IET Research Journals

Research Article

IET Journals
The Institution of Engineering and Technology

ISSN 1751-8644
doi: 0000000000
www.ietdl.org

Analysis of the effect of aging, age, and other factors on iris recognition performance using NEXUS scores dataset

Dmitry O. Gorodnichy[1]*, *Michael P. Chumakov*[2]

[1] *Science and Engineering Directorate, Canada Border Services Agency, Ottawa, Canada*
[2] *Business Application Services Directorate, Canada Border Services Agency, Ottawa, Canada*
* *E-mail: dmitry.gorodnichy@cbsa-asfc.gc.ca (Corresponding author)*

Abstract: The historical NEXUS iris kiosks log dataset collected by the CBSA from 2003 till 2014 has become the focus of scientific attention due to its involvement in the iris aging debate between NIST and University of Notre Dame researchers. To facilitate this debate, this paper provides additional details on how this dataset was collected, its various properties and irregularities, and presents new results related to the effect of aging, age, and other factors on the system performance obtained using the portions of the dataset that have not been previously analyzed. In doing that, the importance of conducting subject-based performance analysis, as opposed to the traditionally done transaction-based analysis, is emphasized. The significance of factor effects is examined. Recommendations on further improvement of the technology are made.

1 Introduction

The biometric kiosks deployed by the Canada Border Services Agency (CBSA) since 2003 for the NEXUS trusted traveller program [1] present one of the longest deployed iris recognition technologies in automated border control to date. The performance log collected from these kiosks provides scientists and developers a unique source of information that can be used to better understand and improve iris technology.

In 2012 a portion of anonymized NEXUS kiosk log data was shared with the NIST scientists for the IREX VI iris aging study, where it was labeled the OPS-XING dataset. The results of this study, which were published in 2013 [3], brought a lot of attention from the scientific community, actively discussed and contested by the scientists from the University of Notre Dame [4]-[10].

One of the key reasons behind the arguments of invalidity of the results obtained by the NIST scientists on the OPS-XING dataset lies in the fact that besides aging, which was the main factor under investigation, the study looked into only one additional factor affecting the system performance – dilation. The log data related to other factors were not made available to the NIST scientists.

Another reason is the fact that the dataset was obtained from the operational system, the full operation of which is not entirely known to external organizations. The dataset contained a number of irregularities – due to human and machine errors, which were not known to the investigators. Full explanation of how the system worked and its performance objectives was not provided.

Finally, the evaluation methodology that was applied in the IREX VI study for analyzing the effect of aging was also put into question. The effect of habituation, even though being admitted by NIST scientists, was not really taken into account.

This paper addresses these three limitations of the IREX VI study. Detailed description of the system operation (Section 2) and dataset irregularities (Section 3) is provided. An alternative methodology based on the use of subject-based metrics, instead of conventionally used transaction-based metrics used in the previous studies, is described and is shown to be more appropriate for the application (Section 4). Finally, new results on the effect of age, aging and other factors, based on the new methodology and the previously unused portions of the dataset are presented (Sections 5 and 6). Recommendations for the improvement of iris recognition performance based on the obtained results conclude the paper.

2 NEXUS system description

The CBSA commenced using iris recognition technology for automated authentication of travellers in airports in 2003, following the launch of a similar iris-enabled registered traveller program in the United Kingdom (UK). First, it was used for CANPASS-Air [2], which is a Canadian program that provides pre-enrolled pre-cleared Canadians expedited passage at arrival in airports for flights within Canada. Later in 2004 the use of iris-enabled identification of travellers was extended to the NEXUS-Air, which is a bi-national, Canada-US program for pre-approved low-risk travellers flying between Canada and the USA [1].

The expedited passage allows NEXUS members to proceed directly to the NEXUS self-serve kiosks, bypassing lengthy queues and interaction with customs border protection (CBP) officers and border services officers (BSO). All kiosks are located in Canadian airports, owned and controlled by the CBSA, with iris biometric data being collected and stored by the CBSA. Kiosks used for travellers arriving to Canada are located in Primary Inspection Area. Kiosks used for travellers leaving Canada to the US are located in a special dedicated lane of the US pre-clearance area. In total, 69 NEXUS kiosks have been installed in Canada in eight Canadian airports: Calgary, Edmonton, Halifax, Montreal, Ottawa, two terminals at Toronto Pearson International Airport, Toronto Billy Bishop (Toronto City Airport), Vancouver and Winnipeg. Of these, 8 kiosks are used in Enrollment Centres and 22 kiosks are used at the US pre-clearance. The same kiosks and iris database are used for both NEXUS-Air and CANPASS-Air programs. The number of CANPASS-Air users (about 2,000 people by 2014) however is significantly less than that of NEXUS-Air (over half a million in 2014).

Two designs (shown in Figure 1) were used for the NEXUS kiosks of the first generation NEXUS system that were deployed from 2003 till 2014, the log of which comprises the OPS-XING dataset: with one-eye LG camera (deployed in 2003) and two-eye Panasonic camera (deployed in 2007).

2.1 System decision logic

At the Enrollment stage, both irises of a traveller are photographed. Image Quality (IQ) control on iris images is performed. Only if their IQ metric is high, will they be enrolled into the system database.

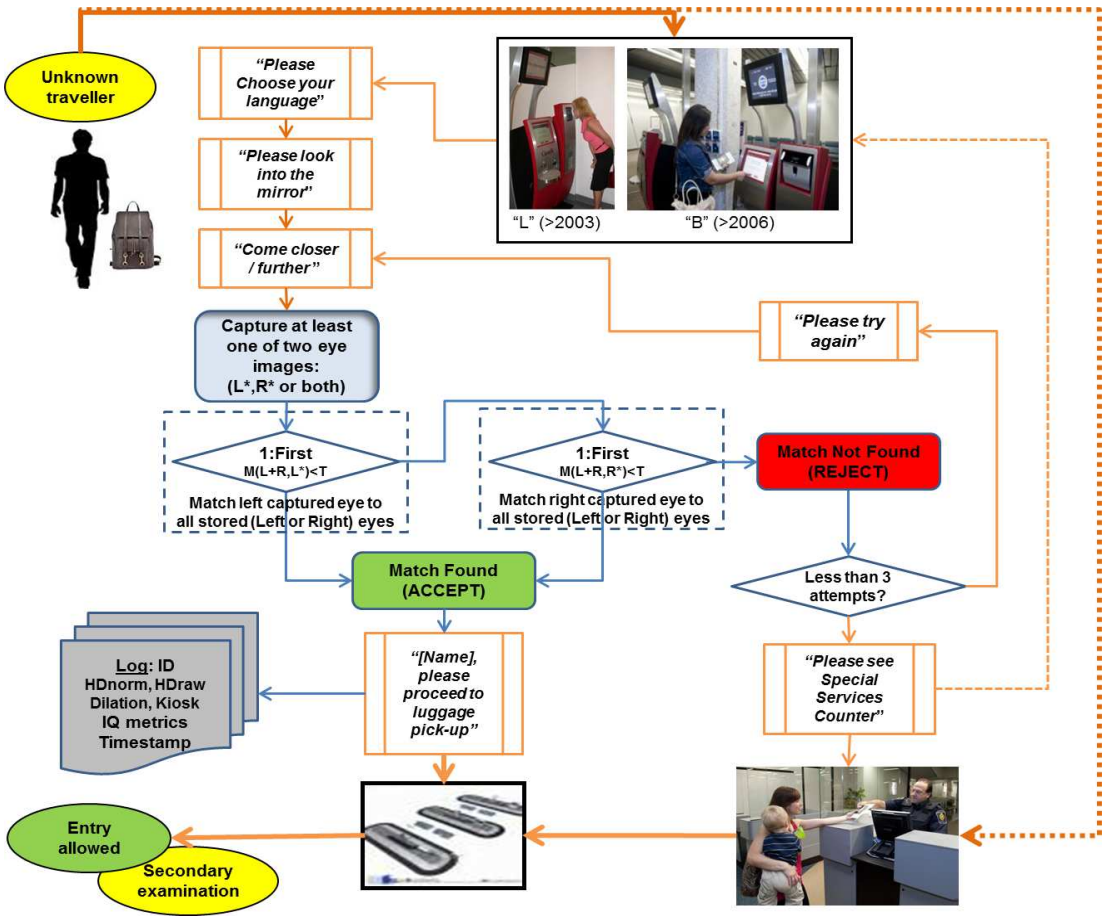


Fig. 1: The workflow and decision logic of the NEXUS kiosks of the first generation, the log of which was used in NIST IREX IV iris aging study. The system decision steps for match and rejection are shown in dark blue arrows. The user's procedural steps are shown in light orange arrows. Dashed orange arrows indicate optional steps for the users.

