Characterization of Biometric Template Aging in a Multiyear, Multivendor 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 needs 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, biometric permanence, to characterize the stability of biometric features. First, we define permanence in terms of the change in a false nonmatch rate (FNMR) over a repeated sequence of enrollment and verification events for a given population. However, since changes in the FNMR are expected to be small, any variability in the biometric capture over time will camouflage the changes of interest. To address this issue, we propose a robust methodology 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 data set 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.**

***Index Terms*— Access control, aging, biometrics, fingerprint identification, measurement errors, measurement uncertainty, system performance.**

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 traits 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, 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 changes 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: for example, 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.

The age progression of biometric traits has perhaps received the most attention within the facial recognition modality. Lanitis and Tsapatsoulis [9] proposed a measure of biometric aging that they termed "aging impact," derived from the homogeneity and dispersion of a collection of templates. Although the primary focus of their work concerned on 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. In subsequent works, the focus was on 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 2-D and 3-D facial biometrics by comparing the genuine acceptance rate at a 0.1% false acceptance rate for short-term intervals (less than three months between enrollment and verification) versus long-term intervals (more than five years between enrollment 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 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 the pupillary diameter [15]. The influence of biometric sample quality on template aging was highlighted by Ryu *et al.* [16] who found that lower sample

Manuscript received December 08, 2017; revised July 08, 2018; accepted July 09, 2018. Date of publication August 29, 2018; date of current version March 8, 2019. This work was supported by the Natural Sciences and Engineering Research Council of Canada under Grant CRD 428240-11. The Associate Editor coordinating the review process was Dario Petri. *(Corresponding author: John Harvey.)*

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

J. Campbell is with Bion Biometrics Ltd., Nepean, ON K2J 5H1, Canada.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIM.2018.2861998

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 the development of recalibration strategies: in the case of a biometric IDMS, it might take the form of a periodic re-enrollment requirement. The chief difficulty in evaluating the 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 the degradation of the particular biometric capture hardware—for example, in the case of the fingerprint modality, scratching or marring of sensor platens might be included. 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 a current presentation and a gallery of previously enrolled templates, each comparison is affected by two systematic terms, which we refer to as *visit biases*.

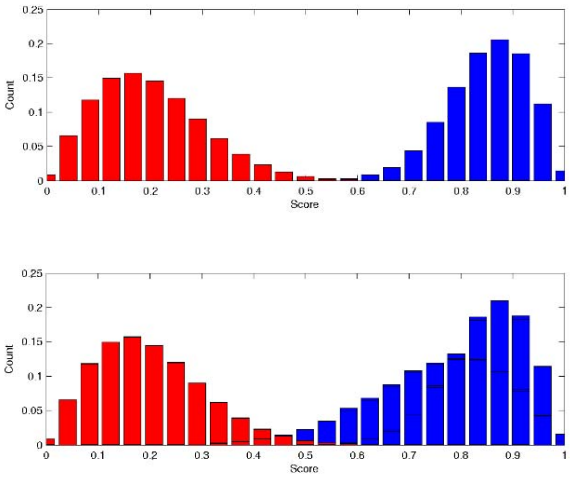
The goal of this paper is to characterize template aging in the fingerprint modality, for a number of commercially available fingerprint sensor devices and technologies, and to understand its impact on the deployment and operation of fingerprint-based IDMSs. First, in Section II, we outline the definition and properties of our metric, biometric permanence, P_B; in Section III, we describe the design of our study, including subject demographics and data collection 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

