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

John Harvey, *Student Member, IEEE,* John Campbell, and Andy Adler, *Senior Member, IEEE*

***Abstract*—Biometric features are known to change over time, presenting a challenge for their use in identity management systems. Viewed as an instrumentation and measurement problem, these changes represent a potential source of measurement or calibration error that need to be addressed at the system level in order to guarantee performance over the lifetime of the system. In this paper we develop a novel metric, which we call biometric permanence, to characterize the stability of biometric features. First, we define permanence in terms of the change in false non-match ratio (FNMR) over a repeated sequence of enrolment and verification events for a given population. However, since changes in FNMR are expected to be small, any variability in the biometric capture over time will camouflage the changes of interest. To address this issue, a robust methodology is proposed that can isolate the visit-to-visit variability, and substantially improve the estimation. We develop and characterize a heuristic statistical model for a biometric capture system, and apply it to a large dataset of fingerprint biometrics collected over a period of seven years on a variety of commercially-available capture devices. We discuss how this methodology can be used to isolate the effect of biometric template aging and to develop system-level strategies for dealing with it.**

I. Introduction

A BIOMETRIC identity management system (IDMS) provides the ability to identify, or to verify the claimed identity of, an individual based on a comparison between a presentation of some biometric trait such as a fingerprint [1], an iris image [2], a pattern of blood vessels [3], or an analysis of gait [4] and a stored record of the same trait commonly known as a *biometric template*.

An assumption underlying the deployment of such 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 [5]. 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 [6], while the onset of degenerative disease, injury, or occupational damage may affect fingerprints [7], [8]. 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 aging or weight gain for example.

The age progression of biometric traits has perhaps received most attention within the facial recognition modality. Lanitis & Tsapatsoulis [9] proposed a measure of biometric aging 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 [10], [11], rather than for biometric IDMSs. Meanwhile Manjani et al. [12] detected aging in 2D and 3D facial biometrics, by 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). Fingerprint aging might be expected to share some of the same physiological factors as face aging – in particular, skin textural changes and loss of tissue elasticity – and has been reported by Uludag et al. [13], who proposed to address its system-level implications via a template update scheme using prototypical templates based on either clustering or on mean feature distance. Aging has also been observed in iris templates [14], where it has been at least partially attributed to age-related changes in pupillary diameter [15]. The influence of biometric sample quality on template aging was highlighted by Ryu et al. [16], who found that lower sample quality (evaluated using the NIST NFIQ measure [17]) was associated with an increased number of matching errors.

In common with many other instrumentation and measurement systems, biometric systems are subject to numerous sources of error. In order to develop strategies to ameliorate such errors, it is useful to separate and characterize them individually [18]. For example, random errors might be addressed by increasing the signal-to-noise ratio (SNR) margin, whereas systematic drifts may require development of re-calibration strategies: in the case of a biometric IDMS, that might take

John Harvey and Andy Adler are with the Department of Systems and Computer Engineering, Carleton University, Ottawa, ON K1S 5B6, Canada e-mail: jharvey@sce.carleton.ca

John Campbell is with Bion Biometrics, Ltd., Nepean, ON, Canada

the form of a periodic re-enrolment requirement. The chief difficulty in evaluating biometric template aging lies in the small effect size and confounding factors including physical environment (particularly temperature and humidity [19]), operator and/or subject acclimation [20], and degradation of the particular biometric capture hardware – in the case of the fingerprint modality, this might include scratching or marring of sensor platens for example. In the context of a longitudinal study, these sources of error are essentially systematic in the sense that they affect all biometric presentations under a particular set of test conditions: since biometric comparisons necessarily involve both a current presentation and a gallery of previously enrolled templates, each comparison is affected by two such systematic terms, which we refer to as *visit biases*.