Because of image quality control, in some cases only one eye can be enrolled, and in some rare cases none of the eyes can be enrolled. Travellers have also a choice of opting out from enrolling their iris images. For travellers enrolling the iris, instructions are provided on how to use the kiosks, among which is the recommendation to remove eye-glasses and contact lenses of any type. However, it is not known how closely these recommendations are followed.

At the time of crossing the border, referred to as the Passage stage, the system is configured to search for the identity of the captured eyes using a 1-to-First search using the decision tree shown in Figure 1. Once the system captures images of a person's eyes, it tries to authenticate a person using the Left eye only. If the Left eye is not matched, the Right eye is used. In both cases, the match is performed against all images (i.e., both left and right images) stored in the database until the first image with a matching score below the threshold is found. This is due to the fact that first generation of NEXUS kiosks used single-eye iris cameras, which captured an eye of person without knowing whether it was a left or right eye.

When a traveller is rejected by the system (which happens because of one of two reasons: either IQ of live image is poor, or no matching image is found in the database), s/he is asked to try again, with the total of three attempts allowed in a single passage session with the kiosk. When a traveller is accepted, her/his attempt number at a given session is recorded.

A passage session ends either because of traveller's inactivity or the maximum number of capture attempts is reached, after which the system resets into the initial state with the "Welcome. Please choose your language" message. Travellers who are not recognized within a single session receive the "Please visit Special Services Counter" message. At the same time, they are also allowed to initiate additional passage session using the same or different kiosks, which they can do as many times as they want. Similarly, they are also allowed to proceed to Special Services Counter any time they experience a problem with the kiosk.

It is possible that some travellers, particularly those who experienced rejection problems in the past, have proceeded directly to the Special Services Counter without initiating a single session with the kiosks. There is no data left in the system log about these travellers. The data from travellers who used the system but were rejected was also not logged. This presents a critical limitation of the OPS-XING dataset made from historical NEXUS log data. By the design, this dataset is biased towards better performing users, as it contains mainly the data from travellers who did not experience problems with the system and does not contain any rejected transactions. Nevertheless, even with this limitation, this dataset presents a unique and very valuable source for investigation of iris biometrics properties and limitations, specifically related to age and aging, which becomes particularly important now with iris modality becoming increasingly used in many government and United Nations programs [20, 21] and the ongoing debate related to the tolerance of iris biometrics to aging [4]-[19].

2.2 Iris recognition algorithm: Matching formula and threshold

NEXUS kiosks use Daugman's original iris recognition algorithm [22, 23]. The same version of the algorithm is used throughout the entire life-cycle of the system. Since its deployment in NEXUS system, iris technology has improved [24, 25], including more precise pupil and iris circles interpolation, better masking bits for occluding parts of the iris region due to eyelashes, specular reflections, boundary artifacts of hard contact lenses, and the use of both real and imaginary bits of the iris code. To our understanding, however,

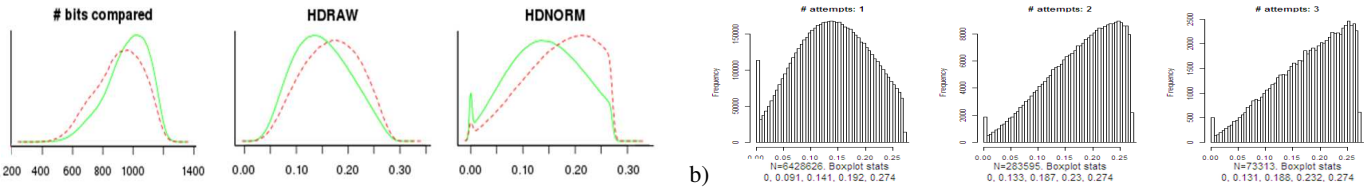


Fig. 2: Distribution of Number of Bits Compared, HDRAW and HDNORM scores in the OPS-XING dataset.
a Left-eye scores (solid green) vs. right eye (dashed red) scores;
b Histograms of HDNORM scores for different number of attempts. Minimum, 25%, 50%, 75%, quartile and maximum values are shown at the bottom of each histogram.

these later improvements of the algorithm are not implemented in the version that was used in the collection of the OPS-XING data.

Iris images are compared using the Hamming Distance (HD), which is a dissimilarity score between the corresponding iris templates (IrisCodes). The score $HD = 0$ means perfect match. A high score (i.e., $HD > T_{HD}$) results in reject. The value of the threshold T_{HD} is automatically selected by the algorithm based on the theoretical prediction of the False Accept Rate for a given number of entries in the data-base, slightly decreasing every year as the number of enrolled NEXUS members grew: from 0.282672 in 2006 (when the logging of the system commenced) to 0.271534 in 2014 (when the logging finished).

The Hamming Distance is computed in two steps. First, the raw Hamming Distance $HDRAW$ is computed as the fraction of bits that disagree between two irises. Then, the normalized Hamming Distance $HDNORM$ is computed from $HDRAW$ following the normalization rule that gives less weight to comparisons performed on heavily occluded irises, using the following formula:

$$HDNORM = 0.5 - (0.5 - HDRAW)\sqrt{\frac{Nbits}{< Nbits >}}, \quad (1)$$

where $Nbits$ is the number of bits used in comparison and where $< Nbits >$ is a vendor defined constant equal to 911, which, according to the original algorithm [23], represents the average number of bits compared.

Figure 2-a shows $HDNORM$, $HDRAW$ and $Nbits$ score distributions in the OPS-XING dataset. It is noted that, in contrast to $HDNORM$ scores distributions, the $HDRAW$ scores distributions have much less visible artifacts due to score truncation and censoring, and are unimodal (i.e., have only one maximum). This makes analysis of $HDRAW$ scores using statistical techniques easier.

We also note that the actual average value of $Nbits$ is 954, which is higher than $< Nbits >$ constant used in the normalization formula (Eq. 1).

2.2.1 Observation related to score normalization : Through our analysis, the value of using the normalization step (1) for the NEXUS application has been questioned in a number of ways. Besides producing non-unimodally distributed values (seen in Figure 2-a), which complicates modeling the system performance using statistical methods, it also contributes to higher false reject rates for travellers with occluded iris.

A number of ways are seen to further improve the matching formula for the application. This includes post-processing score normalization described in [26], the use of conditional normalization formula (conditioned on additional image quality metrics such as contrast and/or person's age), which are analyzed further in the paper, or not applying the normalization formula (Eq. 1) at all. These however are outside of the scope of this paper.

In this paper it is the importance of analyzing $HDRAW$ in combination with image quality metrics, as opposed to analyzing $HDNORM$ scores only as done in the past, that is emphasized.

2.2.2 Observation related to the correlation between matching score and number of attempts : In our analysis, in addition to the matching score ($HDRAW$ and $HDNORM$), we also use the number of recorded attempts ($\#Attempts$) as one of the important kiosk performance metrics. There exists a subtle relationship between the two, illustrated in Figure 2, which shows the distribution of matching scores for different number of attempts and the corresponding five-number statistics for $HDNORM$.

On one hand, it is seen that the larger the number of attempts, the larger (worse) the matching scores, as reported in [8]. On the other hand, a higher matching score does not necessarily mean that a person gets rejected (as long as the matching score is less than the threshold, a person is accepted). Similarly, recognition from a single attempt does not necessarily mean that a person has not tried and was already rejected multiple times during other sessions that were not logged. Therefore, using both metrics in the analysis provides richer complementary evidence for the results obtained.

3 OPS-XING dataset

The OPS-XING dataset, a part of which was used in the IREX VI evaluation by NIST [3, 5] and the evaluations conducted by UND [6, 7, 10], consists of over a quarter billion of matching and image quality metrics that were recorded during Enrollment and Passage transactions by the NEXUS system. These metrics are listed in the Table 1. The metrics that were shared with NIST and UND and used in previous research [3]-[10] are marked bold.