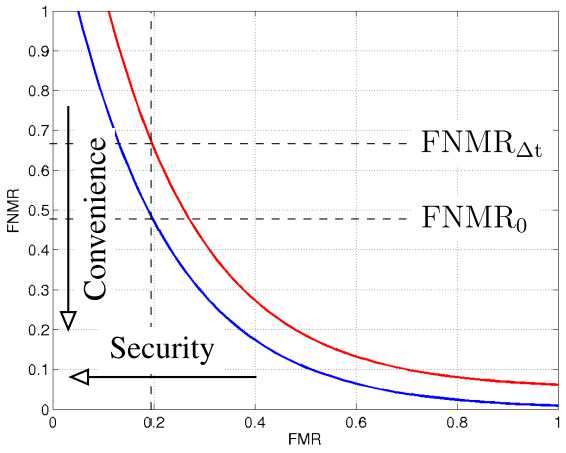
Here, we expand on [21] in which we proposed a measure called *biometric permanence*, $P_B(\Delta t)$, at a given elapsed time Δt, as follows:

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

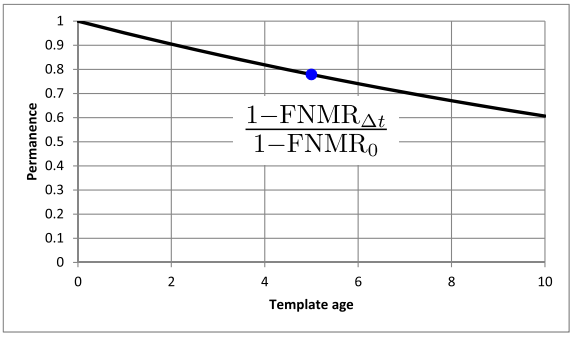
based on the change in the false nonmatch rate (FNMR) at a given false match rate (FMR) [22]. The definition was motivated by operational considerations, that is, template aging will manifest itself as a decrease in the security and/or convenience provided by the biometric system at a given operating point. Since it is usually operational security that is of primary concern, it is natural to fix the FMR and consider the change in the FNMR. A schematic overview of the development, starting from the empirical match scores, is shown in Fig. 1. At time zero, we enroll 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 provide a collection of labeled (i.e., genuine or imposter) biometric match scores whose distributions may or may not be completely separable at some decision threshold θ. Later some time, new



(a)



(b)



(c)

Fig. 1. Overview of the method. (a) Empirical match score distributions immediately after enrollment (top) and after some time interval (bottom). (b) Change in classification accuracy represented on a DET curve: arrows indicate the directions of increasing security and convenience. (c) Permanence P_B derived from the change in the FNMR at a fixed FMR according to (1).

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 the FNMR is expressed as a permanence value for the enrollment-verification time interval.

As well as reflecting the underlying performance degradation mechanism, other desirable features of this formulation are as follows.

1) Permanence P_B increases toward unity as the $\text{FNMR}_{\Delta t}$ tends toward FNMR_0; this case would correspond to a perfectly permanent template.
2) Permanence P_B decreases toward zero as the $\text{FNMR}_{\Delta t}$ increases toward unity; a biometric template might be said to be completely impermanent at this point.

In the pathological case where $\text{FNMR}_{\Delta t} < \text{FNMR}_0$, 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 this paper, 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 (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 Section V.

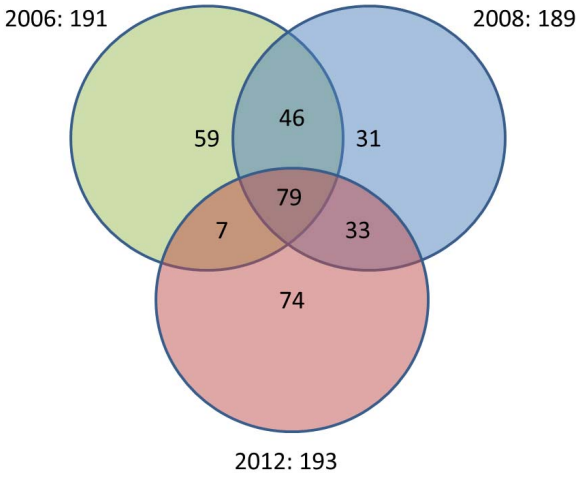


Fig. 2. Overlap of participants between data collection phases (the 2013 collection is omitted for clarity; it overlaps almost completely with 2012).