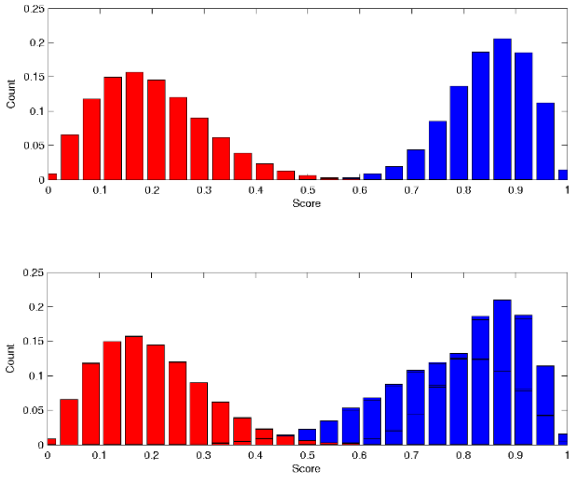
The goal of this work is to characterize template ageing in the fingerprint modality, for a number of commercially-available fingerprint sensor devices and technologies, and to understand its impact on the deployment and operation of fingerprint-based IDMSs. First, in Section II we outline the definition and properties of our metric, Biometric Permanence, P_B; next in Section III we describe the design of our study, including subject demographics and data colletction protocols. Section IV proposes a heuristic model for the study data and describes, with select results, the methodology used to estimate P_B. Finally, in Section V we attempt to justify, through further data analysis, the key assumptions underlying the methodology.

II. Biometric Permanance

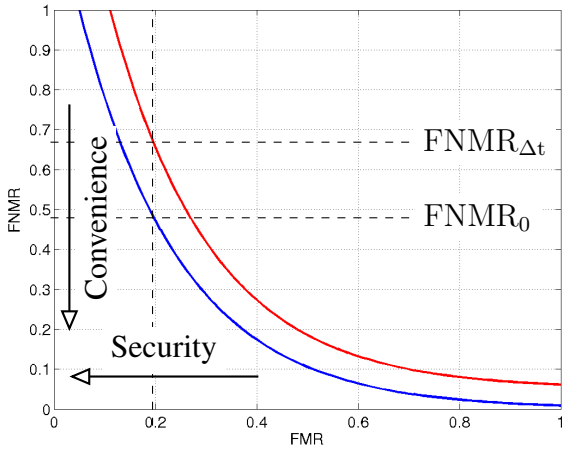
Here we expand on [21], in which we proposed a measure called *biometric permanence*, $P_B(\Delta t)$, at a given elapsed time Δt, as follows

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

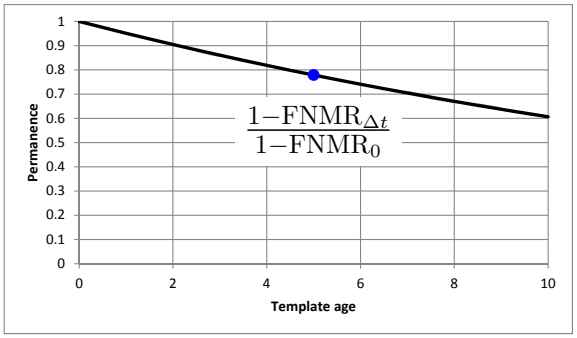
based on the change in false non-match rate (FNMR) at a given false match rate (FMR) [22]. The definition was motivated by operational considerations i.e. that template ageing will manifest itself as a decrease in the security and/or convenience provided by the biometric system at a given operating point. Since it is usually operational security that is of primary concern, it is natural to fix FMR and consider the change in FNMR. A schematic overview of the development, starting from the empirical match scores, is shown in Fig. 1. At time zero, we enrol subjects into a biometric IDMS, generating a set of biometric templates. At the same time, we capture an (independent) set of baseline verification images. These images are compared against the templates to give a collection of labelled (i.e. genuine or imposter) biometric match scores, whose distributions may or may not be completely separable at some decision threshold θ. Some time later, new verification images are obtained, and the corresponding genuine and imposter match scores are evaluated again. Changes in the match score distributions will be manifested in a shift of the decision error tradeoff (DET) curve i.e. a change in the FNMR at a given FMR. This change in FNMR is expressed as a permanence value for the enrolment-verification time interval. As well as reflecting the underlying performance degradation



