Template Aging:
A Study of the NIST Biometric Score Set Release 1

Michael E. Schuckers, Nicole L. Lopez
Department of Mathematics, Computer Science and Statistics, St. Lawrence University, Canton, NY 13617, USA
schuckers@stlawu.edu

1. Template Aging

For bio-authentication devices the topic of template aging is an important one. Mansfield and Wayman have defined template aging as " the increase in error rates caused by time related changes in the biometric pattern, its presentation, and the sensor," [4]. Understanding such an effect is important both for policy and for system design. To date, the work that has been done in this area has been descriptive of the template aging effect rather than inferential. [1, 3, 6] have all presented descriptive evidence of template aging. Here we focus on studying error rates (both FAR and FRR) over time and determining whether or not the impact of time is statistically significant. We apply our methodology to the National Institute for Standards and Technology (NIST) Biometric Score Set Release 1 (BSSR1), which was recently made public. The results of our regression analysis of the match scores and our logistic regression of the decision data indicate that there is a significant effect due to the changes in time for some matching algorithms at some thresholds. However, the template aging effect is not consistent across all modalities and all thresholds.

2. Data: NIST BSSR1

"Biometric Scores Set - Release 1 (BSSR1) is a set of raw output similarity scores from two [c.] 2002 face recognition systems and one [c.] 2004 fingerprint system, operating on frontal faces, and left and right index live-scan fingerprints, respectively. The release includes true multimodal score data, i.e. similarity scores from comparisons of faces and fingerprints of the same people," [5]. We focus here on the smallest of the three databases - 517 individuals - since it provides match scores for two different modalities as well as data information which is integral to a template aging analysis. We will refer to this database as the "Fing-Face" database. Because the date of each image collection is recorded in the database, it is possible to know the length of time between the collection of the images in each case, [2]. In some cases "enrollment" images were collected after the images to which they were compared. To remedy this we took the absolute value of the difference in days. In the Fing-Face data, all cross-comparisons were completed. There are four complete data sets in BSSR1: two facial algorithm comparisons, C and G, and a fingerprint algorithm, V, applied to both right index and left index fingers. We will refer to these data sets as Face-C, Face-G, FP-LI-V and FP-RI-V, respectively. We will refer to genuine scores as those produced by two images of the same individual, while the imposter scores are those produced by two images from different individuals. Decision score data are the zero's and one's produced when we set a threshold for a particular distribution and determine which values yield errors (one's) and which yield correct decisions (zero's).

3. Methods

We used two different methodologies here each appropriate to the two type of data we were considering. First, for the match score data, we used linear regression to determine if there was a significant relationship between the match scores - both imposter and genuine - and time. Figures 1 and 2 graph the scores over time along with the fitted regression lines. Because error rates involve not the mean but some percentile of the match scores, it is important to look at the decision level data to determine whether or not the error rates are changing. To that end we considered a logistic regression of the decision scores against time for evaluating changes in the error rates. For each data, we considered a wide range of thresholds to produce the decision scores.

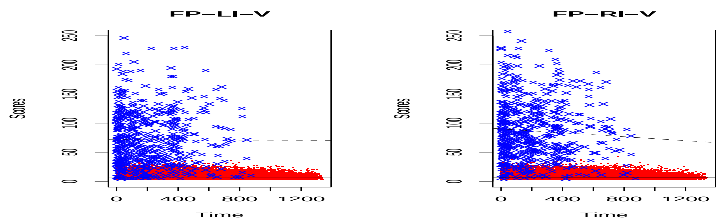

Figure 1. Match Scores versus Time (in days) for a Fingerprint Algorithm matching. Left Index fingers (LI) and Right Index Fingers (RI) Red indicates Imposter scores, Blue indicates Genuine scores

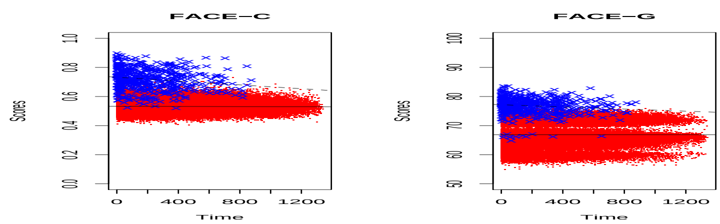

Figure 2. Match Scores versus Time (in days) for a Two Face Recognition Systems Face-C and Face-G. Red indicates Imposter scores, Blue indicates Genuine scores

4. Results

4.1. Regression Results

Overall there were some significant - p-value < 0.01 - trends in the match scores. Among the imposter distributions, Face-C and FP-LI-V yielded significant negative relationships between match scores and time. Hence Face-G and FP-RI-V showed no significant relationship between mean imposter scores and time. For the genuine match scores, both of facial recognitions systems, Face-C and Face-G, yielded significant negative relationships between match scores and time, while neither of the fingerprint data possessed a significant relationship.

4.2. Logistic Regression Results

Here the outcomes are more complex because they depend explicitly on the threshold to determine the error rates. In what follows, we specify the average error rate at which significant trends were observed. For Face-C, we studied the impact of thresholds from 0.545 to 0.625 in increments of 0.005. We observed no significant changes in the FRR and significant decreases in the FAR when the average error rate was more than 0.01. The range of thresholds studied under Face-G was from 65 to 75.5 in increments of 0.5. FAR significantly increased over time for the lowest threshold, 65, which gave an average FAR of approximately 0.81, while FRR significantly increased when the average FRR was more than approximately 0.20. Among the fingerprint data, only the FAR for FP-V-LI showed significance. The FAR decreased significantly when the average error rate was more than approximately 0.300.

References

[1] D. M. Blackburn, M. Bone, and P. J. Phillips. Facial recognition vendor test 2000: Evaluation report. http://www.frvt.org/DLs/FRVT2000.pdf, 2001.

[2] P. Grother. Personal correspondence. 2004.

[3] J. R. Kitchel and S. J. Elliot. Effects of template aging on facial recognition. http://www.tech.purdue.edu/it/ resources/biometrics/ posters/Kitchel-Template-Aging-At-Airport.ppt, 2003.

[4] T. Mansfield and J. L. Wayman. Best practices in testing and reporting performance of biometric devices. on the web at www.cesg.gov.uk/site/ast/biometrics/media/BestPractice.pdf, 2002.

[5] National Institute of Standards and Technology. Biometric Score Set Release 1. http://www.itl.nist.gov/iad/894.03/biometricscores/, 2004.

[6] P. J. Phillips, P. Grother, R. J. Micheals, D. M. Blackburn, E. Tabassi, and M. Bone. Facial recognition vendor test 2002: Evaluation report. http://www.frvt.org/DLs/FRVT2002EvaluationReport.pdf, 2003.