Presented at IEEE Computer Society Biometrics Workshop, June 17, 2012.

Analysis of Template Aging in Iris Biometrics

Samuel P. Fenker and Kevin W. Bowyer

Department of Computer Science and Engineering
Univ. of Notre Dame, Notre Dame IN 46556

sfenker@nd.edu, kwb@cse.nd.edu

Abstract

It has been widely believed that biometric template aging does not occur for iris biometrics. We compare the match score distribution for short time-lapse iris image pairs, with a mean of approximately one month between the enrollment image and the verification image, to the match score distributions for image pairs with one, two and three years of time lapse. We find clear and consistent evidence of a template aging effect that is noticeable at one year and that increases with increasing time lapse. For a state-of-the-art iris matcher, and three years of time lapse, at a decision threshold corresponding to a one in two million false match rate, we observe an 153% increase in the false non-match rate, with a bootstrap estimated 95% confidence interval of 85% to 307%.

1. Introduction

Mansfield and Wayman [1] define biometric template aging as follows – "Template ageing refers to the increase in error rates caused by time related changes in the biometric pattern, its presentation, and the sensor." While template aging has received substantial research attention for some biometric modalities, such as face [2,3], it has received little or no attention for other modalities, such as iris and fingerprint. This paper presents results of the most extensive experimental investigation to date of template aging for iris biometrics.

In their 1987 iris biometrics patent, Flom and Safir [4] asserted that the texture features of the iris were relatively stable over time – "… the significant features of the iris remain extremely stable and do not change over a period of many years." However, they also acknowledged that the iris texture is subject to change over time, and that re-enrollment after some period of time might be needed – "Even features which do develop over time … usually develop rather slowly, so that an updated iris image will permit ID for a substantial period ...". Daugman's 1994 iris biometrics patent [5] asserted that the iris texture is stable over a person's life – "The iris of every human eye has a unique texture of high complexity, which proves to be essentially immutable over a person's life."

We know of no studies with experimental results that support the conclusion that template aging does not occur for iris biometrics. Even so, the "essentially immutable over a person's life" view has been the prevailing view. It is commonly repeated in research papers, e.g., "… the iris is highly stable over a person's lifetime …" [6] and in popular references such as Wikipedia, "A key advantage … is … template longevity, as, barring trauma, a single enrollment can last a lifetime" [7].

We present results of experimental investigation into template aging in iris biometrics. We analyze datasets involving one, two, and three years of time lapse between acquisition of enrollment and verification images. We find that there is a template aging effect, and that it results in an increase in the false non-accept rate with increasing time since enrollment, over the length of time covered in our dataset. We use bootstrap methods to estimate 95% confidence intervals for the change in false non-match rate (FNMR) with increase in time lapse. For a state-of-the-art iris matcher, and three years of time lapse, at a decision threshold corresponding to a one in two million false match rate (FMR) we observe an 153% increase in the false non-match rate (FNMR), with a 95% confidence interval of 85% to 307%.

2. Related Work

An important element of related background is that human perception of similarity in iris texture patterns implies nothing about closeness of iris biometric match. Recent work by Hollingsworth et al. [8] demonstrates this. In one experiment, they showed subjects pairs of iris images that were either from identical twins or from unrelated persons. In another experiment, the pairs of images were either from the left and right eyes from the same person, or from unrelated persons. In both cases, subjects were highly accurate at categorizing pairs of images that belong together versus pairs that do not. This shows that humans can perceive a similar in the iris texture pattern between left and right eyes of the same person, or between eyes of identical twins. However, using the same images, the average biometric match score was no closer for identical twins, or the left and right eyes of same person, than it was for unrelated persons. This shows that humans perceive similarities in iris texture that are not reflected in iris biometric match scores. Thus human perception of a stable iris texture pattern over time does not necessarily imply anything about iris biometric template aging. Studies of iris biometric template aging must be done in the context of biometric match scores.

