Exploratory Analysis of an Operational Iris Recognition Dataset from a CBSA Border-Crossing Application

Estefan Ortiz
Department of Computer Science & Engineering
University of Notre Dame
Notre Dame Indiana 46556
eortiz@nd.edu

Kevin W. Bowyer
Department of Computer Science & Engineering
University of Notre Dame
Notre Dame Indiana 46556
kwb@cse.nd.edu

Abstract

This paper presents an exploratory analysis of an iris recognition dataset from the NEXUS border-crossing program run by the Canadian Border Services Agency. The distribution of the normalized Hamming distance for successful border-crossing transactions is examined in the context of various properties of the operational scenario. The effects of properties such as match score censoring and truncation, same-sensor and cross-sensor matching, sequence-dependent matching, and multiple-kiosk matching are illustrated. Implications of these properties of the operational dataset for the study of iris template aging are discussed.

1. Introduction

The Canadian Border Services Agency (CBSA) has run the very successful NEXUS border-crossing program for the past decade. The program allows low-risk travellers to save time through the use of self-serve kiosks at designated Canadian points of entry [1]. The NEXUS program uses iris recognition technology to recognize a person for a transaction at border-crossing kiosks. In addition, the program has hundreds of thousands of enrolled users, who have collectively logged millions of border-crossing transactions over the last decade. Thus NEXUS is a prime example of a successful, large-scale, long-term operational application of iris recognition.

The IREX VI report released by the United States National Institute of Standards and Technology (NIST) analyzes a dataset from the NEXUS program in its study of iris recognition aging [2]. IREX VI estimates a linear, mixed-effects regression model in an effort to study the effect of iris ageing on genuine match scores. Contrary to other research in iris template ageing [3-6], IREX VI concludes that there is "no evidence of a widespread iris ageing effect. Specifically, the population statistics (mean and variance) are constant over periods of up to nine years" [2].

We perform an exploratory analysis of a NEXUS iris recognition dataset in order to better understand how the dataset might be appropriately analyzed in studying iris template aging. The dataset that we obtained from CBSA is a superset of the dataset analyzed in IREX VI. It contains data for border-crossing transactions that have occurred since IREX VI. It also contains additional meta-data for each border-crossing transaction. Our analysis shows that various assumptions needed for the IREX VI regression analysis are not satisfied.

2. Dataset

The dataset received from the CBSA consists of enrollment data and border-crossing data. The enrollment data contains anonymized ("Fake_ID") records of persons enrolled in the NEXUS program. The total number of unique IDs is 705,553. There are 688,166 left irises enrolled, and 682,724 right irises enrolled, for a total of 1,370,890 enrolled irises.

The enrollment data spans the period from 2003 to 2014. The LG 2200 iris sensor, indicated by "L" within the dataset, was initially used for enrollment. Enrollment was later transitioned to the Panasonic BM-ET 330 sensor, indicated by "B" within the dataset. There are 86,043 enrollment instances made with "L", and 1,284,847 enrollment instances made with "B". All probe images for border crossing transactions were acquired using sensor "B". Thus, an "L" match is a cross-sensor match, and a "B" match is a same-sensor match. In general, cross-sensor matches have higher Hamming distances (HD) than same-sensor matches [7].

The border crossing data is a record of successful match attempts using the NEXUS system. Table 1 summarizes the meta-data elements present for each of the border-crossing transactions. In the context of studying iris aging, the two primary elements to be examined are the normalized Hamming distance (HDNORM), and the number of days elapsed between the probe image and the enrolled image (ELAPSED_TIME). Additionally, a match (successful transaction) is based on the normalized Hamming distance and the threshold (THD) that was in use at the time of the recorded transaction.

In this operational scenario, the matching of a probe iris against the set of enrolled irises is done in "1-to-first" manner [8]. That is, the probe iris is matched to each enrolled iris until the first below-threshold match is found. Once the first below-threshold match is found, the probe is not matched against any further enrolled irises. This is different from a 1-to-N search. In a 1-to-N search, the probe is matched against all enrolled irises and the best match is kept, provided that it is below threshold. A 1-to-first search is faster than a 1-to-N search, but also has a higher probability of generating a false-match result.

Table 1: Meta-data describing border-crossing transactions.

