Biometric Aging

Effects of Aging on Iris Recognition

Author(s)

Keith Browning

Nicholas Orlans

Approved for Public Release

Distribution Unlimited

Case Number 13-3472

Sponsor: MITRE Innovation Program

Department No.: G020, G025

Project No.: 51MSR609-EA

Document No: MTR140308

Location : McLean, VA

Clarksburg, WV

MITRE

This page intentionally left blank.

Table of Contents

List of Figures

List of Tables

This page intentionally left blank.

Introduction

The fundamental question being asked is, does iris recognition performance remains constant and persist well over time? Or conversely, is the iris structure susceptible to irreversible changes that cause greater dissimilarity as time intervals between recognition events increase? Medical literature reports that the pupil size decreases slightly as people age [1]. However, the iris is naturally protected within the body, with protective coverings (eyelids and cornea) that prevent the physical surface "wear and tear", and loss of moisture experienced by fingerprints and the skin and soft tissue on the face. There are certainly disease and vision related ailments that occur with age; the eye can suffer presbyopia, cataracts and other diseases. But the fundamental *assumption* for iris recognition is that the structure and patterns in the elastic connective tissue of the iris remain stable and permanent throughout life.

The belief that iris features are permanent and stable was asserted by the originator of the first patented iris recognition algorithm. John Daugman [2] studied iris images from ophthalmologists spanning 25 years, and found no noticeable changes in iris patterns.[1] Daugman's patent [3] states "the iris of every human eye has a unique texture of high complexity, which proves to be essentially immutable over a person's life. No two irises are identical in texture or detail, even in the same person." As commercial iris sensors have only existed since the 1990s and significant collections for military or border management have only been around since the early 2000, there is limited longitudinal data has been available for study. Daugman's assertion has never been empirically validated against a large population observed over 5 years or more. This work is a step toward that goal.

Iris recognition has provided high speed and accurate biometric identification for variety of government applications. The permanence and invariance of biometric features over a significant span of time is an important property for identification systems. "Biometric aging" and "template aging" refer to the degradation of recognition performance over time. The assumption that there are no changes to the iris would make the iris ideal for essentially a life time of identification and/or authentication. If changes are observed, however, applications are prone to error over time, and more frequent re-enrollment would be required to maintain accuracy and reference to the most recent state of the iris.

A 2011 study by Fenker and Bowyer [4] suggests the eye may change over time. Fenker and Bowyer analyzed recognition scores observed over a two year time lapse in a study involving two different iris algorithms with 43 subjects. The study found that the authentic distribution shifted so as to increase the false non-match rate (FNMR) for long-term (two to three years) matches relative to short-term (less than six months) matches.

The Fenker and Bowyer study used the IrisBEE matcher [5], a Daugman-derivative iris algorithm with a hamming distance type score (where small values indicate greater similarity), and the Neurotechnology VeriEye matcher (where large values indicate greater similarity) [6]. The study found that the false reject rate increased by 157% at a threshold of 0.28 and increased 305% at 0.34. For the Neurotechnology VeriEye matcher, Fenker and Bowyer observed the false reject rate increased by 195% at a threshold of 30 and increased 457% at a threshold of 100.

[1] Note these images would not have not been near infrared images.

In an expanded 2012 study involving 322 subjects over a three year period, Fenker and Bowyer [7] reported a 153% increase in false non-match rate.

The Stability of Iris Matches over Time on DoD Data

Since 2003 to present, the Department of Defense (DoD) has collected iris images from base workers and detainees during contingency operations in Afghanistan and Iraq. The time span between collection of the same subject is not regularly spaced, rather the times are determined by security events such as job application and contract periods, detainee transfers, or background investigations.

A set of over seven million of the DoD iris images, indexed by subject using combined ten-print fingerprint and iris identification in the DoD Automated Biometric Identification System (ABIS), were rendered anonymous by extraction from the original submission transaction files. The image data set was provided, along with the identity index, to the National Institute for Standards and Technology (NIST) and MITRE for research and calibration purposes. A subset of the DoD/NIST 2004-2011 iris collection provided to MITRE contains 3.5 million images and represents 1.4 million different individuals. Of the 1.4 million individuals, the data set includes 285,616 people who were encountered two or more times where both iris pairs of iris images were available.

Laboratory collected data for academic studies generally does not contain compound effects from different device manufacturers and conditions, operator variability, and variable subject cooperation. In contrast, the operational data set included multiple challenging data quality issues that make the isolated measurement of aging a challenge. Dates of transaction capture (DOC) are (usually) automatically generated by the collection systems, but dates of birth (DOB) are more often entered manually, and contain errors. Examples are non-existent DOB (DOB = NULL), invalid dates (e.g., February 30, 1900), conflicting DOB dates for the same person, and conflicting estimated dates (e.g., 1/1/19xx DOB). As accurate dates are critical to longitudinal analysis, transactions that lacked plausible DOB and DOC values were excluded from the experiment.

