FRVT 2006: Quo Vidas Face Quality[☆]

J. Ross Beveridge[*,a], Geof H. Givens[b], P. Jonathon Phillips[c], Bruce A. Draper[a], David S. Bolme[a], Yui Man Lui[a]

[a]*Department of Computer Science, Colorado State University*
[b]*Department of Statistics, Colorado State University*
[c]*National Institute of Standards and Technology*

Abstract

This paper summarizes a study of how three state-of-the-art algorithms from the Face Recognition Vendor Test 2006 (FRVT 2006) are effected by factors related to face images and the people being recognized. The recognition scenario compares highly controlled images to images taken of people as they stand before a camera in settings such as hallways and outdoors in front of buildings. A Generalized Linear Mixed Model (GLMM) is used to estimate the probability an algorithm successfully verifies a person conditioned upon the factors included in the study. The factors associated with people are: gender, race, age and whether they wear glasses. The factors associated with images are: the size of the face, edge density and region density. The setting, indoors versus outdoors, is also a factor. Edge density can change the estimated probability of verification dramatically, for example from about 0.15 to 0.85. However, this effect is not consistent across algorithm or setting. This finding shows that simple measurable factors are capable of characterizing face quality; however, these factors typically interact with both algorithm and setting.

Key words: face recognition, generalized linear mixed models, image covariates, biometric quality

[☆]The work was funded in part by the Technical Support Working Group (TSWG) under Task T-1840C. PJP was supported by the Department of Homeland Security, Director of National Intelligence, Federal Bureau of Investigation and National Institute of Justice. The identification of any commercial product or trade name does not imply endorsement or recommendation by Colorado State University or the National Institute of Standards and Technology.

[*]Principal corresponding author.

1. Introduction

Frontal face recognition algorithms have matured considerably in recent years. The results from the Face Recognition Vender Test 2006 (FRVT 2006) showed that it is possible to achieve verification rates of 0.01 at a false accept rate of one in a thousand for frontal face images taken with controlled lighting and plain backgrounds [1]. Unfortunately, this result is not robust to changes in imaging conditions. Moving the imaging location to a hallway or outside, dramatically reduces verification rates.

Intuitively, one suspects that face recognition algorithms succeed on high-quality images, but perform less well on images of lower quality. The challenge is to identify factors that characterize "high-quality" images. This leads to efforts to quantify biometric quality in general [2] and face image quality in particular [3, 4, 5, 6, 7]. In biometric quality research, one goal is to find measurable properties of an image that are predictive of match performance. A second goal is to find universal quality measures. A universal quality measure is a measurable property of an image that is predictive of performance for a wide class of matchers.

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 covariates on three of the better performing algorithms in FRVT 2006. The statistical analysis technique for measuring the effect of these covariates is generalized linear mixed modeling (GLMM).

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.

Person specific covariates are attributes of the person being recognized, such as age, gender, or race. Person specific covariates can be transitive properties of people, 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. 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 and region density measures, to be introduced shortly, are measurable covariates. 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.

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.

The primary findings, described in Section 4, are for potential quality measures or actionable covariates. These covariates are edge density, region density, face size, and setting. Edge density may be thought of as a proxy for focus, although it also responds to other important aspects of face images. Edge density can exert a dramatic influence over face recognition performance. Region density does not have as intuitive an interpretation, but it is modestly predictive of performance. Setting is where an image is taken – either outdoors or indoors.

Being able to analyze the performance of three algorithms in two settings– outdoors and in hallways, allows us to examine the universality of quality measures in terms of both algorithms and settings. Generally, the potential quality measures exhibit strong interactions with both algorithm and setting. The one exception to this is region density measured over query images, which when our full statistical model is used has no significant interaction with either setting or algorithm. At a minimum, what these results underscore is the importance of studying potential quality measures carefully and techniques that account for multiple factors. In general, we expect that choices of face quality measures will have to be qualified with respect to scenario and algorithm.

The need to qualify quality measures does not, however, diminish the need for rigorous studies to reveal the factors and combinations of factors that influence performance. The study presented in this paper quantifies the relationship between factors both outside user control and under user

control that influence face recognition performance, and understanding these relationships is important both to algorithm developers and anyone involved with deploying facial biometrics.

The remainder of this paper is organized as follows. Section 2 provides a brief review of our prior work using GLMM models to study how multiple factors influence face recognition performance as well as a brief review of the FRVT 2006. Section 3 provides a detailed account of the specific GLMM that is the basis for the study presented. In particular, Section 3.1 details the specific covariates used in the model and Section 3.2 explains the model itself, including the process of selecting significant covariates and covariate interactions. Section 4 summarizes four of the most significant findings of our study. These involve sensitivity of the three algorithms to where images are acquired and the size of the face in images. It also involves the response of algorithms to region and edge density in images. These last two factors are easily measured properties of face images. Definitions for region and edge density appear in Appendices B and C. Findings for age, gender and race are included in Appendix A.

The details presented in Section 3 are critical for anyone wanting to understand precisely how our analysis has been carried out, but it is not necessary to read Section 3 to understand the practical implications of the results presented in Section 4.

2. Background

The study presented here is an expansion of a previously published study [8] that considered a single face recognition algorithm created by fusing similarity scores from the three individual algorithms studied here. This difference means, principally, that in the results presented below we are able to make observations about how covariates influence individual algorithms. As a consequence, most of the results presented here are new. However, where we think the comparison is helpful we relate the findings for individual algorithms back to the findings for the single fused algorithm.

More generally, the study we are presenting here is the most recent in a series of examinations carried out using linear and generalized linear models to relate covariates to the performance of face recognition algorithms [9, 10]. This study also represents a review of algorithm performance on the FRVT 2006. Some of the prior papers to include performance results from the FRVT

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

2006 and Face Recognition Grand Challenge which proceeded it are [1, 11, 12].

The remainder of this background section will briefly summarize FRVT 2006 and specifically the uncontrolled controlled lighting experiment that is the basis for our study. It will also provide a brief overview of the statistical model developed in Section 3.

2.1. FRVT 2006

The FRVT 2006 was an independent evaluation of face recognition algorithms administered by the National Institute of Standards and Technology (NIST) [1]. The FRVT 2006 was the latest in a series of U.S. Government sponsored challenge problems and evaluations designed to advance automatic face recognition [13] [14] [15].

This paper analyzes performance on the FRVT 2006 *very high-resolution* image set. The very high-resolution images were acquired with a 6 Megapixel 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 [16]. 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.

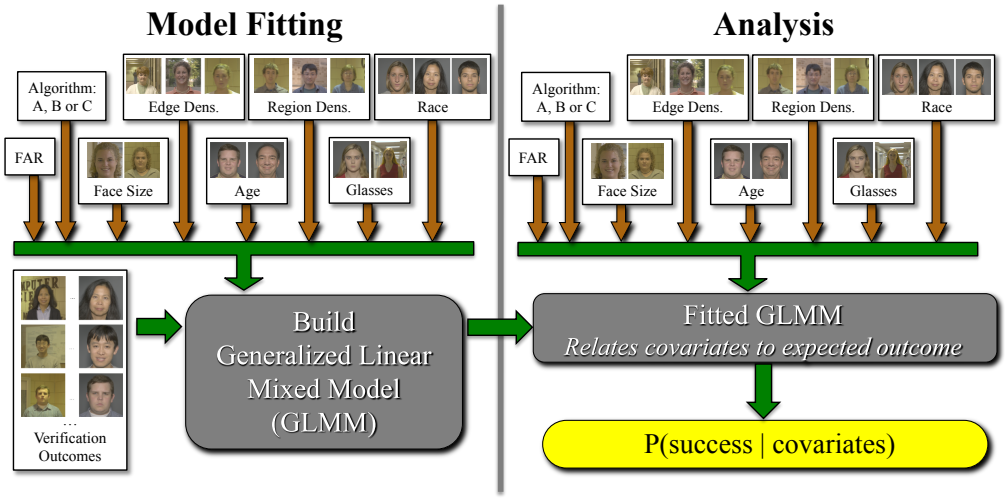

Figure 2: A GLMM is first fit to our data and then used to analyze the manner in which covariates influence the probability an algorithm successfully verifies a person.

