1-1-2010

Facial Recognition System Screening Evaluation Methodology For Complexion Biases

Harold Harris

Eastern Illinois University

This research is a product of the graduate program in [Technology](#) at Eastern Illinois University. [Find out more](#) about the program.

Recommended Citation

Harris, Harold, "Facial Recognition System Screening Evaluation Methodology For Complexion Biases" (2010). *Masters Theses*. 703.
<http://thekeep.eiu.edu/theses/703>

This Thesis is brought to you for free and open access by the Student Theses & Publications at The Keep. It has been accepted for inclusion in Masters Theses by an authorized administrator of The Keep. For more information, please contact tabruns@eiu.edu.

THESIS MAINTENANCE AND REPRODUCTION CERTIFICATE

TO: Graduate Degree Candidates (who have written formal theses)

SUBJECT: Permission to Reproduce Theses

The University Library is receiving a number of request from other institutions asking permission to reproduce dissertations for inclusion in their library holdings. Although no copyright laws are involved, we feel that professional courtesy demands that permission be obtained from the author before we allow these to be copied.

PLEASE SIGN ONE OF THE FOLLOWING STATEMENTS:

Booth Library of Eastern Illinois University has my permission to lend my thesis to a reputable college or university for the purpose of copying it for inclusion in that institution's library or research holdings.

Author's Signature

Date

I respectfully request Booth Library of Eastern Illinois University **NOT** allow my thesis to be reproduced because:

Author's Signature

Date

This form must be submitted in duplicate.

Facial Recognition System Screening Evaluation Methodology for

Complexion Biases

(TITLE)

BY

Harold Harris

THESIS

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF

Masters of Science in Technology

IN THE GRADUATE SCHOOL, EASTERN ILLINOIS UNIVERSITY
CHARLESTON, ILLINOIS

2010

YEAR

I HEREBY RECOMMEND THAT THIS THESIS BE ACCEPTED AS FULFILLING
THIS PART OF THE GRADUATE DEGREE CITED ABOVE

THESIS COMMITTEE CHAIR DATE

Sept 5/2010

THESIS COMMITTEE MEMBER DATE

9/1/10

THESIS COMMITTEE MEMBER DATE

DEPARTMENT/SCHOOL CHAIR DATE
OR CHAIR'S DESIGNEE

Sept. 1, 2010

THESIS COMMITTEE MEMBER DATE

THESIS COMMITTEE MEMBER DATE

Running head: METHODOLOGY

Facial Recognition System Screening Evaluation Methodology for
Complexion Biases

Harold Harris
Eastern Illinois University
School of Technology

Abstract

Over the years, facial recognition systems (FRS) have come under scrutiny from watchdog groups who voice their complaints concerning the potential existence of FRS biases towards certain cultures of people when deployed in security screening situations. Psychology has used the term "other-race effect" to describe this occurrence in humans. Researchers of FRS have used the theory behind the other-race effect to describe the occurrence of bias associated with algorithms tested during the Face Recognition Vendor Test of 2006. In this thesis, we develop a scientific methodology based on the factors of illumination, distance, and angle to evaluate whether or not a chosen system exhibits a significant bias towards a light complexion 3D facial model over a medium complexion 3D facial model. In our particular test, chosen parameters, and statistical software (DOE PRO), our methodology shows that there is no statistically significant bias between the sample mean of a light complexion 3D facial model and the sample mean of a medium complexion 3D facial model. Identical statistical results can not be guaranteed for all FRS, although the methodology can be applied to any FRS to determine if a complexion bias exists.

Contents

Abstract.....	ii
Contents.....	iii
List of Figures.....	iv
Acknowledgements.....	v
Introduction.....	1
Statement of the Purpose.....	3
Statement of the Problem.....	3
Assumptions.....	4
Delimitations.....	4
Limitations.....	5
Hypothesis.....	5
Definition of Terms.....	5
Literature Review.....	7
Research Methodology.....	17
Research Results.....	26
Discussion.....	29
Summary.....	31
Recommendations.....	33
Appendix.....	35
Reference.....	41

List of Figures

Figure 1: 1:one relationship.....	2
Figure 2: 1:many relationship.....	2
Figure 3: Two 3D mannequin models.....	19
Figure 4: Three distance levels.....	20
Figure 5: Three angle levels.....	21
Figure 6: Verilook options screen.....	22
Figure 7: Lab setup.....	23
Figure 8: Doe Pro regression table.....	24

Acknowledgements

I would like to show my gratitude to those who have made valuable contributions to this thesis. Their support and guidance throughout this process helped to make this thesis possible.