For the purposes of this study a subset of 14,227 persons were selected where the individuals had been encountered at least three or more times and with time spans of at least three years. After data examination, transactions involving nine of these persons were eliminated due to malformed transactions, leaving 14,218 persons with three or more encounters. The set includes 892 individuals who were encountered three or more times over a time span of at least five years. A summary of the number of repeat encounters is shown in Table 1 below:

Table 1 – Number of Encounters

Number of Persons	Number of Encounters with Same Person
2	14
1	13
12	11
23	10
93	9
198	8
521	7
1064	6
2332	5
4323	4
5649	3
14218	**58754**

All pair-wise ordered combinations were considered for each subject. For example a subject with three encounters results in 3 choose 2 (or 3 pairs), and a subject with four encounters results in 4 choose 2 (or 6 pairs) for each subject. All the combinations for the test set resulted in 91,004 combinations of encounters (182,008 iris pair combinations), representing 117,508 distinct iris images.

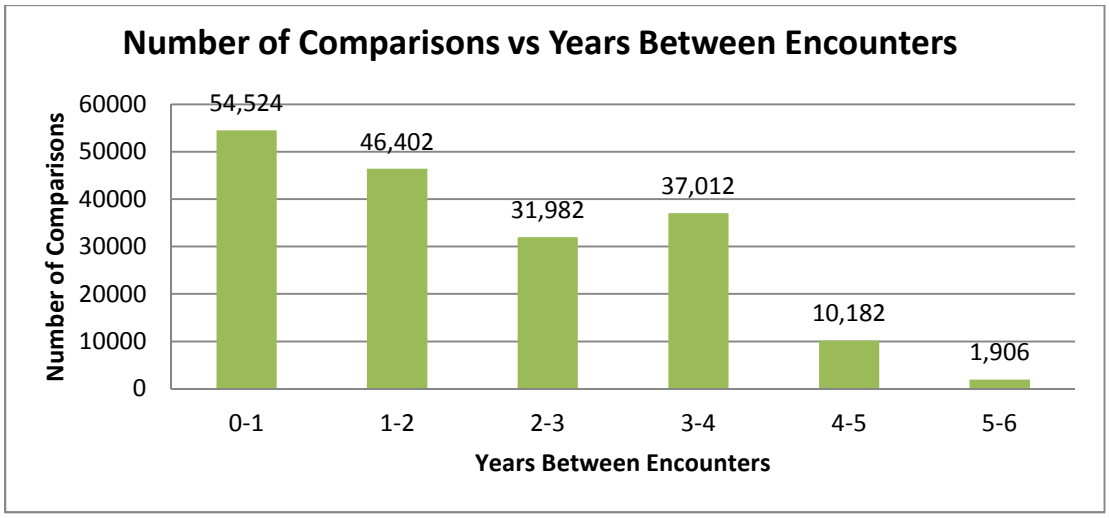

Figure 1 – Number of Comparisons vs. Years Between Encounters

Eye Position Designation Errors

Eye position (left or right) designations in the original data set were unreliable for at least 20% of the images. Therefore, MITRE staff performed human reviews to correct mislabeled iris positions and confirm left/right designation. The intent was to eliminate eye position designation errors as possible causes for false non-match results and also to isolate measurements for each eye. All same eye position iris images for the same person were also subjected to human review to confirm eye positions for all non-matching cases.

Agreement between high match scores for two iris algorithms and the human-reviewed designations was considered high confidence in matching left/right eye designations.

Multiple Iris Image Sources

Multiple iris imagers (Securimetrics PIER 4.2 and PIER 4.2, L-1 Identity Solutions HIIDE, CrossMatch SEEK, CrossMatch IScan, and others) were represented in the data set, used in multiple collection software application systems including the Biometrics Automated Toolset (BAT), the Biometric Identification System for Access (BISA), the L-1 Identity Solutions HIIDE, the Defense Manpower Data Center DBIDS, and the CrossMatch Technologies SEEK, LSMS, and MOBS systems). The different sensors introduce images of different sizes, and different illumination and quality characteristics. On average, the introduction of a new collection device could contribute to a change in pair-wise quality, and hence a change in performance.

Images Excluded for Noise Reduction

Compressed Images Excluded

Approximately 3% of the images in the data set used compression. Although many of the compressed images were successfully matched by both algorithms in the study, transactions containing compressed images were eliminated from the test data set in order to suppress noise and eliminate other possible causes of reduced algorithm performance.

Image Formats Excluded

Iris images with formats other than 640 pixel width, 480 pixel height, and 256 gray scale color were excluded from the study since one algorithm requires the 640 x 480 format.

Poor Quality Images Excluded

Iris images with margin violations (where the iris outer diameter overlaps the image frame edge), severe specular reflections, uneven illumination, poor contrast, defocus blur, motion defocus, or severe iris occlusion (eyelids or eyelashes) were eliminated from the data set without bias toward dates of capture. The purpose for eliminating these images was to remove image quality factors from the data and hence attempt to isolate potential aging effects.

Same Day False Non-Matches

Similar to the methodology employed by Fenker and Bowyer, same day comparisons were excluded in this study. A surprisingly high 6% false non-match rate was noted for images taken on the same day for the same persons.