Meta-data	Description
FAKE_ID	Unique ID for enrolled person
EYE	Eye: L (left) or R (right)
MONTH	Calendar month: 1 to 12
ELAPSED_TIME	Number of days between enrollment and probe images
HDNORM	Successful HD score, with Daugman score normalization
DILATION	Pupil dilation estimate
CAPTURE_NUMBER_ WITHIN_PA	Attempt number on which the transaction was successful
FAKE_KIOSK_ID	Identifier for a particular border crossing kiosk
THD	HD threshold used for match transaction success
MATCHING_MODE	Indicator of left-then-maybe-right or both-left-and-right matching process.

The "capture number within passage attempt" (CAPTURE_NUMBER_WITHIN_PA) is understood as follows. A person using the kiosk acquires an initial left-right pair of iris images. The left iris is matched against enrolled left irises and if a below-threshold match is found, the passage attempt is successful. If no match is found with the left iris, the right iris is matched against the enrolled right irises, and if a below-threshold match is found, the passage attempt is successful. A match on either the left or right iris from the first pair of images is indicated with a 1 within the data. If neither the left nor right iris results in a match on the first attempt, then the person can acquire a second pair of images, and the matching process is repeated. Similarly, a match at this stage is indicated with a 2 within the data. If neither the left nor the right iris from the second set of images results in a match, then the person can acquire a third pair of images. A match from one of this pair of images is a capture number 3 match. If there is no match on the third pair of images, then the passage attempt "times out" and the passage attempt has failed.

The matching mode (MATCHING_MODE) described in Table 1 consists of two different modes. One mode labeled "SEM" indicates that a match was successful by comparing the left eye first, and then the right eye if the left did not match. The second mode, "SEP", corresponds to the situation in which the left and right irises were both matched independently. The NEXUS program started out using the SEM mode, then switched to the SEP mode for a period of time, then switched back to the SEM mode.

Our dataset obtained from CBSA contains a total of 8,900,684 transactions. It does not contain any of the iris images. The metadata for "passage number within attempt", for matching mode, and for kiosk ID was not used in the dataset analyzed for the IREX VI report.

3. HD Distribution Censoring and Truncation

Figure 1 shows a histogram of the overall distribution of normalized Hamming distance scores in the dataset. Two features stand out. On the low end, there is a sharp spike at zero. And on the high end, there is a steep cliff. Each of these features presents a problem for a Gaussian distribution assumption used in linear regression analysis.

The spike at zero results from how scores are computed and recorded in this operational scenario. The raw HD scores are (most likely) normalized according to a Daugman's procedure [9], given here in Equation 1:

$$NHD = 0.5 - \left(0.5 - HD\right) \cdot \sqrt{n/900} \tag{1.}$$

where NHD is the normalized Hamming distance, HD is the raw Hamming distance, n is the number of iris code bits in this particular match, and 900 is an empirical parameter estimated previously.

The purpose of the normalization is to take into account the fact that different match scores are computed based on different numbers of unmasked bits in the iris codes. One side effect of this normalization is that normalized scores can become negative. In this case, the negative values have been recorded as zero. In effect, the tail of the distribution that would have run into negative values has been "bunched up" on the value zero. In the regression analysis literature, this is referred to as "censoring" the data [10-15].

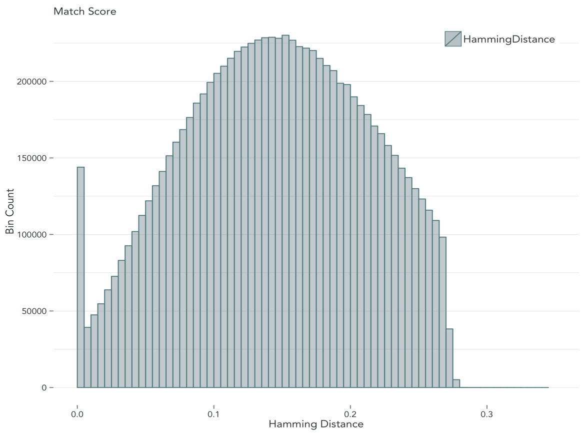

Figure 1: Normalized fractional HD for all transactions.

The steep fall-off at the high end of the distribution is due to the fact that, at any point in time, only scores that fall below the current threshold value (THD) are recorded. Scores that are above threshold, and so would not be accepted as a match that allows a border crossing, are simply not recorded. In the regression analysis literature, this is referred to as "truncation" of the data. In the operational scenario, the match score threshold was reduced at various times as the size of the enrollment database increased. This accounts for the presence of several "steps" in the "cliff" on the high end of the distribution.

The truncation of the distribution is documented in the IREX VI report, but the censoring of the distribution is not. Given that standard linear regression assumes that the distribution is Gaussian, both the censoring and the truncation mean that the Gaussian assumption is not satisfied. Appropriate regression analysis for this data would need to use truncated regression and censored regression techniques to deal with these issues.