First of all, I would like to thank Dr. Chinchilla for introducing me to the area of biometrics and allowing me the opportunity to expand my knowledge through quantitative research. Without his dedication to learning and steadfast support, it would not have been possible for me to complete this daunting task.

I would like to extend a very special thanks to Dr. Izadi and Dr. McDonald who both graciously accepted my invitation to be important contributing members of my defense committee. Their constructive criticism throughout his process has greatly improved the quality of this thesis.

Finally, I would like to thank Kelli Davis for providing her unwaivering support during my pursuit of a master's degree at Eastern Illinois University.

Introduction

Traveling by air, accessing restricted areas, or conducting law enforcement surveillance are just a few examples that require recognition and verification of identity. Earlier airline industry passenger screening methods used the computer-assisted passenger pre-screening system (CAPPS). Although major airlines are still using some form of CAPPS today, civil libertarian groups argue that its practice constitutes racial profiling and is prone to high numbers of innocent people being detained.

Since the tragic events of September 11 2001, interest in the field of biometric identification systems has risen at an astonishing rate. As of 2001, 51 percent of the biometric revenue received was from the physical access control category of authentication while 41 percent of revenue received was from the logical access control category of authentication.

A biometric identification system is considered a process of automated pattern recognition whereby a user's specific physiological or behavioral characteristics are matched against a digitally enrolled template. Most enrollments last the entire life of the system although changes in a user's physiological features may warrant the need for additional enrollments.

Two modes of operation are possible with biometric identification systems. The first mode of operation is verification or authentication. In this

mode, the system attempts to establish a 1:1 relationship with a user who is claiming to be someone previously enrolled in the database. The system simply confirms or denies the 1:1 relationship by comparing the biometric features to a specific digital template stored in the database. Figure 1 illustrates the 1:1 mode of operation in which the user claims an identity.

Figure 1: 1:1 relationship comparing a sample against a stored template.

The second mode of operation is identification. In this mode, the system attempts to establish a 1:many relationship with a user to determine if that user is among the entire enrolled database. There is no claim of identity for a 1:many relationship. This mode can be seen in Figure 2.

Figure 2: 1:many relationship comparing a sample against a database containing many templates.

Accuracy and performance begins to diminish as the 1:many relationship database size increases. This situation requires human intervention exception handling to make a positive identification.

Statement of the Purpose

The purpose of this study is to develop an evaluation methodology that can be successfully applied to face recognition systems. This evaluation methodology, when used with dissimilar complexion 3D facial models and statistical software, can aide in determining whether or not the system is biased.

Statement of the problem

Prior to implementation, the 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 (2002) 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 wheather or not every 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. (2002) 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."

Also, 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 (Kinage & Bhirud, 2008, p. 78). 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 (Gross, Baker, Matthews, & Kanade, 2005).

Assumptions

The subject population being screened is considered cooperative and there are no significant appearance changes between the enrolled template and the newly acquired template for the same subject.

Delimitations

Although there are an infinite number of angle, lighting, and distance factor combinations, the vendor software developer's kit should be consulted before attempting to determine the correct combination for maximizing system performance.

Limitations

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.

Null Hypothesis

Applying our testing methodology to any selected facial recognition system will result in no significant bias in mean scores when comparing dissimilar complexion 3D facial models.

Research Hypothesis

Applying our testing methodology to any selected facial recognition system will result in a significant bias in mean scores when comparing dissimilar complexion 3D facial models.

Definition of Terms

Enrollment: The act of capturing and indoctrinating a user's biometrics into a biometric system.

False Acceptance Rate (FAR): The probability that a biometric system will incorrectly identify or will fail to reject an imposter.

False negative rate (FNR): In a screening type situation a subject is considered legitimate when in reality the subject is most wanted.

False positive rate (FPR): In a screening type situation a subject is considered most wanted when in reality the subject is legitimate.

False reject rate (FRR): The probability that a biometric system will fail to identify an individual who has previously been enrolled.

Matching score: A measurement that determines the degree of similarity in features between a given sample and a stored template.

Operational evaluation: The purpose of this testing method is to determine the performance of a complete biometric system in a specific application environment with a specific target population.

