Presented at the 2008 8th IEEE International Conference on Automatic Face and Gesture Recognition.

Focus on Quality, Predicting FRVT 2006 Performance *

J. Ross Beveridge
Department of Computer Science
Colorado State University
ross@cs.colostate.edu

Geof H. Givens
Department of Statistics
Colorado State University
geof@stat.colostate.edu

P. Jonathon Phillips
National Institute of Standards and Technology

Bruce A. Draper
Department of Computer Science
Colorado State University

Yui Man Lui
Department of Computer Science
Colorado State University

Abstract

This paper summarizes a study carried out on data from the Face Recognition Vendor Test 2006 (FRVT 2006). The finding of greatest practical importance is the discovery of a strong connection between a relatively simple measure of image quality and performance of state-of-the-art vendor algorithms in FRVT 2006. The image quality measure quantifies edge density and likely relates to focus. This effect is part of a larger four-way interaction observed between edge density, face size and whether images are acquired indoors our outdoors. This finding illustrates the broader potential for statistical modeling of empirical data to play an important role in finding and codifying biometric quality measures.

1. Introduction

Understanding factors that influence performance is fundamental to developing, evaluating, and operating face recognition algorithms. This paper describes a statistical analysis that quantifies the effects of multiple factors, covariates, for the Face Recognition Vendor Test 2006 (FRVT 2006). The statistical analysis technique is generalized linear mixed modeling (GLMM).

Previous GLMM work has analyzed how subject covariates, such as gender and age, influence face recognition performance [2, 3]. This paper advances this work by identifying relatively simple image measures that predict dramatic differences in the performance of a state-of-the-art algorithm. The algorithm studied fuses similarity scores from three top performers in FRVT 2006.

Covariates, in the context of this paper, are factors independent of an algorithm that may effect performance; e.g., gender of a person and the size of the face in an image. The goal of covariate analysis is to identify which covariates affect algorithm performance and to quantify those effects. This includes quantifying interactions among covariates.

Subject covariates are attributes of the person being recognized, such as age, gender or race. Subject covariates can be transitive properties of subjects, such as smiling or wearing glasses. *Image covariates* are attributes of the image or sensor, such as size of the face or focus of the camera.

In the field of biometrics, there is considerable interest in identifying good quality measures. Grother and Tabassi [4] define a quality measure as a number that relates an image's quality to a recognition system and should be predictive of performance. Within our framework, we define a quality measures as a covariate that is measurable, is predictive of performance and is actionable.

A measurable covariate can be reliably and consistently computed from an image. The edge density measure to be introduced shortly as a proxy for measuring image focus is a measurable covariate. Other factors that may influence performance, for example hair style, are not easily measured and hence are not good candidates for quality.

An actionable covariate is one over which a biometric application has a degree of control over. For example, potential actionable covariates are size of the face in an image, focus, and whether a person is smiling. Examples of covariates that are not actionable are gender, race, and age.

*The identification of any commercial product or trade name does not imply endorsement or recommendation by the authors or their institutions.

978-1-4244-2154-1/08/$25.00

Figure 1. Examples of controlled lighting, and indoor and outdoor uncontrolled lighting imagery.

Quality measures naturally fit into the GLMM modeling framework. The GLMM quantifies the effect of quality measures and their interactions with other covariates. In addition, actionable covariates do not have to be identified a priori. Rather, one analysis can provide input to assessing impact of quality measures for multiple applications. In applications where the system designers can select a limited number of covariates to manipulate, the model can assist in the selection process.

Our primary finding, described in Section 4.2, is a four-way interaction between focus, face size, and environment. Environment is either outside or indoors in a hallway, and the focus measure is a proxy based upon edge density. Over these four image covariates, the model estimated verification rate varied from 0.1 to 0.9 at a false accept rate of 0.001.

This is a surprising and highly significant scientific finding. The effect of this interaction is greater than the effect of gender, race, and whether a person was wearing glasses. An additional major benefit of the GLMM technique is that our key finding is independent of these other covariates; e.g., the four-way interaction effects performance regardless of gender, race, and wearing of glasses.

All the covariates in the key finding are potentially actionable and hence quality measures. The results of this analysis provide input to algorithm developers about where to concentrate research; and to system designers regarding which image covariates are most important to control.