4. Same- and Cross-Sensor Distributions

Figures 2 and 3 show the normalized HD distribution by sensor. The L (cross-sensor) distribution of scores is shifted toward higher values in general, has a more severe truncation at the high end, and has no noticeable spike at zero. In principle, the spike still exists because the scores are normalized in the same manner. The shift toward higher values results from the cross-sensor nature of the matching [7], and this also reduces the apparent spike at zero.

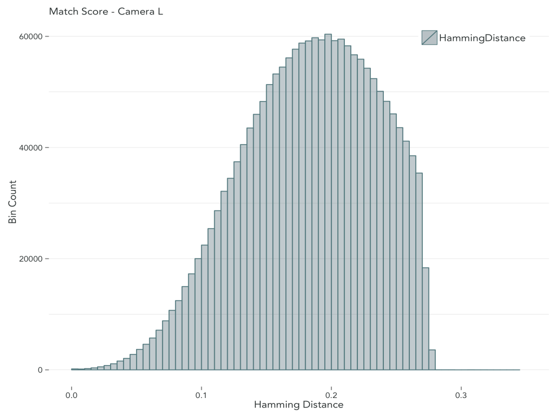

Figure 2: Normalized fractional HD for L transactions.

The B (same-sensor) distribution, in Figure 3, shows the strong spike at 0, along with the sharp drop due to the truncation on the high end. The B transactions comprise the majority of the dataset.

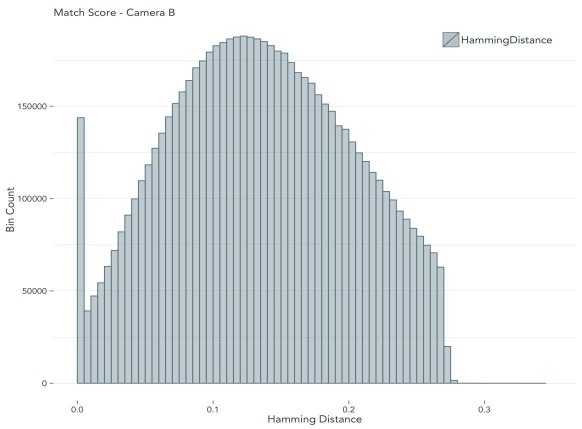

Figure 3: Normalized fractional HD for B transaction

5. Dependence of Left and Right Iris Scores

As mentioned earlier, the typical matching mode ("SEM") in this application is when the probe left iris is matched against the enrolled left irises, and if no below-threshold match is found, then the probe right iris is matched against the enrolled right irises. If no match is found for either the left or the right, the same process is followed with a second set of images, and then possibly also a third set. Thus, in this matching mode, a right iris match is only attempted if the left iris match failed.

The HD distributions for the left and right iris matches for sensor L are shown in Figures 4 and 5, respectively. The corresponding distributions for sensor B are shown in Figures 6 and 7, respectively. Note that the HD distributions for the right eye are clearly much worse than those for the left eye. Because of the truncation, the difference between the left and right iris distributions is not as simple as having the same shape of distribution with a difference in the mean.

An explanatory regression analysis should either treat the left and right scores data separately, or possibly include them as random effects in a mixed-effects regression model. The IREX VI analysis and regression model does not do either. Instead, it lumps left and right scores into the same regression with a binary indicator variable in the fixed effects.

The situation with right-iris match scores in the dataset is actually more complicated than is immediately apparent from Figures 5 and 7. While most right-iris matches were recorded in SEM (match right iris only if left iris fails) matching mode, some were recorded in SEP (left and right both matched independently) matching mode. There are also instances in which a matching mode was not recorded; such instances are indicated as "NA". The HD distributions are quite different between SEM and SEP matching modes, and the NA instances appear to be similar to the SEM matching mode. This is shown in Figure 8, where the aggregate right-iris HD distribution for sensor B is broken out by matching modes SEM, SEP and NA.

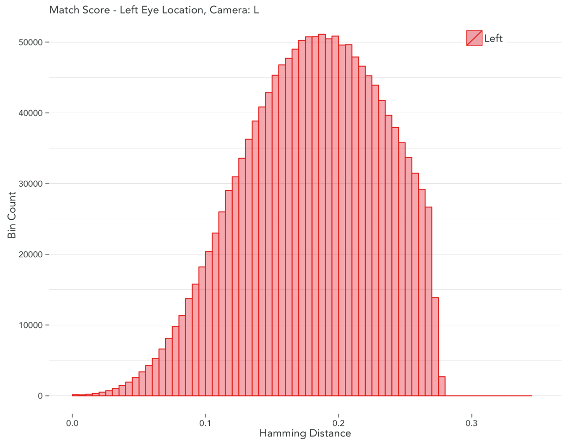

Figure 4: Normalized fractional HD, left eye, sensor L.