Scenario evaluation: The purpose of this testing method is to use a real-world application and population so that the biometric device can be tested on representative users in a realistic environment.

Technology evaluation: The purpose of this testing method is to compare two competing algorithms using a standardized database and a universal sensor.

Threshold: The point at which it becomes reasonably certain that a biometric sample matches a particular reference template.

Literature Review

An algorithm can be defined as a set of mathematical or logical instructions that, after being 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 (as cited by Introna and Wood, 2004, p. 180) explains algorithmic surveillance as computerized surveillance that make 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 CCTV system that is 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 (2004, p. 183) 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 proprietary and 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, 2004, p. 185) identified two categories of algorithms currently in use.

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 (2004, p. 186) explain how reduction effects 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 (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 false acceptance rate to a level where a

higher number of individuals are subjected to governmental scrutiny in an attempt at identifying others.

FRVT 2002 uncovers numerous identification biases but fails to check race due to the fact that the database largely consisted of Mexican non-immigrant visa applicants. Givens et al (as cited by Introna and Wood, 2004, p. 190), using Principal Component Analysis (PCA), also found race biases in FRVT 2002 in addition to the identification biases previously uncovered.

Their results conclude that Asians and African-Americans are easier to recognize than whites, older people are easier to identify than younger people, and other complexion people are easier to identify than clear complexion people.

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 (2002, p. 798) 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.

Furl, Phillips, and O'Toole (2002, p. 800) also believe that a practical reason for studying the other-race effect in algorithms is due to their increased deployment for law enforcement applications and the extent to which accuracy varies for different races of faces.

After researchers analyze several diverse algorithms, they conclude that a small number of them exhibited signs of the other-race effect. This would indicated that experience alone was not the only factor that contributed to face recognition accuracy in different races of faces.

This theory was tested by Feinman and Entwisle (as cited by Furl et al., 2002, p. 799) using 288 African American and Caucasian children from integrated and segregated schools. The subjects where from grade 1-3 and 6. The test was developed using photographs of African American and Caucasian children along with an old/new face recognition task.

Test results revealed a trend toward a higher other-race effect for children of segregated schools than for children of integrated schools. When researchers took into account the neighborhood integration mix from where the children resided,

racial composition proved to be a significant factor. It was determined that children of segregated neighborhoods showed a greater magnitude of the other-race effect than children of integrated neighborhoods.

An identical study of 120 African American and Caucasian adolescents conducted by Cross, Cross, & Daly (as cited by Furl et al., 2002, p. 799) found similar results. The study found that Caucasians from integrated neighborhoods were less likely to exhibit the other-race effect than their counterparts residing in segregated neighborhoods.

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 (2002) 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.

Experiment 1 used East Asian and Western algorithms to match face identities for all available East Asian and Caucasian face pairs taken from the Face Recognition Vendor Test 2006 database. A low false acceptance rate typical for security applications was set for the initial test.

Experiment 2 used the East Asian and Western algorithms as a benchmark against the performance recorded by the humans of East Asian and Caucasian descent. Both categories of humans provided a control variable to which researchers could confirm the presence of the other-race effect for the test sample.

Phillips et al (2002) 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.

Also, the algorithm findings are considered to be analogous to the findings recorded for human face recognition. Phillips et al (2002) maintain that the performance of state-of-the-art face recognition algorithms vary with respect to the demographic origin of the algorithm and the demographic structure of the training population.

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 (2002, p. 2) 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. There are three evaluation methods that can be conducted on a biometric FRS for performance testing: technology, scenario, and operational.

Phillips et al (as cited by Mansfield and Wayman, 2002, p. 3) gives a general definition for the three evaluation methods.

Technology evaluation: The goal of a technology evaluation is to compare competing algorithms from a single technology. Testing of all algorithms is carried out on a standardized database collected by a "universal sensor." Nonetheless, performance against this database will depend upon both the environment and the population in which it is collected.

Scenario evaluation: The goal of scenario testing is to determine the overall system performance in a prototype or simulated application. Testing is carried out on a complete system in an environment that models a real-world target application of interest. Care will be required that data collection across all test systems is in the same environment with the same population.

Operational evaluation: The goal of operational testing is to determine the performance of a complete biometric system in a specific application environment with a specific target population. In general, operational results will not be repeatable because of unknown and undocumented differences between operational environments.

