AC 2011-259: FACIAL RECOGNITION SYSTEM SCREENING EVALUATION METHODOLOGY FOR COMPLEXION BIASES

Rigoberto Chinchilla, Eastern Illinois University

Dr. Rigoberto Chinchilla (PhD in Integrated Engineering, Ohio University) is an Associate Professor in the School of Technology since 2004 and Current Interim Coordinator of Graduate Studies for the School of Technology at Eastern Illinois University. His teaching and research interests include Applied Statistics, Quality Assurance, Computer and Biometric Security, Information Systems, and Automation. Dr. Chinchilla has been a Fulbright scholar, a recipient of a United Nations scholarship, chosen as a Faculty Marshall for the Graduate School, and received an Achievement and Contribution Award as well as the ”Excellence in the Use of Technology ” (research) at EIU. His publications include: ”Ethical and Social Consequences of Biometric Technologies in the USA”, ”Technology in Central America and the Impact on CAFTA” and ”Design of an Industrial Control Laboratory” amongst others. Dr. Chinchilla has been awarded numerous grants and serves in numerous departmental and university committees at Eastern Illinois University.

Mr. Harold Jay Harris, Eastern Illinois University School of Technology

Facial Recognition System Screening Evaluation Methodology for Complexion Biases

Abstract

Over the years, Facial Recognition Systems (FRS) have come under scrutiny from watchdog groups who voice their complaints concerning the potential existence of a FRS bias towards certain cultures of people while such systems are generally deployed in security screening situations. To better understand this potential FRS bias, researchers examine a theory developed from the behavioral sciences known as the "other-race effect." FRS researchers have also used the "other-race effect" theory in an attempt to explain the occurrence of biases associated with algorithms tested during the Face RecognitionVendor Test of 2006. In this paper, we develop a scientific testing methodology based on the factors of illumination, distance, and angle to evaluate whether or not a chosen FRS exhibits a significant bias when presented with two dissimilar three dimensional (3D) facial models for comparison. To test our methodology, we compare a light complexion (3D) facial model with a medium complexion 3D facial model. Our methodology will incorporate a full factorial experiment and Design of Experiments (DOE) Pro statistical software for data processing. Multiple regression and ANOVA are also used to analize the results. This testing methodology has been incorporated in our academic programs and implemented in different Biometric Security and DOE course projects.

Introduction

Authentication is a process in which a conformation, with a high degree of certainty or probability, is made about the identity of an individual. Human beings have unique physical and behavioral attributes that can be used for authentication purposes. Biometrics can be defined as all the authentication techniques that rely on measurable physiological or behavioral human characteristics that can be verified using computers. Authentication can be accomplished by comparing the biometric information an individual presents to an algorithm on the computer against a binary template previously stored in a database. If the algorithm makes a comparison against one and only one template, the authentication process is called verification. In the verification process, the individual is the one who claims a specific identity. Verification applications are typically aimed at granting individuals the right to access a facility or to use a resource.

If the algorithm attempts to authenticate an individual against more than one template to determine whether or not that individual belongs to the algorithm's database, the process is called identification. In the identification process, there is no previous claim of an individual's identity. Identification applications are typically used by forensics, crime investigation and security applications. A biometric can be broadly classified as behavioral (i.e. Signature, Gait, Lip motion) or physiological (i.e. Fingerprints, Iris, Face, Hand geometry, Retina). In order to build a biometric application, the first step is to enroll the potential users of the application into a biometric database. Enrollment is performed by using electronic sensors and complex mathematical algorithms capable of detecting and capturing the physiological or behavioral characteristics of the individual. After the image representing the biometric characteristic of an

individual is captured, a set of vendor dependent algorithms are in charge of processing the image in order to convert it to a template. Quality differences between proprietary algorithms, binary representations and capture sensors may lead to possible inaccuracies of the biometric sample as well to the creation of different representations for the same biometric.