(a)



(b)



(c)

Fig. 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 tradeoff (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. 1

mechanism, other desirable features of this formulation are:

- permanence P_B increases towards unity as $\mathrm{FMNR}_{\Delta t}$ tends towards FMNR_0; this case would correspond to a perfectly permanent template
- permanence P_B decreases towards zero as the $\mathrm{FMNR}_{\Delta t}$ increases towards unity; a biometric template might be said to be completely impermanent at this point

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

In [21], we assumed that the performance degradation would be dominated by changes in the genuine match score distribution, implying that, *for a fixed decision threshold*, the FMR would remain constant while the FNMR degraded: this is generally the most desirable degradation mode for a biometric system since it would result in no loss in security. In the present work, we remove that restriction and allow for variation in both the genuine and imposter scores. In an operational setting, the formulation of P_B according to Eq. 1 then implies adjustment of the decision boundary in order to maintain the desired FMR. We discuss the relative magnitudes of the imposter and genuine distribution variabilities for the devices in our study in Section V.

III. Study Protocol and Demographics

In order to detect biometric template aging, and to evaluate our methodology, we need a dataset of similarity scores evaluated for the same subjects at different enrolment-verification time intervals. Ideally the biometric collection should take place under well-controlled conditions with a consistent protocol, in order to control (so far as possible) for confounding environmental factors.

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 2). The protocol for each subject visit consisted of a sequence of two-finger enrolments, followed by a sequence of single-finger verification presentations [23], [24]. 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.

The study was approved by the Carleton University Research Ethics Board, subject to restrictions on the storage and sharing of personally identifiable information.

Fig. 2: Overlap of participants between data collection phases (the 2013 collection is omitted for clarity; it overlaps almost completely with 2012).

TABLE I: Available 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

Twelve different commercially-available fingerprint sensor devices were obtained, representing multiple vendors and technologies: single-spectral optical, multi-spectral optical and capacitive (Table I). Unfortunately, the contractual terms under which the fingerprint device vendors provided acquisition devices and software to the study do not permit more detailed attribution. 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 II).

TABLE II: 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

IV. Analysis

A. Methodology

As in [21], we seek to evaluate biometric permanence, $P_B(\Delta t)$ according to Eq. 1, where FMR and FNMR are the false match and false non-match rates obtained by binary classification of a set of labeled match scores, each score corresponding to a (generally, vendor-dependent) measure of similarity between a presentation of a subject's biometric at occasion n ("verification"), and a biometric template from a gallery of such templates recorded on occasion m ("enrolment"), with $\Delta t_{n,m}$ being the time interval between enrolment and verification, or *template age*.

Our 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 [23], [24]. 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 for (respectively), the enrolment visit m and the verification visit n;
- 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 modeled as

$$\overline{s}_{nm}^{ji} = s_{nm}^{ji} + a_m + b_n + \overline{W}^{ji} \tag{2}$$

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

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

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

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

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

$$\mathrm{var}\left(\frac{1}{2}\sum_{k=0}^{3}(-1)^k \overline{W}_{ji}^{(k)}\right) = \mathrm{var}\left(\overline{W}_{ji}^{(k)}\right) \tag{4}$$

leaving the signal-to-noise ratio of the measurement effectively unchanged.

Phenomenologically, a_m (defined as a positive constant) would represent an amount by which all enrolments in visit m read "better than" their true value, with b_n being the corresponding amount for verification visit n. This is really the simplest model we can envisage, in which the confounding factors of enrolment and verification are considered to be independent – the extent to which this model is reflected in the real data will determine the success of the method, which we investigate below.

