Empirical Evidence for Correct Iris Match Score Degradation with Increased Time-Lapse Between Gallery and Probe Matches

Sarah E. Baker, Kevin W. Bowyer, and Patrick J. Flynn,
sbaker3, kwb, flynn@cse.nd.edu

University of Notre Dame

Abstract. We explore the effects of time lapse on iris biometrics using a data set of images with four years time lapse between the earliest and most recent images of an iris (13 subjects, 26 irises, 1809 total images). We find that the average fractional Hamming distance for a match between two images of an iris taken four years apart is statistically significantly larger than the match for images with only a few months time lapse between them. A possible implication of our results is that iris biometric enrollment templates may undergo aging and that iris biometric enrollment may not be "once for life." To our knowledge, this is the first and only experimental study of iris match scores under long (multi-year) time lapse.

Key words: iris biometrics, enrollment template, template aging, time-lapse, match distribution stability

1 Introduction

The iris biometrics research community has accepted the premise that the appearance of the iris is highly stable throughout most of a person's life. Daugman stated the assumption this way-"As an internal (yet externally visible) organ of the eye, the iris is well protected and stable over time"[1]. The assumption is repeated in similar form in recent academic references: "[the iris is] stable over an individual's lifetime‘"[3], "the iris is highly stable over a person's lifetime"[5], "[the iris is] essentially stable over a lifetime"[4]. While the basic assumption is broadly accepted as valid and commonly re-stated, we know of no experimental work that establishes its validity. This paper describes our experimental evaluation of the extent to which this assumption is true in terms of practical application in biometrics.

We formulate an experimental test of the long-term stability of iris texture in iris biometrics as follows. Assume that a person has an iris image acquired at one point in time for enrollment, and at a later point in time has another image acquired for recognition. The result of matching the two iris images is reported as a fractional Hamming distance, a value between 0 and 1 that indicates the fraction of iris code bits that do not match. A fractional Hamming distance of

0 indicates a perfect match, and a distance of 0.5 indicates random agreement. The "stable over a lifetime" concept can be tested by comparing the Hamming distance of image pairs acquired with different time lapses.

To investigate this question experimentally, we use a set of iris images acquired at the University of Notre Dame [6][10][8], and a modified version of the open source "ICE baseline" iris code matcher[7][9][13]. Comparing matching scores between images taken a few months apart with scores between images taken approximately four years apart, we find that there is a statistically significant difference in the average Hamming distance between short-time-lapse matches and long-time-lapse matches. This suggests that the "lifetime enrollment" concept may not be valid. This would also suggest that time lapse between images should factor into a decision about match quality, and that guidelines are needed for time between re-enrollment.

1.1 Related Work

Gonzalez et al. report an effect of time separation on iris recognition [11] that may initially seem similar to this paper. However, their work is based on comparing matches between images acquired at the same acquisition session with those acquired with at most three months time lapse. They report a higher match statistic for images from the same session than those across sessions. They note little change in match statistics when comparing matches with short time lapses, between two weeks and three months. In this paper, we eliminate matches between images acquired at the same session as we expect they would be unfairly similar. Additionally, we focus on the effect of time-lapse between gallery and probe images and same-session images are not used as both the gallery and the probe in a real world scenario. We do not note significant differerences in average Hamming distance for images with a few months time lapse. However, at four years time lapse, we do observe a significant difference.

2 Experimental Methods and Materials

2.1 Experimental Materials

The iris images analyzed in this study were acquired using an LG 2200[2], and the acquisition protocol is the same as that used in the collection of images for the Iris Challenge Evaluation[8]. A small subset of people have participated in data collections from spring of 2004 through spring of 2008. We know of no other iris image data set that has four years of time-lapse data available.

Our data set consists of images acquired approximately weekly during each academic semester. At each acquisition session, six images of each iris are acquired from each subject. Some images were discarded from our data set due to poor quality.

