Parametrically Controlled Synthetic Imagery Experiment for Face Recognition Testing

Nicholas M. Orlans
The MITRE Corporation
7515 Colshire Drive
McLean, VA 22102
1.703.883.7454
norlans@acm.org

Alan T. Piszcz
The MITRE Corporation
7515 Colshire Drive
McLean, VA 22102
1.703.883.7124
apiszcz@mitre.org

Richard J. Chavez
The MITRE Corporation
7515 Colshire Drive
McLean, VA 22102
1.703.883.3682
rchavez@mitre.org

ABSTRACT
The use of synthetic imagery for testing biometric systems is relatively new and in need of further exploration. In this paper, we describe methods and procedures for using synthetic images generated from shape and texture data to refine and extend the current state of the art of face recognition performance testing. Two example experiments are presented based on the canonical Facial Recognition Vendor Test 2000—pose experiments and temporal experiments. We demonstrate how the use of synthetically generated face models (and resulting images) can enhance and extend existing test protocols and analysis. These methods and results will be of use to developers and practitioners alike.

Categories and Subject Descriptors
D.2.5 Testing and Debugging (data generators); D.2.8 Metrics (performance metrics)

General Terms
Measurement, Performance, Reliability, Experimentation, Verification.

Keywords
Biometric systems, biometrics testing, face recognition, performance evaluation, synthetic imagery, 3D face modeling.

1. INTRODUCTION
This paper describes the use of synthetic imagery for face-recognition testing. Synthetic imagery is useful for rapidly generating large data sets. Synthetic imagery can also support custom data sets designed to isolate and parametrically control specific test features. The proposed methodology augments the results of the well-known Facial Recognition Vendor Test 2000 (FRVT 2000) [2] by improving the resolution, measurement, and understanding of face-recognition technology tests. Synthetically generated inputs for face-recognition are defined in terms of 3D head and face geometry, image (face) texture, and rendering (lighting) and viewing parameters.

2. BACKGROUND
As developed by Raffaele Cappelli and others, there is precedence for the use of synthetically generated and controlled fingerprint images for fingerprint recognition testing [3]. The Cappelli SFINGE (Synthetic FINgerprint GEnerator) system generates fingerprint images for less cost and effort than is required for collecting consistent test images from live subjects. (SFINGE authors state that one thousand realistic prints can be generated in one hour using a single Pentium IV CPU [6]). Other biometric techniques have not exploited or fully explored the use of synthetic data sources to support development and test. Synthetic data techniques can generate images that mimic, filter, or perturb live data in known ways to augment live and heuristic test methods. In addition to there being no privacy concerns with synthetic data, synthetic data has the potential to augment Best Practices test procedures [1], enhance understanding, and support certain types of results not achievable with real images. Moreover, synthetically generated data can provide better isolation, control, granularity (hence better measurement) of certain known performance factors. Critical performance transition regions can be identified and examined in more detail than by using conventional methods alone.

The Facial Recognition Vendor Test 2000 program (FRVT 2000) was established based on Dr. Jonathon Phillip's FacE REcognition Test (FERET) methodology. FRVT 2000 Evaluation Report, a program performed under joint sponsorship from DoD Counterdrug, DARPA, and NAVSEA, documented eight face recognition experiments. Each experiment addresses different face recognition performance factors. FRVT 2002 experiments extended the original 1994 FERET images to a collection of 1396 images. Of the eight experiments, the pose and temporal experiments were cited as requiring additional research.

Pose experiments address the affect of camera angle on recognition performance. The FRVT 2000 Evaluation Report [2] concluded, "performance is stable when the angle between a frontal gallery and a probe is less than 25 degrees and that performance dramatically falls off when the angle is greater than

45 degrees." As symmetry is assumed the test data contained only 5 reference samples were used to describe the variation of pose angles (15, 25, 40, 45, and 60 degrees). As a consequence there is no information on what happens in the critical pose angles between 25 and 45 degrees.