Baker et al. [9] published the first experimental results on iris biometric template aging. Their study involved 26 irises (13 persons) with images acquired over the time period 2004-2008 using an LG 2200 iris sensor. The LG 2200 was state-of-the-art at the beginning of their data acquisition, but its image capture technology allows for the possibility of interlace artifacts, and the LG 2200 is no longer marketed. Baker et al. used a version of the IrisBEE matcher that was distributed as a baseline matcher in the Iris Challenge Evaluation [10]. This matcher does not have state-of-the-art performance. They compared the authentic and impostor distributions for short-term and long-term matches. Short-term matches were between two images taken in the same academic semester but not on the same day. Long-term matches were between an image taken in spring of 2004 and one taken in spring of 2008. They found no significant change in the impostor distribution for long-term matches compared to short-term matches. However, they found that the authentic distribution for long-term matches shifted in a way that resulted in an increase in the false non-match rate. As an example, they reported that, "at a false accept rate of 0.01%, the false reject rate increases by 75% for long-time-lapse" [9].

Tome-Gonzalez et al. [11] report results comparing matches between iris images acquired in the same session with matches between images acquired with one to four weeks of time lapse. They use two different datasets, each acquired across four weekly sessions, using an LG 3000 sensor. (The LG 3000 sensor, like the LG 2200, is no longer marketed.) They use Masek's iris matcher implementation, which has relatively weak overall performance. At a false match rate of 0.01%, they report false non-match rates of 8.5% to 11.3% for within-session matches, versus FNMRs of 22.4% to 25.8% for across-session matches. This is essentially a same-session versus not-same-session comparison, rather than a longitudinal template aging experiment, but it does show that iris match quality depends on factors that may change with time.

Fenker and Bowyer [12] reported results of an iris biometric template aging study that compared authentic and imposter distributions for a set of 86 irises (43 persons) imaged over a two-year period. They compared short-term matches, between two images taken in one semester but not on the same day, with long-term matches, between an image taken in spring 2008 and an image taken in spring 2010. They found no noticeable change in the impostor distributions. However, they found that the authentic distribution shifted so as to increase the FNMR for long-term matches relative to short-term matches. For the IrisBEE matcher, they found that "The increase in false reject rate ranges from 157% at a threshold of 0.28 to 305% at 0.34" [12]. For the VeriEye matcher, they found that "The observed false reject rate increases from short to long time-lapse by 195% at a threshold of 30 and up to 457% at a threshold of 100" [12].

Rankin et al [13] studied variation in iris appearance over three imaging sessions at three-month intervals. However, they used visible-light illumination rather than near-infrared illumination. We are not aware of any commercial iris biometric systems that use visible-light illumination. The quality of their images and iris matching software was such that "Recognition failure was detected in 21% of intra-class comparisons cases overall, taken at both three and six month intervals" [13]. The combination of relatively short time lapse, use of visible-light images, and use of an iris matcher with poor absolute performance makes it difficult to draw any firm conclusions about iris biometric template aging based on these experiments.

Our results generally extend and improve over previous work in several ways. We use a larger dataset and a follow subjects for a longer period of time. Rather than contrasting just "short-term" and "long-term" results, we present results for multiple lengths of time lapse: one, two and three years. This allows an assessment of whether template aging is cumulative with additional time lapse. Also, rather than presenting only a point estimate of the change in the false non-match rate (FNMR), we use a bootstrap method to compute 95% confidence intervals. This allows an assessment of the degree of uncertainty involved in the estimate of the FNMR.

3. Experimental Materials

Iris images were acquired from spring of 2008 through spring of 2011, using an LG 4000 sensor, following the same acquisition procedure in the same laboratory. Overall, the experiments in this paper use images from 644 irises (322 subjects). The total number of images is 22,156, with 2312 coming from the year 2008, 5859 from 2009, 6215 from 2010, and 7770 images from 2011. The age range for the subjects is 20 to 64 years old. Of the 322 subjects, 177 are male and 145 are female; and 243 are Caucasian, 37 are Asian, 24 are other and 18 unknown. In the template aging experiments reported in this paper, we consider the dataset in terms of one cohort of subjects imaged over three years of time lapse, two cohorts imaged over two years, and three cohorts imaged over one year. (See Table 1.)

Table 1: Experimental Datasets By Period of Time Lapse

Time period	Number of subjects	Number of images	Avg. # short time-lapse matches	Avg. # long time-lapse matches
08-09	88	4,553	11986	30470
09-10	157	8,046	23882	54417
10-11	181	11,734	29120	97879
08-10	40	2097	5829	14282
09-11	124	8,082	18963	66849
08-11	32	2,338	5244	20888

We evaluated four iris biometric matchers available to us in order to select the best-performing one for use in our template-aging experiments. The first matcher is the IrisBEE matcher. The second is the VeriEye SDK (version 2.4) [14]. The third and fourth matchers are other commercially available systems; the specific products are not named here due to restrictions in the license terms for these software packages.