In total, there were 1,370,890 enrollment transactions (recorded from September 2003 till May 2014, from 705,553 travellers – most (662,220) done with dual-eye Panasonic cameras deployed in 2007, others done by single-eye LG cameras) and over 10 million passage transactions (recorded from October 2007 till May 2014, from

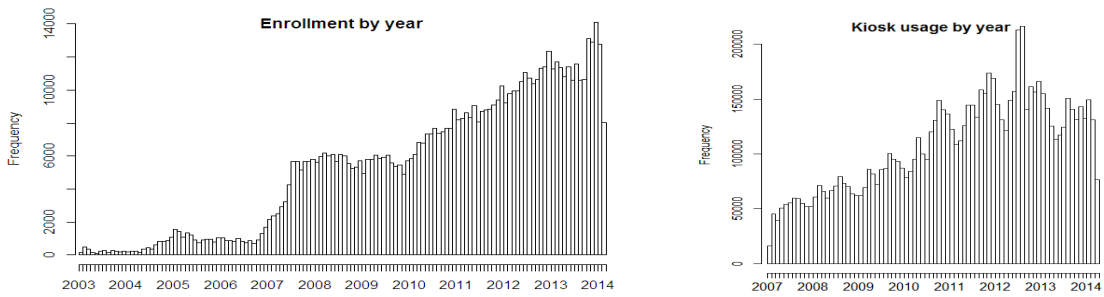


Fig. 3: Number of Enrollment (left) and Passage (right) transactions per month.

Table 1 Metrics recorded in the OPS-XING dataset.

At Enrollment	**FAKE_ID**, age, **EYE** (L-left or R-right), **CAMERA** ('L' for old LG camera, 'B' for new Panasonic camera) ENROLLMENT_DATE (month, **year**, time of the day) IQ metrics: related to localization accuracy — iris center x, iris center y; iris radius; pupil center x, pupil center y; related to dilation — pupil radius, pupil iris ratio (the same as **DILATION**); related to image contrast — iris sclera contrast, iris pupil contrast, average iris intensity, iris texture energy; related to occlusion — iris area, number of bits encoded
At Passage	**FAKE_ID**, **EYE used**, TRANSACTION_DATE (**month**, **year**, time of the day) **ELAPSED_TIME** (the number of days since enrollment) **HDNORM**, HDRAW **CAPTURE_NUMBER_WITHIN_PA** (capture-and-recognize attempts) **FAKE_KIOSK_ID, THD** (matching threshold), **MATCHING_MODE** (SEP: two-eye pilot, SEM: regular one-eye operation), IQ metrics: same as at Enrollment, number of bits encoded, number of bits compared

Note: metrics used in the previous work [3]-[10] are marked bold. The distributions of metric values are shown in Figure 5.

467,314 travellers – all done with dual-eye Panasonic cameras). Distribution of Enrollment and Passage transactions over the years is shown in Figure 3. Seasonal patterns in Passage data can be noticed.

3.1 Aberrations in data

The OPS-XING dataset contains a number of abnormal entries that are not described by the system logic. Mostly caused by human error or temporal experimentation with the system (either by kiosk users or programmers), such aberrations in data may give rise to additional challenges in understanding the technology and arriving to the correct conclusions by external researchers who process this historical dataset. These data aberrations are described below. They needed to be removed or taken into account prior to conducting the analysis.

HD scores higher than threshold: There are 351 passage transactions in two Kiosks that happened with Right eye which have $HDNORM > THD$. These are from the Pilot that was conducted in 2012, in which the first eye is recognized but the second eye is verified as 1:1.

More than three attempts: There are 1495 passage events in which there were more than three attempts. These are due to some users unexpectedly interrupting the operation of the kiosk in the middle of its operation.

Enrollments of left and right eye on different days: Some (14) travellers have eyes enrolled on different years. When there was a problem enrolling an eye image, the older eye image was often kept.

Multiple enrollments (dilation scores) at enrollment: Some (1405) travellers have multiple IQ data (including dilation score) at enrollment transactions for the same eye, due to applying several attempts to enroll the iris.

Other issues: As mentioned above, the system performs a 1-to-First search. In doing that a new probe iris image, which can be either from Left eye (default eye) or Right eye (when Left eye did not find the match) is compared to *all* iris images stored in the enrollment database, including left and right eye images, and sometimes old and new name-records of a person. This results in some unknown number of zero-effort false match scores being recorded as part of the dataset.

A filtered version of that OPS-XING dataset with data aberrations marked or removed (other than the unknown number of false match scores) has been prepared and used in our analysis.

4 Methodology for analyzing the performance of NEXUS kiosks

This section presents one of the key results of our study, which shows that the performance of the system varies considerably among the subjects and that subjects who experience problems with the system use it much less than others. Based on this finding, methodology for subject-based performance analysis is developed to allow one to investigate the factors affecting the system performance. The taxonomy for categorizing such factors is established.

4.1 Variation of performance among subjects

As mentioned in Section 2, the OPS-XING dataset does not contain the data about travellers who were rejected by the kiosks. Therefore, the following two metrics are used to stipulate the number of travellers who have experienced difficulty in using the system, knowing that some of them used the system only once and some used it more than a hundred times, with 942 passages being the largest number of passages for a subject.

- Metric 1: Traveller's average number of $Attempts$ is higher than 1.5 (i.e., s/he is over 50% likely to be rejected by the system from the first attempt).
- Metric 2: Traveller's minimum matching score $HDNORM$ is higher than 0.2.

The first metric relates directly to the border wait time, which is a performance metric that the agency needs to minimize. This metric however may not always show the actual number of attempts taken by a traveller (e.g., as described in Section 2, when a traveller tries different kiosks or different sessions at the same kiosks, the number of attempts from the last session is recorded only). The second metric addresses this issue, as it allows one to estimate the difficulty of using the kiosk under situations when the number of recorded attempts is the same.

As highlighted in Section 2.2.2, the $HDNORM$ metric correlates with the $Attempts$ metric (the more attempts it takes the traveller to be recognized, the worse is the HD value). This allows one to use $HDNORM$ metric as a proxy performance metric for kiosk performance instead of $Attempts$.

Table 2 shows the number of travellers who used the system different number of times and the percentage of them who experienced the "difficulty" using it, where the difficulty is defined using the two metrics described above.

Table 2 Number of travellers as the function of the number of times they used the system and percentage among them experiencing "difficulty".

Times used the system	2+	4+	8+	16+	32+	64+	128+
Number of travellers	383,463	287,472	196,573	119,538	61,332	24,383	6,530
Percentage of them having $HDNORM > 0.2$	4.2%	2.4%	1.3%	0.8%	0.6%	0.3%	0.2%
Percentage of them having $Attempts > 1.5$	3.4%	2.4%	1.3%	0.6%	0.3%	0.12%	0.06%

Note: "Difficulty" is measured by high minimum HD score ($HDNORM > 0.2$) and high average number of attempts ($Attempts > 1.5$). The temporal information (i.e., whether a traveller used the system over a short or long period of time) is not used. More details are provided in [28].

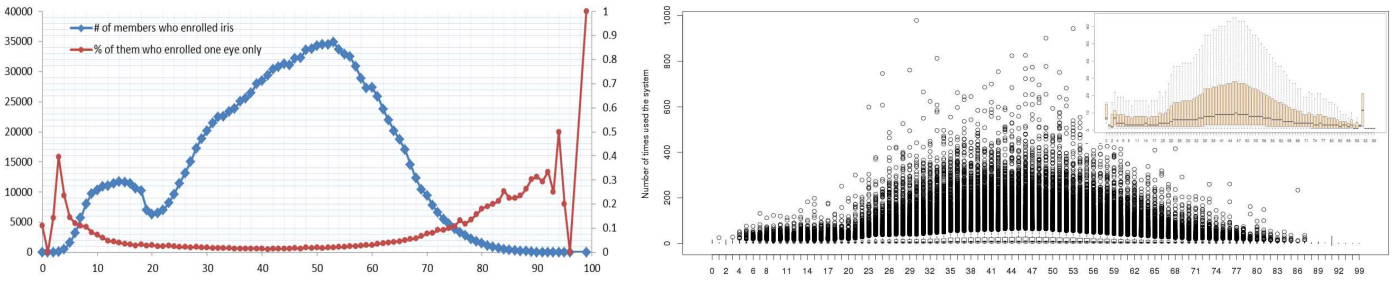


Fig. 4: The number of travellers by age at Enrollment and Passage. Left image shows the number of travellers who enrolled iris (in blue) and the percentage among them who were able to enroll one iris only (in red). The right image shows boxplots summarizing the number of passages for each age. Inset shows 95% truncated boxplots (i.e., with 5% of outliers removed).