Temporal experiments address the affect of temporal gaps on recognition performance. The temporal gap is the time span from when the reference biometric was collected and subsequent recognition attempts. FRVT 2000 images of any given individual spans no more than 2 years, and temporal gaps typically involve other image differences (in addition to the "aging" of the subject) that resulting from the collection of images at different times and places.

The authors demonstrate how FRVT 2000 findings for pose and temporal experiments can be supported with synthetic images by isolating experimental features and creating finely controlled, parametrically generated test probes. Pose angles are compared in one-degree increments. Temporal gaps are interpreted as simulated aging and are compared in five-year increments.

As established by Duane Blackburn and others (in [2], [3], [8]) there are five ideals that should be present for the proper evaluation of biometric systems. Each of the five ideas, summarized below, is addressed in this experiment:

Independent groups—the organization administering and executing the evaluation should have no interests, financial or otherwise, in performance outcomes. The MITRE Corporation is a not-for-profit national resource that provides systems engineering, research and development, and information technology support to the government. In this role, evaluation is a fundamental activity to advise stakeholders in technical capabilities for future systems.

Independent test data—the data used for testing and evaluation should be collected independently and not known to the participants. There was a constructive general exchange of information with Viisage, however, there was no disclosure of the data set used in the experiments. (Note these experiments purposefully did not involve multiple vendors. The experiments were made possible with the knowledge, cooperation, and support of Viisage. The results presented here are to demonstrate complete testing methodology—they should not be interpreted to reflect on vender performance.)

Three Bears Principle—evaluations should not be too hard, nor should they be too easy. Difficulty should be somewhere in the middle, or "just right." To this end, the experiments are based on existing, known, and "fair" operational ranges for the technology.

Repeatable—evaluations should be conducted and documented with enough detail so that others can reproduce statistically similar results. The evaluation methods and test environment is presented in this paper. Additionally, the test data is available by request from the authors.

Know your requirements—evaluations are useful only if they relate to known application requirements. The experiments we conducted, pose and temporal, were selected because they are relevant to a large class of face-recognition identification applications. Moreover, these two experiments were identified in [2] as the most promising areas for additional research and development.

Common Criteria testing stipulates that real images must be used for ultimately determining performance, the collection of face images will continue to present practical problems, including privacy, repeatability, and biases caused by difficult to control lighting or population selection details. Regardless of Common Criteria test requirements, the isolation and parametric control of facial images provides additional understanding to face recognition technology. The eventual goal is to apply that understanding to be able to make corrections and adjustments in images to boost recognition performance across differences in pose, time, and lighting. The basis for such techniques were presented and presented by Blanz and Vetter in [4] in 1999, and the promise of such techniques was also mentioned in FRVT 2002.

3. EXPERIMENT ENVIRONMENT

Generation of synthetic face image galleries was accomplished using the process depicted below (Figure 1). Face models were created using the FaceGen 2.2 modeler. The models were then imported into 3D Studio max using a developed script to provide consistent orientation, rendering, and file nomenclature. The generated image sets were the inputs for face recognition, here the Viisage FaceTools product (version 2.3). Lastly, enrollment and comparison results were reviewed, analyzed, and reported. A description of the primary steps and tools are further discussed in this section.

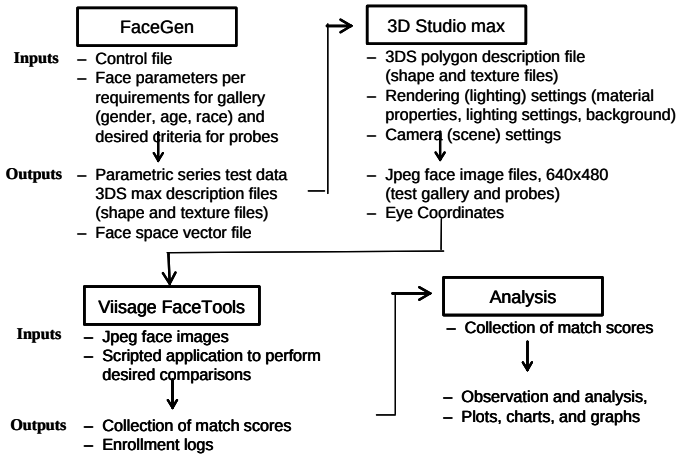

Figure 1. Parametrically driven synthetic face generation environment.