2. FRVT 2006 Overview

The FRVT 2006 was an independent evaluation of face recognition algorithms administered by the National Institute of Standards and Technology (NIST) [11]. The FRVT 2006 was the latest in a series of U.S. Government sponsored challenge problems and evaluations designed to advance automatic face recognition [8] [9] [10].

This paper analyzes performance on the FRVT 2006 *very high-resolution* image set. The very high-resolution images were acquired with a 6 Mega-pixel Nikon D70 camera. Images were captured under three conditions, see Figure 1. All images in the data set are full face frontal. The *controlled illumination* images were taken in studio conditions with lighting that followed the NIST mugshot best practices [6]. The average face size for the controlled illumination images was 400 pixels between the centers of the eyes. The *indoor uncontrolled illumination* images were taken in hallways and indoor open spaces with ambient lighting. The average face size was 190 pixels between the centers of the eyes (this is over the entire dataset). The *outdoor* images were taken outdoors with ambient lighting. The average face size was 163 pixels between the centers of the eyes.

The FRVT 2006 large-scale experiment report [11] presented results matching controlled illumination images to controlled illumination images, and controlled illumination images to indoor uncontrolled illumination images. This is the first article to report results on matching controlled illumination images to outdoor images for the FRVT 2006.

We analyze the performance of an algorithm that is the fusion of three top performers in the uncontrolled illumination experiment in FRVT 2006. The performance of the fusion algorithm was significantly better than the individual algorithms. For the experiment in this paper, the performance of the fusion algorithm was a verification rate of 0.81 at a FAR of 0.001; the verification rates for the three component algorithms was 0.74, 0.69, and 0.66.

The algorithms were fused as follows. For each algorithm, the median and the median absolute deviation (MAD) were computed from 36,602 similarity scores randomly sampled from a total of 37,443,978 scores. Next, similarity scores for each algorithm were standardized by subtracting the median and dividing by its MAD. Formally, if s_k is a similarity score for algorithm k and s_f is a fusion similarity score for $k = 1, 2, 3$, then $s_f = \sum_k (s_k - \text{median}_k)/\text{MAD}_k$ where median_k and MAD_k are the median and MAD for algorithm k.

Analyzing the performance of the fusion algorithm has two benefits. First, the analysis is done on an algorithm that is better than any of the individual algorithms. Second, attention is focused on the effect of covariates on performance. Most people have an understandable predisposition to focus on how well individual algorithms perform and which performs best. In most circumstances, this is very appropriate. However, in our studies it is desirable to concentrate on on how covariates influence performance in general and presenting results on the fused algorithm serves this purpose well. That said, a complementary study of the individual algorithms is underway.

3. Relating Performance to Covariates

Figure 2 illustrates our modeling approach and will be referenced at several points throughout this section. The left side of the figure indicates that modeling begins by relating