Possible or observed reasons for the false non-matches include:

- Incorrect person enrolled for a known identity, but transaction retained

- Poor quality images recognized at the point of capture by the equipment handler resulting in a re-imaging attempt,
- Avoidance or non-cooperative behavior by the subject during iris image acquisition,
- Identification feedback from the multi-modal biometric system indicating an identification with one modality (e.g., fingerprint) and a non-identification result from a different modality (e.g., iris), resulting in a re-imaging attempt, or
- Enrollment on the same day in two different systems that could not exchange enrollment images due to distance, security domain, enrollment file incompatibility, or enrollment purpose.

The diversity of the sources of the DoD data set prevented case by case determinations of the causes of the same day non-matches. However, if one image of two taken on the same day is substandard, the same image would have the same likelihood of failing to match a high quality image taken in the future.

Cross-linked Identities

During analysis of false non-match results between pairs of transactions associated with the same person identity, 860 pairs were not matched by either algorithm for either eye position. It was suspected the transaction pairs represented two different individuals since the differences in the iris images were clearly beyond the changes likely to occur with aging effects, and were more plausibly cases of clerical errors, database errors, or errors in collection and identity verification processes. Ground truth identity determination was not possible beyond subjective human review. However, suspected cross-link cases represent less than 0.3% of the false non-match outcomes. Therefore, the suspect transaction pairs were not excluded from the overall results.

Failures to Enroll

Many of the DoD capture systems employ image quality metrics at the point of capture. Some capture systems also generate enrollment templates for local identification purposes, imposing an additional image quality confirmation before transmission. However, the available data set did not include any image quality scores. The data set also did not include any indicators associated with the failure of the image device to acquire an image for age-related medical conditions such as cataracts, glaucoma, corneal opacity, pupil dilation, pupil constriction, or droopy eyelids. Failure to enroll rate analysis based upon this data set is limited to template creation success or failure of the images that were captured and transmitted by the operational field collection systems to the DoD ABIS system. The extent of images that failed to create templates at the point of collection and hence never made it into the data set is unknown.

Age Relationship to Failure to Enroll (FTE) Rate

There were 116,805 iris images in the data set for which the metadata included both date of birth (DOB) and date of capture (DOC). An additional 174 images had invalid dates of

birth or dates of capture; therefore, age at time of capture could not be calculated. The overall FTE rate was 1.5% (1783 instances). Template generation failure rate was higher for 20-30 year olds than for older persons.

Table 2 – Failures To Enroll Grouped by Age Intervals

Age Bracket Total	*Number of Enrollment Attempts*	*Failures to Enroll*	*Fail Rate*	*Overall FTE Rate*
(0-15]	2	0	0.00%	
(15-20]	2562	38	1.48%	
(20-25]	15113	439	2.90%	
(25-30]	26120	490	1.88%	
(30-35]	22686	406	1.79%	
(35-40]	17967	204	1.14%	
(40-45]	14515	104	0.72%	
(45-50]	10063	56	0.56%	
(50-55]	5411	32	0.59%	
(55-60]	1801	6	0.33%	
(60-65]	395	6	1.52%	
(65-70]	120	0	0.00%	
70+	50	0	0.00%	
	116805	1781		1.5%

The highest FTE rate occurred in the 20-30 year age group. A possible explanation is this age group is likely to include many non-cooperative individuals (e.g., subjects being detained). There is a second increase in FTE for the 60-65 age group, but this is roughly half of the FTE for the 20-30 age group.

Analysis of Algorithms Used In Study

Two different algorithms were used for this study to corroborate findings. The algorithm names (or product names) are redacted to avoid generalized performance comparisons based on the specialized nature of this study. A summary of the overall performances of the two algorithms on the data set used in this study is given in Table 3 – Overall Matching Performance Summary.

Table 3 – Overall Matching Performance Summary

Algorithm	*Result Type*	*Count*	*Result %*
Algorithm A	Total Matches Possible	91247	
	Right or Left Iris True Match	89910	98.5%
	Left Iris True Match	87852	96.3%
	Right Iris True Match	87730	96.1%

	Right and Left Iris True Match	85672	93.9%
Algorithm B	Total Matches Possible	91247	
	Right or Left Iris True Match	89365	97.9%
	Left Iris True Match	88324	96.8%
	Right Iris True Match	88306	96.8%
	Right and Left Iris True Match	87265	95.6%

Algorithm A

Algorithm A is an algorithm that provides a hamming distance type score ranging from 0.0 to 1.0. A lower score indicates higher probability of a match. Genuine (same iris comparison) score distributions for short term (0-1 year time differences) are shown in Figure 2 – Algorithm A Genuine Pair Score Distribution for Short Term (< 1 yr.) Time Difference . Algorithm A produced non-match response with a hamming distance score of 1.0 for 3.9 % of the "genuine" iris image pairs. Algorithm A produced no false match results in 182,494 "imposter" iris image pair comparisons. Algorithm A produced no match scores between 0.34 and 1.0, likely due to a threshold operation to reduce the possibility of producing false matches. A threshold score of 0.34 was used for performance comparisons on genuine image pair subsets binned by time difference.