Three of the FRVT 2006 top performing algorithms were selected for study here. They are the same three algorithms that were used to create the fusion algorithm previously reported on in [8]. In this paper the three algorithms will be designated as A, B and C.

This study, and its predecessor [8], are the first to include results for uncontrolled illumination images taken outdoor. The FRVT 2006 large-scale experiment report [1] presented results matching controlled illumination images to controlled illumination images, and controlled illumination images to indoor uncontrolled illumination images. We will see that setting, in terms of indoors versus outdoors, interacts with many of the covariates in our study and that algorithms behave differently in the two settings.

2.1.1. Statistical Model Overview

Figure 2 provides a pictorial overview of the GLMM used in our study. The figure accentuates two aspects of this analysis. The first is the process of developing the model, which involves a variety of steps including selecting factors to include and fitting the model to the observations. The second step, labeled analysis, involves using the model to summarize how factors and combinations of factors effect an estimate of the probability that an algorithm will successfully verify a person given a specific target and query

Table 1: List of covariates eligible for inclusion in the GLMM. For each, the units or the observed values are indicated. For the query/target pairs of eye distance, edge density and region density variables, polynomial terms up to cubic order and similar cross-products were also eligible. In other words, for any query/target pair of these variables, the following terms were considered: Q^3, Q^2, Q, T^3, T^2, T, Q^2T, QT^2, and QT. The baseline values of each covariate indicate the value for which the effect is included in the 'intercept' term μ in the GLMM linear predictor; see the text. The asterisked terms were standardized so that the baseline value was zero.

Covariate	**Values/Units**	**Baseline**
Algorithm	A, B, C	C
FAR	1/100, 1/1000, 1/10,000	1/1000
Gender	Female (F), Male (M)	Male
Race	Caucasian, East Asian, Hispanic, Other	Caucasian
Age	years	mean
Query Setting	Indoors, Outdoors	Indoors
Target Eye Distance*	pixels	0
Query Eye Distance*	pixels	0
Target Edge Density*	gradient mag. (x8)	0
Query Edge Density*	gradient mag. (x8)	0
Target Region Density*	region count	0
Query Region Density*	region count	0
Person Wears Glasses	No, Yes	Yes
Person Id	Random Effect	N/A

image. In the course of actually developing a study and an associated model this process is not sequential as implied by the figure, but instead involves considerable iteration.

That said, the distinction illustrated by the two halves of Figure 2 provide a useful simplification, and is carried forward into the following two sections, where Section 3 may be thought of as describing the left side of Figure 2, i.e. the process of creating the GLMM, and Section 4 may then be thought of as the process of using the model to shed light on the role covariates play in changing the estimated probability of verification. Also, reiterating what we said in the introduction, if your interest lies primarily in results of our analysis, please skip to Section 4.

3. The Statistical Model

Our study uses a Generalized Linear Mixed Model (GLMM) [17, 18] to relate a collection of covariates to the verification performance of algorithms.

For each algorithm, the available data consisted of $110,514$ records of attempts to verify a pair of query and target images. Each attempt was made on a matching pair of images, i.e., both the query and target images were of the same person. There were 345 distinct people in the study. Combining the results for the three algorithms, the dataset consisted of $331,542$ records. Each record included the person's identity, the verification outcome, and related covariates. The verification outcome is determined by a threshold on the algorithm's similarity score associated with one of three false accept rate (FAR) choices: 1/100, 1/1000 and $1/10,000$.

Below we describe the covariates, the model, and the method for selecting which covariates should be used as predictors in the model.

3.1. Covariates

Table 1 lists the collection of covariates considered for use in the GLMM. Not all of these are used in the final model; our model selection strategy is discussed in Section 3.3. On the other hand, covariates may enter the model in more than one form. In addition to standard terms, covariates may be involved in interactions with other variables. Furthermore, for the query and target covariates related to eye distance, edge density, and region density, a large collection of polynomial terms were also eligible for inclusion. Heuristically, for a query/target pair of any such variable, terms like Q^3, Q^2, Q, T^3, T^2, T, Q^2T, QT^2, and QT were considered, where Q represents a covariate measured on the query image, and T represents a covariate of the target image. Finally, interactions of polynomial terms with other covariates were eligible to the extent allowed by our model selection process described below.

The FAR covariate, which indicates the false accept rate was transformed to $-\log$ FAR. The choice of the face recognition algorithm used for a verification attempt was treated as a factor with three levels, corresponding to algorithms A, B and C. These algorithms have been discussed above. The demographic information for the Gender, Race and Age covariates is available as part of the FRVT 2006 data. The last covariate related to the person, as opposed to the image, indicates whether a person was wearing glasses in the query image.

The next covariate, Query Setting, indicates whether the query image was acquired indoors or outdoors. Recall that the data used in this study represent comparisons between target images acquired under highly controlled

lighting conditions and query images acquired under less controlled conditions. Consequently, every pair of query and target images for which verification is tested is tagged as either indoors or outdoors depending upon the setting in which the query image was acquired. Eye Distance indicates the number of pixels on the face in the target and query images.

Edge Density is a measure of the edge density in the region of the face. This measure has been suggested as a good surrogate for whether an image is in focus [19] and is computed as the average Sobel edge magnitude within an oval defining the region of the face. In our past work [8] we had labeled this covariate as 'FRIFM'. Here, the simpler term 'Edge Density' is used. Details on how Edge Density is computed are included in Appendix B.

Region Density is a count of the number of regions found using a standard open-source region segmentation routine developed by Comaniciu and Meer at Rutgers [20]. The routine was run using its medium sensitivity setting and only regions intersecting the face oval were counted. Details on how Region Density is computed are included in Appendix C.

It is important to recall that the list of eligible covariates is relatively small compared to the variety of eligible terms considered over the course of our model selection process. Also, some eligible covariates were not included at all—in any form—in the final model. Section 3.3 describes the terms used in the final model.

3.2. Our Generalized Linear Mixed Model

One of the useful attributes of our GLMM is that it directly relates covariates to the expected probability of successful verification, or in essence to the expected verification rate. The fact that our model produces estimates of one of the most commonly used performance measures for face recognition, namely verification rate, makes the task of interpreting results much simpler than, for example, analysis based on similarity scores [9].

The word *generalized* in GLMM refers to the sensible assertion that verification outcomes are Bernoulli distributed, and to a resultant approach that fits a nonlinear dependency between predictors and expected outcomes. In contrast, an ordinary linear model assumes Gaussian outcomes and a linear relationship between covariates and expected outcomes.

There are many types of generalized models with a variety of distributional assumptions. Each uses a 'link function' to introduce nonlinearity that is appropriate for the particular distributional assumption. When modeling Bernoulli distributed outcomes, the standard link function is the logit

function (see below). Although our Bernoulli/logit model relates expected verification outcomes to a nonlinear function of the covariates, it provides a linear relationship between log odds ($\log\{p/(1-p)\}$) and the covariates.

