Computer Vision and Image Understanding 113 (2009) 750–762

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Factors that influence algorithm performance in the Face Recognition Grand Challenge ☆

J. Ross Beveridge [a,*], Geof H. Givens [b], P. Jonathon Phillips [c], Bruce A. Draper [a]

[a] Department of Computer Science, Colorado State University, Fort Collins, CO 80523-1873, USA
[b] Department of Statistics, Colorado State Univesity, Fort Collins, CO 80523-1877, USA
[c] National Institute of Standards and Technology, Gaithersburg, MD 20899, USA

ARTICLE INFO

Article history:
Received 21 December 2007
Accepted 30 December 2008
Available online 30 January 2009

Keywords:
Face recognition
Subject covariates
Performance analysis
Statistical modeling

ABSTRACT

A statistical study is presented quantifying the effects of covariates such as gender, age, expression, image resolution and focus on three face recognition algorithms. Specifically, a Generalized Linear Mixed Effect model is used to relate probability of verification to subject and image covariates. The data and algorithms are selected from the Face Recognition Grand Challenge and the results show that the effects of covariates are strong and algorithm specific. The paper presents in detail all of the significant effects including interactions among covariates.

One significant conclusion is that covariates matter. The variation in verification rates as a function of covariates is greater than the difference in average performance between the two best algorithms. Another is that few or no universal effects emerge; almost no covariates effect all algorithms in the same way and to the same degree. To highlight one specific effect, there is evidence that verification systems should enroll subjects with smiling rather than neutral expressions for best performance.

1. Introduction

Over the last two decades the effort to develop effective automatic face recognition has resulted in hundreds if not thousands of papers [1]. Typically, these papers report performance on a single dataset in order to draw comparisons among competing approaches [2]. This type of analysis is valuable when the goal is to conclude that a particular approach is superior to another on a very specific task as exemplified by the dataset. However, this style of analysis tells us little about underlying factors that make recognition easier or harder.

When it is addressed at all, the question of what factors affect recognition performance is almost invariably addressed by dataset partitioning. For example, several carefully constructed datasets have been developed for studying the effects of pose and illumination through partitioned data. Work with the Yale [3] and PIE [4] datasets typically falls into this category. Studies look at relative performance across changes in pose, illumination, or both as exemplified by performance on distinct data partitions. Partitioned dataset analysis has also been applied to the question of whether women or men are easier to recognize [5,6].

Unfortunately, data partitioning is ineffective for answering questions about more than a few factors, or for answering questions about interactions among factors. From a practical perspective, partitioning quickly becomes infeasible due to the combinatorial explosion of partitions over multiple factors. Moreover, it is difficult to control for confounding factors with partitioning schemes. If one skirts around the combinatorial problem by resorting to marginal analysis (i.e., abandoning control via partitioning), control of confounding effects is eliminated altogether. More sophisticated multi-factor statistical techniques provide greater control and permit more thorough evaluation of factor effects.

Generalized linear mixed models (GLMMs) are one such technique. This paper uses GLMMs to provide the largest statistical analysis to date of factors that influence face recognition performance. Our analysis investigates how a set of factors, henceforth called covariates, predict verification rate at various false accept rates. Performance data for three algorithms from the Face Recognition Grand Challenge Experiment 4 are used in our analysis. The covariates include gender, race, age, distance between eyes in pixels, and image focus quality. The complete set of covariates are described below.

Multi-factor statistical analysis represents an important step forward on the path toward developing a mature empirical understanding of what covariates make recognition harder or easier. It is

☆ The work was funded in part by the Technical Support Working Group (TSWG) under Task T-1840C. P.J.P. was supported by the National Institute of Justice. The identification of any commercial product or trade name does not imply endorsement or recommendation by the National Institute of Standards and Technology or Colorado State University.

* Corresponding author.
E-mail address: ross@cs.colostate.edu (J.R. Beveridge).

1077-3142/$ - see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.cviu.2008.12.007

broader in scope, more powerful statistically, more efficient with data, and more precise for inference. With its greater control of confounding factors it also affords greater confidence that the conclusions will generalize to other datasets.

We have previously used GLMM analysis to characterize the performance of face recognition algorithms [7–9]. There are few other instances of these techniques being used to characterize the performance of face recognition algorithms, although it should be noted that Mitra et al. [10] use a GLM with random (rather than mixed) effects to predict the performance of a biometric authentication system in a watch-list scenario from the dataset size. This allows them to predict performance over datasets that are much larger than the tested sets, but provides no information about the effects of specific factors.

The study presented here goes beyond our own prior work in several important ways. First, our earlier work focused only on the FERET data and our own implementations of algorithms included in the original FERET tests. This paper presents results for the two top performing algorithms in the Face Recognition Grand Challenge Experiment 4. Hence, the dataset is newer and more challenging and the algorithms represent the state-of-the-art in the time frame of the Face Recognition Grand Challenge, i.e., 2005–2006. Second, the presence of more imagery per person enables us to design a more data rich study. Finally, the study presented here is our first to include measured properties of the images themselves. As the results will show, some of these are strong predictors of algorithm performance. This opens up a connection between our work and the larger question of how to characterize the quality of a face image in terms of measured properties of that image.

Most of the findings presented in this paper have not been observed before in a major empirical study of face recognition performance. For example, the results summarized in Section 5.2.1 suggest that when a single face image is enrolled it is better to let the person smile than not. This directly contradicts the common wisdom among many who are designing biometric collection protocols. Another new finding, presented in Section 5.2.2, indicates that one of the top performing algorithms is not influenced by elapsed time between images as measured in months. Common wisdom dictates that increasing time between the acquisition of images makes recognition harder. To offer a third example, Section 5.2.3 presents an important interaction between resolution of target and query images that suggests relative size of one versus the other is more important than the absolute number of pixels available. Most discussions of image resolution in the context of face recognition focus on the absolute resolution of individual images rather than the importance of the relative sizes of the images being compared, and consequently ignore what our data suggests may be extremely important.

Other findings presented here bolster results from previous empirical studies. For example, Section 5.3.4 presents findings showing recognition performance is better on Asian subjects when compared to White subjects even though Asians constitute only 26% of the sample population compared to 72% for Whites. This kind of result favoring a non-majority group has been seen before. While it is relatively easy to speculate on why algorithms might behave in this fashion, to our knowledge algorithm developers have yet to really explore in depth what is going on in this situation. Such future work offers the promise of improving overall recognition performance, and it begins when algorithm developers become aware of the phenomenon.

The overall outline of this paper is as follows. The next section reviews the Face Recognition Grand Challenge and explains why Experiment 4 has been selected as the focus of our study. Section 3 motivates our approach. A summary of results appears in this section, highlighting how much influence these covariates exert over recognition performance. Section 4 introduces the formal statistical model, and describes the complete set of covariates examined and the full experimental design. Section 5 presents the results of our analysis and Section 6 summary conclusions.

2. Face Recognition Grand Challenge Experiment 4

Version 2.0 of the Face Recognition Grand Challenge (FRGC) specified six separate experiments [11]. Experiments 1 and 2 compared frontal still images of faces taken under mug shot lighting conditions. Experiments 3, 5 and 6 involved 3-D face data. Our focus is on Experiment 4, which used frontal still 2-D images. Unlike Experiments 1 and 2, however, some of the imagery in Experiment 4 was acquired with uncontrolled lighting. These images were taken in a hallway with subjects standing facing the camera. These uncontrolled conditions make recognition harder, and consequently performance between algorithms varied much more in Experiment 4 than in either Experiments 1 or 2. Fig. 1 shows two controlled and two uncontrolled images from the FRGC Experiment 4 data.

Experiment 4 presents algorithms with problems that range from easy to hard. This is important since the goal of our analysis is to identify covariates or sets of covariates that make recognition easier or harder, and this influence can only be measured using a dataset with sufficient variation in problem difficulty. Experiment 4 is also a good choice because the scenario is of practical importance in that it closely mimics potential implementation protocols.

The performance task for Experiment 4 was verification. Specifically, algorithms were given the problem of deciding whether a person was who they claimed to be. The decision must be made by comparing two images. One image—the target image—was presumed to have been taken earlier and to reside in a database. The other image—the query image—served as the 'claim' and was presumed to have been taken at the time the person attempted to have his or her identity verified.