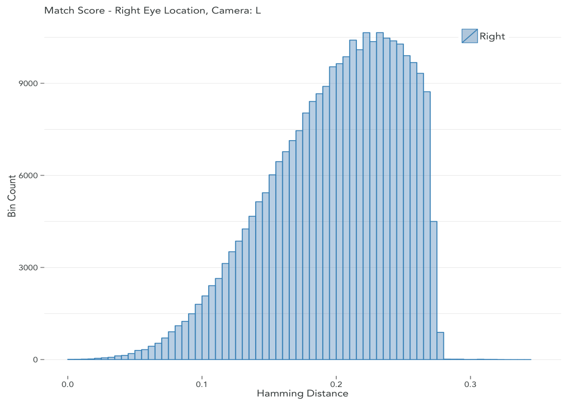

Figure 5: Normalized fractional HD, right eye, sensor L.

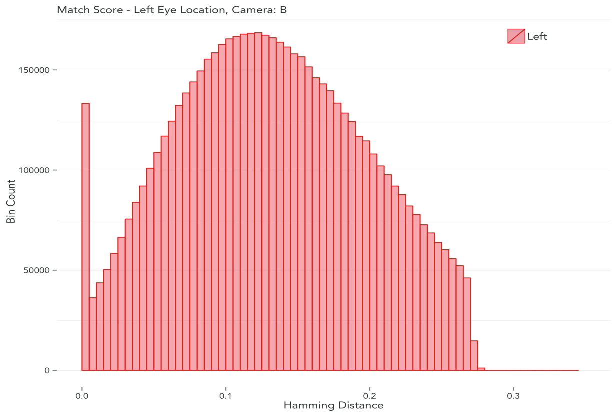

Figure 6: Normalized fractional HD, left eye, sensor B

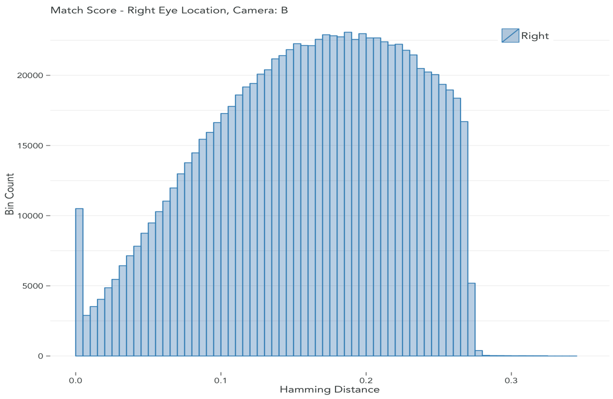

Figure 7: Normalized fractional HD, right eye, sensor B

The different matching modes operated over different time periods. Thus the difference in right iris HD distributions seen in Figure 8 is an inherent confounding factor in any attempt to use the right-iris data to study the effects of time lapse. It is certainly not appropriate to perform an explanatory regression analysis based on left and right iris scores together in a regression model that simply differentiates between the two with an indicator variable in the fixed effects, as is done in IREX VI.