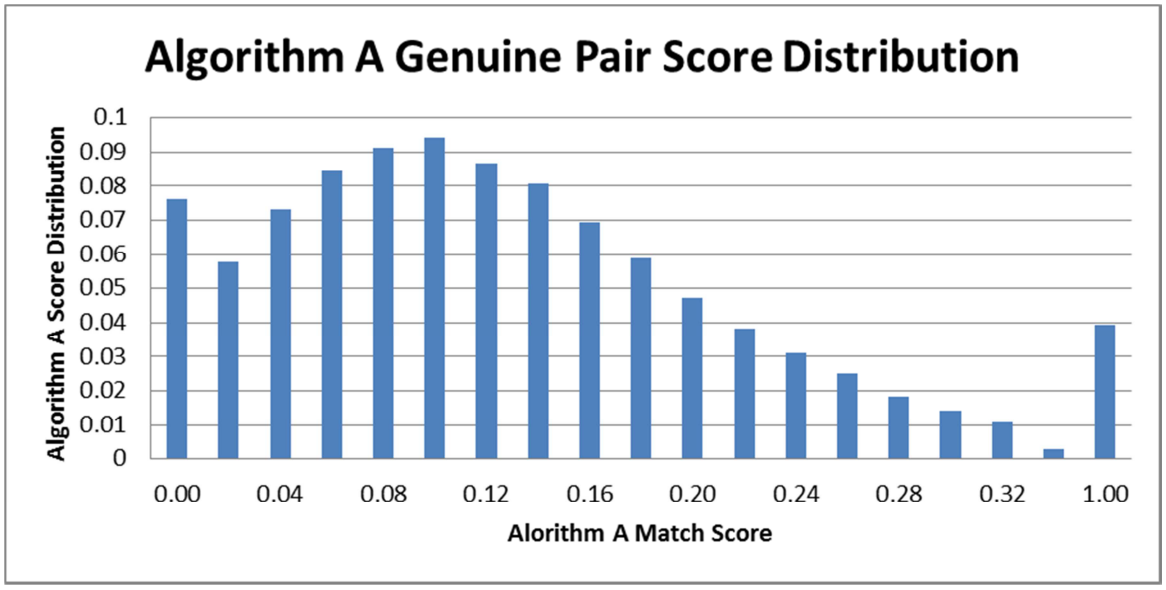

Figure 2 – Algorithm A Genuine Pair Score Distribution for Short Term (< 1 yr.) Time Difference

Score distributions for Algorithm A for additional data subsets of time difference do not appear to have any significant differences as shown in Figure 3 – Algorithm A Genuine Pair Score Distribution for Multiple Time Differences. The data subset for iris pair comparisons with time differences between 5 and 6 years appears to perform slightly better than the lower duration subsets, possibly due to a smaller available set of comparisons (1906 pairs of 182,008 total).

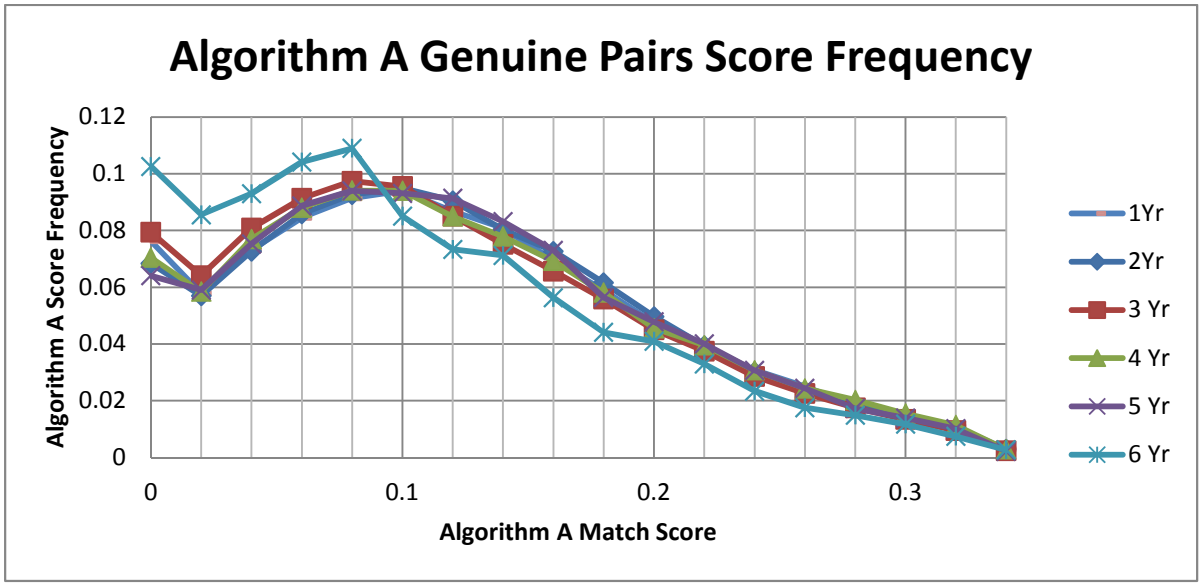

Figure 3 – Algorithm A Genuine Pair Score Distribution for Multiple Time Differences

A comparison of the cumulative false-non-match rates at the thresholds between 0.28-0.34 in Figure 4 – Algorithm A Cumulative False Non-Match Rate vs. Score by Time Difference Subset shows no correlation between overall identification failure rates between time difference subsets. The shortest time difference subset counter-intuitively had slightly higher non-match rates than the longer time difference subsets, and the longest duration subset had the lowest failure rates.