In our model, the verification outcomes are expressed as Bernoulli random variables Y_{iaj} with success probabilities p_{iaj}. The subscripts i, a and j index specific covariates. While our actual model is more complex, this example is sufficient to illustrate key aspects of the GLMM. For 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

$$\begin{array}{lcl} \mu & = & \text{grand mean} \\ \gamma_a & = & \text{effect of level } a \text{ of factor } A \\ \beta & = & \text{effect of continuous covariate } B \\ \gamma_j & = & \text{effect of the } j\text{th FAR level} \\ \gamma_{aj} & = & \text{interaction effect between } A \text{ and FAR} \\ \pi_i & = & \text{person-specific random effect} \end{array}$$

The right hand side of this equation is called the linear predictor. In it, the last term π_i is particularly important. It 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 related to one person to be correlated while outcomes between people 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 [21], and our model takes this into account with the randomized person 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.

The other Greek letters in the linear predictor are non-random parameters which are interpreted analogously to ordinary linear regression, except for the impact of the link function. For example, μ represents the log odds of

verification (aside from person-specific impacts) when A and FAR are set at baseline levels and $B = 0$. A parameter like γ_a indicates a discrete effect on the linear predictor when A moves from its baseline level to the level a.

The effects of variables treated as continuous, such as pixels between the eyes, are also straightforward. The coefficient β indicates the change in the linear predictor associated with a one-unit change in B, which in turn can be directly related to the estimated probability of verification p_{iaj}. In our model, continuous variables are treated in standard units [1], so a one-unit change in B corresponds to a shift of one standard deviation.

For models of verification performance for face recognition algorithms, it is important to include a specific covariate that enables sampling of outcomes at distinct false accept rates (FAR). So, for example, γ_j may parameterize the effect of setting FAR at $\frac{1}{100}$, $\frac{1}{1000}$ or $\frac{1}{10{,}000}$. Considering that we examined only these three FAR values, it would be reasonable to treat the FAR variable as a factor with discrete levels. This matches the γ_j parameterization in the example equation above. However, in much of our past work [9, 10] we have found that the odds of verification are extremely close to log-linear in FAR over the FAR range considered here. Therefore it is also reasonable to treat $-\log$ FAR as a continuous variable with corresponding parameter, say, ϕ. In the model, this would replace γ_j with a term $-\phi$FAR. We have chosen this latter option.

One of the most important aspects of models like ours is the ability to explicitly measure interactions between covariates, as illustrated by the term γ_{aj}. The specific interactions to include in the final model and those to disregard are identified during the overall model selection process summarized in Section 3.3.

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

3.3. Model Selection

Model selection involves searching for a reduced set of covariates and interactions that provide a highly effective but parsimonious model for predict-

[1] The exception is Age which was left encoded as age in years.

ing verification performance. This is fundamentally an optimization problem with the competing objectives of prediction performance and parsimony. In the project described here, there are two important, serious challenges for model selection. Both relate to the large size of the dataset, namely 331,542 observations.

3.3.1. Search - Balancing Effectiveness and Parsimony

The first challenge is that fitting these GLMMs is sufficiently computationally intensive that an exhaustive search over the 2^k-sized space of models is infeasible. Note that k is the number of possible model terms, including polynomial terms and multi-way interaction terms. k is therefore much larger than the number of covariates.

The GLMM developed for the fusion algorithm [8] provided us with a starting point for model selection in this study. To select that model we used a manual, semi-greedy search strategy guided by pre-established principles and by expert judgment. Philosophically, the strategy was most akin to 'backwards elimination' in that the search generally progressed from larger models to simpler ones. Intermittent phases of model expansion ('forward selection') were also used to discourage entrapment in local optima.

In the current model, we sought to account also for algorithm-specific dependencies. Consequently, we added an Algorithm factor to the model and also tested for all possible interaction effects involving Algorithm. An iterative strategy was again employed to select algorithm-specific terms that should be added to the baseline model. In this case, the general philosophy was 'forward selection', i.e., identifying important model additions starting with the simplest necessary Algorithm effects and progressing toward more complex terms.

Adding Algorithm and corresponding interaction terms to the GLMM often better explained verification performance than did the original fusion model. However, variables often have complex relationships and contain partially shared information. As a result, some of the terms that originated from the fusion model were no longer necessary and were removed from our final model. For example, the fusion model included a four way interaction between between Query Setting, Query Eye Distance, Target Edge Density and Query Edge Density. While the new model still includes all of these covariates in lower order interactions, the four way interaction was no longer significant and was therefore removed. Such removals were interspersed throughout the overall selection process.

Notwithstanding such cases, we observed few signs of confounding between predictors, in that the estimated effects of model terms remained fairly unchanged when other terms were added or excluded from the model. An important consequence of this is that the final model choice should be relatively insensitive to the particular strategy of model selection. Furthermore, our approach was sufficiently iterative in both directions ('forwards' and 'backwards') to implicitly cover a large portion of the overall model space. Finally, the nature of GLMMs (and other regression models) and our expansive selection strategy strongly suggest that predictions of verification probabilities will be quite reliable even if parameters for certain covariates are less certain as a consequence of partial covariate confounding.

3.3.2. Significance

The second model selection challenge resulting from the huge dataset size is that standard measures of statistical significance are of little help when sample sizes are so large. Almost any possible covariate or interaction we might add will have an associated p-value that surpasses ordinary standards for statistical significance. Consequently, a different and more practical approach is necessary. We switched from the formal notion of statistical significance to the more useful measure of operational significance.

A covariate or interaction effect is deemed operationally significant if it predicts a change in verification performance equivalent to at least 2 out of 100 people. This degree of change must be attributable to a change in a single factor level, for example to a one-standard deviation change in eye distance, edge density, or region density, or to a corresponding change in an interaction or polynomial term. During the manual model selection process terms were added and removed from the model based on whether they met the 2 in 100 change threshold established by this definition of operational significance.

3.3.3. Selected Terms

We have listed the eligible covariates in Table 1, and have described our method for selecting model terms derived from this list. The final model included the following terms. There were main effects for the covariates: Algorithm, Gender, Race, Age, Glasses, Query Setting, $-\log$ FAR, Query Eye Distance (and its square), Target and Query Edge Density, and Query Region Density. There were also a variety of interactions. Algorithm interacted with Gender, Race, Age, Glasses, Query Setting, $-\log$ FAR, and

Query Edge Density. This means that in each of these cases the effect of the corresponding covariate was different for different algorithms.

There were also many interactions with the Query Setting, namely: Gender, Race, Age, Glasses, Query Eye Distance (and its square), and Target and Query Edge Density. Additionally, Algorithm was involved in a variety of three-way interactions, all of which also involved Query Setting. The three-way interactions for Algorithm involved: Gender $\times$ Setting, Race $\times$ Setting, Age $\times$ Setting, Glasses $\times$ Setting, Query Eye Distance $\times$ Setting, and Target and Query Edge Density $\times$ Setting. These interactions indicate, for example, that although the effect of Gender on verification probability differed for the two Settings, the nature of this difference itself differed between algorithms.

Finally, it is worth noting that Query Region Density was operationally significant in our final model but did not interact with any other term, most notably Algorithm. This finding is discussed later in more depth.

4. Findings

Our major findings, particularly as they relate to properties of face images that might be associated with image quality, are summarized here. Specifically, findings for Region Density, Query Face Size and Edge Density are presented. We begin with the results for false accept rate (FAR) and indoor versus outdoor query images.

4.1. Query Setting and FAR

Figure 3 shows how FAR and Setting effect each of the three algorithms. The vertical axis indicates the probability of verification indicated by our GLMM. The horizontal axis indicates each of three FAR levels: $1/10,000$, $1/1000$ and $1/100$. There are a total of six line plots shown, one for each combination of Algorithm and Setting. Colors red, green and blue are used to indicate algorithms A, B and C. Solid lines indicate results on indoor query images and dashed lines indicate results on outdoor query images.

