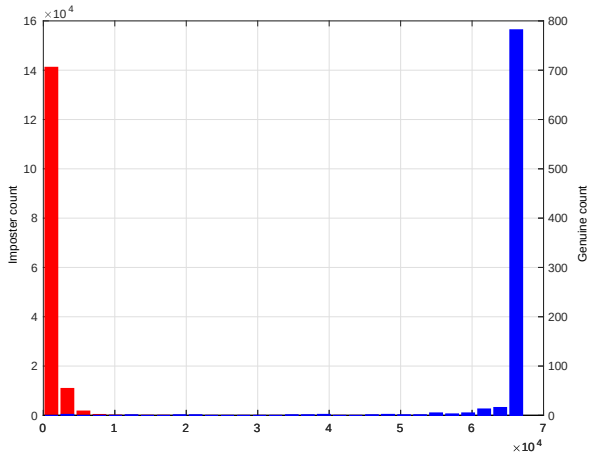


(b) DET curve at $\Delta t = 0$ (blue) and at $\Delta t = 373$ weeks (red)

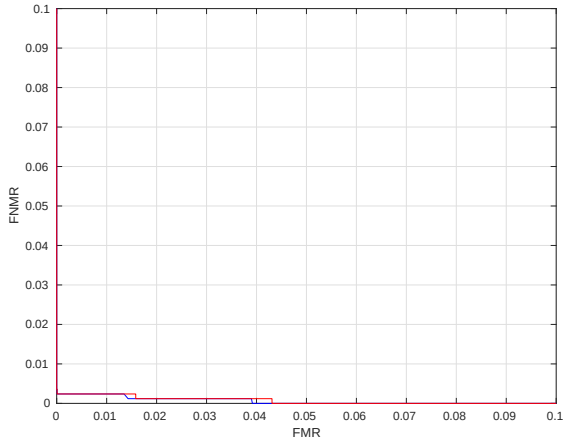


(c) Permanence vs. template age: naïve calculation (red); present method (blue)

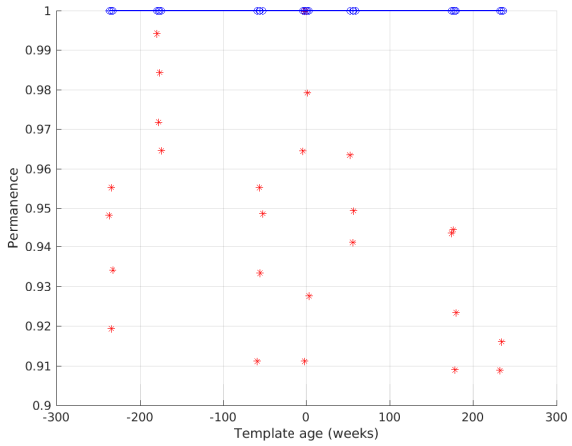
Figure 5.8: Match score distributions, DET, and P_B: Device H (optical)



(a) Genuine (blue) and Imposter (red) score counts at $\Delta t = 0$

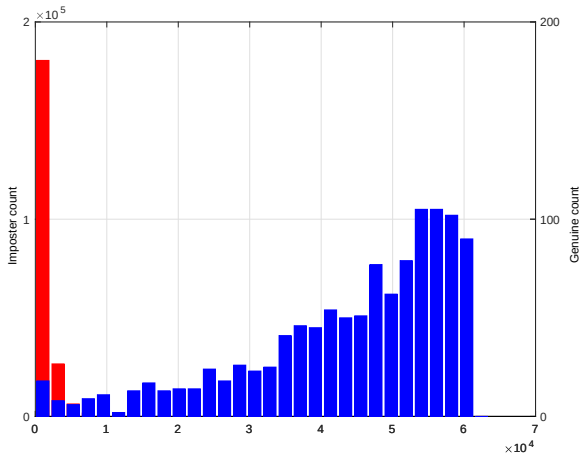


(b) DET curve at $\Delta t = 0$ (blue) and at $\Delta t = 373$ weeks (red)

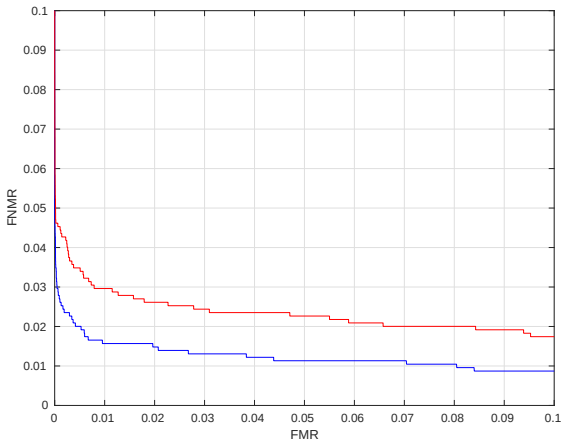


(c) Permanence vs. template age: naïve calculation (red); present method (blue)

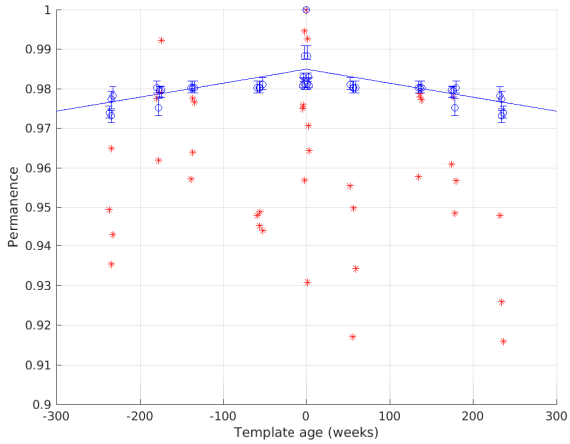
Figure 5.9: Match score distributions, DET, and P_B: Device J (optical)



(a) Genuine (blue) and Imposter (red) score counts at $\Delta t = 0$

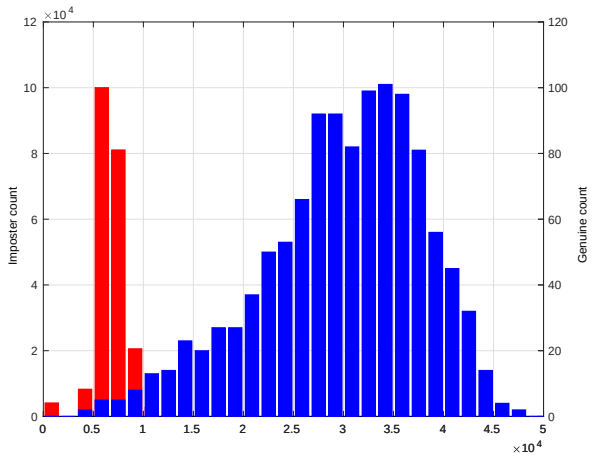


(b) DET curve at $\Delta t = 0$ (blue) and at $\Delta t = 373$ weeks (red)

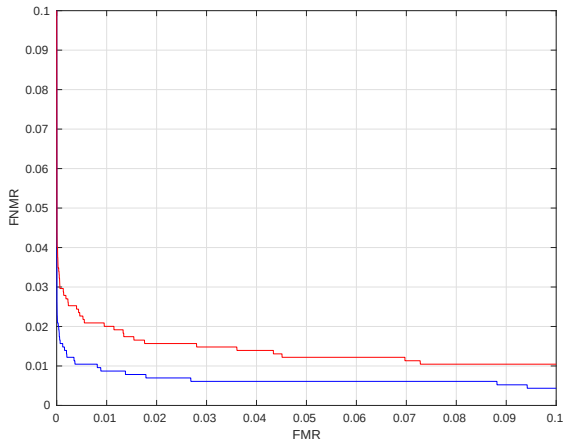


(c) Permanence vs. template age: naïve calculation (red); present method (blue)

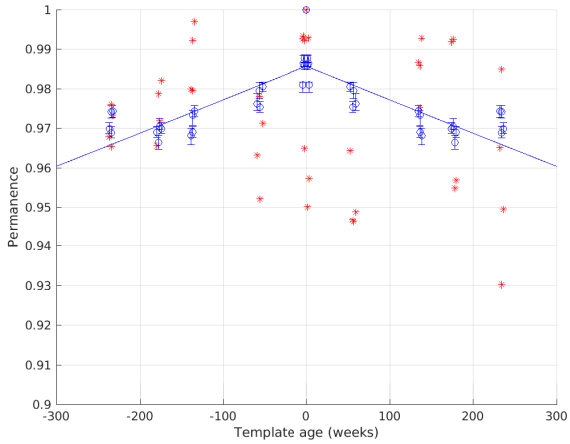
Figure 5.10: Match score distributions, DET, and P_B: Device K (optical)



(a) Genuine (blue) and Imposter (red) score counts at $\Delta t = 0$



(b) DET curve at $\Delta t = 0$ (blue) and at $\Delta t = 373$ weeks (red)



(c) Permanence vs. template age: naïve calculation (red); present method (blue)

Figure 5.11: Match score distributions, DET, and P_B: Device L (capacitive)

5.3 Discussion

The values of P_B derived using the preceding methodology show one of two distinct characteristics: either monotonically decreasing over the course of the study, or constant, depending on the specific device under test. These characteristics seem intuitively reasonable when we consider the baseline (relative template age $\Delta t_{mn} = 0$) genuine and imposter score distributions: those that are essentially separable at $\Delta t_{mn} = 0$ remain so for the duration of the study, while those whose genuine and imposter scores overlap at $\Delta t_{mn} = 0$. In no case did we observe an increasing trend in P_B over time: in this respect, we believe that our methodology exhibits convergent validity with respect to the recorded template ages.

A question that arises in this respect is whether the permanence estimates in Table 5.1 may be used to identify "good" versus "bad" devices. Since our definition of P_B uses the day zero FNMR as a normalizing factor, the answer is generally 'no': a bad device that simply does not get much worse over time may still display high P_B.

For the two devices that showed no change in permanence, the analysis is likely affected by the large class imbalance inherent in such biometric comparisons. That is, for a dataset of K distinct fingers, there are of order K^2 imposter matches but only K genuine matches, which causes the tails of the genuine match score distributions to be much less well defined than those of the imposter distributions. This in turn makes it hard to estimate with confidence the threshold at which to evaluate the corresponding FNMR for the permanence calculation. While the bootstrapping procedure described in 5.2 attempts to ameliorate this effect, if the empirical distributions are separable, then no amount of re-sampling can guarantee that there will be a non-zero FNMR at the chosen FMR. In this regard, a larger study size would have increased the probability of observing ageing behaviour where present.

Since the majority (8 out of 10) devices did show a measurable reduction in permanence over the 7 years, we believe we have observed template ageing over this

time span. A time span of 7 years is broadly in line with common renewal intervals of documents such as biometrically enabled passports (typically either 5 or 10 years), and therefore should be of practical interest to the end users of such technologies. It would be particularly interesting to extend the duration of the study to see whether they eventually showed a similar trend in discriminability.

In the following sections we discuss some other aspects of the data, and their potential impact upon the interpretation of our results.

5.3.1 Time symmetry of the match scores

A key assumption that allows us to substantially remove the visit-to-visit bias factors is that the underlying "true" match scores are time-symmetric: that is, in the absence of these factors, comparisons between a biometric enrolment obtained at time t_1 and a set of verification presentations at later time t_2, and between a biometric enrolment obtained at time t_2 and a set of verification presentations at earlier time t_1, have the same expected match score. ('Expected' because there will still be presentation-to-presentation variability, denoted by the W^{ij} terms in our formalism.) The extent to which this is the case will depend on the algorithm and implementation of the similarity measures used: we might imagine that a simple degree-of-overlap measure to be time-symmetric, whereas a more heuristic matcher might not be. For example, consider the case in which the number of extractable fingerprint minutiae decreases with time, perhaps due to occupational injury or environmental damage; when applied in the reverse time direction, a heuristic might consider the apparent increase in minutiae count to be implausible. Unfortunately such implementation details were not available for the devices in our study.

5.3.2 Constancy of the imposter distributions

Intuitively, we might expect the imposter score distribution to be relatively insensitive to template age, since factors that decrease the similarity between any given pair of subject-fingers may increase the similarity between other such imposter pairs.[1] However this does not allow for gross differences in biometric presentation quality between different pairs of visits. We attempted to quantify the relative contributions of mean changes in imposter scores and those of the genuine match scores as follows.

It is important here to distinguish between statistically significant changes, and changes of significant effect size: since the imposter sample sizes ($\sim K^2$, for a sample of K distinct subject-fingers) are approximately two orders of magnitude larger than those of the genuine matches ($\sim K$, for the same set of subject-fingers), it is almost always possible to reject the null hypothesis that the imposter samples at Δt_{nm} come from the same distribution as those at Δt_{mm}. First we define a discriminability measure Q_{nm} for a pair of visits n, m as the ratio of the difference in sample mean score u between genuine and imposter presentations to the sum of their sample standard deviations s

$$Q_{nm} = \frac{u^G_{nm} - u^I_{nm}}{s^G_{nm} + s^I_{nm}} \tag{5.3}$$

This measure is similar to the Mahalanobis distance familiar from linear discriminant analysis (LDA); the form chosen here is widely used for characterizing the error probability in a binary optical communication channel [23]. We then define the visit-averaged quantities

$$\overline{u}^G = \frac{1}{NM}\sum^N\sum^M u^G_{nm} \qquad \overline{s}^G = \frac{1}{NM}\sum^N\sum^M s^G_{nm} \tag{5.4}$$

$$\overline{u}^I = \frac{1}{NM}\sum^N\sum^M u^I_{nm} \qquad \overline{s}^I = \frac{1}{NM}\sum^N\sum^M s^I_{nm} \tag{5.5}$$

[1] This in fact was an assumption made in the simulations of Chapter 4.

allowing us to express the contributions of the genuine and imposter score variability separately as

$$Q_{nm}^{(G)} = \frac{u_{nm}^{G} - \overline{u}^{I}}{s_{nm}^{G} + \overline{s}^{I}} \qquad\qquad Q_{nm}^{(I)} = \frac{\overline{u}^{G} - u_{nm}^{I}}{\overline{s}^{G} + s_{nm}^{I}} \tag{5.6}$$

i.e. $Q_{nm}^{(G)}$ is the discriminability of the scores between visits nm when the imposter mean and standard deviations are held constant at their visit-averaged values, and $Q_{nm}^{(I)}$ the corresponding discriminabilities this time with the genuine mean and standard deviations held constant. Finally, we evaluate the fractional contribution of the imposter scores to the root mean-square variation in discriminability over the set of visits as

$$\frac{\Delta Q^{(I)}}{\Delta Q} = \sqrt{\frac{\mathrm{var}\left(Q_{nm}^{(I)}\right)}{\mathrm{var}\left(Q_{nm}\right)}} \tag{5.7}$$

where $\mathrm{var}\,(x)$ is the variance of x. Values of $\Delta Q^{(I)}/\Delta Q$ for each of the devices in our study are summarized in Table 5.2.

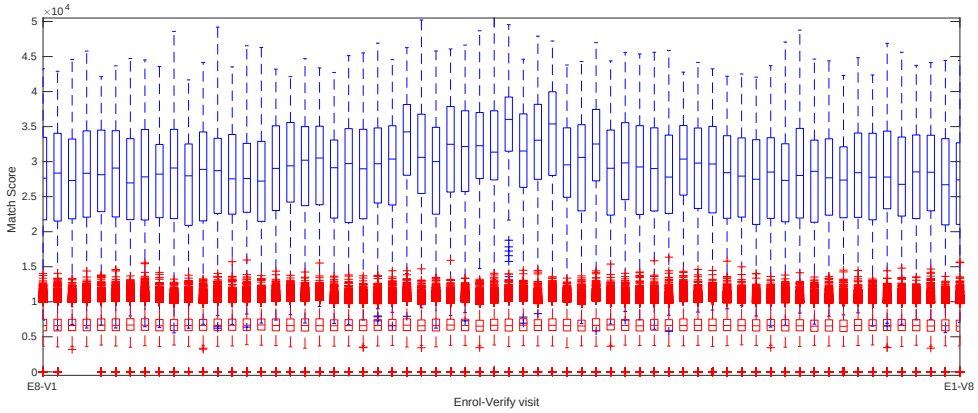
In light of this observed variability in imposter scores, we chose to extend the original method of Chapter 4 to include the imposter matched delta term $\Delta\overline{s}_{nm}^{I}$ in the present work.

The discriminabilities of the devices with the lowest and one of the higher imposter contributions from Table 5.2 were visually examined using box plots (Figs. 5.12a and 5.12b). (The device with the very highest imposter contribution, Device H at 26.45%, was not chosen since its data were only available for six of the eight visits, making direct comparison difficult.) Although these plots confirm clear trends in discriminability, with particularly obvious peaks at each of the $\Delta t_{nm} = 0$ distributions in the case of Device F (Fig. 5.13b), they also highlight a weakness in our treatment: while the "matched delta" methodology seems physically reasonable for the underlying biometric, it does not take into account any thresholding or similar

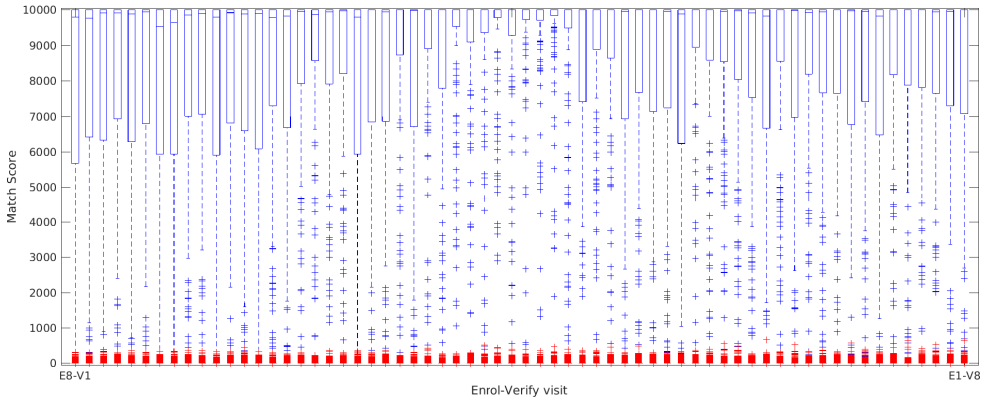
Table 5.2: Relative effect of the imposter distributions to the RMS change in match score discriminability, by device.

ID	$\Delta Q^{(I)}/\Delta Q$ (%)	ID	$\Delta Q^{(I)}/\Delta Q$ (%)
A.	0.40	G.	6.80
B.	12.46	H.	26.45
C.	7.40	J.	1.57
D.	21.12	K.	1.68
F.	0.07	L.	12.49

non-linear processing of the raw match scores. In particular, whereas the box plots of Fig. 5.12a fit well to our assumption that the distributions change in their mean value rather than their shape, those of Fig. 5.12b show distinct limiting behaviour in the - processed - genuine distributions.

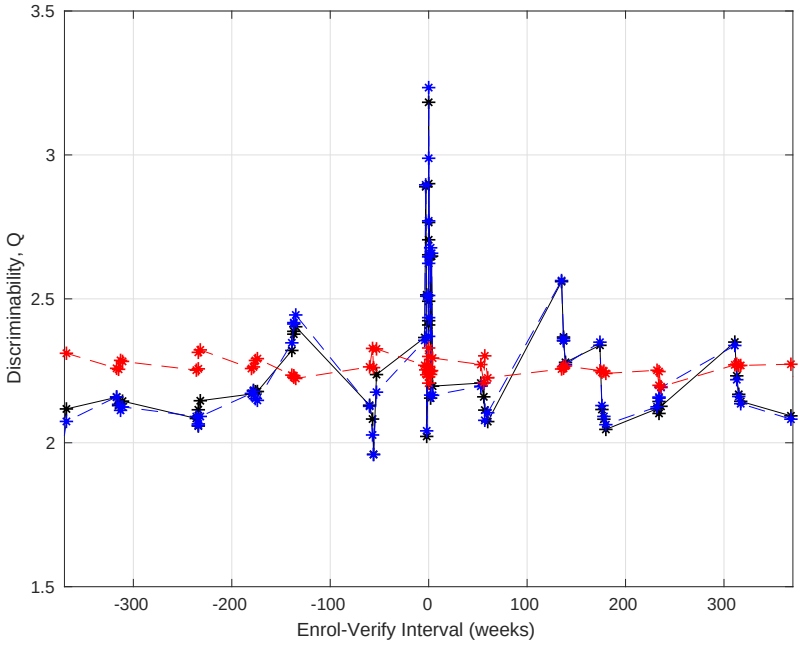


(a) Raw genuine (blue) and imposter (red) match scores: Device L.

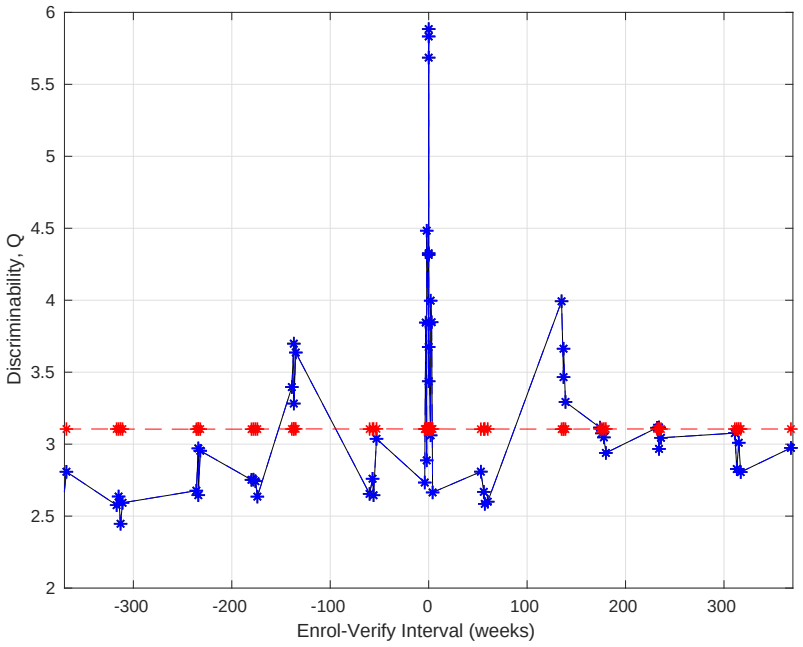


(b) Raw genuine (blue) and imposter (red) match scores: Device F.

Figure 5.12: Box plots of the raw match scores between enrol visit E_m and verify visit V_n. The boxes are plotted from most negative to most positive template age i.e. from 'Enrol 8 – Verify 1' to 'Enrol 1 – Verify 8'. Maximum discriminability occurs around the center of the chart - corresponding to template ages close to zero.



(a) Device L



(b) Device F

Figure 5.13: Binary discriminability Q as a function of template age in weeks. Total discriminability is shown in black; the contributions Q_G (blue) and Q_I (red) are due to changes in the genuine and imposter distributions respectively. Variation of the imposter distribution contributes non-negligibly to the discriminability in Device L but is negligible in the case of Device F.

5.4 Conclusion

We have elaborated a method to isolate and measure changes in biometric system performance over time, using a metric which we call biometric permanence. The method was applied to a dataset spanning several years, and template ageing according to this metric was observed in 8 out of 10 available devices. We have discussed the limits of validity of the underlying assumptions of the methodology, highlighting some device-dependent characteristics of the match score distributions. Because of these factors, it seems appropriate to consider template ageing to be a property of a given biometric system as a whole, rather than a specific physiological mechanism or biometric modality. In order to maintain system performance over life, we recommend that system integrators take such template ageing behaviour into account – for example, by implementing an in-service template update procedure, or a requirement for periodic re-enrolment.

Chapter 6

Biometric quality and classification performance

In this chapter, we compare the available measures of biometric quality on a device-by-device basis, first from the point of view of their assessed quality scores, and then as predictors of classification performance. The goal of the chapter is to understand the relationship between NFIQ-1 and NFIQ-2 scores (and, where available, vendor-specific quality scores obtained during template enrolment), and the relationship between these scores and the measured biometric performance - both in terms of raw match scores and overall classification accuracy.

6.1 History and application of the NFIQ measures

6.1.1 NFIQ-1

The original NFIQ standard, published in 2005, utilizes a single artificial neural network (ANN) classifier that classifies fingerprint images into one of five quality "bins", from 1 (highest) to 5 (lowest). The classification is based on an 11-component feature vector whose values were derived from NIST's own 'MINDTCT' minutia detection

algorithm, and include both features related to the number and quality of the extracted minutiae, and ones derived from an overall image quality map that includes factors such as ridge contrast, direction, flow, and curvature [74].

Since similarity scores necessarily involve pairs of images, training of the ANN was split into two rounds. In the first round, both images in a pair were assigned to the same class; then, in the second round, the predicted classes from the first round were used to adjust the input pattern weights to the classifier.

6.1.2 NFIQ-2

A new NFIQ standard, NFIQ-2 was published in 2016 [73], with a more systematic approach to both feature selection and classifier choice. A large primary feature set was ranked, with redundant features being eliminated based on correlation. Several classifier implementations were evaluated, including support vector machine (SVM), K-nearest neighbour, and random forest, with random forest eventually being selected based in part on its ability to output class probabilities. Similarity score algorithm selection for the classifier training was expanded from three to seven. Continuity with NFIQ-1 was ensured in part by incorporating NFIQ-1 scores into the criteria for labelling of the training data. While NFIQ-1 classifies images into one of five broad quality classes, NFIQ-2 outputs quality values in the range 1 (lowest) to 100 (highest).

6.1.3 Vendor quality metrics

In addition to the open NFIQ metrics, vendors of fingerprint sensing devices may incorporate their own quality assessment algorithms into their devices' software. These are typically used during the enrolment phase, for example by prompting for representation of the finger until a suitably clear record can be obtained; or, if no such record is obtained after a certain number of presentations, declaring a failure to enrol

(FTE) event. The ISBIT BioAPI provides a single-byte field in the fingerprint record of each enrolled finger, with a usable value from 0 to 100 (0 represents lowest quality, 100 represents highest quality) [38]. The field is not mandatory however, and it is up to the vendor to decided if and how it is used.

6.2 Generation of the NFIQ scores

The NFIQ software is most easily built and installed on the Linux operating system. Since the Norwood database is hosted on Windows using Microsoft SQL Server, the procedure used for generating the NFIQ scores was as follows.

First, the the database was queried for a list of candidate image metadata for each device. In the case of EnrolImage records, a candidate is selected if its FingerPosition corresponds to either the FingerPrimary or FingerSecondary of a valid EnrolTemplate record, thus excluding fingers that failed to enrol.

In the case of VerifyImage records, all images are considered candidates. Each list of candidates was then parsed in a Windows batch file and used to construct a sequence of bulk copy ("bcp") instructions that extracted the binary large object (BLOB) fields from the image records and wrote each to a bitmap file on the Windows hard drive. The Windows drive was then CIFS-mounted to the Linux laptop in order to run the NFIQ software, in bulk mode, on the extracted images.

When run in bulk mode, the NFIQ software outputs a plain text list of file names and quality scores (with, optionally, constituent feature scores in the case of NFIQ-2). Since metadata (such as Subject, Visit, and Algorithm primary keys) was included in the candidate file names, it is then possible to create new database tables and import the quality scores back into the database in a way that allows further structured queries.

The same procedure was used for both the original NFIQ-1 and the later NFIQ-2

scores.

6.3 Extraction of vendor quality scores

The original ISBIT protocol provides for vendor-supplied quality metrics to be recorded in the fingerprint template. In fact, since it is a two-finger protocol, there are quality values for both the enrolled images and the template as a whole. In this work, we are only interested in the image quality assessment, where available, since that is what is most directly comparable with the NIST NFIQ measures.

The procedure for extracting vendor-assessed image qualities mirrors that for the NFIQ enrol image scores, with the candidate lists being used this time to identify EnrolTemplate records from which a binary Biometric Information Record (BIR) BLOB is extracted. A small Python program was then written to unpack the binary BIR and extract the quality scores from the individual primary and secondary fingerprint records (Appendix B).

No attempt was made to extract vendor quality scores associated with verification presentations.

6.4 Comparisons of NFIQ1, NFIQ2, and vendor quality

Histograms of the quality scores, by device, are shown in Figure 6.1. For ease of comparison, the NFIQ-1 scores are reversed so as to go from 5 (worst) to 1 (best). Scores for enrolment images are shown in light blue, and for verification images (plotted second) in light brown, becoming dark brown where the colours overlap. Hence blue regions extending above dark brown indicate an excess of enrolment images in the score category, while light brown extending above dark brown indicates an excess of verification images. NFIQ-2 scores for Devices A and E were unavailable at the time of writing.

The vendor quality scores from Devices F and H provide no useful information, being constant over the entire data set. Likely the vendors of these devices simply chose not to evaluate image quality, or not to expose the values via the API.

Device H is interesting for a number of other reasons. Initially, the NFIQ-1 software would not process any of its captured images. Side by side examination of the image metadata against that of successfully processed images suggested this was a result of incorrect image metadata: specifically, that the images were (apparently incorrectly) labelled as having 16-bit depth. Re-exporting the extracted files explicitly as 8-bit uncompressed images using the commercially available ImageMagick *convert* program, viz.

```
convert BMP:"$f" -type palette -depth 8 -compress none BMP:"$f"
```

seemed to fix the issue, and allowed putative NFIQ-1 scores to be obtained for the device. However, when the same "fix" was applied during generation of NFIQ-2 scores, the newer software again failed to recognize them as valid 8-bit (greyscale) bitmaps. Further examination of metadata suggested this was a result of a change in ImageMagick's default BMP save format (from BMP3 to BMP4) that could apparently be remedied by making the save format even more explicit

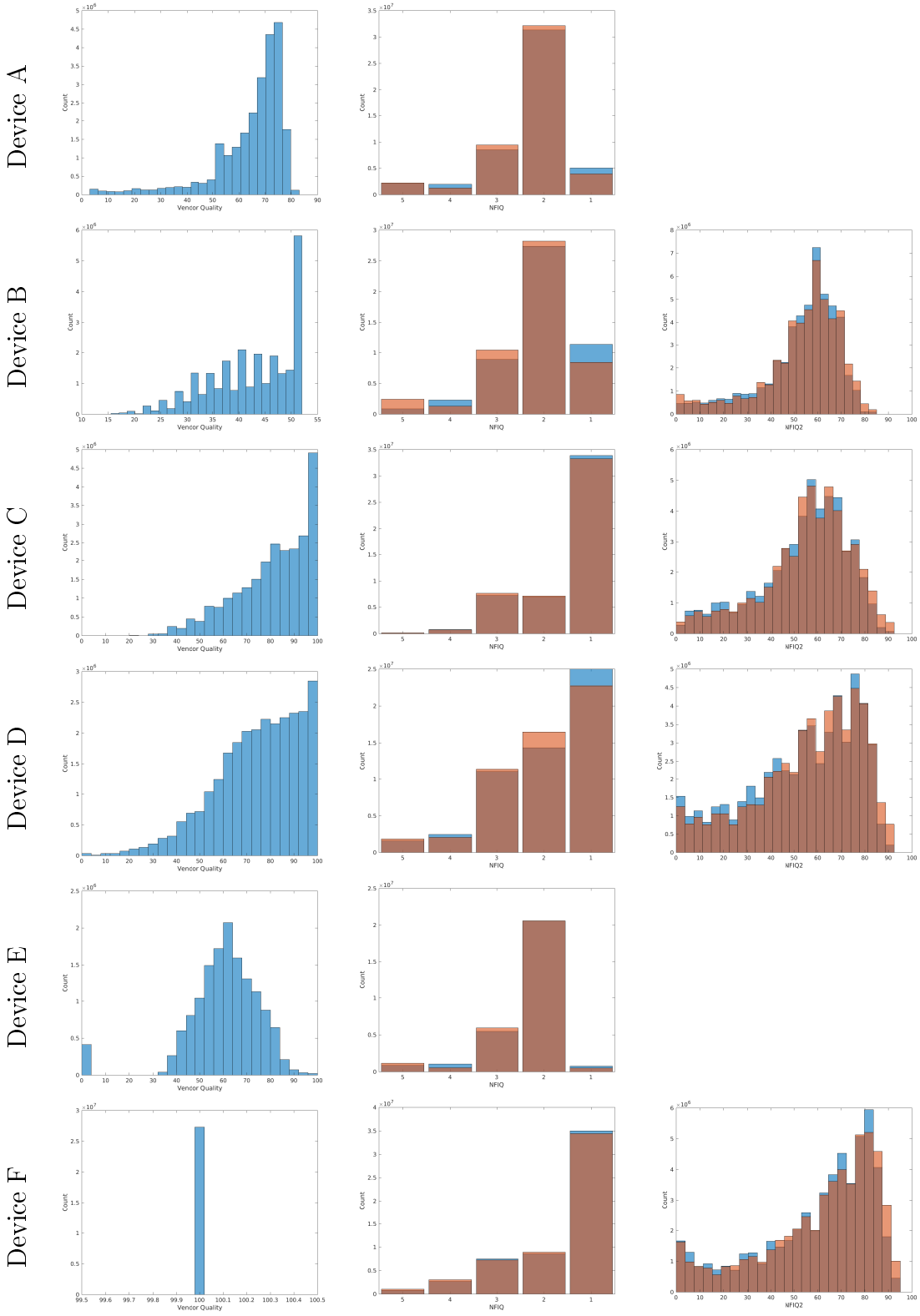
```
convert BMP:"$f" -type palette -depth 8 -compress none BMP3:"$f"
```

allowing putative NFIQ-2 scores to be obtained for the device as well. However, the scores so obtained are markedly different from those from the other devices in the study: in particular, NFIQ-1 scores show a large fraction of low quality (score 5) images, whereas NFIQ-2 scores are rather tightly clustered around the high quality quartile (scores 50-100).

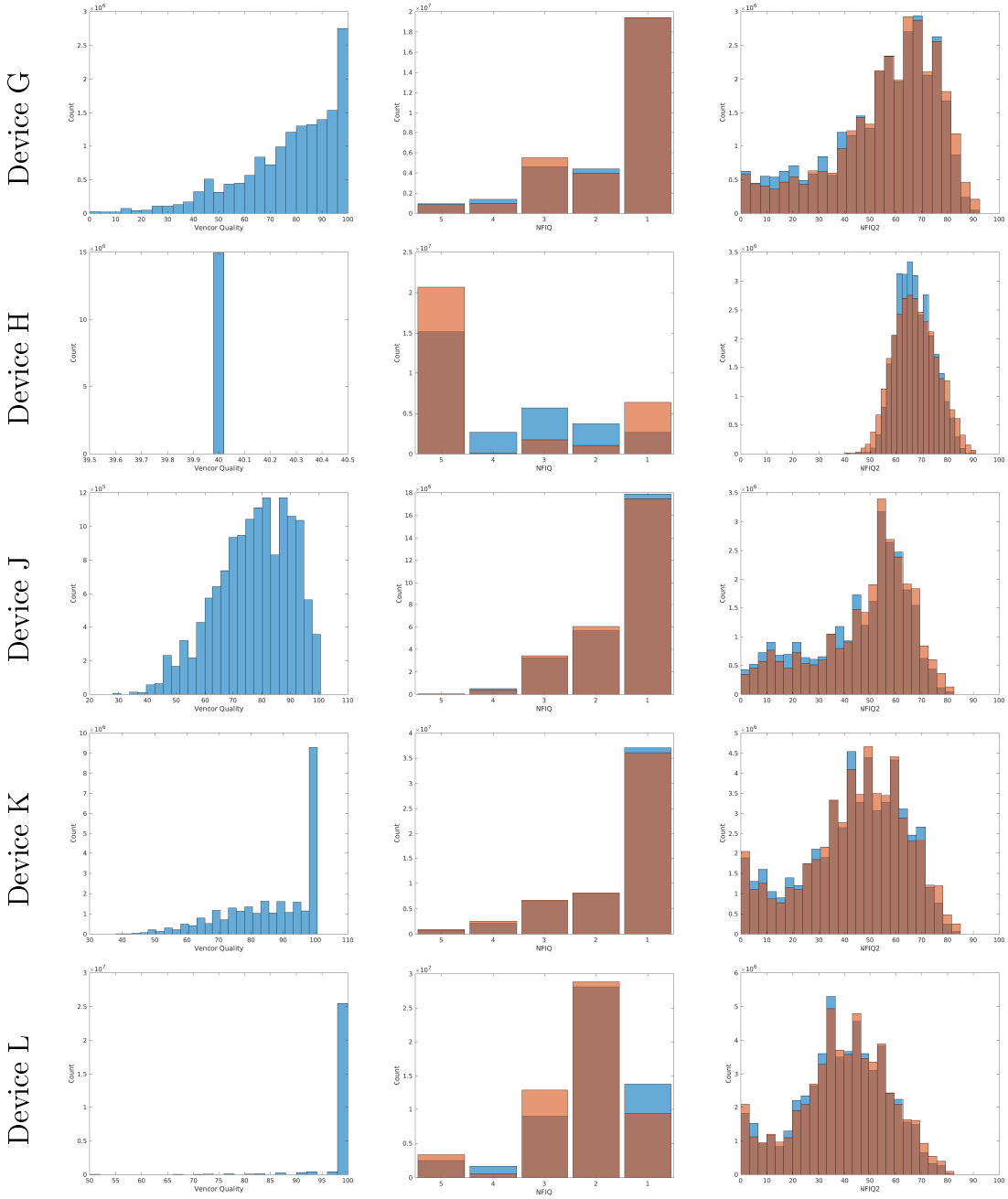
Device H is the only multi-spectral capture device in the study, and it seems likely that the processing of its images into a single greyscale bitmap makes it in some way

unsuitable for the NFIQ algorithms. For this reason, results for this device should be regarded as of questionable value.