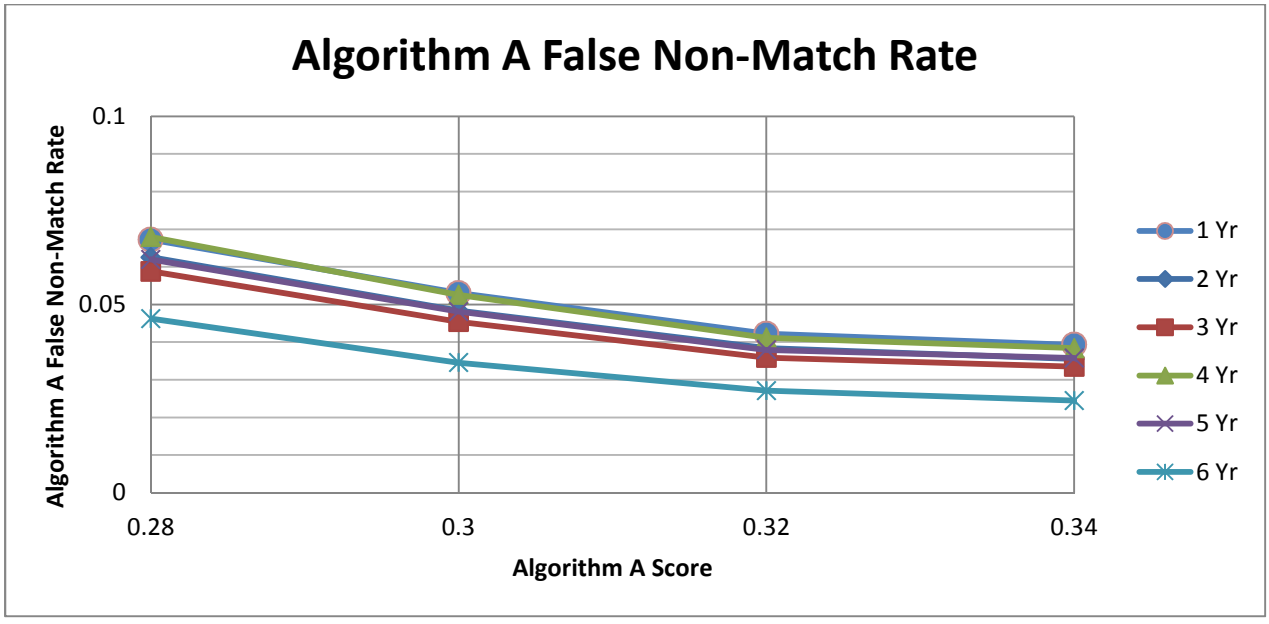

Figure 4 – Algorithm A Cumulative False Non-Match Rate vs. Score by Time Difference Subset

By picking a fixed threshold match score of 0.34 the false non-match rate for Algorithm A can be more easily compared for multiple time difference subsets as shown in Figure 5 – Algorithm A False Non-Match Rate at 0.32 Score vs. Time Difference (Yr.).