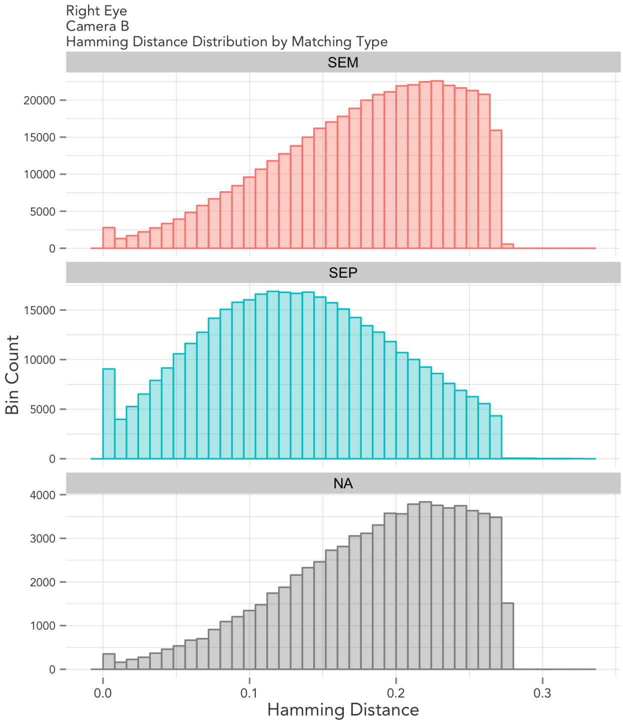

Figure 8: Right-iris, sensor-B HDs by Matching Mode

6. Sequential Dependence of Capture Number

It was mentioned earlier that up to three image captures could be made within one passage attempt. If no match results from the first pair of left and right iris images, a second pair is acquired, etc. Figures 9, 10 and 11 show the HD distributions for the first-, second- and third-attempt for sensor B. These are components of the overall distribution shown in Figure 3.

By definition, a second-attempt HD is recorded only in instances where matching the first-attempt image resulted in above-threshold HDs and so no HD was recorded for the first-attempt image. The fact that the second-attempt distribution in Figure 10 is clearly worse than the first-attempt distribution in Figure 9 suggests that the failure to find a first-attempt match was not due to a "random" problem. Rather, it suggests that there is some underlying process that persists across multiple image acquisitions. People whose first image does not result in a match are a sub-group that has a worse Hamming distance distribution for the second-attempt image than the overall group had for the first-attempt image.

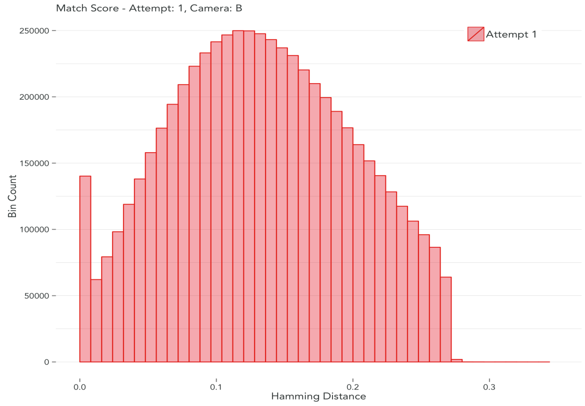

Figure 9: Distribution for first-attempt sensor B matches

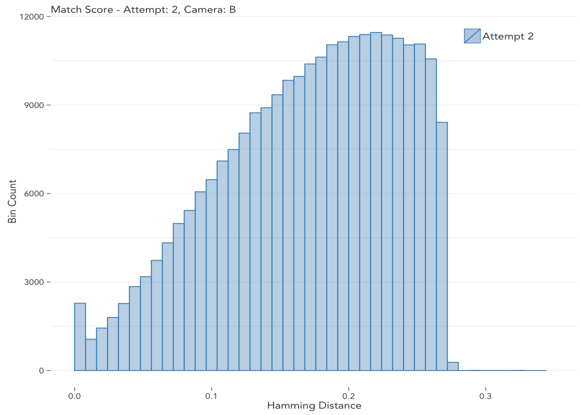

Figure 10: Distribution for second-attempt sensor B matches.

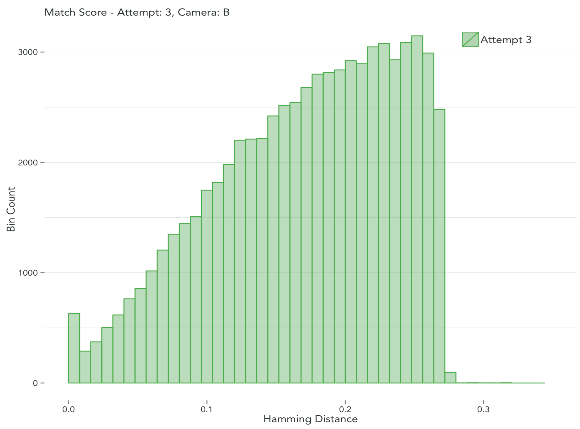

Figure 11: Distribution for third-attempt sensor B matches.

It is not clear that there is any significant difference in the HD distributions for second-attempt images, shown in Figure 10, and for third-attempt images, shown in Figure 11. Sensor L data shows the same pattern of differences in HD distributions across first-, second- and third-attempt images. These plots are shown in Figures 12, 13 and 14.