We compare two types of matches: (1) matches between two images both acquired in the same semester (but not on the same day) and (2) matches between one image from spring 2004 and one image from spring 2008. We found 13 subjects in the data set with both spring 2004 and spring 2008 images of each iris. For these 26 "iris subjects", we used 1236 images from 2004 and 573 images from 2008 for a total of 1809 images. This data set contains eight males and five females between the ages of 24 and 56. Three of these subjects are Asian and ten are Caucasian. Four of these subjects wore contacts and nine did not; no subjects wore glasses for this acquisition.

All images used in our experiments were acquired by the same LG2200 camera. They were also acquired in the same studio using the same acquisition procedure, computer system, digitizer board, driver software, and application software[6][10].

Our iris segmentation technique employs encoding and matching, we used software based on the open source IrisBEE[8]. This software uses one dimensional log-Gabor wavelets to create a 240x20x2-bit iris code and contains improvements to the segmentation as described in [6].

2.2 Experimental Method

Our null hypothesis and alternative hypothesis are stated as follows.

Null Hypothesis: The fractional Hamming distance for iris code matches between images taken a longer time apart is not greater than that for matches between images taken a shorter time apart.

Alternative Hypothesis: The fractional Hamming distance for iris code matches between images taken a longer time apart is greater than that for matches between images taken a shorter time apart.

We consider two experimental scenarios to test the null hypothesis, an "all-irises" test and an "iris-level" test. The experimental results and conclusions are similar for both formulations. The "all-irises" scenario combines the set of images from all 26 "iris-subjects" and is explained as follows.

For each iris we have multiple images. Each such image is considered as a gallery image in succession. For each gallery image, all other images are considered as probe images. Each match between a gallery and a probe image results in a Hamming distance. This HD is placed in either a short-time-lapse set or a long-time-lapse set, depending on the time elapsed between the gallery and the probe image. The process is repeated for every image of that iris subject, yielding a set of short-time-lapse HDs and a set of long-time-lapse HDs. These sets are each averaged, yielding a short-time-lapse mean HD and a long-time-lapse mean HD for that iris subject.
We introduce the following notation:

We have a set of iris images: $\mathcal{I} = \{I_1, I_2, \ldots I_n\}$
Each image in our set has a subject ID including a left-right indicator and a date:
$\forall\ I \in \mathcal{I}$, I.id = SubjectID (i.e. 02463L)
I.date = Date of Image
For each unique subject S,
$\mathcal{I}_S = \{I \in \mathcal{I} \mid I.id = S\}$
For each $I \in \mathcal{I}_S$, we obtain the set of images within a short-time lapse, $\mathcal{I}_{S*S}$:

$\mathcal{I}_{S*S} = \{I' \in \mathcal{I}_S \mid |I'.date - I.date| < T_d\}$

and we obtain the set of images taken after a long-time lapse, $\mathcal{I}_{S*L}$:

$\mathcal{I}_{S*L} = \mathcal{I}_S - \mathcal{I}_{S*S}$

where T_d is a time difference threshold. We use $T_d = 6$ months.

We also define sets of Hamming distances as follows:

$\mathcal{D}_{S*S} = \{HD(I, I') \mid I \in \mathcal{I}_S,\ I' \in \mathcal{I}_{S*S}\}$

$\mathcal{D}_{S*L} = \{HD(I, I') \mid I \in \mathcal{I}_S,\ I' \in \mathcal{I}_{S*L}\}$

$\mu_{S*S} = \frac{\sum \mathcal{D}_{S*S}}{||\mathcal{D}_{S*S}||}$(mean short-time-lapse match score)

$\mu_{S*L} = \frac{\sum \mathcal{D}_{S*L}}{||\mathcal{D}_{S*L}||}$(mean long-time-lapse match score)

The difference between the means $(\mu_{S*L} - \mu_{S*S})$ is computed, and the process is repeated for every iris subject, yielding a set of differences between mean HDs.