It is observed that travellers who experience "difficulty" in using the system use it much less than those who do not. Therefore, any performance evaluation results obtained by aggregating transaction metrics, such as those obtained in previous analysis of the OPS-XING dataset [3]-[10], will be highly skewed towards "better" performing subjects. In order to provide an objective picture of the system performance quality, *subject-based performance* analysis is required.

In contrast to the *transaction-based analysis*, established by the ISO and currently used by industry [34], which answers the question: *"How many times did the system reject a person?"*, the *subject-based analysis* answers the question: *"How many persons were rejected by the system?"*

4.2 Subject-based performance analysis

Subject-based variation of biometric performance is well studied for voice and face modalities [30, 31]. It has been much less documented and analyzed for the iris modality. The major first evidence of subject-based variation of biometric performance in iris systems was presented in our earlier work in 2011 [27] and has become since then an important guiding principle for us in performing evaluation of biometric systems.

As a general rule for conducting subject-based analysis the following approach is used. All performance metrics X that are computed for a population are computed using the averages obtained separately for each individual (Eq. 2), as opposed to using averages over all individuals of the entire population (Eq. 3), done in the transaction-based analysis.

$$< X >_{subject-based} = \frac{\sum_{s \in \ subjects} (< X_s >)}{\#\ \text{subjects}} \tag{2}$$

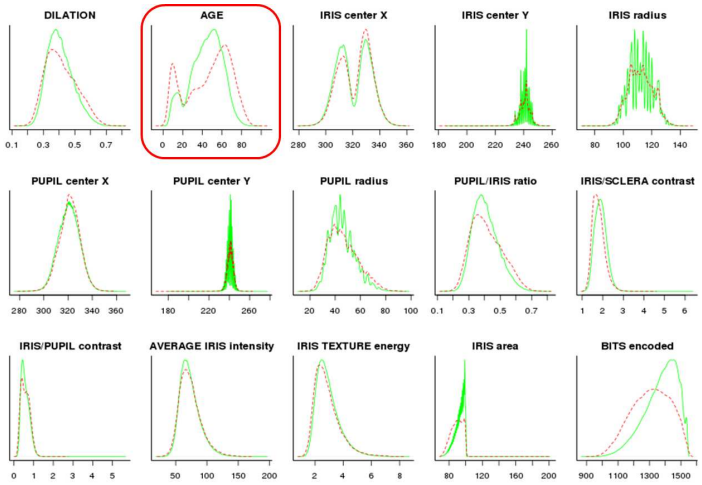
$$< X >_{transaction-based} = \frac{\sum_{s \in \ transaction} X_t}{\#\text{transactions}} \tag{3}$$

In general, one should expect transaction-based metrics to be different from subject-based one, skewed towards the average metrics of the most frequently observed subjects. By conducting subject-based analysis, one is able to better decipher the factors that negatively affect the system performance. These factors are categorized and analyzed next.

4.3 Factor categorization

From an operational perspective, it is important to distinguish factors by their prime cause. Using the approach that we first developed for video surveillance applications [29], the factors that effect biometric systems performance are classified into one of three types according the "technology-process-subject" factor triangle:

- *Technology-related factors.* This group of factors relate to the general limitations of the technology. They affect all users regardless of the process and user-specific characteristics. Any improvement of the system performance due to these factors requires contacting a vendor and potentially replacing the technology. Aging (i.e., deterioration of the technology performance with time) is an example of technology-related factor.
- *Process-related factors.* The second group of factors relate to the conditions in which the technology is used. It is normally the responsibility of the organization deploying the technology to make sure



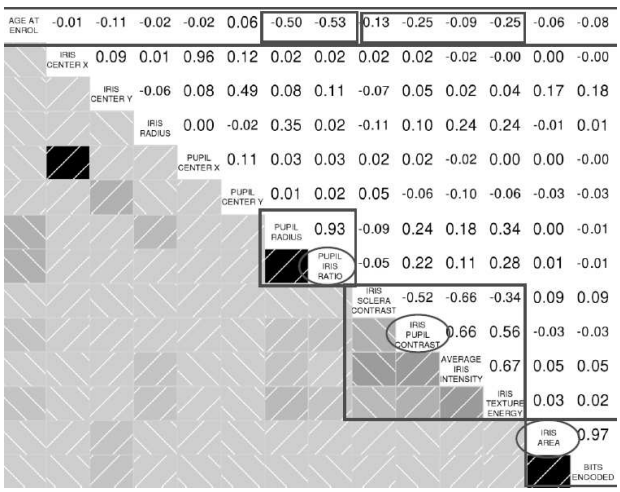


Fig. 5: Analysis of scores at Enrollment: Relative distribution of Age and Image Quality scores for "two-eye" (solid green) vs. "one-eye" (dashed red) enrollments (shown at left); Correlation of Age and Image Quality scores (shown at right). Data from new "B" cameras are used.

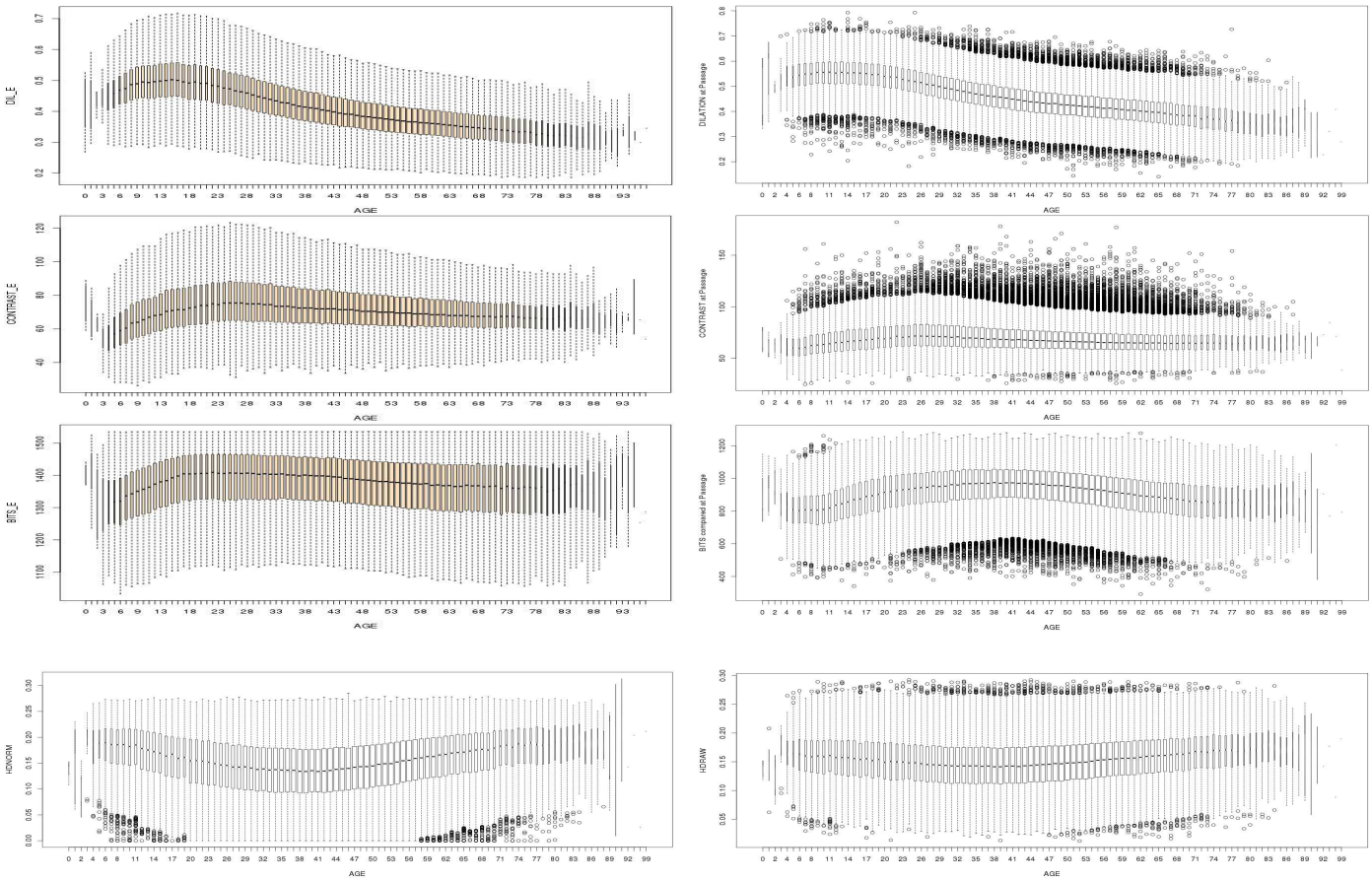


Fig. 6: Variation of Image Quality and Matching Scores by age. Boxplots on the top show the distribution of Dilation, Contrast, and Number of Bits Encoded / Compared scores for each age – at Enrollment (left) and Passage (right). Boxplots at the bottom show the distribution of HDNORM and HDRAW scores at Passage. Box width is proportional to the population size. Data from new "B" cameras are used.

that the technology is used under the conditions where it works the best. Kiosk location is a prime source of process-related factors, potentially leading to worse image quality and performance for all users.