Verification performance on Experiment 4 can be summarized in two simple ways. When a single performance number is reported for an algorithm, the number reported is the verification rate at a false accept rate (FAR) of 1 in 1000. A FAR of 1 in 1000 was selected because it is the standard operating point for reporting verification rates in the FRGC [12]. To report verification rates over a range of false accept rates, ROC curves are used. Average

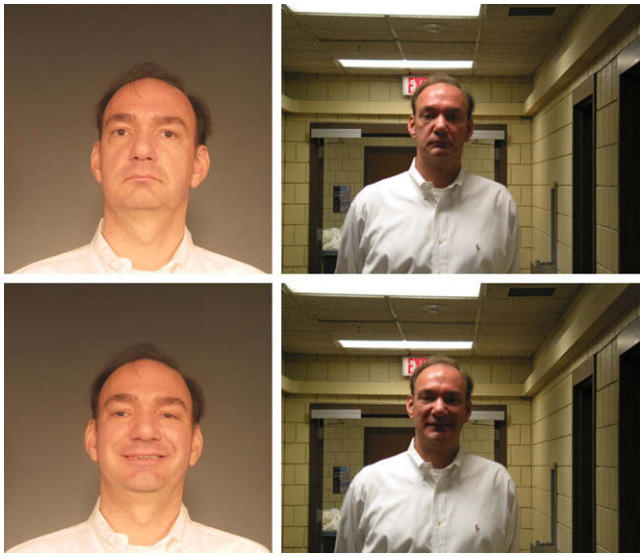

Fig. 1. Sample imagery from the FRGC Experiment 4 data.

verification rates for different algorithms at the 1 in 1000 false accept rate ranged from about 10% up to about 75% [12].

The dataset for Experiment 4 contains images of 466 people. These images were acquired over multiple sessions during the Fall 2003 and Spring 2004 Semesters at the University of Notre Dame. During a session, a subject would sit while four still digital photographs were taken under controlled lighting conditions. Two images were taken with controlled lighting coming from two professional photographer's lamps, one to either side of the camera. Two more images were taken with a third photographer's lamp added in the center. Subjects were also asked to stand, typically in a hallway, and face a camera. Two more photographs were then taken. These hallway images were taken in uncontrolled lighting conditions which were highly variable and often quite poor. Examples of controlled lighting and uncontrolled lighting images are shown on the left and right of Fig. 1, respectively.

The reason for taking pairs of photographs in each setting was to explore neutral versus smiling expressions. The labeling of an image as smiling or neutral is based on the instructions provided in the FRGC collection protocol. Under each lighting configuration, subjects were instructed to adopt two facial expressions: one neutral, the other smiling. The quality and degree of each expression depends on each subject's normal manifestation of each expression and willingness to follow instructions. Over the FRGC dataset there is a notable distinction between the two expressions. Neutral and smiling expressions are illustrated in the top and bottom row of Fig. 1, respectively.

FRGC Experiment 4 compares target images taken under controlled lighting to query images taken under uncontrolled lighting. This models a scenario in which subjects are enrolled under controlled conditions, but must be verified under uncontrolled conditions. The target set for Experiment 4 contains 16,029 controlled lighting images; the query set contains 8014 uncontrolled lighting images. Comparing every query image to every target image produces over 128 million similarity scores per algorithm.

Twelve algorithms were tested in the FRGC. Three of these algorithms, designated A, B and C, have been selected for study in this paper. Algorithm A uses principal component analysis (PCA), and was used as a baseline in FRGC. It performs relatively poorly. Algorithms B and C were submitted by Carnegie Mellon University (CMU) and the New Jersey Institute of Technology (NJIT), respectively. These algorithms were chosen for study here because their performance represented a breakthrough on the FRGC Experiment 4 data. Roughly speaking, a 30% verification rate at a FAR of 1 in 1000 represented median performance for algorithms on Experiment 4. Algorithm B and C roughly doubled this rate. Both research groups have subsequently published descriptions of their approaches [13,14].

The results from NJIT were provided to NIST for analysis in January 2005 and the results from CMU were provided in August 2005. The analysis in this paper reflects properties of these two algorithms on the submission dates. In the remainder of the paper the algorithms are designated by A, B, and C to emphasize the analysis, not the properties of a particular instantiation of an algorithm.

3. Motivation for covariate analysis

Our goal is to measure whether, and to what degree, a covariate or combination of covariates influences the performance of face recognition algorithms. Our analysis fits a generalized linear mixed model (GLMM) to a carefully designed dataset built from the FRGC Experiment 4 data. The details of our design, model, and results are presented in the following sections. Before taking up the full model in detail, it is helpful to illustrate that the Experiment 4 dataset contains important and scientifically interesting performance information that can be extracted using a GLMM.

Fig. 2 provides a simple visual summary of performance variations at the standard false accept rate of one in one thousand, i.e., FAR = 0.001, estimated from the GLMM developed in this paper. In this plot, the vertical axis shows verification rates associated with changing values of subject covariates such as age, gender and race. The horizontal axis shows verification rates associated with changing values of image-derived covariates such as image focus, resolution and rotation. The specific subject and image covariates being varied are summarized in Table 1 and are explained Section 4.2.

The three crosses in Fig. 2 indicate the range of predicted verification rates associated with changes to the subject and image covariates. The center of each cross is located at the predicted overall verification rate for the algorithm at FAR = 0.001. To avoid hidden extrapolation with the continuous image covariates, we found the vector described by the marginal 5th or 95th percentiles of each covariate, then shrunk this vector back towards the mean until its Mahalanobis distance was within the 95% joint probability region of the appropriate Chi-squared distribution. The range of performance shown in the figure relates only to this range of covariate variation.

Fig. 2 shows that the estimated verification rate varies enormously depending upon the values of the covariates used in our model. This performance variation is observed both within and between algorithms, and is affected by both subject and image covariates. The overlap of the crosses for Algorithms B and C further suggests covariates influence expected performance at least as much as choice of algorithm. Indeed, model results shown later will illustrate that C does not dominate B: there is a substantial minority of cases where the relative performance of B and C is reversed.

Also observe in Fig. 2 that the extent of performance degradation associated with Algorithm B is smaller than for Algorithm C, particularly with respect to variation in subject covariates. This is evidence for a covariate-algorithm interaction and is our first hint that covariates influence algorithms differently. As we present the results of our full model, the presence of covariate-algorithm interactions will become an important theme leading us to conclude that there are few universal covariates. In other words, we have found practically no covariate that influences all three algorithms in the same manner and to the same extent.

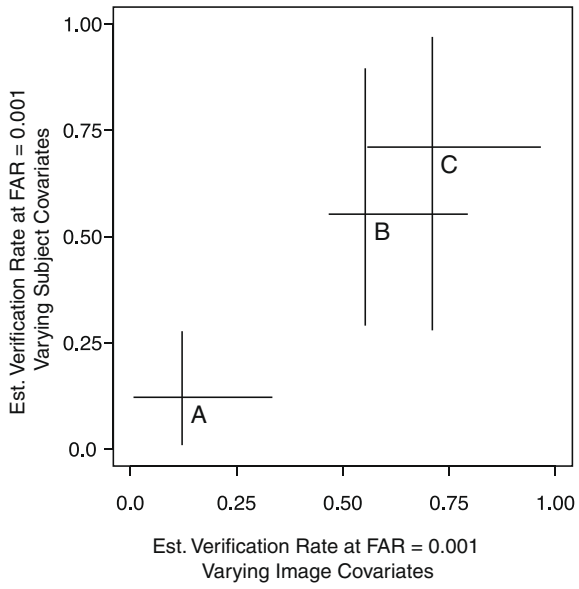

Fig. 2. Performance range associated with varying subject and image-derived covariates, for Algorithms A, B, and C.

Table 1
Subject and image covariates used in our analysis. For discrete covariates, the second column lists covariate levels and proportional representations in the dataset. For quantitative covariates, this column shows the mean and range of values observed. Capital Q and T are used as shorthand for query and target image, respectively. Facial expressions were categorized as neutral (N) or smiling (S). Details about the covariates are included with the discussion of covariate results in Section 5.

	Values, or median and range
Subject covariates	
Gender	M (0.55), F (0.45)
Race	White (0.66), Black (0.01), Asian (0.26), Hispanic (0.04), Unknown (0.03)
Query wearing glasses	N (0.83), Y (0.17)
Age	21, (17, 69)
Expression Q:T	NN (0.25), SN (0.25), NS (0.25), SS (0.25)
Elapsed months, Q to T	0.9, (0, 7.9)
Image covariates	
Image size ratio, Q/T	0.55, (0.39, 0.81)
Query image resolution (pixels between eyes)	142, (104, 200)
Tilt difference, $\|Q - T\|$ (°)	0.03, (0, 0.41)
Focus difference, $Q - T$	−1.2, (−54.6, 45.8)
Focus average	42.4, (21.7, 63.8)
Fragmentation	8.0, (7.7, 12.9)
Novelty difference, $Q - T$	−0.04, (−0.26, 0.17)
Novelty average	0.08, (0.01, 0.23)

In general, while the importance of covariates is evident to many, through statistical modeling this paper goes further than most by:

1. identifying which covariates matter,
2. quantifying how different covariates affect performance,
3. measuring interactions between covariates, and
4. quantifying how covariates affect algorithms differently.

4. Methods

4.1. Overview

We use empirical performance and covariate data associated with people and imagery to fit a model relating covariate values to the probability that a person will be correctly verified. The purpose of the model is to quantify how changes in covariates alter the probability that a person will be correctly verified. This section sketches how this is done with additional detail coming in Sections 4.2 and 4.3.