3.1 FaceGen

FaceGen is a modeler for generating and manipulating human faces. It provides control of shape, texture, expression, phones, and accessories (hair, glasses). The FaceGen modeler is from Singular Inversions Inc., Toronto Canada. The MITRE experiment used version 2.2 of the software, which provides improved face texturing features from the previous version. FaceGen runs on most Windows platforms and is also available as a software development kit, allowing for programmatic control and morphing of face models.

3.2 3D Studio max

3D Studio max is a full-featured modeling, animation, and rendering environment from Discreet, Montreal Canada, owned by Autodesk Inc.. The MITRE experiment used version 5.1 of the

product. A script was developed using the built-in scripting language, MaxScript, to automate the import, rendering, and file nomenclature of the synthetic data sets.

3.3 Viisage FaceTools

Viisage FaceTools is a face recognition product developed and sold by Viisage Inc., Littleton MA. The FaceTools product line uses an Eigenface implementation, a technique originally introduced by Turk and Pentland in [9]. Viisage provides a suite of face recognition products and system integration services and has also participated in both FRVT 2000 and FRVT 2002 test programs.

4. POSE EXPERIMENT

As established by FRVT, pose experiments study the effects of face orientation (or viewing angle) on face recognition. The MITRE synthetic pose experiment used one hundred randomly generated faces produced with the FaceGen modeler. Each face was subsequently rendered in one-degree horizontal pose increments from –60 to +60 degrees, generating a total of 121 pose variations for each face. A subset of pose experiment face images (5-degree angle increments) is shown below (Figure 2). The actual images are separate 640x480 jpeg files. The choice of image size corresponds to camera recommendations in the FaceTools SDK documentation.

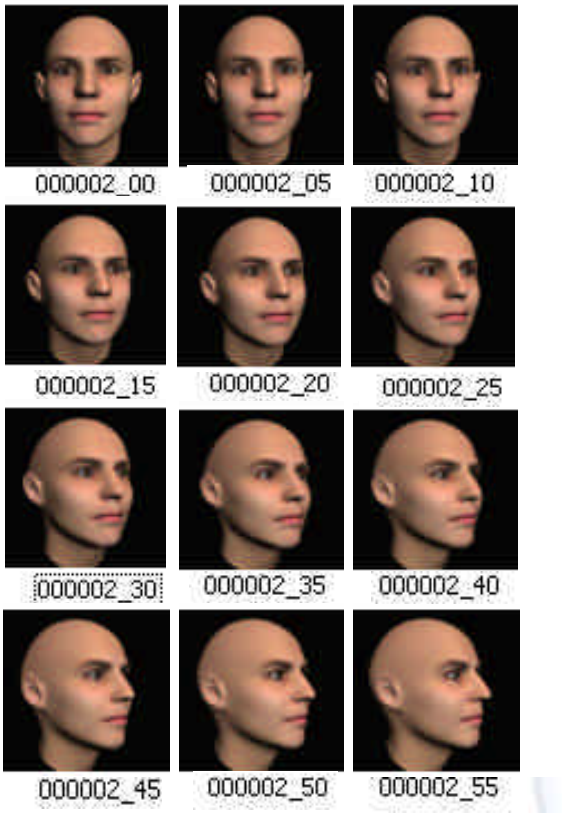

Figure 2. Example faces for pose experiment.