B. Data analysis

In our procedure, the averaged "matched deltas" $\Delta\overline{s}_{nm}^{ji}$ from Eq. 3 are averaged again across a particular pair of enrolment and verification visits m, n to give mean genuine and imposter score offsets $\Delta\overline{s}_{nm}^{G}$ and $\Delta\overline{s}_{nm}^{I}$ for the visit pair. We then aggregate the corresponding zero-time genuine and imposter scores $\{\overline{s}_{kk}^{ii}\}, \{\overline{s}_{kk}^{j\neq i}\}; k \in 1 \ldots N$ and use these aggregate distributions shifted by the respective mean offsets $\Delta\overline{s}_{nm}^{G}$, $\Delta\overline{s}_{nm}^{I}$ to evaluate P_B according to Eq. 1 at time interval Δt_{mn}. We use bootstrap resampling [25] of the aggregate distributions in order to estimate 95% confidence intervals for P_B, as follows. First we arrange the aggregate genuine and imposter scores into a vector $\left(\overline{s}_{kk}^{ii}, \overline{s}_{kk}^{j\neq i}\right)$ 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 aggregate vector is then resampled, with replacement, $n_B = 1000$ times with sampling weights inversely proportional to class size in order to remove class imbalance.

Since an offset $\Delta\overline{S}^I$ to the imposter distribution is exactly equivalent to a shift in the threshold $\theta \rightarrow \theta + \Delta\overline{S}^I$ for the chosen FMR, we just need to evaluate $1 - \mathrm{FNMR}$ (or, equivalently, the true match rate TMR) at a set of thresholds $\theta_{nm} = \hat{\theta}_0 + \Delta\overline{s}_{nm}^{I} - \Delta\overline{s}_{nm}^{G}$. 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

$$\hat{P}_B = \frac{\left|\{\overline{s}_{kk}^{ii} : \overline{s}_{kk}^{ii} > \hat{\theta}_0 + \Delta\overline{s}_{nm}^{I} - \Delta\overline{s}_{nm}^{G}\}\right|}{\left|\{\overline{s}_{kk}^{ii} : \overline{s}_{kk}^{ii} > \hat{\theta}_0\}\right|} \tag{5}$$

TABLE III: 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

evaluated for each enrolment-verification visit pair n, m (Fig. 3).

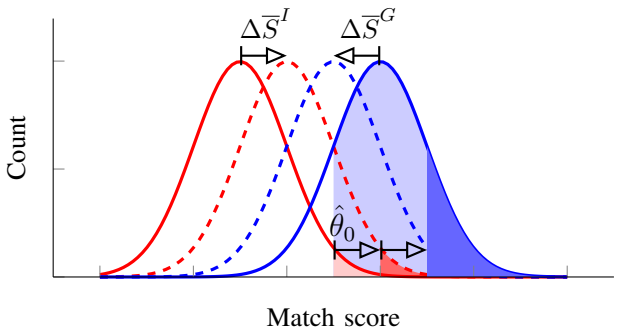


Fig. 3: A shift in the mean imposter score results in a shift in the estimated decision threshold $\hat{\theta}_0$ for a specified FMR (red area) – and a corresponding change in the achievable TMR (blue area) for the shifted genuine scores.

C. Results

Representative results of this procedure are shown graphically in Fig. 4, with comparison to a "naive" evaluation that does not attempt to account for visit bias.

In the top and middle rows of Fig. 4 we see the evolution of the typical observed aging behavior of the devices in our study. First, we note that the baseline ($\Delta t = 0$) score distributions Fig. 4a, Fig. 4d 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 tradeoff (DET) curves Figs. 4b, 4e 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. 4c,4f we see the permanence P_B according to Eq.1 decrease monotonically away from template age $\Delta t = 0$.

Two of the available devices (B and J) did not show this typical behavior. Instead, they showed well-separated genuine and imposter score distributions at $\Delta t = 0$ (Fig. 4g) 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 $\mathrm{FNMR} = 0$ at $\mathrm{FNMR} = 0$ (Fig. 4h) and correspondingly no discernable change in permanence P_B in Fig. 4i.