The raw data for our analysis are based on pairs of matching query and target images along with an indication of whether the person pictured was correctly verified using a particular algorithm at a specific FAR setting. The response variable is binary: it equals 1 if the query was correctly verified and 0 otherwise. The other data used are covariate measurements. Covariate measurements may, as in the case of gender, have been recorded along with the original data. Covariates may also be automatically computed off of the imagery itself, as in the example of estimates of image focus. We discuss covariates further below. The match similarity score itself is not directly analyzed, but it is used for determining whether verification was successful at a given FAR.

A generalized linear mixed model (GLMM) with logit link is used because our response variable is binary and we have repeated measures on individuals. Except for a random effect for subject, all other effects (e.g., for gender) are treated as estimable fixed effects. Inclusion of the random subject effect in the model is important because it is well known that some people are inherently harder to recognize than others [15,8,9] and because we view the Experiment 4 individuals as a random sample from a larger population of interest.

The details of GLMM modeling are too complex to fully explain here; see [16,17]. In general, it is important to understand that model is closely related to the notion of odds. In our context, let p_i be the verification probability for a person of type i, as identified by their covariates. The odds of verification for type i is $p_i/(1-p_i)$. A GLMM models the log odds with a linear predictor that is a function of the covariates.

The effects of covariates in a GLMM are most easily expressed in terms of odds ratios. If we want to compare the ease of verification of type i to that of type j, we may use the ratio of their odds. The *odds ratio for verification, type i relative to type j* is

$$OR = \frac{p_i/(1-p_i)}{p_j/(1-p_j)}. \quad (1)$$

An odds ratio of 1.0 indicates that there is no difference between the odds for i and j. We use estimates of both odds ratios and verification probabilities to describe our results, and for GLMMs we adopt the SAS convention that odds ratios are calculated from the fixed-effect portion of the model. Section 4.2 will provide additional detail pertinent to the specific model developed here.

As the reader considers the results from our GLMM, it is critical to understand that a GLMM quantifies effects over a set of covariates and that the result should never be confused with a 'marginal analysis' obtained simply by separately tabulating verification outcomes split by the levels of each covariate (e.g., split by gender). Marginal analysis fails to control for the other covariates, whereas the GLMM analysis provides an estimate of the effect of each covariate controlled for all other variables in the model. For this reason, we view the GLMM approach as providing a more appropriate means of isolating the true effects of each covariate.

4.2. The covariates and statistical model

As already suggested by Fig. 2 above, covariates fall broadly into two categories. Subject covariates are properties of the person. Examples of subject covariates include gender and race. A person's expression, smiling versus neutral, is also regarded as a subject covariate. Image-derived covariates are properties of the image, not the person. For example, the extent to which an image is in focus is an image covariate. Some useful covariates do not fit neatly into this simple dichotomy. For example, the distance between a person's eyes (measured in pixels) depends both upon the size of the person's face and the distance between their face and the camera. However, under the data collection protocol for Experiment 4, distance between eyes tells us more about the distance to the camera, and consequently the number of pixels on-face, than it does about the intrinsic geometry of the person's face.

The specific covariates used in the statistical model developed here are summarized in Table 1. The subject covariates for the most part represent demographic information collected along with the images. The image covariates address basic properties of the imagery such as the resolution of the face, tilt of the face and relative focus of the image. These later take on extra practical significance since they represent factors over which those deploying a system might exert some control.

The false accept rate as a covariate merits special attention. We collected data for verification at eight different FAR settings: 1/10, 1/100, 1/200, 1/350, 1/700, 1/1000, 1/2500, 1/5000 and 1/10,000 and 1/10,000. These choices are roughly evenly distributed on a log scale. In a previous study of verification performance on the FERET dataset, we observed a nearly linear relationship between the negative log of the false accept rate and the probability of cor-

rect verification [9]. A similar relationship has been observed for the FRGC data, thus supporting the appropriateness of our decision to model log odds of verification.

Now we can more precisely describe how response variable values were determined. Algorithms produce similarity measures between query and target images. The set of similarity scores can be divided into two groups: match scores, which are computed between 'matching' query and target images of the same subject, and non-match scores, where the query and target images are of different subjects and therefore should not match. To set the false accept rate (FAR), the non-match scores are sorted from greatest to least, and thresholds are chosen to create specific FARs. For example, the similarity score at the 99th percentile of this list would—if used as a verification threshold—yield a FAR of $\frac{1}{100}$ because only 1% of non-match scores would exceed the threshold and thereby be designated (incorrectly) as matches. We used this approach to find similarity score thresholds corresponding to each of the eight FAR settings listed above. For each pair of query and target images for a person (i.e., match pairs), the response variable is 1 if and only if the similarity score between these two images is greater than the verification threshold for the FAR setting at which the image pair is tested. To reiterate, only match pair responses are directly analyzed in our model.

4.2.1. The GLMM

To estimate the effect of these covariates on performance, we fit a generalized linear mixed model (GLMM) with Bernoulli response and random effects for subjects. This class of models is a generalization of the ordinary linear regression (OLR) model. There are several key differences between the GLMM and OLR frameworks which are discussed below as we describe our approach.

A single case, or "trial", is the unit of analysis. It consists of a verification outcome (correct or incorrect) between two face images of the same person and a set of covariates describing the person, features of the images in the pair, and aspects of the verification attempt. All these covariates are described above.

For subject s, let $Y_{i_s,s}$ denote the (binary) outcome of a trial, where i_s indexes the trials for subject s. The outcome of each trial is a Bernoulli random variable, but the probability of success may differ for each trial. For any trial, let $\pi_{i_s,s}$ be the modeled probability of successful verification so $E(Y_{i_s,s}|s) = \pi_{i_s,s}$ is the corresponding Bernoulli parameter.

Our GLMM models the verification probability, $\pi_{i_s,s}$, as a function of the covariates. Unlike the linear relationship required with OLR, our GLMM uses a nonlinear function, g, called the link function. Our link function is the logit, namely $g(\pi) = \log(\pi/(1-\pi))$, which converts Bernoulli probabilities to the log odds scale.

The GLMM asserts that the covariates for each trial are linearly related to the logit of the Bernoulli verification probability for that trial. However, unlike the OLR, the GLMM is constructed so that the linear predictor comprises two parts: a fixed component and a random one. The fixed part models the effects of the covariates and interaction among covariates. The random component can be thought of as a subject effect. It arises because the data include multiple trials for each person due to the existence of multiple face images for each person. Since people exhibit random variation in ease of recognition (i.e., some people are inherently harder to verify than others even after compensating for measured covariates), modeling this effect as a random component can adjust for the person-specific variation in order to permit fully generalizable conclusions.

The fixed part of the GLMM is based on covariates encoded in a matrix of the form $\mathbf{X}^T$, where the columns of $\mathbf{X}^T$ are constructed in the standard manner from the covariates and interactions in the model. These matrices vary by image pair and hence are indexed as $\mathbf{X}^T_{i_s,s}$. The corresponding effects are parameterized with a vector $\boldsymbol{\beta}$. Random effects—one per subject—are represented as γ_s. The nature in which these γ_s terms appear in the model introduces a statistical dependence between outcomes of multiple trials for the same person, while retaining independence between people.

Putting this all together, the relationship between the verification probabilities and the covariates is written as

$$\text{logit}(\pi_{i_s,s}) = \beta_0 + \mathbf{X}^T_{i_s,s}\boldsymbol{\beta} + \gamma_s, \tag{2}$$

where $\boldsymbol{\beta}$ represents the fixed effects of the covariates and γ_s is a random effect indexed by subject and assumed to have mean zero and variance σ^2. Overall, the Bernoulli assumption provides $\text{var}(Y_{i_s,s}|\gamma_s) = \pi_{i_s,s}(1-\pi_{i_s,s})$ and γ_s contributes additional variance of $\text{var}(\gamma_s) = \sigma^2$.

Thus, the mean verification probability for an outcome $Y_{i_s,s}$ is $E(Y_{i_s,s}|\gamma_s) = \pi_{i_s,s}$, as defined in (2) above. Like OLR, hypothesis tests and confidence intervals for the effects of covariates in a GLMM can be constructed from the estimates of $\boldsymbol{\beta}$ and corresponding estimates of precision.

The specific GLMM for which we present results below was fit using the GLIMMIX procedure in SAS 9.1 with conjugate gradient optimization [18]. Additional details may be found through the SAS website.[1] The model includes additive effects for each of the covariates/predictors enumerated in Table 1 as well as for FAR and algorithm (A, B, and C). Interactions with algorithm were also used for FAR and all the subject and image covariates except tilt difference. A three-way interaction between FAR, algorithm, and query eye distance was also included. Model selection was carried out manually, mainly using a backwards elimination philosophy starting from much richer models than the final choice described here.

4.3. FRGC Experiment 4 data

The dataset used in this study contains a total of 134,760 outcomes. Our subject list was chosen as the Experiment 4 subjects for which there were at least 16 target images and 8 query images per subject. Imagery was randomly sub-sampled to these levels for subjects with more abundant imagery. Thus, for each subject there were 128 possible target and query pairings for a total of 128 outcomes.