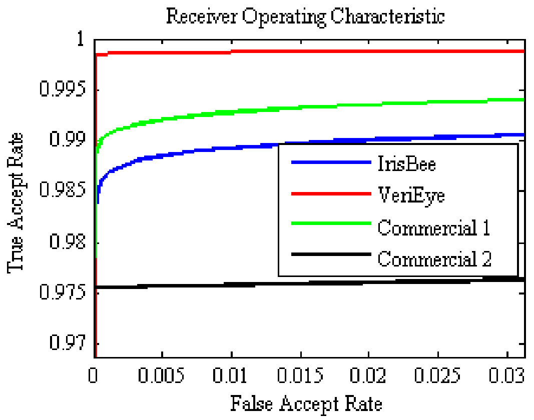

Figure 1. Selection of a best-performing matcher to use in template-aging experiments.

Using the entire image dataset described above, an all-vs-all matching experiment was performed with each of the four matchers. The results involve over 285 million comparisons. Figure 1 shows the ROC curves. The VeriEye matcher performs the best of the four matchers considered. Based on this result, the VeriEye matcher was selected for use in the remaining experiments.

Note that, unlike the typical Hamming distance matcher, the VeriEye matcher generates a similarity match score that ranges from 0 to 9443, where 0 is a non-match and 9443 is an exact match; that is, an image matched against itself.

4. Experimental Method

We define a short time-lapse match to be a comparison of two images acquired on different days but within a few months of each other. For the experiments described in this paper, we have a set of short time-lapse matches for images acquired in the spring semester of each of 2008, 2009 and 2010.

We define a long time-lapse match to be a comparison of two images acquired in different years. For example, we have a set of long time-lapse matches between an image acquired in spring 2008 and another image acquired in spring 2009. We have three different one-year time-lapse datasets, two different two-year time-lapse datasets, and one three-year time-lapse dataset.

We consider the datasets in terms of causal longitudinal groups. The short time-lapse spring 2008 dataset is the baseline for the one-year 2008-2009 dataset, for the two-year 2008-2010 dataset, and the three-year 2008-2011 dataset. The short time-lapse 2009 dataset is the baseline for the one-year 2009-2010 dataset and the two-year 2009-2011 dataset. The short time-lapse spring 2010 dataset is the baseline for the one-year 2010-2011 dataset. Table 1 summarizes the number of subjects, images, and short- and long-time-lapse matches in these datasets.

We compared the impostor distributions and the authentic distributions for each short time-lapse baseline to those of each of its long time-lapse datasets. We found that the impostor distributions show no noticeable difference between short and long time-lapse. That is, the impostor distribution for matches between images taken on average about one month apart does not differ noticeably from the impostor distribution for images taken on average about one, two or three years apart. However, the authentic distributions do show a noticeable change. The authentic distribution for a long time-lapse dataset has a generally higher FNMR than its corresponding short time-lapse authentic distribution. To investigate this effect, we compare the FNMR for the short and long time-lapse data, across a range of possible decision threshold values.

Also, we calculate a bootstrap 95% confidence interval [15] for the estimated change in FNMR. Given an experiment with N participating subjects, the bootstrap randomly selects with replacement N subjects from this pool. For each subject, with M corresponding short time-lapse comparisons and K long time-lapse comparisons, we randomly select with replacement M short time-lapse comparisons and K long time-lapse comparisons. This process is repeated 1000 times, and the distributions of these experiments are used to estimate confidence intervals for the change in FNMR. For each bootstrap sample, the FNMR is calculated for both the short- and long-time-lapse comparisons, and the change in FNMR is computed. The mean FNMR over 1000 bootstrap samples is calculated, and the confidence interval limits estimated using the 26th and 975th-ranked bootstrap samples.

Statistics are noted for the various datasets for a particular decision threshold value based on the false match rate (FMR) of the all-vs.-all experiment shown in Figure 1. For the VeriEye matcher, in general, at a low enough value for the decision threshold, the FMR is 100% and the FNMR is 0%. As the value for the decision threshold is increased, the FMR decreases and the FNMR increases. For the experiment shown in Figure 1, a decision threshold of 580 corresponds to an approximately 1 in 2 million FMR. At this threshold, there is sufficient false-non-match data in the various datasets to justify computing bootstrap estimates of the change in the FNMR between short and long time lapse. At the decision threshold corresponding to a 1 in 1 million FMR, there generally are too few FNM results to allow a reliable estimation of the FNMR.