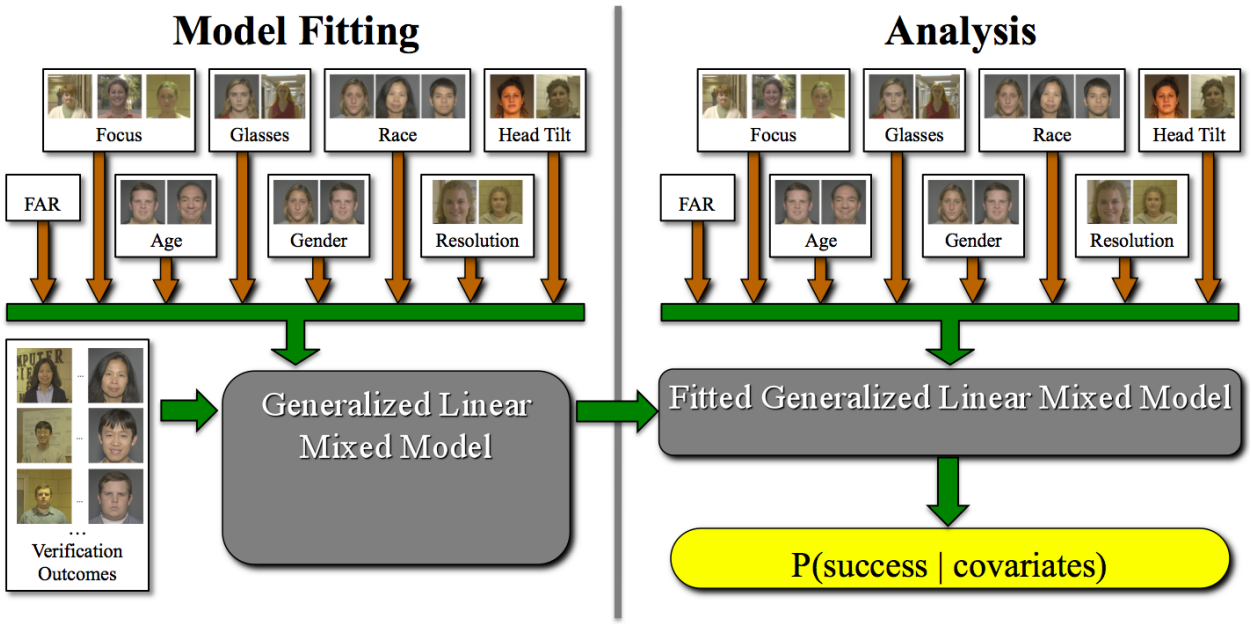

Figure 2. Schematic illustrating the overall flow of information for estimation and uses of the generalized linear mixed model (GLMM). The left panel illustrates the process of fitting the model, whereas the right panel illustrates prediction of performance.

covariates to face recognition outcomes, in this case verification outcomes as defined in Section 3.1. The right side of the figure indicates a fitted statistical model provides a quantitative basis for relating covariates to the probability that a face will be correctly verified given a set of covariates. Section 3.2 discusses the principal covariates used in our model and Section 3.3 introduces the statistical model.

3.1. Verification Outcomes

The performance variable in our analysis is whether or not a matching pair of images, two images of the same person, are correctly verified at one of three possible false accept rates (FARs). The FARs are $\frac{1}{100}$, $\frac{1}{1,000}$ and $\frac{1}{10,000}$. These FAR settings are assigned to match pairs at random, balancing the total number of samples associated with each.

The connection between FAR and verification success or failure is established through the population of non-match scores derived from the images used in the study. Put simply, the population of non-match scores provides us the match score threshold that yields each of the three FARs. A match pair is recorded as a successful verification if it's match score is higher than the corresponding threshold.

3.2. The Covariates

As illustrated on the left side of Figure 2, GLMM analysis begins with us fitting a model whose inputs are the verification outcomes and covariates associated with the match pairs. There are $110,514$ match pairs derived from a population of 345 distinct people. For the controlled lighting there are between 16 and 32 images per person. For the indoor and outdoor images there are between 4 and 16 images per person. For subject covariates such as gender and race there is only one value per match pair. For other match pairs there are two values; e.g., the size of the face in each image.

In this paper we report findings for gender, race, size of the face, degree of focus of a face, wearing glasses, whether images were taken indoors or outdoors, and FAR. Our model had 50 covariates, but these seven covariates produced the most interesting scientific effects.

As is the case with virtually all face recognition applications, a measure of focus has to be computed post hoc from the face in an image. Krotkov [5] advocated the average edge density in an image as an effective after-the-fact measure of focus, showing it did a superior job of predicting quality of focus when compared to other measures including those based on the spectral energy in an image. The edge-density measure is not perfect, as it is sensitive to environmental factors which give rise to high edge density.

Only the face if of interest, so the focus measure we introduce in this analysis is the Face Region In Focus Measure (FRIFM). Figure 3 illustrates how this measure is computed. First, a face is transformed to a standard size. Second, a Sobel edge mask is applied to the image to derive edges. The FRIFM is then simply the average Sobel edge magnitude within an oval defining the region of the face.

Figure 3. Face Region In Focus Measure (FRIFM) values are computed by summing edge density within an oval face mask.

3.3. The Statistical Model

The word *generalized* in generalized linear mixed model refers to the sensible assertion that verification outcomes are Bernoulli distributed, rather than normally distributed as in ordinary linear models. Through a link function (canonically, the logit function in the present case) transforming the mean response, this model allows one to relate verification outcomes to a linear function of the covariates.