Otherwise, there is a good deal of similarity between the NFIQ-1 and NFIQ-2 score distributions. For example, Devices B, D G all all show similar low-quality tails. For Devices F and J, NFIQ-2 appears to classify more images as low quality, producing slightly bimodal distributions where the NFIQ-1 scores are unimodal. Perhaps the closest correspondence is for Device L, where bimodal distributions are observed for both NFIQ-1 and NFIQ-2 scores.



(a) Vendor **(b)** NFIQ1 **(c)** NFIQ2



(a) Vendor **(b)** NFIQ1 **(c)** NFIQ2

Figure 6.1: Histograms of extracted quality scores, by device. Scores for enrolment images are shown in light blue, and for verification images (plotted second) in light brown, becoming dark brown where the colours overlap. *NFIQ-2 scores for Devices A and E were unavailable at the time of writing.*

6.5 Effect of biometric quality on match score

Training of both NFIQ-1 and NFIQ-2 classifiers is based on match score. Having generated NFIQ scores and imported them into the Norwood database as described in the preceding section, it becomes possible to label biometric match scores with the NFIQ-1, NFIQ-2, and (where available) vendor-assessed quality values of their constituent enrolment and verification images, and explore the relationship between quality and match score for the Norwood dataset. Since the demographics and capture protocol of the ISBIT/Norwood study are likely quite different from those used to obtain the majority of the NFIQ training data, this is expected to be a useful test of the broader applicability of the NFIQ measures.

In the case of enrolment, the ISBIT/Norwood protocol involves three presentations of each of two fingers, and hence associates six fingerprint image records with each enrolment template. While match scores are provided separately for each enrolled finger, it is not possible to break these down to a single enrolment image: indeed, the enrolment protocol does not specify how the three single-finger presentations are to be used, so that vendors may (for example) have chosen a single preferred presentation, or have aggregated presentations at either the image level or at the feature level.

Both NFIQ-1 and NFIQ-2 scores are returned, and stored in the Norwood database, as integers[1] but are cast to floats for the purpose of averaging. In the case of NFIQ-1, for which the quality score is a coarse ordinal scale from 1 (highest) to 5 (lowest), the averaging provides somewhat finer granularity.

Best, average, and worst enrolment image qualities are compared for NFIQ-1 in Figure 6.2 and for NFIQ-2 in Figure 6.3. NFIQ-2 results are presented as simple scatter plots, while for the NFIQ-1 results, the coarse score classes, especially for the non-averaged cases, favour the use of box plots.

Although the protocol for verification also involves multiple presentations, each

[1] As *tinyint* in the case of NFIQ-1, and as *smallint* in the case of NFIQ-2

results in a individual match score which we can identify uniquely with a single verification image. No averaging is employed in this case. Best, average, and worst enrolment image qualities for NFIQ-1 and NFIQ-2 are compared in Figure 6.4.

Naïvely, one might expect (or at least hope) that genuine match score would be simply correlated with the quality of either or both the enrolment and verification images. Perhaps unsurprisingly, the relationship is apparently not so straightforward.

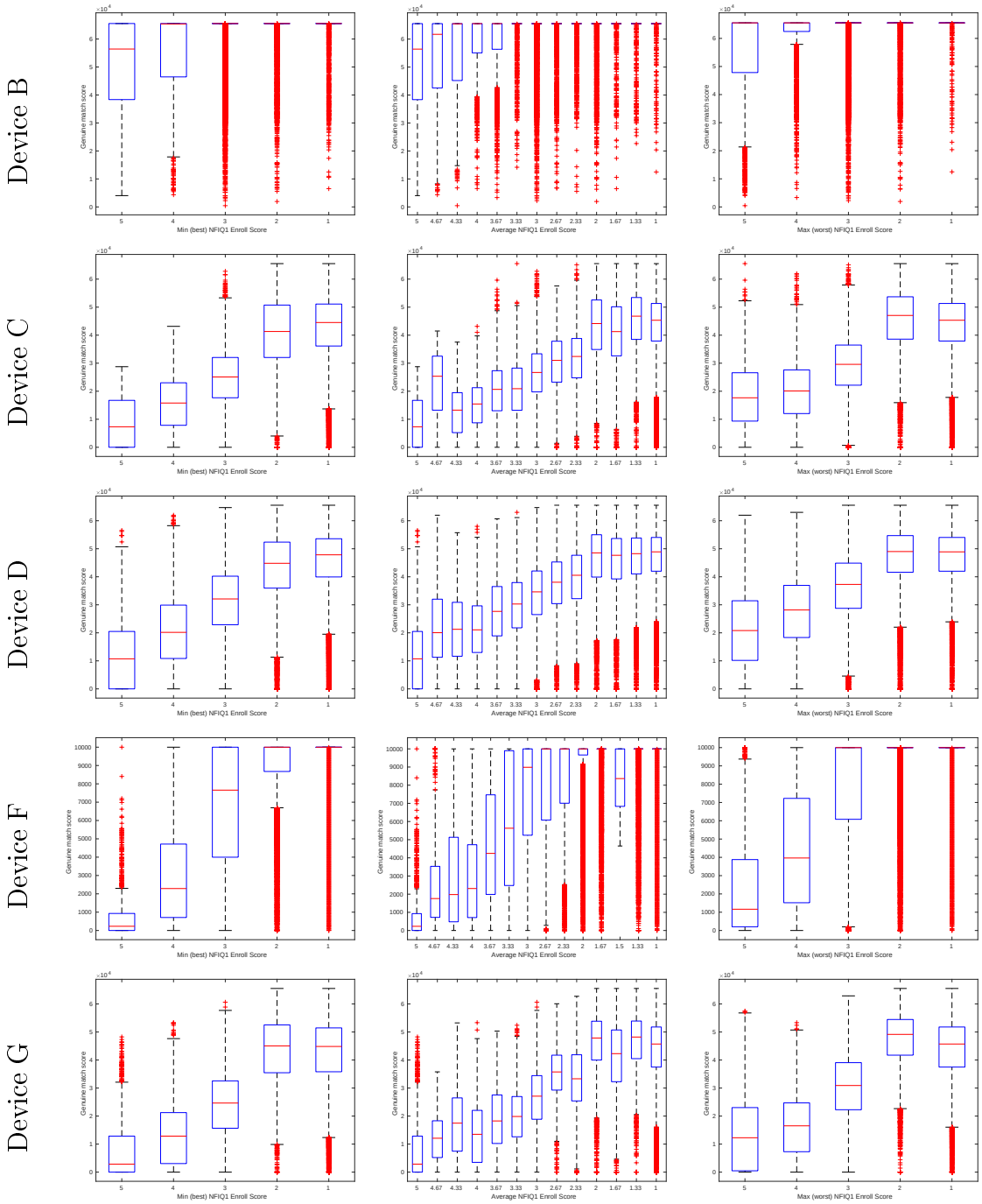
Perhaps the most obvious way in which the NFIQ-1 box plots deviate from an ideal quality-to-match score relationship is that they often display long tails of low match score for the high quality classes. These tails are especially apparent in Devices B, F, and J. Box plots for imposter match scores were omitted for clarity.

For NFIQ-2, the figures (6.3 and 6.4) are in the form of scatter plots of match score versus NFIQ-2 quality for both genuine and imposter scores. Genuine match scores generally show a positive correlation with NFIQ-2 quality score, while (as might be expected) imposter match scores are generally uncorrelated. However there is a rather broad spread of match scores for the same assessed quality. In the case of Device F, there is evidence of a subset whose NFIQ-2 quality score does not correlate at all to match score. In devices K and L, we see some evidence of elevated (false) match scores in the case of high quality imposters.

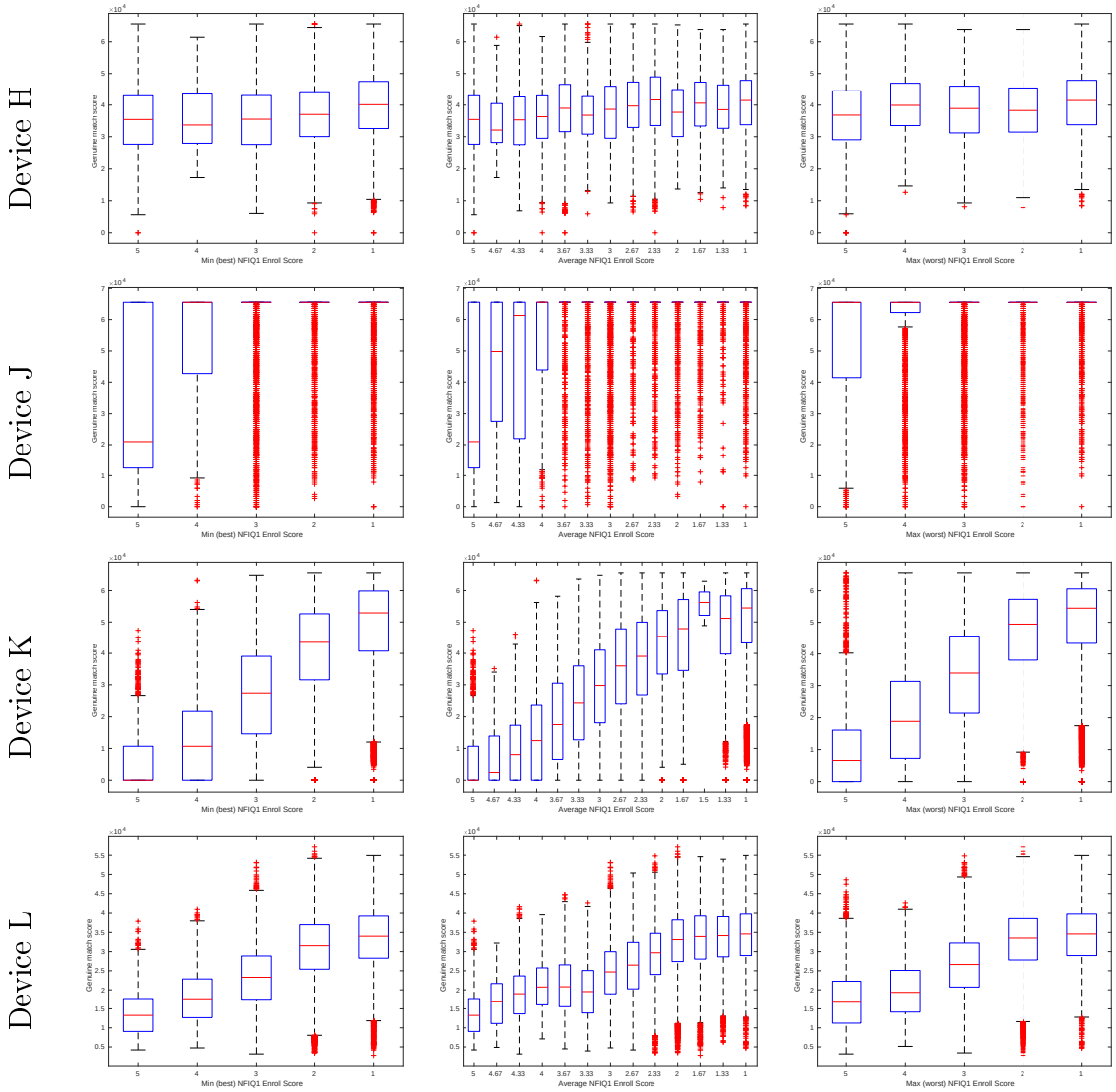
So far, results have been presented separately for enrolment quality and verification quality. A natural question that arises is how the qualities combine in a single match.

In order to probe this, genuine match scores were plotted against various composite enrolment-verification quality indices. Figure 6.5 shows two such indices: firstly the geometric mean of the (3-presentation average) enrolment and verification NFIQ-2 scores; and secondly the reciprocal sum given by

$$\frac{2}{\frac{1}{NFIQ2_{enrol}} + \frac{1}{NFIQ2_{verif}}}$$

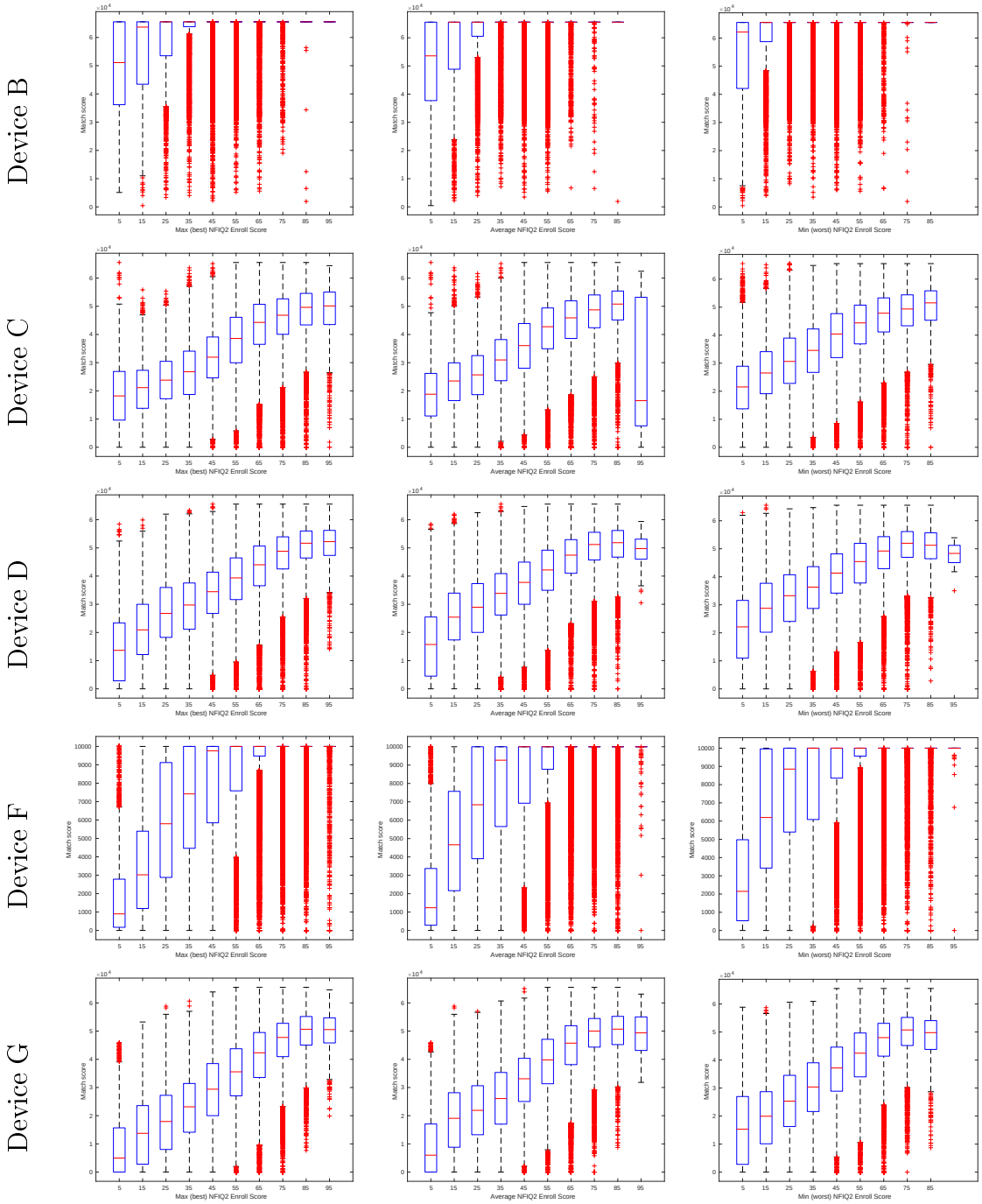


(a) Best **(b)** Average **(c)** Worst

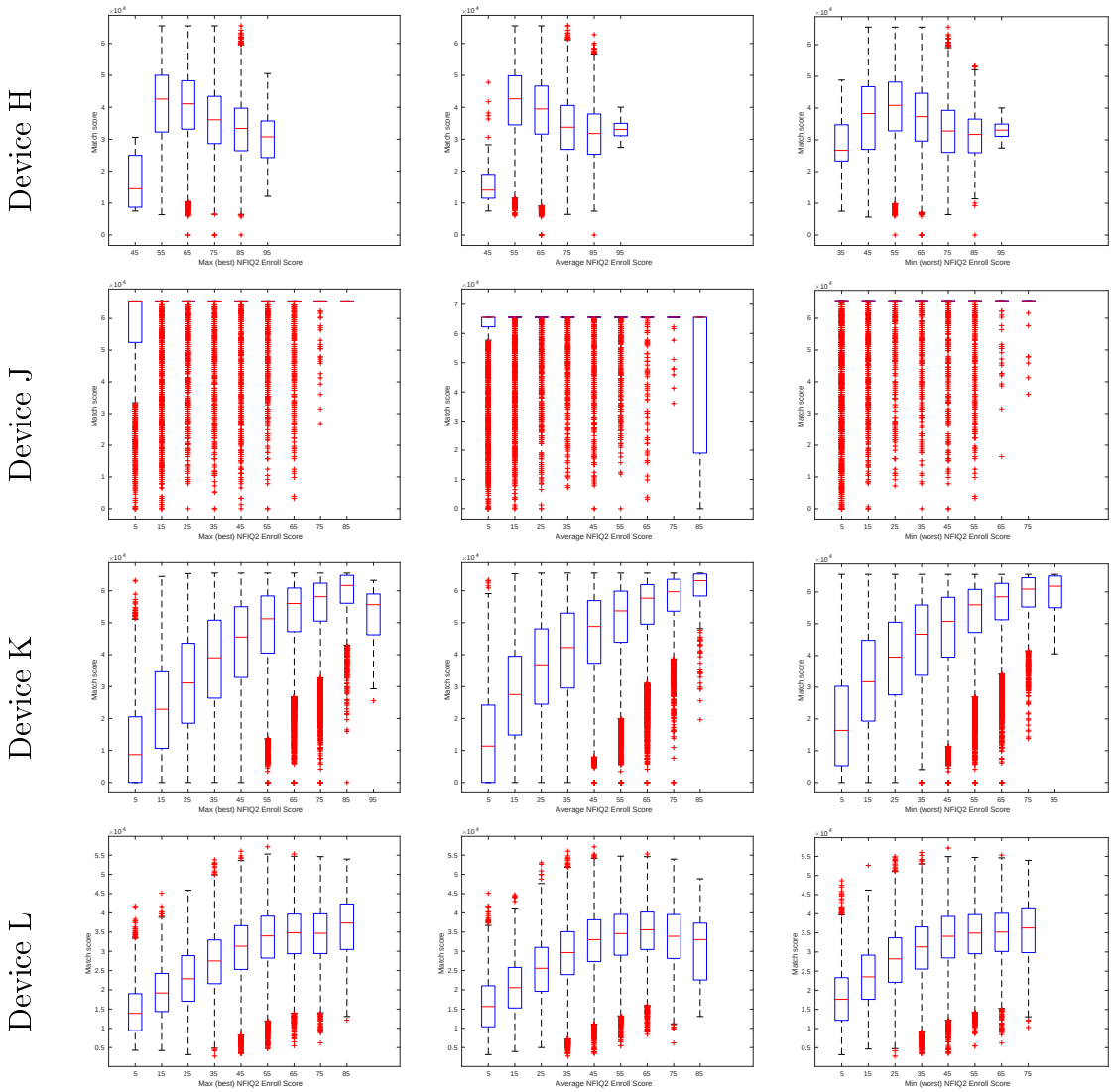


(a) Best **(b)** Average **(c)** Worst

Figure 6.2: Comparison of genuine match score versus enrol NFIQ-1 for the three presentations of each enrolment event; for ease of comparison with the later NFIQ-2 results, the NFIQ-1 scores are reversed so as to go from 5 (worst) to 1 (best).

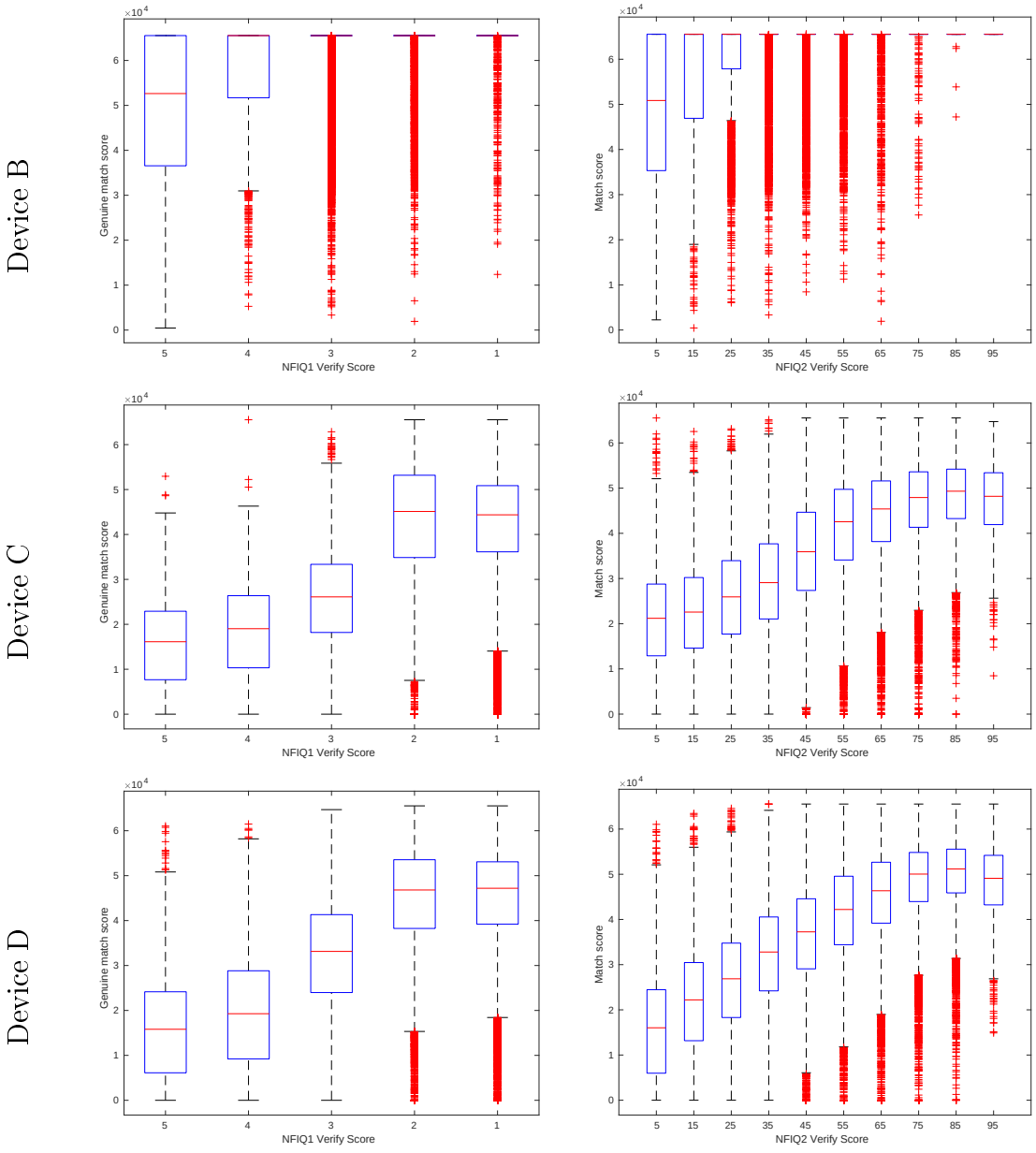


(a) Best **(b)** Average **(c)** Worst

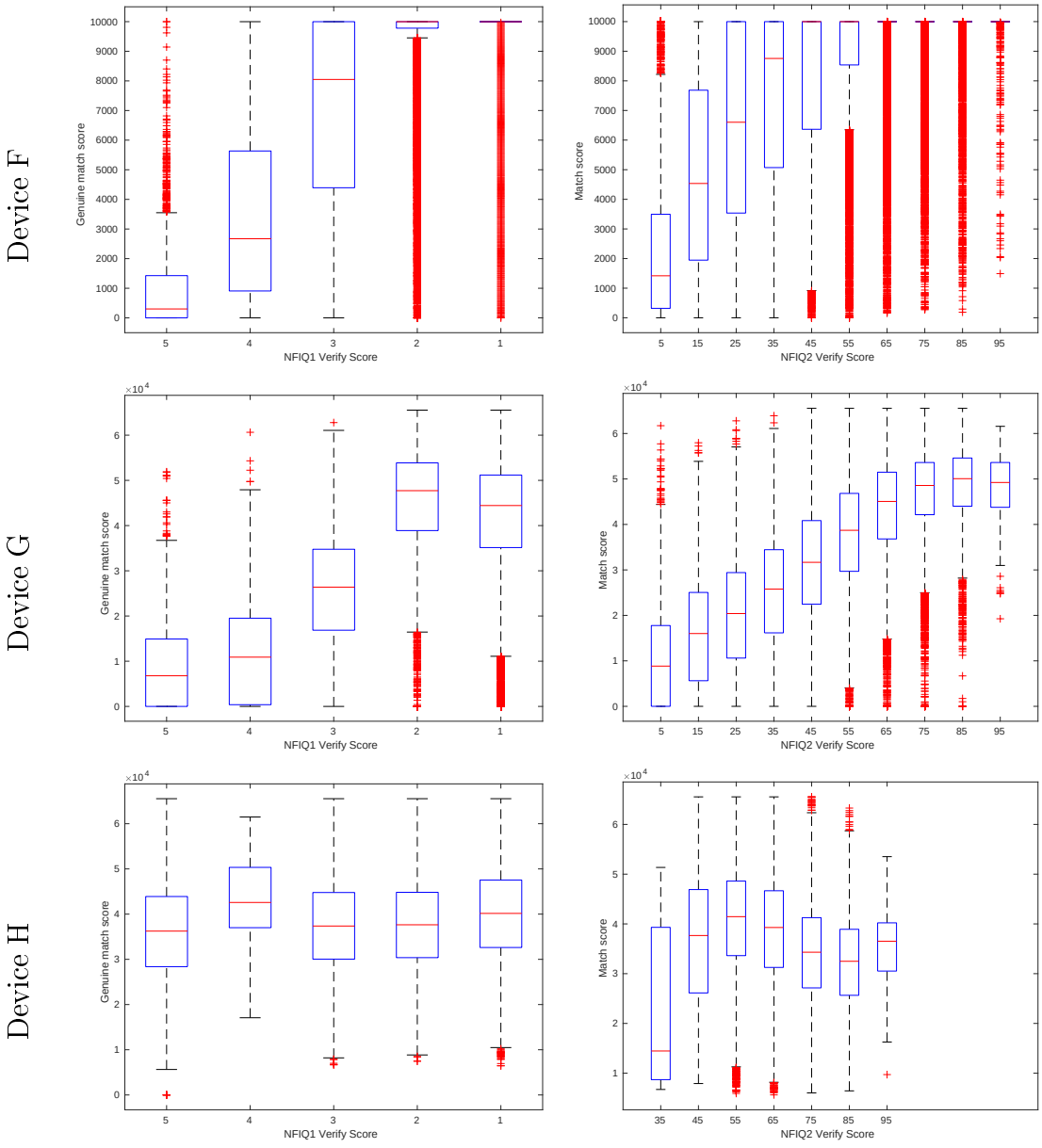


(a) Best **(b)** Average **(c)** Worst

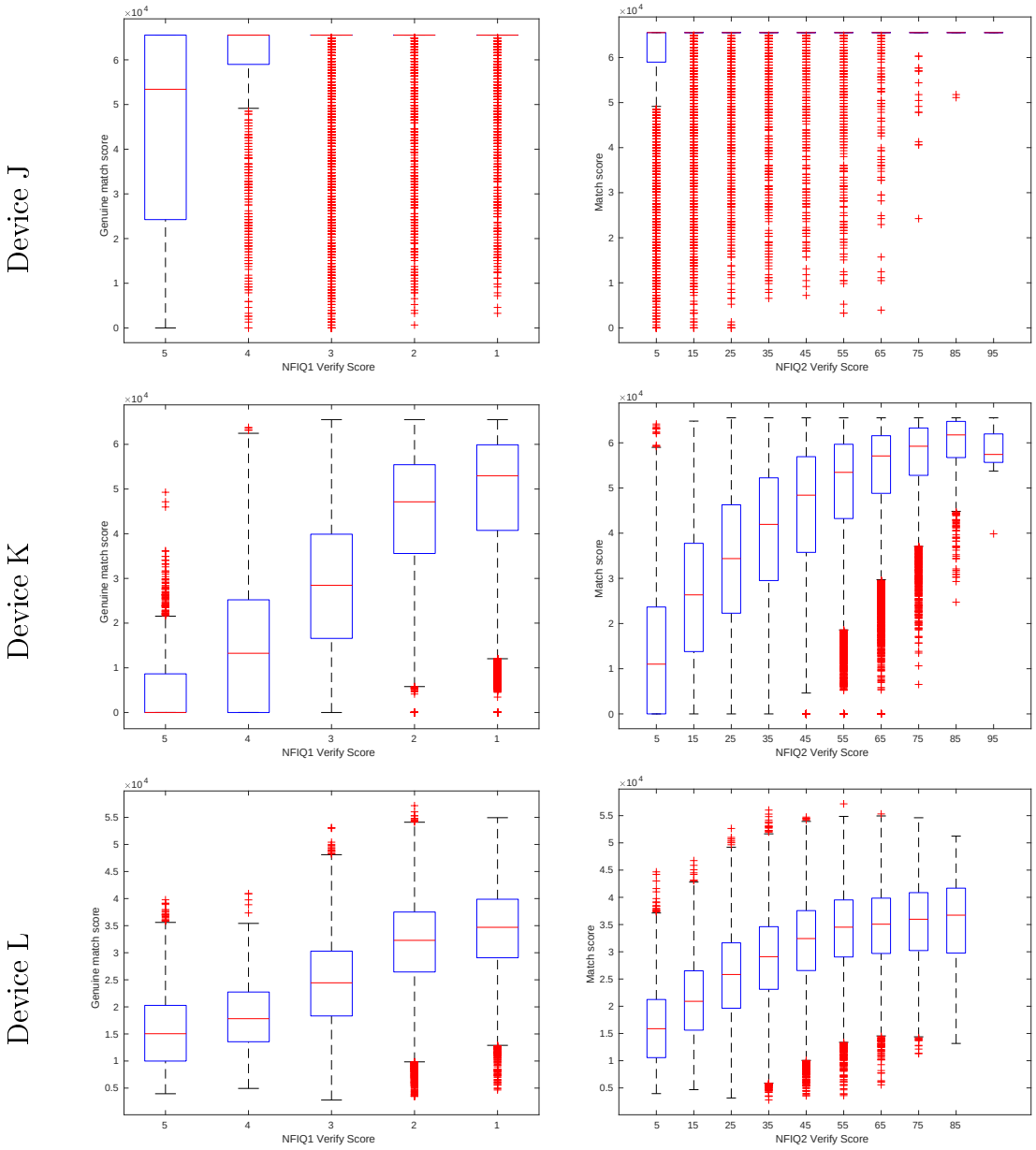
Figure 6.3: Comparison of genuine match score versus enrol NFIQ-2 for the three presentations of each enrolment event.



(a) NFIQ1 **(b)** NFIQ2



(a) NFIQ1 **(b)** NFIQ2



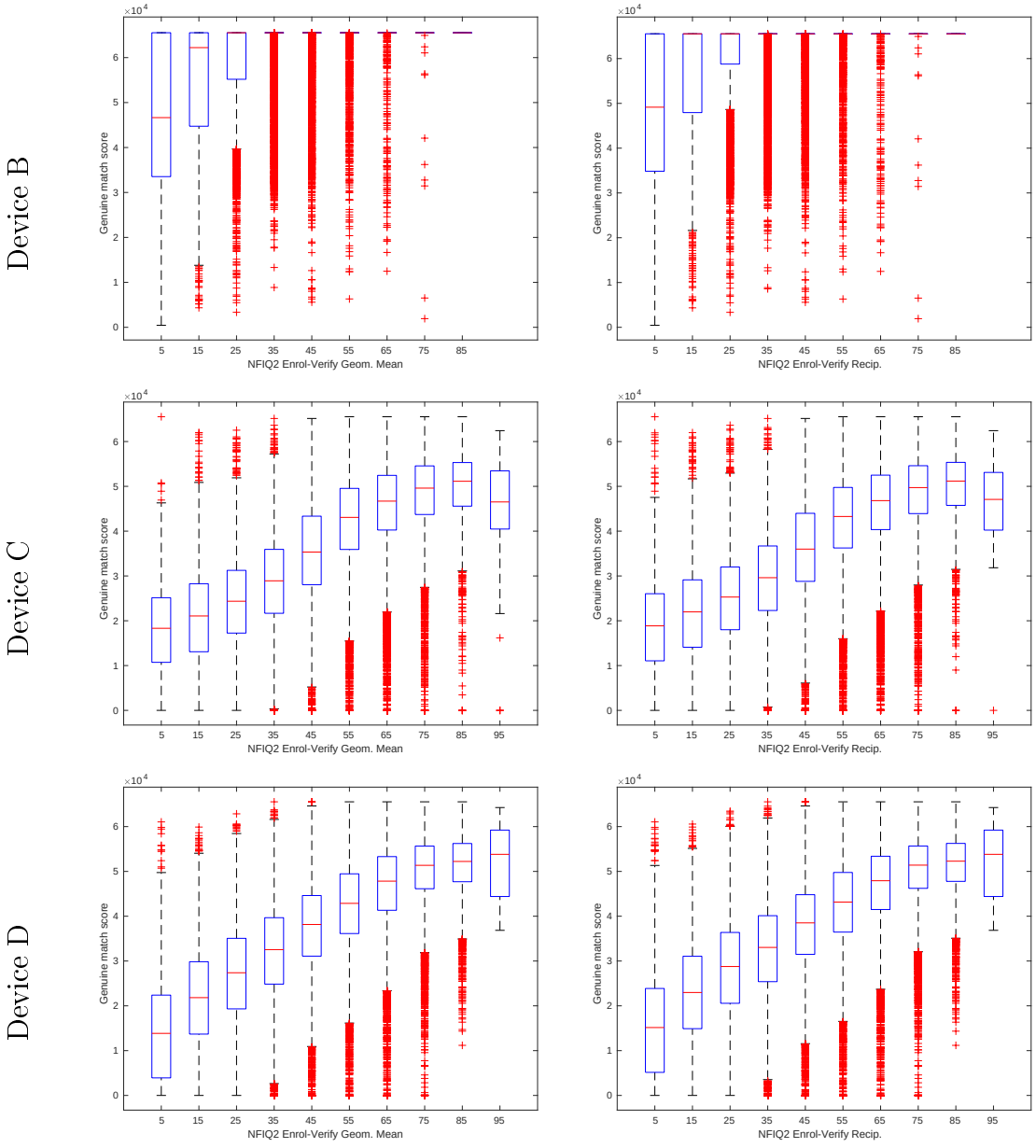
(a) NFIQ1 **(b)** NFIQ2

Figure 6.4: Comparison of verification NFIQ1 and NFIQ2 scores for genuine matches; for ease of comparison, the NFIQ-1 scores are reversed so as to go from 5 (worst) to 1 (best).