During the testing phase input images are collected, which will be later used for template generation at enrollment and score matching against a defined threshold. The collection of images can be used immediately for an online enrollment or identification attempt, or used later for an offline enrollment and identification.

In planning the biometric evaluation, one should first have a solid idea of what the evaluation is trying to determine. Next, one needs to decide the appropriate evaluation method to be used. Mansfield and Wayman (2002, p. 9) maintain that these two decisions will help form the essential basis for developing an appropriate test protocol, help to pinpoint appropriate environmental controls, and help establish appropriate test sizes.

Also during the evaluation planning stage, additional attention should be given to the significant factors that will influence the performance measurement. As Mansfield and Wayman (2002, p. 9) point out, these factors can be implicitly or explicitly divided into one of four categories.

1. Factors that are incorporated into the structure of the experiment as independent variables so that we can observe the effect the factors have on the response.
2. Factors controlled to become part of the experimental conditions that are unchanged throughout the evaluation;
3. Factors that are considered to be nuisance variables are randomized out of the experiment;
4. Factors judged to have negligible effects will be ignored.

Due to the fact that evaluation performance measurements can vary with the application, environment, and population, these aspects should be taken into account and decided in advance. User and environmental factors that should be taken into consideration while evaluating face recognition systems are ethnic origin, profiles, distance to the camera and lighting levels.

Mansfield and Wayman (2002, p. 10) maintain that while attempting to determine which factors to control, a conflict may arise between the need for internal validity (i.e. that performance differences are due only to the independent variable(s) recorded in the study) and external validity (i.e. that the results truly represent performance on the target application).

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.

Research Methodology

In preparation for the research methodology, the following three courses were taken at the college level: (a) research methods with an emphasis on statistics; (b) biometrics with an emphasis on verification, identification, and facial recognition; and (c) industrial design of experiments with an emphasis on a full factorial and the use of different factors that can affect the process and the output of that process.

One mannequin head stand and two female gender mannequin heads are 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.

Illumination

To determine the optimal illumination setting and backdrop color, a light meter (Sekonic L-308s) was purchased. Several light meter readings are taken from Willard airport in Savoy Illinois, Coles County Illinois airport, and the Secretary of State's office in Coles County 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.

Distance

Several steps are followed to determine the distances between the camera and mannequin 3D model. First, measurements were taken in 1 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.

Finally, measurement points 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

Angle levels are determined by creating a circle several inches larger than the diameter of the mannequin 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 mannequin head, measured down from the bridge of the nose, is directly above the 0 degree starting point placed on the mannequin 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 mannequin head and web camera are level on a horizontal and vertical plane.

This thesis offers an evaluation methodology that can be used to determine whether or not a FRS exhibits a biased response when given multiple dissimilar complexion subjects. Experiment setup involves two 3D life size facial models and the FRS.

Experiment setup

For the two 3D facial models, two mannequin heads were used; one light complexion and one medium complexion (see Figure 3).

Figure 3: Two 3D mannequin models showing two dissimilar skin complexions.

During the experiment, both 3D heads are supported by an adjustable mannequin floor stand measuring 53 inches from the base to the bridge of the 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, three distances, and three angles 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 models and web camera are set to 14 inches for D1, 31 inches for D2, and 40 inches for D3 to allow for adjustments on three levels (see figure 4).

Figure 4: Three distance levels associated with the FRS being evaluated.

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 model angle adjustments on 3 levels (see Figure 5).

Figure 5: Three angle levels associated with the FRS being evaluated.

Verilook 3.1 manufactures the face recognition algorithm currently being using for this study. The Verilook application has three operation modes:

1. Enrollment. This mode processes the face image, extracts features, and writes them to the database.
2. Enrollment with feature generalization. The mode generates the generalized face features collection from a number of face templates of the same person. Each face image is processed and features are extracted. Then collections of features are analyzed and combined into a generalized features collection, which is written to the database. The face recognition quality increases if faces are enrolled using this mode.
3. Matching. This mode performs new face image matching against face templates stored in the database.

Verilook face recognition GUI application displays the digital captures as a numeric score. Figure 6 is a screen shot

of the Verilook user defined settings. At this point, it should be noted that the manufacturer default settings that appear in figure 6 are used for this experiment. Settings contained in the a) face detection and extraction options box, b) matching options box, c) enrollment options box, and d) generalization options box can all be reconfigured to form additional factors.