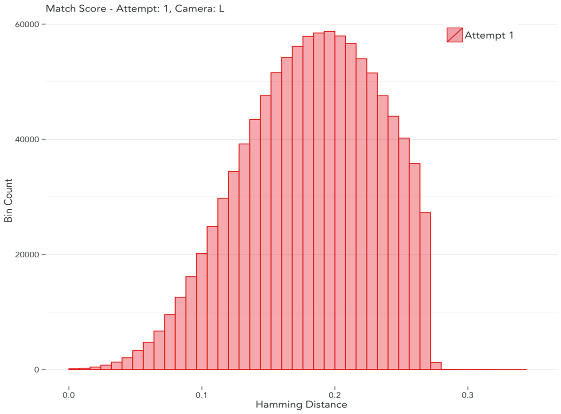

Figure 12: Distribution for first-attempt sensor L matches.

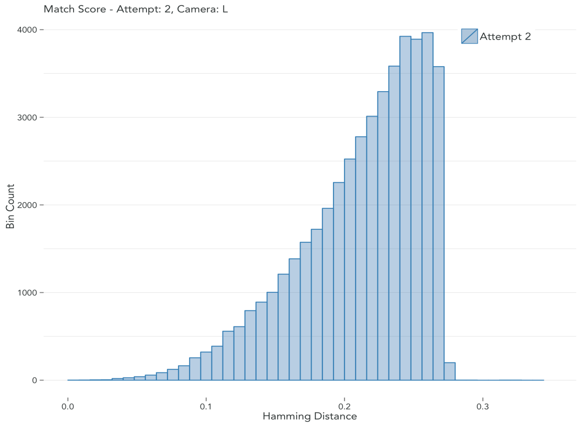

Figure 13: Distribution for second-attempt sensor L matches.

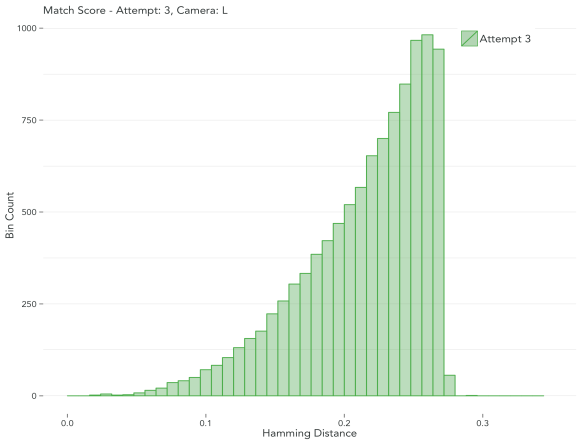

Figure 14: Distribution for third-attempt sensor L matches.

It is clear that match scores are not identically distributed across first, second and third attempts. This means that it is not appropriate to mix the scores together without distinction in a regression analysis, as done in IREX VI. Much like the differences in the left and right iris distributions, a regression analysis should split out the data by attempt number, or include the attempt number as a random effect in a mixed-effects model.

7. HD Distribution Across Kiosks

There are 69 different possible "Fake Kiosk ID" values that can be associated with transactions in the dataset. Each Fake Kiosk ID corresponds to a particular iris-recognition kiosk in the NEXUS program. The 69 kiosks are distributed across eleven different border-crossing locations: Toronto Terminal 1, Toronto Terminal 3, Ottawa, Vancouver, Montreal, Calgary, Edmunton, Halifax, Fort Erie, Winnipeg, and Billy Bishop. (The Billy Bishop Toronto City Airport is Canada's ninth-busiest airport, located on the Toronto Islands [16].)

In previous sections we considered the overall dataset as a composite of a left-iris dataset and a right-iris dataset, a composition of a sensor L dataset and a sensor B dataset, and as a composite of first-, second- and third-attempt datasets. Similarly, the dataset can be considered as a composite of datasets for the various kiosks.

Border-crossing traffic is not evenly distributed across the kiosks. Figure 15 shows this with the distribution of number of sensor B (same-sensor) transactions across the kiosks. The most frequently represented kiosk for both sensor B and L transactions is OK16, and 5 of the 10 most frequent kiosks for B are also in the 10 most-frequently-represented kiosks for sensor L transactions.

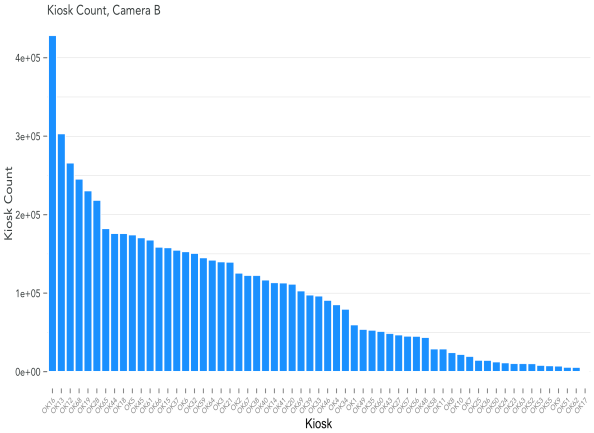

Figure 15: Total transactions per kiosk for the B transactions.