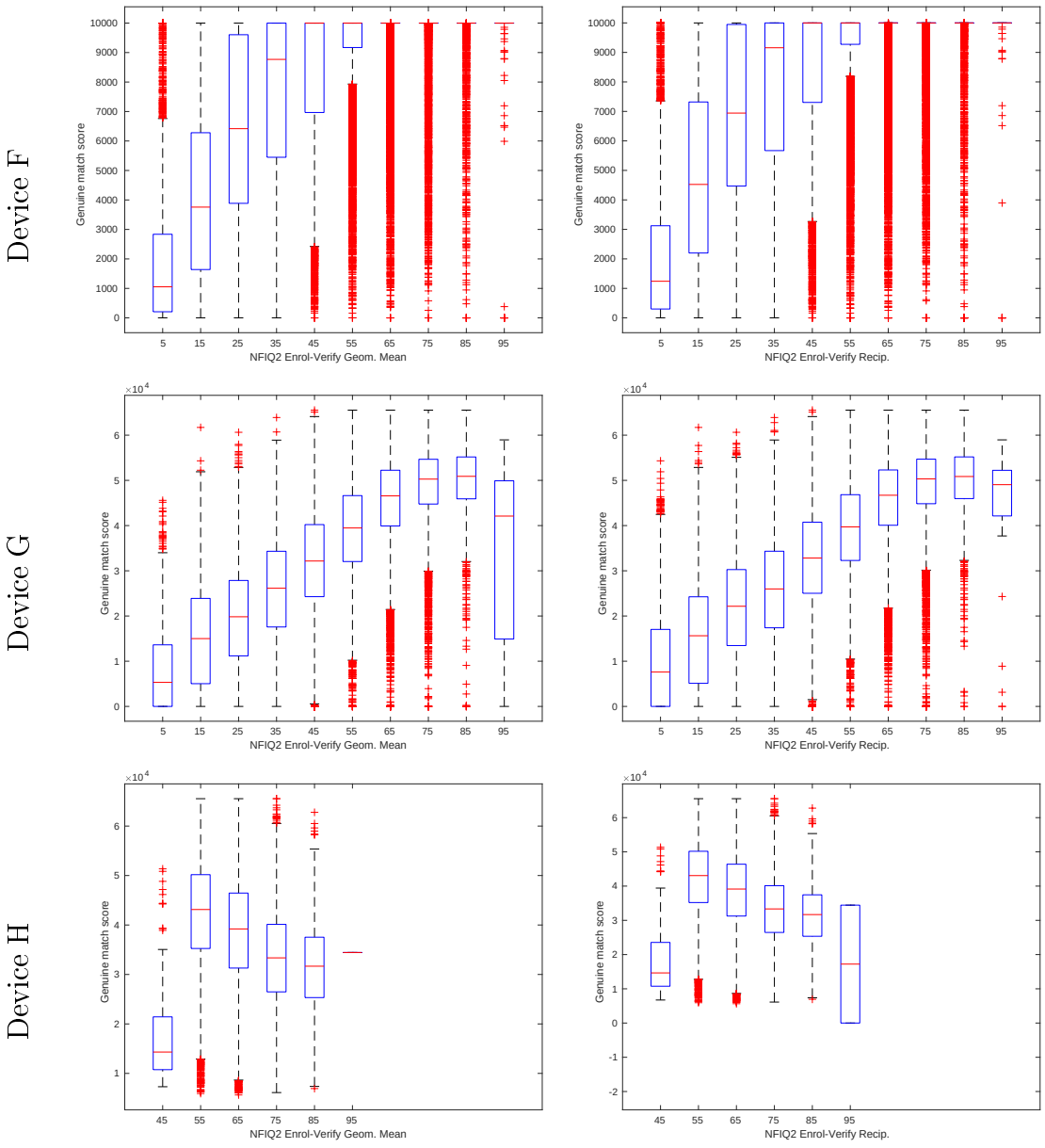
Compared with the individual results (Figures 6.3 and 6.4) it appears that combining the qualities of both (average) enrolment and verification results in a stronger correlation with match score - in particular, it reduces the number of low quality – high match score outliers.

In all these results, the behaviour of Device H is clearly anomalous: this is the multispectral device whose NFIQ processing was troublesome and, as noted in the previous section, should be treated as unreliable.

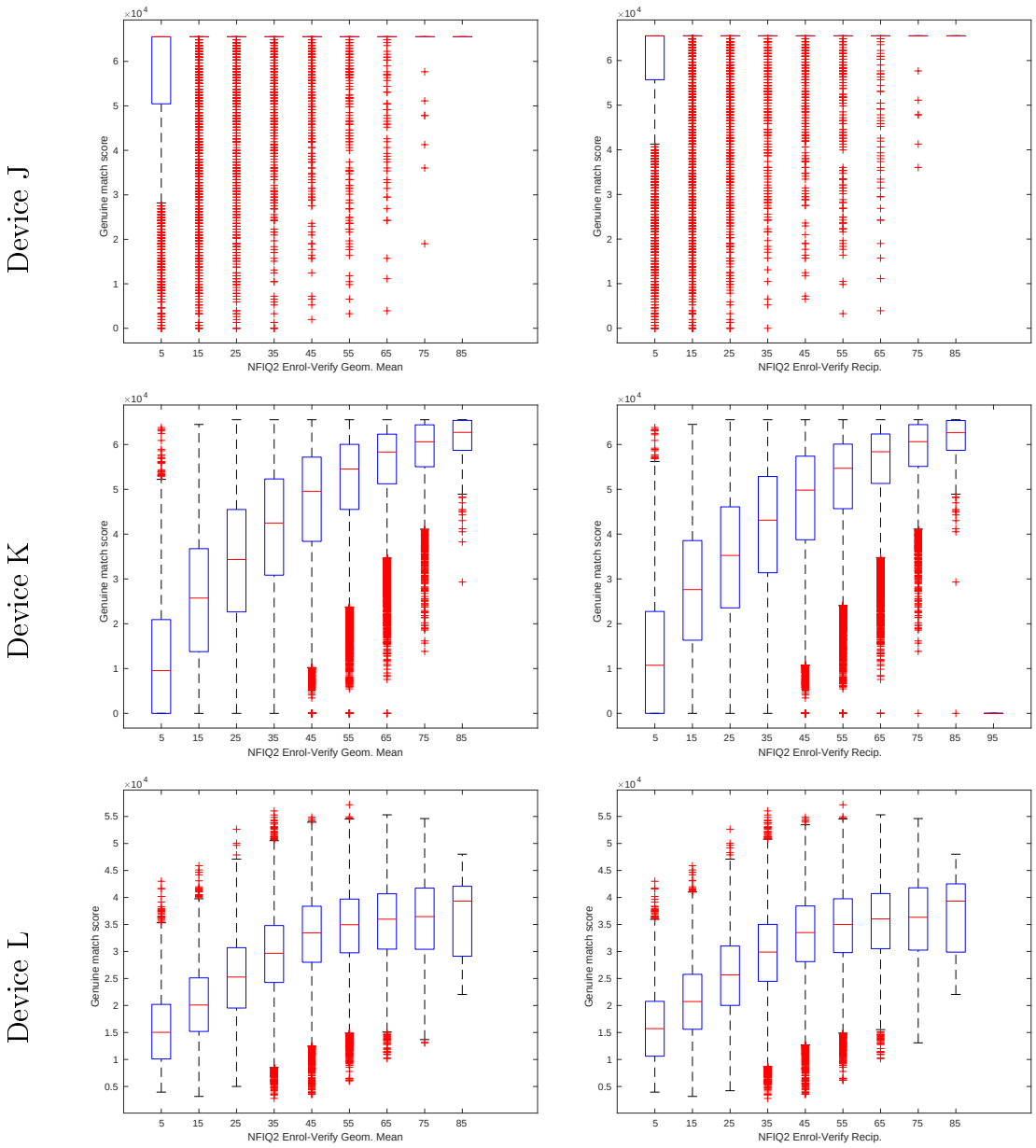
In the following sections we look more directly at classification performance, by means of Decision Error Trade-off (DET) curves.



(a) Geometric mean **(b)** Reciprocal sum



(a) Geometric mean **(b)** Reciprocal sum



(a) Geometric mean **(b)** Reciprocal sum

Figure 6.5: Genuine match scores versus composite Enrol-Verify NFIQ-2 score: geometric mean and reciprocal sum.

6.6 Effect of biometric quality on classification accuracy

The previous section examined the relationship between biometric quality and match score for individual match pairs. Here, we consider its impact on overall classification accuracy of a biometric IDMS.

The classification accuracy was explored using Decision Error Trade-off (DET) curves, broken down as before by device. The procedure for evaluating classification performance versus quality was based on selecting subsets of the available data based on their quality indices, and constructing DET curves for each subset. Based on the preceding results, the arithmetic mean of the three enrolment image qualities was used in all cases.

In the quality screening application, the idea would be to reject some low quality presentations: during enrolment (perhaps forcing enrolment of a different - higher quality - finger) and/or during verification (for example, declining to attempt a match of a poorly presented or occluded finger). In order not to diminish significantly the overall convenience of the system, it would be desirable to limit to a small fraction the number of such rejected presentations. Thus we seek a quality threshold that is effective at improving overall classification accuracy, while limiting the increase in failure to enrol (FTE).

When comparing quality metrics in this application, we should ideally do so at the same fraction of rejected presentations i.e. compare the increase in classification performance for the same decrease in convenience. Since NFIQ-1 classifies images into one of five discrete quality classes, it provides little scope for precise control of the rejected fraction: at best, we can really only reject all of the lowest quality class - with the caveat that this may correspond to a quite widely varying fraction of cases from device to device. For example, referring to the histograms of Figure 6.1, we see that Devices C and J classify less than 1 % of images at the lowest NFIQ-1 quality (Class 5) whereas Devices B and L have up to 6 % in this class. (Device H apparently

Table 6.1: Excluded percentages x of the lowest quality NFIQ-1 class (Class 5) and the corresponding thresholds (lower x^{th} percentiles) for exclusion from NFIQ-2

Device ID	NFIQ-1		NFIQ-2	
	Enrol %	Verify %	Enrol Thresh.	Verify Thresh.
B.	1.7	4.8	6.0	13.0
C.	0.2	0.3	2.0	2.0
D.	2.9	3.4	4.0	6.0
F.	1.5	1.9	1.3	2.0
G.	3.1	2.8	7.0	6.0
H.	51.0	69.1	67.3	72.0
J.	0.2	0.2	1.3	0.0
K.	1.4	1.7	1.7	1.0
L.	4.6	6.1	5.0	7.0

shows anomalously high (50-70 %) in NFIQ-1 Class 5, but we believe those results to be erroneous, and due to the particular biometric capture technology of the device.) In contrast, NFIQ-2 has a much more expressive quality scale, from 1-100, which allows quite precise rejection below (or above) a given quality percentile.

In order to compare the efficacy of NFIQ-2 versus NFIQ-1 for this application, we therefore first evaluated the change in overall classification performance obtained by eliminating the lowest NFIQ-1 quality class (from both enrolment and verification), noting for each device the percentage of enrolments and verifications removed. Then we repeated the evaluation, thresholding the NFIQ-2 scores at the same percentiles (Figure 6.6). Table 6.1 summarizes the percentiles and thresholds.

Although of little practical value, it is also scientifically interesting to consider the effect of excluding the *highest* quality matches. Table 6.2 shows the corresponding percentages of NFIQ-1 Class 1 enrolments and verifications, and NFIQ-2 thresholds by device for this case. Figure 6.7 presents results for each device side-by-side, with the baseline DET curve in black, the results with lowest quality matches removed in green, and those with highest quality scores removed in red.

Taken device by device, we see that there is very little to distinguish the performance of NFIQ-2 versus NFIQ-1 at the same percentage of rejected presentations. If

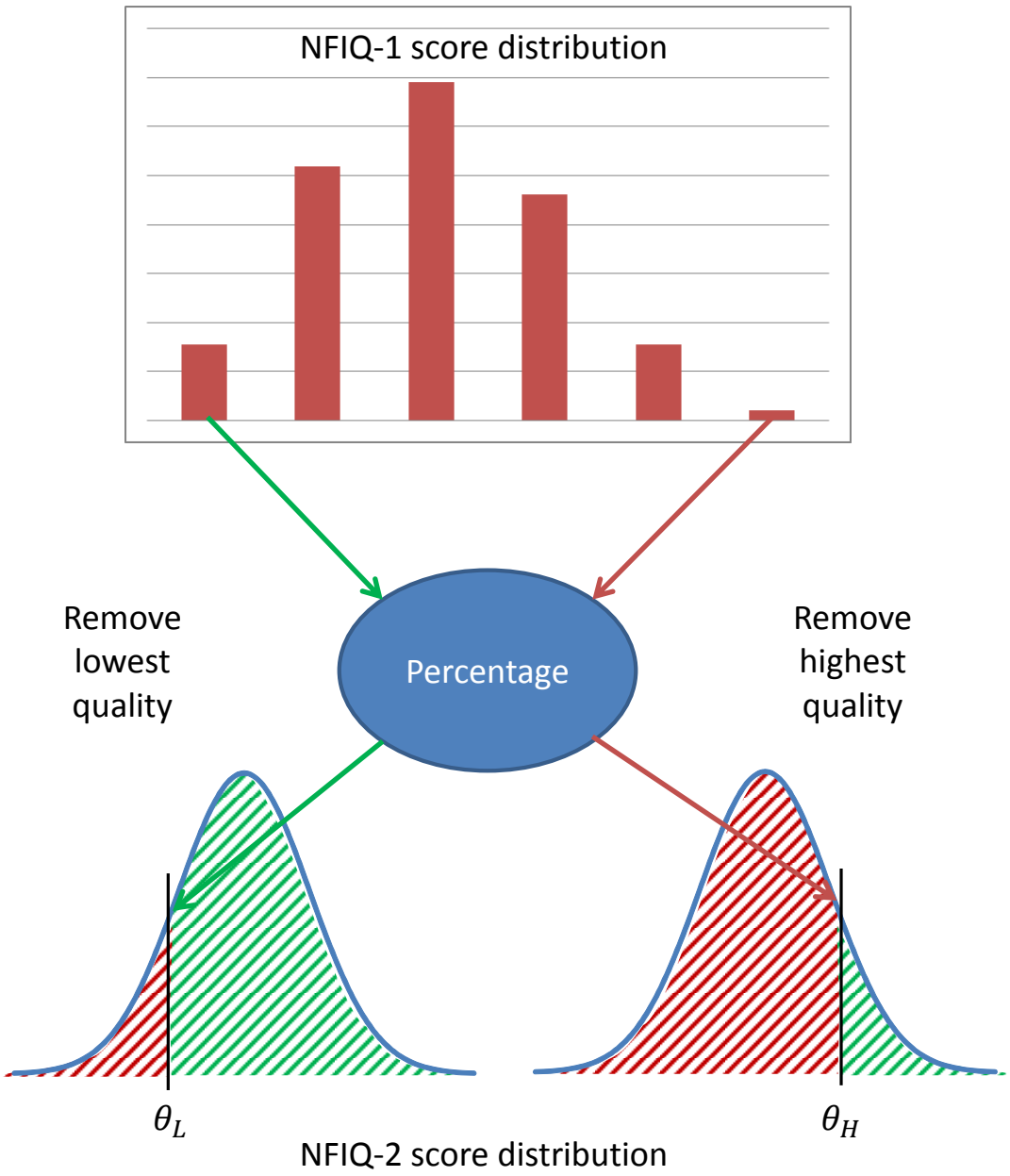


Figure 6.6: Schematic representation of the conversion of NFIQ-1 class percentages into NFIQ-2 thresholds. The correspondence is applied separately for enrolment and verification quality scores.

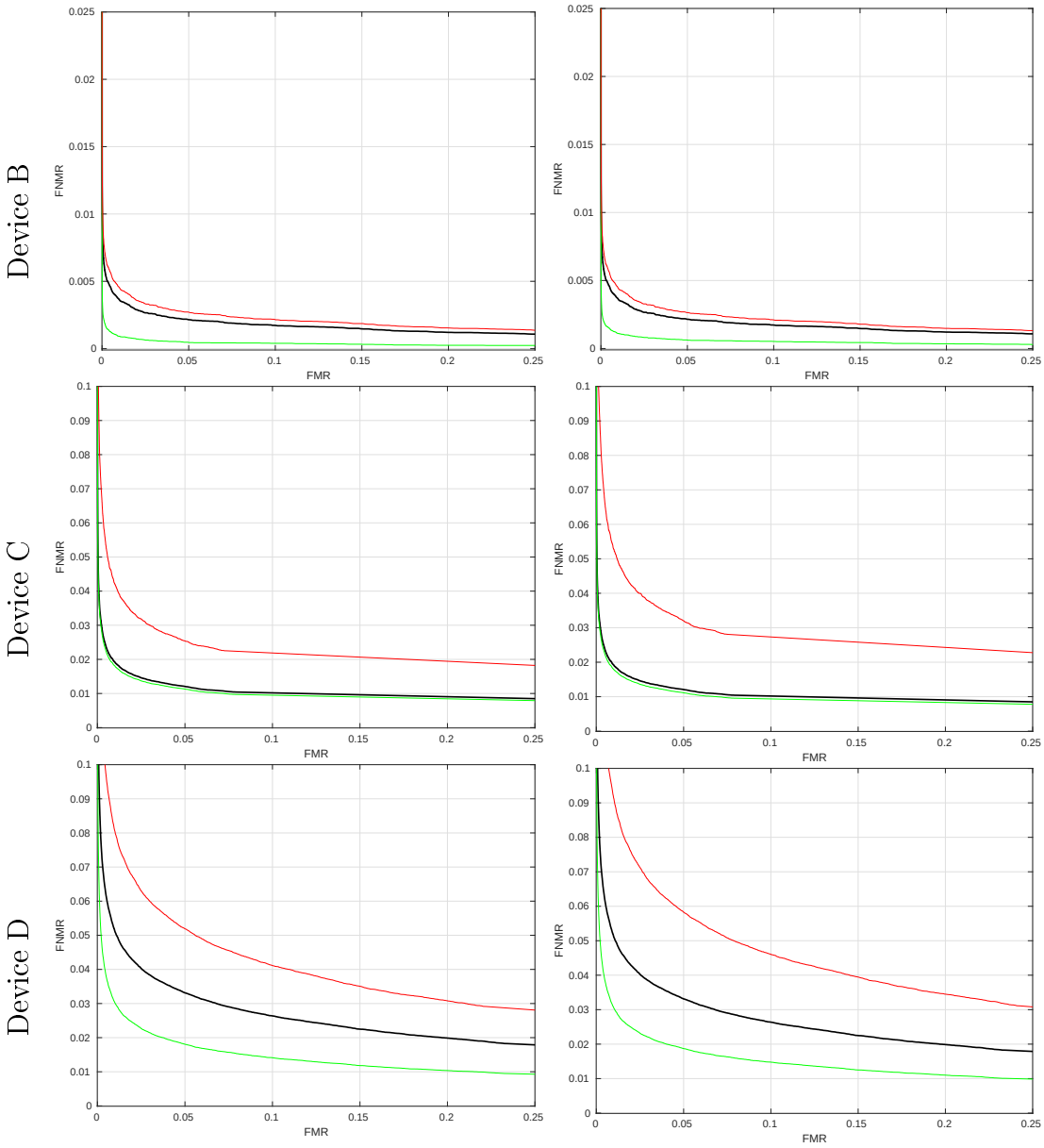
Table 6.2: Included percentages x of the highest quality NFIQ-1 class (Class 1) and the corresponding thresholds (upper x^{th} percentiles) for inclusion from NFIQ-2

Device ID	NFIQ-1		NFIQ-2	
	Enrol %	Verify %	Enrol Thresh.	Verify Thresh.
B.	93.2	83.4	71.3	68.0
C.	46.7	31.9	56.7	50.0
D.	75.8	58.2	73.3	65.0
F.	52.8	37.0	68.7	59.0
G.	50.7	36.8	60.0	54.0
H.	94.4	78.8	79.3	74.0
J.	51.9	36.2	52.3	47.0
K.	46.9	33.6	44.7	39.0
L.	92.6	83.0	63.3	57.0

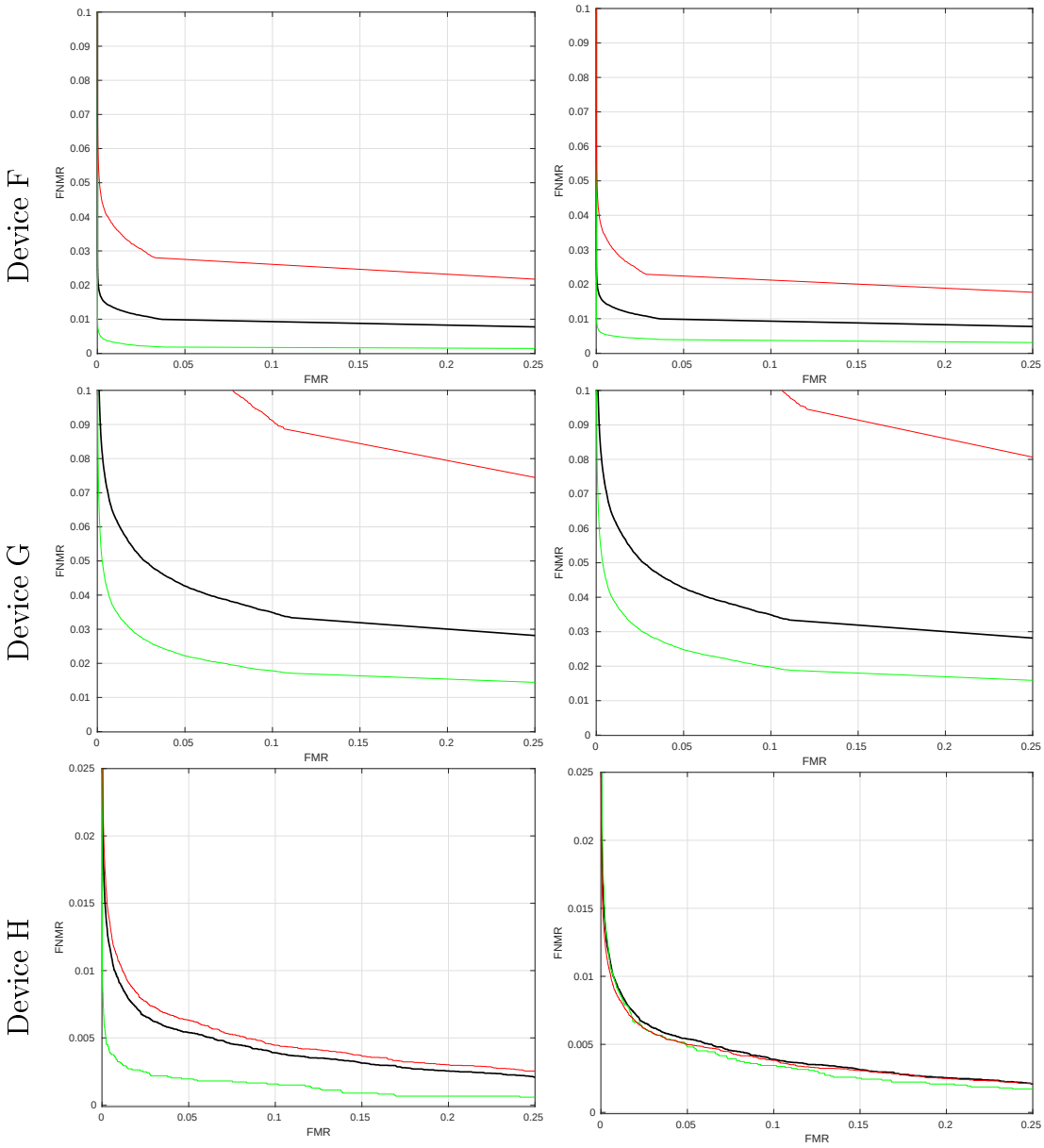
a difference can be noted anywhere, it is for the operationally less interesting scenario in which we have removed a fraction of the highest quality presentations (red curves): for this case, the overall classification performance appears to degrade slightly less when the quality is thresholded using NFIQ-2 rather than NFIQ-1.

In an operational context, it would be preferable to use a fixed device-dependent threshold that results in an acceptably low increase in FRR. From this point of view, the NFIQ-2 measure is superior to NFIQ-1 since it provides the necessary fine-grained quality control. Figure 6.8 compares the available improvement in classification performance for NFIQ-1 (discarding the entire lowest quality class) versus NFIQ-2 (discarding, respectively, the lowest fifth, tenth, fifteenth, and twentieth quality percentiles). Also shown, where available, are results using quality scores obtained from the vendors' enrolment records - although these are not strictly comparable since no attempt was made to threshold the corresponding verification qualities.

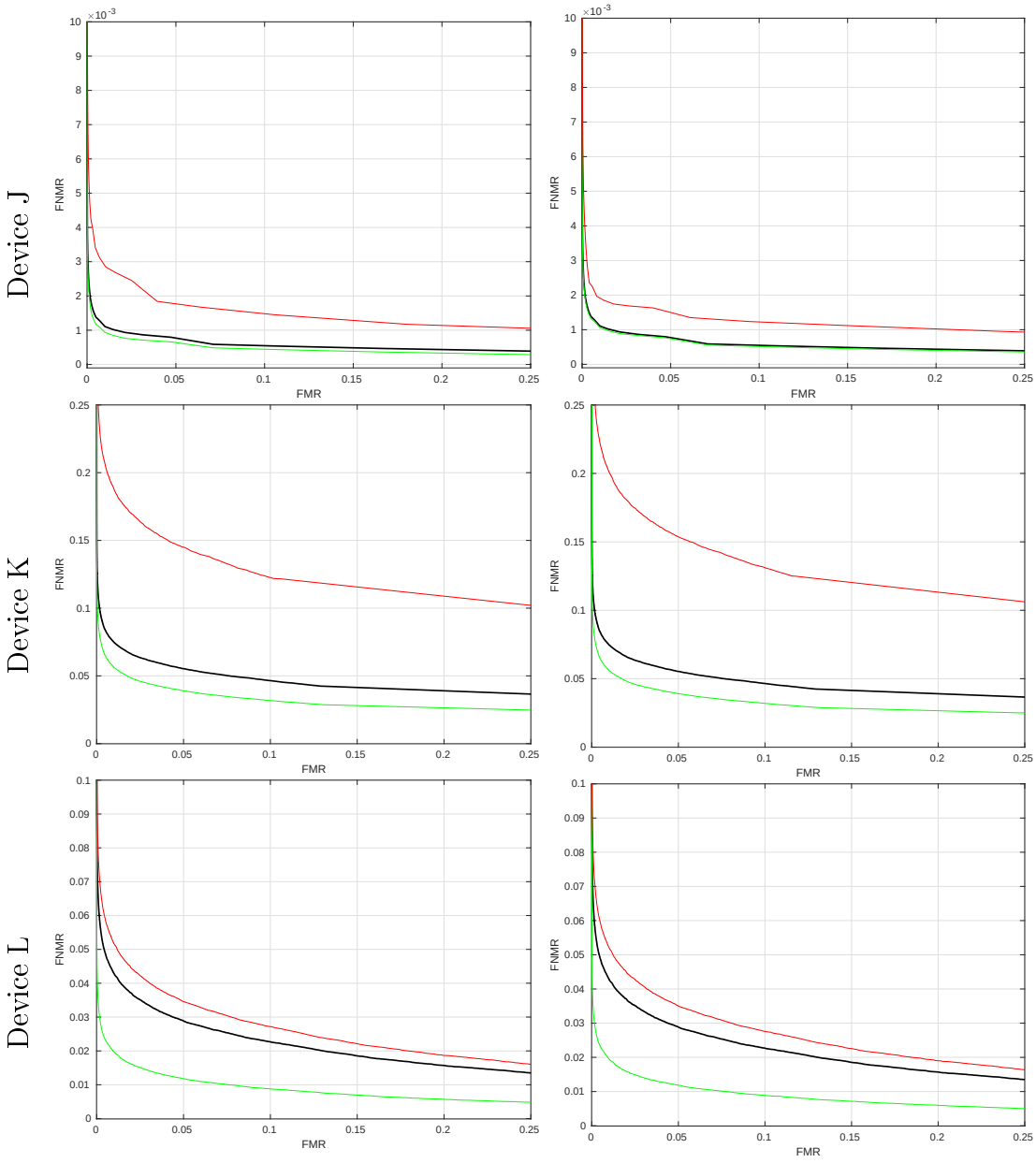
Note that DET curves versus vendor quality are not available for Devices F and H, since the vendor supplied qualities are not useful (see Figure 6.1). We also failed to obtain useful NFIQ-2 scores from Devices A and E. Results for the full datasets are shown in black, while those with the lower quality matches removed are shown in green, and those with the higher quality matches removed are in red. Curves were



(a) NFIQ1 **(b)** NFIQ2



(a) NFIQ1 **(b)** NFIQ2



(a) NFIQ1 **(b)** NFIQ2

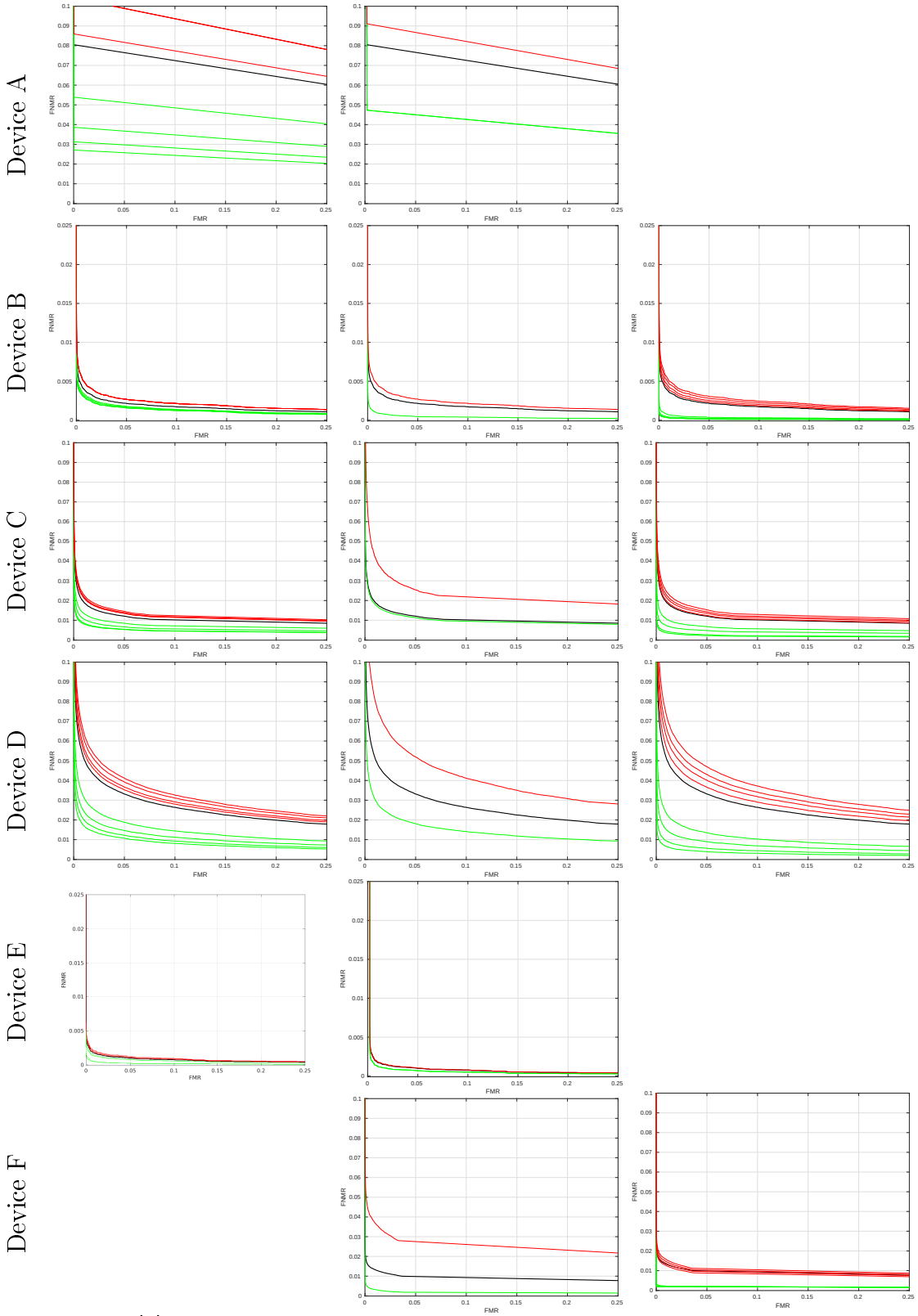
Figure 6.7: Comparison of classification performance by device versus quality for the NFIQ measures. For NFIQ-1, the DET is evaluated for NFIQ with all results (black), with lowest quality class removed (green) and with the highest quality class removed (red), noting in each case the percentage of cases removed. Corresponding NFIQ-2 results are obtained by thresholding the data at the same percentages as those recorded for NFIQ-1.

generated in MATLAB, using the perfcurve() function from the MATLAB Statistics and Machine Learning toolbox. Plots are scaled on a per-device basis to reflect the differing baseline performance of the devices.

In almost all cases, we can see that each of the three quality indices (vendor provided, NFIQ-1, and NFIQ-2) is effective to some extent at predicting classification accuracy. Vendor provided scores for devices B, E, and L appear to provide only marginal discrimination, while that for Device K actually appears to provide weak negative discrimination i.e. its overall classification performance *d*ecreases slightly when the reported lowest quality matches are removed, and *i*ncreases upon removal of the reported highest quality matches.

For Device C, NFIQ-1 appears to be effective at identifying the highest quality images, since removing score 1 cases clearly decreases overall classification accuracy. However it does not seem to provide good identification of (the more practically significant) low quality images - the DET with lowest quality (score 5) matches removed is indistinguishable from the baseline.

The device for which NFIQ-2 seems least effective is Device H, which is the multispectral device whose NFIQ-2 scores are probably unreliable, as discussed in the preceding sections. Otherwise, NFIQ-2 appears to be very effective at predicting match performance - most usefully, we observe decreasing returns after removal of the fifth percentile, suggesting that substantial benefit may be obtained by only light filtering of the IDMS cohort.



(a) Vendor **(b)** NFIQ1 **(c)** NFIQ2

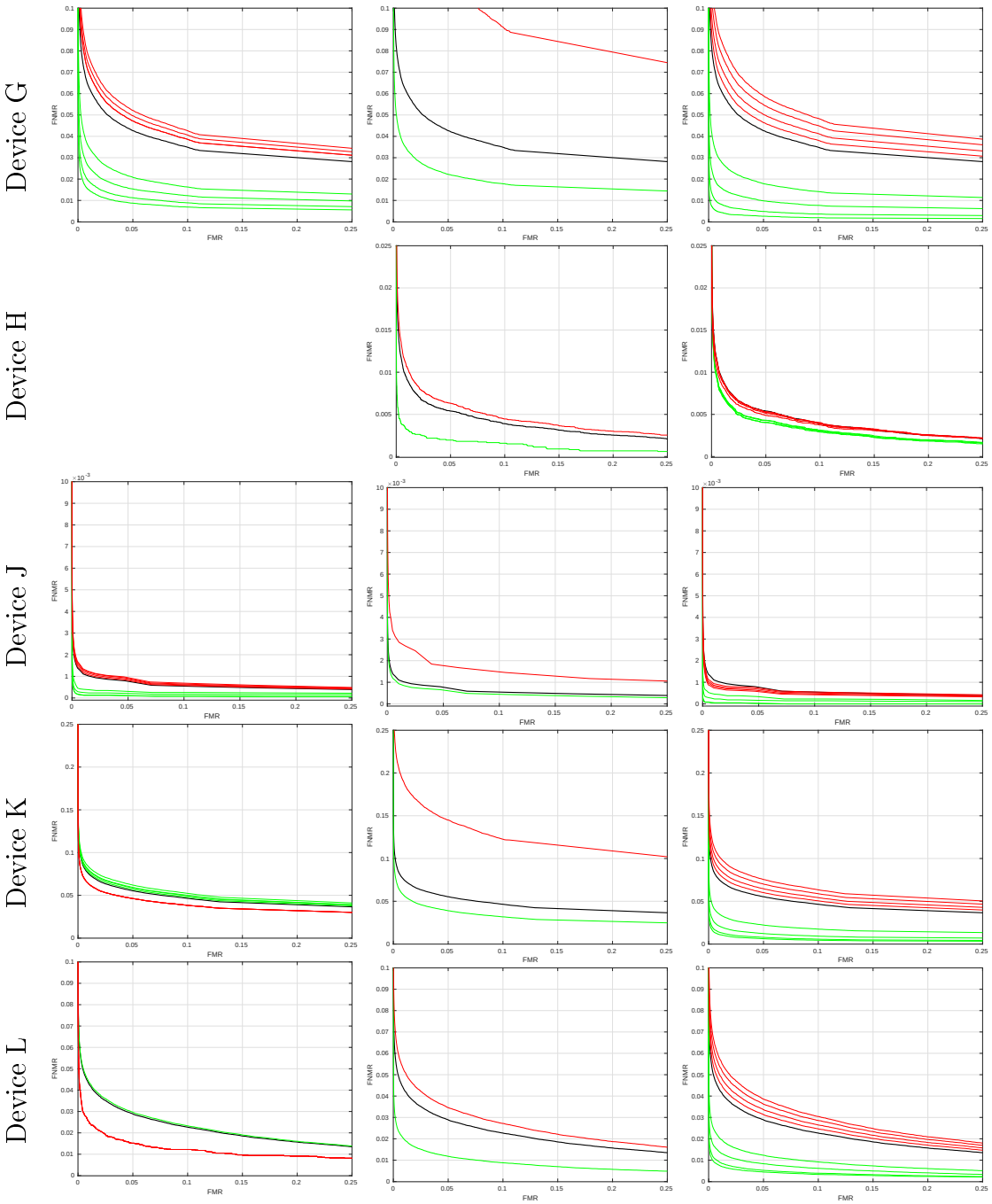


Figure 6.8: Decision Error Trade-off (DET) curves by device: results for the full datasets are shown in black, while those with the lower quality matches removed are shown in green, and those with the higher quality matches removed are in red. For NFIQ-1, classes 5 and 1 respectively were removed; for NFIQ-2, successively larger quality percentiles from 5% to 20% and from 80% to 95% were removed.

6.7 Discussion

The goal of the chapter was to understand the relationship between NFIQ-1 and NFIQ-1 scores (and, where available, vendor-specific quality scores obtained during template enrolment), and the relationship between these scores and the measured biometric performance - both in terms of raw match scores and overall classification accuracy.

We have seen that both NFIQ-1 and NFIQ-2 are effective predictive tools for genuine match score, across a broad range of fingerprint capture devices with different (vendor-dependent) image and match score processing.