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

III. STUDY PROTOCOL AND DEMOGRAPHICS

In order to detect biometric template aging and to evaluate our methodology, we need a data set of similarity scores evaluated for the same subjects at different enrollment-verification time intervals. Ideally, the biometric collection should take place under well-controlled conditions with a consistent protocol, in order to control (as far as possible) 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 (see Fig. 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 enrollment were right and left index in the first instance; however, if either of these was unavailable (or failed to enroll), 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's finger were captured during each enrollment and further six images (in two distinct three-presentation verification attempts) were captured per enrolled finger during each verification such that a typical visit resulted in 18 single-finger images per subject per device. In each subject visit, the order in which devices were presented for both enrollment and verification was randomized under software control in order to counterbalance the subject and operator acclimation.

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

Twelve different commercially available fingerprint sensor devices were obtained, representing multiple vendors and technologies: single-spectral optical, multi-spectral optical, and capacitive (see 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 the best of our knowledge, all of the optical sensors are based on frustrated total internal reflection. Age of the participants at the time of the most recent collection ranged from 15 to 70 years. In excess of 15 000 ISO/IEC standards-compliant two-finger biometric enrollment 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

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

transactions, with approximately 900 000 genuine (same subject, same finger) matches (see Table II).

IV. ANALYSIS

A. Methodology

As in [21], we seek to evaluate biometric permanence $P_B(\Delta t)$ according to (1), where the FMR and the FNMR are the false match and FNMRs 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 enrollment 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 discussion, 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 enrollment, as described in [23] and [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} enrollment 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:

1) 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 the enrollment visit m and the verification visit n;
2) 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 enrollment 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

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

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

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

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 root-mean-square (rms) basis such that

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

leaving the SNR 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 enrollment 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 (3) are averaged again across a particular pair of enrollment and verification visits m, n to provide 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 (1) at the 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 $(\overline{s}_{kk}^{ii}, \overline{s}_{kk}^{j\neq i})$ 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, respectively. The aggregate vector is then resampled, with replacement, $n_B = 1000$ times with sampling weights inversely proportional to the class size in order to remove the class imbalance.

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

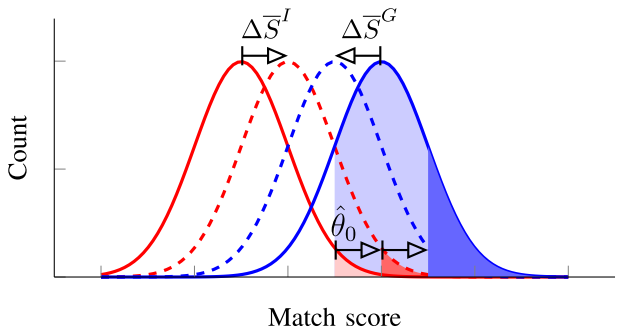


Fig. 3. A shift in the mean imposter score results in a shift in the estimated decision threshold $\hat{\theta}_0$ for a specified FMR (red area)—a corresponding change in the achievable TMR (blue area) for the shifted genuine scores.

chosen FMR, we just need to evaluate $1 - \text{FNMR}$ (or, equivalently, the true match rate TMR) at a set of thresholds $\theta_{nm} = \hat{\theta}_0 + \Delta\overline{s}^I_{nm} - \Delta\overline{s}^G_{nm}$. 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

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

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

C. Results

Representative results of this procedure are shown graphically in Fig. 4 with a 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. 4(a) and (d) 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 DET curves in Fig. 4(b) and (e) 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 Fig. 4(c) and (f), we see the permanence P_B according to (1) decreases 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 $\Delta t = 0$ [see Fig. 4(g)] 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 $\text{FNMR} = 0$ at $\text{FNMR} = 0$ [see Fig. 4(h)] and correspondingly no discernable change in permanence P_B in Fig. 4(i).