The 8 FAR settings listed above were allocated evenly across the 128 query target image pairings per subject, yielding 16 trials at each FAR. For each trial, the binary response variable described above recorded whether an algorithm did or did not successfully verify the person at the indicated FAR. Due to the nature of Experiment 4 imagery, query and target facial expressions were also perfectly balanced with respect to subject, FAR, and algorithm. Constructing our dataset in this fashion also balanced the number of outcomes per subject.

Since the Notre Dame image data used in the FRGC includes many more target and query images for some subjects than for others, there is a trade-off to be made between the total number of subjects included in the study and the number of possible outcomes per subject. The specific choice of 128 outcomes per subject allows us to include 351 of the 466 subjects in our study while at the same time generating more than 100 outcomes per subject. This reduction in the number of subjects means our results are not exactly equivalent to those reported in the standard FRGC Experiment 4 ROCs. However, we have checked mean performance on our 351 subjects versus the original ROCs, and performance is comparable.

A question arose regarding how to handle training as a possible covariate. Version 2.0 of the FRGC data included training and vali-

[1] The URL is http://support.sas.com/rnd/app/papers/glimmix.pdf.

dation partitions and the performance results reported here and elsewhere are for the FRGC validation partition. While the training and validation partitions include no images in common, of the 351 subjects included in our study, 133 subjects were included in the FRGC v2.0 training partition. Unfortunately, because the FRGC protocol did not require participants to specify their training sets, the exact images used in training is unknown and we could not include a training covariate in our study.

The complete list of face pairs, covariates and algorithm outcomes used in this study are available through our website.[2] Some additional details concerning the data preparation are also provided at this site.

5. Results

5.1. Overview

Performance varied widely among algorithms. Over all FAR settings and subjects, the rates of correct verification were 15%, 58%, and 70% for Algorithms A, B and C, respectively. At a FAR of 0.001, these rates were 12%, 55%, and 71%. As expected, verification rates increased with increasing FAR.

Model fit appeared adequate, with an estimated over-dispersion parameter of 1.36 and the estimated variance of the random effects being 2.16 on the logit scale [19,20]. In the following sections, we discuss many specific findings about covariate effects. All of these effects were found to be highly statistically significant using traditional testing methods. In part, this is due to the very large sample size. Therefore, in choosing which effects to include in a final model, we required the effect to exhibit both statistical significance and scientific importance as judged by whether the magnitude of the estimated effect was large enough that verification outcomes would be expected to change by at least several people per 100.

In the following sections, predictions of verification probability require specification of the levels of all covariates in the model (whereas odds ratios do not, because they are comparative). For such predictions, all unmentioned variables are set at their baseline levels, which are the first categories listed in Table 1 for qualitative variables and which equal the marginal sample means for quantitative variables. The baseline FAR is 0.001.

5.2. Surprises

This section highlights several findings that are new and unexpected. A key aspect of some of these surprising results relates to asymmetries in the effects of covariates.

5.2.1. Expression: smiling versus neutral

Performance varied strongly for different facial expressions. Not surprisingly, the odds of a subject being correctly verified increased when the expressions in the target and query images matched: verification was easier for smiling queries paired with smiling targets, and for neutral queries paired with neutral targets.

Surprisingly, there was an asymmetry of effect when expressions differed. Performance for smiling query images paired with neutral target images was inferior (for Algorithms B and C) to that for neutral query images paired with smiling target images. The magnitude of this effect and the degree of asymmetry varied substantially between algorithms.

Estimated odds ratios for verification for each pairing of facial expressions, relative to neutral–neutral pairings are given in Table 2. The associated estimated probabilities of verification are also shown. Expressions were either neutral (N) or smiling (S) and are listed with the query image first. Thus, NS indicates a neutral expression in the query image and a smiling expression in the target image.

Table 2
Estimated odds ratio for verification for each expression pairing, relative to NN, for each algorithm. Expressions were either neutral (N) or smiling (S) and are listed with the query image first. Also shown are the estimated probabilities of verification.

Expressions	Odds ratios			p (verification)		
	A	B	C	A	B	C
NN	1.00	1.00	1.00	0.065	0.644	0.850
NS	0.31	0.84	0.64	0.021	0.602	0.783
SN	0.48	0.74	0.45	0.033	0.571	0.718
SS	1.17	1.44	1.18	0.075	0.723	0.870

These results challenge conventional wisdom. It is common for operators collecting facial biometrics, namely images, to instruct people not to smile. For example, the Canadian policy for passport photos is that no smiles are allowed. Specifically "Applicant must show a neutral facial expression (no smiling, mouth closed) and look straight at the camera." [3]

In these instructions is a presumption that a stored image with a neutral expression will be most universally recognizable. Our findings directly contradict this presumption and are consistent with earlier work by Yacoob and Davis [21]. All three algorithms have a higher estimated probability of verification when the subject is smiling in the query and target images. If we assume for the moment that we can not control the expression of the subject in the query image and only one image may be enrolled, then a smiling target image is better than one with a neutral expression. Further, as observed in FRGC Experiment 1 versus Experiment 2 [11] and also by Faltemier et al.[22], enrolling multiple target images per subject (e.g., one neutral and one smiling) is better still.

5.2.2. Elapsed time

One of the most commonly reported covariates contributing to recognition difficulty is the elapsed time between when query and target images are acquired [23–25]. This elapsed time is one of the covariates in our study. The total elapsed time for image pairs in our dataset is relatively modest, and skewed toward smaller values. Specifically, the minimum, quartiles, and maximum values are 0, 4, 28, 52, 84 and 238 days, respectively.

Nevertheless, there is sufficient variety in the data to test for an effect, and the odds ratios for verification associated with a 30-day increase in elapsed time are 0.85, 0.99, and 0.89, respectively, for A, B, and C. Thus, Algorithm B is essentially unaffected by elapsed time (up to eight months), whereas A and C suffer substantially with increasing elapsed time. Fig. 3 shows estimated verification probabilities as a function of elapsed time for each algorithm.

It is surprising that Algorithm B is essentially unaffected by elapsed time. Obviously this may not hold for longer time periods, but insensitivity to elapsed time even over the modest range of 8 months is noteworthy.

5.2.3. Image resolution: pixels between eyes

The pixel coordinates for the center of the left and right eyes were collected by hand at the University of Notre Dame. The distance between the eyes, measured in pixels, indicates the resolution at which the face has been captured. There is considerable variation in this variable among the imagery. For target images typical distances between the eyes ranged between 220 and

[2] Follow the link for this paper at: http://www.cs.colostate.edu/~ross/research.

[3] From http:www.ppt.gc.ca/cdn/photos.aspx?lang=eng.

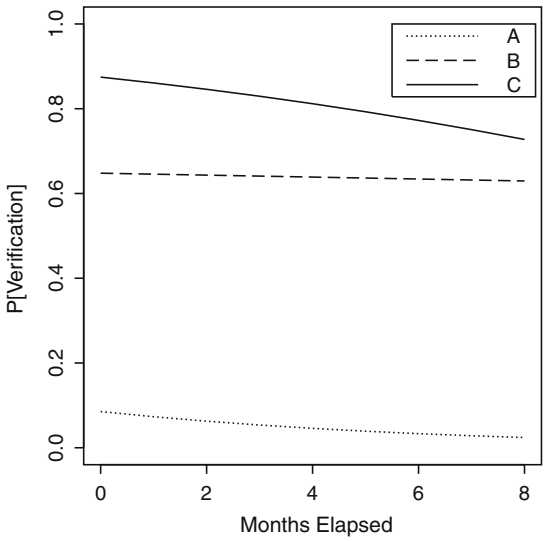

Fig. 3. Elapsed time and estimated verification probabilities for each algorithm.

320 pixels. The range for query images was 120–180 pixels.

Our main analysis fit model terms for the query image eye distance and for the ratio of eye distances, query relative to target. For our data set, the query-target eye distance ratio ranges from about 0.45 to 0.75; the query images are always smaller than the target images due to the uncontrolled query imaging protocol. The ratio has a strong association with performance variations and it interacts strongly with algorithm. The odds ratio for verification associated with an increase of one standard deviation in the query-target eye distance ratio are 0.66, 1.16, and 1.24, respectively, for Algorithms A, B and C. Thus, the performances of Algorithms B and C improve with increasing relative size of the query image, whereas the performance for Algorithm A degrades markedly.

It is important to note that for GLMMs the interpretation of the query/target eye distance ratio is conditioned on holding all other predictors fixed, including the query image eye distance itself. This complicates interpretation. To better understand how performance varies with query and target image sizes, we also fit a secondary model replacing the terms in query eye distance and eye distance ratio. The replacement terms represented a cubic polynomial in query and target eye distances. This polynomial included all terms having total exponent of 3 or less (i.e., terms like Q^2 and Q^2T^1 were allowed but not Q^3T^3).