- *Subject-related factors.* The last group of factors relate to particular characteristics of a person or group of people that make some travellers more vulnerable in operating biometrics systems than others. This includes person's gender, age and other subject-specific physiological and behavioural peculiarities such eye colour, size or shape of pupil, medical conditions, including wearing contact lenses. If such factors are detected, they can be used to improve the performance of the system by either alerting a user (e.g., by automatically detecting contact lenses and asking a user to remove them), or by allowing different thresholds for users of different groups (e.g., for the elderly).

In the following the effect of these three groups of factors is examined, using the enrollment data and then using the passage data.

5 Analysis of Enrollment data

Enrollment data allows one to examine subject-related factors, specifically the affect of age on image quality. It does not require subject-based metrics, because all enrolled travellers have exactly one enrollment transaction.

5.1 Young and elderly have worse image quality and are harder to enroll

Figure 4 shows the number of NEXUS members who enrolled iris and the percentage among them who could enroll one iris only for each age: from newborn to 100-year old people. A dip at 19-20 years of age is explained by the NEXUS program rules where children 18 and under are free to enroll at no charge with parents.

Two important observations are made. First, it is seen that the majority of enrolled travellers are between 30 and 60 years old, and almost all of them ($> 98\%$) were able to enroll both irises.

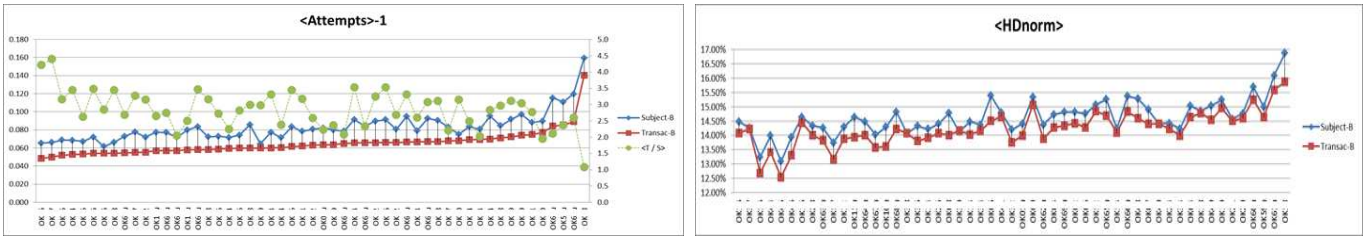


Fig. 7: Effect of kiosk location: Performance of NEXUS kiosks measured by the average number of attempts (left) and the average matching score (right), using transaction-based (in red) and subject-based (in blue) metrics, sorted from best performing to worst performing. The average number of transactions per subject (T/S) is shown (in green). Kiosks numbers are obscured to protect airport identities.

Second, it is seen that the ability to enroll both eyes is much worse for young travellers and and diminishes steadily with age for older travellers. This is an indication that image quality of these age groups is worse than that for middle-age group. This conjecture is validated next.

In order to remove the factors due to camera quality, the data from travellers enrolled with new ("B") cameras are used only. These data counts for over 95% of the dataset. The boxplots for Dilation, Contrast, and Number of Bits Encoded values at Enrollment in these data, for each age, are shown in Figure 6.

It is observed that dilation monotonically decreases with age for adults, which supports the conclusions from [3]. However, it is also observed that other IQ metrics also slightly decrease with age for adults. The decrease of all IQ metrics for young people is also observed. This explains the lower number of iris enrollments for older and young users.

In order to further examine the relationship between traveller's Age and IQ metrics at Enrollment, we plot in Figure 5 the correlation of Age and IQ metrics, and the distribution of Age and IQ metric scores at Enrollment - for cases where both irises were captured vs. those cases where where only one iris was captured.

The observation is that older (over 60 years) and younger (under 15 years) users are harder to enroll, i.e., have more "one eye only" enrollments. Three distinct IQ metric groups are also observed – related to Dilation (pupil-iris ratio), Contrast, and Openness, of which the Dilation group correlates with Age the most (at 0.53).

6 Analysis of Passage data

6.1 Variation of performance by kiosk location

As pointed out by the UND researchers in [7, 10], the NEXUS system performance varies among airports. Using subject-based analysis we can now further quantify this observation, while demonstrating the importance of applying such analysis for the NEXUS application.

Figure 7 shows the performance of all kiosks measured by the average number of additional attempts ($Attempts - 1$) and the average matching score (HDNORM), computed using transaction-based and subject-based metrics, sorted from worst to best. The average number of transactions per subject (T/S) is shown as well.

It is observed that some kiosks perform 10-20% better than others, according to both metrics. Furthermore, it is seen that performance reported using subject-based metrics is always worse than that reported using transaction-based metrics, sometimes by more than 30%. Kiosks with higher Transactions per Subject ratio (T/S) report better averaged performance, which is not surprising taking into account the finding presented earlier that people who use the system more frequently tend to have better matching scores.

It is also observed that the variation in kiosk performance within the same airport and the same direction of border crossing is less than that across different airports or different direction of border crossing. We use this finding later, when we need to minimize the effect of kiosk location on the system performance.

To further quantify the difference in performance due to kiosk location, we apply T-test [32] on the $HDRAW$ scores measured at different kiosks. The application of T-test is justified in this case, because we have over a thousand points measured at each kiosk and the distribution of $HDRAW$ scores is unimodal as highlighted earlier in Section 2.2 (Figure 2). Table 3 shows the result. It shows the 95% confidence intervals for the difference in the kiosk average HDRAW score computed for two better performing (in green) and two worse performing (in red) kiosks. Kiosks are chosen so that to have different traffic densities (one has much higher traffic than the other). Results are obtained using both subject-based and transaction-based metrics. The HDRAW scores are shown in grayed area, the number of transactions and subjects (T / S) for each kiosk are shown on the margin. The 95% confidence intervals for the score difference are shown in the middle part of the table.

Table 3 The difference in the average of the HDRAW score.

			t-b	s-b	t-b	s-b	T / S
			OK		OK		
		<HDraw>	0.15844	0.15918	0.16271	0.16842	
t-b	OK	0.13882	[0.0190, 0.0201]		[0.0205, 0.0272]		48099
s-b		0.14164		[0.0167, 0.0183]		[0.0220, 0.0315]	15184
t-b	OK	0.13904	[0.0190, 0.0198]		[0.0203, 0.02698]		95642
s-b		0.14171		[0.0169, 0.0180]		[0.0220, 0.0314]	36506
T / S			209222	80582	29327	27294	

Note: HDRAW score is computed for two better performing and two worse performing kiosks using subject-based (s-b) and transaction-based (t-b) metrics.

It is observed that the difference in system performance due to different kiosk location can be as high as 15%. This confirms that kiosk location is one of the most important factors affecting iris recognition performance.

6.2 Variation of performance by age

This section presents the main finding of our analysis related to the demographic bias of the iris biometrics, i.e., that iris biometrics performs worse for certain age groups. The existence of a demographic bias in other biometric modalities (face, fingerprint) has been reported previously and has become the basis for the development of new ISO guidelines on mitigating such biases [35]. Nothing however has been reported so far on the existence of a demographic bias in iris systems.