We have shown how both may be used to validate or sanitize fingerprint presentations by rejecting a fraction of the lowest indicated quality, in order to improve the overall classification performance of a biometric IDMS with minimal degradation in convenience (FTE). When compared at the same fraction of rejected presentations (i.e. the same increase in FTE), NFIQ-1 and NFIQ-2 perform similarly in terms of the potential decrease in FNMR at a given FMR. However, the more expressive 1-100 quality scale of NFIQ-2 gives it the significant advantage of allowing the system integrator to choose a more precise quality threshold for exclusion.

Chapter 7

Identification and demographics of a biometric menagerie, and its effect on classification performance and template ageing

In the previous chapter, it was shown that both NFIQ-1 and NFIQ-2 measures are effective at identifying low quality fingerprints, and how selective rejection of such fingerprints can improve the overall classification performance of a fingerprint-based IDMS.

The goals of this chapter are twofold: first, to explore the demographics of fingerprint quality, in particular to see if we can identify any common demographic traits among the subjects whose fingerprints are identified by NFIQ as low quality. We attempt to put these individuals in the context of Doddington's biometric "zoo" [17]. Secondly, we re-visit the topic of template ageing (Chapter 5), and examine whether a relatively small cohort of low fingerprint quality individuals disproportionately affects template ageing behaviour.

The demographics of NFIQ-1/NFIQ-2 quality for the ISBIT/Norwood data set as a whole was reported recently by Rong (Roy) Yang in "Effects of sensors, age, and gender on fingerprint image quality" [85], Chapter 5 "Results and Discussions" and will not be revisited here. Briefly, it was found that both NFIQ-1 and NFIQ-2 quality scores were generally higher for males than for females, and for younger than for older subjects – with the exception of the multi-spectral Device H for which the opposite behaviour was observed.

7.1 Revisiting Doddington's zoo

7.1.1 Identification of a common Goat subset

Doddington's original classification of biometric "goats" was based on sorting in increasing order of genuine match score, and taking the lowest 2.5th percentile of individuals [17]. In our scenario, we must deal with match scores from multiple devices, whose vendor-specific scoring algorithms do not permit an overall ordering. Instead, we developed the following classification scheme.

First, choose a common target FNMR across all devices: preferably this should be small enough to be operationally reasonable, yet large enough that it gives a measurable count of false non-matches on every device[1]. More precisely, we would like to have enough false non-matches to make a low-variance estimate of the FNMR. Then, for each device, determine the decision threshold corresponding to the target FNMR, and evaluate a confusion matrix at that threshold for each subject-finger, taken over all the available visit pairs. Note that there is some multiplicity here since the protocol for a verify visit consists of two verification attempts of three presentations each: this helps to fill out the confusion matrices in spite of the relatively

[1]The number of genuine matches per visit pair is relatively small due to the limited number of subjects, and not all devices were present in every visit

small number of visit pairs.

At this point, Device J was eliminated because it was not possible to determine a suitable threshold for the target 5 % FNMR, and Device A because, although a putative threshold was determined, the actual FNMR at that threshold differed significantly from the target FNMR, for reasons as yet undetermined.

The true match (TM), false non-match (FNM), false match (FM) and true non-match (TNM) counts for the remaining devices were then aggregated, and the subject-fingers sorted in order by number of false non-matches (FNM), from highest to lowest. We then defined as the "goat set" the set of subject fingers that together accounted for 50 % of the total false non-matches. Note that although FNMR values in excess of 55 % are observed for individual subject-fingers, the overall FNMR for each device, as well as for the aggregate subject-finger set as a whole, is constrained to the target 5 % (aggregate TM = 722895, FNM = 38081).

The choice of 50 % is somewhat arbitrary, however it results in a conveniently-sized subset of 29 out of the 879 available unique subject-fingers (Table 7.1). For a key to the numerical finger position identifiers used, see Table 7.2.

A question that arises naturally is whether the "Goat set" established by this procedure reflects an aggregation of substantially disjoint sets of high FNMR matches from each separate device, or whether it indicates a set of subject-fingers that is intrinsically susceptible to false non-matches regardless of device, and whose members in some sense therefore provide less *information* about the identity of the subject.

To explore this question, the confusion matrices of the individual devices were ordered by FNM in the same manner as for the aggregate set, and their own goat sets constructed as the subject-fingers that accounted for 50 % of the total false non-matches of the device. A search was then made for each of the aggregate Goat subject-fingers in each of the individual devices' Goat sets.

Figure 7.1 shows the counts of how many devices each Goat appears in, clearly

Table 7.1: Subject-fingers with the highest number of false non-matches over all devices and visits, and which together constitute 50% of the false non-matches when the overall FNMR is constrained to 5 %: we refer to these as the Goat set.

SID	BirthYear	EthnicOrigin	Gender	Finger	TM	FNM	FM	TNM	Cum. FNM	% of tot.	FNMR	FMR
159	1939	North America	Female	7	1222	1360	2392	514206	1360	4	52.7%	0.5%
10	1961	Europe	Female	2	1017	1213	6815	465895	2573	7	54.4%	1.4%
10	1961	Europe	Female	7	963	1194	4177	468323	3767	10	55.4%	0.9%
163	1954	Philippines	Female	1	1607	971	363	35431	4738	13	37.7%	1.0%
231	1952	North America	Female	7	307	969	2982	310538	5707	15	75.9%	1.0%
196	1946	Asia	Female	7	330	886	5967	315413	6593	18	72.9%	1.9%
231	1952	North America	Female	2	565	847	3466	348168	7440	20	60.0%	1.0%
232	1962	Europe	Female	2	437	846	3818	327955	8286	22	65.9%	1.2%
134	1948	North America	Male	2	1226	766	3254	413226	9052	24	38.5%	0.8%
232	1962	Europe	Female	7	655	757	3632	340346	9809	26	53.6%	1.1%
114	1942	North America	Female	7	1577	713	6808	455371	10522	28	31.1%	1.5%
38	1952	North America	Female	7	1748	706	4754	486538	11228	30	28.8%	1.0%
67	1949	North America	Female	7	2237	545	3566	523566	11773	32	19.6%	0.7%
59	1944	Europe	Male	2	2215	542	2910	525803	12315	33	19.7%	0.6%
153	1965	Europe	Female	2	2320	542	5409	532009	12857	35	18.9%	1.0%
18	1961	North America	Female	2	2416	530	3034	543108	13387	36	18.0%	0.6%
59	1944	Europe	Male	7	1939	495	6090	485243	13882	37	20.3%	1.2%
34	1945	Europe	Male	2	513	489	2155	310379	14371	39	48.8%	0.7%
240	1971	North America	Female	7	1286	478	4362	419814	14849	40	27.1%	1.0%
19	1952	Asia	Female	7	2341	457	7727	517367	15306	41	16.3%	1.5%
38	1952	North America	Female	2	2393	439	8844	528604	15745	42	15.5%	1.6%
34	1945	Europe	Male	7	752	434	2122	327457	16179	43	36.6%	0.6%
36	1955	North America	Female	7	2126	412	7455	511419	16591	45	16.2%	1.4%
146	1942	North America	Male	2	2153	379	3564	499018	16970	46	15.0%	0.7%
60	1950	Europe	Female	2	2508	378	5593	533907	17348	47	13.1%	1.0%
91	1964	Asia	Female	2	2067	363	5410	472004	17711	48	14.9%	1.1%
57	1958	Europe	Female	7	2416	356	5396	520075	18067	49	12.8%	1.0%
163	1954	Philippines	Female	7	2490	343	7274	528729	18410	49	12.1%	1.4%
60	1950	Europe	Female	7	2580	336	10096	532304	18746	50	11.5%	1.9%

Table 7.2: Finger positional identification ("FingerPosition") numbers

	Right hand	Left hand
Thumb	1	6
Index	2	7
Middle	3	8
Ring	4	9
Pinky	5	10

showing that the sets are far from disjoint. For example Subject 159 Finger 7 (left index), which was responsible for more false matches than any other finger in the aggregate set, is also in the Goat set of 9 out of 10 of the individual devices. Only one of the aggregate Goats (Subject 36 Finger 7) is not visible in the goat set of more than a single device.

7.1.2 Demographics of the Goat subset

Having identified a subset of subject-fingers that contribute disproportionately to the FNMR, it becomes natural to ask whether there is anything demographically interesting about this subset.

The ISBIT/Norwood database provides a limited amount of self-reported demographic information: birth year at first enrolment (an obvious proxy for subject age), sex[2], self-reported ethnic origin, and a single field indicating whether, at first enrolment, the individual considered their fingers to have been subject to "manual or chemical exposure". While recruitment criteria ensured a good sex balance between subjects ($M = 180$, $F = 178$), it proved harder to ensure population-representative subject ages: in practice, the distribution is somewhat bimodal, with an older cohort carried over from the original ISBIT study and a younger cohort recruited in the later phases largely from the Carleton student population. No attempt was made to balance the manual or chemical exposure classes.

[2]Nominally "gender", however no choice beyond 'M' or 'F' was provided

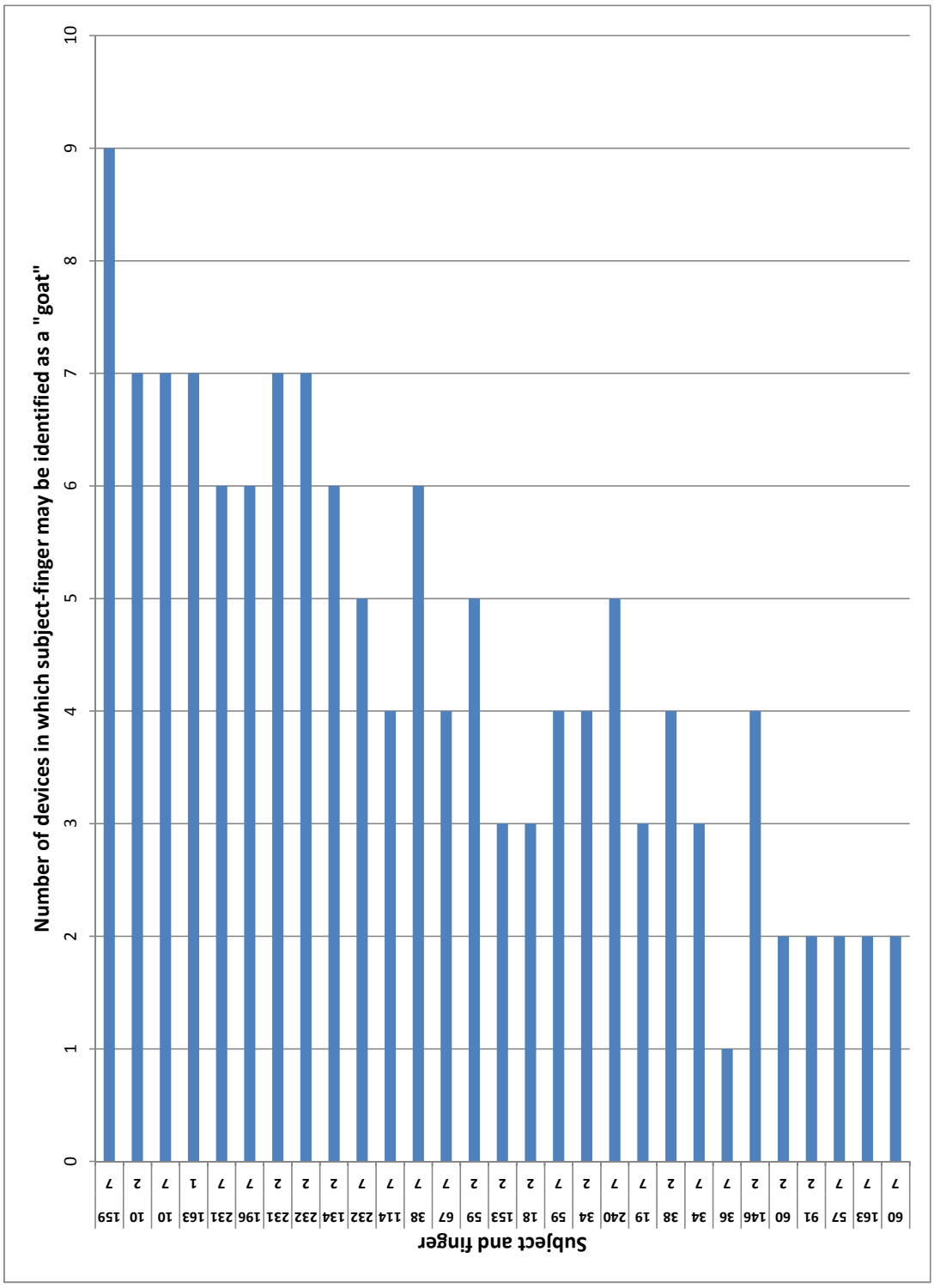


Figure 7.1: Consistency of the aggregate Goats across devices. For example, Subject 159 Finger 7 (left index), which was responsible for more false matches than any other finger in the aggregate set, is also in the goat set of 9 out of 10 of the individual devices. Only one of the aggregate goats (Subject 36 Finger 7) is not visible in the goat set of more than a single device.

When comparing the demographics of the Goat subset to those of the subject set as a whole, a subtlety of the data collection protocol must be addressed: that of the preferred fingers for enrolment (Chapter 3). In order to provide as many valid matches as possible, it was desirable to enrol, if possible, the same fingers across all devices and all visits: the right and left index fingers (finger positions 2 and 7) were selected by default. However if a preferred finger was unavailable (because of injury for example), or failed to enrol on a particular device, enrolment of other fingers was attempted in a pre-defined order. Because the study spanned multiple devices and multiple visits over which the successfully enrolled fingers might change, the number of distinct enrolled subject-fingers is not simply equal to twice the size of the subject set. In fact, over the 358 subjects we enrolled 879 distinct subject-fingers; approximately 75 % of subjects enrolled only two fingers each, with approximately a further 14 % enrolling a third finger at some point, and two particularly egregious subjects enrolling, respectively, seven and eight fingers (Figure 7.2).

In the present context, what is more significant than the number of enrolled subject fingers is their demographic distribution. For example, while there is almost perfect sex-balance among *subjects*, females are apparently over-represented ($M = 410$, $F = 469$) in the set of enrolled *subject-fingers* (Figure 7.3) – and hence also over-represented in the set of available biometric matches. If we compare the demographics of the Goats against that of the subjects, we risk confounding the effect of the demographic variable with that of multiplicity of enrolment; on the other hand, when multiplicity of enrolment results from FTE events, it would not be surprising to find some of those fingers among the Goat set – since fingers that are hard to enrol might be expected to be hard to match. FTEs are believed to contribute the majority of the multiple enrolments in the ISBIT/Norwood data: only three of the subjects whose fingers appear in the Goat set were recorded as having had unavailable fingers for one or more visits of the study.

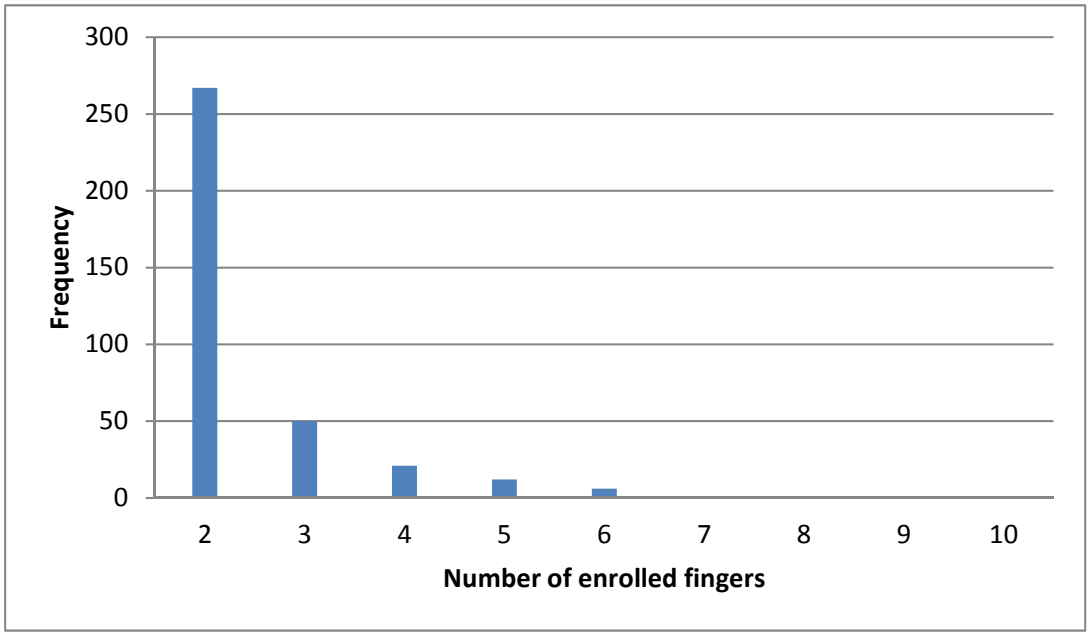


Figure 7.2: Distribution of numbers of enrolled fingers per subject. Most subjects (approximately 75 %) enrolled just two fingers over the course of the study, however a few subjects enrolled additional fingers as a result of unavailability of a previously-enrolled finger, or failure to enrol a preferred finger on a given device.

In the remainder of this section we use the demographics of the 879-member enrolled subject-finger set as the baseline for the comparisons.

Figures 7.4 – 7.7 show the empirical distributions of these demographics for the aggregate Goat subset identified in the preceding section, compared with the subject-finger cohort as a whole. We see compelling evidence that the Goats come predominantly from the older individuals (birth years before 1970 - so aged at least 35-45 during the phases of the study) and predominantly from female subjects.

We may implement Fisher's exact test for significance on the sex-imbalance of the Goats as follows [35]. Suppose we are given a population of size N comprising K females and $N - K$ males. Under the null hypothesis H_0, we pool the males and females together and then select, randomly and without replacement, a sample of size n, finding that we have k females and $n - k$ males. The probability mass function of

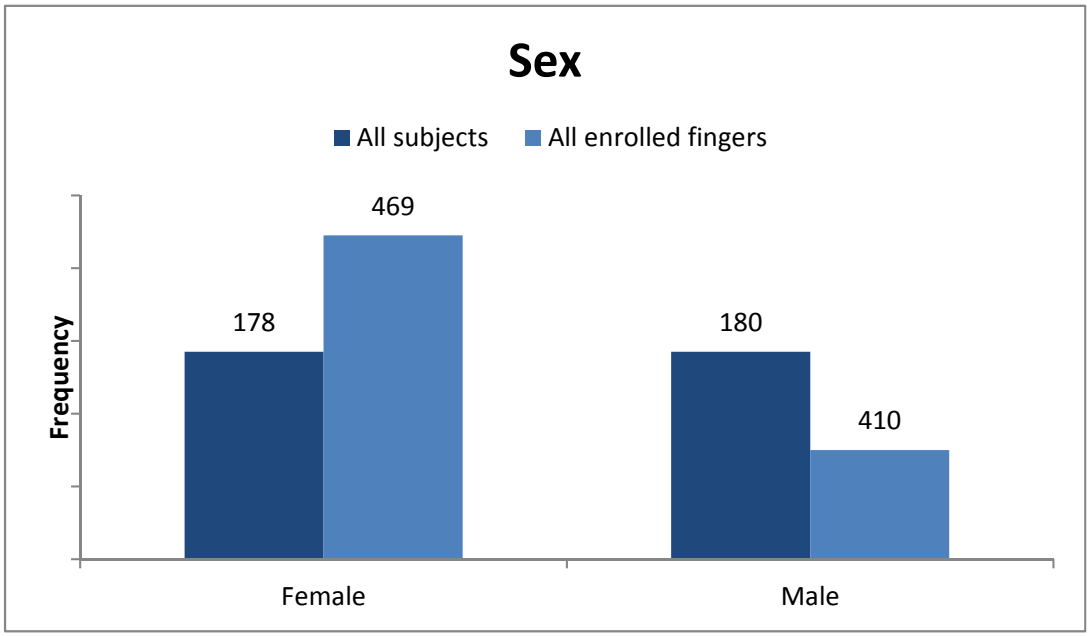


Figure 7.3: Distribution of subject sex for the set of enrolled subject-fingers (light blue) compared to the subject set as a whole (dark blue). The difference arises from a larger number of multiple finger enrolments among female subjects.

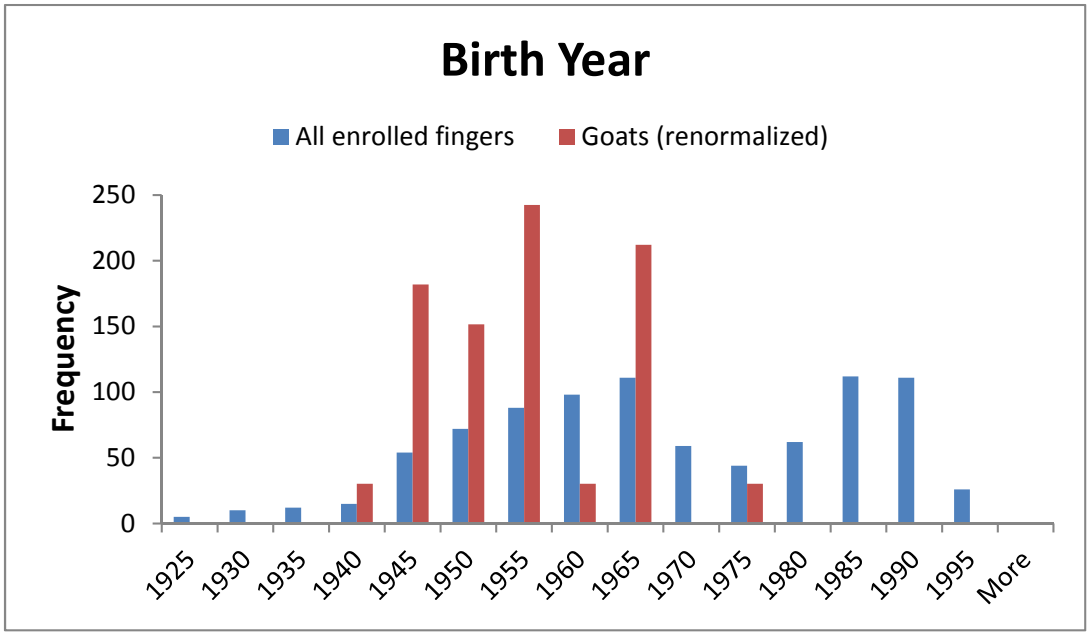


Figure 7.4: Distribution of subject birth year for the Goat subset (red) compared to the enrolled set as a whole (blue). Almost all of the goats appear to come from the older population.

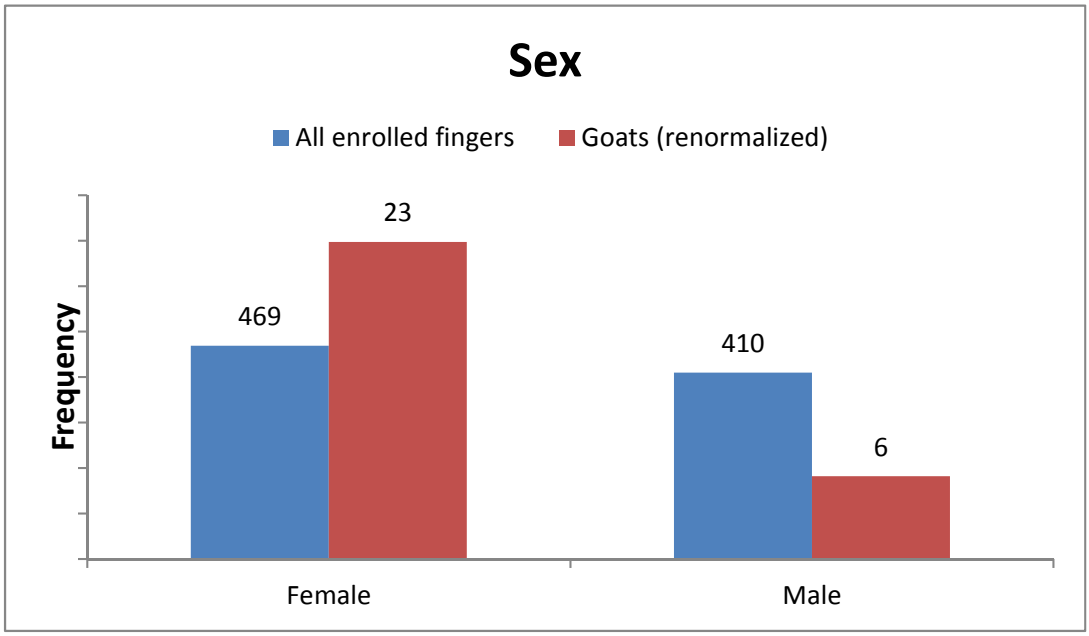


Figure 7.5: Distribution of subject sex for the Goat subset (red) compared to the enrolled set as a whole (blue), with raw counts shown at the top of the columns: while the study is closely sex-balanced overall, the fingers of female subjects dominate the Goats.

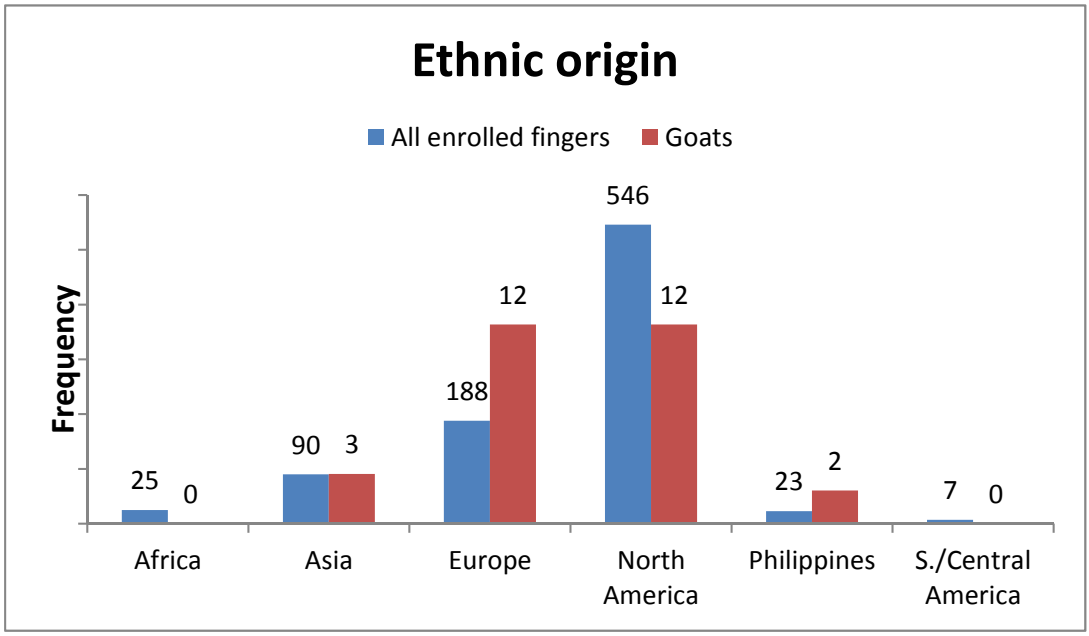


Figure 7.6: Distribution of subject ethnic origin for the Goat subset (red) compared to the enrolled set as a whole (blue), with raw counts shown at the top of the columns.

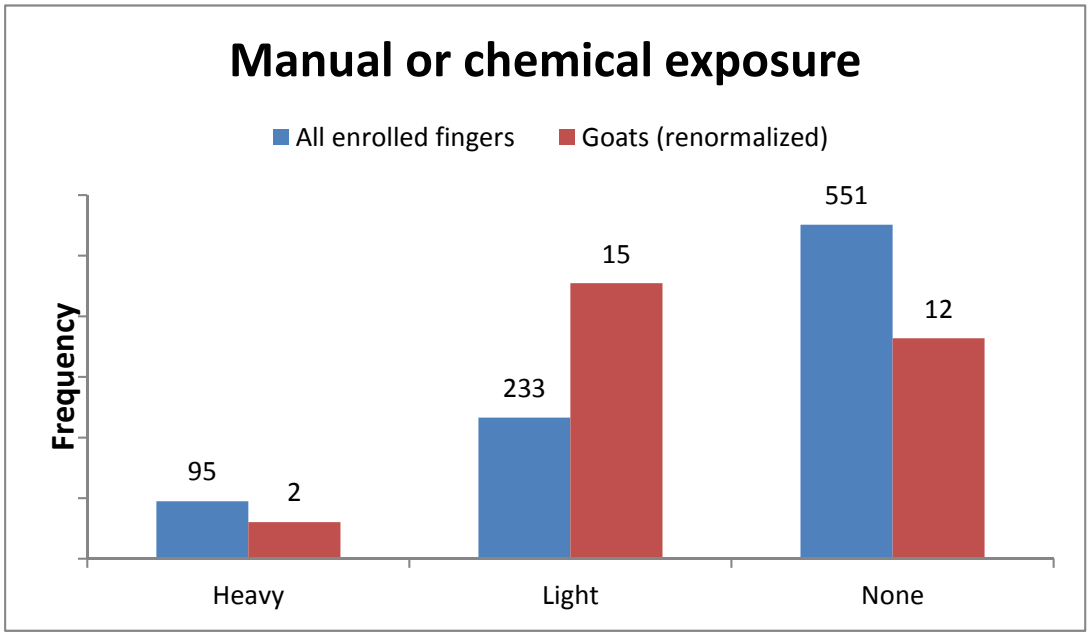


Figure 7.7: Distribution of subject manual or chemical exposure for the Goat subset (red) compared to the enrolled set as a whole (blue), with raw counts shown at the top of the columns: 'Light' exposure appears to have more effect than 'Heavy' exposure.

the number of females selected, k, is given by

$$p_X(k) = Pr(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}} \tag{7.1}$$

and we may write the p-value as

$$p = 1 - \mathcal{H}(N, K, n, k) \tag{7.2}$$

where $\mathcal{H}$ is the cumulative hyper-geometric distribution.

In the case of the sex ratio of the Goat subset, we have $k = 23$ female subject-fingers out of a total of $n = 29$ categorized Goats. The number of distinct subject-fingers in the ISBIT/Norwood data set is $N = 879$, of whom $K = 469$ belong to females: the p-value is evaluated as $p = 0.0008$ (Table 7.3). We may therefore reject the null hypothesis that the Goats come from the identical sex distribution as the enrolled set as a whole.

Table 7.3: Fisher exact test for the significance of subject sex.

	All	Goats	Expected under H_0
Female	469	23	15.47
Male	410	6	13.53
Total	879	29	p-value: 0.0008

The other demographic variables each have more than two categories, making exact tests of this form unfeasible, and we fall back to chi-squared tests of significance. That is, we calculate expected category counts for the Goat subset assuming the same proportions as the enrolled set as a whole, and then evaluate the χ-squared probability to determine whether the expected and observed counts differ significantly. The results of these tests are summarized in Tables 7.4, 7.5, 7.6. Consistent with visual examination of the histograms (Figures 7.4 - 7.7), there is strong evidence in favour of the effect of subject birth year ($p = 0.00024$), and weak evidence in favour of the effect of ethnic origin ($p = 0.0619$). The evidence for the effect of manual or chemical exposure is a little harder to interpret: the χ-squared $p = 0.00885$ allows us to reject the null at the Bonferroni-corrected significance level $\alpha = 0.0125$, however the exact *nature* of the relationship is not clear, since 'Light' exposure appears to have more effect than 'Heavy' exposure. This may be because the available exposure categories were not sufficiently objective – or may reflect the fact that the chi-squared test is not really appropriate for ordinal (as opposed to strictly categorical) data.

An interesting open question is the extent to which demographic differences in biometric match performance are intrinsic, and to what extent they reflect training bias in the development of the matching algorithms. For example, algorithms trained predominantly on latent fingerprint data from forensic databases might be expected to be biased in favour of younger males. On the other hand, it is not unreasonable to suppose that younger individuals in general might have better fingerprints (as a result of less opportunity for physical damage, and of physiological factors such as skin elasticity) than older individuals.

Table 7.4: χ-squared goodness-of-fit test for the significance of subject birth year at time of enrolment.

	All	Goats	Expected under H_0
1925	5	0	0.16
1930	10	0	0.33
1935	12	0	0.40
1940	15	1	0.49
1945	54	6	1.78
1950	72	5	2.38
1955	88	8	2.90
1960	98	1	3.23
1965	111	7	3.66
1970	59	0	1.95
1975	44	1	1.45
1980	62	0	2.05
1985	112	0	3.70
1990	111	0	3.66
1995	26	0	0.86
More	0	0	0.00
Total	879	29	p-value: 0.00024

Table 7.5: χ-squared goodness-of-fit test for the significance of subject self-reported ethnic origin.

	All	Goats	Expected under H_0
Africa	25	0	0.82
Asia	90	3	2.97
Europe	188	12	6.20
North America	546	12	18.01
Philippines	23	2	0.76
S./Central America	7	0	0.23
Total	879	29	p-value: 0.0619

Table 7.6: χ-squared goodness-of-fit test for the significance of subject manual or chemical exposure.

	All	Goats	Expected under H_0
Heavy	95	2	3.13
Light	233	15	7.69
None	551	12	18.18
Total	879	29	p-value: 0.0088

In addition to algorithm training, bias might be introduced at the feature extraction phase, either physically (for example, sensor devices whose size and resolution is better matched to a specific demographic) or via the choice of which features to extract for template generation.

7.1.3 NFIQ quality of the Goat subset

In the preceding Chapter, we considered the relationship between NFIQ quality and overall classification performance, but did not attempt to investigate whether low quality is a more significant factor for false matches or for false non-matches. Having now identified a subset of fingers that contribute disproportionately to the FNMR, we can approach the relationship from the other side and examine the relationship more specifically between false non-matches and quality.

Figures 7.8 and 7.9 compare the mean NFIQ-1 and NFIQ-2 scores for the Goat subset versus all subject-fingers at enrolment and verification, broken down by the available devices. With the exception of Device H (whose NFIQ results we believe to be erroneous) we observe consistently worse quality (NFIQ-1 higher and NFIQ-2 lower) amongst the Goats.

Although the set of Goat fingers is relatively small (29), the sets of images for which we have NFIQ scores are quite large, even when broken down by device. For enrolment, there would be in the ideal case three images of each finger for each of eight visits, giving a total of $29 \times 24 = 696$ samples per finger per device[3]. In cases where devices were removed in later visits the sets are somewhat smaller with the minimum sample size being 328 images, for Device J. Verification image sets are a factor two larger since the protocol demands two verification attempts, of three presentations each. The baseline (all enrolled finger) sets are substantially larger again - typically around 9000 images per device for enrolment and 18,000 per device for verification.

[3]The largest we see in practice is 648, because not all of the goats were enrolled at every visit on every device

Hence the standard errors associated with the mean quality values are typically very small, and the large numbers of degrees of freedom justify using z-tests in preference to t-tests for the difference of mean quality scores, where the standard normal test statistic z^* is given by

$$z^* = \frac{M_A - M_G}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_G^2}{n_G}}} \tag{7.3}$$

in which the subscripts G and A denote the Goat set and the set of all enrolled fingers respectively and M and s are sample means and standard deviations.

One-tailed tests for the difference of means were performed for all cases, equivalent to the null hypothesis that Goat qualities are drawn from the same distribution as the population as a whole versus alternate hypothesis that Goats have lower biometric quality[4]. The p-values are summarized in Tables 7.7 and 7.8: taking a Bonferroni-corrected significance level $\alpha = 0.00125$ (there are a total of 40 tests: 11 devices in the NFIQ-1 tests and 9 in the NFIQ-2 tests, times 2 for enrolment and verification[5]) we may reject, at 5 % family-wise error rate (FWER), the null hypothesis in all cases except for the NFIQ-2 scores on troublesome Device H (which appear to be better for the Goats).

Although the focus of this section is NFIQ, it is interesting to compare the vendor-reported quality metrics for Goats versus subject-fingers as a whole, where available. The vendor-supplied metric sets are somewhat smaller since they provide only a single quality value over the three captured images, but sample sizes are at least 110 for the Goat set and more than 2000 for the baseline. Figure 7.10 shows the mean vendor-reported enrolment image quality by device, together with the mean number of extracted minutiae. Aside from Devices F and H, whose vendor-reported quality

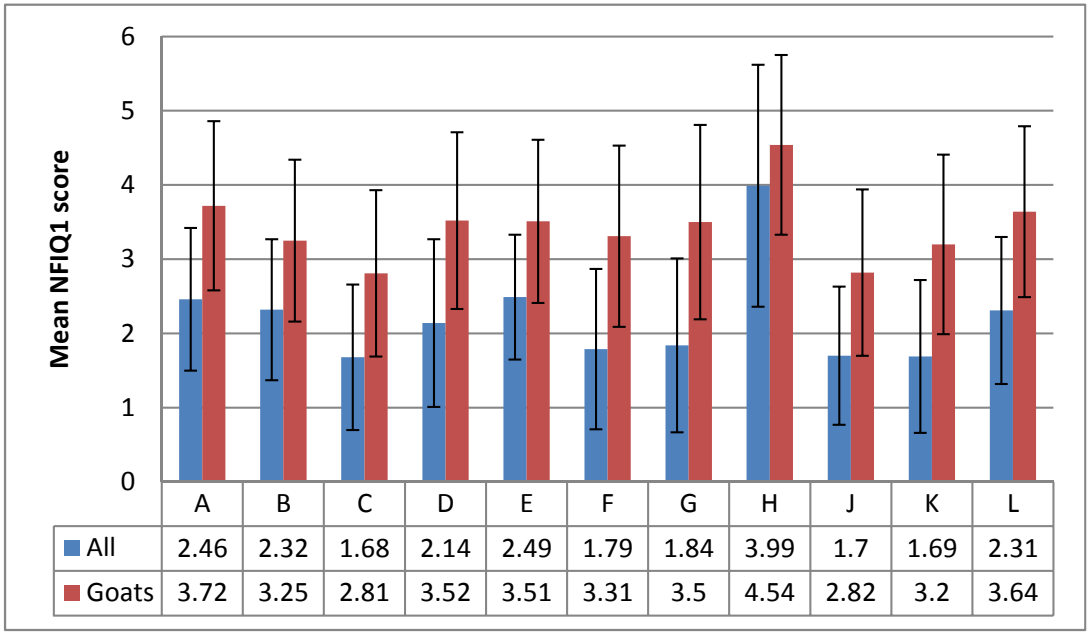
[4]For the case of NFIQ-1, whose scale is reversed, we modified the numerator of the statistic to $M_G - M_A$ so that all tests were right-tailed

[5]This correction is likely over-conservative since we expect correlation between NFIQ-1 and NFIQ-2 scores, and between enrolment and verification scores of the same finger.