Before we say more about the specific effects presented, let us use this plot to give some background on how we are presenting information. Plots of this kind are a visually useful way to present covariate effects, since one can readily look at the slope of the line to see the relative magnitude of an effect. It is important to understand that the lines are connecting specific estimated probabilities of correct verification coming out of the GLMM. So, for example, the solid red line at the top of the plot connects three estimated

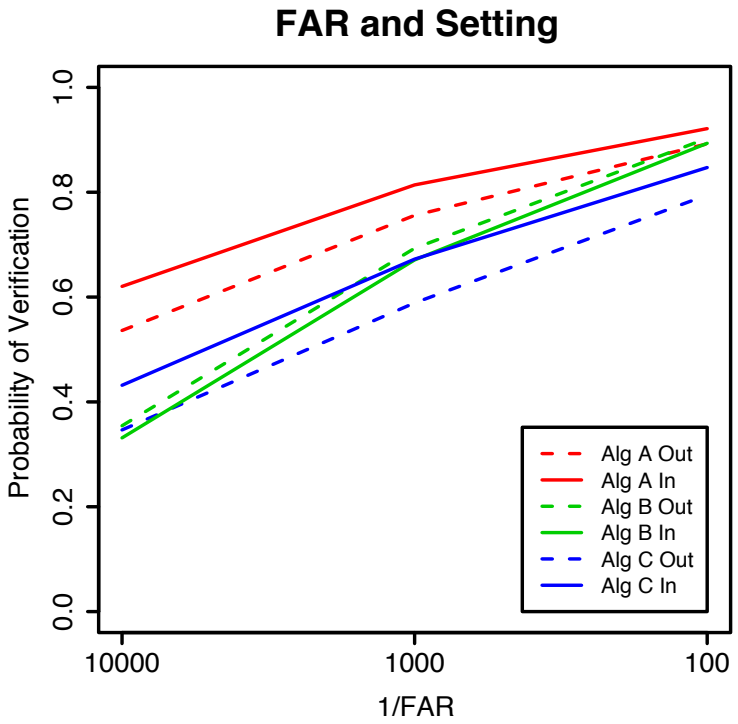

Figure 3: Response of three algorithms to FAR for the indoor and outdoor query images. In this case there is no interaction between FAR and Setting and the results are shown on one plot for convenience. What is evident is that while outdoor images are harder for Algorithms A and C, there is little difference and even a slight advantage to being outdoors for Algorithm B.

probabilities of verification generated for algorithm A indoors at FARs of $1/10,000$, $1/1000$ and $1/100$.

Whenever covariate effects are presented for one or several covariates, the results are always in the context of having controlled for the other covariates in our model. Moreover, those other covariates are given the baseline values indicated in Table 1. So, for example, the probability of verification of about 0.81 for algorithm A on indoor images and FAR = $1/1000$ is for Caucasian males of average age wearing glasses in the query image.

Returning to the question of what these results tell us about FAR, Setting and Algorithm, the first observation is that the probability of verification

increases with FAR. This is, of course, a mathematical necessity growing out of the relationship between false accepts and true accepts.

More interestingly, the plot lines for algorithms A and C exhibit similar slopes both for indoor and outdoor query images, although they are offset from each other . This suggests first that FAR and Setting are influencing these two algorithms in a similar fashion even though there is a notable difference in absolute level of performance between the two algorithms. Overall, algorithm A has a higher expected probability of verification than algorithm C.

It also appears that algorithm B is not behaving in the same fashion as algorithms A and C when it comes to the manner in which it is influenced by either FAR or Setting. Most important from an operational standpoint is the fact that algorithm B does slightly better on query images acquired outdoors versus query images acquired indoors. It has been commonly assumed that the outdoor imagery is harder, and indeed this is what we discovered in our earlier study of the fusion algorithm. Now it appears that although two out of three algorithms find outdoor images more difficult, it is not universally true that moving outdoors makes recognition harder.

Another effect of interest is that algorithm B exhibits a much higher degree of sensitivity to the choice of FAR. It's estimated probabilitiy of verification for the most difficult situation, FAR = 1/10,000 is as low as that seen for any combination of Algorithm and Setting. At the other extreme, it's estimated probability of verification for the easiest situation, FAR = 1/100 is nearly equal to the best achieved by any combination of Algorithm and Setting.

One other observation about FAR and Setting deserves mention. While for convenience of presentation we have chosen to combine the results for FAR and Setting on one plot, in the course of developing our GLMM we found no significant interaction between these two covariates. This is interesting especially in light of our previous results for the fusion algorithm, where an interaction between FAR and Setting was found. The likely explanation for the difference is as follows. Since algorithms A and C respond to indoor versus outdoor query images differently than algorithm B, and since algorithm B is more sensitive to changes in FAR, it makes sense that an algorithm combining the properties of all three would be sensitive to interactions between FAR and Setting.

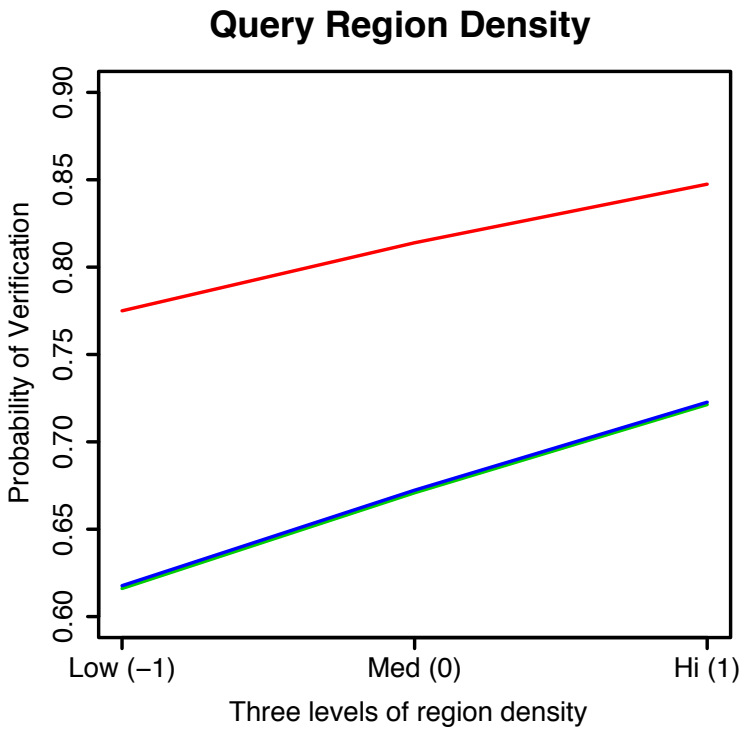

Figure 4: Response of three algorithms to Query Region Density. Three curves are shown, red, green and blue for algorithms A, B and C respectively. The algorithm B and C curves lie essentially on top of each other. There is no interaction between region density and Setting or Algorithm, hence both influence all three algorithms in the same manner.

4.2. Region Density

Figure 4 shows how Query Region Density effects the probability of verification. Region density is a count of the number of individual regions found on the face by a region segmentation algorithm. It is a measure of local homogeneity over the face. The details of how region density is computed are included in appendix C.

For all three algorithms, the probability of verification increases for face images with a higher region density. In addition, the magnitude of the increase is enough to be of practical interest. For example, probability of verification for algorithm A goes from 0.77 for images with low region density to 0.83 for images with high region density. A similar trend is observed

for the other two algorithms. Indeed, the trend is not merely similar for the other two algorithms, there is no interaction effect between region density and algorithm and hence our GLMM model tells us that algorithms are all influenced in the same fashion.