The fifty "best" faces, as determined by enrollment performance, were used for subsequent pose comparisons (please refer to the enrollment section for information on the enrollment process). Each of the 120 rotated poses was compared against the base frontal image for each face. Recognition performance by pose angle shows (Figure 3) the average match score for each pose angle. The horizontal axis of the plot is the pose angle and the vertical axis is distance scores, thus small scores represent the closest matches (e.g. 0.0 is an exact match or self match). Note that performance degrades in a sharp curve as pose angle increases. Based on the vendor's recommended threshold of 0.75, recognition performance falls off completely in the neighborhood of five to eight degrees (for this synthetic data). Note that the distant eye effectively disappears from the image between 35 and 40 degrees. Poses between 35 and 60 degrees have the highest scores on the recognition curve, indicating the least precision.

The results from this pose experiment differs significantly from the previously mentioned FRVT 2000 summary result that "performance is stable when the angle between a frontal gallery and a probe is less than 25 degrees and that performance dramatically falls off when the angle is greater than 45 degrees." The performance curve also reveals some asymmetry between the left and right pose angles. We speculate the asymmetry is due to the other great challenge in face recognition—lighting. Future work is necessary to address the nuances of lighting in face recognition with real, purely synthetic, and processed (hybrid) images. Analogous to the parametric control of pose, parametrically controlled lighting experiments should be useful utilities for determining sensitivity thresholds for face recognition.

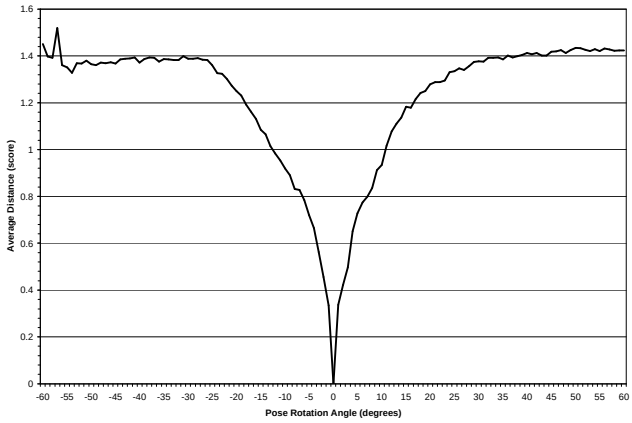

Figure 3. Recognition performance by pose angle.

5. TEMPORAL EXPERIMENT

Due to the logistic difficulties for collecting large, controlled (natural) data sets over extended time periods, temporal face recognition tests to date have been somewhat limited. FRVT 2000 temporal data spanned only two years yet provided enough information for the authors to identify temporal experiments as an area needing additional research. The MITRE synthetic aging experiment used simulated aging as defined within the FaceGen modeler. Fifty random faces were used and each face was "aged" from age 20 to 60 in five year increments, generating a total of nine temporal images for each face. A representative set of synthetically aged face images is shown (Figure 4). The actual images are separate 640x480 jpeg files.

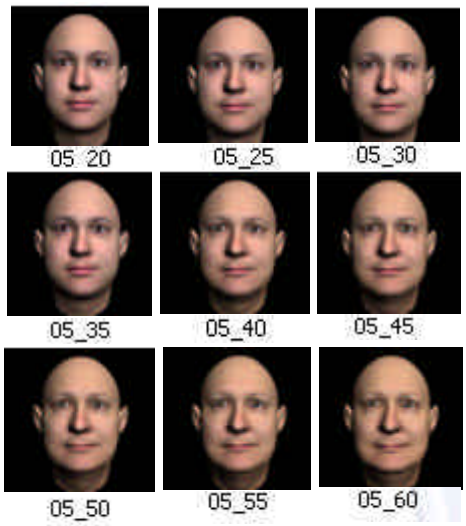

Figure 4. Example faces for temporal experiment.

Cross comparisons were performed for all 450 images (nine temporal instances of fifty faces), allowing us to view performance using any age face as the base reference. For each simulated age, the average comparison scores against the other ages were averaged across the fifty faces. While the statistical significance of this result in light of real data results reported in FRVT 2002 is uncertain due the dependency and differences in the respective galleries, the initial qualitative results appear consistent with FRVT conclusions.