Specifically, the verification outcomes are expressed as Bernoulli random variables Y_{iaj} with success probabilities p_{iaj}. The subscripts indicate specific covariates, and here only a sufficient number of covariates has been used to suggest the form of GLMM. In this example, a GLMM may be defined by the following equation:

$\log\left(\frac{p_{iaj}}{1-p_{iaj}}\right) = \mu + \gamma_a + \beta B + \gamma_j + \gamma_{aj} + \pi_i$ where

μ	=	grand mean
γ_a	=	effect of level a of factor A
β	=	effect of continuous covariate B
γ_j	=	effect of the jth FAR level
γ_{aj}	=	interaction effect between A and FAR
π_i	=	subject-specific random effect

The last term, π_i, is a random variable having a Normal(0,σ^2) distribution. This term is associated with the word *mixed* in GLMM because it means that the linear predictor contains both fixed and random effects. The random effect parameterizes the extra-Bernoulli variation in verification outcomes associated with unexplained difficulty or ease of recognizing various people. It also allows outcomes within subject to be correlated while outcomes between subjects remain independent.

In practical terms, the presence of a random effect to account for differences in recognition difficulty between people is very important. It is well understood that some people are harder to recognize than others [1], and our model takes this into account with the random subject effect. It is called a random effect because we do not care precisely who is difficult and who is easy; all that we care about is that some people are harder than others to recognize. Accounting for this variation reduces the unexplained variation that would otherwise weaken our ability to detect how other covariates influence performance.

While we are the first group to our knowledge to have introduced GLMMs to the task of evaluating biometric algorithms, these models are well-known and increasingly used by statisticians. Their use has grown over roughly the last 20 years as reliable and efficient computational strategies have been developed for fitting them.

In our context, one of the useful attributes of the GLMM is that it directly relates covariates to the expected probability of successful verification, or in essence to the expected verification rate. This aspect is highlighted on the right hand side of Figure 2. The direct mapping between the output of our statistical model and one of the most commonly used performance measures for face recognition makes the task of interpreting results simpler compared to, for example, analysis based on similarity scores.

4. Findings

Our major findings are summarized here. Limited space has led us to omit many details, and as we prepare this work for archival publication, details including a summary of the statistical model selection process will be added. This topic is especially noteworthy because it requires a careful mixture of quantitative analysis and expert judgment.

Because of the massive sample size of our dataset, $110,514$ match pair observations, almost all effects pass common tests of statistical significance. Thus, the common notion of statistical significance (i.e., that an observed effect is too large relative to estimated precision to be attributable to chance alone) is not particularly relevant here because the precision is extremely fine due to sample size.

We turn, therefore, to a notion of operational or scientific significance. Specifically, an effect is considered scientifically significant if it is statistically significant and it leads to a change in verification performance equivalent to at least 2 out of 100 people. The five findings that follow pass this test and are particularly notable.

4.1. Findings 1 to 4

Finding 1: FAR for Indoor and Outdoor Images Figure 4 shows the estimated probability of successful verification as a function of the FAR, separately for indoor and outdoor query image locations. The fact that the probability of verification increases with increased FAR is a mathematical necessity. Also, for all FAR settings, verification

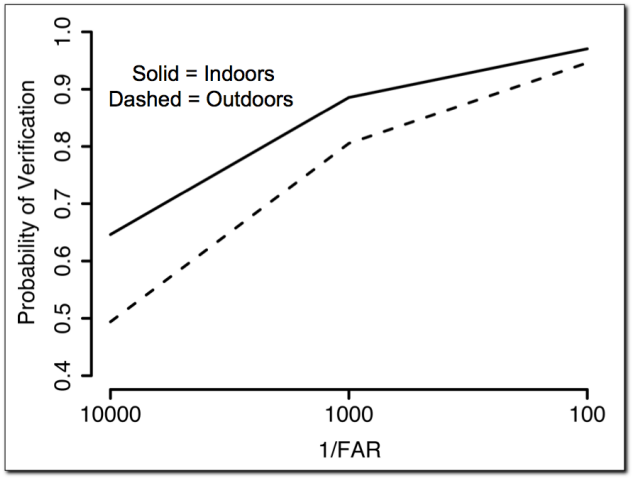

Figure 4. Estimated probability of successful verification for outdoor and indoor query images at 3 false accept rates.