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

V. 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.

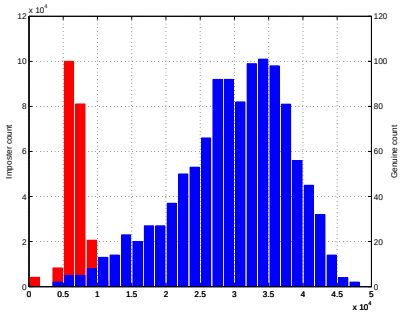
For the two devices that showed no change in permanence, the analysis is likely affected by the large class imbalance inherent in such biometric comparisons. That is, for a dataset of K distinct fingers, there are of order K^2 imposter matches but only K genuine matches, which causes the tails of the 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 IV-C attempts to ameliorate this effect, if the empirical distributions are seperable, 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 aging behavior 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 aging 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.

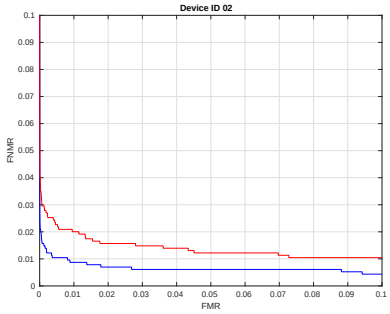
In the following sections we discuss some other aspects of the data, and their potential impact upon the interpretation of our results.

A. Time symmetry of the match scores

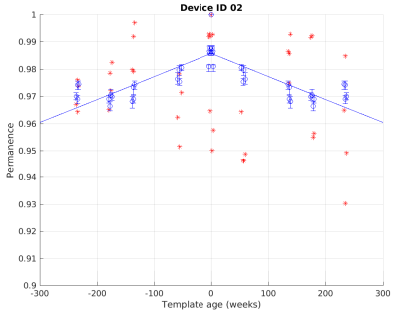
A key assumption that allows us to substantially remove the visit-to-visit bias factors is that the underlying "true" match scores are time-symmetric: that is, in the absence of these factors, comparisons between a biometric enrolment obtained at time t_1 and a set of verification presentations at later 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



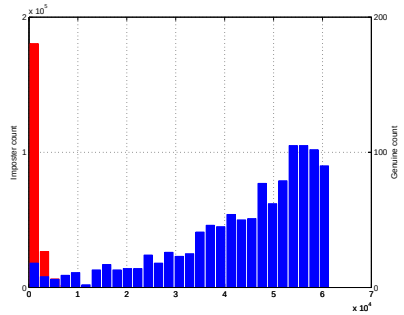
(a) Genuine (blue) and Imposter (red) score counts at $\Delta t = 0$



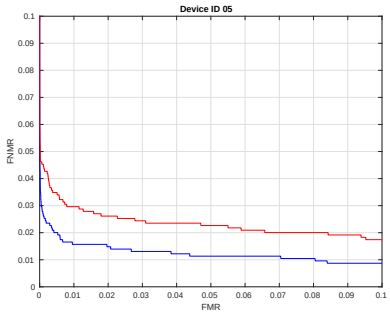
(b) Decision error tradeoff (DET) curve at $\Delta t = 0$ (blue) and at $\Delta t = 373$ weeks (red)



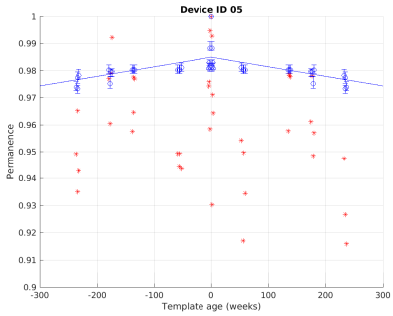
(c) Permanence versus template age: "naive" calculation (red); present method (blue)



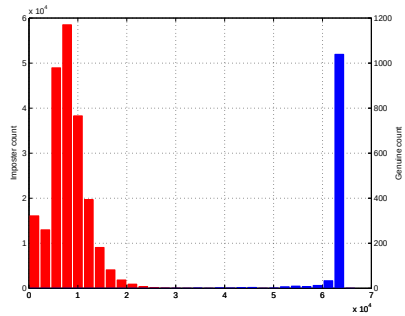
(d) Genuine (blue) and Imposter (red) score counts at $\Delta t = 0$