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

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

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 remain 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 $\Delta t_{mn} = 0$ become increasingly difficult to classify. 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.

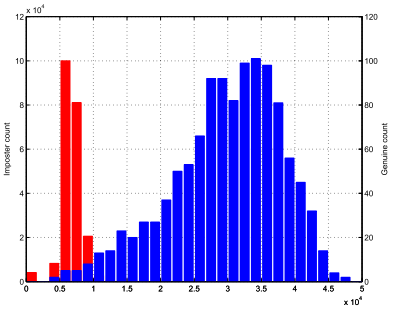
For the two devices that showed no change in permanence, the analysis is likely affected by the large-class imbalance inherent in such biometric comparisons. That is, for a data set of K distinct fingers, there are about K^2 imposter matches with only K genuine matches, which causes the tails of the 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 the corresponding FNMR for the permanence calculation is evaluated. While the bootstrapping procedure described in IV-C attempts to ameliorate this effect, if the empirical distributions are separable, then no amount of resampling can guarantee that there will be a nonzero 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 that 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.

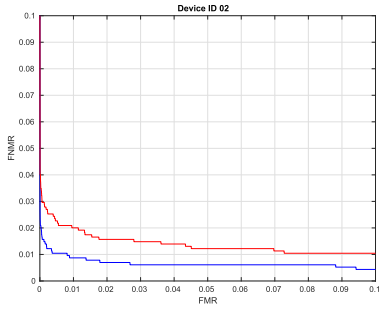
In the remainder of this section, we discuss some other aspects of the data and their potential impact upon the interpretation of our results.

A. Time Symmetry of the Match Scores

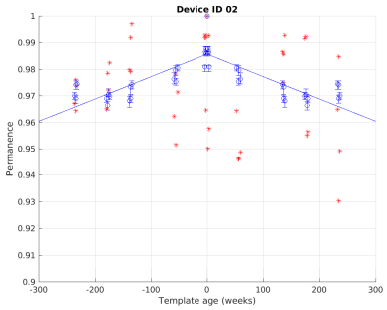
A key assumption that allows us to substantially remove the visit-to-visit bias factors is that the underlying "true" match scores are time-symmetric: that is, in the absence of these factors, comparisons between a biometric enrollment obtained at time t_1 and a set of verification presentations at later time t_2, and between a biometric enrollment 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



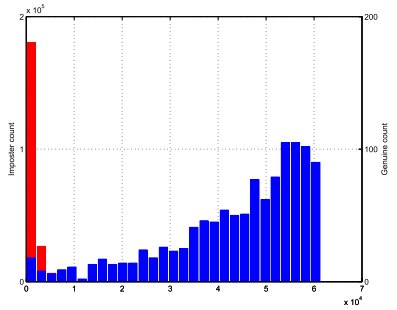
(a) Genuine (blue) and Imposter (red) score counts at $\Delta t = 0$



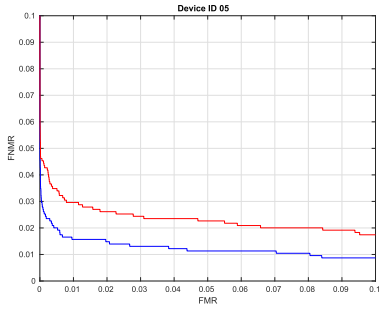
(b) Decision error tradeoff (DET) curve at $\Delta t = 0$ (blue) and at $\Delta t = 373$ weeks (red)



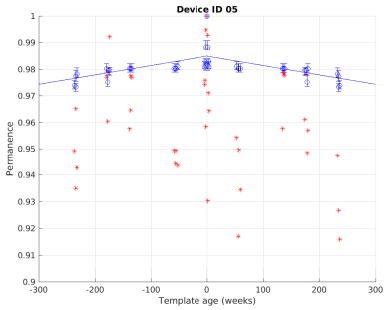
(c) Permanence versus template age: "naive" calculation (red); present method (blue)