By examining the Passage statistics for each age (shown in Figure 4), it is noted that middle-aged travellers use the system much more often than young and elderly travellers. At the same time, as highlighted earlier (Section 5.1), middle-aged travellers have better quality enrollment images and therefore should be expected to have better performance at Passage. Hence subject-based analysis, introduced in Section 4.2, needs to be applied in order to objectively measure the effect of age on the technology performance. This is done below. In meanwhile, knowing the high interest in using iris biometrics for humanitarian and national ID programs [21], we can confirm (from our Enrollment and Passage age statistics) that iris biometrics is as successfully used by young children and youth as it is by elderly.

Figure 6 shows boxplots of IQ scores (Dilation, Contrast, Number of Bits Compared) at Passage. The bottom of the figure shows the boxplots of matching scores (HDNORM, HDRAW) for each age group in the OPS-XING dataset: from newborn to 99-year old persons that have used the kiosks. Data are taken for all kiosks and all cameras.

As with enrollment data, variation of image quality scores among different age groups is observed. The increased (worse) matching scores for young and elderly travellers are also observed. In the following we further quantify the variation of the system performance due to age, and compare it to that due to other factors.

Figure 8-a plots the average HDNORM and DIL scores as a function of AGE computed using generalized additive model (GAM) regression [32] for three largest Canadian airports (Toronto Terminal 1, Vancouver, and Montreal). Subject-based analysis is conducted separately for each airport for travellers enrolled with old ('L') and new ('B') cameras. The number of subjects at each airports for each camera is indicated on the top of which graph. The gray area shows 95% confidence interval. Large gray areas for travellers of over 80 years of age indicate that there is not sufficient data to reliably compute the function.

A clear drop in average matching scores (i.e., better performance) for middle-aged travellers is observed at each airport: from 0.18 (for those younger than 15 years and older than 80 years) to less than 0.14 for 40-year old travellers. This is in contrast to average Dilation, which monotonically decreases with age: from 0.55 (at 15 years of age) to 0.35 (at 80 years of age). This is an indication that Dilation is *not* the only factor that contributes to worsening of the matching score. Other Image Quality metrics are also likely affecting the result.

It is also noted that kiosks in Vancouver airport have been relocated during the period of data collection, resulting in their improved

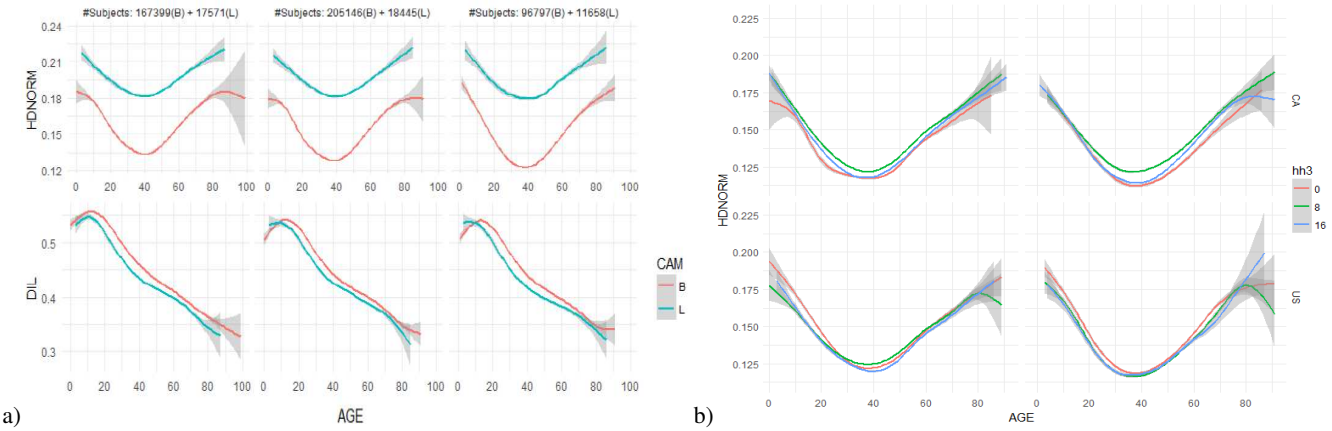


Fig. 8: Effect of age.
a Average HDNORM and DIL computed for subjects enrolled with old ('L') and new ('B') cameras at three largest airports using generalized additive models (GAM) regression.
b Average HDNORM computed for subjects enrolled with new ('B') cameras at different times of day: 0:00-8:00 (in red), 8:00-16:00 (in green), 16:00-24:00 (in blue).

performance (which was noted in [10]). This however did not affect the result related to the variation of system performance by age. It is also seen that variation due to age is larger than that due to kiosk location.

6.3 *Age vs. time of day and time of year*

The data used in the previous experiment is further split into three subsets, corresponding to three different times of day (morning, mid-day, evening), using Left-eye transaction data from travellers enrolled with 'B' cameras. Figure 8-b shows the results for two airports. Bottom row shows results for kiosks in US pre-clearance area, top row for kiosks in the arrival area.

A slight increase in matching scores for all ages at mid-day, i.e., during the brightest time of the day, is seen in two areas. This is consistent with earlier results suggesting that iris recognition produces poorer match scores when passage image acquisition takes place in strong sunlight, and is an indication that kiosks in those two airports are likely located where a large amount of sunlight comes through the windows. Critically however, it is seen that performance variation due to day time difference is much less than that due to age difference.

In another experiment, some consistent increase in HDNORM during December - January was also observed, supporting earlier such funding in [10]. In contrast to [10] however, where such variation is explained by the effect of season on eye dilation, we are inclined to think that most likely this is due to the subject-based performance variation, as more people travel and use the technology during the holiday season, including those who do not travel often and who (based on the results presented above) have a higher risk of experiencing the difficulty in using the system. In either case, the effect of time of the year is also seen to be much less than that of age and kiosk location.

6.4 *Age vs. aging*

To address the debate between NIST and UND researchers related to the effect of aging, we compare this effect to that of age and other factors. In order to do that, we apply *generalized additive mixed models* (GAMm) regression [33] to compute average HDRAW scores as a function of age (AGE) and aging (measured by the number of days since enrollment, ELAPSED_TIME) using Left-eye passage data from all kiosks for all users enrolled with 'B' cameras.

In contrast to *generalized additive models* (GAM) used earlier (Figure 8), *generalized additive mixed models* allow one to include random effects, which in this case are kiosk location (FAKE_KIOSK_ID) and person's physiology (FAKE_ID), in addition to fixed effect (AGE and ELAPSED_YEAR). The 'GAMm' function from the 'mgcv' R package is used for this purpose [33].

Once the predictive model is computed, it is applied to compute the expected average HDRAW scores for a grid of age-aging values, where age is incremented by 5 years, and aging (ELAPSED_TIME) by 100 days. The result is shown in Figure 9-a. The following observations are made.

First, for all ELAPSED_TIME groups (i.e., along the horizontal axis), the relationship between the matching score and age is exactly the same as found earlier (seen in Figure 8): the matching score is the lowest at 35-40 years of age and monotonically increases as one moves further away (left or right) from the middle.

Second, for most age groups (i.e., along the vertical axis), aging has no negative effect on matching scores. It is only for 55-65 age group, where slightly increased (worse) matching scores with aging are observed. Critically, the variation in matching score due to aging is much less than that due to age difference.

To explain the observed improvement of HDNORM score with ELAPSED_TIME, we offer the following four reasons: 1) habituation (travellers learn how to make the machine work better for them, e.g., by opening wider their eyes), 2) the improved positioning of the kiosks (as in Vancouver, found in [10]), and 3) the use of transaction-based metrics (which show 'better' results for travellers who use the

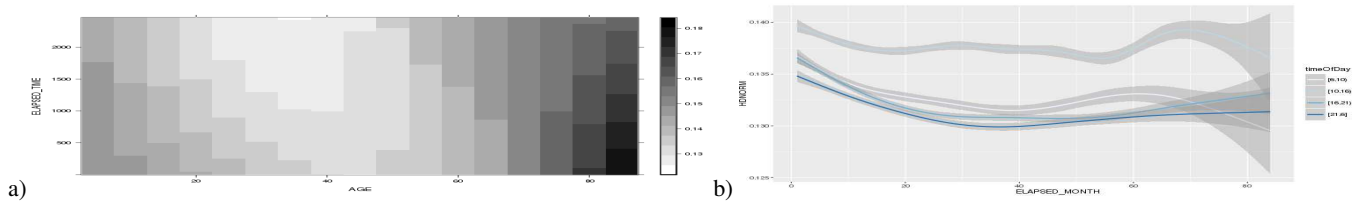


Fig. 9: Effect of aging.
a Average HDRAW as a function of AGE and the number of days since enrollment (ELAPSED_TIME) computed using generalized additive mixed models (GAMm) regression. Kiosk fake id (FAKE_KIOSK_ID) and traveller's fake id (FAKE_ID) are treated as random effects, while AGE and ELAPSED_TIME are treated as fixed effects.
b Average HDNORM as a function of the number of months since enrollment (ELAPSED_MONTH) computed using generalized additive model (GAM) regression at different times of day of the passage. Data from a single airport, where variation due to kiosk location is small, are used.

system more often), and 4) the reduction over time in the threshold for recording a match score, THD, means that subjects who use the system over a period of years are able to record a higher score in the earlier years of using the system than they are able to record in the later years.