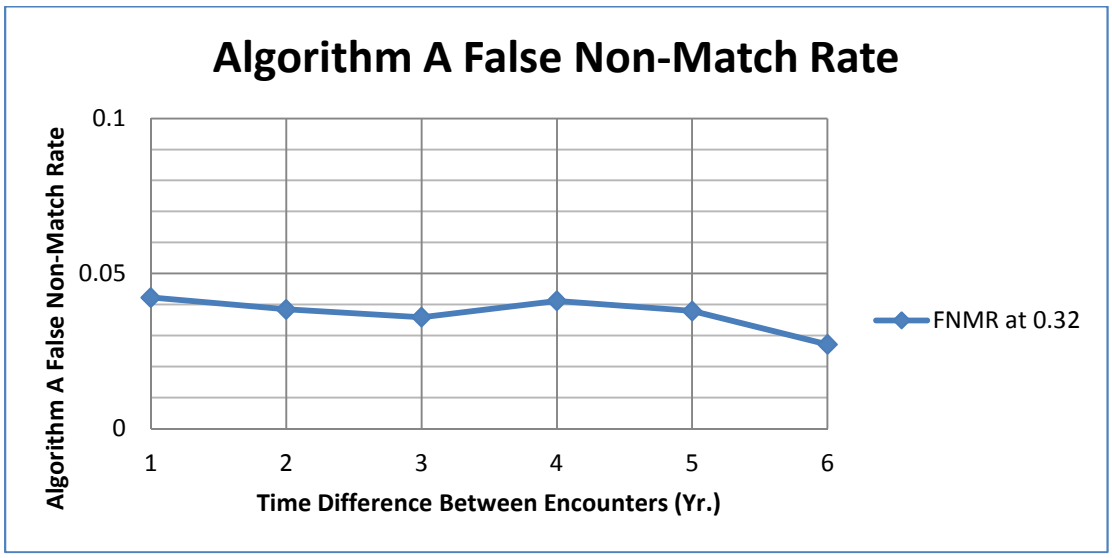

Figure 5 – Algorithm A False Non-Match Rate at 0.32 Score vs. Time Difference (Yr.)

Algorithm B

Algorithm B produces a match score between 0 and 1000 where higher scores indicate higher probability of a match. Algorithm B produced match scores below 80 for two imposter image pairs. A threshold score of 100 was used for performance comparisons on genuine image pair subsets binned by time difference.

Genuine (same iris comparison) score distributions for short term (0-1 year time differences) are shown in Figure 6 – Algorithm B Genuine Pair Match Score Distribution for Short Term (< 1 yr.) Subset. Algorithm B produced non-match response with a score less than 100 for 3.8 % of the "genuine" iris image pairs. Algorithm B produced no false match results with scores over 80 in 182,494 "imposter" iris image pair comparisons, but produced two false match results at scores between 50 and 80 for imposter iris image pairs composed of images of opposite iris positions for the same person.

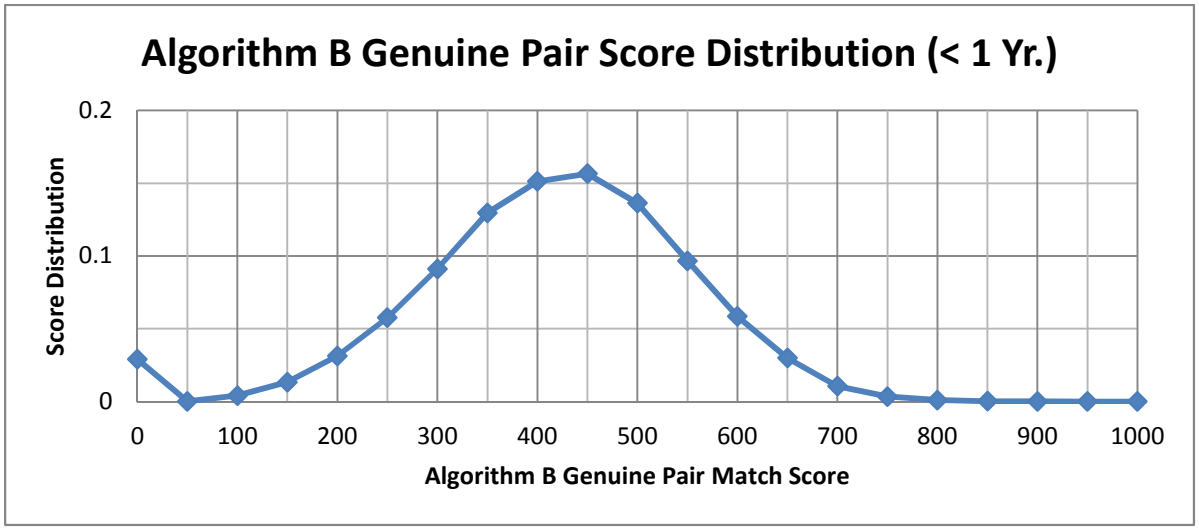

Figure 6 – Algorithm B Genuine Pair Match Score Distribution for Short Term (< 1 yr.) Subset

Score distributions for Algorithm B for additional data subsets of time difference do not appear to have any significant differences as shown in Figure 7 - Algorithm B Genuine Pair Score Distribution for Time Difference Subsets. The data subset for iris pair comparisons with time differences between 5 and 6 years appears to perform slightly better than the lower duration subsets, possibly due to a smaller available set of comparisons (1906 pairs of 182,008 total).