We consider two tests of the null hypothesis using these differences. For the sign test, we consider the null hypothesis that a positive difference occurs equally as often as a negative difference. The alternative hypothesis is that the more prevalent, a positive difference, occurs more often. Using a one-tailed Student's t test on the difference of means, we consider the null hypothesis that the mean of the N differences is zero. The alternative hypothesis is that the mean of the differences is greater than zero.

The "iris-level" scenario involves tests performed on each iris separately, yielding 26 different p values. For each iris subject, S, the short-time-lapse set, $\mathcal{D}_{SS}$, and the long-time-lapse set, $\mathcal{D}_{SL}$, used in the "all irises" experiment give two samples of HDs. To test our null hypothesis, we test if these two samples are from a distribution with $\mu_{SS} = \mu_{SL}$.

2.3 Possible Sources of Change in Match Quality

We consider four factors other than time-lapse that itself could conceivably cause poorer quality matches with longer time lapse.

1. The number of bits used in comparisons can affect the match distribution. If two images are masked in such a way that few bits are left to be used in the comparison, the Hamming Distance may be lower than it ought to be[12]. To control for differences in the number of bits used in a match, we implemented score normalization as suggested by Daugman[12]. Across our data, 5400 was the average number of bits used and was used as the scaling parameter in the normalization step.

2. It has been shown that the pupil to iris ratio affects the match distribution[13]. When two images of irises with largely dilated pupils are compared, the Hamming distance is greater than two irises with less dilated pupils. Similarly, as the difference in dilation between the two irises increases, the match distribution shifts in the positive directiont[13]. To account for any effects of pupil dilation, we consider the difference in pupil dilation between irises as a factor in the experiment below.

3. The presence of contact lenses can adversely affect match quality[15]. We performed a manual, retrospective check for contact lenses in all images used in this study. Four subjects wore contact lenses in both years and nine did not wear them in either year. No subjects appear to have begun to wear contacts in 2008 when they did not in 2004, or to have changed the type of contacts they wore.

4. Poor image quality and segmentation affect match quality[16]. We manually inspected every image and its segmentation produced by our segmenter. Approximately 7% of the images acquired for these subjects were discarded due to poor quality and an additional 24% were discarded due to poor segmentation.

3 Results

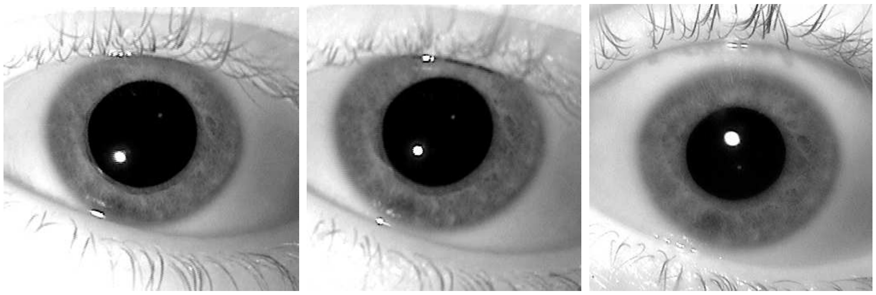
For every iris-subject, we computed the mean Hamming distance and the standard deviation for the short-time-lapse matches and the long-time-lapse matches. In 23 of the 26 irises, μ_{SL} was greater than μ_{SS} matches. The difference in mean HDs for the two sets of time lapse, $\mu_{diff} = \mu_{S*L} - \mu_{S*S}$, was computed for each iris. We also found the difference in the average number of bits used where $Bits_{diff} = B_{S*L} - B_{S*S}$, where B_{S*L} is the average number of bits used in long-time-lapse matches and B_{S*S} is the average number of bits used in short-time-lapse matches. This data is shown in Table 1.