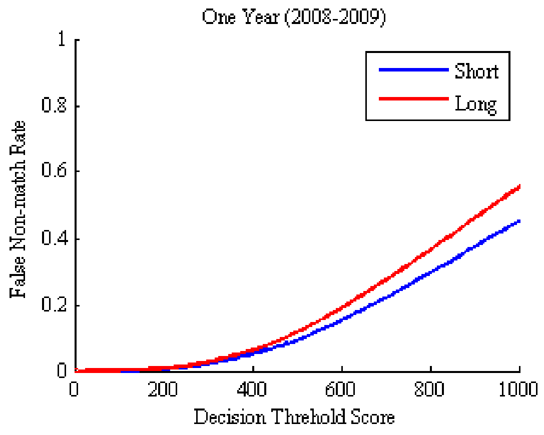

Figure 2 – Short- and Long-Time-Lapse FNMR for 2008-2009 Cohort.

5. Results

We first consider results for the three one-year time-lapse datasets, then for the two two-year time-lapse datasets, and then for one three-year time-lapse dataset.

5.1. One-Year Time Lapse Results

We have three datasets representing one-year time lapse. As summarized in Table 1, these represent 176 irises (88 subjects) imaged in 2008-2009, 314 irises (157 subjects) imaged in 2009-2010, and 362 irises (181 subjects) imaged in 2010-2011. The average number of days between acquiring a pair of images in the short-time-lapse match group is approximately 39 days, and the average number of days between acquiring a pair of images in the long-time-lapse match group is approximately 360 days. Thus the average “aging” between the two authentic distributions is approximately 11 months.

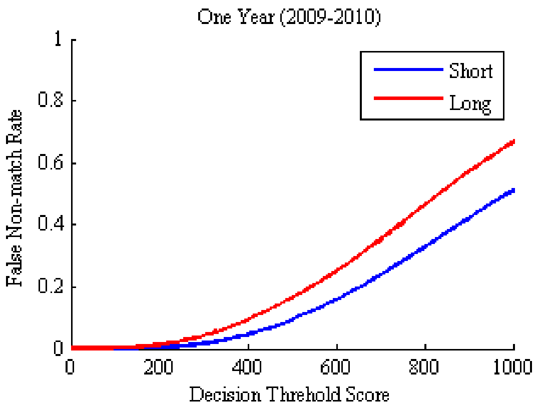

Figure 3 – Short- and Long-Time-Lapse FNMR for 2009-2010 Cohort.

The FNMR as a function of the decision threshold for the 2008-2009 dataset is shown in Figure 2. The FNMR starts at zero for a low decision threshold value and increases as the decision threshold increases. Correspondingly, the FMR starts out high and decreases as the decision threshold increases. It is clear from the curves in Figure 2 that the FNMR curves for short and long time-lapse begin to separate as soon as the FNMR is measurable and that the gap between them increases for higher decision thresholds.

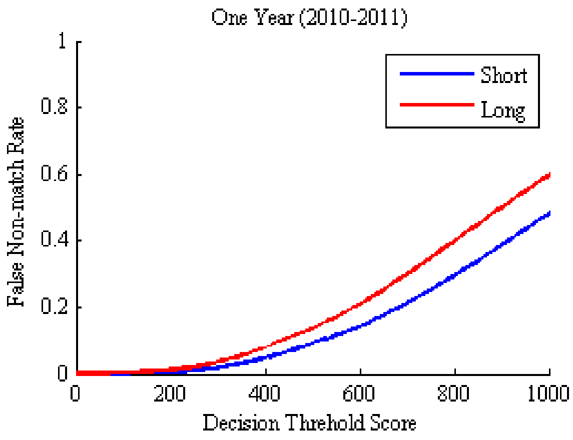

Figure 4 – Short- and Long-Time-Lapse FNMR for 2010-2011 Cohort.

At the decision threshold corresponding to 1 in 2 million FMR, the bootstrap mean increase in the absolute FNMR is 0.035, with a 95% confidence interval of 0.00723 to 0.0627. This corresponds to a mean percent increase in FNMR of 27%, with a confidence interval of 5% to 61%.