Biometric systems compare templates based on probabilistic processes. When an individual wants to access a facility, a biometric sample is provided resulting in the creation of a sample template. This template is then compared with the stored template in the algorithm's database. In biometrics, a score is a number that results from the statistical comparison of two templates. The score represents the probability that two templates belong to the same individual. The biometric systems administrator has to setup a score threshold to which the samples will be compared. Typically, if the statistical score from the sample template is greater than the score threshold, the biometric system concludes that the sample-template and the one stored in the database belong to the same individual. If the sample score is below the score threshold, the biometric system concludes that the two templates are statistically different and the individual does not belong to the database.

Biometric systems are not 100% accurate. Biometric systems accuracy during the template comparison process of authentication depends on external variables, namely, temperature, training level of the enrollment process technicians, physical condition of the individual to be authenticated, etc. Biometric systems accuracy is also dependent on internal variables such as quality of the equipment and the proprietary algorithms being used. Most biometric systems derive their fundamental accuracy from the following parameters[1]:

- False Match Rate (FMR): Is the probability that an imposter will be accepted as a genuine user by incorrectly judging a match in his or her enrollment template
- False Non-Match Rate (FNMR): Is the probability that a genuine user will be rejected by incorrectly judging a mismatch in his or her enrollment template
- Failure To Enroll (FTE): Is the probability that a given user will be unable to enroll in a biometric system

FMR and FNMR are dependent variables and their relationship to one another can be described by the Receiving Operating Characteristic Curve (ROC) shown in Figure 1.

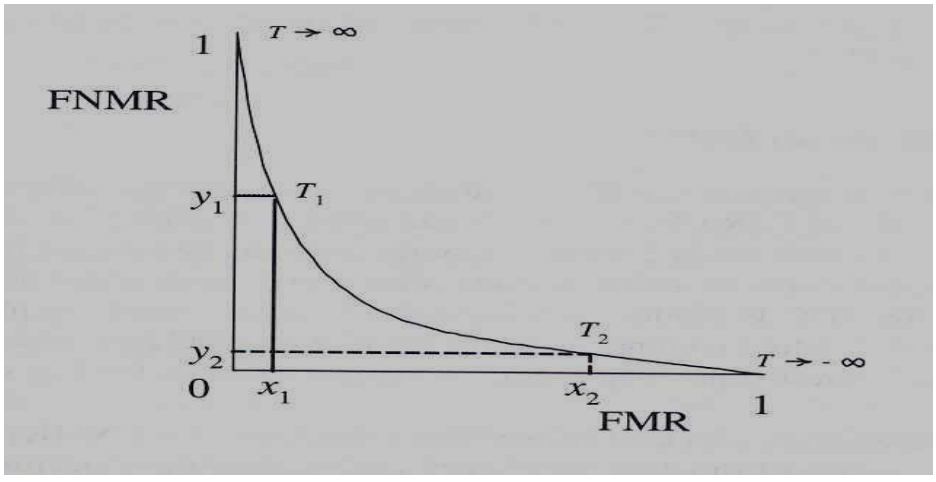

Figure 1: ROC curve[1]

By looking at Figure 1, it can be concluded that the lower the probability that imposters can be accepted as a genuine users (i.e. "x_1", low FMR implies high security), the higher the probability that genuine users will be rejected (i.e. "y_1", high FNMR implies inconvenience for genuine users). Conversely the higher the probability that imposters can be accepted as genuine users (i.e. "x_2", high FMR implies low security) the lower the probability that genuine users will be rejected (i.e. "y_2", low FNMR implies convenience for genuine users). In biometric systems, a trade-off between security and convenience is always present; any setup of the operating point (i.e. "T_1" or "T_2") will inherently modify the relationship between the FMR and the FMNR. Accuracy and performance may diminish as the one-to-many relationship database size increases, this situation may require human intervention via exception handling to make a positive identification.