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.

Figure 6: Screen shot of the Verilook options screen.

Six replications are performed on each 3D model, producing a total of 162 scores per 3D facial model (see Appendix A). For each data collection iteration, one 3D facial model is placed on the mannequin floor stand in front of a blue background. A web camera, extending 53 inches from the floor, is then placed in front of the 3D model. Figure 7 illustrates the lab setup.

Figure 7: Three lab setup views.

After the initial enrollment process, each random matching iteration response is scored and recorded in the appropriate matrix cell similar to the one located in appendix B. Both matrixes are constructed over a period of four days. Once the light complexion and medium complexion 3D facial model response scores are recorded into the matrix, all scores are entered into Doe Pro statistical software for the purpose of conducting statistical analysis.

Doe Pro for excel is a tool for creating two and three level designs using full factorial and fractional factorial. Doe Pro also includes design analysis featuring Y-hat multiple response

regression modeling, analysis of variance (ANOVA), and multiple plots. Figure 8 illustrates the Doe Pro initial regression table. The regression equation is a statistical technique used to explain or predict the behavior of a dependent variable. In the initial design sheet, items in red are important factors and should be retained. Items in black are considered insignificant factors and should be deleted. This is done by removing the "X" in the active column to the right of the factor.

Figure 8: Doe Pro regression table.

The following three statistical testing techniques are utilized to make comparisons between the light complexion 3D model and the medium complexion 3D model: (a) t-test, (b) multiple regression, and (c) ANOVA.

The t-test, developed by W.S. Gossett, is a tool used to evaluate the hypothesis about group level differences in outcomes. Calculation of the test requires the following components:

1. The average of both samples represented.
2. The standard deviation of both averages.
3. The number of observations in both populations.

Once the t-statistic is calculated, compare the t-value with a standard table of t-values to determine whether the t-statistic reaches the threshold of statistical significance.

Multiple regression is a tool that allows researchers the ability to examine the effect of many different factors on some outcome simultaneously. The general idea of multiple regression is to learn more about the relationship between several independent or predictor variables and a dependent variable. In addition, regression can be used to examine how much a

particular set of independent variables explains differences in some outcome.

ANOVA is a tool that can be used to examine differences among the sample means of several different groups at once, to determine the existence of statistically significant differences between them. ANOVA can also be considered a statistical technique for assessing how nominal independent variables influence a continuous dependent variable.

To determine whether or not the difference in mean response is based on the 3 factor 3 level matrix and not based on chance, an initial test for the degrees of freedom approximation will be calculated. This test will be performed with zero score outliers ($n = 27$) and without zero score outliers ($n_1 = 24$, $n_2 = 23$) in order to determine whether the results are consistent over multiple runs.

Finally, the results from the degrees of freedom test will be used to calculate a two sample T-procedural test for significant and confidence intervals.

Research Results

Appendix C illustrates the Doe Pro $\hat{Y}$ results for the light complexion and medium complexion 3D facial models respectively. When conducting a side-by-side comparison using

the $\hat{Y}$ multiple response prediction equations, we find only slight differences in the sign and coefficient for each factor. This is illustrated by the two matrices located in appendix B and in the $\hat{Y}$ -hat equations given below.

$$\begin{aligned}\hat{Y} = & 137.5 + 3.93A - 17.91B - 4.14C - 7.44AB - 5.069AC \\ & + 5.792BC - 6.64A^2 - 15.26B^2 - 39.59B^2 - 47.92C^2\end{aligned}$$

Light Complexion

$$\begin{aligned}\hat{Y} = & 141.98 + 12.139A + 10.52B + 7.25C - 7.97AB + 5.86AC \\ & - 8.4BC - 7.125A^2 - 39B^2 - 62C^2\end{aligned}$$

Medium Complexion

Doe Pro also gives us the percentage contribution of our three factors in the form of an ANOVA table. The two tables shown in appendix D represent the single and combined contributions. As observed from the light complexion 3D facial model ANOVA table, angle has the highest percentage contribution followed by distance and light. The medium complexion 3D facial model ANOVA table contains a percentage mix similar in order to the light complexion model.

One notable difference between the two ANOVA tables is that the medium complexion ANOVA has slightly higher contribution percentage figures with respect to angle and distance than the light complexion ANOVA. The remaining two-way iterations (AB,

AC, BC) and the remaining three-way iteration (ABC) factor contributions for each model are roughly identical.

