

How Features of the Human Face Affect Recognition: a Statistical Comparison of Three Face Recognition Algorithms

Geof Givens
Statistics Department
Colorado State University
Fort Collins, CO

J Ross Beveridge and Bruce A. Draper
Computer Science Department
Colorado State University
Fort Collins, CO

November 14, 2003

Abstract

Recognition difficulty is statistically linked to 11 subject covariate factors such as age and gender for three face recognition algorithms: principle components analysis, an interpersonal image difference classifier, and an elastic bunch graph matching algorithm. The covariates assess race, gender, age, glasses use, facial hair, bangs, mouth state, complexion, state of eyes, makeup use, and facial expression. We use two statistical models. First, an ANOVA relates covariates to normalized similarity scores. Second, logistic regression relates subject covariates to probability of rank one recognition. These models have strong explanatory power as measured by R^2 and deviance reduction, while providing complementary and corroborative results. Some factors, like changes to the eye status, affect all algorithms similarly. Other factors, such as race, affect different algorithms differently. Tabular and graphical summaries of results provide a wealth of empirical evidence from which many conclusions may be drawn. Plausible explanations of many results can be motivated from knowledge of the algorithms. Other results are surprising and suggest a need for further study.

1 Introduction

Many algorithms have been proposed for human face recognition [19, 3, 20], spawning a new industry [15]. Scientists working with these systems know that some people are harder to recognize than are others. Surprisingly, however, few studies have been published looking at what attributes make a subject easier or harder to recognize for even a single class of face recognition algorithm.

This paper examines how subject factors affect recognition difficulty for three well known algorithms. The first is a principal components (PCA) algorithm [8]. The second is an interpersonal image difference classifier (IIDC) [11]. The third, an elastic bunch graph matching (EBGM) algo-

rithm, uses localized landmarks and Gabor jets [14]. The PCA algorithm is chosen because it is a de facto standard. The IIDC and EBGM algorithms performed well in the original FERET evaluations [16] and represent qualitatively distinct alternatives to PCA.

Our study uses 2,144 images from the FERET data set [16, 4]: two images for each of 1,072 human subjects. This represents most but not all of the subjects in FERET. Earlier studies included more images and subjects, but doing so imbalances the number of image pairs per subject. It also introduces image pairs taken on different days and such temporal separation is understood to make recognition more difficult.

The covariates in our study measure race, gender, age, glasses use, facial hair, bangs, mouth state, skin complexion, state of eyes, makeup use, and facial expression. These covariates were not collected with the FERET data, so it was necessary for us to reconstruct them from visual inspection of the images.

Two statistical models are used in this study. The first is an ANOVA relating covariates to normalized pairwise image distance scores. Inferences from this model are based on the belief that subjects are more easily recognized when the distance between the two images of a subject is small. Obviously this is a heuristic, since the top ranked choice of a nearest neighbor classifier depends on how other images distribute themselves in the immediate vicinity of the image being tested. Our results below suggest that this heuristic is generally sound, though not perfect.

The second model is a generalized linear model that uses logistic regression to relate subject covariates to probability of rank one recognition. For each subject and each image of that subject, a gallery of 1,072 images is sorted by increasing distance (decreasing similarity) to the probe image. Ideally the first image in the sorted gallery is of the same subject as the probe image. When this is true, a nearest neighbor classifier recognizes this subject “at rank one”. If an algorithm has trouble recognizing a subject, then the

matching image will be found at some inferior position in the sorted gallery, and the rank of this position is called the recognition rank.

The advantage of using logistic regression is that the probability of rank one recognition relates directly to recognition rate, and recognition rate is a nearly universal performance measure for face recognition algorithms. One disadvantage arises because all three of the algorithms studied here recognize most subjects correctly. Thus, only a limited portion of our data is directly informative about rank one recognition failure. In contrast, all subjects contribute similarity scores to the linear model which are (in a rough sense) equally informative.

Both analyses have value. Where the results agree, conclusions are even more convincing; disagreements highlight interesting cases where image distance is not directly linked to rank one recognition performance. An important strength of both models is their multivariate nature. This allows for the interpretation of subject factors to be controlled for other covariates, thereby eliminating the variable confounding or surrogacy that invalidates simpler one-way analyses that examine the effect of a single covariate in isolation.

The major conclusions of this study are summarized in Figures 1 and 2, and Tables 1 and 2. We find that older subjects and non-Caucasian subjects are easier to recognize, regardless of the algorithm. Subjects with their eyes closed are easier to recognize for PCA and IIDC, but harder for EBG. We believe this is the first example of a large, controlled study showing a clear algorithm-covariate interaction. Moreover, this result is plausible: EBG uses eyes as landmarks, and it may not be able to localize them as well when the eyes are closed. Other results confirm common sense: subjects who change across the pair of images, for example by altering their expression or opening or closing their eyes, are harder to recognize than subjects who are more consistent.