As expected, recognition continues to degrade over time (differences to the reference image increases), and this result is shown (Figure 5). One of the more interesting results of FRVT 2002 was that older (males) were more recognizable than younger ones (and in particular younger females). While we did not partition temporal data by gender, note that where the base reference age is 40 (Figure 6), there is slightly better performance for subsequent aging. That is, the average distance from the reference image to the subsequent five year increments is less than the respective distances observed when the reference age is 20 (Figure 5).

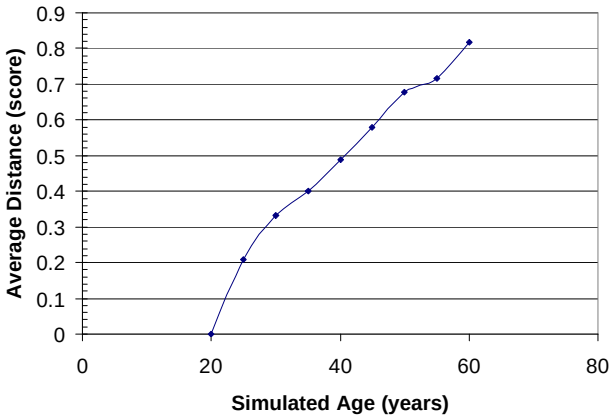

Figure 5. Recognition performance by simulated age (reference age 20).

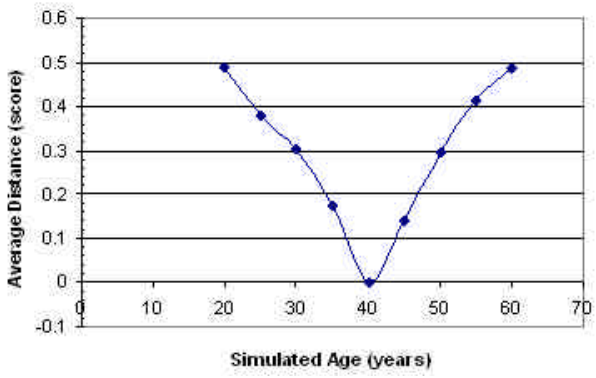

Figure 6. Recognition performance by simulated age (reference age 40).

6. ENROLLMENT EXPERIMENT

The 12,100 pose experiment faces were enrolled into FaceTools using a modified version of the automatic enrollment sample application provided by Viisage. The modified version provided more complete logging of the enrollment process, and recorded the following information for each image: image identifier, enrollment success (with return code from eye find routine), coordinates of right and left eye locations, eye spacing, time stamp.

Enrollment performance for face recognition is the rate of successful enrollment as reported by the software's segmentation and eye finding processes.[1] The enrollment rate for the temporal data with gray backgrounds was observed to be 90.2%, which was less than expected. Enrollments were then subsequently compared using white, gray, and black backgrounds, and we observed the best automated enrollment performance occurred with the black background. The comparison of enrollment performance with the different backgrounds is shown below (Table 1). The percentages represent the successful enrollment rate (that is, the combination of both correct and incorrect enrollments). As the black background result differs with NIST best practices for mugshot capture [7], and the vendor's recommendation, it is a detail that merits further exploration.

Table 1. Successful enrollments by image background

White	Grey	Black
89.3%	90.2%	97.7%

Besides background color and contrast, other possible factors adversely affecting synthetic enrollments are eye quality, eye

[1] The related, more commonly used metric associated with acquisition and enrollment of biometrics is *failure to acquire*. For live tests, the failure to acquire rate is the rate that, for any reason, an image is unobtainable. For our methodology, however, we are also interested in determining the more subtle counterpart metric, *false enrollments* (images that were acquired and enrolled but with incorrect eye finds).

contrast, and baldness of subjects. The relatively abrupt image transitions from background across to the (hairless) head, and from face to the eyes is visibly different with synthetic images, hence the software techniques used to detect these boundaries are likely to perform differently. The decision to not include hair models in the rendered images was based on the fact that the eigenface implementation used by Viisage masks out these regions prior to template generation.