We found the average pupil to iris ratio in 2004 and the average ratio in 2008 for each of the irises. In 23 of the 26 irises, the average ratio was smaller in 2008 than in 2004. For every match, we computed the difference in the pupil to iris ratio of the two matched images. For each iris we determined the average ratio difference of short-time-lapse matches, $\Delta\mathcal{P}_{S*S}$ and the average ratio difference of long-time-lapse matches, $\Delta\mathcal{P}_{S*L}$. We found $\Delta\mathcal{P}_{S*L}$ was greater than $\Delta\mathcal{P}_{S*S}$ in 22 of the 26 irises. This change in pupil to iris ratio difference may account for an increase in the HD for long-time-lapse matches. However, we observe no correlation between $\Delta\mathcal{P}_{S*L} - \Delta\mathcal{P}_{S*S}$ and $\mu_{S*L} - \mu_{S*S}$ (see Table 1.)

Across all matches, we determined the mean Hamming distance for a long-time-lapse was 0.230, whereas the mean HD for a short-time-lapse was 0.212. However, we found the nonmatch mean HD was 0.447 for a long-time-lapse and 0.446 for a short-time-lapse. These results indicate a time-lapse effect on match scores, but a negligible effect on nonmatch scores. Fig. 2 clearly indicates the shift in the match distribution for long-time-lapse matches and the consistency within the nonmatch distributions.

3.1 All Irises Test

The difference, $\mu_{S*L} - \mu_{S*S}$, was positive for 23 of the 26 irises with an average difference of 0.0165. In a random sample, we would expect the average HD for long-time-lapse matches to be worse for 13 irises and better for 13 irises. We applied a sign test to test the null hypothesis that the number of positive differences is not statistically significantly greater than the number of negative



(a) Enrollment Image from Spring 2004 (b) Verification Image from Spring 2004 (c) Verification Image from Spring 2008

Fig. 1. Subject 04233 Left iris- HD for spring 2004 gallery versus spring 2004 probe was 0.156. HD for spring 2004 gallery versus spring 2008 probe was 0.285.

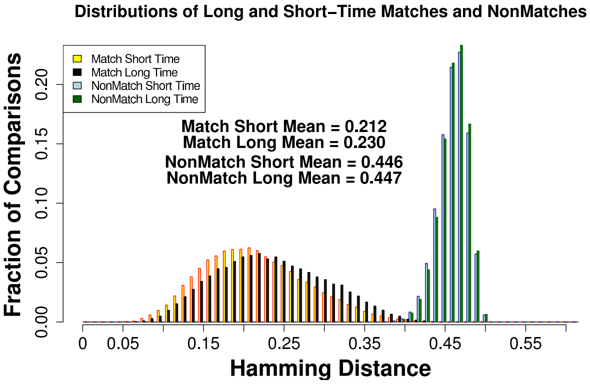


Fig. 2. We observe no change in the non-match distribution, but a significant shift to the right for long-time-lapse matches.

differences. With a sign test statistic value of z = 4.1184, we reject the null hypothesis at a significance level of 5% (p = 0.0001).

A histogram representing this sample of differences of mean Hamming distances is shown in Fig. 3. We applied a chi-square goodness-of-fit test to the sample of 26 differences of means. The null hypothesis that this sample is from a normal distribution cannot be rejected at a 5% significance level. Since this data is approximately normal, we can use a t-test to compare the difference of means.

We applied a one-tailed paired Student's t test to test the null hypothesis that this difference-of-means sample comes from a distribution with a mean of zero. The alternative hypothesis is that the difference distribution has a mean greater than zero, which would mean that the long-time-lapse HDs are on average

Table 1. Average Hamming distance and standard deviation for short-time-lapse and long-time-lapse matches and the change in mean Hamming distances, bits used, and pupil to iris ratio for all 26 irises.