It would appear that in Query Region Density we have found the exception to the rule that simple measures of image properties will not universally predict behavior across sets of face recognition algorithms. However, while this result is important, it has a significant caveat. We took the additional step of fitting alternative GLMMs with effects for Query Region Density and for various subsets of the covariates included in the full model. This revealed that the universality of Query Region Density as a predictor of performance emerges only after controlling for a wide variety of other covariates related to characteristics of the person and the image.

Without such controls, Query Region Density no longer exhibits such universality because it partially subsumes effects for the uncontrolled variables. For example, after removing covariates related to people the effect of Query Region Density becomes dependent upon Algorithm. Since algorithms clearly treat types of people differently, this suggests that Query Region Density compensates for some of the variation in verification performance that would otherwise be attributable to sensitivities of algorithms to various types of people. Similarly, after removing the image related covariates, the effect of Query Region Density becomes dependent upon Query Setting. This has the analogous implication that Query Region Density likely captures some aspects of Setting, thereby compensating for some of the variation in verification performance that would otherwise be attributable to whether query images were taken indoors versus outdoors. This raises the intriguing possibility that other universal measures of quality might exist and be discovered if sufficient other confounding variables are taken into account.

Overall, we think the Query Region Density result is important and warrants further investigation. It seems unlikely, as the suite of algorithms we consider expands and the number of data sets we analyze grows, that the Query Region Density will remain as clean a result as it appears in this study. In other words, it is our expectation that it will exhibit interactions with algorithms and probably with other factors. That said, this is a strong result and one we hope that algorithm developers will begin to look into more closely. It's worth pointing out that there is no particular reason to think a priori that a high region density should be better than a low region density. We also examined some of the images that lie at the extremes of

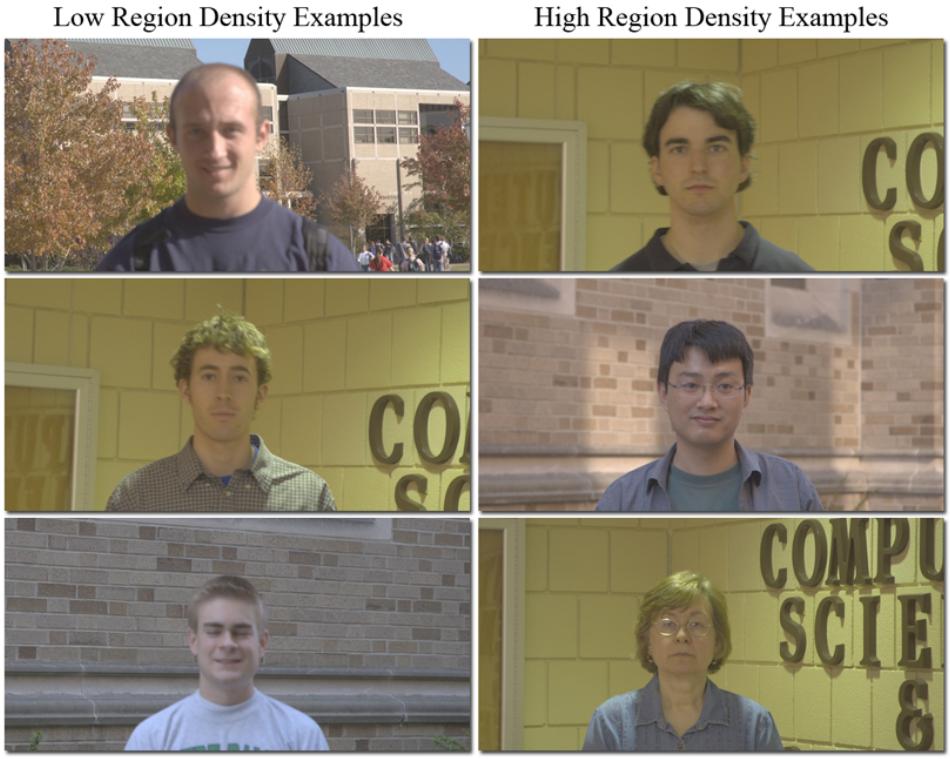

Figure 5: A sampling of low and high Query Region Density images.

this measure, and we present a sampling in figure 5.

4.3. Query Face Size

Face size, as measured by the number of pixels between the eyes, was recorded for both target and query images. The average target and Query Eye Distances are 448 and 168 pixels respectively for those specific FRVT 2006 images included in this study. The associated standard deviations are 39 and 34 respectively. Given the relatively large face sizes in the target images, the variation is comparatively small and while we looked we did not find a significant target face size effect.

In contrast, as shown in Figure 6, the size of the query face images matters a great deal in terms of probability of verification. Note that the three probability of verification levels shown for each Algorithm and Setting combination in Figure 6 are associated with the mean Query Eye Distance minus

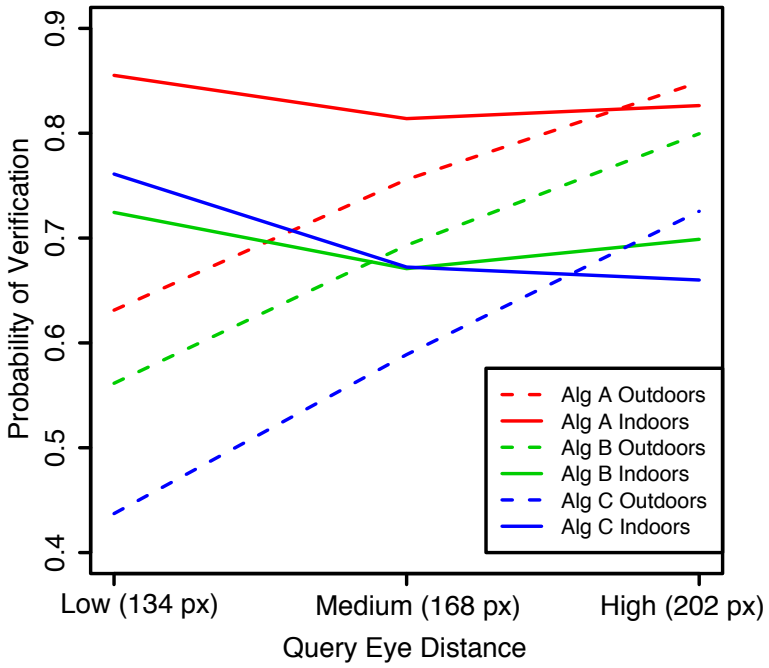

Figure 6: Response of three algorithms to Query Eye Distance. Probability of verification is plotted for low, medium and high values, which correspond to 134, 168, and 202 pixels between the centers of the eyes.

one standard deviation, the mean Query Eye Distance, and the mean Query Eye Distance plus one standard deviation. In other words, Low, Medium and High on the horizontal axis corresponds to 134, 168, and 202 pixels between the eyes.

Several aspects of this finding are striking. First, there is a very significant interaction between Setting and Query Eye Distance. For outdoor query images, the probability of verification goes from a low of 0.64 to a high of 0.86 for algorithm A. Moreover, while the absolute levels for probability of verification differs between algorithms, the overall direction and magnitude of the affect is the same for all three algorithms on outdoor query images.

The influence of Query Eye Distance on indoor images is markedly different. Overall, changes in Query Eye Distance do not greatly alter the

estimated probability of verification. Further, the effect differs somewhat between algorithms. For example, for algorithm C larger face size is associated with a monotonic decrease in verification probability. In contrast, for algorithms A and B, there is actually a small curve in the shape of the response that even makes it impossible to support a simple statement such as small faces are better.