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 square roots and n_1 and n_2 represent the number of runs for the light and medium complexion 3D facial models respectively. This formula is used to calculate the degrees of freedom (dof) of approximately 52.02 with zero score outliers.

Results of the two sample t-procedure tests calculated from the formulas for significant and confidence interval are expressed as

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

for 26 dof. Results from the significance test is $P > 0.25$. for the 26 dof.

The same degrees of freedom formula is used without outliers resulting in a dof of 43.6. The result for the two-sample t-procedure test of significance is $0.20 < P < 0.10$ for 22 degrees of freedom. The two sample t-procedure with a confidence interval of 95% is $0.29 \pm (4.24)$.

The t-procedure 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. These research results are in support of our null hypothesis stating that there is no significant bias in mean scores.

Discussion

With respect to appendix B, there are observed similarities and slight differences in row combination iterations when comparing the light complexion matrix and the medium complexion matrix. Most notable similarities are among the zero score outliers. The light complexion matrix zero score outliers occur during the 3 factor combination when illumination is at its lowest (L1), distance is at the closest (D1), and angles are at 15 degrees in both directions (A1, A3). This can be seen using the 3 factor location combination from appendix A.

The light complexion matrix zero outlier also occurs during the combination when illumination is at its lowest (L1), distance from the model is at its furthest (D3), and angle is set to 15 degrees right (A1). It should also be noted that even though the light complexion matrix combination L1,D3,A3 is not a zero outlier, it does contain one of the lowest overall means when compared to the other matrix combination means. The same conclusion can be made for the light complexion matrix combination L2,D1,A1 and matrix combination L3,D1,A1.

When conducting a row by row comparison of the medium complexion matrix zero outlier scores to the light complexion matrix zero outlier scores, we see two identical factor combinations. The light and medium complexion matrix zero score outliers occur at the 3 factor combination L1, D1 and A1 and A3.

Slight deviations from the light complexion matrix occur in the medium complexion matrix at 3 factor combination L1,D3,A1. It is at this combination point where the medium complexion mean score is near its lowest. It should also be noted that where the light complexion 3 factor combination located at L2,D1,A1 is near its lowest mean score, the identical medium complexion 3 factor combination is at zero.

Locate on the high score end of the spectrum, the similarities between the light and medium complexion matrices

center on the following 3 factor combinations of scores:

L1,D2,A2:L2,D2,A2:L3,D1,A2:L3,D2,A2.

Also, the slight differences observed between the light and medium complexion matrices are found among the following 3 factor combinations of scores:

L2,D1,A2:L2,D3,A2:L3,D3,A2.

Overall, there is no extreme mismatching of scores when comparing the light complexion 3 factor combinations against the medium complexion 3 factor combinations.

Summary

In an article from Technology and Liberty (2002), the ACLU contends that the use of intrusive surveillance technologies must be subjected to a process of thoughtful and deliberate scrutiny. As stated earlier in the study, being able to scrutinize and inspect facial recognition systems to determine how they operate is unlikely for the ordinary person.

The ACLU (Technology and Liberty, 2002) also adds that a study conducted by the Department of Defense found the following:

The Department of Defense found very high error rates even under ideal conditions, where the subject is staring directly into the camera under bright lights.

The study also found very high rates of both "false positives" and "false negatives." That suggests that if installed in airports, these systems would miss a high proportion of suspects included in the photo database, and flag huge numbers of innocent people thereby lessening vigilance, wasting precious manpower resources, and creating a false sense of security.

Companies and governmental entities, which already have FRS installed, should consider having their systems evaluated for performance by a reputable 3rd party research group utilizing industry best practices.

One important benefit to this thesis may be to provide these entities with valuable information into how demographic and environmental conditions may lead to false negatives. The research group could also determine, through statistical analysis, whether or not an in service FRS is significantly effected by these conditions, which may result in a bias response. This thesis presents a system testing methodology based on best practices to detect complexion biases in FRS.

Findings for this thesis would suggest that there is no conclusive evidence to indicate a bias associated with the current facial recognition system under evaluation. 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 were not manipulated during testing. Although our testing methodology did not show a significant bias, changing the FRS default setting 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.