$Iris$	μ_{S*L}	std	$\|\|\mathcal{D}_{S*L}\|\|$	μ_{S*S}	std	$\|\|\mathcal{D}_{S*S}\|\|$	μ_{diff}	$Bits_{diff}$	$\Delta\mathcal{P}_{S*L} - \Delta\mathcal{P}_{S*S}$
02463L	0.1843	0.0404	1419	0.1847	0.0418	1219	−0.0004	61.5	0.0044
02463R	0.2056	0.0377	987	0.1952	0.0375	1008	0.0104	41.0	−0.0014
04233L	0.1977	0.0402	2254	0.1795	0.0420	2108	0.0183	46.3	0.0281
04233R	0.1752	0.0353	2372	0.1712	0.0398	2080	0.0040	33.8	0.0034
04261L	0.1584	0.0378	156	0.1463	0.0403	224	0.0121	−87.7	0.0318
04261R	0.1408	0.0255	127	0.1381	0.0302	176	0.0027	−8.1	0.0168
04385L	0.2316	0.0409	2676	0.1999	0.0383	1628	0.0317	−91.0	0.0289
04385R	0.2288	0.0387	960	0.2140	0.0455	960	0.0148	−74.8	0.0335
04397L	0.1398	0.0363	1983	0.1365	0.0355	1323	0.00330	21.8	0.0219
04397R	0.1441	0.0266	2311	0.1380	0.0290	1496	0.0061	1.2	0.0087
04470L	0.2470	0.0524	479	0.2377	0.0499	572	0.0093	8.9	0.0010
04470R	0.2401	0.0468	689	0.2403	0.0521	518	−0.0002	97.2	0.0084
04537L	0.2131	0.0463	733	0.1991	0.0433	825	0.0140	−24.2	0.0530
04537R	0.1933	0.0387	805	0.1829	0.0442	864	0.0104	−69.0	0.0578
04629L	0.2947	0.0440	1246	0.2691	0.0485	1056	0.0255	−97.2	−0.0170
04629R	0.2994	0.0438	964	0.2678	0.0476	840	0.0316	−171.8	−0.0072
04815L	0.2769	0.0491	2236	0.2336	0.0555	1474	0.0433	−51.6	0.0531
04815R	0.2524	0.0493	2922	0.2263	0.0517	2294	0.0260	−26.8	0.0392
04851L	0.3101	0.0364	1356	0.2619	0.0478	1122	0.0481	−213.8	0.0203
04851R	0.3307	0.0357	2229	0.3092	0.0485	1755	0.0215	−47.5	0.0099
04870L	0.2460	0.0594	1477	0.2432	0.0564	880	0.0028	43.1	0.0228
04870R	0.2607	0.0539	1556	0.2647	0.0520	855	−0.0041	57.1	0.0267
04888L	0.2594	0.0373	147	0.2376	0.0454	162	0.0218	−74.8	0.0268
04888R	0.2261	0.0371	216	0.2168	0.0407	252	0.0093	6.0	0.0007
04917L	0.2380	0.0397	2022	0.2050	0.0406	1386	0.0330	−250.6	−0.0095
04917R	0.2342	0.0395	2419	0.2000	0.0347	1768	0.0343	−246.8	0.0007
All	0.2302	0.0663	28845	0.2118	0.0632	36741	0.0184	−41.3	

greater than the short-time-lapse HDs. The null hypothesis was rejected at a 5% significance level (p=0.00001.)

To confirm that there is no significant effect from the number of bits used in matches, we applied a Student's t test to the distribution of $Bits_{diff}$. The null hypothesis was that the mean of this sample was zero. We failed to reject the null hypothesis at a 1% significance level (p = 0.0285). Thus, across all irises, there was no significant change in the number of bits used.