The relative frequency of use of different kiosks is not constant over time. Figure 16 shows the variation in kiosk usage for the top ten kiosks by number of transactions grouped by the year the transaction was recorded for years 2009, 2010, 2011, and 2012. Figure 16 demonstrates that there is also a variation in kiosk usage over time that an explanatory regression model and analysis should consider when examining time dependent relationships between covariates and the normalized HD.

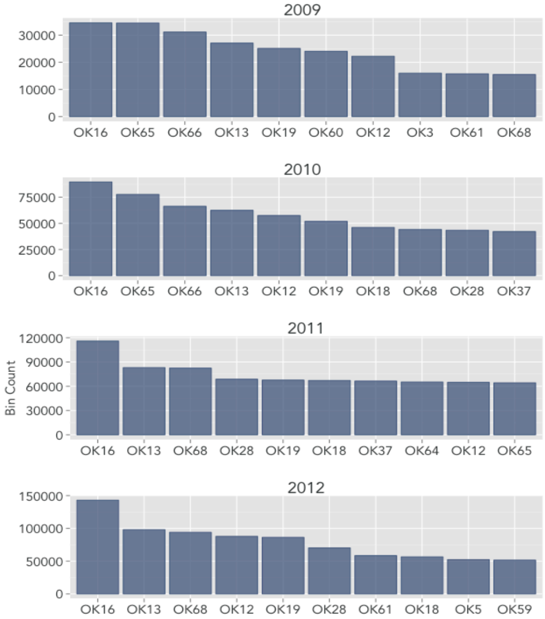

Figure 16: Ten most-frequently-used kiosks by year, 2009-11.

In addition to kiosk usage, Figure 17 displays the HD distribution for top ten most frequently used kiosk of sensor B. The distributions seem similar but not identical in shape across the kiosks. Small differences may be seen in the distributions shown in Figure 17. The distributions may vary between kiosks due to, for example, lighting differences that cause a difference in pupil dilation. This suggests that a proper regression analysis should take into account the kiosk at which a transaction takes place.

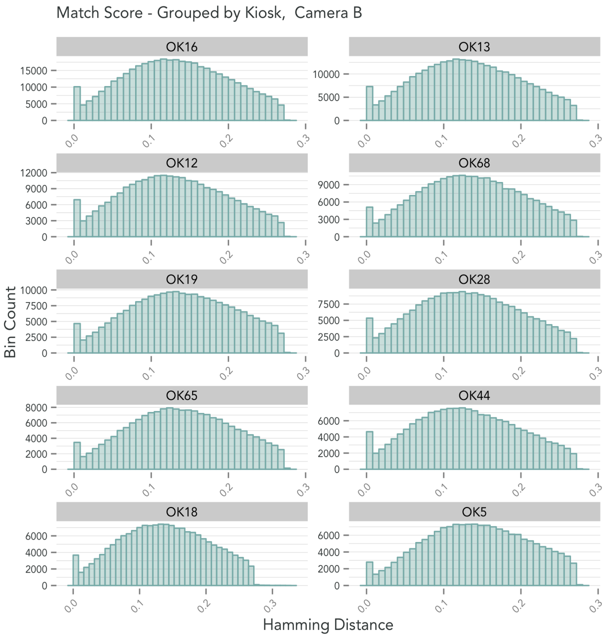

Figure 17: HDs for most-used kiosks for B sensor.

8. Discussion

The NIST IREX VI report [2] on iris recognition aging relies heavily on the "OPS-XING" dataset, which is drawn from logs recorded by the CBSA's operational border-crossing program [1]. We have obtained a dataset from CBSA that is a superset of the dataset analyzed in the IREX VI report. This dataset contains meta-data from the match score logs that was not discussed in the dataset analyzed in IREX VI. Our analysis documents a number of issues that were not brought to light in the IREX VI report.

One issue documented in our results that does not appear in IREX VI is that the distribution of match scores exhibits a "censoring" effect on the low end of the distribution. This is in addition to the well-documented "truncation" effect on the high end of the distribution. Both of these effects mean that the Hamming distance distribution does not follow a Gaussian distribution, as is assumed by the regression analysis performed in IREX VI.