Statement of the problem

Prior to implementation, a Biometric Information System (BIS) must undergo a training process aimed at optimizing the ability of the system to recognize the face. This optimization process has come under scrutiny with what is known as the "other-race effect". Phillips, Jiang, Narvekar, Ayyad, and O'Toole[2] suggest that "our ability to perceive the unique identity of other-race faces is limited relative to our ability to perceive the unique identity of faces of our own race." At this point, it is not clear whether or not every United States (US) based FRS is calibrated properly to minimize or prevent biases associated with complexion. During the training process, the database demographics used to improve the accuracy may be dissimilar from the demographics where the algorithm will be deployed. Phillips et al.[3] maintain that "understanding the stability of algorithm performance for populations of faces that vary in demographics is critical for predicting face recognition accuracy when application venues vary in their demographic structure."

The ability of the face recognition algorithm to discriminate noisy image data into several classes (persons) can also influence accuracy. The noise can be attributed to varying conditions of illumination, pose, and distance[4] Although extensive research has focused on frontal face recognition with adequate illumination, less has been done on algorithm biases resulting from the manipulation of pose, illumination, and distance from the image[5]

Biometric Face recognition Algorithms

An algorithm can be defined as a set of mathematical or logical instructions that, once coded into binary machine language, can be executed by the computer. The coded algorithms are then linked together to form software that control hardware functions. Norris and Armstrong, cited by Introna and Wood[6] explain algorithmic surveillance as computerized surveillance that makes use of step-by-step instructions to compare the captured data to other data and provide matches. An example of algorithmic surveillance would be a closed circuit television system capable of capturing faces with a camera and comparing those faces with known offenders. In recent years, automated systems designed for the direct monitoring of humans based on a unique physical trait has seen a dramatic rise in all aspects of our economy. Of particular interest is the

silent technology of the face recognition algorithm and society's inability to examine its inner workings for potential biases.

As Introna and Wood[6] point out, even if the code was open for inspection it would be impossible to follow in operation as it flows through multiple layers of translations for its execution. It should also be noted that most algorithms are based on sophisticated methods that can be interpreted and understood by a limited number of experts. These factors only add to the obscurity of the algorithms and increase the need for society to question their legitimacy. According to Gross et al (as cited by Introna and Wood[6]) identifies two categories of algorithms currently in use: Image template algorithms and Geometry feature-based algorithms.

Image template algorithms: use a template-based method to calculate the correlation between a face and one or more standard templates to estimate the face identity. These standard templates tend to capture the global features of a gallery of face images. Thus, the individual face identity is the difference between (or deviation from) the general or 'standard' face. Geometry feature-based algorithms capture the local facial features and their geometric relationship. They often locate anchor points at key facial features (eyes, nose, mouth, etc), connect these points to form a net and then measure the distance and angle of the net to create a unique face 'print'. Even though both approaches are distinct, they do share one aspect known as reduction. Reduction is the method by which algorithms make efficient use of processing power and storage space by reducing the face image into a numeric representation as small as 84 bytes. At this point, the algorithm discards certain information for the sake of retaining others. Introna and Wood[6] explain how reduction affects the performance of algorithms.

Template based algorithms: In these algorithms certain biases become built into the standard template. It obviously depends on the gallery used to create the standard template as well as the range of potential variations within a population. Feature based algorithms: These algorithms do not have an initial bias. However, because of the reduction the 'face prints' generated are in close proximity to each other. Thus, as the gallery database increases more and more face prints are generated in ever diminishing proximity, thereby making the discrimination required for the recognition task more difficult. It also makes the system depend on good quality images. In addition to this it will tend to be better at identifying those that are more distinctive, or less similar, to those already in the database. In either case, there would be an expectation of bias results emerging from the reduction process. Even though the Facial Recognition Vendor Test[7] (FRVT) was conducted to evaluate the efficiency and effectiveness of an assortment of algorithms, it was not intended to uncover inherent biases.