Table 7.7: One-tailed tests of significance for the difference of mean NFIQ-1 between Goats and subject-fingers as a whole.

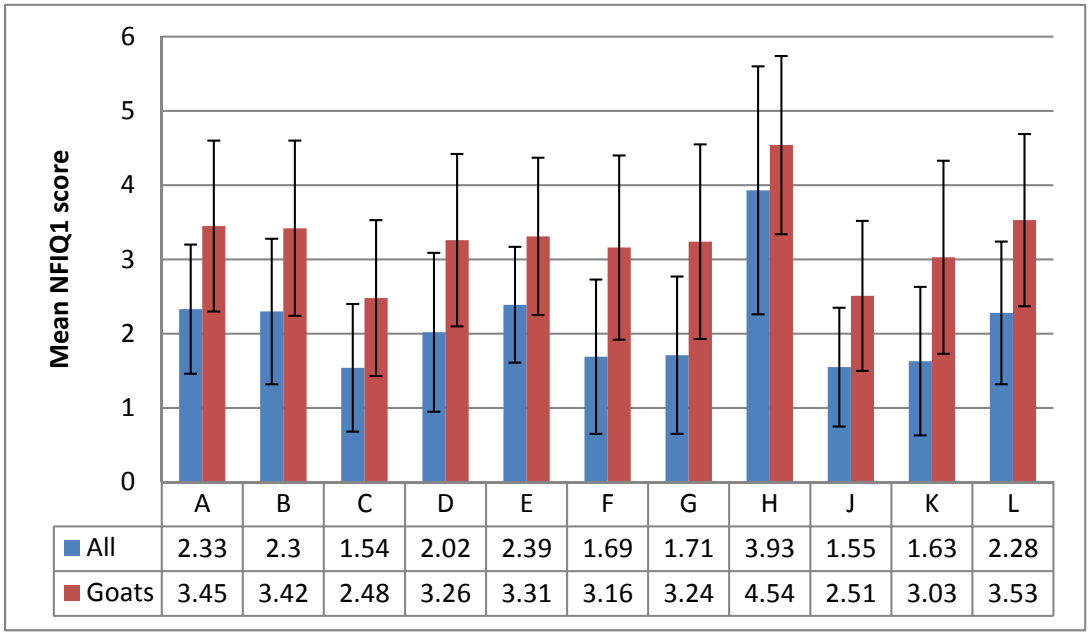
Device	Enrolment			Verification		
	$M_G - M_A$	SE	p-value	$M_G - M_A$	SE	p-value
A	1.260	0.050	$< 10^{-15}$	1.120	0.038	$< 10^{-15}$
B	0.930	0.051	$< 10^{-15}$	1.120	0.037	$< 10^{-15}$
C	1.130	0.055	$< 10^{-15}$	0.940	0.039	$< 10^{-15}$
D	1.380	0.049	$< 10^{-15}$	1.240	0.034	$< 10^{-15}$
E	1.020	0.057	$< 10^{-15}$	0.920	0.039	$< 10^{-15}$
F	1.520	0.049	$< 10^{-15}$	1.470	0.035	$< 10^{-15}$
G	1.660	0.060	$< 10^{-15}$	1.530	0.043	$< 10^{-15}$
H	0.550	0.058	$< 10^{-15}$	0.610	0.041	$< 10^{-15}$
J	1.120	0.063	$< 10^{-15}$	0.960	0.043	$< 10^{-15}$
K	1.510	0.049	$< 10^{-15}$	1.400	0.037	$< 10^{-15}$
L	1.330	0.046	$< 10^{-15}$	1.250	0.033	$< 10^{-15}$

values are invariant, we again see evidence of lower quality amongst the Goats by these metrics. Table 7.9 summarizes results of one-tailed tests by device: we accept the hypothesis that Goats have lower vendor quality in all but 2 of the 9 applicable cases, but only find evidence for lower minutia count in Devices A, B and E.



	A	B	C	D	E	F	G	H	J	K	L
All	2.46	2.32	1.68	2.14	2.49	1.79	1.84	3.99	1.7	1.69	2.31
Goats	3.72	3.25	2.81	3.52	3.51	3.31	3.5	4.54	2.82	3.2	3.64

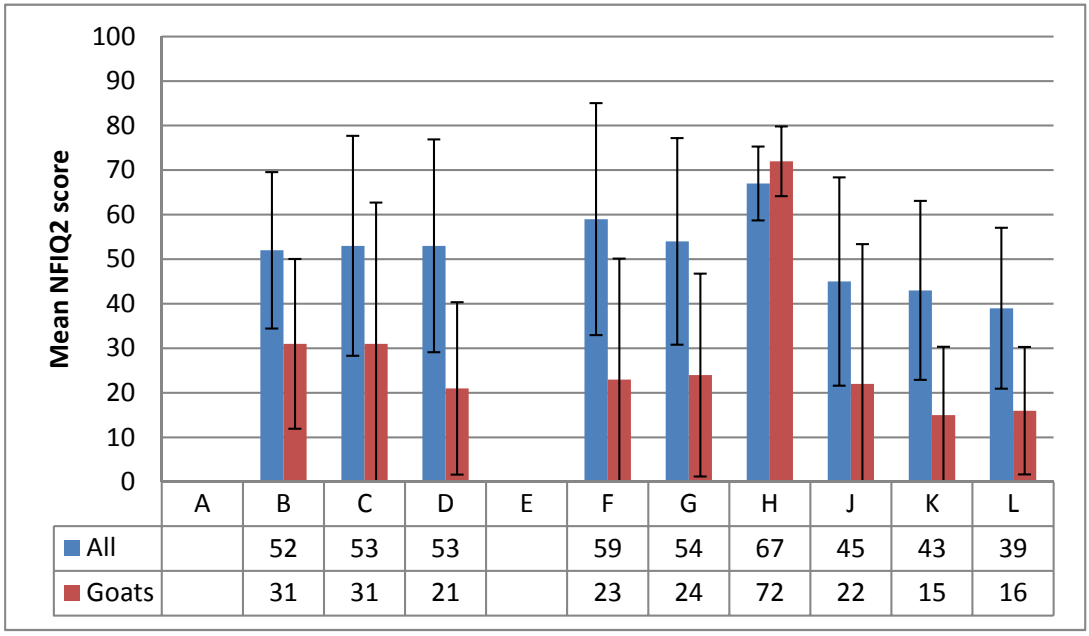
(a) Enrolment



	A	B	C	D	E	F	G	H	J	K	L
All	2.33	2.3	1.54	2.02	2.39	1.69	1.71	3.93	1.55	1.63	2.28
Goats	3.45	3.42	2.48	3.26	3.31	3.16	3.24	4.54	2.51	3.03	3.53

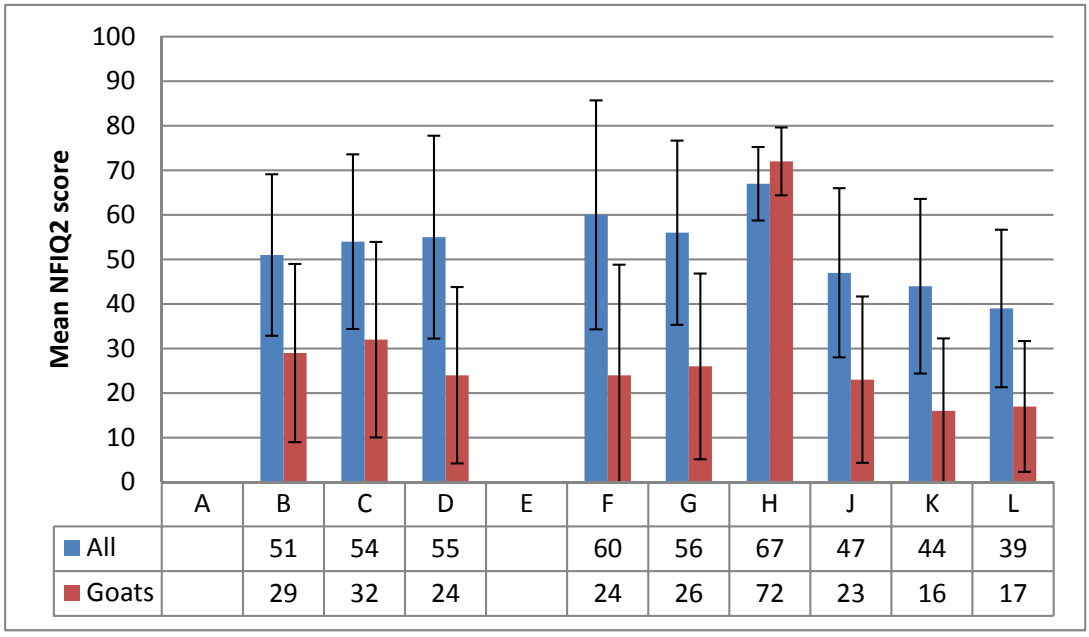
(b) Verification

Figure 7.8: Comparison of mean NFIQ1 score (lower = better quality) by device for the Goat subset (red) versus the enrolled set as a whole (blue). Error bars at ±1 standard deviation for each device.



	A	B	C	D	E	F	G	H	J	K	L
All		52	53	53		59	54	67	45	43	39
Goats		31	31	21		23	24	72	22	15	16

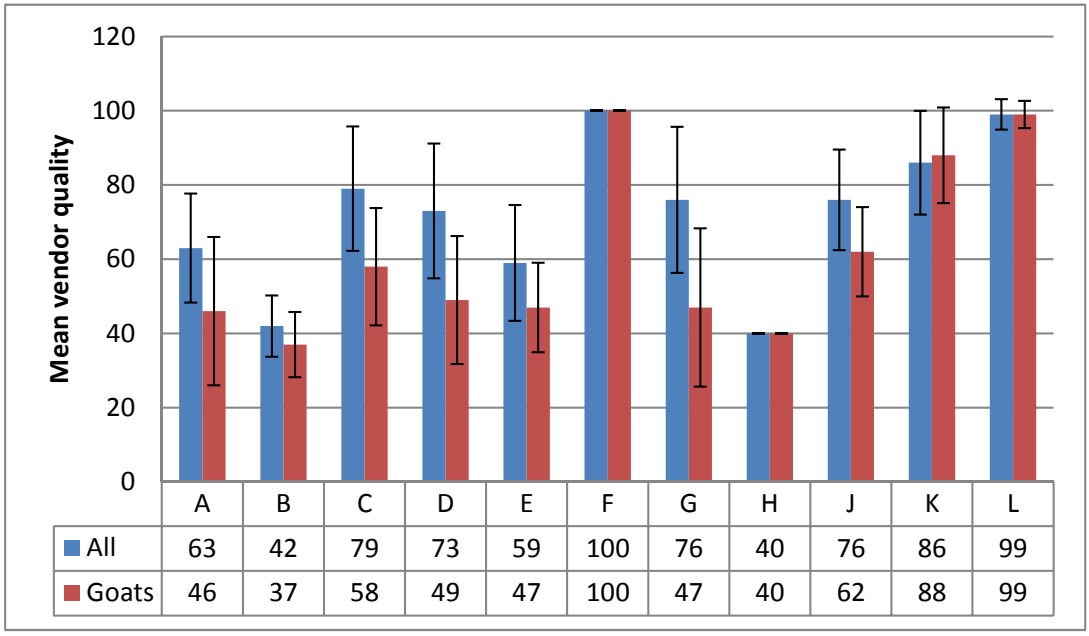
(a) Enrolment



	A	B	C	D	E	F	G	H	J	K	L
All		51	54	55		60	56	67	47	44	39
Goats		29	32	24		24	26	72	23	16	17

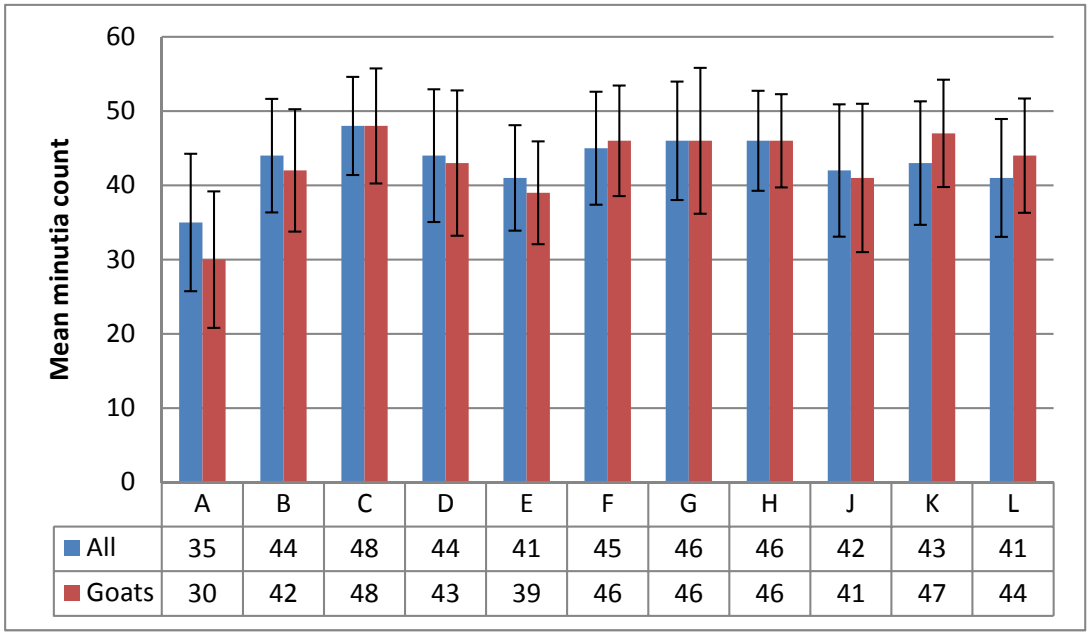
(b) Verification

Figure 7.9: Comparison of mean NFIQ2 score (higher = better quality) by device for the Goat subset (red) versus the enrolled set as a whole (blue). Error bars at ±1 standard deviation for each device.



	A	B	C	D	E	F	G	H	J	K	L
All	63	42	79	73	59	100	76	40	76	86	99
Goats	46	37	58	49	47	100	47	40	62	88	99

(a) Enrolment quality



	A	B	C	D	E	F	G	H	J	K	L
All	35	44	48	44	41	45	46	46	42	43	41
Goats	30	42	48	43	39	46	46	46	41	47	44

(b) Minutia count

Figure 7.10: Comparison of mean vendor-reported enrolment score (higher = better quality) and mean extracted minutia count by device for the Goat subset (red) versus the enrolled set as a whole (blue). Error bars at ±1 standard deviation for each device.

Table 7.8: One-tailed tests of significance for the difference of mean NFIQ-2 between Goats and subject-fingers as a whole.

Device	Enrolment			Verification		
	$M_A - M_G$	SE	p-value	$M_A - M_G$	SE	p-value
A	0	N/A	N/A	0	N/A	N/A
B	21	0.888	$< 10^{-15}$	22	0.631	$< 10^{-15}$
C	22	1.542	$< 10^{-15}$	22	0.811	$< 10^{-15}$
D	32	0.809	$< 10^{-15}$	31	0.586	$< 10^{-15}$
E	0	N/A	N/A	0	N/A	N/A
F	36	1.106	$< 10^{-15}$	36	0.717	$< 10^{-15}$
G	30	1.046	$< 10^{-15}$	30	0.687	$< 10^{-15}$
H	-5	0.369	1.000	-5	0.254	1.000
J	23	1.757	$< 10^{-15}$	24	0.796	$< 10^{-15}$
K	28	0.642	$< 10^{-15}$	28	0.477	$< 10^{-15}$
L	23	0.593	$< 10^{-15}$	22	0.428	$< 10^{-15}$

Table 7.9: One-tailed tests of significance for the difference of mean vendor-reported enrolment image quality and minutia count between Goats and subject-fingers as a whole.

Device	Vendor quality			Minutia count		
	$M_G - M_A$	SE	p-value	$M_G - M_A$	SE	p-value
A	17	1.518	$< 10^{-15}$	5	0.737	5.74×10^{-12}
B	5	0.711	1.05×10^{-12}	2	0.669	0.001
C	21	1.348	$< 10^{-15}$	0	0.712	0.500
D	24	1.232	$< 10^{-15}$	1	0.750	0.091
E	12	1.083	$< 10^{-15}$	2	0.678	0.002
F	0	0.000	N/A	-1	0.508	0.975
G	29	1.687	$< 10^{-15}$	0	0.859	0.500
H	0	0.000	N/A	0	0.493	0.500
J	14	1.182	$< 10^{-15}$	1	0.993	0.157
K	-2	0.920	0.985	-4	0.556	1.000
L	0	0.261	0.5	-3	0.530	1.000

7.2 Wolves, Lambs and Sheep

The enumeration of a Goat subset is relatively straightforward, since both enrolment and verification images come from the same individual. Doddington also attempted to classify those matches that contribute disproportionately to the false match rate (FMR), each in this case involving a pair of distinct individuals whom he labelled the 'Lamb' and the 'Wolf' - the former being the preyed upon (impersonated) and the latter the predator (impersonator). Implicit in this classification is that the match process be asymmetric [75].

We attempted to identify an aggregate Lamb-Wolf subset in the same way as we did for the Goats: that is, with the same consideration as before of finding a reasonable operating point across all the available devices, choose a target FMR (in our case, we chose 0.05 %), interpolate a set of decision thresholds for that FMR across the various devices, and then find all the imposter match pairs whose scores exceed those thresholds. Then aggregate the results across all devices by subject-finger, order them from largest to smallest number of false matches, and select the subset accounting for 50 % of the total.

For the ISBIT-Norwood dataset, this results in a Lamb-Wolf subset of 2512 enrol-verify subject-finger pairs out of a total of 260719 imposter matches. Note that the match software never attempts a "wrong finger" match so, with a high probability, these are all different subject - same finger matches[6]. The first thirty such match pairs (ordered from highest to lowest aggregate number of false matches) are shown in Table 7.10. Interestingly, the first two entries show exact symmetry i.e. they correspond to false matches in which the same subject-fingers, subject 291 and subject 167 finger position 7 (left index), play the roles of Wolf and Lamb interchangeably.

Nevertheless, we may partition the Lamb-Wolf set according to which role each

[6]Although fingers would, as a result of subject and/or operator inattention, infrequently be mislabelled during enrolment or verification, manual scrubbing of each day's data acquisition by an experienced supervisor is believed to have substantially removed these mislabellings

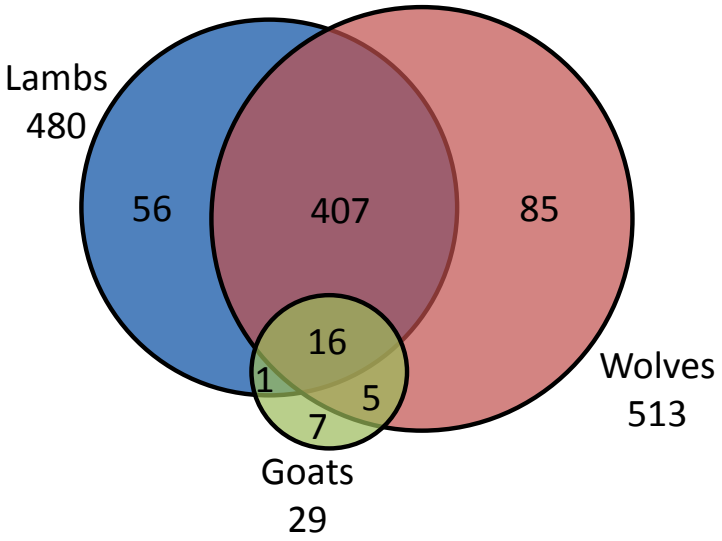


Figure 7.11: Venn diagram illustrating the overlap between the identified Goats, Lambs, and Wolves. By elimination, 302 of the 879 subject-fingers are not within the union of the sets and may thus be identified as Sheep.

subject-finger plays in the match, reducing the 2512 Lamb-Wolf match pairs to sets of 480 distinct Lamb subject-fingers and 513 distinct Wolf subject-fingers respectively. The overlap between Goats, Lambs, and Wolves is represented graphically in Figure 7.11. By elimination, 302 of the 879 subject-fingers are not within the union of the sets and may thus be identified in Doddington's classification as Sheep.

7.2.1 Demographics of the Lamb and Wolf subsets

Since the Lamb and Wolf subsets identified in the preceding section substantially sample the entire set of enrolled subject fingers (representing respectively 480 and 513 of the 879 distinct subject-fingers), we would not expect demographic differences to so pronounced as for the much smaller Goat subset. However we include the demographic analysis here for completeness (Figures 7.12 - 7.15). A Fisher exact test was performed as before for the sex distributions - while the Lambs show essentially no evidence of an effect ($p = 0.47$) there is a small amount of evidence that males are over-represented among the Wolves - that is, with a post-hoc assignment of males

Table 7.10: *only first 30 rows shown* Subject-fingers *pairs* with the highest number of false matches over all devices and visits, and which together constitute 50% of the false non-matches when the overall FMR is constrained to 5 %: we refer to these as the Lamb-Wolf set.

Verify				Enrol									
Id	BirthYear	EthnicOrigin	Gender	Id	BirthYear	EthnicOrigin	Gender	Finger	FM	TNM	Cum. FM	% of tot.	FMR
291	1980	North America	Male	167	1954	Europe	Male	7	380	1204	380	0	24.0%
167	1954	Europe	Male	291	1980	North America	Male	7	242	1342	622	1	15.3%
214	1987	North America	Male	64	1965	Europe	Male	2	191	2185	813	1	8.0%
86	1968	North America	Female	141	1968	North America	Female	2	176	2776	989	1	6.0%
111	1948	North America	Female	184	1973	North America	Female	2	169	2687	1158	1	5.9%
64	1965	Europe	Male	214	1987	North America	Male	2	158	2218	1316	1	6.6%
89	1969	North America	Female	126	1959	North America	Female	2	157	2795	1473	1	5.3%
89	1969	North America	Female	275	1986	Asia	Male	2	153	1431	1626	2	9.7%
184	1973	North America	Female	111	1948	North America	Female	2	139	2685	1765	2	4.9%
141	1968	North America	Female	86	1968	North America	Female	2	138	2814	1903	2	4.7%
213	1989	North America	Male	177	1982	North America	Male	7	135	945	2038	2	12.5%
275	1986	Asia	Male	89	1969	North America	Female	2	131	1453	2169	2	8.3%
184	1973	North America	Female	166	1957	North America	Female	2	127	2761	2296	2	4.4%
141	1968	North America	Female	167	1954	Europe	Male	7	126	2826	2422	2	4.3%
86	1968	North America	Female	330	1958	North America	Female	7	125	1459	2547	3	7.9%
9	1974	North America	Female	165	1973	North America	Female	2	123	837	2670	3	12.8%
223	1990	North America	Female	89	1969	North America	Female	2	122	2254	2792	3	5.1%
53	1966	Europe	Male	127	1956	Asia	Male	7	120	2832	2912	3	4.1%
186	1963	North America	Male	320	1960	North America	Female	7	116	1468	3028	3	7.3%
86	1968	North America	Female	1	1968	Europe	Male	7	115	2837	3143	3	3.9%
165	1973	North America	Female	9	1974	North America	Female	2	111	849	3254	3	11.6%
223	1990	North America	Female	201	1988	North America	Female	2	111	1797	3365	3	5.8%
167	1954	Europe	Male	141	1968	North America	Female	7	111	2841	3476	3	3.8%
320	1960	North America	Female	186	1963	North America	Male	7	110	1470	3586	4	7.0%
32	1959	North America	Female	147	1969	North America	Male	7	104	1264	3690	4	7.6%
46	1986	North America	Male	33	1984	North America	Male	2	103	473	3793	4	17.9%
89	1969	North America	Female	223	1990	North America	Female	2	101	2275	3894	4	4.3%
156	1956	North America	Female	149	1975	Europe	Female	7	100	2044	3994	4	4.7%
167	1954	Europe	Male	330	1958	North America	Female	7	97	1487	4091	4	6.1%
185	1973	North America	Male	89	1969	North America	Female	2	97	2855	4188	4	3.3%

Table 7.11: Fisher exact test for the significance of subject sex.

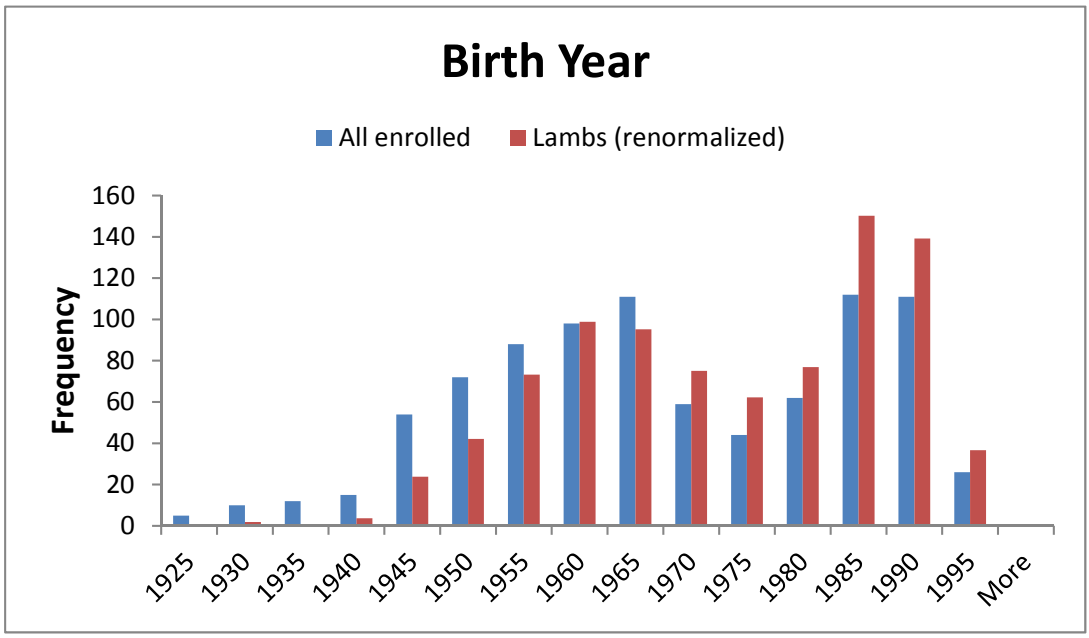
	All	Lambs	Expected under H_0	Wolves	Expected under H_0
Female	469	256	256.11	266	273.72
Male	410	224	223.89	247	239.28
Total	879	480	p-value: 0.47	513	p-value: 0.13

Table 7.12: χ-squared goodness-of-fit test for the significance of subject birth year at time of enrolment.

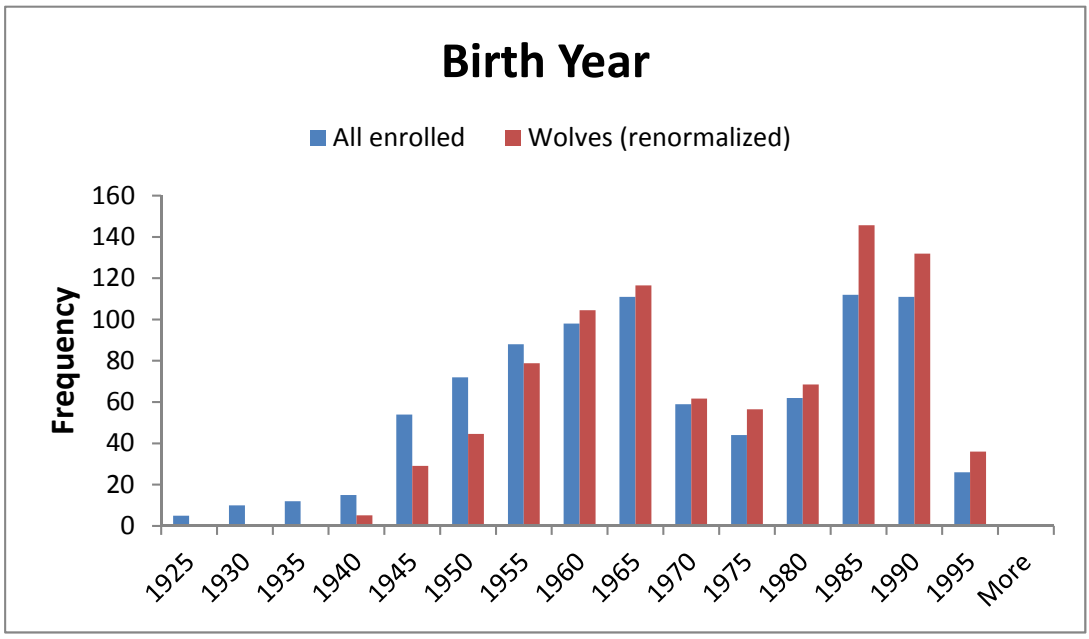
	All	Lambs	Expected under H_0	Wolves	Expected under H_0
1925	5	0	2.73	0	2.92
1930	10	1	5.46	0	5.84
1935	12	0	6.55	0	7.00
1940	15	2	8.19	3	8.75
1945	54	13	29.49	17	31.52
1950	72	23	39.32	26	42.02
1955	88	40	48.05	46	51.36
1960	98	54	53.52	61	57.19
1965	111	52	60.61	68	64.78
1970	59	41	32.22	36	34.43
1975	44	34	24.03	33	25.68
1980	62	42	33.86	40	36.18
1985	112	82	61.16	85	65.37
1990	111	76	60.61	77	64.78
1995	26	20	14.20	21	15.17
More	0	0	0	0	0
Total	879	480	p-value: 2.58×10^{-7}	513	p-value: 2.50×10^{-5}

as the "success" category, Fisher's test gives $p = 0.13$, consistent with Figure 7.13 in which the proportion of females appears to go down and that of males to go up relative to the baseline of all enrolled fingers (Table 7.11).

As for the Goats, chi-squared goodness-of-fit tests were conducted for birth year (a subject age proxy), ethnic origin, and manual or chemical exposure (Tables 7.12, 7.13, 7.14). Of these, only birth year shows a significant effect, with both Lambs ($p = 2.58 \times 10^{-7}$) and Wolves ($p = 2.50 \times 10^{-5}$) being skewed towards younger individuals. There is weak evidence ($p = 0.11$) for some dependence of Lambs on ethnic origin - probably reflecting an unexpectedly low number of Lambs of Philippino origin.

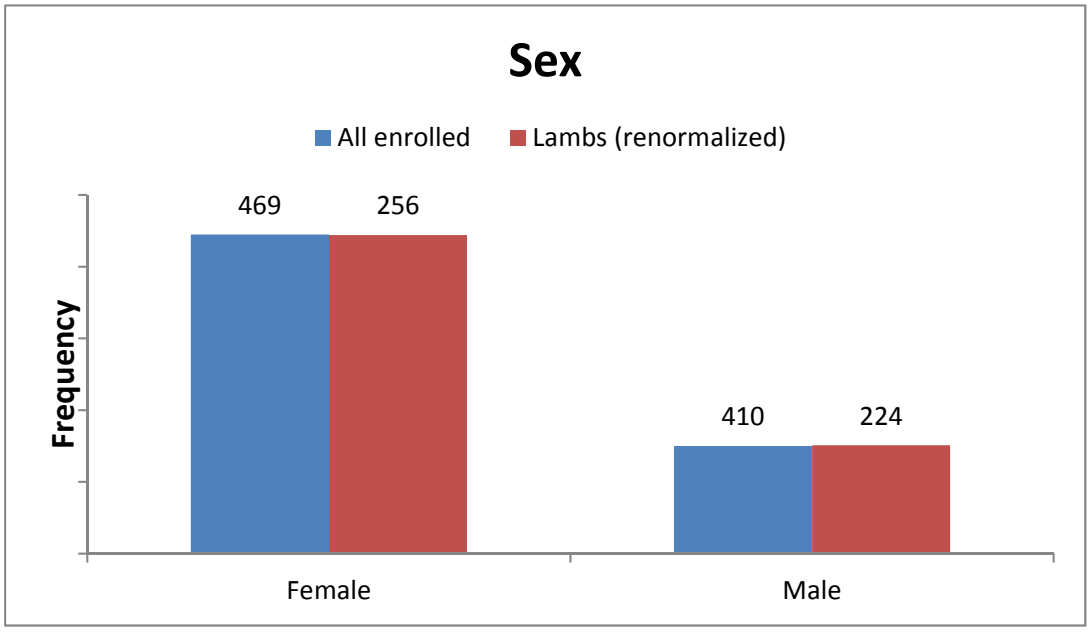


(a) Lambs

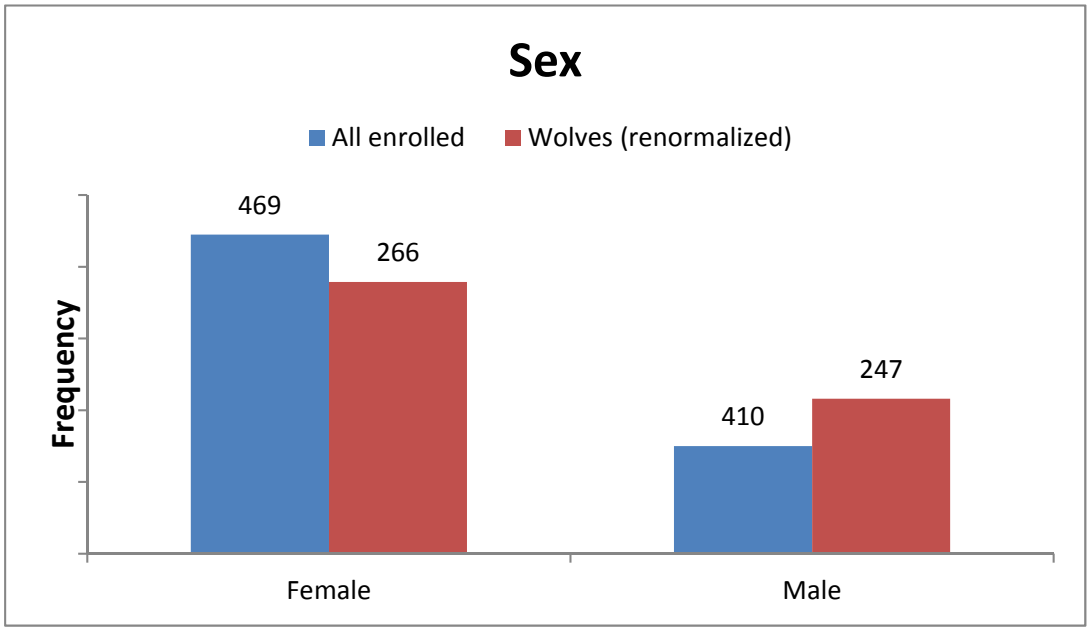


(b) Wolves

Figure 7.12: Distribution of subject birth year for the Lamb and Wolf subsets (red) compared to the enrolled set as a whole (blue). There is evidence for a bias towards younger individuals in both cases.