2 The Subject Covariates

Below is a list of the covariate factors and their levels. For each factor, one level was designated as a baseline, indicated with an asterisk in the list below. These covariates were assigned by hand by a single person viewing the images. Thus the ratings have a subjective component, but they do not introduce inter-viewer variability. Gender, skin, glasses, and bangs were easy for our viewer to judge. Race, facial hair, expression, mouth and eyes were somewhat harder, although the viewer was still confident in his judgments. Age and makeup were reported to be difficult to estimate. When the evidence for a factor was inconclusive, the default value was selected.

Each individual image was rated. Age, Race, Gender, and Skin ratings were constrained to remain constant within

each same-subject image pair; the other covariate ratings could differ between the two images of a subject. No subjects among the 1,072 we examined changed their use of glasses changed between images.

Age {Young*, and Old}. Old was assigned to subjects judged to be at least 40.

Race {White*, African-American, Asian, Other}. The "Other" category was used for Arab, Indian, Hispanic, mixed race, and any subject that did not fit into the other three categories.

Gender {Male*, Female}. Self-explanatory.

Skin {Clear*, Other}. The Other category included wrinkles, freckles, etc.

Glasses {Yes, No*}. Self-explanatory.

Facial Hair {Yes, No*}. There were many men who had thin beards or were not clean shaven. Any visible facial hair triggered a Yes rating.

Makeup {Yes, No*}. A Yes was only assigned if it was obvious that a subject was wearing makeup. The most obvious feature to look for was the shade of the lips, however the eyes and general appearance also influenced the decision.

Bangs {Yes, No*}. Bangs was set to Yes if the subject's hair was visible in the masked/normalized image. This included hair that came down over the forehead and hair that sometimes covered the sides of the face. In some cases hair was barely visible around the edge of the image; these cases were assigned No.

Expression {Neutral*, Other}. Neutral referred to a natural, relaxed face. The other category were mostly smiles, but included all non-neutral expressions.

Mouth {Closed*, Other}. Closed was typically associated with a relaxed, neutral expression. When subjects had a mostly neutral expression with their mouth open they were assigned Other, as in cases with visible teeth or smiles, indescribable expressions, and closed mouth smiles.

Eyes {Open*, Not Open}. Open eyes were associated with relaxed open eyelids, with the person staring directly into the camera. Not Open states included closed eyelids and eyes that were half open, that looked somewhere other than directly at the camera, or that in some other way did not appear relaxed.

3. Algorithms

The PCA algorithm is based on Turk and Pentland's original algorithm [8], with one twist. The similarity measure is the cosine of the angle between two images after they have been projected into the PCA subspace and whitened. To be

more specific, each dimension of subspace is scaled by the inverse of its sample standard deviation. This is a whitening transformation, since it gives the training data unit sample variance in all directions. This whitened cosine measure is a refinement of one proposed by Moon and Phillips in the original FERET study [12]. When PCA is used with this measure, it is competitive with both IIDC and EBGm on the FERET data [7].

The IIDC is based on an algorithm developed by Moghaddam and Pentland[10]. A detailed description of our implementation may be found in [18]. The algorithm uses PCA to generate a parametric characterization of two spaces. The first is the space of intrapersonal image differences, i.e. differences between same-subject image pairs. The second is the space of interpersonal images, or differences between images of different people. These two classes of image differences are assumed to be Gaussian. Two variants of the algorithm employ different classification rules: a *maximum a posteriori* (MAP) and *maximum likelihood* (ML) classifier. Experiments suggests there is little difference between the two on the FERET data, so ML is used here.

The EBGm algorithm is based on an approach from University of Southern California[14]. The algorithm locates 52 landmarks per image, including the eyes, nose, and mouth. These landmarks are located based on templates extracted by hand from model images. Our implementation uses 70 such templates for each landmark, drawn from 70 hand chosen images. Gabor jets are extracted at each landmark and used to form face graphs; the algorithm measures similarity between face graphs by comparing the corresponding Gabor jets. A detailed description of our implementation may be found in [1].

These three algorithms are qualitatively different, making comparisons between them interesting. PCA is arguably the simplest, being a nearest neighbor classifier in a subspace explicitly based on the variance in the training data. IIDC is a parametric algorithm that operates on image differences rather than images. Finally, EBGm is a localized method using distinct facial landmarks. It therefore emphasizes some face regions over others by design.

3.1 Image Normalization

Our work on FERET has made us acutely aware of how important image normalization is to recognition performance. The imagery in this study has been subjected to the same preprocessing as used by NIST in the original FERET study. First, faces are translated to the center of the image based on hand-selected eye coordinates. Next the image is cropped using an elliptical mask such that only the face from forehead to chin and cheek to cheek is visible. Histogram equalization is applied to the unmasked region of the image, and

finally pixel values are scaled to have a mean of zero and a standard deviation of one.