Contour plots of estimated verification probability for this model are shown in Fig. 4. The ellipses indicate the region of covariates space where 95% of our samples fall. The contours are labeled with the predicted verification rate, i.e., 70 indicates a predicted probability of correct verification of 0.70. We found (in both models) a significant three-way interaction between FAR, query eye distance, and algorithm (see discussion of FAR later); therefore these contour plots are made specifically for the baseline FAR of 0.001.

In Fig. 4, the effect of increasing query eye distance is negative for Algorithm A but positive for B and C. Concentrating on the latter algorithms, we notice that the contours are generally diagonal at the same angle, suggesting that the ratio of image sizes is a key variable: performance is roughly constant for a fixed ratio, regardless of

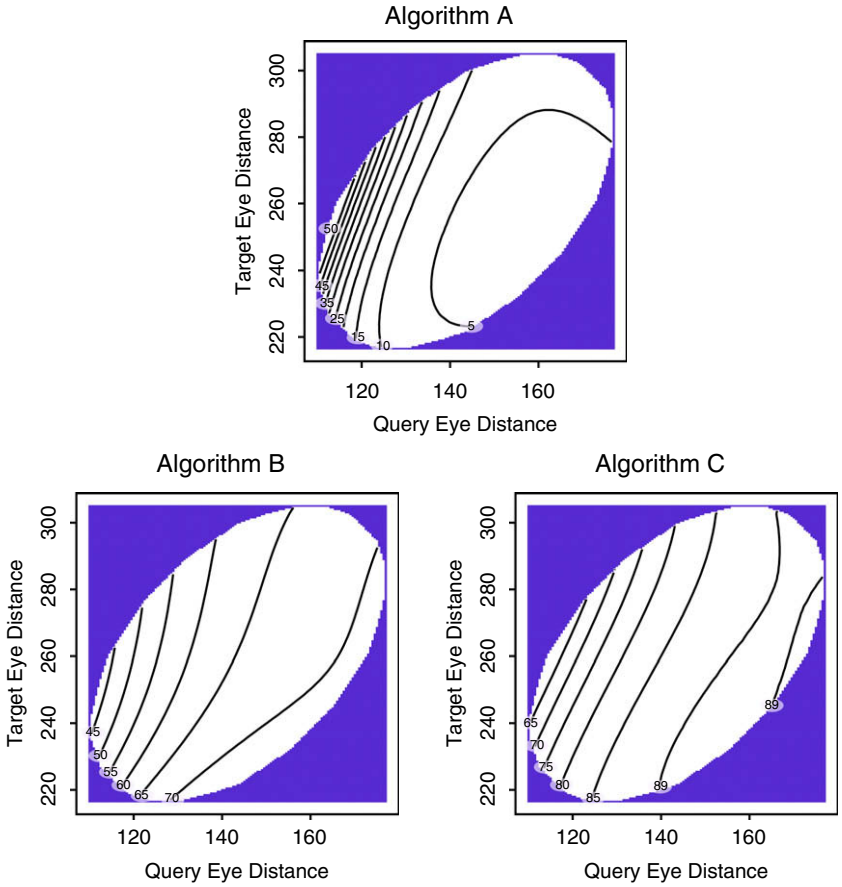

Fig. 4. Contours of estimated verification probabilities for each algorithm, against query and target eye distances.

whether this ratio is achieved with large images or smaller ones. Finally, comparing the central and right panels of Fig. 4, the performance benefit of increasing query image size vanishes for Algorithm C above about 140 pixels between the eyes in the query image, whereas the benefit continues (albeit more slowly) for Algorithm B.

5.2.4. Face tilt difference

The pixel coordinates for the left and right eyes also indicate the extent to which the face is tilted side to side. The tilt of most faces lies in the range −7 to 0° for target images and −6 to 2° for query images. A slight tendency for the right eye to be higher in the image than the left eye is evident; the median tilt is 3 and 2° for the target and query images, respectively.

It is reasonable to ask whether a face rotated in a query image relative to the target image might complicate recognition. To test for this effect, our study considered the absolute in-plane rotational difference, i.e., angle, between line segments connecting the eyes in the query and the target image. A zero on this covariate tells us the face tilt is the same in the target and query image. Departure from zero tells us the face is rotated between the query and target images.

We found the effect of tilt difference to be significant but statistically indistinguishable among the three algorithms. Increasing tilt difference degrades performance. The odds ratio for verification associated with increasing the tilt difference by 1 standard deviation (about 3°) is 0.89. It is worth drawing special attention to the fact that this is one of the few effects that did not involve any algorithm interaction, and thus tilt difference seemed to affect all algorithms in the same fashion and to the same extent.

5.2.5. Image focus

The extent to which an image is, or is not, in focus also influenced verification performance. Unfortunately, we possess no perfect knowledge of focus and must infer it after the fact from the imagery itself. While imperfect, estimates of focus can be derived using the Tenengrad function measure as defined in "Active Computer Vision by Cooperative Focus and Stereo" by Eric Krotkov [26]. This approach essentially convolves the image with an edge detector (Sobel) and then measures edge density. Focus is measured by the sum of the pixel gradient magnitudes.

This measure of focus is restricted to only pixels lying on the face. The face region is defined by the eye coordinates and a canonical oval template. A choice arose in designing this study whether to measure focus over pixels in the original images, or in images down-sampled to a standard 130×150 size. Both options were investigated, and it was found that focus computed over the down-sampled images was a more useful predictor of verification performance. Thus, our focus variables were computed over down-sampled imagery.

In our main analysis, we fit model terms for the difference in focus (query minus target) and the mean focus (over query and target). We found significant effects for both variables and for their interactions with algorithm. To better understand the nature of the relationships, we fit a secondary model with cubic polynomial terms in query and target focus analogous to the case described above for eye distance. The estimated verification probabilities from this model are shown with contour plots in Fig. 5. Overall, there are strong effects for these focus variables, and a strong interaction with algorithm. There is also a strong asymmetry in the effect of focus. Blurry query with sharp target is not the same as sharp query with blurry target.

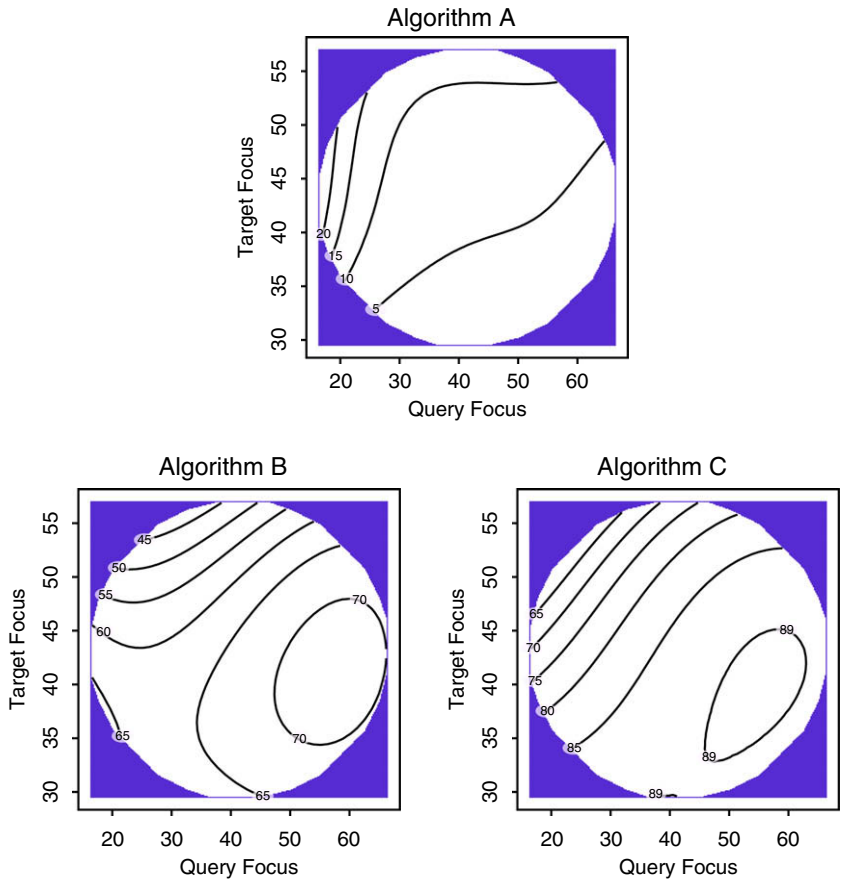

Fig. 5. Contours of estimated verification probabilities for each algorithm, against query and target focus.