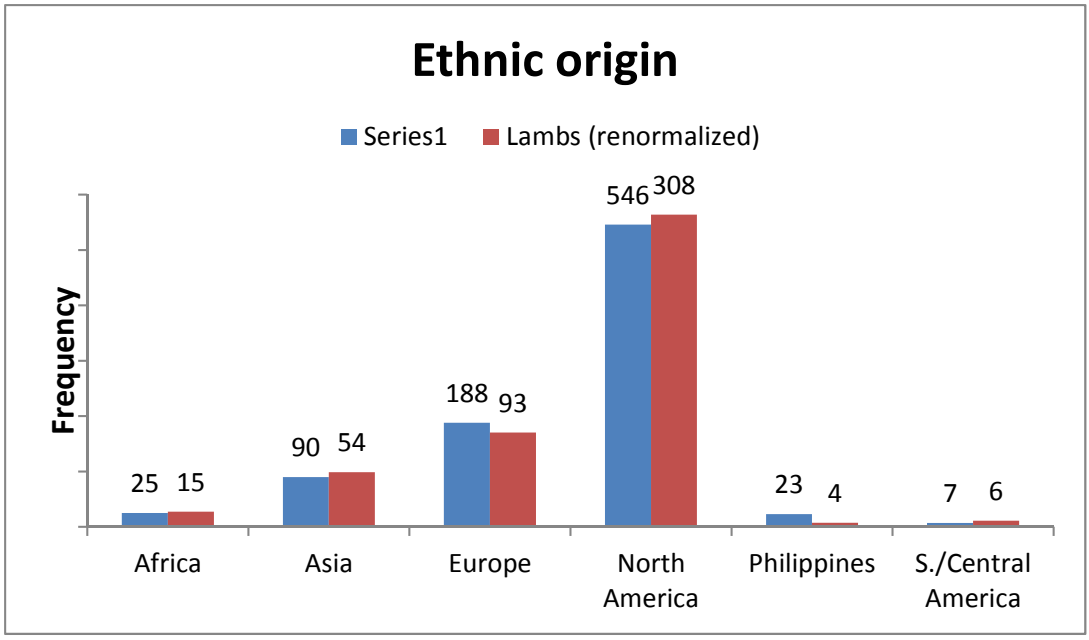


(a) Lambs

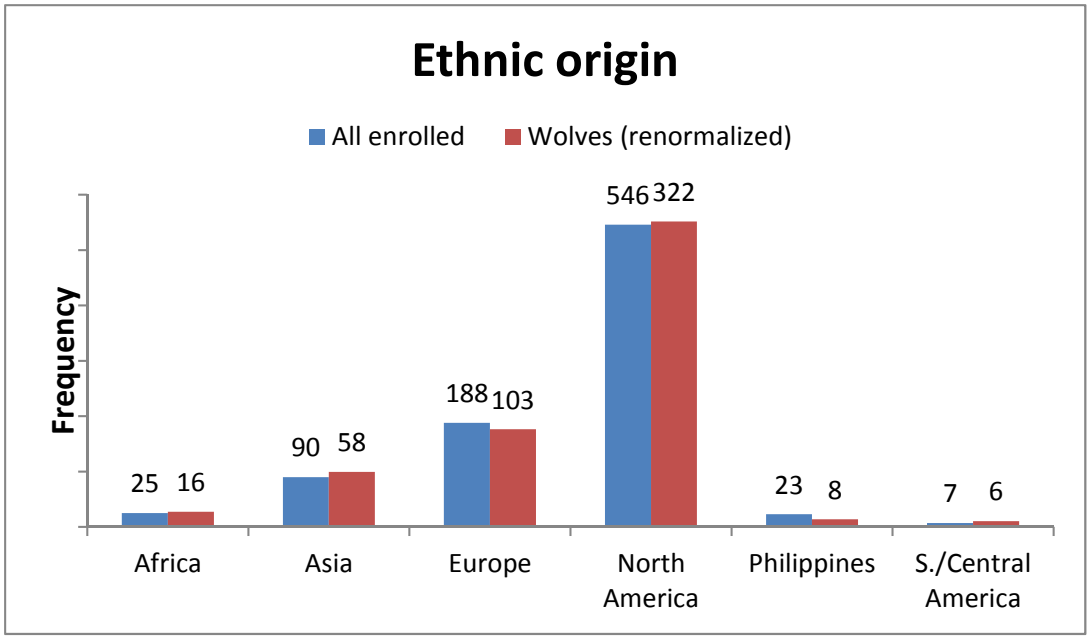


(b) Wolves

Figure 7.13: Distribution of subject sex for the Lamb and Wolf subsets (red) compared to the enrolled set as a whole (blue). There is weak evidence for an over-representation of males among the Wolves.

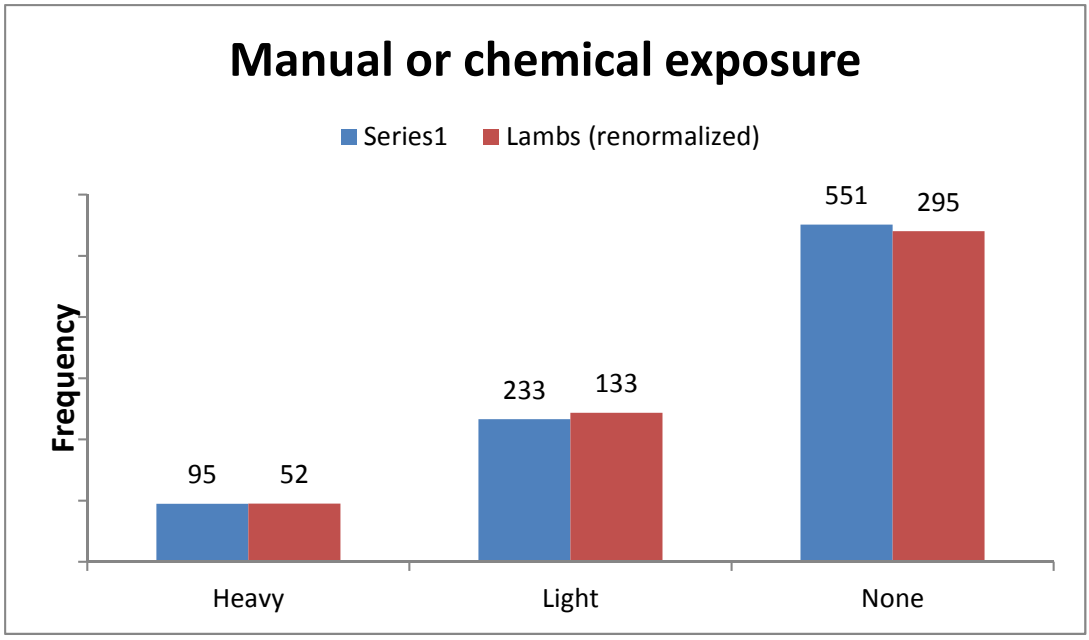


(a) Lambs

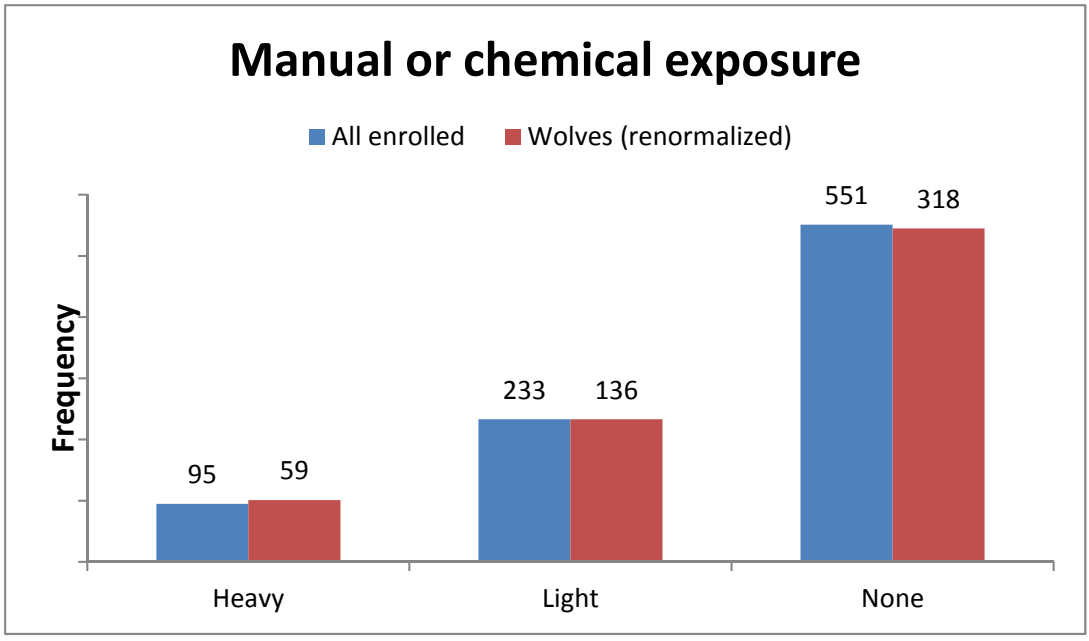


(b) Wolves

Figure 7.14: Distribution of subject ethnic origin for the Lamb and Wolf subsets (red) compared to the enrolled set as a whole (blue). There is little evidence of an effect.



(a) Lambs



(b) Wolves

Figure 7.15: Distribution of subject manual or chemical exposure for the Lamb and Wolf subsets (red) compared to the enrolled set as a whole (blue). There is little evidence of an effect.

Table 7.13: χ-squared goodness-of-fit test for the significance of subject self-reported ethnic origin.

	All	Lambs	Expected under H_0	Wolves	Expected under H_0
Africa	25	15	13.65	16	14.59
Asia	90	54	49.15	58	52.53
Europe	188	93	102.66	103	109.72
North America	546	308	298.16	322	318.66
Philippines	23	4	12.56	8	13.42
S./Central America	7	6	3.82	6	4.09
Total	879	480	p-value: 0.11	513	p-value: 0.52

Table 7.14: χ-squared goodness-of-fit test for the significance of subject manual or chemical exposure.

	All	Lambs	Expected under H_0	Wolves	Expected under H_0
Heavy	95	52	51.88	59	55.44
Light	233	133	127.24	136	135.98
None	551	295	300.89	318	321.57
Total	879	480	p-value: 0.83	513	p-value: 0.87

7.2.2 NFIQ quality of the Lamb and Wolf subsets

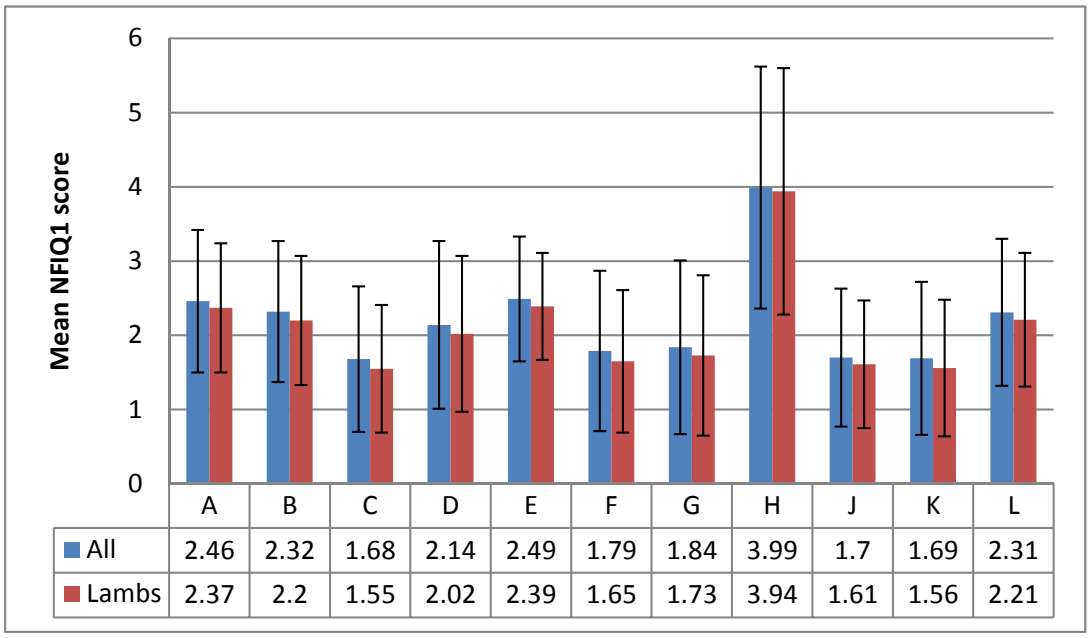
Figures 7.16 and 7.17 compare the mean NFIQ-1 and NFIQ-2 scores for the Lamb (enrol-side) and Wolf (verify-side) subsets versus all subject-fingers broken down by the available devices. The differences in mean biometric quality are smaller than those observed in the Goats, as one might expect given the much greater overlap between the Lamb/Wolf subsets and the set of all enrolled fingers. More significantly, there appears to be evidence that both Lambs and Wolves have *higher* biometric quality on average than the set of all enrolled fingers.

Table 7.15: One-tailed tests of significance for the difference of mean NFIQ-1 between Lambs/Wolves and subject-fingers as a whole.

Device	Lambs			Wolves		
	$M_A - M_L$	SE	p-value	$M_A - M_W$	SE	p-value
A	0.09	0.015	6.33×10^{-10}	0.08	0.009	$< 10^{-15}$
B	0.12	0.015	$< 10^{-15}$	0.15	0.010	$< 10^{-15}$
C	0.13	0.015	$< 10^{-15}$	0.07	0.009	3.75×10^{-14}
D	0.12	0.017	3.34×10^{-12}	0.15	0.011	$< 10^{-15}$
E	0.1	0.015	3.79×10^{-12}	0.09	0.009	$< 10^{-15}$
F	0.14	0.016	$< 10^{-15}$	0.12	0.011	$< 10^{-15}$
G	0.11	0.021	4.89×10^{-8}	0.09	0.013	2.23×10^{-12}
H	0.05	0.030	0.050	0.06	0.022	0.003
J	0.09	0.017	2.97×10^{-8}	0.06	0.010	2.06×10^{-9}
K	0.13	0.016	$< 10^{-15}$	0.12	0.010	$< 10^{-15}$
L	0.1	0.015	1.93×10^{-11}	0.09	0.010	$< 10^{-15}$

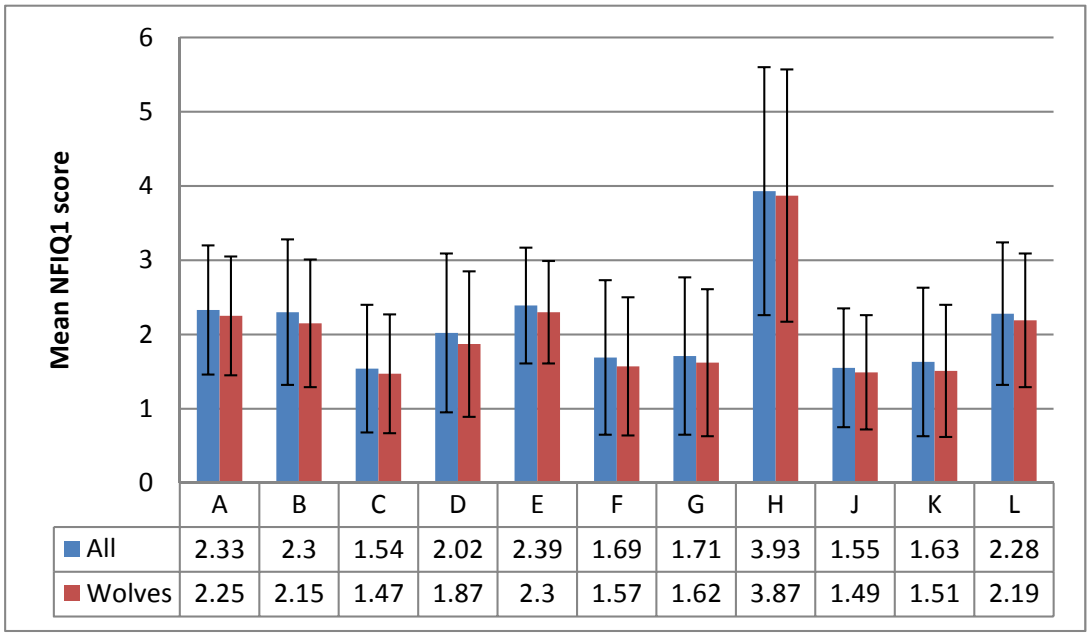
Table 7.16: One-tailed tests of significance for the difference of mean NFIQ-2 between Lambs/Wolves and subject-fingers as a whole.

Device	Lambs			Wolves		
	$M_L - M_A$	SE	p-value	$M_W - M_A$	SE	p-value
A	0	N/A	N/A	0	N/A	N/A
B	2	0.27	4.66×10^{-14}	3	0.19	$< 10^{-15}$
C	2	0.36	1.96×10^{-08}	1	0.22	1.94×10^{-06}
D	2	0.37	3.52×10^{-08}	2	0.24	$< 10^{-15}$
E	0	N/A	N/A	0	N/A	N/A
F	3	0.39	1.15×10^{-14}	3	0.27	$< 10^{-15}$
G	2	0.42	9.50×10^{-07}	1	0.26	4.76×10^{-05}
H	0	0.15	0.500	0	0.10	0.500
J	1	0.41	0.007	1	0.24	2.08×10^{-05}
K	2	0.31	7.21×10^{-11}	1	0.21	9.15×10^{-07}
L	2	0.28	9.14×10^{-13}	1	0.19	9.23×10^{-08}



	A	B	C	D	E	F	G	H	J	K	L
All	2.46	2.32	1.68	2.14	2.49	1.79	1.84	3.99	1.7	1.69	2.31
Lambs	2.37	2.2	1.55	2.02	2.39	1.65	1.73	3.94	1.61	1.56	2.21

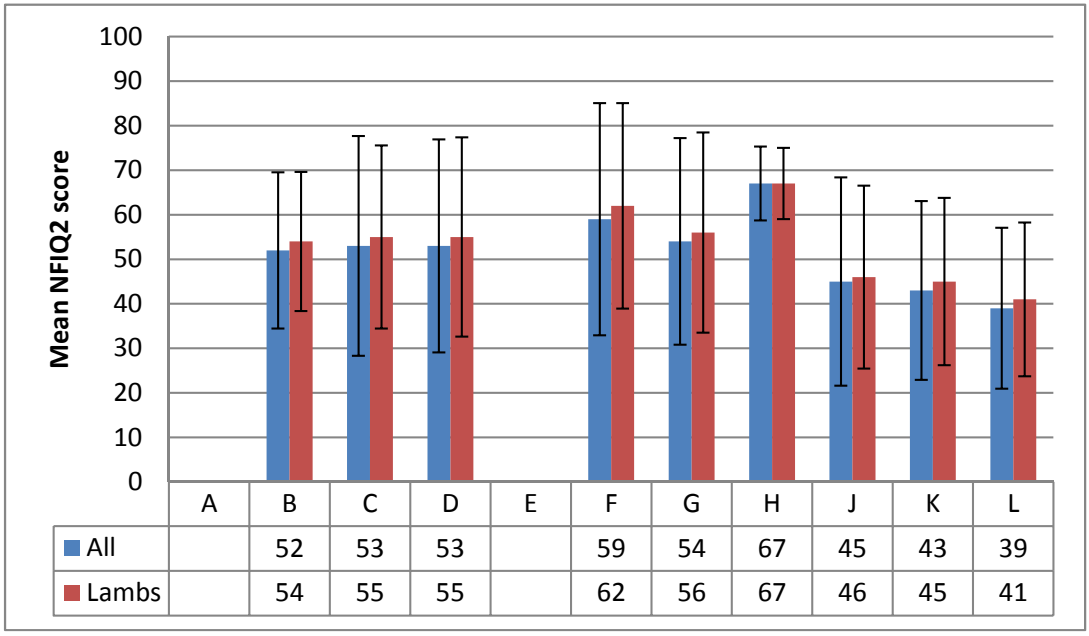
(a) Lambs



	A	B	C	D	E	F	G	H	J	K	L
All	2.33	2.3	1.54	2.02	2.39	1.69	1.71	3.93	1.55	1.63	2.28
Wolves	2.25	2.15	1.47	1.87	2.3	1.57	1.62	3.87	1.49	1.51	2.19

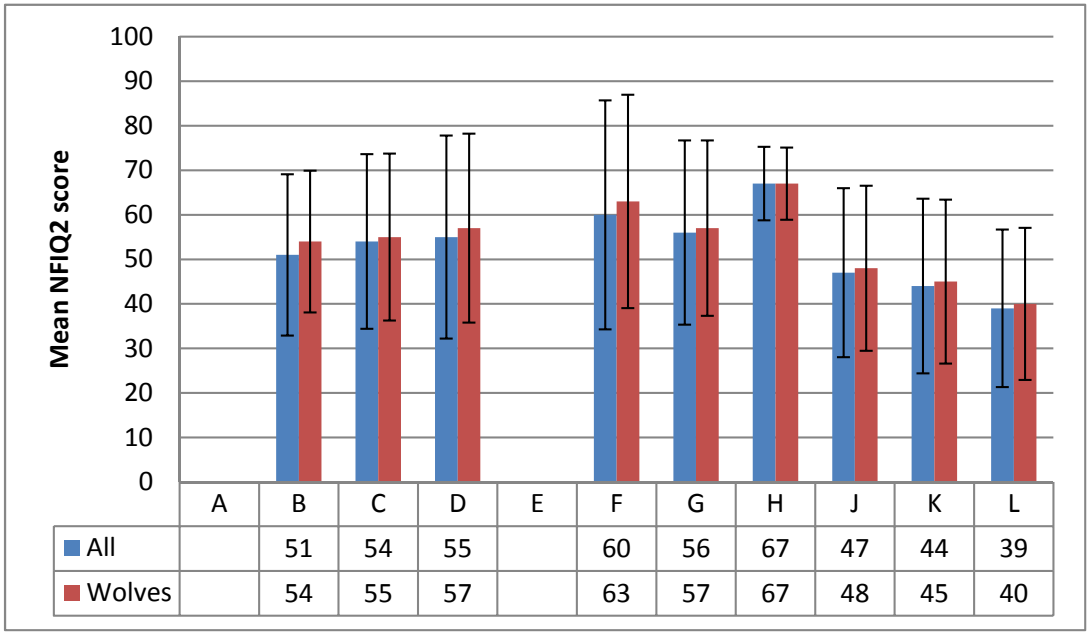
(b) Wolves

Figure 7.16: Comparison of mean NFIQ1 score (lower = better quality) by device for the Lamb and Wolf subsets (red) versus the enrolled set as a whole (blue). Error bars at ±1 standard deviation for each device.



	A	B	C	D	E	F	G	H	J	K	L
All		52	53	53		59	54	67	45	43	39
Lambs		54	55	55		62	56	67	46	45	41

(a) Lambs



	A	B	C	D	E	F	G	H	J	K	L
All		51	54	55		60	56	67	47	44	39
Wolves		54	55	57		63	57	67	48	45	40

(b) Wolves

Figure 7.17: Comparison of mean NFIQ2 score (higher = better quality) by device for the Lamb and Wolf subsets (red) versus the enrolled set as a whole (blue). Error bars at ±1 standard deviation for each device.

7.3 Effect of Goats on biometric permanence

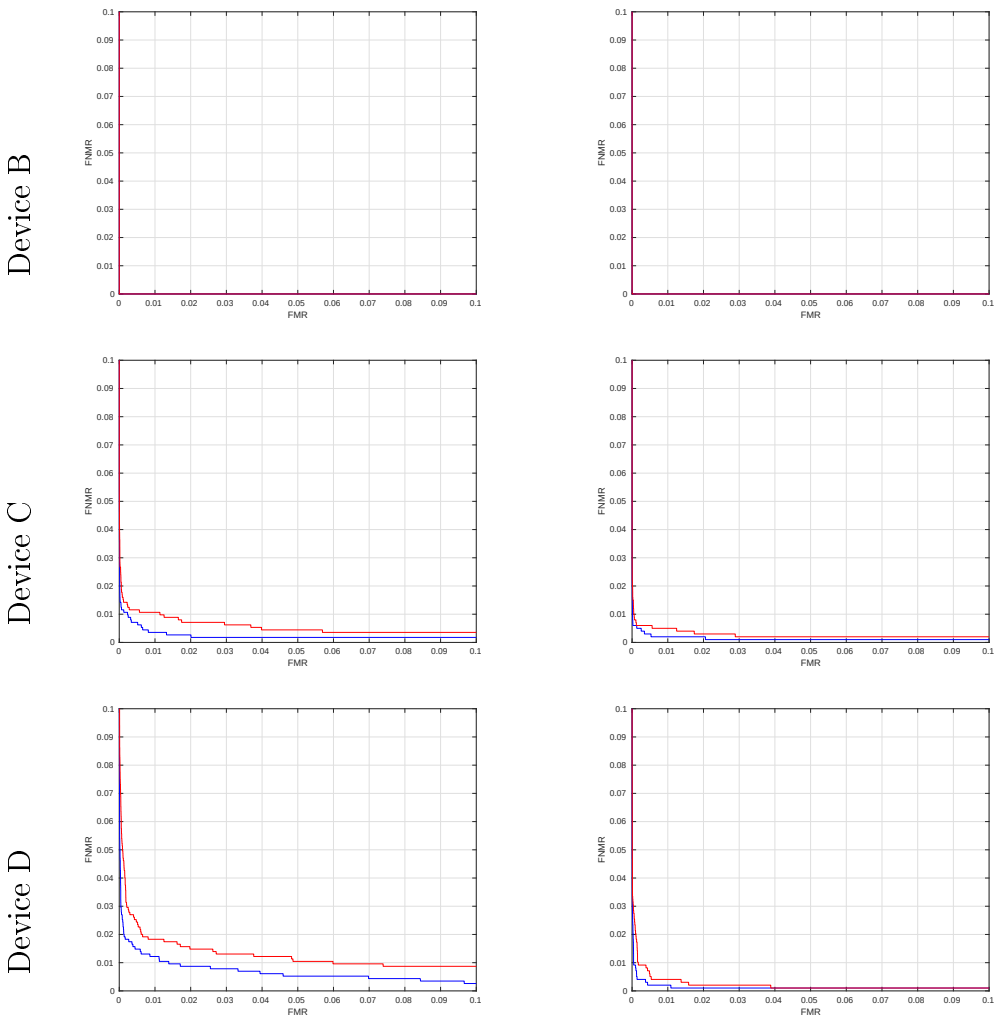
Having demonstrated the existence of a set of Goats (common subject-fingers whose low genuine match scores disproportionately affect the overall FNMR, regardless of device) it is natural to ask what effect they have on the overall classification performance of the IDMS. This is a subtly different question from that addressed in Chapter 6, where a given *fraction* of the lowest quality fingerprints was removed from the data sets of each device individually, regardless of which subject fingers they belonged to.

In particular, we would like to revisit the question of biometric template ageing (Chapters 4 and 5), and explore the impact of the Goat set on biometric permanence, P_B, which we defined in terms of the change of FNMR over time at a specified FMR. Once could imagine a number of distinct modes of behaviour:

1. The classification performance at both time $t = 0$ and $t = \Delta t$ is dominated by the Goats. Removing the Goats improves classification performance at all times but makes the system more sensitive to changes, over time, in the discriminability of the remaining subject fingers: as a result, although the classification performance increases, the biometric permanence P_B *decreases*

2. The classification performance at both time $t = 0$ and $t = \Delta t$ is dominated by the Goats, but changes over time more or less uniformly for all fingers; removal of Goats improves the performance but the permanence remains the same.

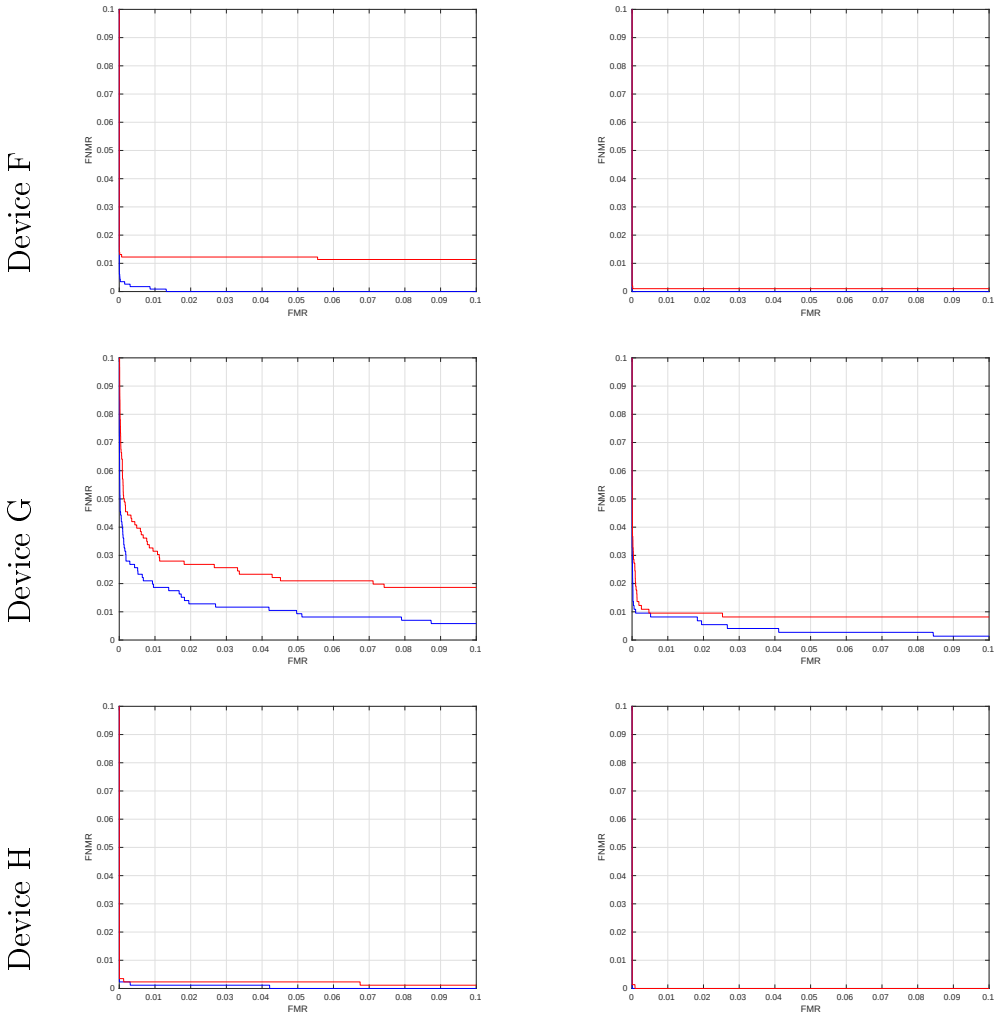
3. The Goats are subject-fingers whose genuine match scores at $t = 0$ dominate the region of the probability distribution just above the threshold for the reference FMR, and are most likely to slip below the threshold as the template ages. The remaining fingers start further above the threshold and take longer to fall below it, so that the biometric permanence *increases*.



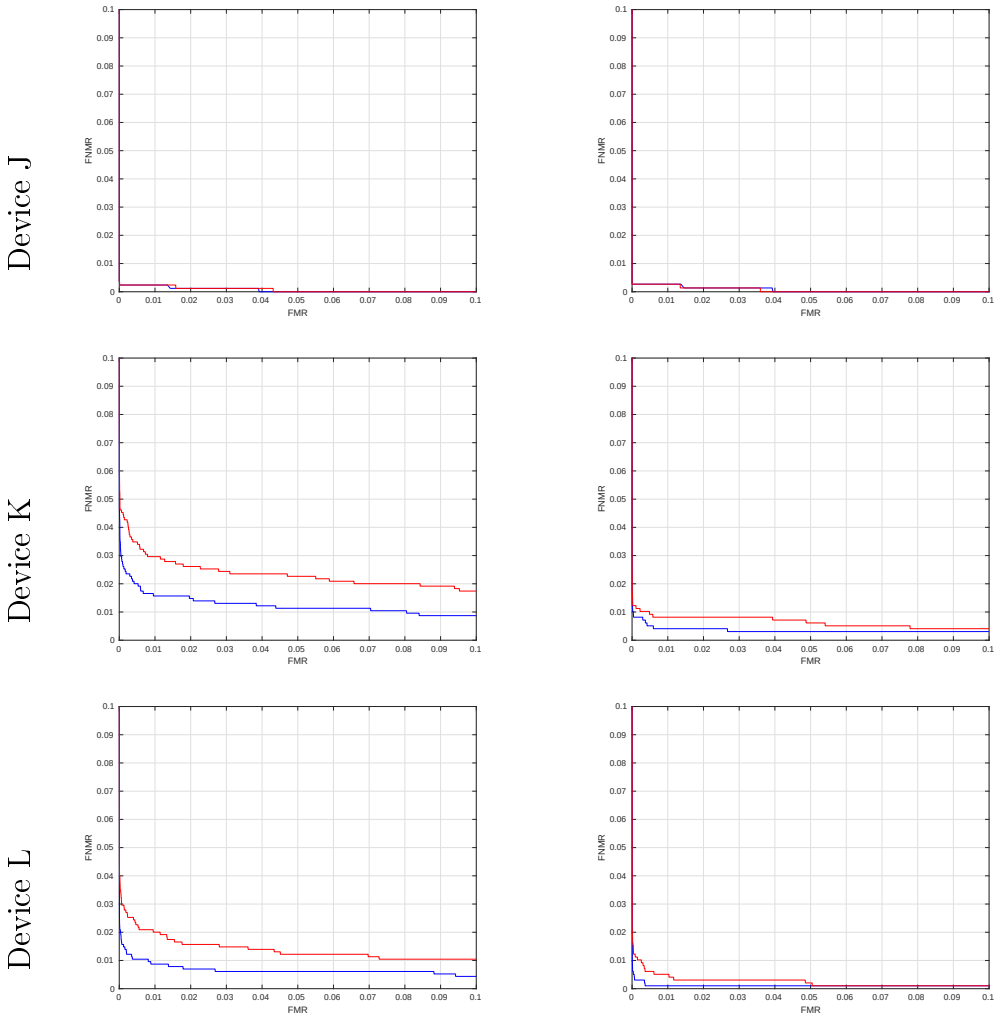
(a) All fingers **(b)** Excluding Goats

In order to investigate this, datasets were prepared with labelled genuine and imposter match scores, with the Goat subject fingers that were identified in the preceding section removed. Per-device DET curves were then evaluated in the same manner as in Chapter 5 for the filtered and unfiltered subject-finger sets, both for $t = 0$ and $\Delta t = 373$ weeks. We observe that, in all cases where the classification errors are not too few to resolve, both the $t = 0$ and $\Delta t = 373$ weeks performance is improved by the exclusion of Goats (Figure 7.18).

Finally, we re-applied the methodology described in Chapter 4 to evaluate bootstrap-resampled confidence 95 % intervals for the permanence measure, P_B for the IS-



(a) All fingers **(b)** Excluding Goats



(a) All fingers **(b)** Excluding Goats

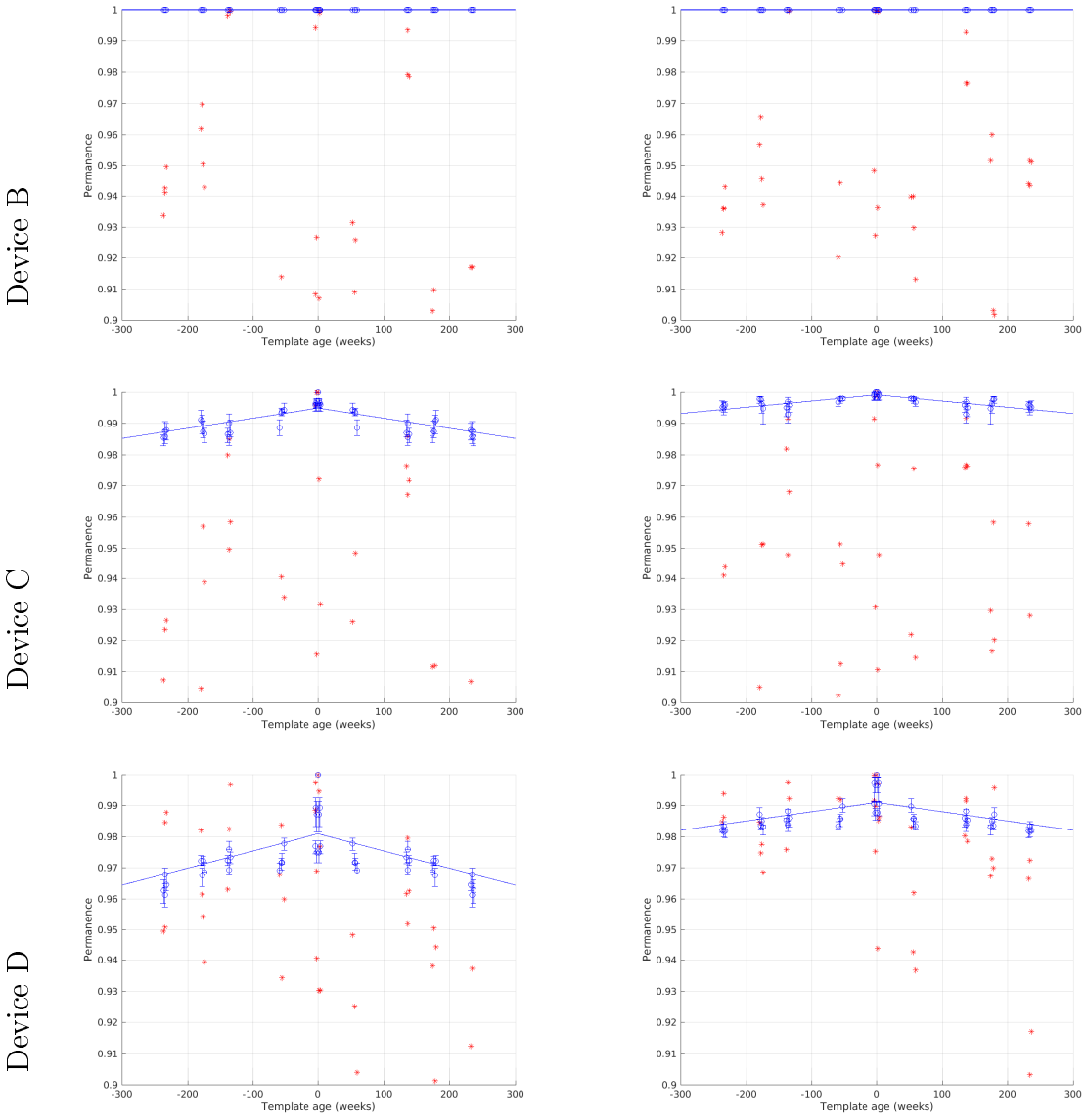
Figure 7.18: DET curves by device at $\Delta t = 0$ (blue) and at $\Delta t = 373$ weeks (red) for the complete subject-finger set, and for the set with Goats removed.