The FNMR as a function of the decision threshold for the 2009-2010 dataset is shown in Figure 3. Again, the FNMR curves separate as soon as the FNMR is measurable and the gap between them increases as the decision threshold increases (and so the FMR decreases). At the 1-in-2M-FAR threshold, the mean absolute increase in the FNMR is 0.086, with 95% confidence interval of 0.063 to 0.112. The corresponding percent mean increase in the FNMR is 60%, with a 95% confidence interval of 40% to 84%.

The FNMR as a function of decision threshold for the 2010-2011 dataset is shown in Figure 4. Again, the FNMR curves for short and long time lapse separate as soon as the FNMR is measurable, and the gap between them increases as the decision threshold increases. At the decision threshold value corresponding to a 1-in-2M-FMR, the mean absolute increase in FNMR is 0.063, with 95% confidence interval of 0.041 to 0.086. The corresponding percent increase in the FNMR is 49%, with a 95% confidence interval of 29% to 73%.

If no template aging effect exists, then the FNMR curves for the short and long time-lapse datasets should be essentially the same. Instead, in each of Figures 2, 3 and 4, the FNMR curve for long time lapse runs above that for short time lapse, and there is an increasing gap for higher decision threshold values. These results across the three different one-year time-lapse datasets clearly show the existence of a template aging effect. At a decision threshold corresponding to a 1-in-2M FAR rate, for the three different one-year datasets, the increase in the mean FNMR is estimated as 27%, 60% and 49%. In none of the three cases does the lower limit of the 95% confidence interval include zero change in the mean FNMR.

5.2. Two-Year Time Lapse Results

We have two datasets representing two-year time lapse. These represent 80 irises (40 persons) in the 2008-2010 dataset and 248 irises (124 persons) in the 2009-2011 dataset. The average number of days between acquiring a pair of images in the two-year time-lapse data is 732 days. Thus the average time-lapse is about 23 months.

The FNMR as a function of the decision threshold for the two-year, 2008-2010 dataset is shown in Figure 5. As is the case with the three one-year datasets, the FNMR curves separate as soon as the FNMR is measurable and the difference between them grows with increasing decision threshold value. At the decision threshold corresponding to a 1-in-2M FMR, the mean absolute increase in the FNMR is 0.11, with a 95% confidence interval of 0.056 and 0.168. This corresponds to a mean percent increase in the FNMR of 82%, with a 95% confidence interval 38% to 150%.

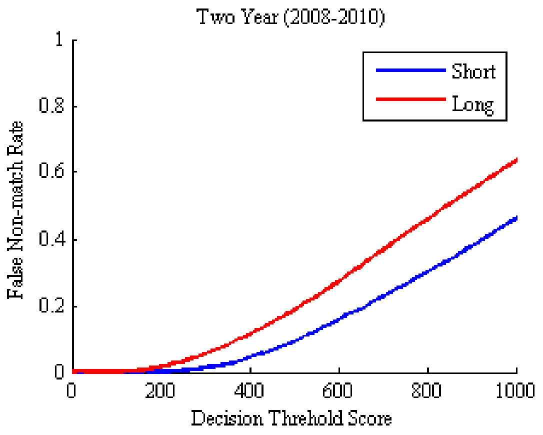

Figure 5 – Short- and Long-Time-Lapse FNMR for 2008-2010 Cohort.

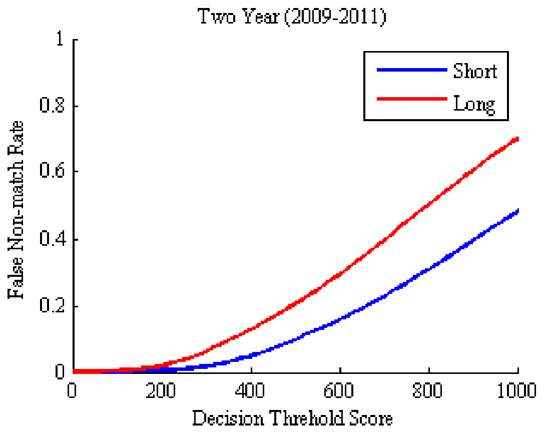

Figure 6 – Short- and Long-Time-Lapse FNMR for 2009-2011 Cohort.