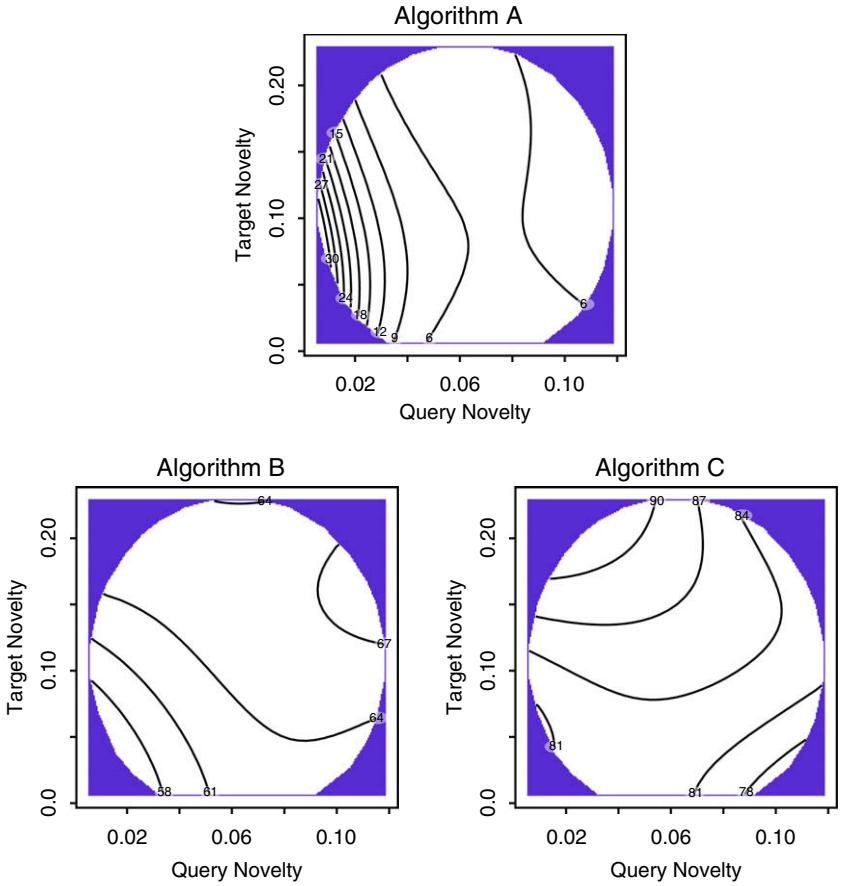

Fig. 6. Contours of estimated verification probabilities for each algorithm, against query and target novelty.

For Algorithm A, performance was best when the target was much sharper than the query. For Algorithm B, performance was best when the query was somewhat sharper than the target. There was an optimal focus level for the target, which is in the middle of the possible range. In other words, for fixed target focus, B preferred the query as sharp as possible, but for fixed query focus, B preferred target focus around 40. The conclusion is surprisingly similar for Algorithm C, except that for blurry queries lower target focus was preferred.

Ignoring Algorithm A, our findings suggest an important quality measure: target images should aim to have a focus score around 40, not substantially higher or lower.[4] Query images should have the highest possible focus score, regardless of target focus.

5.2.6. Novelty

Some face images are more unusual or novel than others. One-way of quantifying this idea is to build a face-space using standard PCA techniques [27,28], and then measure the novelty of new images relative to the standardized representation. We have done this by first selecting a set of imagery, not part of our study, that represents what might be thought of as normal variation.

In particular, 496 Notre Dame face images of 67 subjects were selected from the Spring 2003 dataset. Only controlled lighting imagery was used, and the set of 67 subjects is disjoint from the subjects included in our covariate analysis. Standard geometric normalization was carried out so that novelty was always measured using 130 by 150 pixel images with cropped faces in standard position. A PCA subspace was then constructed from the Spring 2003 imagery and all images in our study were projected into this subspace. Images which are novel are not well encoded by the PCA subspace, and thus part of their content is lost.

Formally, let V be a vector representing a geometrically normalized face image. Further rescale V such that it is of unit length: $|V| = 1$. Now let W be V after projection into the PCA subspace. Novelty is

$$\text{novelty} = 1 - |W|. \tag{3}$$

If the PCA space perfectly encodes the new image V, then novelty is zero. Otherwise, novelty is a value between zero and one indicating the degree to which V is unlike our original set of images. For the query images, the minimum, quartiles and maximum novelty values are 0.0031, 0.047, 0.061, 0.075 and 0.222, respectively. For the target images, these values are 0.0026, 0.068, 0.106, 0.144 and 0.278.

In our main analysis we fit model terms for the difference (query minus target) and mean (over query and target) novelty scores. We found significant effects for both variables and their interactions with algorithm. To better understand the nature of our findings, we fit a secondary model with cubic polynomial terms for query and target novelty, analogous to the case described above for eye distance. The estimated verification probabilities from this model are shown in the contour plots given in Fig. 6.

The contour graphs in Fig. 6 show how query and target image novelty interact to affect performance. For Algorithm A, performance was best when query novelty was low. Target novelty appeared not to affect Algorithm A much.

Algorithm B performed best when query and target novelties were high, and worst when they were low. In contrast, Algorithm

[4] The output of the Sobel mask was used directly, so the average gradient magnitude is over estimated by a factor 2. Hence, 40 corresponds to an average gradient magnitude of 20 grey-levels.

C performed best when target novelty was high but query novelty was low. Notice that novelty difference operated roughly symmetrically on Algorithm B (in that it did not matter much which image was more novel), whereas it operated highly asymmetrically on Algorithm C (with negative values of the difference strongly preferred).

5.2.7. Fragmentation

Another covariate considered in our study is the degree of fragmentation or homogeneity in the face. The region segmentation algorithm developed by Meer [29] was used to segment the face region as defined by the same clipping oval used above for the focus measure. Our model included the signed difference between the number of detected regions in the query and target images. Algorithm-specific effects were found. The odds ratios for verification associated with a 1-standard-deviation increase in this covariate are 1.12, 1.00, and 0.85, respectively, for A, B, and C. Notice that performance of Algorithm C was degraded when the number of regions in the query image increases relative to the number of regions in the target image, while B was indifferent to increased fragmentation.

5.3. Mounting evidence

Some of our findings corroborate previous results from our own studies [6,30,31] and those of other researchers [21,5]. While not entirely new, some of these results are still of considerable practical importance. For example, while not surprising, our results further confirm that comparing images of a person with glasses and without glasses significantly reduces performance. Perhaps more interestingly, the outcome that performance is higher for East Asian than for Caucasian subjects is consistent with other studies and raises important practical questions for algorithm designers concerning how algorithms handle individuals from distinct groups.

5.3.1. False accept rate

Fig. 7 summarizes how varying FAR affected the probability of verification. Generally, verification performance improved significantly for higher FAR settings, and this is entirely to be expected. However, note that the algorithms responded differently to changes in FAR. In other words, there is a clear interaction between FAR and algorithm.

Also note the inclusion in Fig. 7 of three curves for each algorithm. These curves are for different distances between eyes in the query images: high, medium and low. Our model revealed a three-way interaction between FAR, algorithm and query image eye distance. Only for Algorithm C does the relationship between FAR and performance seem unaffected by query image eye distance.

In terms of verification probabilities, Algorithm B was relatively insensitive to eye distance for the lowest FAR settings, and then performed better at higher FAR settings when eyes in the query images were farther apart. Algorithm A was always better when eyes in the query images were closer together, and Algorithm C was always better when the query image eyes were further apart. For Algorithm A, the range of performance variation as eye distance varies diminished somewhat with increasing FAR, but the range was unchanged for Algorithm C.

The right panel of Fig. 7 repeats the graph on the log odds scale. In this panel, model fits are straight lines because our GLMM is linear in log odds. Presence of parallel lines in the log odds plot indicates a lack of an interaction effect. The right panel is thus included here to underscore the basis for concluding that the relationship between FAR and verification performance for Algorithm C is unaffected by query image eye distance.

5.3.2. Age

Most subjects in the dataset are young. The minimum, quartiles, and maximum values for age are 17, 19, 21, 24 and 69. Out of 351 subjects, only 10 are over 40 years of age. Consequently, while results are presented over the full range of ages, the data is sparse for subjects older than 40 and attention is best focused on the results for subjects younger than 40.

That said, this study corroborates results from past studies indicating that performance improves for older subjects [6]. The odds ratios for verification associated with a 1-year increase in age are 1.01, 1.08, and 1.05 for Algorithms A, B, and C, respectively. For a 1-decade increase in age, these become 1.13, 2.06, and 1.60, respectively. Fig. 8 shows the estimated probability of verification for each algorithm as a function of age.

5.3.3. Gender

There are 159 female and 192 male subjects in our study. The estimated probabilities of verification are summarized in Table 3. Men were easier to recognize than women for Algorithms B and C, while women were easier to recognize for Algorithm A.