Table 7.17: Estimated 95% confidence intervals for permanence, P_B after 7 years, by device, with and without the Goat subset. N_s is the count of distinct subject-fingers in the sets, and is intended to give an indication of the fraction of fingers that would be excluded.

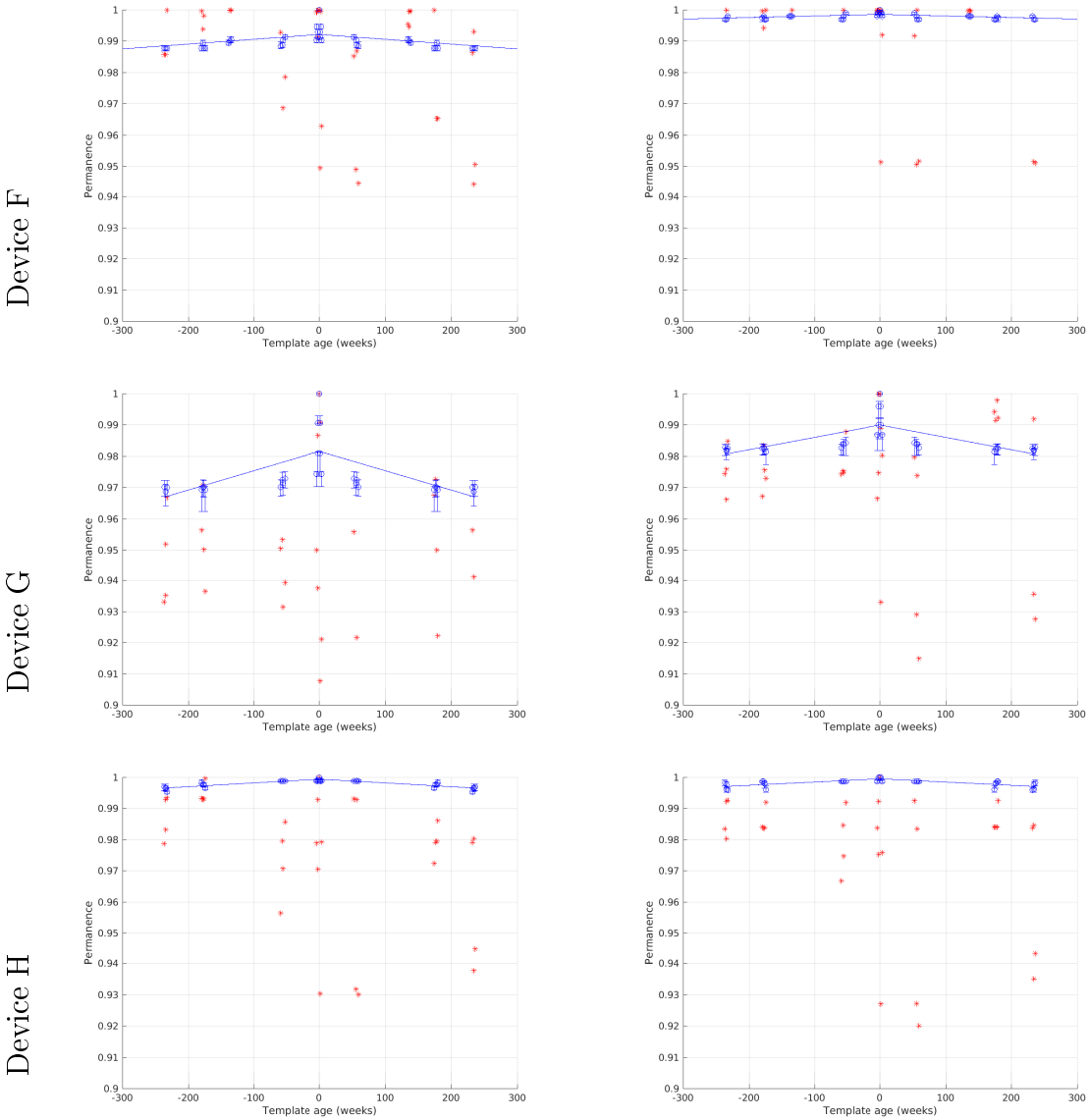
Device	With Goats		Without Goats	
	N_s	95 % CI for P_B	N_s	95 % CI for P_B
A	N/A		N/A	
B	355	1.000000, 1.000000	355	1.000000, 1.000000
C	504	0.980619, 0.985318	490	0.989762, 0.993654
D	288	0.958358, 0.963932	268	0.978285, 0.981307
E	N/A		N/A	
F	216	0.985664, 0.987388	201	0.996197, 0.997152
G	216	0.954134, 0.962193	201	0.972423, 0.976295
H	288	0.993765, 0.996223	272	0.994554, 0.996988
J	576	1.000000, 1.000000	552	1.000000, 1.000000
K	216	0.970424, 0.974079	201	0.992524, 0.994692
L	216	0.952864, 0.957523	201	0.971115, 0.975103

BIT/Norwood dataset with and without the Goat subject-fingers (Figure 7.19). The results are represented graphically in Figure 7.20 and enumerated in Table 7.17. For all cases in which measurable template ageing is observed (i.e. where the permanence $P_B < 1$), removal of the Goats increases the mean estimated biometric permanence: only in the case of the multi-spectral Device H do the confidence intervals overlap.

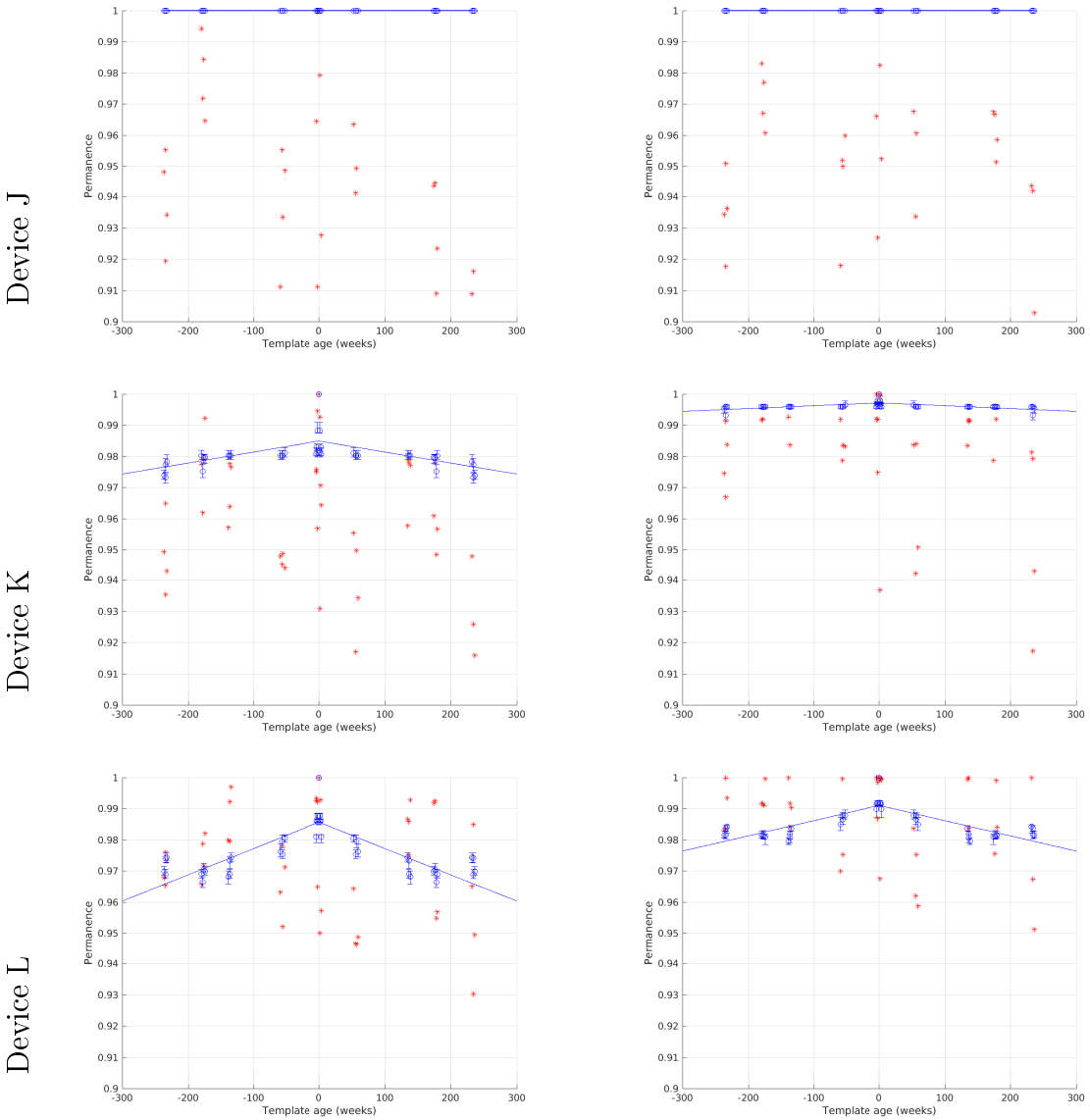
No attempt was made to repeat the analysis with either Lambs or Wolves removed, since these sets are much larger and excluding them would unreasonably diminish the set of available matches.



(a) All fingers **(b)** Excluding Goats



(a) All fingers **(b)** Excluding Goats



(a) All fingers **(b)** Excluding Goats

Figure 7.19: Biometric permanence P_B curves by device at $\Delta t = 0$ (blue) and at $\Delta t = 373$ weeks (red) for the complete subject-finger set, and for the set with Goats removed.

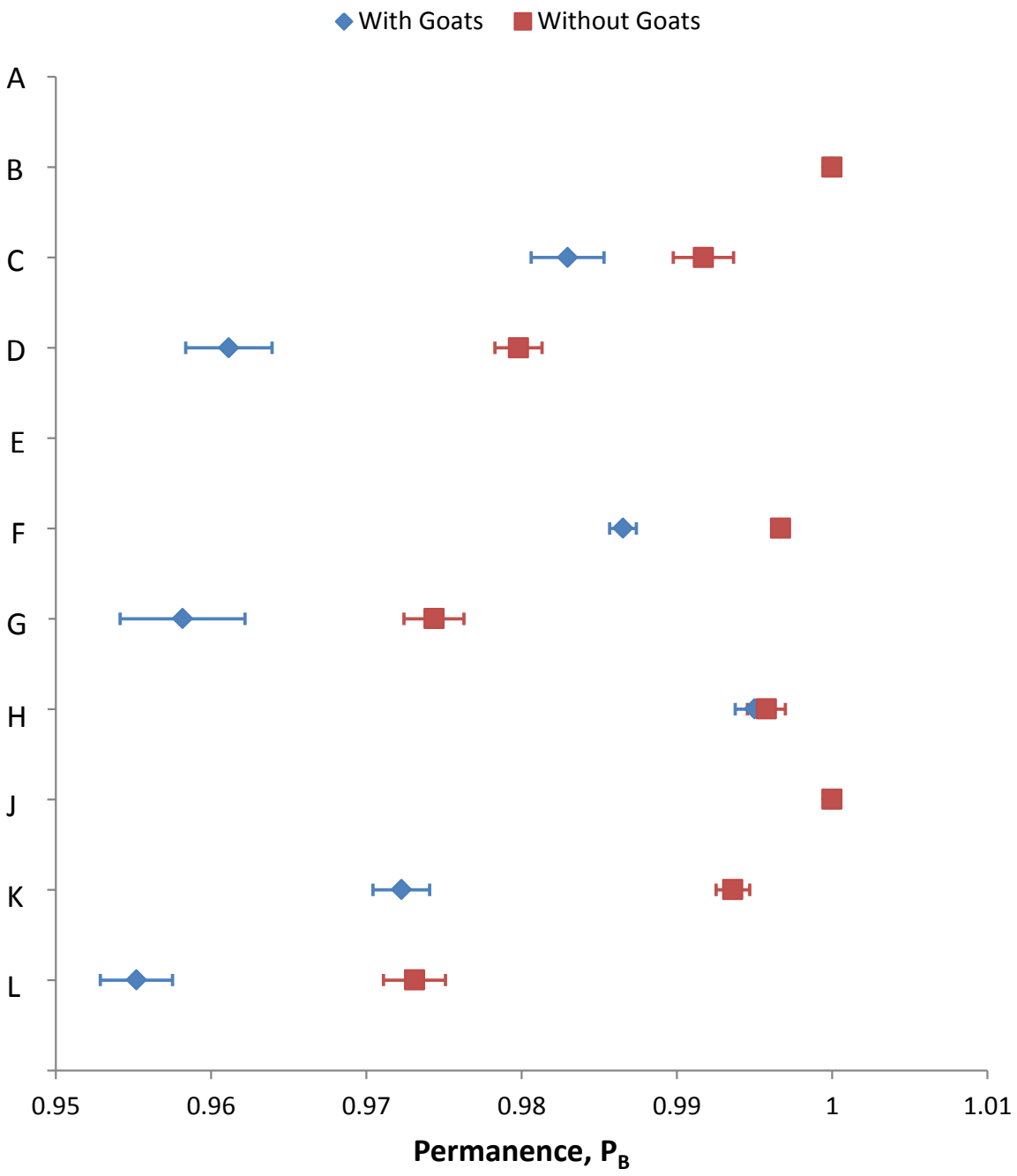


Figure 7.20: Graphical representation of the estimated 95% confidence intervals for permanence, P_B after 7 years, by device with and without the Goat subset.

7.4 Discussion

In this chapter, we applied a Doddington-like classification scheme to the subject-fingers of the ISBIT/Norwood fingerprint database, and were able to identify sets of Goats, Lambs, and Wolves. We observed that these sets are substantially common across the available fingerprint capture devices so that, suggesting that they reflect intrinsic properties of the underlying biometric rather than extrinsic capabilities of the devices themselves. We examined the available demographics of these sets, and found strong evidence that older individuals and females are over-represented in the Goat set (the subject-fingers that contribute disproportionately to the FNMR), while younger individuals are over-represented in both the Lamb and Wolf sets (the subject-fingers whose cross-matches contribute disproportionately to the FMR). Weak evidence was found for males being more common amongst Wolves.

We then applied the NIST NFIQ metrics to the sets, and found that while Goats are, on average, identified as having lower biometric quality than the cohort of subject-fingers as a whole, both Lambs and Wolves are identified as having higher quality. This may hint at a fundamental difficulty in assessing biometric quality in a one-to-many context – that is, to be truly informative about an individual's identity, a biometric should in some sense maximize the distance between individuals' biometric records.

Finally we revisited the concept of biometric permanence and showed that as well as having a disproportionate effect on the classification performance in general, Goats are a significant factor in limiting biometric permanence. This last observation may have implications for the management of long-term IDMSs such as biometric passports, for example suggesting increasingly frequent re-enrolments of older individuals in order to maintain overall IDMS performance.

Chapter 8

Discussion

8.1 Estimation of biometric permanence

In this work we have confirmed the existence of biometric template ageing in a multi-year, multi-vendor fingerprint dataset obtained under perhaps more controlled conditions (as detailed in Chapter 3) than those used in previous studies. The methodology, outlined in Chapter 4, provides first-order elimination of systematic confounding factors in the study, at the expense of imposing an assumption about the time-symmetry of the ageing effect. The amount of ageing observed was somewhat vendor-dependent, and consistent with the baseline performance of the devices, i.e. those with better than average classification performance at time $t = 0$ display a slower decrease in permanence (Chapter 5).

8.1.1 Motivation for the biometric permanence metric

Previous studies of biometric template ageing have reported their results in terms of change in genuine match score [13] and/or the change in a single measure of classification accuracy such as true match rate (TMR) [86]. From an operational point of view however, it is important to consider both security (via FMR or TNMR) and con-

venience (via FNMR or TMR). Uludag et al. for example used changes in the equal error rate (EER) [77]: this measure effectively gives equal weight to both factors. In contrast, our proposed definition of biometric permanence, P_B (Equation 4.1) emphasizes operational security by considering the change in FNMR (convenience) at fixed FMR (security). Furthermore, by normalizing the metric to the time-zero FNMR, we obtain a metric that permits comparisons across biometric systems of varying baseline performance. Some other desirable properties are outlined in Section 4.2.1.

8.1.2 Assumptions and limitations

Although the results we obtained in Chapter 5 seem intuitively reasonable, we were not able to provide an independent evaluation of biometric permanence by way of comparison with our results.

As already noted, a key assumption of the forward-backward "matched delta" methodology is that the change in match score is time-symmetric. The validity of this assumption will depend on how the match scores are evaluated: in Section 5.3.1 we posited a case in which it might be violated. The method also assumes that the confounding effect of gross visit-to-visit variability can be expressed (at least to leading order) in the form of additive bias terms.

We can imagine a "thought experiment" to explore the validity of these assumptions (and that of the underlying model, Equation 4.3). In such an experiment one would perform an ensemble of many studies, taking a "fix – block – randomize" approach to the ensemble design. For example:

- always employ the same test administrator(s) [FIX]
- conduct the data collections under the same environmental conditions (time of year, location, humidity) [FIX] or conduct multiple collections and subdivide the analyses by condition [BLOCK]

- conduct many studies over a variety of ambient conditions and test personnel and aggregate them in a meta-analysis [RANDOMIZE]

If the assumptions of the matched-delta model are valid, then the results naïve application of Equation 4.1 across the ensemble should converge to those of the matched delta method. In 4.3.2 we in fact simulated such an ensemble, randomizing the visit biases a_m, b_n. In principle, one could do the same with real data however the scale of such an experiment is beyond the scope of the present study.

In the simulated results of Chapter 4, it was also assumed that the imposter distributions were constant over time. Subsequent analysis of the experimental data cast doubt on this assumption, at least in the case of some devices in the study, and in Chapter 5 we removed this restriction (Section 5.3.2).

8.1.3 Illustrative application of the metric

To illustrate the potential application of the biometric permanence methodology and results, consider a biometric system based on our Device K. From Table 5.1, we estimated its permanence after 7 years as $P_B = 97.2 \pm 0.19$. Suppose the system is designed to operate at a FMR of 1 % and that the corresponding FNMR upon commissioning the system is found to be 1.5 %: then, using Equation 4.1, we can estimate the attainable FNMR after 7 years if the FMR is to be maintained as

$$\mathrm{FNMR}_{\Delta t} = 1 - P_B\,(1 - \mathrm{FNMR}_0) = 1 - (0.972)(0.985) = 0.043$$

i.e. the FNMR can be expected to degrade from 1.5 % to approximately 4 % over this time period.

Even taking into account the ± 0.19 uncertainty, this value is somewhat larger than that which might be inferred from the DET curve of Figure 5.10b - suggesting that the simple extrapolation overestimates the slope of P_B.

8.2 Validation of NFIQ quality metrics

Next, we applied the NIST NFIQ quality metrics on our labelled data, confirming the predictive capability of these metrics with respect to fingerprint classification accuracy – effectively using our study as an independent validation dataset for NFIQ. We observed essentially equivalent error-rate versus reject ratio from the two available implementations, NFIQ-1 and NFIQ-2, noting that NFIQ-2 is nevertheless preferred from an operational point of view because of its finer-grained scale (Chapter 6). Results for NFIQ were compared with those for vendor-provided quality values, where available.

The novelty of these results really comes from the data rather than from any methodological considerations. Data used in the training and validation of the NFIQ classifiers is dominated by US Government datasets (in the case of NFIQ-1, these were DOS-C, DHS2-C, DHS10, TXDPS, and BEN for training alone, with the addition of VISIT_POE and VISIT_POE_BVA for validation [74]), which might be assumed to share certain common characteristics of acquisition and processing. In contrast, the specifications and data acquisition protocols for our study were developed on behalf of the ILO, with an initial focus on vendor interoperability: it is interesting to explore how well the NFIQ measures perform on an independent dataset such as this.

8.3 Presence of a biometric menagerie

Subject-dependence of the classification performance was also explored, based on a modified version of the criteria originally proposed by Doddington. A relatively small subset of 'Goat'-like subject-fingers was identified, and confirmed to be substantially vendor-independent, suggesting that such fingers (which match poorly against themselves, and hence dominate the FNMR) are intrinsic, rather than a result of specific image capture technologies or feature extraction algorithms. Dependencies on the

available demographic factors were evaluated, with significant effects being found for subject sex and for chronological age at time of enrolment (or Birth Year). 'Lamb' and 'Wolf' subsets (which match too well against one another, together contributing disproportionately to the FMR) were also identified and examined: although representing a small fraction of the available imposter matches, these subsets substantially cover the set of subject-fingers as a whole, and their demographics are, unsurprisingly, not significantly different from it in most cases. The one exception was Birth Year, with both Lambs and Wolves appearing to favour younger individuals. It might be instructive to repeat this analysis with a more exclusive definition of the 'Lamb' and 'Wolf' subsets: they are large because of the choice of target FMR, which in turn was driven by a desire to operate each of the devices somewhere in the "knee" region of its DET curve.

Finally we examined the effect of subject-dependence on template ageing behaviour, and demonstrated (by removing them from the dataset) the importance of Goats in the magnitude of the observed ageing.

In the following sections, we consider the information-theoretic interpretation of template ageing, and its relation to biometric quality.

8.4 Biometrics as a communication channel

An information-theoretic model of a communications channel may be constructed as follows (see for example Cover and Thomas §7.5 [8]).

A message W is drawn from an index set $\{1, 2 \dots M\}$. We transmit a signal $X^n(W)$ which is a sequence of n symbols from the alphabet of some random variable X. A receiver receives signal Y^n related to transmitted signal X^n by a channel transition matrix $p(y^n|x^n)$. The task of the receiver is to infer W from Y^n using some decision rule $\hat{W} = g(Y^n)$.

We can imagine a similar formalism for a biometric system. That is, an individual W is drawn from an an index set $\{1, 2 \dots M\}$ of enrolled subjects. Associated with W is a biometric record $X^n(W)$ which is a collection (or vector) of features, such as a list of fingerprint minutiae. A verifier observes a biometric Y^n related to the record X^n by a biometric transition matrix $p(y^n|x^n)$, and the task of the verifier is to infer the identity W from the observed biometric Y^n using some decision rule $\hat{W} = g(Y^n)$ - or, in the case of a biometric verification system, to estimate the probability $P(W = w|Y^n)$ for some claimed identity w.

8.4.1 Biometric rate and capacity

In this formalism, we see that a biometric modality becomes a *channel* for the transmission of information about an individual's identity. We may then ask about the characteristics of this channel. For example, we might model the disappearance or obfuscation of individual fingerprint minutiae, perhaps as the result of occupational damage, as erasures in an erasure channel; or the uncertainties in minutia location as positional noise in a manner similar to that in an AWGN channel[1]. We might also be able to talk about the *rate* of a biometric system in terms of the cardinality of the subject set and the size of the biometric record:

$$R = \frac{\log_2 M}{n} \tag{8.1}$$

Given this definition of biometric rate, it might then be possible to define a biometric capacity C for a modality as the supremum of all achievable rates for the channel: in principle, this would allow us to compare the fundamental efficiency of different modalities, and to compare implementations within a modality based on how closely their rate approaches C.

[1] For a discussion of these and other standard channel models, see for example Cover and Thomas (op. cit.) §7.1.

Probably the biometric modality whose information-theoretic foundation has been developed most explicitly is that of Daugman's *IrisCodes* [14], whose templates take the form of binary sequences whose match scores are evaluated in terms of Hamming distance [15]. In particular, by developing a hidden Markov model (HMM) for IrisCode generation and tuning its parameters to obtain a best match to the observed FMR of real irises, Daugman et al. were able to obtain an estimate for the capacity of the IrisCode channel as 0.469 bits/bit [16]. A white noise analysis of the encoding procedure estimated a theoretical maximum capacity of 0.566 bits/bit: the difference being attributed to correlation between regions of the real iris.

8.4.2 Biometric "good codes"

Of course, much effort is directed in the field of communications towards the design of good (high rate) channel codes: that is, sets of codewords X^n for which R approaches C. In the case of linear block codes for example, one may imagine an (n, k) code as mapping a k-bit message space into a n-bit codeword space; good codes essentially maximize the minimum distance between codewords, making it easier to place a set of decision boundaries between them at the receiver. In the biometric analogue, codewords are biometric templates and clearly we do not have the same freedom to choose them: we do however have at least some ability to include more informative (and exclude less informative) features. Conceptually, this is similar to how the NFIQ quality metrics attempt to identify high quality fingerprints.

8.5 Information-theoretic interpretation of template ageing

When applied to the problem of biometric template ageing, the information-theoretic approach implies that what we are observing is a decrease in mutual information, over

time, between an individual's identity (as represented by their previously-enrolled biometric template) and their current biometric.

At $t = 0$, we can imagine that a biometric presentation Y is able to provide enough mutual information $I(X|Y)$ to almost completely resolve the uncertainty about the identity X of the individual (Figure 8.1(a)). We can imagine two scenarios by which biometric mutual information might decrease over a time. In the first, (a)–(b), the amount of information in the biometric remains essentially the same, but it becomes less informative about the identity of the individual, leaving a larger unresolved uncertainty $H(X|Y)$ after an interval Δt. In the fingerprint modality, this might occur for example if the number and type of the extracted minutiae remains the same, but their spatial relationship changes over time, perhaps due to morphological changes (stretching, shrinking, distorting) of the fingertip. In the second scenario, (a)–(c), the amount of information $H(Y)$ in the biometric is actually decreasing over time, i.e. $H(Y; \Delta t) < H(Y; 0)$. For fingerprints, that might correspond to a decrease in the number of extractable features (minutiae being obscured or erased by occupational damage or disease), or might represent a more noise-like process in which the mean spatial relationships of minutiae are preserved, but localizing them within a specific presentation becomes more difficult.

In practice, it is likely that template ageing results from some admixture of the two effects, which one may think of as ageing via "different information" versus "less information".

We would expect image- or feature-based biometric quality metrics like NFIQ to be insensitive to changes of the first type. On the other hand, changes of the first type should be amenable to the kinds of template update procedures proposed by [77, 60].

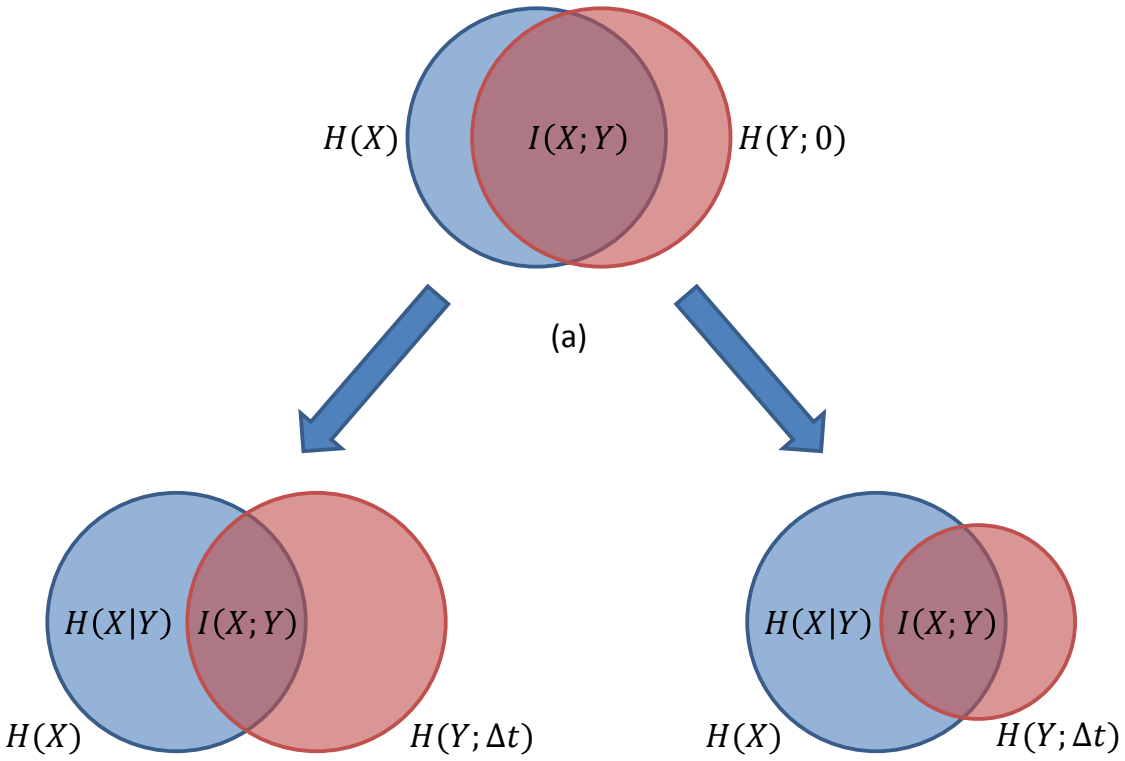


Figure 8.1: Conceptual scenarios for a decrease in biometric mutual information over time. At time zero (a), almost all of the uncertainty $H(X)$ about the individual's identity X is resolved by knowing the biometric Y. In (b), the amount of information $H(Y;\Delta t)$ provided by the biometric after time interval Δt is the same, but less of it is helpful in determining the individual's identity. In (c), the biometric after interval Δt is intrinsically less informative, $H(Y;\Delta t) < H(Y;0)$.

8.6 Suggestions for future work

We have deduced that, at least conceptually, there are two mechanisms for the changes in biometric match score over time. "Ageing via less information" would be asymmetric, and represents (or at least, will be indistinguishable from) chronological ageing of the individual. "Ageing via different information" on the other hand is symmetric.

In the case of genuine matches, it is easy to see how both effects might contribute to decreased match scores over time (i.e. to increased FNMR). The behaviour with respect to imposter matches is less obvious, and may depend on how exactly the particular scoring algorithm evaluates biometric similarity. For example, do two presentations that have low information content (few extractable minutiae), but are nevertheless similar, score higher or lower than two others that have many extractable minutiae but differ with respect to a few of them? Indeed, we saw evidence in Chapter 5 that changes in imposter score distributions for many devices were negligible - but not for all.

The measure of template ageing, P_B (biometric permanence) used in the present work was defined in terms of reduction in FNMR at fixed FMR, and reflects changes in both genuine and imposter match scores - the latter implicitly, by allowing the binary decision threshold to vary in order to maintain the chosen FMR. In future work, it might be instructive to maintain a constant threshold, and attempt to evaluate changes in FNMR and FMR separately. Alternatively, one might consider using variation of the decision threshold for constant FMR to be a proxy for changes in the imposter distributions.

The biometric menagerie classifications of Chapter 7 turned out to have a richer demographic structure than anticipated, and it would be worthwhile to revisit a number of aspects. In particular, the canonical categories of Sheep, Goats, Lambs and Wolves are rather broad, being based either on FNMR or FMR: a more refined categorization such as the one proposed by Yager & Dunstone [84] may yield additional

insight into the interplay of genuine and imposter match scores. At least, the demographic analysis of Lambs and Wolves might be enhanced by tightening the thresholds for inclusion into those subsets, reducing their size and so potentially making them more distinct from the cohort as a whole.

Finally, although the proposed permanence measure P_B and the methodology developed for evaluating it have shown promising results when applied to our moderately large fingerprint dataset, it will be important to evaluate them much more widely - preferably with even larger datasets and over additional biometric modalities.

Appendix A

Sample MSSQL database queries

Listing A.1: Simple query, for the author's own demographic and visit information

```
USE Norwood

SELECT * FROM Subject WHERE Id = 256

SELECT AlgorithmOrderEnrol, AlgorithmOrderVerify, EnrolBeganOn, VerifyEndedOn
FROM dbo.Subject_Visit_Map
WHERE SubjectId = 256
```

Listing A.2: A complicated query, for biometric qualities of match transactions with genuine/imposter class labelling

```
USE Norwood

SELECT eo.AlgorithmId, eo.VisitId, eo.SubjectId, nfiqA.FingerPosition, nfiqA.
    IsSecondary,
CASE WHEN nfiqA.FingerPosition = evq.FingerSecondary THEN evq.SecondaryQuality ELSE
    evq.PrimaryQuality END AS [VendorQ],
(SELECT MIN(v) FROM (VALUES (nfiqA.Score), (nfiqB.Score), (nfiqC.Score)) AS value(v))
     AS [NFIQ(min)],
(SELECT AVG(v) FROM (VALUES (nfiqA.Score), (nfiqB.Score), (nfiqC.Score)) AS value(v))
     AS [NFIQ(avg)],
(SELECT MAX(v) FROM (VALUES (nfiqA.Score), (nfiqB.Score), (nfiqC.Score)) AS value(v))
     AS [NFIQ(max)],
(SELECT MIN(v) FROM (VALUES (nfiq2A.Score), (nfiq2B.Score), (nfiq2C.Score)) AS value(
    v)) AS [NFIQ2(min)],
(SELECT AVG(v) FROM (VALUES (nfiq2A.Score), (nfiq2B.Score), (nfiq2C.Score)) AS value(
    v)) AS [NFIQ2(avg)],
(SELECT MAX(v) FROM (VALUES (nfiq2A.Score), (nfiq2B.Score), (nfiq2C.Score)) AS value(
    v)) AS [NFIQ2(max)],
-- see https://stackoverflow.com/a/6871572/4440445 "SQL MAX of multiple columns?"
vo.SubjectId, vo.FingerPosition, vo.Attempt, vi.Presentation, mp.Score,
CASE WHEN vo.SubjectId = eo.SubjectId AND vo.FingerPosition IN (et.FingerPrimary, et.
    FingerSecondary) THEN 1 ELSE 0 END AS [GEN]
FROM
dbo.EnrolTemplate et INNER JOIN dbo.EnrolOnline eo ON eo.Id = et.EnrolOnlineId
INNER JOIN dbo.EnrolVendorQuality evq ON evq.EnrolOnlineId = eo.Id
INNER JOIN dbo.EnrolNFIQ nfiqA ON nfiqA.EnrolOnlineId = eo.Id
INNER JOIN dbo.EnrolNFIQ nfiqB ON nfiqB.EnrolOnlineId = eo.Id AND nfiqB.
    FingerPosition = nfiqA.FingerPosition
INNER JOIN dbo.EnrolNFIQ nfiqC ON nfiqC.EnrolOnlineId = eo.Id AND nfiqC.
    FingerPosition = nfiqA.FingerPosition
INNER JOIN dbo.EnrolNFIQ2 nfiq2A ON nfiq2A.EnrolOnlineId = eo.Id AND nfiq2A.
    FingerPosition = nfiqA.FingerPosition
INNER JOIN dbo.EnrolNFIQ2 nfiq2B ON nfiq2B.EnrolOnlineId = eo.Id AND nfiq2B.
    FingerPosition = nfiq2A.FingerPosition
```

```
22  INNER JOIN dbo.EnrolNFIQ2 nfiq2C ON nfiq2C.EnrolOnlineId = eo.Id AND nfiq2C.
        FingerPosition = nfiq2A.FingerPosition
23
24  INNER JOIN dbo.MatchPresentation mp ON mp.EnrolOnlineId = eo.Id
25  INNER JOIN dbo.VerifyImage vi on vi.Id = mp.VerifyImageId
26  INNER JOIN dbo.VerifyOnline vo ON vo.Id = vi.VerifyOnlineId AND vo.FingerPosition =
        nfiqA.FingerPosition
27  WHERE
28  nfiqA.Presentation = 1 AND nfiqB.Presentation = 2 AND nfiqC.Presentation = 3
29  AND
30  nfiq2A.Presentation = 1 AND nfiq2B.Presentation = 2 AND nfiq2C.Presentation = 3
31  AND
32  eo.AlgorithmId = 2
33  -- for testing; should be 12 gen and 12 imp:
34  -- AND eo.SubjectId = 256 AND eo.VisitId = 6 AND vo.SubjectId IN (256,1) AND vo.
        VisitId = 6
35  -- ORDER BY vo.FingerPosition, vo.Attempt, vi.Presentation
```