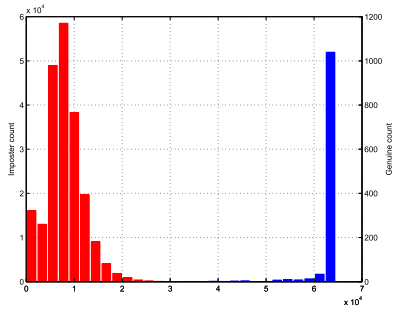
(d) Genuine (blue) and Imposter (red) score counts at $\Delta t = 0$



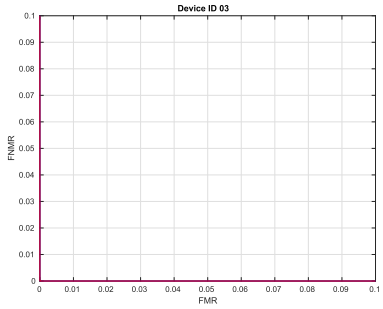
(e) Decision error tradeoff (DET) curve at $\Delta t = 0$ (blue) and at $\Delta t = 373$ weeks (red)



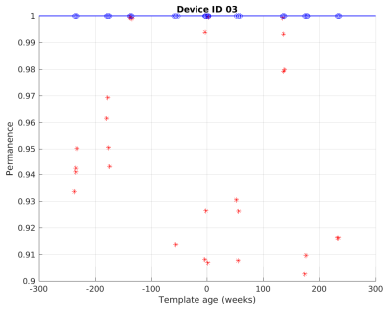
(f) Permanence versus template age: "naive" calculation (red); present method (blue)



(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. The 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.

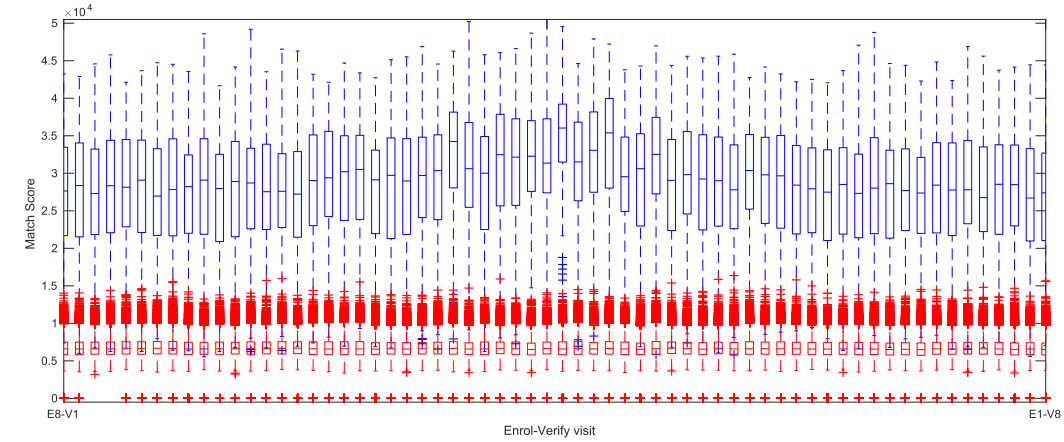
by the W^{ij} terms in our formalism.) The extent to which this is the case will depend on the similarity measure used: we might expect a simple degree-of-overlap measure to be time-symmetric, whereas a more heuristic measure might not be. For example, consider the case in which the number of extractable 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 this paper.