One interesting difference between the previous fusion algorithm study and this study is that for the fusion algorithm there was a significant interaction between Query Eye Distance, Setting and Edge Density. While selecting the model being presented here no significant interaction between Query Eye Distance and Edge Density was found. There is a significant interaction between Algorithm, Setting and Edge Density which we will present next.

4.4. Edge Density

Before proceeding to the results, a comment about nomenclature is appropriate. In our previous study [8] we used the term 'focus' to discuss edge density. We did this because our choice of edge density was motivated by a desire to measure, after the fact, whether an image is in focus and Krotkov [22] suggested edge density is a good surrogate for image focus. Inspection of our own low and high edge density images subsequently lead us to conclude that while often low edge density images are out of focus, not surprisingly, edge density picks up on many other aspects of an image as well. As can be seen in the examples in Figure 7, it appears factors such as strong lighting or hair across the face can result in elevated edge density. Therefore, what was called focus in our previous paper is now described using the more explicit term 'edge density'. The precise way edge density is computed is described in Appendix B.

In our study of the fusion algorithm [8] we concluded that by far the most interesting finding was a four way interaction between Query Face Size, Setting and Edge Density. Here too, we think that the edge density finding is the most interesting. Again we see edge density interacting significantly with Setting. In addition, the dependence on Query Face Size has gone away and a new significant interaction between Algorithm and and Edge Density is evident.

Figure 8 summarizes the results for Query Edge Density, Target Edge Density, Setting and Algorithm. The two rows show results for query images taken indoors versus outdoors and the three columns show results for each of

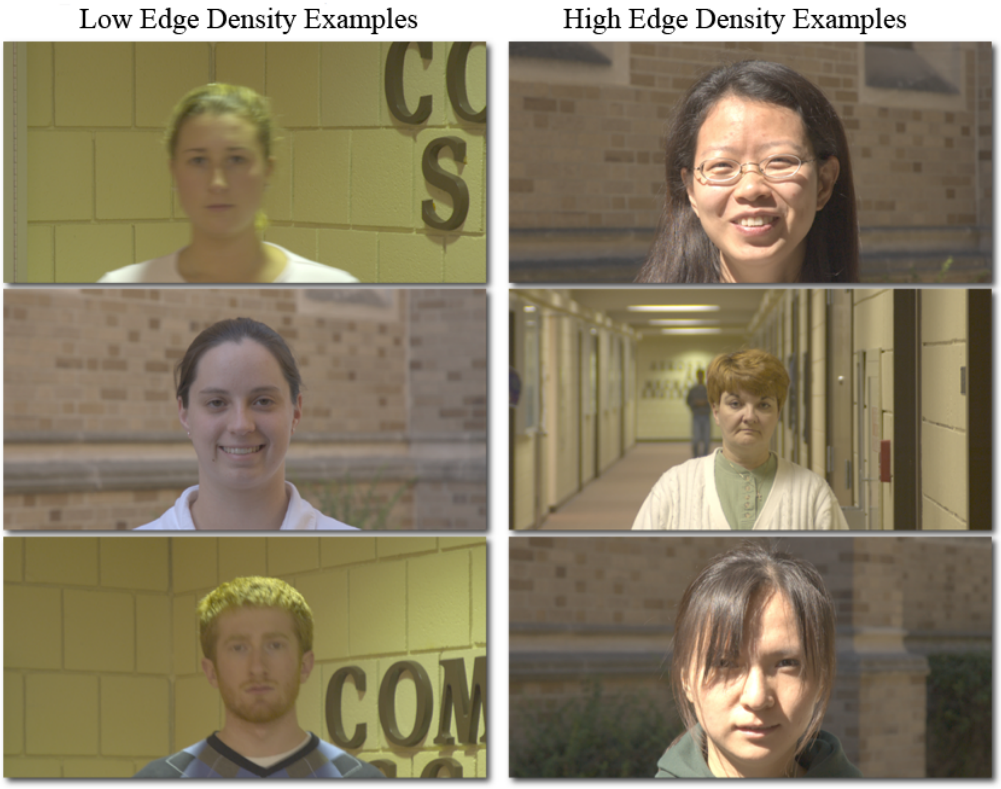

Figure 7: Examples of low and high edge density images.

the three algorithms. Each of the six surface plots indicate the probability of verification using a thermal color encoding. 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.

There are three major aspects of the results shown in Figure 8 to which we wish to draw attention. First, observe the dramatic range in estimated probability of verification associated with different edge query densities. In particular, for the outdoor imagery and algorithm A the estimated probability of verification drops from a high of nearly 0.90 to a low of 0.10 when Query Edge Density increases from 15 to 70. This is an almost astonishing

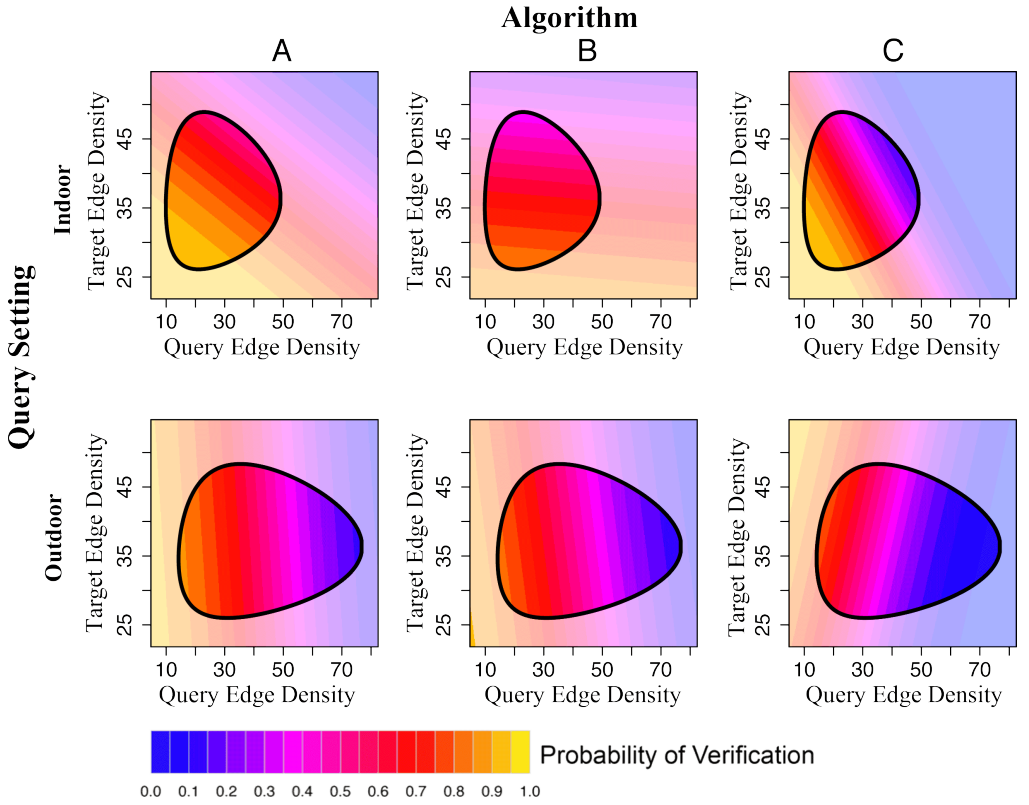

Figure 8: Interaction between Query Edge Density, Target Edge Density, Setting and Algorithm.

level of association between a simple image property and the performance of a face recognition algorithm. In terms of sheer practical significance, it seriously out ranks the rather modest change ascribed to changes in region density. Further, while the result on outdoor query images for algorithms B and C are not identical, they are similar. It is also worth noting that others have also reported a relationship between edge density and face recognition performance [5, 7].