Recommendations

This thesis does present some additional experimental considerations for FRS deployed in diverse demographic locations. The first consideration would be to test the outer limits of complexion to determine whether a significant bias in mean score exists. This testing method would be very helpful in predicting possible biases of FRS deployed in international airports. International airports are locations where the population demographics may service light complexion travelers who are indigenous to Europe or dark complexion travelers who are indigenous to Africa.

It may also be beneficial if a third experiment was conducted using either of the two 3D facial models used in this study. The mean score of the third experiment, if closer in

similarity to either mean score from the initial experiments, may provide additional evidence to support the "other-race effect."

Another consideration is to use live subjects for comparison as opposed to 3D facial models. Introducing live subject testing would be the preferred method for an operational evaluation. This method would also allow for a more accurate assessment of FRS performance under a controlled environment. I will leave these research considerations to other researchers to explore, as the debate over the controversial use of facial recognition systems for screening continues to make headlines.

Appendix A

3 Factors 3 Levels								
			Y1	Y2	Y3	Y4	Y5	Y6
L1	D1	A1	125	73	131	127	145	143
		A2	168	57	71	134	112	163
		A3	52	45	7	177	61	69
L1	D2	A1	53	166	174	187	184	38
		A2	3	66	167	195	151	120
		A3	18	150	148	60	37	132
L1	D3	A1	113	152	138	5	6	109
		A2	129	79	122	126	139	153
		A3	58	197	182	34	50	35
L2	D1	A1	128	183	92	91	106	179
		A2	99	173	28	30	33	56
		A3	75	22	105	93	194	115
L2	D2	A1	196	155	26	169	171	114
		A2	157	15	144	89	178	74
		A3	140	164	59	20	78	81
L2	D3	A1	32	25	133	158	77	181
		A2	130	110	186	23	121	111
		A3	42	9	159	192	72	193
L3	D1	A1	156	141	175	190	80	11
		A2	116	88	41	46	16	84
		A3	149	117	180	170	21	29
L3	D2	A1	172	65	24	17	199	82
		A2	146	94	39	70	107	14
		A3	100	108	154	27	10	86
L3	D3	A1	54	161	123	36	1	64
		A2	198	47	76	8	103	137
		A3	12	96	87	62	44	40

L1 = 160 Lux

L2 = 320 Lux

L3 = 640 Lux

D1 = 14 Inches

D2 = 31 Inches

D3 = 40 Inches

A1 = 15° Right

A2 = 0°

A3 = 15° Left

Appendix C

Light Complexion

Y-hat Model		Score			Active
Factor	Name	Coeff	P(2 Tail)	Tol	
Const		137.57	0.0000		
A	Light	3.935	0.0727	0.2000	X
B	Distance	-17.917	0.0000	0.2000	X
C	Angle	-4.139	0.0153	0.3333	X
AB		-7.444	0.0000	1	X
AC		-5.069	0.0000	1	X
BC		5.792	0.0000	1	X
ABC		-6.646	0.0000	1	X
AA		-15.259	0.0000	1	X
BB		-39.593	0.0000	1	X
CC		-47.926	0.0000	1	X
AAB		5.250	0.0121	0.3333	X
ABB		6.639	0.0016	0.3333	X
ACC		7.931	0.0002	0.3333	X
BBC		9.681	0.0000	0.3333	X
BCC		19.875	0.0000	0.3333	X
R ²		0.9284			
Adj R ²		0.9210			
Std Error		10.1182			
F		126.1199			
Sig F		0.0000			
F _{LOF}		7.1764			
Sig F _{LOF}		0.0000			
Source		SS	df	MS	
Regression		193680.3	15	12912.0	
Error		14947.3	146	102.4	
Error _{Pure}		9432.0	135	69.9	
Error _{LOF}		5515.3	11	501.4	
Total		208627.6	161		