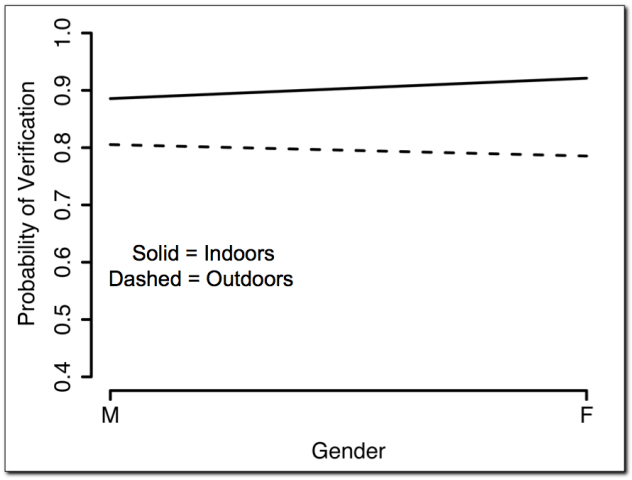

Figure 5. Estimated probability of successful verification for indoor and outdoor query images Male (M) and Female (F) subjects.

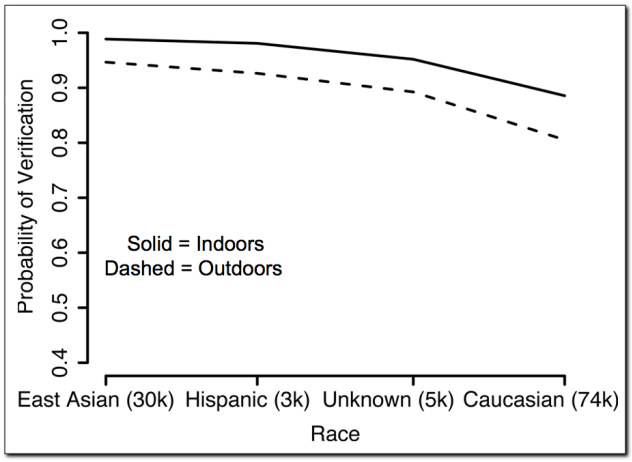

Figure 6. Estimated probability of successful verification for indoor and outdoor query images for subjects of various races.

Figure 7. Estimated probability of successful verification for indoor and outdoor query images for cases when the subject did or did not wear glasses in the query image. Note subjects never wore glasses in the target images.

is easier when the query image is indoors and there is an interaction between FAR and query image location. Specifically, the penalty for outdoor query images is reduced as FAR increases.

It is also important to recognize that the results shown represent effects *after* controlling for the impact of all other covariates in the model. In other words, the model has accounted for other factors and the probability of successful verification shown is an output of the GLMM with covariates not explicitly indicated set to default/nominal values. This is true here and for all the findings which follow.

Finding 2: Gender Figure 5 shows the estimated probability of successful verification as a function of gender for each query image location. The effect of gender on performance is scientifically significant. However, because gender interacts[1] with query image location, there is no significant marginal (i.e., averaged across locations) gender effect. Instead, we see that when men and women are photographed indoors, women are somewhat more likely to be correctly verified. Conversely, when men and women are photographed outdoors, men are slightly more likely to be verified correctly. Furthermore, the penalty for outdoor query images is greater for women than for men.

[1]Line plots are commonly used in this type of analysis to accentuate relationships. Specifically, the observation that the two lines are not parallel is a visual cue reinforcing the conclusion that there is an interaction between gender and environment (indoor/outdoor).

Finding 3: Race Figure 6 shows the estimated probability of successful verification as a function of race for each query image location. Most of the 345 subjects used in this analysis are either East Asian or Caucasian. The actual number of match pairs, i.e. verification outcomes for each of the four races, are indicated along the horizontal axis of the plot. Overall, we would not wish to overly emphasize the result for the Hispanic or Unknown categories due to comparatively low numbers of subjects. However, the distinction between the verification performance for East Asians versus Caucasians is convincing and consistent with previous findings by other studies [2, 3]. For reasons that are still not fully understood, verification performance for East Asian subjects is better.

Finding 4: Glasses When the people in the study were photographed under uncontrolled lighting they were also permitted to wear their glasses. For the controlled imagery people were never permitted to wear glasses. Consequently, some of the comparisons in the study involved people wearing glasses in the query image but not in the target. It is not surprising that glasses make verification much harder. This result is shown in Figure 7. However, it is more surprising that there is a significant interaction between wearing glasses and the query image location. Specifically, for query images without glasses the estimated performance penalty for outdoors query imagery is seen but for query images with glasses the outdoor query location actually improves performance, albeit from a much lower baseline.