In our earlier studies of a PCA algorithm, similar to Algorithm A in this study, men were found to be somewhat easier to recognize as characterized by probability of rank one identification [31] and definitively easier to recognize as characterized by probability of correct verification [9]. Thus, women being easier to recognize for Algorithm A is counter to our previous findings. Obviously Algorithm A is on the whole failing much more often than succeed-

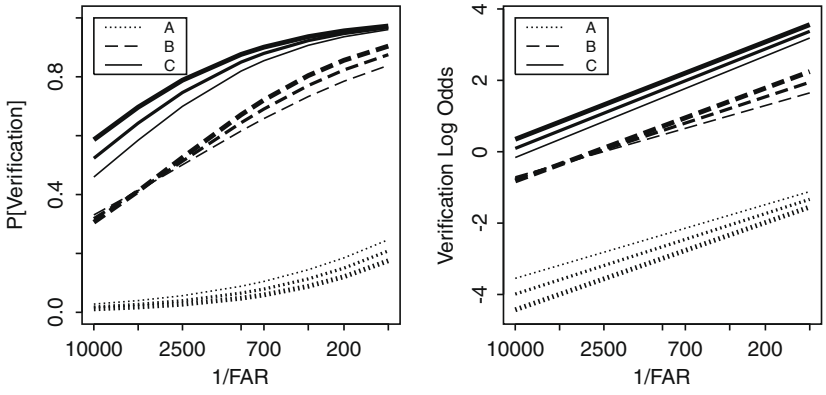

Fig. 7. Estimated verification probability and log odds against FAR. For each algorithm, the heavy, medium, and light line correspond to high, medium, and low values of query eye distance, respectively. The FAR axis labels are inverted to simplify labeling, but FAR increases to the right.

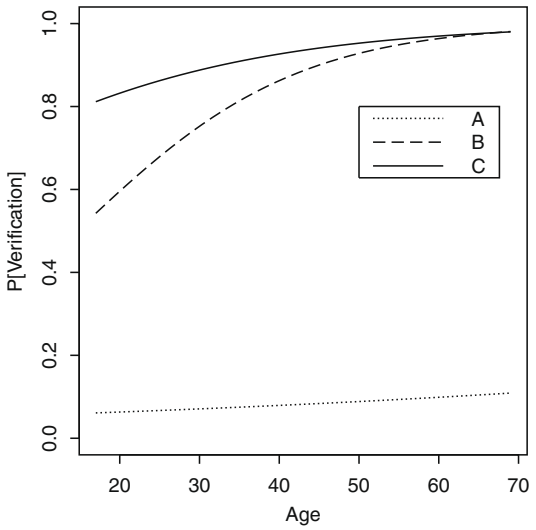

Fig. 8. Age and estimated verification probabilities for each algorithm.

ing for Experiment 4. This is not the case in prior studies, and may be contributing to somewhat atypical behavior on the part of Algorithm A.

More generally, whether men are easier or harder to recognize has been addressed in several prior studies. Notably, marginal analysis for the FaceIt algorithm in 2001 on the AR dataset indicates women being somewhat easier to recognize [5]. For a marginal analysis over a much larger population and multiple algorithms, FRVT 2002 found men typically easier to recognize than women [6]. Discounting Algorithm A based upon its overall poor performance, a trend is emerging suggesting men are easier to recognize than women.

5.3.4. Race

The estimated probabilities of verification for different races are shown in Table 3. Only Caucasian (253) and East Asian (91) subjects are present in large numbers, while Black (4), Hispanic (13), and Unknown (10) are not.

Algorithms responded differently to race, but generally all non-Caucasian races except Black were easier to verify. East Asian seemed easiest, and Black hardest. The performance of Algorithm B improved less for East Asian and Hispanic than that of A and C. However, due to small sample sizes, caution should be used in making inferences about races other than Caucasian and East Asian.

Table 3
Estimated probabilities of verification for departures from baseline associated with gender, race and glasses.

	A	B	C
Gender			
Male	0.065	0.644	0.850
Female	0.101	0.624	0.812
Glasses			
No	0.065	0.644	0.850
Yes	0.030	0.462	0.586
Race			
Caucasian	0.065	0.644	0.850
East Asian	0.119	0.727	0.942
Black	0.025	0.281	0.690
Hispanic	0.151	0.691	0.858
Unknown	0.133	0.436	0.838

This result continues a trend noted in our earlier study using data from the FERET evaluations [31], where we observed algorithms performing better for non-majority races. It is difficult to precisely explain this outcome, but in general it is worth noting that any disadvantage one might presume due to under-representation in algorithm training seems more than compensated for by some factor having to do with a reduced likelihood of confusion for non-majority races. Furthermore, in [7], race effects persisted even after we corrected for under-representation in training and testing.

It is also worth noting that we observed Caucasians being harder to recognize than East Asians in the HCInt portion of FRVT 2002. This was not originally reported as part of FRVT [6] due to a relatively small number of Asians, about 130. However, it is worth mentioning given this new finding for FRGC Experiment 4.

5.3.5. Glasses

Recall that subjects never wore glasses in the target imagery, but subjects wore glasses in 465 of the 2,808 query images. The estimated probabilities of verification with and without glasses in the query image are shown in Table 3. Wearing glasses generally degraded performance, but it affected different algorithms to different extents. In particular, Algorithm C was more sensitive to glasses than were the other two algorithms.

This result highlights something most researchers expected, namely that inconsistent usage of glasses can strongly degrade performance. Our previous study on the FERET data [9] showed a modest tendency for glasses to improve performance, but in that case we were measuring a benefit associated with a subject always wearing the same pair of glasses. Clearly the two situations should not be confused.

5.4. Model calibration and power

A common but often misguided criticism of empirical analyses of performance is that the fitted model is divorced from reality, particularly in the sense that its predictions have some explanatory utility on average but less so for individuals. It is important to examine model fit in light of this concern, and here we present additional results that suggest our model is indeed well calibrated to the observed data and has genuine predictive power.[5]

An elementary check on the power for a model such as ours is to ask how well it predicts individual outcomes. Overall, about 83.5% of model estimates (59,156 of 70,192 for true failures and 53,730 of 64,568 for true successes) would have provided correct classifications of success or failure using an estimated probability threshold of 0.5 for classification.[6] This indicates that as a predictive tool, our model is effectively extracting information from the covariates about verification outcome.

Another valid concern is the behavior of the model over the range of covariates, particularly at the extremes. One way to examine the relationship between what the model predicts and the actual fraction of verification successes to failures is to divide all the observed data into a series of strata based upon the outcome predicted by the model, and then to report the fraction of true verification successes for each stratum. This has been done in Fig. 9. Each row of this table shows a stacked bar chart indicating the numbers of true verification failures and true verification successes. For example, the top bar indicates that about 14,000 observations had an estimated probability of correct verification

[5] The term "prediction" here is used in the sense common to discussions of statistical models, and concerns the observed data. We are not making stronger claims about generalization to other datasets in the manner common in machine learning.

[6] Classification success is even better if one chooses the best possible threshold, knowing that the dataset has 52% failures.

between 0.95 and 1.0. Further, the relative sizes of the black versus brown portions of the bar indicated that the vast majority of these instances were true verification successes.

What we can observe from Fig. 9 is that while estimated verification probabilities range widely—nearly from 0 to 1—this range is not indicative of inappropriate over-fitting. Nearly all the cases having very high estimates of verification probability were true successes, so the model is making accurate predictions. The analogous result holds at the other end of the spectrum. If anything, the model is over-smooth, with fitted verification probabilities not adjusting quite quickly enough to the predictors. This is common in such models.

Another key question in our analysis relates to the explanatory power of our model. Also, what is the relative importance of these covariates compared to between-subject variation that is identifiable but unattributed to our chosen covariates? To answer such questions, we constructed a table comparing results from four models:

(i) the full model,
(ii) a model having only covariate effects with no random effects for subjects,
(iii) a model having only a random effect for subjects with no covariate effects, and
(iv) a constant (null) model which yields estimated verification probability of 0.48 for every observation.

For each observation in the dataset, a predicted verification probability was estimated for each model, along with a confidence interval. If that interval covered 0.5, the model prediction was judged to be inconclusive. If that interval fell entirely on the correct/incorrect side of 0.5, the model was judged to yield a prediction success/failure for that case. Finally, prediction decisions were also forced by examining only the point estimate, ignoring the interval. If the point estimate fell on the correct/incorrect side of 0.5, the model was judged to yield a prediction success/failure for that case. If the point estimate equaled 0.5, model prediction was said to have failed. Note that for the 'forced' case, the constant model (iv) predicts non-verification in every instance; therefore every actual verification in the dataset constitutes a model prediction failure.

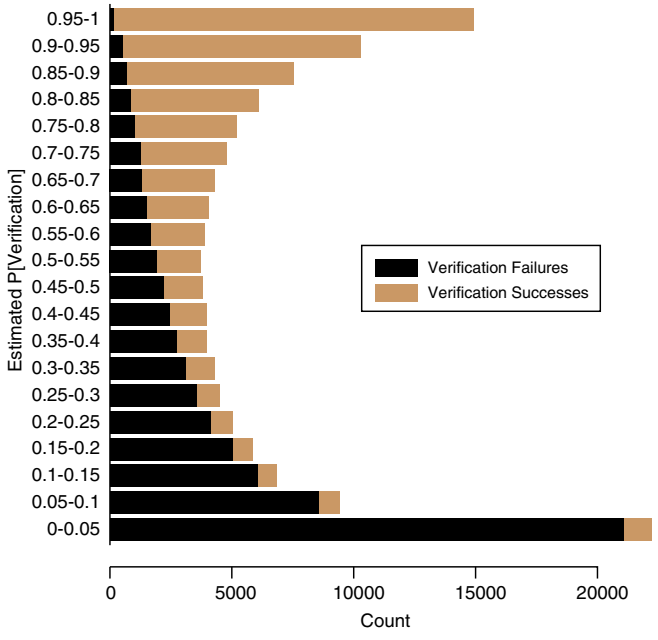

Fig. 9. True verification outcomes and estimated verification probabilities.