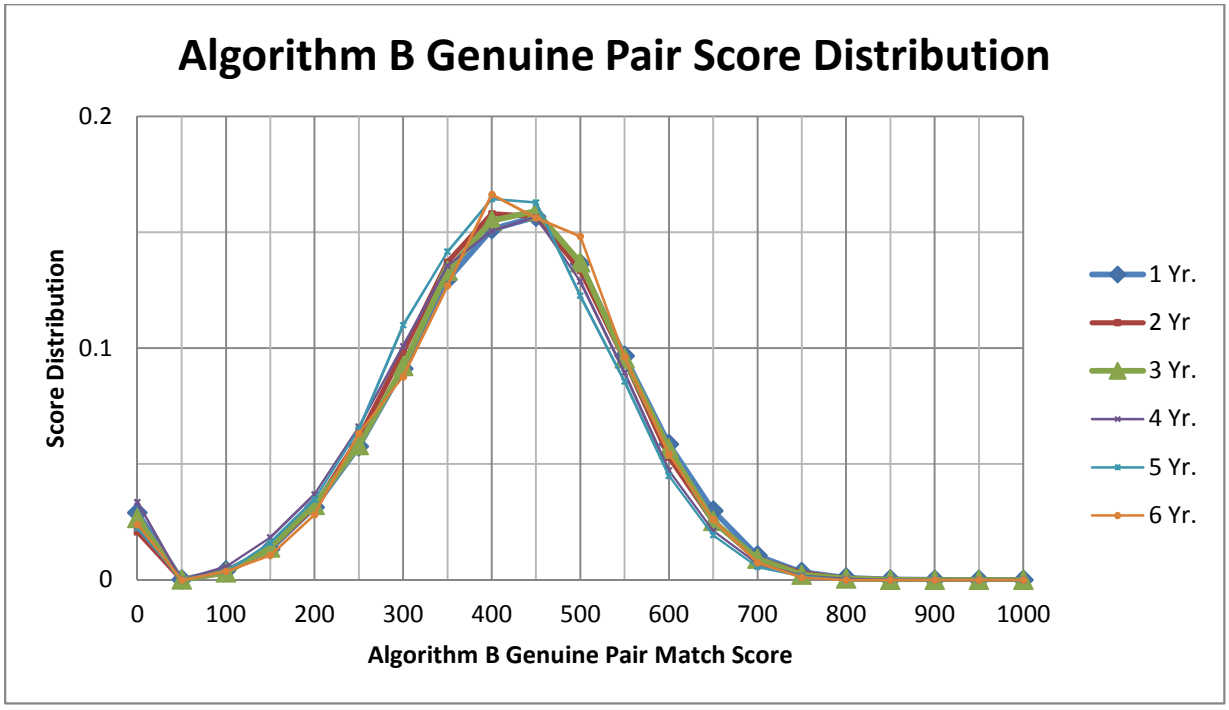

Figure 7 - Algorithm B Genuine Pair Score Distribution for Time Difference Subsets

The cumulative false non-match rate for Algorithm B with threshold scores in the range of 50 to 150 is shown in Figure 8 – Algorithm B Cumulative False Non-Match Rate at

Threshold Scores 0-150. Algorithm B does not exhibit monatomic relationship between false non-match rate and time difference between encounters for the data subsets in this study.

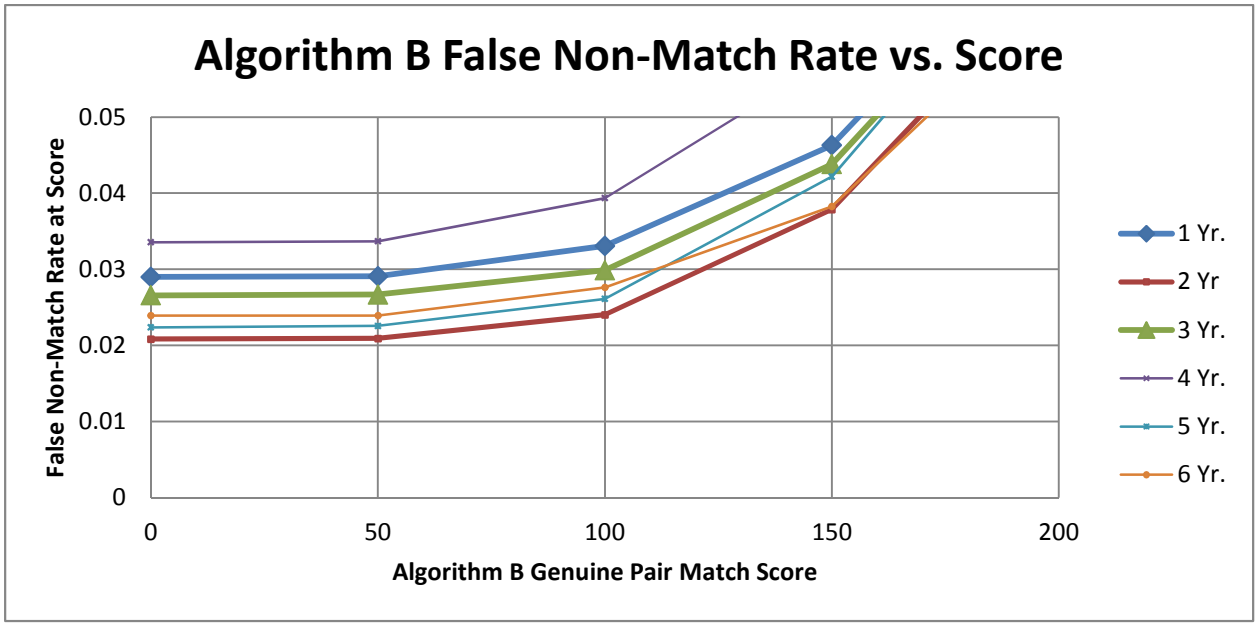

Figure 8 – Algorithm B Cumulative False Non-Match Rate at Threshold Scores 0-150

By picking a fixed threshold match score of 100 the false non-match rate for Algorithm B can be more easily compared for multiple time difference subsets as shown in Figure 9 – Algorithm B Genuine Pair False Non-Match Rate vs. Time Difference.