The FNMR as a function of the decision threshold for the FNMR for the two-year, 2009-2011 dataset is shown in Figure 6. Again, the FNMR for long time-lapse is consistently greater than that for short time-lapse. At the decision threshold corresponding to a 1-in-2M FMR, the mean absolute increase in the FNMR is 0.13, with a 95% confidence interval of 0.095 to 0.167. This corresponds to a mean percent increase in the FNMR of 91%, with a 95% confidence interval of 63% to 127%.

For each of the two two-year time-lapse datasets, as was the case with each of the three one-year time-lapse datasets, the evidence for a template aging effect is clear. For each of the two two-year datasets, the FNMR for long time-lapse is consistently greater than for short time-lapse, across the broad range of possible decision threshold values. At the decision threshold corresponding to a 1-in-2M FMR, the mean FNMR for the two datasets is estimated at 82% and 91%. Each of these two values is greater than any of the three values for the three one-year time-lapse datasets. And, for either of the two-year datasets, the lower limit of the 95% confidence interval does not include zero change in the FNMR.

5.3. Three-Year Time Lapse Results

There are 64 irises (32 subjects) in the 2008-2011 time-lapse dataset. The mean time between image acquisitions for the two images in a long-time-lapse match is 1,068 days. Thus the average time lapse is about 34 months.

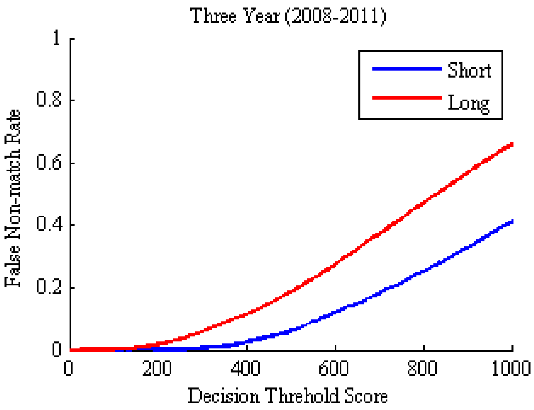

Figure 7 – Short- and Long-Time-Lapse FNMR for 2008-2011 Cohort.

The FNMR as a function of the decision threshold for the three-year, 2008-2011 dataset is shown in Figure 7. Again, as with the one-year and two-year datasets, the FNMR for long time-lapse is consistently greater than it is for short time-lapse. At the decision threshold corresponding to a 1-in-2M FMR, the mean absolute increase in the FNMR is 0.147, with a 95% confidence interval of 0.088 and 0.208. This corresponds to a mean percent increase in the FNMR of 153%, with a 95% confidence interval 85% to 307%. This mean percent increase for three-year time lapse is greater than that for either of the two-year datasets.

5.4. Longitudinal View of 2008 to 2011

. We can also consider the three-year, 2008-2011 results in the context of the one-year, 2008-2009 results and the two-year, 2008-2010 results. These three sets of results share the same short time lapse authentic distribution based on images from spring 2008. The long time lapse authentic distributions represent the spring 2008 subjects who also had iris images acquired in spring 2009, 2010 and 2011, respectively.

Table 2. FNMR Change from 2008-2009 to 2008-2011.

Time period	% Change in FNMR (95% CI)
2008-2009	27% (5%, 61%)
2008-2010	82% (38%, 150%)
2008-2011	153% (85%, 307%)

For the 2008-2009 cohort, the mean percent increase in FNMR is 27%, with a confidence interval of 57% to 61%. For the 2008-2010 cohort, the mean percent increase in the FNMR is 82%, with a 95% confidence interval 38% to 150%. For the three-year, 2008-2011 cohort, the mean percent increase in the FNMR is 153%, with 95% confidence interval for the percent increase is 85% to 307%. This data, summarized in Table 2, clearly shows a pattern of increasing FNMR with increasing time lapse.

5.5. Longitudinal View of 2009 to 2011

The short time lapse authentic distribution for the 2009-2010 and 2009-2011 datasets is composed of matches between images of the same iris taken on different days in spring 2009. The 2009-2010 results compare this authentic distribution to that for matches between an image from the spring of 2009 and an image from the spring of 2010. At a decision threshold representing a 1-in-2M-FMR, this shows a 60% increase in the FNMR. The 2009-2011 results compare to the authentic distribution for matches between an image from spring 2009 and an image from spring 2011. At the same decision threshold, this data shows a 91% increase in the FNMR. Again, the data clearly shows an increasing FNMR with increasing time lapse.