3.2 Iris-Level Test

For each iris subject, we have two samples of Hamming distances, one from long-time-lapse matches, $\mathcal{D}_{S*L}$, and one from short-time-lapse matches, $\mathcal{D}_{S*S}$. These

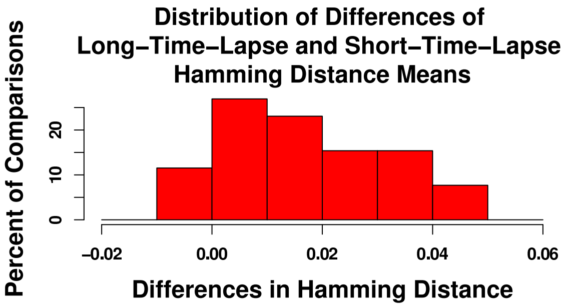


Fig. 3. Distribution of difference of long-time-lapse means and short-time-lapse means

samples were approximately normal, so we applied a one-tailed Student's t test to test the null hypothesis that these two samples of matches come from the same distribution with equal means. The alternative hypothesis is that $\mathcal{D}_{S*L} > \mathcal{D}_{S*S}$. The null hypothesis was rejected for 21 of the 26 irises at a significance level of 0.05.

3.3 Sensor Tests

We have observed that the Hamming distance for long-time-lapse matches is on average larger than that for short-time-lapse matches. One possible cause for this observation would be that there is some subtle change in iris texture over time. However, it is important to note that this is not the only possible cause. For example, if the sensor properties changed over time, this could also produce a change in the imaged texture even if there is no change in the true iris texture.

We performed an experiment with images from the original LG2200 camera used in the acquisition for all images in this paper and a different, rarely-used, LG2200 camera. We tested images from both cameras to determine if the original, well-used, camera and sensor have a degrading effect on match quality. To perform this test, we used two sets of images from Fall 2008 acquired with the original camera as the gallery set and the first probe set, and a third set of images from Fall 2008 acquired with the new, rarely-used camera as the second probe set. We found the matches produced from the two different probe sets were not significantly different. Therefore, we do not see any evidence that the sensor properties have changed enough to explain the time-lapse conclusions we have presented.

4 Discussion and Future Work

We observe an approximate 0.018 increase in Hamming distance for matches with a four years time-lapse. HDs are between 0 and 0.5, so our result represents an approximate $3-4\%$ increase over a four year period. Additionally, at a false accept rate of 0.01%, the false reject rate increases by 75% for long-time-lapse.

The basic results and conclusion presented here run counter to conventional wisdom about iris biometrics. However, we know of no experimental study that has previously tested the "one enrollment for a lifetime" assumption. The previous time variability study referenced in the introduction compared images with less than three months time lapse. Their results show better performance for images acquired in the same session than images acquired across sessions. They also note no significant differences between two weeks to four weeks to two months time lapse. Our results are based on images of the same iris imaged with time lapse as long as four years. With this long-term time lapse we note statistically significant changes in the iris match quality. Upon visual examination of the irises with the largest difference in Hamming distance, we observed no drastic changes in iris textures, suggesting that if the iris aging affect is real, it is based on subtle differences.

In this study, we use the same iris imaging system, and control for contact lenses, pupil dilation, and number of bits in a match. We noted no apparent trend in the change in the number of bits in a match. In 22 of the 26 irises, the difference in the pupil dilation between the images of a match was greater for matches of long-time-lapse than matches of short-time-lapse. However, this change in pupil dilation difference does not correlate with the change in Hamming distance across the two sets of time-lapse. We have considered the major potential complicating factors for an experimental study of this type. However, it is still important for our result to be replicated by other research groups using different and larger data sets with more subjects.

Future work includes investigation into textural changes and pinpointing the location of such changes. Predicting textural or pupil dilation changes may aid in accounting for degradation in the match statistic. While we have observed an increase in Hamming distance and the false reject rate over a four year period, we do not know if this trend is linear or how the match quality will change with eight years, or longer, time lapse.