Taking into account the above conditions including poor image quality due to illumination variations, poor camera angle, and image reduction, the probability of an image match is very low. In order to reduce the effects of these factors, operators may choose to increase the False Acceptance Rate (FAR) to a level where a higher number of individuals are subjected to governmental scrutiny in an attempt at identifying others. The perceived concept behind the template based bias is the ability of the algorithm to recognize a race of faces from its training database more accurately than faces of other dissimilar appearing races. This concept has been given the psychological term "other-race effect" and centers around the hypothesis of the amount of contact or experience one has with the other race face may predict the size of the other-race effect.

Furl, Phillips, and O'Toole[2] point out that the accuracy advantage in recognizing own- versus other races leads to the commonly heard statement about other race faces that "they all look alike to me." The other-race effect suggests that people find it more difficult to recognize the uniqueness of individuals of other race faces than their own race faces. The purpose of their study is to determine how susceptible face recognition algorithms are of exhibiting the other-race effect. Researchers believe that by studying the psychological manner by which face representations are processed and retrieved by humans, they might gain insight into the variety of algorithm training methods being utilized that may result in biased algorithms. After researchers analyze several diverse algorithms, they conclude that a small number of them exhibited signs of the other-race effect. This would indicate that experience alone was not the only factor that contributed to face recognition accuracy in different races of faces.

The basis for a methodology for testing face recognition algorithms for the presence of an other-race effect is due to the manner in which the algorithm is trained and the demographic makeup where it will be used. In other words, the database that is used to optimize the algorithm for accuracy does not necessarily represent the human demographic category where the system will be used. Believing that some of the underlying contributing factors causing the other-race effect in humans may also apply to algorithms, Phillips et al[3] conducted two experiments where performance is compared for algorithms and humans on matching identity on pairs of faces. The identity matching task consists of a human or algorithm being presented with two face images and a response must be given with a degree of confidence indicating whether the faces are identical or different. This is reflective of the biometric verification process.

Phillips el al[3] conclude that "demographic origin of face recognition algorithms and the demographic composition of a test population interact to affect the accuracy of the algorithms." This would indicate algorithm performance variations when deployed over dissimilar population demographics. To help detect some of the pitfalls associated with face recognition, certain biometric best scientific practices have been established as guidelines for conducting technical performance testing. As Mansfield and Wayman[8] explain in version-2 of Biometric Testing Best Practices, technical performance testing involves attempting to determine the underlying causes of error and throughput performance as it relates to decisions involving false positives, false negatives, and failure-to-acquire rates. A document of this nature is necessary due to the high number of conflicting and contradictory biometric testing protocols that have been written in the last decade. It should be noted that the procedures contained in Biometric Testing Best Practices are considered general recommendations and may not be completely followed in all tests. These procedures are established as an initial guide for researchers to consider while formulating their evaluation objective.

Experimental Assumptions and limitations

- The subject population being screened is considered cooperative and there are no significant appearance changes (glasses, hair, aging etc) between the enrolled template and the newly acquired template for the same subject.
- The vendor software developer's kit should be consulted before attempting to determine the correct combination of factors (Illumination, distance and face angle) for maximizing system performance.

- The 3D face models used for this study are only two complexions out of a multitude of complexion variations. Additional testing would need to be performed using the complexion variations from the subject population where the system will be deployed.

Research Methodology

Our research methodology will be capable of showing whether or not there is a significant statistical difference (bias) between mean scores of two dissimilar complexion 3D facial models. In preparation for the research, one 3D facial model stand and two female gender 3D facial models were purchased; one light complexion and one medium complexion. Although the two complexions are not representative of the entire range of possible human complexions, the realistic human complexions and facial features are appropriate for this experiment.

To determine the optimal illumination setting and backdrop color, a light meter (Sekonic L-308s) was purchased. Several light meter readings were taken from Willard airport in Savoy Illinois, Coles County Illinois airport, and the Secretary of State's office in Charleston Illinois. The identical backdrop color used at the Secretary of State's office for identification photos is also used for this experiment. In the lab, illumination levels are manipulated manually to reflect the levels observed at the above locations.