Appendix B

Unpacking the BioAPI Biometric Information Record

Listing B.1: Python library function to unpack ILO BIR

```
from struct import unpack


def read_minutiae(f,mcount):
  minutiae = [];
  minutia = {}
  #print mcount
  for m in range(mcount):
    #print m
    # 2 bits type + 14 bits xpos
    bb, = unpack('>h', f.read(2))
    minutia['type'] = (bb & 49152) >> 14
    minutia['xpos'] = bb & 16383
    # 2 bits reserved + 14 bits ypos
    bb, = unpack('>h', f.read(2))
    minutia['reserved'] = (bb & 49152) >> 14
    minutia['ypos'] = bb & 16383
    # 1 byte angle
    minutia['angle'], = unpack('B', f.read(1))

    # append to minutiae list
    minutiae.append((minutia['xpos'], minutia['ypos'], minutia['angle'], minutia['
        type']))

  return minutiae

def unpack_BIR(f):
  # read and unpack the ``0paque biometric data`` header
  BIR = {}
  BIR['format'], = unpack('4s', f.read(4))
  BIR['version'], = unpack('4s', f.read(4))
  BIR['length'], = unpack('>h', f.read(2))

  BIR['reserved'], = unpack('>h', f.read(2))

  BIR['hsize'], = unpack('>h', f.read(2))
  BIR['vsize'], = unpack('>h', f.read(2))

  BIR['hres'], = unpack('>h', f.read(2))
  BIR['vres'], = unpack('>h', f.read(2))

  BIR['fingers'], = unpack('B', f.read(1))
  BIR['views'], = unpack('B', f.read(1))
```

```
43  # print BIR
44
45    return BIR
46
47  def unpack_finger_records(f):
48
49    fingerPri = {}
50    # read and unpack the primary fingerprint record
51    fingerPri['finger'], = unpack('B', f.read(1))
52    fingerPri['view_imp'], = unpack('B', f.read(1))
53    fingerPri['quality'], = unpack('B', f.read(1))
54    fingerPri['mcount'], = unpack('B', f.read(1))
55    # read (and discard) primary minutiae
56  # print fingerPri
57    #read_minutiae(f, fingerPri['mcount'])
58    for i in range(0,fingerPri['mcount']):
59      minutia = unpack('5B', f.read(5))
60  #   print i, minutia
61
62    fingerSec = {}
63    # read and unpack the secondary fingerprint record
64    fingerSec['finger'], = unpack('B', f.read(1))
65    fingerSec['view_imp'], = unpack('B', f.read(1))
66    fingerSec['quality'], = unpack('B', f.read(1))
67    fingerSec['mcount'], = unpack('B', f.read(1))
68  # print fingerSec
69    # read (and discard) primary minutiae
70    #read_minutiae(f, fingerSec['mcount'])
71    for i in range(0,fingerSec['mcount']):
72      minutia = unpack('5B', f.read(5))
73  #   print i, minutia
74
75    return fingerPri, fingerSec
76
77  def unpack_BIR_HEADER(f):
78    # read and discard 8 byte length added by bcp bulk copy
79    bcp = f.read(8)
80
81    # read and unpack ILO 16 byte BioAPI BIR header
82    BIR_HEADER = {}
83    BIR_HEADER['length'], = unpack('<i', f.read(4))
84    BIR_HEADER['version'], = unpack('B', f.read(1))
85    BIR_HEADER['type'], = unpack('B', f.read(1))
86    BIR_HEADER['format'], = unpack('4s', f.read(4))
87    BIR_HEADER['quality'], = unpack('B', f.read(1))
88    BIR_HEADER['purpose'], = unpack('B', f.read(1))
89    BIR_HEADER['factors'], = unpack('<i', f.read(4))
90    # print BIR_HEADER
91
92    return BIR_HEADER
```

Listing B.2: Python program to extract minutia counts and vendor-assessed biometric quality

```
1  import sys
2  import os
3  import bioapi
4  import csv
5
6
7
8  datadir = sys.argv[1]
9
10 for filename in os.listdir(datadir):
11
12   if not filename.endswith(".dat"):
13     continue
```

```
14
15    enrolOnlineId,algorithmId,fingerPrimary,fingerSecondary,remainder = filename.split(
          "_")
16
17    # read and unpack the ILO SID template blob extracted by bcp from MS SQL Server
18    with open(os.path.join(datadir, filename)) as f:
19      try:
20        hdr = bioapi.unpack_BIR_HEADER(f)
21      except:
22        print "Error␣reading␣BIR_HEADER"
23        pass
24      try:
25        bir = bioapi.unpack_BIR(f)
26        f1,f2 = bioapi.unpack_finger_records(f)
27        with open('vendor_mcount_quality.csv', 'ab') as csvfile:
28          writer = csv.writer(csvfile, delimiter=',',quotechar='"', quoting=csv.
                QUOTE_MINIMAL)
29          writer.writerow([enrolOnlineId,algorithmId,hdr['quality'],f1['finger'],f1['
                mcount'],f1['quality'],f2['finger'],f2['mcount'],f2['quality']])
30      except:
31        print "Unexpected␣error:", sys.exc_info()[0]
32        pass
```

Appendix C

Bootstrap resampled confidence interval (CI) for P_B

Listing C.1: Bootstrap resampled confidence interval (CI) for the permanence

```
1
2  function [pNaive, pMatched, age] = ...
3      matchedDelta3(genNxN, impNxN, ageNxN, fmrRef)
4  %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
5  %
6  % PURPOSE:  re-implementation of John Campbell's 'matched delta' method
7  %           for estimating biometric permanence from a set of match
8  %           scores
9  %
10 % AUTHOR:   John Harvey
11 %           Carleton University Dept. of Systems and Computer Engineering
12 %
13 % DATE:     June 2017
14 %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
15
16 global bootopts;
17 bootopts = statset('UseParallel',true);
18
19 nboot = 1000;                    % bootstrap samples
20 nbootg = nboot;                  % bootstrap samples for deltaGen
21 nbooti = 30;                     % bootstrap samples for deltaImp
22
23 nE = size(ageNxN,1);             % number of enrol visits
24 nV = size(ageNxN,2);             % number of verify visits
25
26 % average the subject-finger scores over presentations
27 genPresMean = squeeze(nanmean(genNxN,4));
28 %genPresStd = squeeze(nanstd(genNxN,0,4));
29 impPresMean = squeeze(nanmean(impNxN,4));
30 %impPresStd = squeeze(nanstd(impNxN,0,4));
31
32 nGen = size(genPresMean,3);      % actual number of gen samples
33 nImp = size(impPresMean,3);      % actual number of imposter samples
34
35 % initialize arrays for the matched delta reference distributions and
36 % forward-backward scores
37 gen0 = zeros(nE * nGen, 1);
38 imp0 = zeros(nE * nImp, 1);
39 deltaGen = zeros(nE,nV);
40 deltaImp = zeros(nE,nV);
41
42
43 % evaluate the "naive" permanence and form the baseline (age = 0)
```

```
% aggregate distributions and matched deltas
pNaive = ones(nE, nV);
for i = 1:nE

    tmr0 = tmrAtfmr(squeeze(genPresMean(i,i,:)), squeeze(impPresMean(i,i,:)), fmrRef,
         nboot);

    % aggregate the baseline "zero time" distributions
    gen0((i-1)*nGen+1:i*nGen) = genPresMean(i,i,:);
    imp0((i-1)*nImp+1:i*nImp) = impPresMean(i,i,:);

    for j = 1:nV
        if (j ~= i)
            tmr = tmrAtfmr(squeeze(genPresMean(i,j,:)), squeeze(impPresMean(i,j,:)),
                fmrRef, nboot);
            pNaive(i,j,:) = tmr/tmr0;
        end

        if (j > i) % (skip lower triangle because of symmetry)
            % form the forward-backward matched deltas
            deltaGen(i,j) = delta( genPresMean(i,j,:), genPresMean(j,i,:),
                genPresMean(i,i,:), genPresMean(j,j,:), nbootg);
            deltaImp(i,j) = delta( impPresMean(i,j,:), impPresMean(j,i,:),
                impPresMean(i,i,:), impPresMean(j,j,:), nbooti);
        end
    end
end

% evaluate the matched deltas - note that we can use the symmetry to
% reduce the number of evaluations
inds = triu(ageNxN) ~= 0.;
PbCI = permMatchedDelta(gen0, imp0, deltaGen(inds)', deltaImp(inds)', fmrRef, nboot);
% short-circuit the acual evaluation (for quick testing of loops):
%PbCI = [ones(size(inds(inds))) ones(size(inds(inds))) ones(size(inds(inds)))]';

% flip and stitch to get a plottable permanence versus age
tmp = sortrows([ageNxN(inds)' ; PbCI]')';
age = [-fliplr(tmp(1,:)) 0. tmp(1,:)];
pMatched = [fliplr(tmp(2:end,:)) [1.;1.;1.] tmp(2:end,:)];

end


function d = delta(sij, sji, sii, sjj, nboot)

% global bootopts;
%
% X = squeeze(sij) + squeeze(sji) - squeeze(sii) - squeeze(sjj);
%
% bootstat = bootstrp(nboot,@nanmean,X,'Options',bootopts);
%
% d = 0.5 * mean(bootstat);

% JH 13-Jul-17: it doesn't make sense to bootstrap the mean, since we know
% that the sample mean is a _sufficient statistic_ for the population mean
d = 0.5 * nanmean(sij + sji - sii - sjj);

end


function tmr = tmrAtfmr(gen, imp, fmrRef, nboot)

global bootopts;

nGen = size(gen,1);
nImp = size(imp,1);

% construct a single score vector and class labels
```

```
scores = [imp ; gen];
labels = [false(nImp,1) ; true(nGen,1)];

% construct a vector of observation weights - we can use this to remove the
% inherent class imbalance i.e. sample the genuines much more often than
% the imposters
weights = [0.5/nImp * ones(nImp,1) ; 0.5/nGen * ones(nGen,1)];

bootfun = @(X,L)(truematch(X, L,fmrRef));
bootstat = bootstrp(nboot,bootfun,scores,labels,'Weights', weights, 'Options',
    bootopts);

% Estimate the TMR at the estimated threshold
tmr = mean(bootstat);

end



function PbCI = permMatchedDelta(gen, imp, deltaGen, deltaImp, fmrRef, nboot)

global bootopts;

nGen = size(gen,1);
nImp = size(imp,1);

% construct a single score vector and class labels
scores = [imp ; gen];
labels = [false(nImp,1) ; true(nGen,1)];

% construct a vector of observation weights - we can use this to remove the
% inherent class imbalance i.e. sample the genuines much more often than
% the imposters
weights = [0.5/nImp * ones(nImp,1) ; 0.5/nGen * ones(nGen,1)];

bootfun = @(X,L)(permbio(X, L, deltaImp - deltaGen, fmrRef));
[ci, bootsam] = bootci(nboot,{bootfun,scores,labels},'Weights',weights,'Options',
    bootopts);

PbCI = [mean(bootsam) ; ci];


%% plot the before-and-after DET curves
figure()

nboot = 0;

[X0,Y0,~] = perfcurve(labels, scores, 1, 'XCrit', 'fpr', 'YCrit', 'fnr', ...
    'TVals', 'all', 'Weights', weights, 'NBoot', nboot, 'Options', bootopts);
if nboot > 0
    errorbar(X0(:,1),Y0(:,1),Y0(:,1)-Y0(:,2),Y0(:,3)-Y0(:,1), 'b');
else
    plot(X0,Y0,'b');
end

hold 'on'; grid 'on';

scores = [imp + deltaImp(end) ; gen + deltaGen(end)];
[XN,YN,~] = perfcurve(labels, scores, 1, 'XCrit', 'fpr', 'YCrit', 'fnr', ...
    'TVals', 'all', 'Weights', weights, 'NBoot', nboot, 'Options', bootopts);
if nboot > 0
    errorbar(XN(:,1),YN(:,1),YN(:,1)-YN(:,2),YN(:,3)-YN(:,1), 'r');
else
    plot(XN,YN,'r');
end

xlabel('FMR'); ylabel('FNMR');
axis([-1e-5 .1 0 .1]);
```

```
174
175 end
176
177
178
179 function r = truematch(X, L, fmrRef)
180
181 % paramaterized bootfun for permNaive, implementing the TMR at reference
182 % FMR
183
184 %% find the sample threshold for the reference FMR
185 t0 = prctile(X(L==0), 100 .* (1. - fmrRef));
186
187 r = sum(X(L==1) > t0) / sum(L==1);
188
189 end
190
191
192
193 function Pb = permbio(X, L, delta, fmrRef)
194
195 % paramaterized bootfun for permMatchedDelta, implementing the Pb biometric
196 % permanence metric
197
198 %% find the sample threshold for the reference FMR
199 t0 = prctile(X(L==0), 100 .* (1. - fmrRef));
200
201 %% find the sample permanence at each delta
202 c0 = sum(X(L==1) > t0);
203 Pb = sum(X(L==1) > t0+delta)/c0;
204
205 % note that since 1-FNMR = TMR = fraction of genuine scores > threshold,
206 % Pb reduces to the ratio of the above-threshold count at threshold 't'
207 % to that at threshold 't0', where 't' and 't0' are thresholds
208 % corresponding to the chosen reference FMR at some age and at 'age0'
209
210 end
```

Appendix D

A note on the Rayleigh synthetic match score distributions

In Chapter 4, the choice was made to use Rayleigh distributions for the synthetic genuine and imposter match scores. The initial justification for this choice was simply that they "look about right": for the imposter scores, we sought a well-known distribution that had a hard cut-off at a score of zero, with positive skew to represent a long tail of high-scoring imposters (potential false matches). For the genuine scores, a "flipped" Rayleigh shape similarly provided a hard cut-off maximum score – normalized to one in our simulations – with a long tail of low-scoring genuines (potential false non-matches).

No attempt was made to match the synthetic distributions to the actual score distributions of the devices in our study. In fact, the actual distributions vary significantly, reflecting differences in the scoring algorithms between them (Figure 5.3 – 5.11).

One nice consequence of the choice of Rayleigh distributions is that the tail integrals (for the FMR and FNMR) in our classification model remain obtainable in closed form after the addition of noise, as will be shown in the following section. This

provides some additional – albeit post-hoc – justification for their choice.

D.1 Tail integral for the FMR

In the model of Section 4.2.2, a true biometric match score s is drawn from a genuine (G) or imposter (I) score distribution, to which is added zero-mean Gaussian noise W. The probability density function (pdf) of the sum of these random variables is the convolution $p_W * p_{I,G}$ of their pdfs where

$$f * g := \int_{-\infty}^{\infty} f(x')g(x - x')dx' \tag{D.1}$$

For the pdf of the imposter scores, we have from Equation 4.5

$$p_I(s) = \frac{s}{\beta_I^2} e^{-s^2/2\beta_I^2}; s \geq 0 \tag{D.2}$$

while the pdf of the noise may be written as

$$p_W(s) = \frac{1}{\sqrt{2\pi}\sigma} e^{-s^2/2\sigma^2} \tag{D.3}$$

so that the pdf of the sum becomes

$$p_{I+W}(s) = \frac{1}{\sqrt{2\pi}\sigma} \int_0^{\infty} \frac{s'}{\beta_I^2} e^{-s'^2/2\beta_I^2} e^{-(s-s')^2/2\sigma^2} ds' \tag{D.4}$$

The "trick" at this point is to realize that it is not necessary to evaluate this convolution in closed form, in order to evaluate the tail integral. In particular, we may write the resulting FMR as

$$Pr_{I+W}\{s > \theta\} = Pr_{W+I}\{s > \theta\} = \int_{\theta}^{\infty} \int_{-\infty}^{\infty} p_W(s')p_I(s - s')ds'ds \tag{D.5}$$

which, changing the order of integration, becomes

$$Pr_{I+W}\{s > \theta\} = \int_{-\infty}^{\infty} p_W(s') \int_{\theta}^{\infty} p_I(s - s') ds ds' \tag{D.6}$$

in which the inner integral (with a further change of variables $s - s' \rightarrow s$) is now just

$$\int_{\theta}^{\infty} p_I(s - s') ds = \int_{\theta - s'}^{\infty} p_I(s).ds \tag{D.7}$$

which is just the probability that the noise-free imposter score is greater than $\theta - s'$. For the chosen Rayleigh distribution, this is easy to evaluate since

$$\frac{d}{ds} e^{-s^2/2\beta_I^2} = -\frac{s}{\beta_I^2} e^{-s^2/2\beta_I^2} \tag{D.8}$$

so that

$$\int_{\theta}^{\infty} p_I(s - s') ds = \left[e^{-s^2/2\beta_I^2} \right]_{\infty}^{(\theta - s')} = e^{-(\theta - s')^2/2\beta_I^2} \tag{D.9}$$

Plugging this back into D.6, we have

$$Pr_{I+W}\{s > \theta\} = \frac{1}{\sqrt{2\pi}\sigma} \int_{\theta}^{\infty} e^{-s'^2/2\sigma^2} e^{-(\theta - s')^2/2\beta_I^2} ds' \tag{D.10}$$

in which the integrand is just a product of two exponentials.[1] From here on in, it is just a matter of completing the square - that is,

$$\frac{s'^2}{2\sigma^2} + \frac{(\theta - s')^2}{2\beta_I^2} = \frac{\beta_I^2 s'^2 + \sigma^2(\theta - s')^2}{2\sigma^2\beta_I^2} \tag{D.11}$$

$$= \frac{\beta_I^2 + \sigma^2}{2\sigma^2\beta_I^2} \left[s'^2 - \frac{2\sigma^2\theta}{\beta_I^2 + \sigma^2} s' + \frac{\sigma^2\theta^2}{\beta_I^2 + \sigma^2} \right] \tag{D.12}$$

$$= \frac{\beta_I^2 + \sigma^2}{2\sigma^2\beta_I^2} \left[\left(s' - \frac{\sigma^2\theta}{\beta_I^2 + \sigma^2} \right)^2 + \frac{\sigma^2\theta^2}{\beta_I^2 + \sigma^2} - \frac{\sigma^4\theta^2}{\beta_I^2 + \sigma^2} \right] \tag{D.13}$$

[1] If we had preserved the original order, taking the inner integral over the noise distribution, the integrand would have been the product of the Rayleigh distribution with an error function

giving

$$Pr_{I+W}\{s > \theta\} = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\theta^2}{2\beta_I^2}\left(1-\frac{\sigma^2}{\beta_I^2+\sigma^2}\right)} \int_{-\infty}^{\infty} e^{-\frac{\beta_I^2+\sigma^2}{2\sigma^2\beta_I^2}\left(s'-\frac{\sigma^2\theta}{\beta_I^2+\sigma^2}\right)^2} ds' \tag{D.14}$$

in which the standard definite integral may be evaluated from tables, yielding

$$Pr_{I+W}\{s > \theta\} = \sqrt{\frac{\beta_I^2}{\beta_I^2+\sigma^2}} e^{-\theta^2/2\left(\beta_I^2+\sigma^2\right)} \tag{D.15}$$

By inspection, (D.15) implies that the noisy score remains Rayleigh distributed, with parameter $\beta_I^2 \to \beta_I^2 + \sigma^2$, i.e.

$$p_{I+W}(s) = \frac{s}{\beta_I^2+\sigma^2} e^{-s^2/2\left(\beta_I^2+\sigma^2\right)}; s \geq 0 \tag{D.16}$$

D.2 Tail integral for the FNMR

Evaluation of the FNMR from the tail of the genuine distribution proceeds using all the same steps as the previous section, but backwards and in heels.

Bibliography

[1] A. Adler, R. Youmaran, and S. Loyka. Towards a measure of biometric information. In *2006 Canadian Conference on Electrical and Computer Engineering*, pages 210–213, 2006. DOI: 10.1109/CCECE.2006.277447.

[2] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman. Eigenfaces vs. fisherfaces: recognition using class specific linear projection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 19(7):711–720, July 1997. ISSN: 0162-8828. DOI: 10.1109/34.598228.

[3] A. Bertillon. *Signaletic Instructions Including the Theory and Practice of Anthropometrical Identification*. Werner Company, 1896.

[4] H. Bose. *Hints on Finger-Prints, with a Telegraphic Code for Finger Impressions*. Thacker, Spink and Company, Calcutta/Simla, 1917.

[5] M. E. Brockly. *The Role of Test Administrator and Error*. Master's thesis, Purdue University, West Lafayette, IN, 2013.

[6] C. Camden. Elizabethan chiromancy. *Modern Language Notes*, 62(1):1–7, 1947. ISSN: 01496611.

[7] L. Chen, J. Wang, S. Yang, and H. He. A finger vein image-based personal identification system with self-adaptive illuminance control. *IEEE Transactions on Instrumentation and Measurement*, 66(2):294–304, 2017. ISSN: 0018-9456. DOI: 10.1109/TIM.2016.2622860.

[8] T. M. Cover and J. A. Thomas. *Channel capacity*. In *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, New York, NY, USA, 2006. Chapter 7. ISBN: 0471241954.

[9] T. M. Cover and J. A. Thomas. *Entropy, relative entropy, and mutual information*. In *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, New York, NY, USA, 2006. Chapter 2. ISBN: 0471241954.

[10] D. R. Cox. Biometrika: The first 100 years. *Biometrika*, 88(1):3–11, Feb. 2001. ISSN: 0006-3444. DOI: 10.1093/biomet/88.1.3.

[11] N. J. Crane, E. G. Bartick, R. S. Perlman, and S. Huffman. Infrared spectroscopic imaging for noninvasive detection of latent fingerprints. *Journal of Forensic Sciences*, 52(1):48–53, 2007. DOI: 10.1111/j.1556-4029.2006.00330.x.

[12] H. Cummins. Finger prints and attempted fraud. *New Orleans Medical and Surgical Journal*, 94:82–86, 1942.

[13] A. Czajka. Influence of iris template aging on recognition reliability. In M. Fernández-Chimeno, P. L. Fernandes, S. Alvarez, D. Stacey, J. Solé-Casals, A. Fred, and H. Gamboa, editors, *Biomedical Engineering Systems and Technologies*, pages 284–299, Berlin, Heidelberg. Springer Berlin Heidelberg, 2014. ISBN: 978-3-662-44485-6.

[14] J. Daugman. New methods in iris recognition. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 37(5):1167–1175, 2007. ISSN: 1083-4419. DOI: 10.1109/TSMCB.2007.903540.

[15] J. Daugman. Probing the uniqueness and randomness of iriscodes: results from 200 billion iris pair comparisons. *Proceedings of the IEEE*, 94(11):1927–1935, 2006. ISSN: 0018-9219. DOI: 10.1109/JPROC.2006.884092.

[16] J. Daugman. Information theory and the iriscode. *IEEE Transactions on Information Forensics and Security*, 11:400–409, 2016.

[17] G. Doddington, W. Liggett, A. Martin, M. Przybocki, and D. Reynolds. SHEEP, GOATS, LAMBS and WOLVES A statistical analysis of speaker performance in the nist 1998 speaker recognition evaluation. In *International Conference on Spoken Language Processing*, 1998.

[18] M. Drahansky, M. Dolezel, J. Urbanek, E. Brezinova, and T. Kim. Influence of skin diseases on fingerprint recognition. *Journal of Biomedicine and Biotechnology*, 2012. Article ID 626148.

[19] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification (2nd Edition)*. Wiley-Interscience, New York, NY, USA, 2000. ISBN: 0471056693.

[20] S. Elliott, E. Kukula, and N. Sickler. The challenges of the environment and the human/biometric device. In *Proc. International Workshop on Biometric Technologies*, 2004.

[21] D Exline, C Wallace, C. Roux, C. Lennard, N Nelson, and P Treado. Forensic applications of chemical imaging: latent fingerprint detection using visible absorption and luminescence. English. *Journal of forensic Science*, 48(5):1047–1053, 2003. ISSN: 0022-1198.

[22] S. P. Fenker and K. W. Bowyer. Experimental evidence of a template aging effect in iris biometrics. In *Proc. 2011 IEEE Workshop on Applications of Computer Vision (WACV) (WACV '11)*, pages 232–239, 2011.

[23] W. Freude, R. Schmogrow, B. Nebendahl, M. Winter, A. Josten, D. Hillerkuss, S. Koenig, J. Meyer, M. Dreschmann, M. Huebner, C. Koos, J. Becker, and J. Leuthold. Quality metrics for optical signals: Eye diagram, Q-factor, OSNR, EVM and BER. In *2012 14th International Conference on Transparent Optical Networks (ICTON)*, pages 1–4, 2012. DOI: 10.1109/ICTON.2012.6254380.

[24] J. Galbally, S. Marcel, and J. Fierrez. Biometric antispoofing methods: a survey in face recognition. *IEEE Access*, 2:1530–1552, 2014. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2014.2381273.

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

[26] P Grother, J. Matey, E Tabassi, G W. Quinn, and M Chumakov. IREX-VI temporal stability of iris recognition accuracy. *NIST Interagency Report 7948*, July 2013.

[27] P. Grother and E. Tabassi. Performance of biometric quality measures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(4):531–543, 2007. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2007.1019.

[28] P. Grother, W. Salamon, and R. Chandramouli. Biometric Data Specification for Personal Identity Verification. ST Special Publication 800-76-2, NIST, 2013.

[29] P. Grother, W. Salamon, and R. Chandramouli. Machine Readable Travel Documents Part 9: Deployment of Biometric Identification. Doc 9303 Part 9, International Civil Aviation Organization, 2015. Seventh Edition, 2015.

[30] J. Harvey, J. Campbell, and A. Adler. Characterization of biometric template aging in a multiyear, multivendor longitudinal fingerprint matching study. *IEEE Transactions on Instrumentation and Measurement*:1–9, 2018. ISSN: 0018-9456. DOI: 10.1109/TIM.2018.2861998.

[31] J. Harvey, J. Campbell, S. Elliott, M. Brockly, and A. Adler. Biometric permanence: definition and robust calculation. In *2017 Annual IEEE International Systems Conference (SysCon)*, pages 1–7, 2017. DOI: 10.1109/SYSCON.2017.7934760.

[32] E. Henry. *Classification and Uses of Finger Prints*. Making of modern law. George Routledge and Sons, 1900.

[33] Sir W. J. Herschel, Bart. *The Origin of Finger-Printing*. Humphrey Milford / Oxford University Press, 1916.

[34] H. Hofbauer, I. Tomeo-Reyes, and A. Uhl. Isolating iris template ageing in a semi-controlled environment. In *2016 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–5, 2016. DOI: 10.1109/BIOSIG.2016.7736929.

[35] J. I. Hoffman. Chapter 13 - hypergeometric distribution. In J. I. Hoffman, editor, *Biostatistics for Medical and Biomedical Practitioners*, pages 179 –182. Academic Press, 2015. ISBN: 978-0-12-802387-7. DOI: https://doi.org/10.1016/B978-0-12-802387-7.00013-5.

[36] ICAO. Deployment of Biometric Identification and Electronic Storage of Data in eMRTDs. Doc 9303, International Civil Aviation Organization, (Seventh Ed.) 2015. Machine Readable Travel Documents.

[37] ILO. ILO Seafarers' Identity Documents Biometric Testing Campaign Report Part I. No. 185, International Labour Organization, 2004. Seafarers' Identity Documents Convention (Revised), 2003.

[38] ILO. The standard for the biometric template required by the Convention. No. 185, International Labour Organization, 2006. Seafarers' Identity Documents Convention (Revised), 2003.

[39] ISO. Information technology – Security techniques – Biometric information protection. Standard 24745, International Organization for Standardisation, Geneva, Switzerland, 2011.

[40] ISO. ISO/IEC 19795-1:2006 - Information technology – Biometric performance testing and reporting – Part 1: Principles and framework. Technical report, International Organization for Standardisation, 2016.

[41] A. Jain, A. Ross, and S. Prabhakar. An introduction to biometric recognition. *Biometrics*, 2004.

[42] A. K. Jain, P. Flynn, and A. A. Ross. *Handbook of Biometrics*. Springer Publishing Company, Incorporated, 1st edition, 2010. ISBN: 1441943757, 9781441943750.

[43] A. Juels and M. Sudan. A fuzzy vault scheme. *Designs, Codes and Cryptography*, 38(2):237–257, 2006. ISSN: 1573-7586. DOI: 10.1007/s10623-005-6343-z.

[44] A. Juels and M. Wattenberg. A fuzzy commitment scheme. In *Proceedings of the 6th ACM Conference on Computer and Communications Security*, CCS '99, pages 28–36, Kent Ridge Digital Labs, Singapore. ACM, 1999. ISBN: 1-58113-148-8. DOI: 10.1145/319709.319714.

[45] S. Kirchgasser and A. Uhl. Template ageing in non-minutiae fingerprint recognition. In *2017 5th International Workshop on Biometrics and Forensics (IWBF)*, pages 1–6, 2017. DOI: 10.1109/IWBF.2017.7935091.

[46] A. W. K. Kong, D. Zhang, and M. Kamel. *An introduction to the iriscode theory*. In *Handbook of Iris Recognition*. M. J. Burge and K. W. Bowyer, editors. Springer London, London, 2013, pages 321–336. ISBN: 978-1-4471-4402-1. DOI: 10.1007/978-1-4471-4402-1_16.

[47] A. Lanitis and N. Tsapatsoulis. On the analysis of factors influencing the performance of facial age progression. In *2016 4th International Conference on Biometrics and Forensics (IWBF)*, pages 1–6, 2016. DOI: 10.1109/IWBF.2016.7449697.

[48] A. Lanitis and N. Tsapatsoulis. Quantitative evaluation of the effects of aging on biometric templates. *IET Computer Vision*, 5(6):338–347, 2011. ISSN: 1751-9632. DOI: 10.1049/iet-cvi.2010.0197.

[49] P. Lee, H. Guan, A. Dienstfrey, M. Theofanos, B. Stanton, and M. T. Schwarz. Forensic Latent Fingerprint Preprocessing Assessment. Technical report NISTIR 8215, Natitonal Institute of Standards and Technology, 2018.

[50] M. Li and P. Vitányi. *An Introduction to Kolmogorov Complexity and Its Applications*. Texts in Computer Science. Springer New York, 2009. ISBN: 9780387498201.

[51] N. A. Makhdoomi, T. S. Gunawan, and M. H. Habaebi. Gait recognition and effect of noise on the recognition rate. In *2013 IEEE International Conference on Smart Instrumentation, Measurement and Applications (ICSIMA)*, pages 1–4, 2013. DOI: 10.1109/ICSIMA.2013.6717927.

[52] D. Maltoni, R. Cappelli, and D. Meuwly. *Automated fingerprint identification systems: from fingerprints to fingermarks.* In *Handbook of Biometrics for Forensic Science*. M. Tistarelli and C. Champod, editors. Springer International Publishing, Cham, 2017, pages 37–61. ISBN: 978-3-319-50673-9. DOI: 10.1007/978-3-319-50673-9_3.

[53] D. Maltoni, D. Maio, A. Jain, and S. Prabhakar. *Handbook of fingerprint recognition*. Springer Science & Business Media, 2009.

[54] I. Manjani, H. Sumerkan, P. J. Flynn, and K. W. Bowyer. Template aging in 3D and 2D face recognition. In *2016 IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–6, 2016. DOI: 10.1109/BTAS.2016.7791202.

[55] Y. Matveev. The problem of voice template aging in speaker recognition systems. In *Proc. 15th International Conf. Speech and Computer (SPECOM) 2013*, Pilsen, Czech Republic, 2013.

[56] C. Z. Mooney, R. D. Duval, and R. Duvall. *Bootstrapping: A nonparametric approach to statistical inference*, number 94-95. Sage, 1993.

[57] J. Neyman and E. S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 231(694-706):289–337, 1933. ISSN: 0264-3952. DOI: 10.1098/rsta.1933.0009.

[58] G. Panis, A. Lanitis, N. Tsapatsoulis, and T. F. Cootes. Overview of research on facial ageing using the FG-NET ageing database. *IET Biometrics*, 5(2):37–46, 2016. ISSN: 2047-4938. DOI: 10.1049/iet-bmt.2014.0053.

[59] C. J. Polson. Finger prints and finger printing: an historical study. *Crim. L. & Criminology*, 41(5):690–704, 1951.

[60] A. Rattani, G. L. Marcialis, and F. Roli. An experimental analysis of the relationship between biometric template update and the Doddington's Zoo: a case study in face verification. In *ICIAP*, 2009.

[61] A. P. Rebera and E. Mordini. *Biometrics and ageing: social and ethical considerations*. In *Age Factors in Biometric Processing*. Security. Institution of Engineering and Technology, 2013, pages 37–62. DOI: 10.1049/PBSP010E_ch3.

[62] A. Ross and A. Jain. Information fusion in biometrics. *Pattern Recognition Letters*, 24(13):2115 –2125, 2003. ISSN: 0167-8655. DOI: https://doi.org/10.1016/S0167-8655(03)00079-5. Audio- and Video-based Biometric Person Authentication (AVBPA 2001).

[63] A. Ross, A. Rattani, and M. Tistarelli. Exploiting the "doddington zoo" effect in biometric fusion. *2009 IEEE 3rd International Conference on Biometrics: Theory, Applications, and Systems*:1–7, 2009.

[64] J. Ryu, J. Jang, and H. Kim. Analysis of effect of fingerprint sample quality in template ageing. In *NIST Biometric Quality Workshop II, November 2007*. NIST, 2007.

[65] W. J. Scheirer and T. E. Boult. Cracking fuzzy vaults and biometric encryption. In *Biometrics Symposium, 2007*, pages 1–6, 2007. DOI: 10.1109/BCC.2007.4430534.

[66] W. J. Scheirer, W. Bishop, and T. E. Boult. *Beyond PKI: the biocryptographic key infrastructure*. In *Security and Privacy in Biometrics*. P. Campisi, editor. Springer London, London, 2013, pages 45–68. ISBN: 978-1-4471-5230-9. DOI: 10.1007/978-1-4471-5230-9_3.

[67] N. C. Sickler and S. J. Elliott. An evaluation of fingerprint quality across an elderly population vis-à-vis 18–25 year olds. In *Security Technology, 2005. CCST '05. 39th Annual 2005 International Carnahan Conference on*, pages 68,73, 2005. DOI: 10.1109/CCST.2005.1594817.

[68] D. P. Sidlauskas and S. Tamer. *Handbook of Biometrics*. In A. K. Jain, P. Flynn, and A. S. Ross, editors. Springer Publishing Company, Incorporated, 2008. Chapter Hand Geometry Recognition. ISBN: 978-0-387-71040-2.

[69] N. Singh, R. Khan, and R. Shree. Applications of speaker recognition. *Procedia Engineering*, 38:3122 –3126, 2012. ISSN: 1877-7058. DOI: https://doi.org/10.1016/j.proeng.2012.06.363. International Conference on Modelling Optimization and Computing.

[70] G. S. Sodhi and J. Kaur. The forgotten Indian pioneers of fingerprint science. *Current Science*, 88(1):185–191, 2005. ISSN: 00113891.

[71] R. F. Stewart, M. Estevao, and A. Adler. Fingerprint recognition performance in rugged outdoors and cold weather conditions. In *IEEE Int. Conf. Biometrics: Theory, Applications and Systems (BTAS09)*, pages 28–30, Washington DC USA, 2009.

[72] E. Tabassi. Image specific error rate: a biometric performance metric. *2010 20th International Conference on Pattern Recognition*:1124–1127, 2010.

[73] E. Tabassi. NFIQ 2.0 : Design, implementation and performance evaluation. In *Proc. of the International Biometrics Performance Conference*. NIST, 2016.

[74] E. Tabassi, C. L. Wilson, and C. I. Watson. Fingerprint image quality. Technical report, National Institute for Standards and Technology, 2004.

[75] M. N. Teli, J. R. Beveridge, P. J. Phillips, G. H. Givens, D. S. Bolme, and B. A. Draper. Biometric zoos: theory and experimental evidence. In *2011 IEEE International Joint Conference on Biometrics, IJCB 2011, Washington, DC, USA, October 11-13, 2011*, pages 1–8, 2011. DOI: 10.1109/IJCB.2011.6117479.

[76] M. Trokielewicz, A. Czajka, and P. Maciejewicz. Cataract influence on iris recognition performance. In *Proc. SPIE*, volume 9290, 2014. DOI: 10.1117/12.2076040.

[77] U. Uludag, A. Ross, and A. Jain. Biometric template selection and update: a case study in fingerprints. *Pattern Recognition*, 37:1153–1542, 2004.

[78] US GAO. Technology Assessment: Using Biometrics for Border Security. Technical report, US Government Accountability Office, Washington, DC, 2002.

[79] J. W. Carls, R. A. Raines, M. Grimaila, and S. Rogers. Biometric enhancements: template aging error score analysis. In *Proc. 8th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2008)*, pages 1–8, Amsterdam, The Netherlands, Sept. 2008. DOI: 10.1109/AFGR.2008.4813331.

[80] J. L. Wayman. *Multifinger penetration rate and ROC variability for automatic fingerprint identification systems*. In *Automatic Fingerprint Recognition Systems*. N. Ratha and R. Bolle, editors. Springer New York, New York, NY, 2004, pages 305–316. ISBN: 978-0-387-21685-0. DOI: 10.1007/0-387-21685-5_15.

[81] M. Wittman, P. Davis, and P. J. Flynn. Empirical studies of the existence of the biometric menagerie in the FRGC 2.0 color image corpus. In *2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06)*, pages 33–33, 2006. DOI: 10.1109/CVPRW.2006.71.

[82] A. Wolf. Template aging in speech biometrics. *Proc. Biometrics: Challenges Arising from Theory to Practice (BCTP)*, 2004.

[83] P. M. Woodward. *Probability and Information Theory, with Applications to Radar*. Pergamon Press, Oxford, 2nd edition, 1964.

[84] N. Yager and T. Dunstone. The biometric menagerie. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(2):220–230, 2010. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2008.291.

[85] R. Yang. *Effects of Sensors, Age, and Gender on Fingerprint Image Quality*. Master's thesis, Carleton University, Ottawa, 2018. DOI: https://doi.org/10.22215/etd/2018-13180.

[86] S. Yoon and A. K. Jain. Longitudinal study of fingerprint recognition. *Proceedings of the National Academy of Sciences*, 112(28):8555–8560, 2015. ISSN: 0027-8424. DOI: 10.1073/pnas.1410272112.