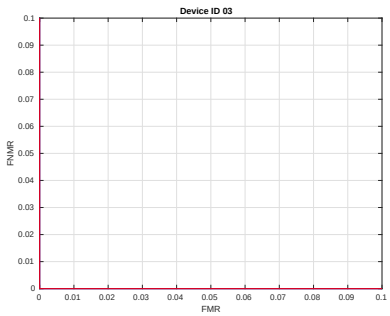
(e) Decision error tradeoff (DET) curve at $\Delta t = 0$ (blue) and at $\Delta t = 373$ weeks (red)



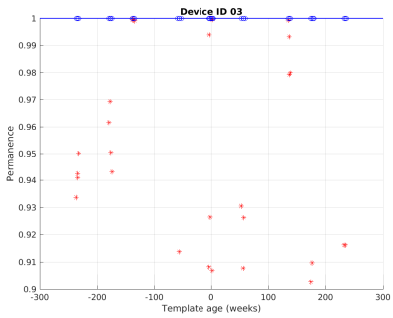
(f) Permanence versus template age: "naive" calculation (red); present method (blue)



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



(h) Decision error tradeoff (DET) curve at $\Delta t = 0$ (blue) and at $\Delta t = 373$ weeks (red)



(i) Permanence versus template age: "naive" calculation (red); present method (blue)

Fig. 4: *(select results)* (a)-(c): Device L (capacitive); (d)-(f): Device K (optical); (g)-(i) Device B (optical). The histograms (column 1) are scaled to account for the large class imbalance between Genuine and Imposter. DET curves (column 2) are generated using the "matched delta" methodology described in the text. The permanence results (column 3) 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; the solid lines represent simple best fits to the data and are intended only as an aid to visualization.

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.

B. 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

[1]This in fact was an assumption made in our previous work [21].

define a discriminability measure Q_{nm} for a pair of visits n, m as the ratio of the difference in sample mean score between genuine and imposter presentations to the sum of their sample standard deviations

$$Q_{nm} = \frac{m^G_{nm} - m^I_{nm}}{s^G_{nm} + s^I_{nm}} \tag{6}$$

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 [26]. We then define the visit-averaged quantities

$$\overline{m}^G = \frac{1}{NM}\sum^N\sum^M m^G_{nm} \quad \overline{s}^G = \frac{1}{NM}\sum^N\sum^M s^G_{nm} \tag{7}$$

$$\overline{m}^I = \frac{1}{NM}\sum^N\sum^M m^I_{nm} \quad \overline{s}^I = \frac{1}{NM}\sum^N\sum^M s^I_{nm} \tag{8}$$

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

$$Q^{(G)}_{nm} = \frac{m^G_{nm} - \overline{m}^I}{s^G_{nm} + \overline{s}^I} \qquad Q^{(I)}_{nm} = \frac{\overline{m}^G - m^I_{nm}}{\overline{s}^G + s^I_{nm}} \tag{9}$$

i.e. $Q^{(G)}_{nm}$ 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 $Q^{(I)}_{nm}$ 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

$$\frac{\Delta Q^{(I)}}{\Delta Q} = \sqrt{\frac{\mathrm{var}\left(Q^{(I)}_{nm}\right)}{\mathrm{var}\left(Q_{nm}\right)}} \tag{10}$$

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 IV.

In light of this observed variability in imposter scores, we chose to extend the original method of [21] to include the imposter matched delta term $\Delta\overline{s}^I_{nm}$ in the present work.

The discriminabilities of the devices with the lowest and one of the higher imposter contributions from Table IV were visually examined using box plots (Figs. 5a and 5b). (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. 6b), 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 non-linear processing of the raw match scores. In particular, whereas the box plots of Fig. 5a fit well to our assumption that the distributions change in their mean value rather than their shape, those of Fig. 5b show distinct limiting behaviour in the - processed - genuine distributions.

TABLE IV: 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

VI. 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 aging 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 aging 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 aging behavior into account – for example, by implementing an in-service template update procedure, or a requirement for periodic re-enrolment.