Several steps are followed to determine the distances between the camera and the 3D facial models. First, measurements were taken in one inch increments extending the center length of a rectangular table supporting the web camera is used. To ensure that the table is not being positioned at an angle, the table is position 49 inches from a wall running parallel with the tables. Next, optimal distance levels are determined through several repetitions during the initial FRS enrollment and matching process. The main objective is to determine the minimum and maximum distances while maintaining recognition. This methodology of measurement is recommended for all systems being tested. Finally, distance points for our experiment are marked at 14 inches, 31 inches, and 40 inches. The 31 inch factor level, although not centered between the minimum and maximum factor levels, is chosen because it allows for the optimal score when combined with the optimal angle and optimal illuminance levels. It should be noted that these distance levels are exclusive to the current system being evaluated and should be reconfigured for each FRS being evaluated.

Angle levels are determined by creating a circle several inches larger than the diameter of the 3D facial model stand base. A compass is then used to mark off the appropriate degrees for this experiment. Using the same method for determining the optimal distances, the optimal angles of 0 degrees, 15 degrees right and 15 degrees left are chosen through several repetitions during the initial FRS enrollment and matching process. These outer factor levels are also chosen because the angles are not uncommon or too extreme for use with a cooperative subject. A plumb-bob is utilized to ensure that the 3D model, measured down from the bridge of the nose, is directly above the 0 degree starting point placed on the 3D model stand base. Again, it should be noted that these angle levels are exclusive to the current system being evaluated. A six inch level was used to ensure that the 3D facial model and web camera are level on a horizontal and vertical plane.

Experiment setup

For the two 3D facial models, two mannequin heads were used; one light complexion and one medium complexion. During the experiment, both 3D facial models are supported by an adjustable floor stand measuring 53 inches from its base to the bridge of the 3D facial model nose. This distance is fixed throughout the experiment. The system includes an automatic adjusting web camera, computer, and face recognition algorithm used for digital captures. Factors under consideration are illumination, distance, and angle. Three illuminations levels (L1,L2 and L3), three distances (D1,D2 and D3), and three angles(A1,A2 and A3) are all combined and adjusted at random for each 3D facial model. Illumination levels are set to 160 lumens for L1, 320 lumens for L2, and 640 lumens for L3 to allow for illumination adjustments on 3 levels. Distances between the 3D facial models and the web camera are set to 14 inches for D1, 31 inches for D2, and 40 inches for D3. This allows for distance adjustments to be made on three levels. The floor stand base is marked at 15 degrees right for A1, 0 degrees for A2, and 15 degrees left for A3 to allow for 3D facial model angle adjustments on 3 levels.

Verilook 3.1[9] manufactures the face recognition algorithm currently being using for this study. Verilook face recognition GUI application displays the digital captures as a numeric score. At this point, it should be noted that the manufacturer default settings are used for this experiment. The testing methodology being presented here is best suited for FRS where the optional user defined settings have already been set to a level that fits the user's security needs.

Six replications are performed on each 3D model, producing a total of 162 scores per 3D facial model. For each data collection iteration, one 3D facial model is placed on the 3D model floor stand in front of a light blue background similar to the one used in a typical governmental office. A high definition web camera, extending 53 inches from the floor, is then placed in front of the 3D facial model. After the initial enrollment process, each randomized matching iteration response is scored and recorded in the appropriate three-factor-three-levels matrix cell. This process is carried out over a period of several days. Once the light complexion and medium complexion 3D facial model response scores are recorded into the three-factors-three-levels matrix, all scores are entered into DOE PRO[10] statistical software for the purpose of conducting statistical analysis.

DOE PRO[10] for excel is a tool for creating two and three level designs using full factorial and fractional factorial experiments. DOE PRO[10] also includes design analysis featuring like multiple response regression modeling, analysis of variance (ANOVA), and factor interaction plots. t-test, multiple regression, and ANOVA techniques are utilized to make comparisons between the light complexion 3D facial model and the medium complexion 3D facial model. The t-test for statistical significance is performed using traditional statistical calculation techniques.