B. Constancy of the Imposter Distributions

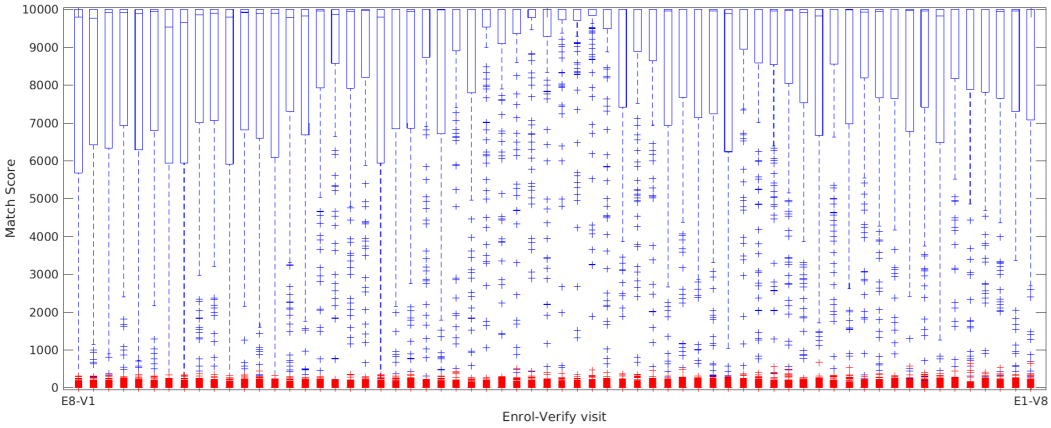
Intuitively, we might expect the imposter score distribution to be relatively insensitive to the template age, since factors that decrease the similarity between any given pair of subject-fingers may increase the similarity between other such imposter pairs.[1] However, this does not allow for gross 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 the changes of a 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

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



(a)



(b)

Fig. 5. Box plots of the raw match scores between enroll visit E_m and verify visit V_n. The boxes are plotted from most negative to most positive template age, i.e., from "enroll 8—verify 1" to "enroll 1—verify 8." Maximum discriminability occurs around the center of the chart–corresponding to the template ages close to zero. (a) Raw genuine (blue) and imposter (red) match scores: Device L. (b) Raw genuine (blue) and imposter (red) match scores: Device F.

of visits n, m as the ratio of the difference in sample mean score between the 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; 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

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

that is, $Q^{(G)}_{nm}$ 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 $Q^{(I)}_{nm}$ is the corresponding discriminability with constant genuine mean and standard deviations. Finally, we evaluate the fractional contribution of the imposter scores to the rms 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 the 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 [see Fig. 5(a) and (b)]. (The device with the very highest imposter contribution, Device H at 26.45%, was not chosen since its data were only available