Another issue documented in our results that does not appear in IREX VI is that the match scores for the right iris were collected under two different matching modes. These two different matching modes result in very different HD distributions. Additionally, the matching modes were used in different time periods of the operational scenario. Thus the mean right-iris Hamming distances has a strong source of variation over time that has nothing to do with iris recognition aging. This indicates that the underlying assumptions for a linear mixed-effects model do not support mixing the left and right iris data into the same regression model with an indicator variable in the fixed effects, as is done in IREX VI.

Our results also indicate that there is a substantial difference in the distributions of match scores between images acquired on a first attempt and those acquired on the second or third attempt. This issue is not examined or addressed in the IREX VI report. The HD distribution for images acquired on a second or third image acquisition is much worse than the distribution for first-attempt images. The IREX VI regression analysis implicitly assumes that the HD distribution is the same for first-, second- and third-attempt images.

One last issue documented in our results that does not appear in IREX VI is that the overall dataset is the result of transactions recorded at over 60 different kiosks distributed across 11 different border-crossing locations. The frequency of kiosk use varies between kiosks, and the pattern of variation between kiosks changes over time.

Our analysis demonstrates that there are multiple relevant factors in the OPS-XING dataset that are not taken into account in the IREX VI analysis. These factors suggest that the regression analysis used as an explanatory model in IREX VI is not appropriate.

Future work will include statistical tests to confirm the observed differences in HD distributions, an analysis of the change in mean HD over time for a cohort of subjects enrolled in the same year. In addition future work will examine a regression analysis applied to left iris match scores only instead of mixing left and right iris score data. And an additional analysis of variation in mean HD across kiosks will be examined in future research.

Acknowledgements

We would like to thank Michael Chumakov of the Canadian Border Services Agency for detailed explanations in response to numerous queries about the meaning of the meta-data in the NEXUS border-crossing logs.

References

[1] Canada Border Services Agency. Nexus. 2004-2014, http://www.cbsa-asfc.gc.ca/prog/nexus/menu-eng.html.

[2] P. Grother, J. R. Matey, E. Tabassi, G. W. Quinn and M. Chumakov, IREX VI: Temporal Stability of Iris Recognition Accuracy, NIST Interagency Report 7948, version dated July 24, 2013. http://biometrics.nist.gov/cs_links/iris/irexVI/irex_report.pdf

[3] M. Fairhurst and M. Erbilek. Analysis of physical ageing effects in iris biometrics. *IET Computer Vision* volume 5, pages 358–366, 2011.

[4] E. Ellavarason, and C. Rathgeb, Template Ageing in Iris Biometrics: A Cross-Algorithm Investigation of the ND-Iris-Template-Ageing-2008-2010 Database, Hochschule Darmstadt, Technical Report Nr. HDA-da/sec-2013-001, March 2013

[5] S. P. Fenker, Estefan Ortiz and K. W. Bowyer, Template Aging Phenomenon In Iris Recognition, *IEEE Access* 1, 266-274, May 16, 2013.

[6] A. Czajka, Influence of iris template ageing on recognition reliability, Communications in Computer and Information Science, Volume 452, 2014, pp. 294-299.

[7] R. Connaughton, A. Sgroi, K.W. Bowyer, P.J. Flynn, A multi-algorithm analysis of three iris biometric sensors, *IEEE Transactions on Information Forensics and Security* 7 (3), 919-931.

[8] Michael Chumakov, CBSA, personal communication.

[9] J. Daugman. New methods in iris recognition. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 37(5):1167–1175, October 2007.

[10] Joreskog, K. G. (2002, December, 3). Censored variables and censored regression. Retrieved March 16, 2015, from http://www.ssicentral.com/lisrel/techdocs/censor.pdf.

[11] R. Breen, ed. Regression models: Censored, sample selected, or truncated data. No. 111. Sage, 1996.

[12] T. Amemiya, "Tobit models: a survey." Journal of econometrics 24.1 (1984): 3-61.

[13] Truncated Regression, UCLA Institute for Digital Research and Education, http://www.ats.ucla.edu/stat/stata/dae/truncreg.htm

[14] T. Amemiya, "Regression analysis when the dependent variable is truncated normal." *Econometrica: Journal of the Econometric Society* (1973): 997-1016.

[15] J. Rawlings, S. Pantula, and D. Dickey. Applied regression analysis: a research tool. Springer-Verlag, 2nd edition, 1998.

[16] Billy Bishop Toronto City Airport, http://www.portstoronto.com/Airport.aspx