Research Results

While conducting a side-by-side comparison using the multiple response prediction equations, we find only slight differences in the sign and coefficient for each factor. The Light complexion multiple regression equation is:

$$\hat{Y} = 137.5 + 3.93L - 17.91D - 4.14A - 7.44LD - 5.069LA$$
$$+ 5.792AD - 6.64LDA - 15.26L^2 - 39.59D^2 - 47.92A^2$$

The medium complexion multiple regression equation is:

$$\hat{Y} = 141.98 + 12.139L + 10.52D + 7.25A - 7.97LD + 5.86LA$$
$$- 8.4AD - 7.125LDA - 7.97L^2 - 39D^2 - 62A^2$$

Ỹ represents the multiple regression mean score response taken from the three factors in study: Illumination Level (L), Distance (D) and Angle (A) and their respective interactions.

DOE PRO[10] also provides the percentage contribution of our three factors in the form of an ANOVA table as follows

	Score					
Source	**SS**	**df**	**MS**	**F**	**P**	**% Contrib**
Illumination(L)	28499.8	2	14249.9	203.958	0.000	13.66%
Distance (D)	56579.6	2	28289.8	404.912	0.000	27.12%
Angle(A)	83266.9	2	41633.5	595.899	0.000	39.91%
LD	6146.3	4	1536.6	21.993	0.000	2.95%
AL	3699.8	4	925.0	13.239	0.000	1.77%
AD	15599.6	4	3899.9	55.819	0.000	7.48%
LDA	5403.6	8	675.4	9.668	0.000	2.59%
Error	9432.000	135	69.867			4.52%
Total	208627.605	161				

ANOVA – Light Complexion

	Score					
Source	**SS**	**df**	**MS**	**F**	**P**	**% Contrib**
Illuminat (L)	10313.6	2	5156.8	55.291	0.000	3.58%
Distance (D)	81452.3	2	40726.2	436.664	0.000	28.25%
Angle(A)	148131.0	2	74065.5	794.126	0.000	51.37%
LD	5054.1	4	1263.5	13.547	0.000	1.75%
AL	4407.0	4	1101.7	11.813	0.000	1.53%
AD	20128.2	4	5032.1	53.953	0.000	6.98%
LDA	6256.8	8	782.1	8.386	0.000	2.17%
Error	12591.000	135	93.267			4.37%
Total	288334.000	161				

ANOVA: Medium Complexion

On the above ANOVA tables, the Sum of Squares terms (SS) and the Mean Square (MS) are used to build the "F" model distribution. which gives zero probability (Column "P") to the factors shown in the table which making L,D,A and the interactions LD, AL,DA and LDA significant enough to be part of the multiple regression equation. Nevertheless, comparisons between the two regression equations show that the interaction contribution for each model is roughly identical. The ANOVA output analysis shows that the distance factor in the multiple regression equations is almost similar for each complexion. Therefore, in this particular experiment, the distance seems to be a non-contributing factor for the complexion differences. The ANOVA tables also show that illumination(L) and angle(A) are the two factors that we can observe significant differences in percentage contributions. Nevertheless, the overall score represented by the regression equation output $\hat{Y}$ will be analyzed to show the expected statistical differences.

Finally, the t-test is performed in order to determine if there is any statistically significant difference between the two complexion scores.

Results from the degrees of freedom formula

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2},$$

are calculated where s_1^2 and s_2^2 represent the sample variances and n_1 and n_2 represent the number of runs for the light and medium complexion 3D facial models respectively. .

Results of the two sample t-test calculated from the formulas for statistical significance and confidence intervals are expressed as

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \text{ and } \left(\bar{x}_1 - \bar{x}_2\right) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Where $\bar{x}_1$ and $\bar{x}_2$ are the sample mean scores for each complexion. The t-test of significance gives a probability of $P > 0.25$ for a confidence level $\alpha = 0.05$

The result for the two-sample t-test of significance gives a probability "P" between 0.20 and 0.10 ($0.20 < P < 0.10$). The two sample t-test gives a confidence interval for the mean difference between the two complexion scores as $0.29 \pm (4.24)$, at 95% confidence. In this particular FRS, the t-test values indicate no statistically significant bias between the mean scores recorded from the light complexion 3D facial model and from the medium complexion 3D facial model.