4.2. Finding 5: The Face Size, Focus and Environment Interaction

This is by far the most interesting finding. It provides an excellent example of the detection and interpretation of multi-factor interactions. It also demonstrates a very strong linkage between easily measured aspects of image quality and probability of successful verification. The overall result is summarized in Figure 8. A careful exploration of this figure is given in the following paragraphs.

First note the three columns of plots in Figure 8 corresponding to the resolution of the query images as measured by the distance between the eyes measured in pixels. The median distance between eyes for the columns labeled Small, Medium and Large are 137, 164 and 210 pixels respectively. The break points between the columns fall at 150 and 185 pixels between the eyes. Query image resolution is the first of the four interacting covariates.

Next focus on the upper and lower rows of plots. The upper row is for query images acquired indoors and the lower row is for query images acquired outdoors. The second of the four covariates participating in this interaction is the distinction between indoor versus outdoor imagery.

The six plots shown share the same x- and y-axes. The horizontal axis shows FRIFM for the query image and the vertical axis shows FRIFM for the target image. Note the overall broader range of FRIFM values for the query imagery compared to the target imagery. This makes sense considering the control exerted over the acquisition of the target images compared to the query images.

The estimated probabilities of successful verification shown in the six plots are color-coded using a standard cold to warm pseudo-color mapping. Each of the six plots has been further refined to indicate approximately which regions the response surface correspond to the available data. To put this another way, interior to the regions bounded by the black outlines are portions of the surface where about 95% of all our observations lie. In order to avoid accidental extrapolation, it is important to restrict our attention to the interior of these regions.

Several conclusions are striking when one studies Figure 8. First, there is a very large variation in predicted performance. For the indoor images, the probability of verification values range from around 0.7 up to greater than 0.95. For the outdoor images the range in probability of verification values over an astonishingly large range, from a low 0.1 to a high of nearly 0.9.

Second, note that query FRIFM scores do not range as high for the indoor query images than for the outdoor query images. This may be suggestive of some relationship between query image location and FRIFM.

Third, it is surprising that lower FRIFM values are associated with higher estimated probability of correct verification. This suggests that the three algorithms prefer images somewhat out of focus. Were we studying older whole image matching algorithms, this finding would not seem so surprising. It has been fairly well established that techniques such as PCA do marginally better when images are smoothed [7]. However, here we are looking at state-of-the-art commercial algorithms, and it is less obvious that they should share this preference for reduced focus.

It is also important to note that while FRIFM is a good surrogate for a true measure of focus, it may also reflect other image attributes such as harsh lighting, hairs across the forehead, etc. Because of the importance of our findings with respect to FRIFM, we have visually inspected about 50 images with very high and very low FRIFM scores. Six of these images are shown in Figure 9. Overall, it is our judgment that in the majority of cases FRIFM is responding to what we as human judges would call focus. However, it is also clear that other factors are at work as well. Notice, for example, the glasses and overall strong shadowing of the face in the woman shown in the upper right of Figure 9. Also notice the hair coming down across the face combined with strong lighting in the woman shown in the lower right of Figure 9. Low FRIFM scores also seem to be produced