For the pose database, we expected to see enrollment performance degrade as function of increased pose angle—the same basic trend as anticipated for actual recognition performance. However, this expected result was not observed. Rather, we observed enrollment performance, when viewed as a function of pose angle (Figure 7), to contain local inflection points that are not easily explained. We also plotted the eye locations of the images that did enroll and found there to be a surprisingly high incidence of false enrollments. While these results may not be indicative of performance on real images, it is worth noting that the enrollment error rates, particularly the false enrollment rate, represent an important aspect of overall performance and is not individually reported in most biometric testing.2 The observed enrollment performance caused us to question the validity of subsequent recognition performance results. We therefore confirmed the results of the recognition performance experiments using supervised enrollment of the images (as opposed to automated enrollment). Supervised enrollment requires inspection of the images and manual designation of eye coordinates.

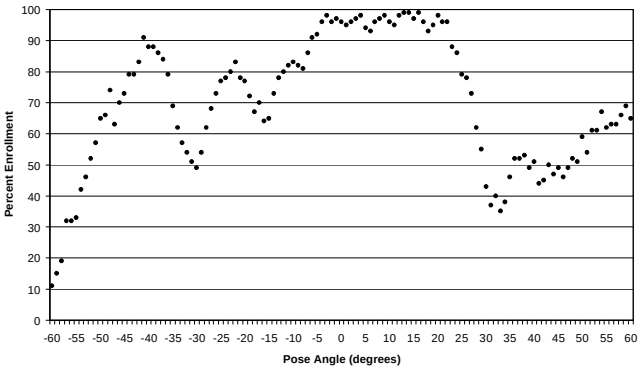

Figure 7. Successful enrollments by pose age angle.

7. INDIVIDUALITY OF GALLERY

An important corollary question that arises with synthetic images is, how unique are the generated faces? The FaceGen modeler uses a statistical appearance model to formalize their concept of gender, age, race, and "random" faces. The issue of validating that a large gallery of synthetic faces is in fact a meaningful representation of a natural population is a difficult challenge. We examined the gallery uniqueness question for our gallery of 100 faces from the pose experiment by performing exhaustive cross comparisons between all the images. The scores were mapped into color regions and plotted in a graphical matrix shown below (Figure 8).

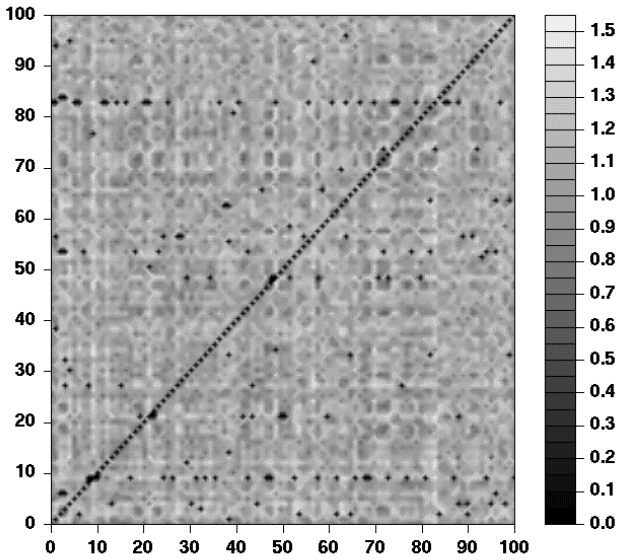

Figure 8. Individuality of the synthetic gallery.

Black dots are assigned to all scores under the threshold of 0.7. Lighter tones represent the more distant matches, and the middle tones depict middle ranges. The visual effect is that all images match themselves (along the diagonal), and there is only a relative small of similar faces. From the raw data, we tabulated the cross over match rate to be 2.9%, and it is evident the confusion that does occur is symmetric. Overall the synthetic gallery exhibits reasonably good separation.