Educational methodology

During the last three semesters, we have incorporated this testing methodology in courses like "Design for Quality" and "Biometric Security". Several design of experiments techniques (DOE) are introduced to the student in order to test significant factor in the authentication process. The testing methodology has also been used in the "Biometric Security" course final projects. The testing methodology presented can be summarized as follows:

- Introduce the student to the basics of Design of Experiments (DOE)
- Review of statistical Techniques: Confidence Interval, t- test, Regression and ANOVA
- Introduce the student to different randomization techniques
- Discuss the rationale for choosing the factors to be tested in the experiment
- Build the experiment set-up
- Introduce the student to the face recognition software and related equipment
- Conducting the experiment in order to minimize noise and/or bias
- Discuss the result appropriately

At this point several thesis and research projects have been conducted with great success. The possibility to offer services to private or governmental institutions in order to test if biases are present in their FRS remains a possibility.

Conclusion

One important benefit of this paper may be to provide private or governmental institutions with valuable information into how demographic and environmental conditions may lead to false negatives. The research could also determine, through statistical analysis, whether or not an in-service FRS is significantly effected by internal and external conditions, resulting in a bias response toward different complexions. This paper presents a system testing methodology, based on best practices, to detect complexion biases in FRS.

In our particular testing system no biases were found. Nevertheless, we are confident that our methodology will detect biases if biases are present. Although earlier studies have shown the presence of the other-race effect, our study yielded modest comparative results. It is important to note that our FRS default settings where not manipulated during testing. Although our testing methodology did not show a significant bias, changing the FRS default settings may result in a different outcome.

Default settings are typically changed by the customer during the initial FRS testing phase. This provides the customer more flexibility in meeting their security needs. The methodology of testing if biases are present in a FRS has been incorporated successfully in our graduate courses. This methodology gives our graduates the opportunity to move from a pure theoretical statistics concept to an applied statistics project

References

[1] Bolle, R. M., Connell, J. H., Pankanti, S., Ratha, N. K., & Senior, A. W. (2004). *Guide to Biometrics.* Hawthorne, NY: Springer-Verlag.

[2] Furl, N., Phillips, P. J., & O'Toole, A. J. (2002). Face recognition algorithms and the other-race effect : Computational mechanisms for a developmental contact hypopaper. *Cognitive Science, 26,* 797-815.

[3] Phillips, P. J., Jiang, F., Narvekar, A., Ayyad, J., & O'Toole, A. J. (2002). An Other-Race Effect for Face Recognition. *Journal of Late X Class Files, 1*(8), 1-7.

[4] Kinage, K. S., & Bhirud, S. G. (2008). Racial inconsistency in face recognition. *SPIT-IEEE Colloquium and International Conference, 1,* 78-81.

[5] Gross, R., Baker, S., Matthews, I., & Kanade, T. (2005). Face recognition across pose and illumination. *Handbook of Face Recognition,* 193-216.

[6] Introna, L. D., & Wood, D. (2004). Picturing algorithmic surveillance: The politics of facial recognition systems. *Surveillance and Society,* 177-198.

[7] P.J. Grother, R.J. Micheals and P. J. Phillips Face Recognition Vendor Test 2002 Performance Metrics, Proceedings 4th International Conference on Audio Visual Based Person Authentication, 2003.

[8] Mansfield, A. J., & Wayman, J. L. (2002). Best practices in testing and reporting performance of biometric devices. *Centre for Mathematics and Scientific Computing*.

[9] VERILOOK 3.1 by Neurotechnology: http://www.neurotechnology.com/verilook.html

[10] DOE PRO by Sigma Systems: http://www.sigmazone.com/doepro.htm