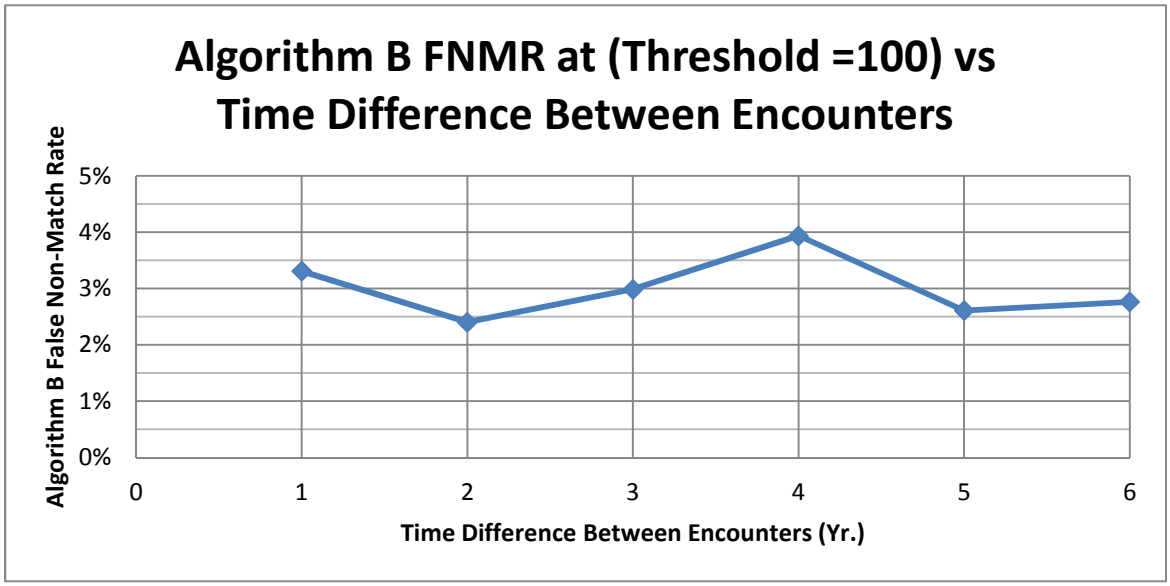

Figure 9 – Algorithm B Genuine Pair False Non-Match Rate vs. Time Difference

Match Score Trends per Individual

Match scores for genuine iris pairs were grouped per person and per eye location in order to determine whether decreasing match performance over time could be observed for a significant large subset of people in the study. The match scores for each person were compared for relative positive or negative trends in match scores over time. Continuous decreases in iris matching performance were observed for only a small fraction of the individuals in the test set. As the images from a small number (3-5) of individual changes were visually examined and these were believed to be due to medical conditions.

Conclusions

This study found no clear and compelling correlation between the time interval between recognition (alone) and iris matching performance. While some performance differences were observed on aggregated populations, they did not follow a clear trend of greater dissimilarity as the time intervals increased that would be consistent with "aging".

The DoD data is irregular and much less controlled than the Notre Dame laboratory collection, but it does span more total time and represents a significantly larger population than previous iris aging studies. The DoD data represented the greatest temporal span that was available to conduct the study.

Some effort was made to suppress noise (and isolate time), however, many primary performance factors are present in the DoD data, and hence the precise cause of each individual change in performance is not easily attributable. Other known, primary performance factors include the use of different sensors, different illumination, different gaze angles and occlusion, different pupil dilations, and compression.

A very small group of individuals may have suffered medical conditions that resulted in a continued reduction in performance with time; however, these were exceptions and not visible in the aggregated population.

References

[1] J. Birren, "Age Change in Pupil Size," *J. Gerontology,* vol. 5.3, pp. 216-221, 1950.

[2] J. Daugman, "Recognizing Persons by Their Iris Patterns," in *BIOMETRICS: Personal Identification in Networked Society*, Klewer Academic Publishers, 1999, pp. 103-121.

[3] J. Daugman, "Biometric personal identification system based on iris analysis". Patent U.S. Patent 5,291,560, 1 March 1994.

[4] S. P. Fenker and K. W. Bowyer, "Experimental Evidence of a Template Aging Effect in Iris Biometrics," in *IEEE Computer Society Workshop on Applications of Computer Vision*, 2011.

[5] P. Phillips, "FRVT 2006 and ICE 2006 Large-Scale Experimental Results," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 32, pp. 832-846, 2010.

[6] "NeuroTechnology VeriEye," NeuroTechnology, [Online]. Available: http://www.neurotechnology.com/verieye-technology.html. [Accessed 10 May 2013].

[7] S. P. Fenker and K. W. Bowyer, "Analysis of Template Aging in Iris Biometrics," in *IEEE Computer Vision and Pattern Recognition Workshop (CVPRW)*, 2012.