Even if the "lifetime enrollment" concept is disproved, it is not necessarily a major barrier to practical deployment of iris biometrics systems. It would mean that consideration should be paid to the time-lapse between image acquisitions in quantifying a match statistic. One possible reconciliation for match quality degradation is to re-enroll a subject with every verification scenario. However, this requires routine verifications as a long time lapse between enrollment and verification will result in an increased false reject rate. Another possibility is to require a re-enrollment session for every subject after a set time frame. The necessary time frame may be difficult to determine. If the time frame is too long the iris match quality may degrade beyond the accept rate before re-enrollment. A third possibility is to report the time lapse between the enrolled and the verification images as well as the match statistic. If further research shows a possible prediction of changes in the match statistic with increased time lapse, we may be able to normalize the statistic based upon this lapse. We suggest these possible considerations but recognize that much further research is needed before making a recommendation.

5 Acknowledgement

This work is supported by the National Science Foundation under grant CNS01-30839, by the Central Intelligence Agency, by the Intelligence Advanced Research Projects Activity and by the Technical Support Working Group under US Army contract W91CRB-08-C-0093. The opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of our sponsors.

References

1. John Daugman, “How Iris Recognition Works,” *IEEE Trans. Circuits and Sys. for Video Tech.,* 14(1):21-30, 2004.
2. LG. http://www.lgiris.com/, accessed August 2008.
3. J. Thornton, M. Savvides, V. Kumar. “A Bayesian Approach to Deformed Pattern Matching of Iris Images,” *IEEE Trans. Pattern Anal. and Mach. Intell.*, 29(4):596-606, April 2007.
4. K. Miyazawa, K. Ito, T. Aoki, K. Kobayashi, H. Nakajima. “An Effective Approach for Iris Recognition Using Phase-Based Image Matching,” *IEEE Trans. Pattern Anal. and Mach. Intell.*, 30(10):1741-1756, Oct. 2008.
5. D. Monro, S. Rakshit, D. Zhang. “DCT-Based Iris Recognition,” *IEEE Trans. Pattern Anal. and Mach. Intell.*, 29(4):586-595, April 2007.
6. X. Liu, K. Bowyer, P. Flynn. “Experiments with an improved iris segmentation algorithm,” *Fourth IEEE Workshop on Automatic Identification Technologies*”, 118-123, Oct 2005.
7. Xiaomei Liu. “Optimizations in Iris Recognition.” PhD Dissertation, University of Notre Dame, 2006.
8. National Institute of Standards and Technology. Iris Challenge Evaluation, 2006, http://iris.nist.gov/ice.
9. K. Hollingsworth, K. Bowyer, P. Flynn, “The Best Bits in an Iris Code,” *IEEE Trans. Pattern Anal. and Mach. Intell.*, in press.
10. J. Phillips, K. Bowyer, P. Flynn, X. Liu, T. Scruggs “The Iris Challenge Evaluation 2005” *2008 IEEE Conf. on Biometrics: Theory, Applications, and Systems.*
11. P. Tome-Gonzalez, F. Alonso-Fernandez, J. Ortega-Garcia, “On the Effects of Time Variability in Iris Recognition” *2008 IEEE Conf. on Biometrics: Theory, Applications and Systems.*
12. John Daugman, “New Methods in Iris Recognition,” *IEEE Trans. Sys., Man, and Cyber.* 37(5):1167-1175, Oct 2007.
13. Karen Hollingsworth, Kevin Bowyer, Patrick Flynn, “Pupil Dilation Degrades Iris Biometric Performance,” *Computer Vision and Image Understanding,* in press.
14. K.W. Bowyer, K.P. Hollingsworth, and P.J. Flynn. “Image Understanding for Iris Biometrics: A Survey.” *Computer Vision and Image Understanding*, 110(2):281-307, 2008.
15. Sarah Ring and Kevin Bowyer, “Detection of Iris Texture Distortions by Analyzing Iris Code Matching Results” *2008 IEEE Conf. on Biometrics: Theory, Applications, and Systems.*
16. Nathan Kalka, J. Zui, N. Schmid, B. Cukic, ”Image Quality Assessment for Iris Biometric”, SPIE 6202: Biometric Technology for Human Identification III: D1-D11, 2006.