Medium Complexion

Y-hat Model		Score			Active
Factor	Name	Coeff	P(2 Tail)	Tol	
Const		141.98	0.0000		
A	Light	12.139	0.0000	0.3333	X
B	Distance	10.528	0.0000	0.3333	X
C	Angle	7.250	0.0154	0.2000	X
AB		-7.972	0.0000	1	X
AC		5.861	0.0004	1	X
BC		-8.403	0.0000	1	X
ABC		-7.125	0.0004	1	X
AA		-7.972	0.0007	1	X
BB		-39.000	0.0000	1	X
CC		-62.000	0.0000	1	X
AAC		-7.250	0.0107	0.3333	X
ACC		-5.278	0.0619	0.3333	X
BBC		10.625	0.0002	0.3333	X
BCC		7.792	0.0062	0.3333	X
R ²		0.9037			
Adj R ²		0.8945			
Std Error		13.7447			
F		98.5178			
Sig F		0.0000			
F _{LOF}		13.5630			
Sig F _{LOF}		0.0000			
Source		SS	df	MS	
Regression		260563.2	14	18611.7	
Error		27770.8	147	188.9	
Error _{Pure}		12591.0	135	93.3	
Error _{LOF}		15179.8	12	1265.0	
Total		288334.0	161		

Appendix D

Anova - Light Complexion

ANOVA TABLE

Source	SS	df	Score			% Contrib
			MS	F	P	
Light	28499.8	2	14249.9	203.958	0.000	13.66%
Distance	56579.6	2	28289.8	404.912	0.000	27.12%
Angle	83266.9	2	41633.5	595.899	0.000	39.91%
AB	6146.3	4	1536.6	21.993	0.000	2.95%
AC	3699.8	4	925.0	13.239	0.000	1.77%
BC	15599.6	4	3899.9	55.819	0.000	7.48%
ABC	5403.6	8	675.4	9.668	0.000	2.59%
Error	9432.000	135	69.867			4.52%
Total	208627.605	161				

Anova - Medium Complexion

ANOVA TABLE

Source	SS	df	Score			% Contrib
			MS	F	P	
Light	10313.6	2	5156.8	55.291	0.000	3.58%
Distance	81452.3	2	40726.2	436.664	0.000	28.25%
Angle	148131.0	2	74065.5	794.126	0.000	51.37%
AB	5054.1	4	1263.5	13.547	0.000	1.75%
AC	4407.0	4	1101.7	11.813	0.000	1.53%
BC	20128.2	4	5032.1	53.953	0.000	6.98%
ABC	6256.8	8	782.1	8.386	0.000	2.17%
Error	12591.000	135	93.267			4.37%
Total	288334.000	161				

References

- Bolle, R. M., Connell, J. H., Pankanti, S., Ratha, N. K., & Senior, A. W. (2004). *Guide to Biometrics*. Hawthorne, NY: Springer-Verlag.
- Chirillo, J., & Blaul, S. (2003). *Implementing Biometric Security*. Indianapolis: Wiley.
- Furl, N., Phillips, P. J., & O'Toole, A. J. (2002). Face recognition algorithms and the other-race effect: Computational mechanisms for a developmental contact hypothesis. *Cognitive Science*, 26, 797-815.
- Gross, R., Baker, S., Matthews, I., & Kanade, T. (2005). Face recognition across pose and illumination. *Handbook of Face Recognition*, 193-216.
- Introna, L. D., & Wood, D. (2004). Picturing algorithmic surveillance: The politics of facial recognition systems. *Surveillance and Society*, 177-198.
- Kinage, K. S., & Bhirud, S. G. (2008). Racial inconsistency in face recognition. *SPIT-IEEE Colloquium and International Conference*, 1, 78-81.
- Lu, X., & Jain, A. K. (2004). Ethnicity identification from face models. *Proceedings of the SPIE*, 5404, 114-123.
- Luis-Garcia, R., Alberola-Lopez, C., & Aghzout, O. (2003). Biometric identification systems. *Signal Processing*, 83, 2540-2557.

Mansfield, A. J., & Wayman, J. L. (2002). Best practices in testing and reporting performance of biometric devices.

Centre for Mathematics and Scientific Computing.

Phillips, P. J., Flynn, P. J., Scruggs, T., Bowyer, K. W., Chang, J., Marques, J., et al. (2005). Overview of the face recognition grand challenge. *National Institute of Standards and Technology*, 1-8.

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.

Phillips, P. J., Moon, H., Rizvi, S. A., & Rauss, P. J. (2000). The Feret evaluation methodology for face-recognition algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(10), 1090-1104.

Technology and Liberty. (2002, February 21). *ACLU opposes use of face recognition software in airports due to ineffectiveness and privacy concerns*. Retrieved April 27, 2010, from American Civil Liberties Union Web Site: <http://www.aclu.org/technology-and-liberty/aclu-opposes-use-face-recognition-software-airports-due-ineffectiveness-and-p>