Table 3. FNMR Change from 2009-2010 to 2009-2011.

Time period	% Change in FNMR (95% CI)
2009-2010	60% (40%, 84%)
2009-2011	91% (63%, 127%)

6. Summary and Discussion

Iris biometric technology is currently in successful operational use in large-scale applications, such as the UAE entry watch list [16]. India is on the way to enrolling over one billion persons in a biometric identification system that uses iris and fingerprint [17]. As iris biometric applications continue to increase and grow, it is important that the basic phenomena of iris matching become better understood. In this paper, we have presented the most detailed experimental analysis to date of the phenomenon of template aging in iris biometrics.

Using a large dataset of iris images acquired over a three-year period, we have analyzed cohorts of irises with images acquired over one, two and three years. We find clear and conclusive evidence that template aging does occur in iris biometric matching. Specifically, the experimental evidence indicates that the false non-match rate increases with increasing time between acquisition of the enrollment image and the image to be recognized. In our results, the false non-match rate increases by greater than 50% with two years of time lapse.

The demonstration that template aging does occur for iris biometrics does not necessarily hinder the deployment of practical iris biometric systems. In situations where the biometrics is used for access, the user may simply need to be re-enrolled in the system after some determined period of time. Moreover, once the fact of template aging for iris biometrics is acknowledged, research effort may be focused on reducing the magnitude of the effect. By analogy, research in face recognition has focused on reducing the effects of face aging for some time and in the results of the recent NIST Multiple Biometric Evaluation it appears that substantial progress has been made by one face recognition system [18].

To facilitate further research on iris biometric template aging, the iris image dataset for the 2008 through 2010 images used in this study is available to the research community. Additional information and the release agreement for this dataset are available at http://www.nd.edu/~cvrl/.

References

[1] A.J. Mansfield and J.L. Wayman, Best Practices in Testing and Reporting Performance of Biometric Devices, Aug 2002.

[2] A. Lanitis, A survey of the effects of aging on biometric identity verification, *International Journal of Biometrics* 2 (1), 2010, 34 -52..

[3] N. Ramanathan, R. Chellappa, and S. Biswas, Computational methods for modeling facial aging: A survey, *Journal of Visual Languages and Computing*, 20 (3), June 2009.

[4] L. Flom and A. Safir Iris Recognition System, U.S. Patent 4,641,349, 1987.

[5] J. G. Daugman, Biometric personal identification based on iris analysis, U.S. Patent 5,291,560, 1994.

[6] D. Monro, S. Rakshit, D. Zhang. "DCT-Based Iris Recognition," *IEEE Trans. PAMI*, vol. 29 (4), pp. 586-595, Apr. 2007.

[7] Iris Recognition, Wikipedia, March 17, 2011.

[8] K. Hollingsworth, Genetically identical irises have texture similarity that is not detected by iris biometrics, *Computer Vision and Image Understanding* 114, 1493-1502, 2011.

[9] S. Baker, K. W. Bowyer, and P. J. Flynn, "Empirical Evidence for Correct Iris match Score Degradation With Increased Time-Lapse Between Gallery and Probe Matches," in *Proc. Int. Conf. on Biometrics*, pp. 1170-1179, 2009.

[10] National Institute of Standards and Technology. Iris Challenge Evaluation, 2006, http://www.nist.gov/itl/iad/ig/ice.cfm.

[11] P. Tome-Gonzalez, F. Alonso-Fernandez and J. Ortega-Garcia, On the effects of time variability in iris recognition, *IEEE International Conference on Biometrics Theory, Applications and Systems*, 2008.

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

[13] D. Rankin, B. Scotney, P. Morrow and B. Pierscionek. "Iris recognition failure over time: the effects of texture", *Pattern Recognition* 45, 145-150, 2012.

[14] http://www.neurotechnology.com/verieye.html. Accessed November 2011.

[15] M.E. Schuckers, Computational Methods In Biometric Authentication, Springer, 2010.

[16] A.N. Al-Raisi and A.M. AL-Khouri, Iris recognition and the challenge of homeland and border control security in UAE, *Telematics and Informatics* 25, 117-132, 2008.

[17] Unique Identification Authority of India. http://uidai.gov.in.

[18] Multiple Biometric Evaluation: Report on the Evaluation of 2D Still-Image Face Recognition Algorithms, NIST IR 7709, 2010.