Table 4
Percentages of correct model prediction (i.e., explanatory power) for four models described in the text.

Model	Success (%)	Failure (%)	Inconclusive (%)	Forced (%)
(i) Full	79	13	8	84
(ii) Only covariates	73	22	4	76
(iii) Only subjects	58	26	16	67
(iv) Constant	52	48	0	52

Table 4 summarizes the explanatory power of the models by tabulating the prediction successes and failures described above, expressed as percentages of total attempts, for each of the four models. The odds ratio for correctly predicting the verification outcome using our model, compared to model (iv), is 4.85. This shows the substantial explanatory power of our model.

To further interpret these results, begin with the constant model (iv). The predominant (52%) outcome in the dataset is non-verification. Thus, a monkey could correctly predict 52% of cases correctly simply by guessing 'non-verification' in every case. This corresponds to the final row of the table. Now, if we model unspecified between-subject variation, we can improve to 67% correct prediction using the random intercept model (iii). The covariates clearly carry a signal beyond subject-to-subject variation, because when they are added to the random intercept model, correct predictions increase from 67% to 84%. However, there is a non-negligible portion of between-subject variation that is unexplained by the covariates, evidenced by the drop from 84% to 76% when random intercepts are omitted from the full model. These findings support the view that the magnitude and importance of covariate-related performance variations are at least as great as those attributable to unexplained variation between subjects.

6. Conclusions

One of the most important results of this analysis was quantifying the impact of covariates on performance. Fig. 2 shows the spread in estimated verification probability from about 0.4 to 0.9 for algorithms B and C when covariates such as gender, age, glasses are taken into account. This is a very large range of possible outcomes, and it overshadows somewhat the difference between algorithms themselves.

It is also important to note that as often as not, and particularly for covariates derived directly by measuring properties of the imagery, algorithms respond differently. This has important implications for work on image quality metrics. At least for the cases considered here, in particular resolution, focus and novelty, it is wrong to assume a universal link between these image properties and algorithm performance.

The asymmetric nature of the smiling versus neutral expression effect is important to those charged with establishing enrollment protocols. For example, recall that predicted verification rates increase from 0.72 to 0.78 for Algorithm C when smiling instead of neutral faces are enrolled in the target set. We repeat here our conclusion that the best practice is to store both a smiling and a neutral expression image. However, if forced to store only a single image, our results strongly argue for a smile.

Three other findings deserve a comment. First, as in previous studies, younger adults are harder to recognize than older adults. This finding is one of the few to appear consistent in all studies, and it is rapidly gaining stature as an accepted fact. The second finding is that males appear easier to recognize than females. As

discussed above, the evidence from prior studies is not as consistent as for age. However, it is our judgment that the weight of evidence is shifting in favor of the view that males are somewhat more easily recognized. Finally, as in past studies, East Asians are showing up as more easily recognized than are Caucasians in datasets with a majority of Caucasian subjects. It will be interesting as more studies are conducted to see what happens on a majority East Asian dataset with a minority subset of Caucasians.

The results from this paper suggest possible new research directions. One direction is covariate-based fusion of algorithms. In covariate-based fusion, the values of covariates would be estimated from the pair of images being compared and then based on their values, the most effective algorithm would perform the matching. Alternatively, the weighting of an algorithm's response would change based on the estimated covariates. Another research direction could address causation. Our current study reports observed changes due to covariates; however, the analysis does not attempt to explain the cause of the effect. Answering the underlying cause of the effects will assist in designing more robust face recognition algorithms. Also, as systems are fielded in different operating environments, it will assist in predicting performance of operational systems under a wide variety of conditions.

References

[1] W. Zhao, R. Chellappa, A. Rosenfeld, P.J. Phillips, Face recognition: a literature survey, ACM Computer Surveys 35 (2003) 399–458.

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

[3] K. Lee, J. Ho, D. Kriegman, Acquiring linear subspaces for face recognition under variable lighting, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (5) (2005) 684–698.

[4] R. Gross, S. Baker, I. Matthews, T. Kanade, Face recognition across pose and illumination, in: S.Z. Li, A.K. Jain (Eds.), Handbook of Face Recognition, Springer-Verlag, 2004, pp. 193–216.

[5] R. Gross, J. Shi, J.F. Cohn, Quo Vadis Face Recognition? The Current State of the art in Face Recognition, Tech. Rep. TR-01-17, Carnegie Mellon University, June 2001.

[6] P. Phillips, P. Grother, R. Micheals, D. Blackburn, E. Tabassi, J. Bone, FRVT 2002: Evaluation Report, Tech. Rep., Face Recognition Vendor Test, 2002. Available from: <http://www.frvt.org>.

[7] Geof H. Givens, J. Ross Beveridge, Bruce A. Draper, David Bolme, A statistical assessment of subject factors in the PCA recognition of human faces, in: CVPR 2003 Workshop on Statistical Analysis in Computer Vision Workshop, IEEE Computer Society, 2003, (online only).

[8] Geof H. Givens, J. Ross Beveridge, Bruce A. Draper, David Bolme, Using a generalized linear mixed model to study the configuration space of a PCA + LDA human face recognition algorithm, in: Proceedings of Articulated Motion and Deformable Objects, 3rd International Workshop (AMDO 2004), 2004, pp. 1–11.

[9] Geof H. Givens, J. Ross Beveridge, Bruce A. Draper, 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 Workshop: In Conjunction with CVPR, 2005, (electronic only).

[10] S. Mitra, M. Savvides, A. Brockwell, Statistical performance evaluation of biometric authentication systems using random effects models, IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (4) (2007) 517–530.

[11] 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: Conference on Computer Vision and Pattern Recognition, IEEE Computer Society, 2005, pp. 947–954.

[12] P. Phillips, P. Flynn, T. Scruggs, K. Bowyer, W. Worek, Preliminary face recognition grand challenge results, in: Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition, Southampton, UK, 2006.

[13] C. Xie, M. Savvides, B.V. Kumar, Redundant class-dependence feature analysis based on correlation filters using frgc2.0 data, in: Proceedings of IEEE Computer Vision and Pattern Recognition Conference, vol. 2, 2005, pp. 266–273.

[14] C. Liu, Capitalize on dimensionality increasing techniques for improving face recognition grand challenge performance, IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (5) (2006) 725–737.

[15] 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 International Conference on Spoken Language Processing (ICSLP), Syndey, Australia, 1998, (paper 608).

[16] N.E. Breslow, D.G. Clayton, Approximate inference in generalized linear mixed models, Journal of the American Statistical Association 8 (1993) 9–25.

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

[18] R.C. Littell, G.A. Milliken, W.W. Stroup, R. Wolfinger, SAS System for Mixed Models, SAS Publishing, Cary NC, 1996.

[19] SAS Institute, The GLIMMIX Procedure, June 2006. Available from: <http://support.sas.com/rnd/app/papers/glimmix.pdf>.

[20] P. McCullagh, J.A. Nelder, Generalized Linear Models, second ed., Chapman and Hall, New York, NY, 1989.

[21] Y. Yacoob, L. Davis, Smiling faces are better for face recognition, in: International Conference on Face Recognition and Gesture Analysis, 2002, pp. 52–57.

[22] T. Faltemier, K. Bowyer, P.J. Flynn, Using a multi-instance enrollment representation to improve 3D face recognition, in: IEEE Conference on Biometrics: Theory, Applications and Systems, 2007, pp. 1–6.

[23] P. Phillips, H. Moon, S. Rizvi, P. Rauss, The FERET evaluation methodology for face recognition algorithms, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (10) (2000) 1090–1104.

[24] D.M. Blackburn, M. Bone, P.J. Phillips, Facial Recognition Vendor Test 2000: Executive Overview, Tech. Rep., Face Recognition Vendor Test, 2000. Available form: <http://www.frvt.org>.

[25] P. Phillips, P. Grother, R. Micheals, D. Blackburn, E. Tabassi, J. Bone, FRVT 2002: Overview and Summary, Tech. Rep., Face Recognition Vendor Test, 2002. Available from: <http://www.frvt.org>.

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

[27] M. Kirby, L. Sirovich, Application of the Karhunen–Loeve procedure for the characterization of human faces, IEEE Transactions on Pattern Analysis and Machine Intelligence 12 (1) (1990) 103–107.

[28] M.A. Turk, A.P. Pentland, Face recognition using eigenfaces, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 1991, pp. 586–591.

[29] D. Comaniciu, P. Meer, Mean shift: a robust approach toward feature space analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (2002) 603–619.

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

[31] 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 of IEEE Computer Vision and Pattern Recognition, 2004, pp. 381–388.