References

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

[2] Y. Gong, D. Zhang, P. Shi, and J. Yan, "High-speed multispectral iris capture system design," *IEEE Transactions on Instrumentation and Measurement*, vol. 61, no. 7, pp. 1966–1978, July 2012.

[3] 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*, vol. 66, no. 2, pp. 294–304, Feb 2017.

[4] 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)*, Nov 2013, pp. 1–4.

[5] "Deployment of biometric identification and electronic storage of data in eMRTDs," International Civil Aviation Organization, Doc 9303, (Seventh Ed.) 2015, Machine Readable Travel Documents.

[6] D. P. Sidlauskas and S. Tamer, *Handbook of Biometrics*, 2008, ch. Hand Geometry Recognition.

[7] 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.

[8] H. Cummins, "Finger prints and attempted fraud," *New Orleans Medical and Surgical Journal*, vol. 94, pp. 82–86, 1942.

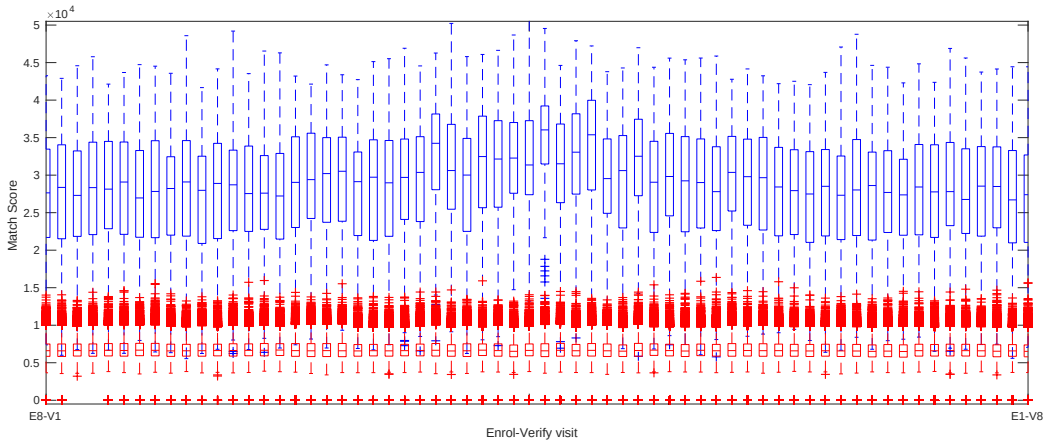
[9] A. Lanitis and N. Tsapatsoulis, "Quantitative evaluation of the effects of aging on biometric templates," *IET Computer Vision*, vol. 5, no. 6, pp. 338–347, November 2011.

[10] ——, "On the analysis of factors influencing the performance of facial age progression," in *2016 4th International Conference on Biometrics and Forensics (IWBF)*, March 2016, pp. 1–6.

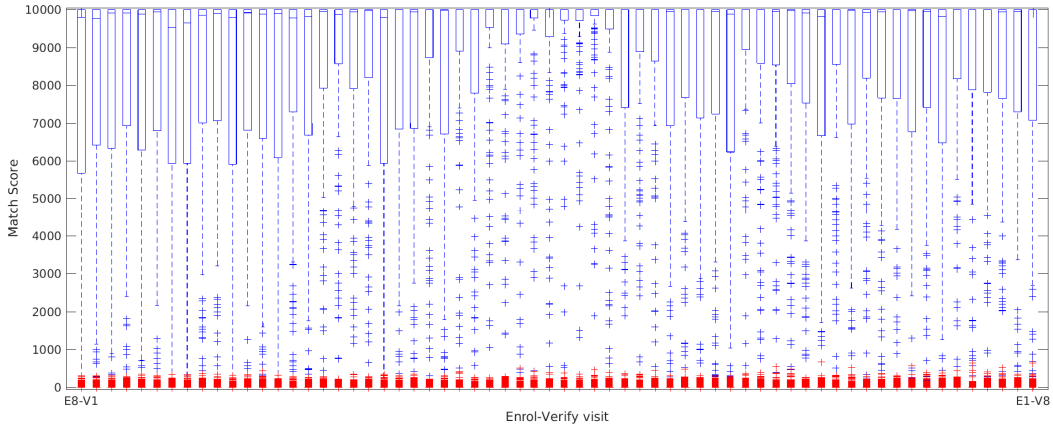
[11] G. Panis, A. Lanitis, N. Tsapatsoulis, and T. F. Cootes, "Overview of research on facial ageing using the fg-net ageing database," *IET Biometrics*, vol. 5, no. 2, pp. 37–46, 2016.