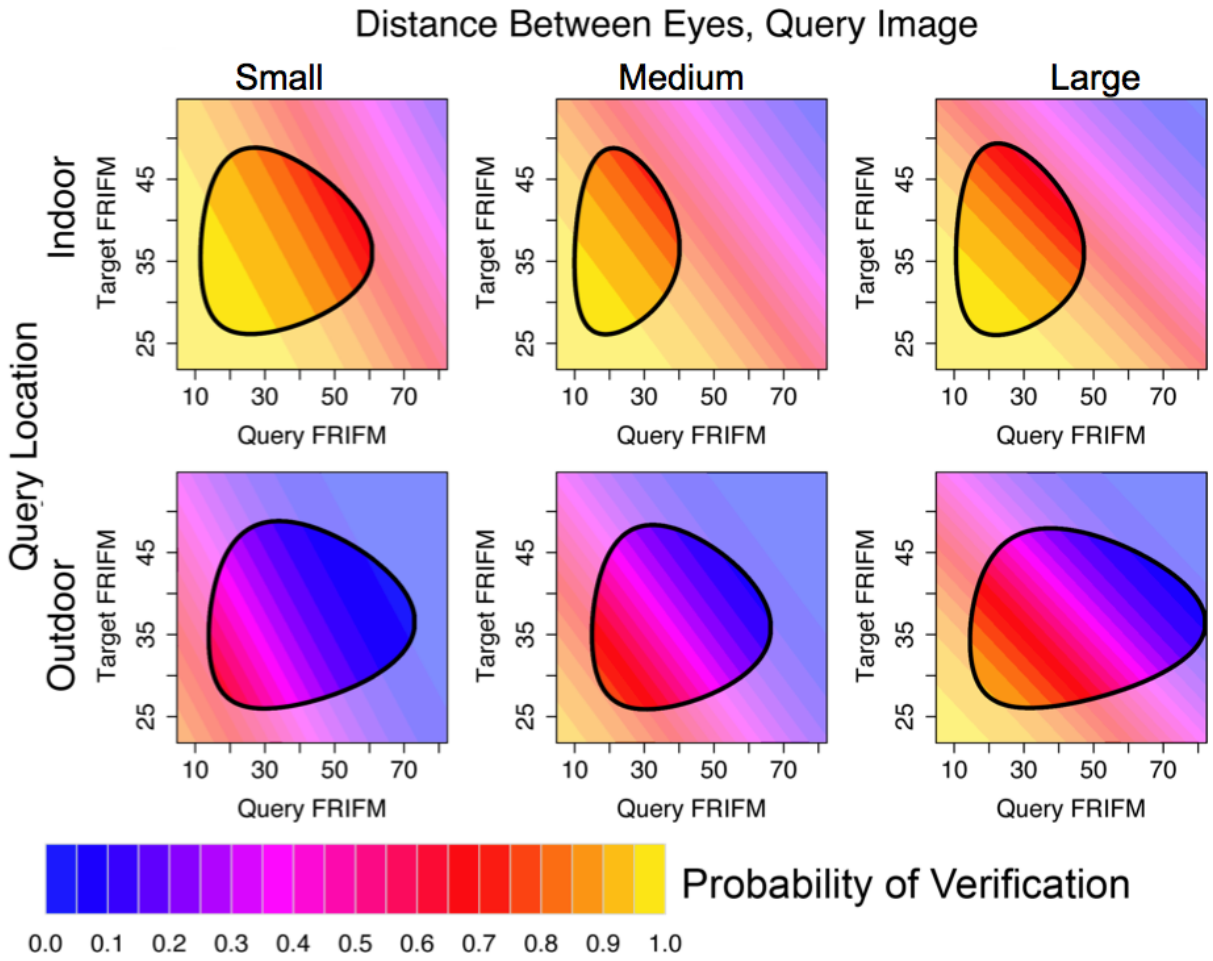

Figure 8. Estimated probability of successful verification split by query image resolution, query image location, and FRIFM for both query and target images. See the text for a full explanation of this figure.

by people whose overall complexion and facial appearance is relatively uniform, for example the woman shown in the middle left image of Figure 9.

Taking a broader perspective, it is clear that a single simple quality measure computed post hoc from images is highly correlated with probability of successful verification by state-of-the-art vendor algorithms. From a practical standpoint, such findings may be very valuable for enabling progress toward a better overall measure of face image quality. In terms of algorithm development and improvement, it is important to better understand *why* images with high edge density (i.e., high FRIFM scores) confound algorithms.

5. Conclusions

A tremendous amount of effort will be spent in the near future trying to better characterize the quality of face images in terms of successful biometric matching. For example, in the past two years NIST has sponsored two workshops on biometric quality. Our work has demonstrated that statistical modeling provides an excellent means of explicitly establishing connections between easily measured properties of imagery and predicted probability of verification success.

Furthermore, the generalized linear mixed model allows us to explicitly to control for confounding covariates, allow for subject-specific performance variation, and capture interactions between covariates. The importance of analysis at this level is illustrated by the complex four-way interaction presented above. Because the GLMM controls for confounding covariates, the four-way interaction effects performance regardless of the other covariates. In the case of this analysis, regardless of gender, race, and wearing of glasses.