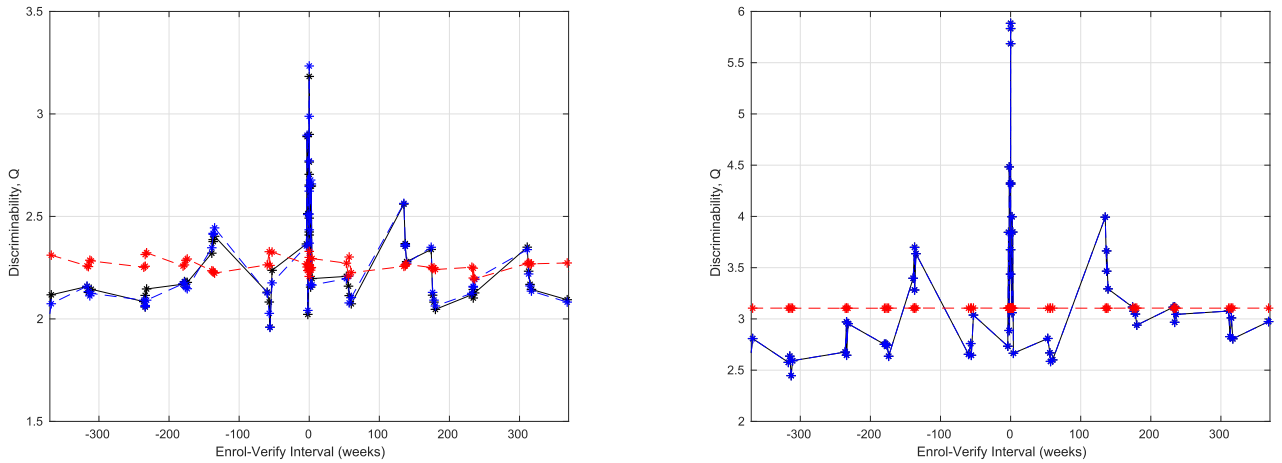


Fig. 6. Binary discriminability Q as a function of the template age in weeks. Total discriminability is shown in black; the contributions Q_G (blue) and Q_I (red) are due to changes in the genuine and imposter distributions, respectively. Variation of the imposter distribution contributes nonnegligibly to the discriminability in the Device L but is negligible in the case of the Device F. (a) Device L. (b) Device F.

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

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 the Device F [see Fig. 6(b)], 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 nonlinear processing of the raw match scores. In particular, whereas the box plots of Fig. 5(a) fit well to our assumption that the distributions change in their mean value rather than their shape, those of Fig. 5(b) show distinct limiting behavior in the—processed—genuine distributions.

VI. CONCLUSION

We have elaborated a method to isolate and measure changes in the biometric system performance over time, using a metric which we call biometric permanence. The method was applied to a data set 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 the 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-enrollment.

REFERENCES

[1] D. Maltoni, D. Maio, A. Jain, and S. Prabhakar, *Handbook of Fingerprint Recognition*. Berlin, Germany: Springer, 2009.
[2] Y. Gong, D. Zhang, P. Shi, and J. Yan, "High-speed multispectral iris capture system design," *IEEE Trans. Instrum. Meas.*, vol. 61, no. 7, pp. 1966–1978, Jul. 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 Trans. Instrum. Meas.*, 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 *Proc. IEEE Int. Conf. Smart Instrum., Meas. Appl. (ICSIMA)*, Nov. 2013, pp. 1–4.
[5] *Deployment of Biometric Identification and Electronic Storage of Data in eMRTDs*, 7th ed., document 9303, International Civil Aviation Organization, 2015.
[6] D. P. Sidlauskas and S. Tamer, "Hand geometry recognition," in *Handbook of Biometrics*. Berlin, Germany: Springer, 2008.
[7] M. Drahansky, M. Dolezel, J. Urbanek, E. Brezinova, and T. H. Kim, "Influence of skin diseases on fingerprint recognition," *J. Biomed. Biotechnol.*, vol. 2012, Feb. 2012, Art. no. 626148. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3359776/
[8] H. Cummins, "Finger prints and attempted fraud," *New Orleans Med. Surg. J.*, vol. 94, pp. 82–86, 1942.
[9] A. Lanitis and N. Tsapatsoulis, "Quantitative evaluation of the effects of aging on biometric templates," *IET Comput. Vis.*, vol. 5, no. 6, pp. 338–347, Nov. 2011.
[10] A. Lanitis and N. Tsapatsoulis, "On the analysis of factors influencing the performance of facial age progression," in *Proc. 4th Int. Conf. Biometrics Forensics (IWBF)*, Mar. 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, May 2016.
[12] I. Manjani, H. Sumerkan, P. J. Flynn, and K. W. Bowyer, "Template aging in 3D and 2D face recognition," in *Proc. IEEE 8th Int. Conf. Biometrics Theory, Appl. Syst. (BTAS)*, Sep. 2016, pp. 1–6.
[13] U. Uludag, A. Ross, and A. Jain, "Biometric template selection and update: A case study in fingerprints," *Pattern Recognit.*, vol. 37, no. 7, pp. 1533–1542, 2004.
[14] S. P. Fenker and K. W. Bowyer, "Experimental evidence of a template aging effect in iris biometrics," in *Proc. IEEE Workshop Appl. Comput. Vis. (WACV)*, Jan. 2011, pp. 232–239.