[12] 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)*, Sept 2016, pp. 1–6.

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

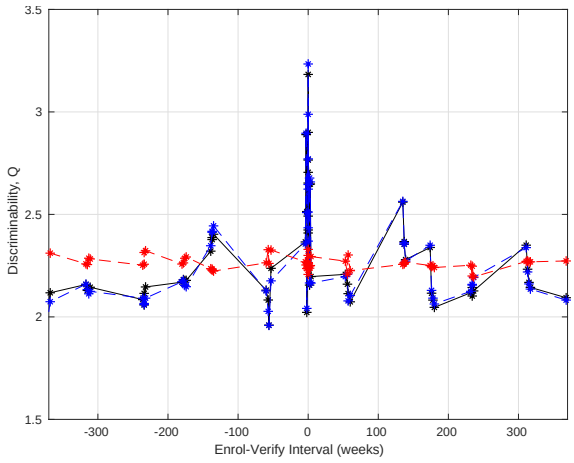


(a) Raw genuine (blue) and imposter (red) match scores: Device L.

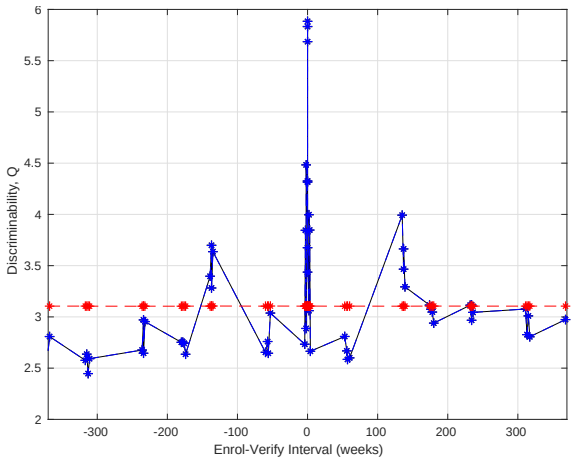


(b) Raw genuine (blue) and imposter (red) match scores: Device F.

Fig. 5: 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

Fig. 6: 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.

[14] 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)*, 2011, pp. 232–239.

[15] 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)*, Sept 2016, pp. 1–5.

[16] 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.

[17] E. Tabassi, C. L. Wilson, and C. I. Watson, "Fingerprint image quality," NIST, Nistir 7151, 2004.

[18] E. P. Kukula, M. J. Sutton, and S. J. Elliott, "The human–biometric-sensor interaction evaluation method: Biometric performance and usability measurements," *IEEE Transactions on Instrumentation and Measurement*, vol. 59, no. 4, pp. 784–791, April 2010.

[19] 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.

[20] M. E. Brockly, "The role of test administrator and error," Master's thesis, Purdue University, West Lafayette, IN, December 2013.

[21] J. Harvey, J. Campbell, S. Elliott, M. Brockly, and A. Adler, "Biometric permanence: Definition and robust calculation," in *IEEE Systems Conference 2017*, April 2017.

[22] M. Gamassi, M. Lazzaroni, M. Misino, V. Piuri, D. Sana, and F. Scotti, "Quality assessment of biometric systems: a comprehensive perspective based on accuracy and performance measurement," *IEEE Transactions on Instrumentation and Measurement*, vol. 54, no. 4, pp. 1489–1496, Aug 2005.

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

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

[25] C. Z. Mooney, R. D. Duval, and R. Duvall, *Bootstrapping: A nonparametric approach to statistical inference*. Sage, 1993, no. 94-95.

[26] 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)*, July 2012, pp. 1–4.