The second observation is that the role played by edge density in the query and target imagery is fundamentally different for the indoor query images. In other words, while a simple statement to the effect that low edge density is good in outdoor query images pretty much summarizes the findings for outdoor images, no analogously simple statement can be made for the

indoor query images. For the indoor query images, density measured in the target image plays an important role for algorithms A and B. In particular, for algorithm A, edge density in the target and query images play roughly equal roles and it is advantageous to have low edge density in both the target and the query image. This is not the case for algorithm B, where the relative importance of edge density in the target and query images is opposite to that observed in the other five cases. In other words, edge density in the target image is strongly associated with performance while edge density in the query imagery is nearly irrelevant.

The third observation concerns the overall nature of the interactions summarized in Figure 8. At the outset of this paper we stated that we found it extremely unlikely that simple measurable properties of imagery would ever yield significant and universal predictions of face recognition reliability. Looking over the results summarized in Figure 8 , it is now possible to explore that claim further in light of hard quantitative evidence, and the whole matter turns on the term 'universal'. The results for edge density are indeed highly encouraging to those of us interested in finding simple measurable image properties that yield significant predictions of face recognition reliability. What these results undermine is any confidence in thinking of such measures as independent of other factors. In particular, here the incredibly key role played by Setting is made explicit, and no where is this more evident in the complete reversal between the role played by query and target image edge density for algorithm B moving between indoor and outdoor query images.

5. Conclusion

Researchers in fingerprint recognition have reported success in their efforts to develop universal measures of biometric quality [2, 23, 24]. It is therefore an interesting question to ask why it has been difficult to find universal quality measures for face recognition. At least two aspects of face recognition evident in this study come to mind. The first is that the variety of algorithms and presumably their features used to encode faces is much greater than in fingerprints. Second, the external sources of variability are intrinsically larger.

This first observation is backed up by what we've just seen about algorithms A, B, and C and the marked difference in how they have responded to Setting or Edge Density. One has to infer from these results that there

are substantial internal differences in how these algorithms encode faces and perform recognition.

The second issue is the variability of face images. The characterization of the variability is arguably harder to pin down since ultimately the variability is tied to an application. In our results, we found the shift from acquiring images indoors to outdoors showed a significant change in the observed behavior of the three algorithms. Further, there is no simple universal generalization that captures how the change in Setting influences the behavior of the algorithms.

It therefore seems fair to conclude that it is essential to continue the exploration of factors that matter in terms of improving face recognition performance, it is problematic at best to commit ourselves prematurely to an overly simplistic interpretation of what constitutes a high-quality face image. For specific scenarios and classes of algorithms, trends will emerge as additional studies extend our understanding of the factors are associated with successful face recognition.

A. Age, Gender, Race and Glasses

Figure 9 shows results for Age, Gender, Race and Glasses. The Age finding further corroborate those of previous studies that older people are more easily recognized. The distribution of young versus old people in the data set used here is highly skewed towards younger people and consequently this is not the best data set in which to carry out a careful study of age. Note therefore our plot in figure 9 has been truncated at age 30 in order to avoid suggesting findings based on only a handful of people.

Perhaps the most interesting aspect of the findings with respect to Gender is the lack of consistency between Algorithms and Settings. Algorithm B is particularly striking in terms of what happens when trying to recognize women indoors versus outdoors. Note the estimated probability of verification indoors is about 0.75 and it drops to about 0.62 outdoors. In light of the fact that algorithm B performs essentially equally well indoors and outdoors for men, the result for women is intriguing. The interaction between Setting and Gender is much less notable for algorithms A and C.

All three algorithms found East Asians easier to recognize than Caucasians. We have observed this sort of result before, where an ethnic group comprising less than 50% of the data set exhibits an advantage in terms of estimated probability of verification [10]. Note that the 30k and 74k shown on

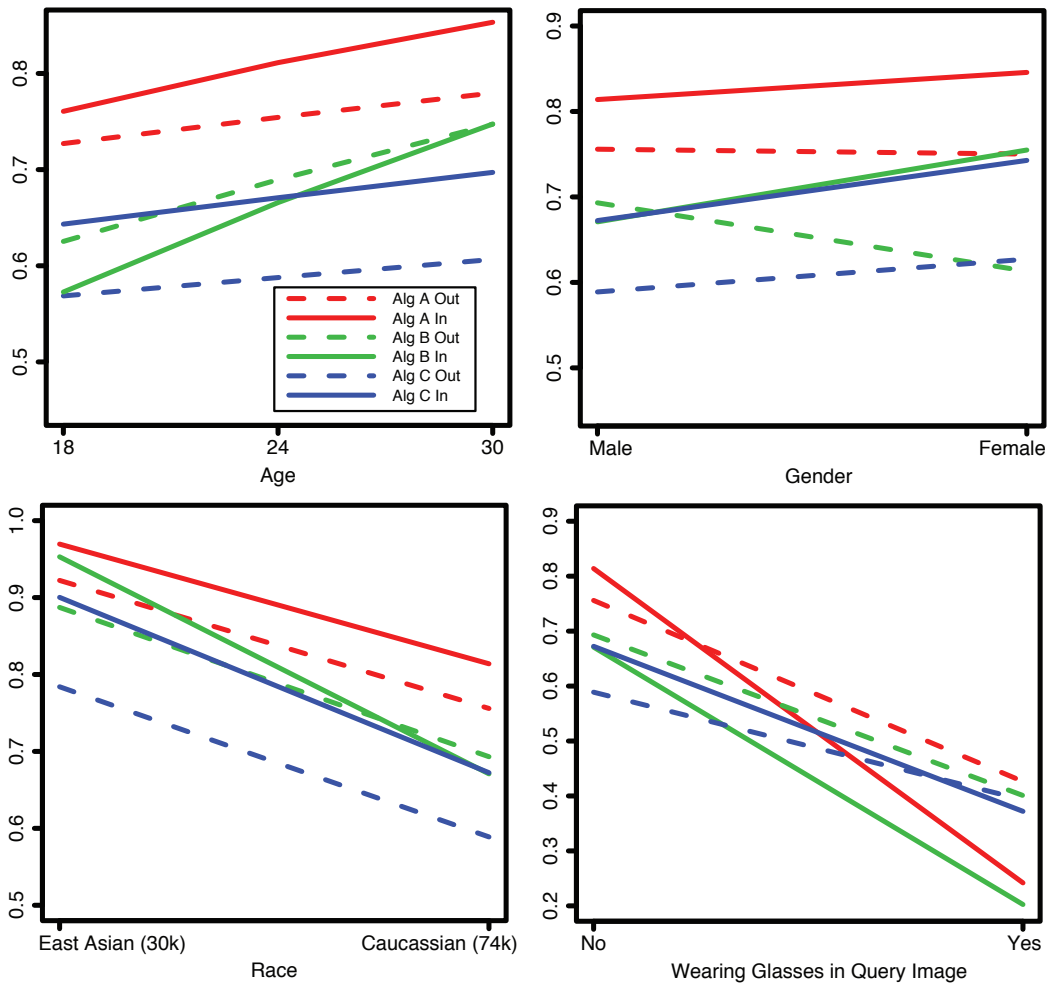

Figure 9: Covariate effects for Age, Gender, Race and Glasses. The vertical axis is estimated probability of verification and the legend is the same for all four plots.

the horizontal axis indicate the number of distinct observations, i.e. matched pairs, for each of the two groups. It is of some interest that we observe that algorithm B, which unlike algorithms A and C, was developed in Asia experiences a more marked decrease in estimated probability of verification when shifting from East Asians to Caucasians.