[15] H. Hofbauer, I. Tomeo-Reyes, and A. Uhl, "Isolating iris template ageing in a semi-controlled environment," in *Proc. Int. Conf. Biometrics Special Interest Group (BIOSIG)*, Sep. 2016, pp. 1–5.
[16] J. Ryu, J. Jang, and H. Kim, "Analysis of effect of fingerprint sample quality in template aging," in *Proc. 2nd NIST Biometric Qual. Workshop*, Nov. 2007, pp. 7–8.
[17] E. Tabassi, C. L. Wilson, and C. I. Watson, "Fingerprint image quality," NIST, Gaithersburg, MD, USA, Tech. Rep. 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 Trans. Instrum. Meas.*, vol. 59, no. 4, pp. 784–791, Apr. 2010.
[19] S. Elliott, E. Kukula, and N. Sickler, "The challenges of the environment and the human/biometric device," in *Proc. Int. Workshop Biometric Technol.*, 2004, pp. 38–44.
[20] M. E. Brockly, "The role of test administrator and error," M.S. thesis, Technol. Leadership Innov., Purdue Univ., West Lafayette, IN, USA, Dec. 2013.
[21] J. Harvey, J. Campbell, S. Elliott, M. Brockly, and A. Adler, "Biometric permanence: Definition and robust calculation," in *Proc. Annu. IEEE Int. Syst. Conf.*, Apr. 2017, pp. 1–7.
[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 Trans. Instrum. Meas.*, vol. 54, no. 4, pp. 1489–1496, Aug. 2005.
[23] *ILO Seafarers' Identity Documents Biometric Testing Campaign Report Part I*, Int. Labour Org., Geneva, Switzerland, 2004, p. 185.
[24] *The Standard for the Biometric Template Required by the Convention*, Int. Labour Org., Geneva, Switzerland, 2003, p. 185.
[25] C. Z. Mooney, R. D. Duval, and R. Duvall, *Bootstrapping: A Nonparametric Approach to Statistical Inference*, New York, NY, USA: SAGE, 1993, nos. 94–95.
[26] W. Freude *et al.*, "Quality metrics for optical signals: Eye diagram, Q-factor, OSNR, EVM and BER," in *Proc. 14th Int. Conf. Transparent Opt. Netw. (ICTON)*, Jul. 2012, pp. 1–4.

John Harvey (S'16) received the bachelor's degree in electrical and information science from the University of Cambridge, Cambridge, U.K. He is currently pursuing the Ph.D. degree with the Systems and Computer Engineering Department, Carleton University, Ottawa, ON, Canada.

Following his bachelor's degree, he worked in telecoms R&D specializing in nonlinear fiber optics and error correcting codes. His current research interests include biometric information and security.

John Campbell has over 15 years of experience in biometric technologies. He has worked for the Canadian Department of National Defence (DND) and as a Consultant with Transport Canada, Public Safety and CBSA. He has provided consulting and advice on biometric systems to international organizations such as the International Labour Organization, the International Organization for Migration, and the Secretariat of the Pacific Community, and to multiple government agencies in Canada, USA, U.K., New Zealand, Russia, Indonesia, Malaysia, and others.

Andy Adler (SM'15) received the B.A.Sc. degree (Hons.) in engineering physics from The University of British Columbia, Vancouver, BC, Canada, in 1990, and the Ph.D. degree in biomedical engineering from the École Polytechnique de Montréal, Montreal, QC, Canada, in 1995.

He has taught and researched at the University of Ottawa, Ottawa, ON, Canada, and worked in senior technology positions at BioDentity (now cryptometrics), AiT (now 3M), and DEW Engineering (now ActivCard), Ottawa, ON, Canada; and at CIL Explosives (now Orica), Montreal. He also held post-doctoral positions at McGill University and the University of Colorado Health Sciences Center, Denver, CO, USA. He is currently a Canada Research Professor in biomedical engineering with the Department of Systems and Computer Engineering, Carleton University, Ottawa. His current research interests include biomedical measurement and robust data analysis.