In order to place the effect of Aging in context with other factors, we compare it to that of time of day. Figure 9-b shows the average HDNORM computed using generalized additive model regression on the data taken from a single airport (which has little variation among its kiosks) as a function of Aging (ELAPSED_MONTH) for four different times of day (morning, mid-day, evening and night). It is observed that the effect of aging is less than that of time of day of passage transaction, which in turn (as discussed earlier) is less than the effect of age and kiosk location.

To conclude, taking into account the results from previous sections, where it was shown that age correlates with IQ metrics, particularly, with Dilation and (to lesser degree) with Contrast, it can be stated that "aging problem" is not about "whether a biometrics modality changes in time" (yes, it does), but rather about "whether the technology can deal with the changes due to aging". Evidently, iris biometrics can deal with changes due to aging quite well, at least over the range of years analyzed in this study (which is seven years). At the same time, it is seen that, as with all other biometric modalities, its performance is affected by sensor quality, capture conditions (lighting), and also by person's age (when comparing technology performance for travellers of different age groups).

6.5 Factor significance

Once the effect of certain factors (explanatory variables) on the performance of the system (response variables) is hypothesized through the observations of descriptive statistics (Figures 8-9), it is possible to apply analysis of variance to obtain the values of statistical significance for each factor and their combination [32]. This is done below, where a combination of subject-related (age), technology-related (aging), and process-related (time of day and time of year) factors are examined for statistical significance.

To avoid the variation due to kiosk location, the data from kiosks in a single airport (where variation due to kiosk location is small) are used. Age is presented as 9-level factor (each level representing a decade), aging is 8-level factor (each level representing a year since enrollment), time of day and time of year are presented as 4-level factors (as done in previous sections). Table 4 shows the result, as produced by running the analysis of variance in R language [32]. The plots showing 95% confidence level intervals on matching score differences for all pair-wise combinations of factors values are presented in Figure 10.

It is seen that all listed factors are statistically ($> 99.9\%$) significant, with a combination of age and aging being less significant than other factors. From a practical point of view however, the important question is not which factors affect the technology performance but to what degree they affect it.

Critically, for an organization deploying the technology it is important to know whether any action is required to improve the system performance. According to the "technology-process-subject" factor triangle (described in Section 4.3), three possible types of actions are possible: replacing the technology, improving the process, or implementing subject-based customization of the decision rules or procedures. As presented in the concluding section, the recommendations related to these actions appear evident from the results presented in this paper without the need of doing more detailed statistical analysis.

Table 4 Analysis of variance in matching scores due to various factors.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
AGE	8	16	2.015	476.87	< 0.0000000000000002	***
ELAPSED_YEAR	7	0	0.024	5.76	0.0000010803	***
timeOfYear	3	0	0.043	10.07	0.0000012431	***
timeOfDay	3	1	0.257	60.91	< 0.0000000000000002	***
AGE:ELAPSED_YEAR	46	0	0.007	1.73	0.0015	**
timeOfYear:timeOfDay	9	0	0.026	6.19	0.0000000091	***

Note: The last column shows the probability $Pr(> F)$ of having the same mean output value despite the change in input factor value.

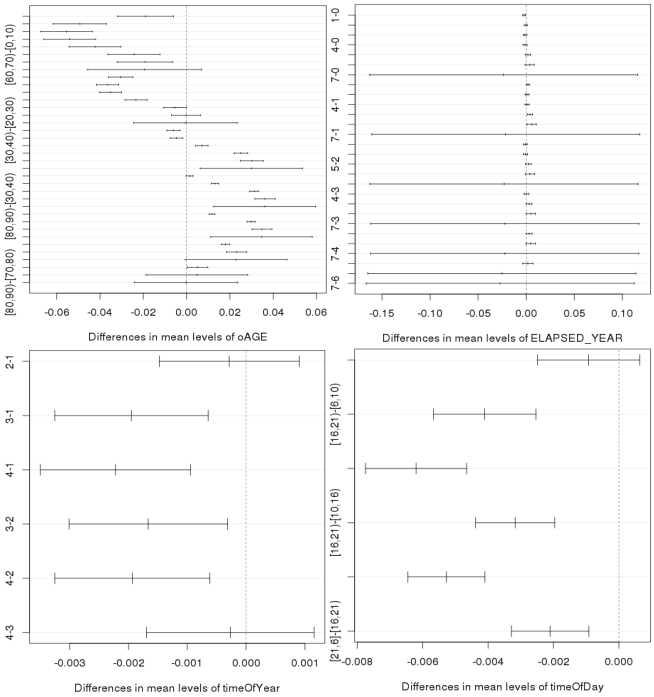


Fig. 10: The 95% confidence level intervals on matching score difference for all pair-wise value combinations of four factors: clockwise – age (9-level factor), aging (8-level factor), time of day (4-level factor) and time of year (4-level factor).

7 Conclusions

Iris biometrics was introduced to automated border control as an extremely robust biometrics [22]. Results obtained from a watch-list screening border application in United Emirates [23] have solidified this belief. When later the University of Notre Dame researchers published results showing that iris performance varied over time [12–14], it brought a lot of concern from the technology users, including many government organizations who actively rely on iris technology in their operations [16]-[19]. To address these concerns, NIST undertook an effort to better understand the effect of aging and other factors on iris biometrics [3]. This effort opened a whole new range of questions related to the factors that effect iris recognition and the ways iris biometrics is evaluated [5]-[10].

Thanks to the efforts of NIST and UND scientists, our understanding about the properties and limitations of iris biometrics and current evaluation practices has improved significantly. The results presented in this paper further contribute to these efforts. Three major conclusions from the obtained results are made.

First, in the applications where the use of technology is not mandatory, as in automated border control [28], it should be expected that subjects who experience problems using the system will use it less than those who do not experience problems. Hence, the performance of biometric systems in such applications, if measured using traditional transaction-based metrics, may show unrealistic "overly optimistic" results. Therefore, the use of subject-based metrics introduced in this paper should be used when analyzing and reporting the performance of such systems.

Second, in relationship to the aging debate [4], where the CBSA-collected OPS-XING dataset played a very important role, it is concluded that the effect of aging is negligible, compared to that of other factors such as kiosk location, time of day, and person's age.

While the effect of kiosk location and time of day on system performance has been already uncovered by the UND researchers [10]

using the previous releases of the OPS-XING dataset, the discovery of the effect of person's age on system performance was made possible only now, using the previously unused portions of the dataset. It is shown that older (over 60 years of age) and young (under 20 years of age) travellers are disadvantaged by the system. The system log shows worse image quality and matching scores for these groups, compared to that of middle-aged travellers. The variation of system performance due to age difference is larger than that due to light changes or different kiosk location.

In a society concerned with providing equal quality services to its all demographic groups (see [36]), this finding may help to adjust its technology settings so that to mitigate the demographic bias exhibited by the iris recognition technology. A new guidelines document is being prepared by ISO in this regard [35].

To conclude, it may still be possible theoretically to improve the results of the analysis conducted on the OPS-XING dataset (e.g., by applying non-linear mixed-effect models [32]). From practical perspective however, this additional effort appears of little importance, since none of analyzed factors appeared to effect significantly the system performance, and critical recommendations related to auditing and improving iris recognition systems can be made based on the results already obtained. These are listed below.

Using the "technology-process-subject" factor categorization triangle, described in Section 4.3, the first step for improving iris recognition performance is seen in optimizing the kiosk placement (Process factor). Then the performance can be further optimized by applying different matching decision or process rules for different age group populations (Subject factor). For example, a higher threshold or a larger number of attempts may be allowed for old and young subjects, or a score normalization formula can be further improved to take into account person's age and other image quality metrics, as discussed in Section 2.2. This will mitigate the demographic bias exhibited by the system. However, no action in relationship to aging-related concerns (Technology factor) appears to be needed.