Perhaps the least surprising finding in our entire study is that the estimated probability of verification drops precipitously for people who were photographed without glasses during enrollment, i.e. during the acquisition of the target images, and who then chose to wear glasses in the query images. For algorithm A, the drop for indoor query images goes from an estimated probability of verification of 0.82 down to an estimated probability of verification of 0.26.

Somewhat more surprising, however, is the fact that going outdoors actually improves the estimated probability of verification for people wearing glasses. Again using algorithm A to illustrate, the estimated probability of 0.26 climbs to 0.48 when the person is outdoors. The one last comment we would make about glasses is that algorithm C, which on the whole has the lowest level of performance, demonstrates considerably less sensitivity to glasses. So much so that, for the indoor query images with people wearing glasses algorithm C has a higher estimated probability of verification then either algorithms A or B.

B. Computation of Edge Density

Edge density as discussed in this paper is based on a measure that has been shown to perform well at estimating the quality of focus of an image [22]. Computing the measure is simple. First, the original image (Figure 10a) is normalized using code from the CSU Face Recognition Evaluation System[25]. This code geometrically registers the face and uses an elliptical mask centered on the face to remove non-face pixels (Figure 10b).

The edge density is produced by first finding the image derivative in both the horizontal (D_x) and vertical (D_y) dimensions by convolving the image with Sobel filters. The Sobel filters are unnormalized which produces values that are 8 times larger than a standard image derivative. The edge magnitude (E_m) is computed as the magnitude of the gradient (Figure 10c).

$$E_m = \sqrt{D_x^2 + D_y^2} \tag{1}$$

Edge Density is the average edge magnitude for all of the unmasked pixels on the face.

C. Computation of Region Density

The face region density is an estimate of grayscale homogeneity of the face region. Like Edge Density, images are first processed using the normalization code from the CSU Face Recognition Evaluation System[25] to produce a gray scale image geometrically normalized. An elliptical mask centered on the face is used to remove non-face pixels (Figure 10b).

To identify contiguous regions in the resulting face image, the Rutgers image segmentation algorithm[20] was run on the normalized images to produce a labeled segmentation. The segmentation algorithm produces three outputs which refer to as Low (under), Medium (quant), and High (over). The Medium setting was used for our model (Figure 10d). The Region Density used in the GLMM is the count of distinct regions found within the masked face oval.

References

[1] P. J. Phillips, W. T. Scruggs, A. J. O'Toole, P. J. Flynn, K. W. Bowyer, C. L. Schott, M. Sharpe, FRVT 2006 and ICE 2006 large-scale results, IEEE Transactions on Pattern Analysis and Machine Intelligence (in press). 2, 5, 6

[2] P. Grother, E. Tabassi, Performance of biometric quality measures, IEEE Trans. Pattern Analysis Machine Intelligence 29 (2007) 531–543. 2, 24

[3] A. Zhang, R. Blum, Image quality estimation using edge intensity histogram and a mixture model, in: Proceedings: Image Understanding Workshop, Morgan Kaufmann, 1998. 2

[4] P. J. Phillips, E. Newton, Meta-analysis of face recognition algorithms, in: Proc. 5th International Conference on Automatic Face and Gesture Recognition, 2002, pp. 235–241. 2

[5] T. Boult, Beyond image quality – failure analysis from similarity surface techniques, in: Biometric Quality Workshop, 2006. 2, 23

a.

b. c. d.

Figure 10: This figure illustrates the computation of the Edge Density and Region Density covariates. (a) Is the source image. (b) The geometrically normalized and masked image. (c) The edge magnitude image used to compute Edge Density. (d) The segmented image used to compute Region Density.

[6] P. Wang, Q. Ji, Performance modeling and prediction of face recognition systems, in: IEEE Conference on Computer Vision and Pattern Recognition, Vol. 2, 2006, pp. 1566 – 1573. 2

[7] F. Weber, Some quality measures for image faces and their relationship to recognition performance, in: Biometric Quality Workshop, 2006. 2, 23

[8] J. R. Beveridge, G. H. Given, P. J. Phillips, B. A. Draper, Y. M. Lui, Focus on quality, predicting FRVT 2006 performance, in: 2008 8th IEEE International Conference on Automatic Face and Gesture Recognition, 2008. 4, 6, 9, 12, 21

[9] Geof H. Givens, J. Ross Beveridge, Bruce A. Draper, Patrick Grother, 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, 2004, pp. 381–388. 4, 9, 11

[10] J. R. Beveridge, G. H. Givens, P. J. Phillips, B. A. Draper, Factors that influence algorithm performance in the face recognition grand challenge, Computer Vision and Image Understanding (2009) (in press). 4, 11, 25

[11] A. J. O'Toole, P. J. Phillips, F. Jiang, J. Ayyad, N. Pénard, H. Abdi, Face recognition algorithms surpass humans matching faces across changes in illumination, IEEE Trans. PAMI 29 1642-1646 (2007) 1642–1646. 5

[12] A. O'Toole, H. Abdi, F. Jiang, P. J. Phillips, Fusing face recognition algorithms and humans, IEEE Trans. on Systems, Man & Cybernetics Part B 37 (2007) 1149–1155. 5

[13] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, W. Worek, Overview of the face recognition grand challenge, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005, pp. 947–954. 5

[14] P. J. Phillips, P. J. Grother, R. J. Micheals, D. M. Blackburn, E. Tabassi, J. M. Bone, Face recognition vendor test 2002: Evaluation report, Tech. Rep. NISTIR 6965, National Institute of Standards and Technology, http://www.frvt.org (2003). 5

[15] P. J. Phillips, H. Moon, S. Rizvi, P. Rauss, The FERET evaluation methodology for face-recognition algorithms 22 (2000) 1090–1104. 5

[16] 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). 5

[17] N. E. Breslow, D. G. Clayton, Approximate inference in generalized linear mixed models, J. Amer. Statist. Assoc. 8 (1993) 9–25. 7

[18] R. Wolfinger, M. O'Connell, Generalized linear models: a pseudo-likelihood approach, Journal of Statistical Computation and Simulation 48 (1993) 233–243. 7

[19] E. P. Krotkov, Active Computer Vision by Cooperative Focus and Stereo, Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1989. 9

[20] D. Comaniciu, P. Meer, Robust analysis of feature spaces: Color image segmentation, in: 1997 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1997. Proceedings., 1997, pp. 750–755. 9, 28

[21] G. Doddington, W. Liggett, A. Martin, M. Przbocki, 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), Syndey, Australia, 1998, pp. 1351–1354. 10

[22] E. P. Krotkov, Active Computer Vision by Cooperative Focus and Stereo, Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1989. 21, 27

[23] J. Fierrez-Aguilar, Y. Chen, J. Ortega-Garcia, A. Jain, Incorporating image quality in multi-algorithm fingerprint verification, in: Advances in Biometrics, Vol. Volume 3832/2005, Springer Berlin / Heidelberg, 2005, pp. 213–220. 24

[24] J. O.-G. Fernando Alonso-Fernandez, Julian Fierrez, J. Gonzalez-Rodriguez, H. Fronthaler, K. Kollreider, J. Bigun, A comparative study of fingerprint image-quality estimation methods, IEEE Transactions on Information Forensics and Security 2 (4) (2007) 734–743. 24

[25] J. R. Beveridge, D. Bolme, B. A. Draper, M. Teixeira, The CSU face identification evaluation system, Machine Vision and Applications 16 (2) (2005) 128–138.
URL http://dx.doi.org/10.1007/s00138-004-0144-7 27, 28