Finally, this study has helped identify where some of the most important short term gains in performance may be achieved. In particular, it seems that improvements or compensation for the FRIFM factor should improve perfor-

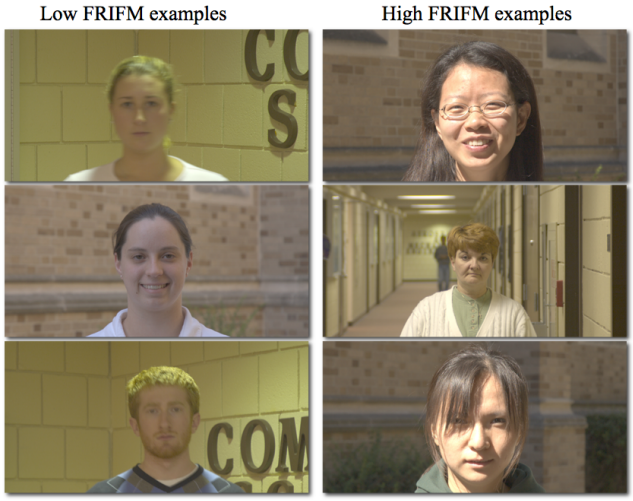

Figure 9. Examples of three images with very low and three images with very high FRIFM scores.

mance. Also, high resolution outdoor query images lead to much better performance than do low-resolution outdoor query images and this finding has practical importance because image resolution is a factor that may be easily increased in many common circumstances. With further analyses of richer datasets that lie ahead, we anticipate identifying additional strategies for improving algorithm performance on the basis of a growing understanding of the impact of subject covariates and image quality on performance.

6. Acknowledgements

This work was funded in part by the Technical Support Working Group (TSWG) under Task T-1840C. P. Jonathon Phillips was supported by the National Institute of Justice.

References

[1] G. Doddington, W. Liggett, A. Martin, M. Przbocki, and D. Reynolds. Sheep, Goats, Lambs and Wolves - A Statistical Analysis of Speaker Performance in the NIST 1998 Speak Recognition Evaluation. In *5th Unternational Conference on Spoken Language Processing (ICSLP)*, page paper 608, Syndey, Australia, November 1998.

[2] Geof H. Givens, J. Ross Beveridge, Bruce A. Draper, and P. Jonathon Phillips. Repeated Measures GLMM Estimation of Subject-Related and False Positive Threshold Effects on Human Face Verification Performance. In *Empirical Evaluation Methods in Computer Vision Workhsop: In Conjunction with CVPR 2005*, pages electronic–only, June 2005.

[3] Geof H. Givens, J. Ross Beveridge, Bruce A. Draper, Patrick Grother, and P. Jonathon. Phillips. How Features of the Human Face Affect Recognition: a Statistical Comparison of Three Face Recognition Algorithms. In *Proceedings: IEEE Computer Vision and Pattern Recognition 2004*, pages 381–388, 2004.

[4] P. Grother and E. Tabassi. Performance of biometric quality measures. *IEEE Trans. Pattern Analysis Machine Intelligence*, 29:531–543, 2007.

[5] E. P. Krotkov. *Active Computer Vision by Cooperative Focus and Stereo*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1989.

[6] R. M. McCabe. Best practice recommendation for the capture of mugshots version 2.0. http://www.nist.gov/itl/div894/894.03/face/face.html, 1997.

[7] H. Moon and P. J. Phillips. Computational and performance aspects of PCA-based face-recognition algorithms. *Perception*, 30:303–321, 2001.

[8] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek. Overview of the face recognition grand challenge. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 947–954, 2005.

[9] P. J. Phillips, P. J. Grother, R. J. Micheals, D. M. Blackburn, E. Tabassi, and J. M. Bone. Face recognition vendor test 2002: Evaluation report. Technical Report NISTIR 6965, National Institute of Standards and Technology, 2003. http://www.frvt.org.

[10] P. J. Phillips, H. Moon, S. Rizvi, and P. Rauss. The FERET evaluation methodology for face-recognition algorithms. *IEEE Trans. PAMI*, 22:1090–1104, October 2000.

[11] P. J. Phillips, W. T. Scruggs, A. J. O'Toole, P. J. Flynn, K. W. Bowyer, C. L. Schott, and M. Sharpe. FRVT 2006 and ICE 2006 large-scale results. Technical Report NISTIR 7408, National Institute of Standards and Technology, 2007.