Acknowledgment

This work was initiated and partially funded by the Canadian Safety and Security Program (CSSP) managed by the Defence Research and Development Canada, Centre for Security Science (DRDC-CSS), as part of the CSSP-2013-CP-1020 ("ART in ABC") project [28] led by the CBSA. It has also contributed to the DRDC-funded CBSA-led CSSP-2015-TI-2158 ("Roadmap for Biometrics at the Border") project deliverables related to the Gender-Based Analysis Plus (GBA+) [36]. Feedback from Kevin Bowyer, Adam Czajka, Patrick Grother, and Jim Matey on iris technology related matters, and assistance from Jordan Pleet and Rafael Kulik on statistical matters are gratefully acknowledged.

Dedication

Dmitry O. Gorodnichy dedicates this paper to the memory of his father, the Doctor of Science of the Ukrainian Academy of Sciences, Oleg P. Gorodnichy (Gorodnichii).

8 References

1 Canada Border Services Agency. NEXUS Air: http://www.cbsa-asfc.gc.ca/prog/nexus/air-aerien-eng.html.
2 Canada Border Services Agency. CANPASS Air: http://www.cbsa-asfc.gc.ca/prog/canpass/canpassair-eng.html.
3 Grother P., Matey J.R., Tabassi E., Quinn G.W., Chumakov M.: IREX VI. Temporal Stability of Iris Recognition Accuracy, NIST Interagency Report 7948, 2013.
4 IET Biometrics Journal, Iris Ageing Debate in IET Biometrics: http://www.theiet.org/resources/irisageing.cfm. Accessed: Sept. 2015 - Nov. 2017.
5 Grother P., Matey J.R., Quinn G.W.: IREX VI: mixed-effects longitudinal models for iris ageing: response to Bowyer and Ortiz, IET Biometric, Volume:4, Issue:4, 2015.
6 Bowyer K., Ortis E. : Critical examination of the IREX VI results, IET Biometric, Volume:4, Issue:4, 2015.
7 Ortis E., Bowyer K.: Exploratory Analysis of an Operational Iris Recognition Dataset from a CBSA Border-Crossing Application, IEEE Computer Society Biometrics Workshop, June 2015.
8 Czajka A., Bowyer K.: Statistical Evaluation of Up-to-Three-Attempt Iris Recognition, IEEE International Conference on Biometrics Theory, Applications and Systems (BTAS 2015).
9 Kuehlkamp A., Bowyer K.: An Analysis of 1-to-First Matching in Iris Recognition, IEEE Workshop on Applications of Computer Vision, March 2016.
10 Ortiz E., Bowyer K.: Pitfalls In Studying Big Data From Operational Scenarios, IEEE International Conference on Biometrics Theory, Applications and Systems (BTAS 2016).
11 Wild P., Ferryman J., Uhl A., "Impact of (segmentation) quality on long vs. short-time span assessments in iris recognition performance", IET Biometrics, vol. 4, no. 4, 2015.
12 Baker S., Bowyer K., Flynn P.: Empirical evidence for correct iris match score degradation with increased time-lapse between gallery and probe matches. Proc. International Conference on Biometrics (ICB), pages 1170-1179, 2009.
13 Baker S., Bowyer K., Flynn P., Phillips J.: Empirical Evidence for Increased False Reject Rate with Time Lapse in ICE 2006, NIST Interagency Report 7752, 2011.
14 Fenker S., Ortis E., Bowyer K.: Template Aging Phenomenon in Iris Recognition, IEEE Access (Volume: 1), Page(s): 266 - 274, 16 May 2013.
15 "Researchers reawaken iris-ageing debate", Accessed: 30 November 2015 http://www.planetbiometrics.com/article-details/i/3439/desc/researchers-reawaken-iris-ageing-debate.
16 "Aged eyes prevent iris recognition. Healthy Seniors", 3/7/2012. http://www.healthyolderpersons.org/news/aged-eyes-reventiris-rec.
17 "Aging process confounds iris recognition biometrics". Homeland Security Newswire, 5/31/2012. http://www.homelandsecuritynewswire.com/dr20120531-aging-process-confounds-iris-recognition-biometrics.
18 "Researchers question long-term reliability of iris recognition". Third Factor, 7/17/2012. http://www.thirdfactor.com/2012/07/17/researchers-question-long-term-reliability-of-iris-recognition.
19 Browning K., Orlans N.: Biometric Aging Effects of Aging on Iris Recognition. Case Number 13-3472, 2014. The MITRE Corporation. https://www.mitre.org/sites/default/files/publications/13-3472-biometric-aging-iris-recognition.pdf
20 Christian Rathgeb, A biometric for life potential for a lifetime breeder document, International Biometric Performance Testing Conference (IBPC), 2014.
21 International Joint Conference on Biometrics (IJCB) 2014 Keynote speaker presentations: http://www.ijcb2014.org/Keynote_Speakers.html (S. Lenharo "Brazilian National Biometric Selection: New and Legacy Challenges", V.S. Madan "Digital ID for Benefit and Service Delivery to Billion Plus People", S. Braiki "The UAE Population Register and ID Card Program: Achievements and the Challenges", W.G. McKinsey ("The Challenges of NGI").
22 Daugman J.: How iris recognition works. IEEE Transactions on Circuits and Systems for Video Technology, 14:21, 2002.
23 Daugman J.: Probing the Uniqueness and Randomness of IrisCodes: Results From 200 Billion Iris Pair Comparisons, Proceedings of the IEEE (Volume: 94, Issue: 11), 2006.
24 Daugman J.: New Methods in Iris Recognition. IEEE Transactions on Systems, Man and Cybernetics, Part B, Vol. 37, No. 5, October 2007.
25 Daugman J.: Information Theory and the IrisCode, IEEE Transactions on Information Forensics and Security, 2015.
26 Gorodnichy D., Hoshino R.: "Score Calibration for Optimal Biometric Identification", Proc. Canadian Conference on Artificial Intelligence (AI 2010), Ottawa, Lecture Notes in Artificial Intelligence, Springer, 2010.
27 Gorodnichy D.: "Multi-order biometric score analysis framework and its application to designing and evaluating biometric systems for access and border control", Proc. IEEE SSCI Workshop on Computational Intelligence in Biometrics and Identity Management (CIBIM), April 2011.
28 Gorodnichy D.: "ART in ABC: Analysis of Risks and Trends in Automated Border Control". Technical Report DRDC-RDDC-2016-C324 (Full report): http://cradpdf.drdc-rddc.gc.ca/PDFS/unc256/p804885_A1b.pdf. Technical Report DRDC-RDDC-2016-C143D (Executive Summary): http://cradpdf.drdc-rddc.gc.ca/PDFS/unc229/p803869_A1b.pdf, 2016.
29 Gorodnichy D., Bissessar D., Granger E., Laganiere R.: "Recognizing people and their activities in surveillance video: technology state of readiness and roadmap", Proc. 13th Conference on Computer and Robot Vision (CRV), Victoria, 2016. Online: http://www.videorecognition.com/doc.
30 Doddington G., Liggett W., Martin A., Przybocki M., Reynolds D.: "Sheep, goats, lambs and wolves: A statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation", Proc. 5th International Conference of Spoken Language Processing, ICSLP 98.
31 Poh N.: IEEE IJCB Tutorial "System Design and Performance Assessment: A Biometric Menagerie Perspective", IJCB 2014 conference. http://ijcb2014.org/Tutorials.html.
32 Grolemund G., Wickham H. : "R for Data Science" , Publisher: O'Reilly, 2017. First Edition.
33 Wood, S.N.: "Generalized Additive Models: An Introduction with R". Chapman and Hall/CRC, 2006
34 ISO/IEC 19795-5, Information Technology – "Biometric Performance Testing and Reporting Part-5: Grading scheme for Access Control Scenario Evaluation".
35 ISO/IEC TR 22116, Information technology – "Identifying and mitigating the differential impact of demographic factors in biometric systems": https://www.iso.org/standard/72604.html
36 Treasury Board Secretariat of Canada, Gender-Based Analysis Plus. https://www.tbs-sct.gc.ca/hgw-cgf/oversight-surveillance/tbs-pct/gba-oacs-eng.asp.