8. CONCLUSIONS AND FUTURE WORK

The techniques presented here demonstrate that different test factors can be isolated to assess their parametric influence on face recognition processes. Moreover, eye finding and other sub-processing tasks critical to overall face recognition performance can be instrumented along with the primary test factors. Examples of other important sub-processing tasks are light balancing, light normalization, and pose estimation (from image data).

In addition to pose and temporal experiments addressed here, accounting for differences in lighting across environments (particularly indoor versus outdoor) is an area identified as requiring new or more robust processing techniques. As presented

[2] FERET test data provided eye coordinates as inputs for most the experiments. FRVT 2000 and FRVT 2002 were fully automatic. Real applications generally are not provided the benefit of face registration information, and consequently performance is an unknown combination of face detection and subsequent recognition processing. Face detection, registration, and segmentation is sufficiently challenging that it may benefit from isolated testing for some applications. Sometimes this information is reported as a *failure to acquire* rate. The counterpart *false enrollment rate*, however, is seldom reported and not sufficiently addressed in current biometric testing.

in Blanz and Vetter's morphable models [4], adjustments to face models (and resulting two dimensional images) are possible providing there are known exemplars within the data. The application to face recognition is to use these parametric representations of lighting and pose (view) angle as the basis for image reconstruction. The goal of the reconstructed image is to remove (or adjust for) any lighting and pose differences that may exist between reference and target images.

While qualitative aspects of sensitivity testing can be achieved with these techniques, the problem of validating that large galleries (10,000 or more) of synthetic faces are representative of a natural population remains a challenge to be addressed.

9. ACKNOWLEDGMENTS

We thank the MITRE Technology Program for supporting us and funding this work. We thank the Viisage Corporation for their support, interest, and their cooperation in this experiment. We also extend thanks and appreciation to Duane Blackburn for providing helpful comments, correcting errors, and for sharing general suggestions for establishing longer term objectives for utilizing synthetic imagery. Helpful review was also offered by Joseph Marques, for which we are thankful.
The authors are responsible for any ambiguities or errors that may remain.

10. REFERENCES

[1] Biometrics Working Group, CESG/NPL. January 2000. "Best Practices for Biometric Testing." Version 1.0. http://www.cesg.gov.uk/technology/biometrics/index.htm

[2] D. Blackburn, M. Bone, and Dr. P.J. Phillips, 2000, "Face Recognition Vendor Test 2000 Evaluation Report," DoD Counterdrug, DARPA, and NAVSEA, February 16, 2001.

[3] D. Blackburn, "Evaluating Technology Properly— Three Easy Steps to Success," Corrections Today, July 2001, Vol 63 (1).

[4] V. Blanz and T. Vetter. A Morphable Model for the Synthesis of 3D Faces. In Proceedings of Computer Graphics SIGGRAPH, Los Angeles, CA, August 8-13, 1999 (pp. 187-194). ACM Press, New York, NY.

[5] R. Cappelli, "SfinGe: Synthetic Fingerprint Generator," in proceedings 12 th CardTech/SecurTech (CTST2002), April 2002.

[6] Computer Science Department, University of Bologna, "Synthetic Fingerprint Generator (SfinGe) Home," Version 2.5, http://bias.csr.unibo.it/research/biolab/sfinge.html (January 2 2003).

[7] National Institute of Standards and Technology, Best Practice Recommendation for the Capture of Mugshots, version 2.0, September 23, 1997 (http://ww.itl.nist.gov/iaui/894.03/face/bpr_mug3.html).

[8] P.J. Phillips, A. Martin, C.L. Wilson, and M. Przybocki, "An Introduction to Evaluating Biometric Systems," *IEEE Computer*, February 2000, pp. 56-63, 2000.

[9] M. Turk and A. Pentland, "Eigenfaces for recognition", Journal of Cognitive Neuroscience. V. 3, pp. 71-86, 1991.