3.2 Training

EBGM has no training phase, beyond the person who hand picks the landmark templates. Both PCA and IIDC use training data to automatically construct subspaces for recognition. Ideally, we would train and test these algorithms on different image sets. Unfortunately, only two images are available for most FERET subjects, and there are insufficient data to support disjoint training and test sets. Consequently, PCA and IIDC are trained on the complete set of 2,144 images. This is the best choice given the limits of the FERET data, in that it removes concerns that observed effects might be due to training inequities.

3.3 Data Coding

All three algorithms are run on 2,144 images, generating three 2,144 by 2,144 similarity matrices. The images are partitioned into two sets: the first image of each subject and the second image of each subject. Which image is first is arbitrary, but the partition plays an important role in computing recognition rank. For each subject, the first image is treated as a probe image, and the second images is treated as part of the gallery¹. The gallery is sorted by decreasing similarity relative to the probe image, and the position, i.e. rank, of the second image of the probe subject is recorded. This is the recognition rank for the probe image. This process is repeated reversing the role of probe and gallery images. Thus, each subject generates two observations in our dataset: one where the first image is the probe, and one where the second image is the probe. The resulting pairs of recognition ranks are used by the generalized linear model.

There are two response variables for each observation: recognition rank and normalized similarity score. For the algorithms studied here, the similarity scores are symmetric, and therefore identical for a given subject. However, the normalization of similarity scores is probe specific, creating asymmetric normalized similarity scores.

To normalize similarity scores, all 1,072 similarity scores between the probe image and gallery images (only one of which matches the probe) are pooled. Normalized similarity scores are calculated by subtracting the sample mean similarity score, and dividing the result by the sample standard deviation. This operation gives the 1,072 scores for a probe image a sample mean of zero and a sample variance of one. Since all three algorithms recognize most subjects successfully at rank one, normalized similarity scores

¹The use of the terms probe, probe set and gallery here are the same as in the original FERET evaluation.

between a probe and the matching image of the same subject usually exceed 3.

Normalization is used to place similarity scores from different algorithms on a common footing. The process is imperfect, however. The mean normalized scores (across all probes) between a probe and its match were 7.39, 3.24, and 4.50, for PCA, IIDC, and EBGm, respectively, and the standard deviations were 2.62, 1.11, and 1.33. We easily equilibrate the absolute levels of normalized similarity scores across algorithms in our modeling framework. Interpretation of our models is complicated (but not invalidated) by the differing standard deviations. Overall, the normalization is useful and does, to a first order, allow scores to be compared.

Each response is associated with a probe image and a gallery image, and these images may differ with respect to the facial hair, makeup, bangs, expression, mouth, and eyes factors. We coded three predictors for each factor, indicating that (i) both images have the baseline level, (ii) both images have the non-baseline level, or (iii) one baseline image and one non-baseline image is in the pair. Coding (i) is treated as the baseline level of each predictor.

4 Analysis and Results

4.1 Linear and Generalized Linear Models

Most readers will already be familiar with the linear modeling framework; ANOVA is reviewed in [13]. Here we will introduce generalized linear models [9]. Let Y be a random variable representing a response, i.e. a quantity measured to evaluate the performance of a single attempt by a single algorithm faced with a single recognition task.

Let \mathbf{X} denote a vector of independent variables with which we hope to predict the response. Here these are the subject covariates. The covariate factors are categorical predictors and they contribute vectors of binary indicator variables to \mathbf{X} in the usual ANOVA fashion [13].

An experiment generally consists of n trials, resulting in a dataset of observations $\{(y_1, \mathbf{x}_1), \dots, (y_n, \mathbf{x}_n)\}$. We write the p components of the the i th predictor vector as $\mathbf{X}_i = (\mathbf{X}_{i,1}, \dots, \mathbf{X}_{i,p})$ and use upper case to denote random variables and lower case to denote observed values.

A generalized linear model consists of three parts: a link function, a linear predictor, and a distributional model. The link function, g , is a possibly nonlinear, monotonic real-valued function of $\mu_{Y_i|\mathbf{X}_i}$, the conditional mean of the response given the predictors for the i th subject. It links the conditional mean to the predictors according to

$$g(\mu_{Y_i|\mathbf{X}_i}) = \mathbf{x}_i\boldsymbol{\beta} = \beta_0 + \beta_1x_{i,1} + \dots + \beta_px_{i,p} \quad (1)$$

if there are p predictor variables in \mathbf{X} . The right side of (1) is called the linear predictor. The vector of parameters

$\boldsymbol{\beta}$ plays a role analogous to ordinary linear regression parameters: each β_j describes the magnitude and direction of relationship between the $g(\mu_{Y_i|\mathbf{X}_i})$ and the j th predictor variable.

The conditional distribution of Y_i given \mathbf{X}_i is assumed to be $Y_i | \mathbf{X}_i \sim f(y; \mu_{Y_i|\mathbf{X}_i})$. The Y_i are conditionally independent.

The simplest generalized linear model takes $g(z) = z$ and $f(y; \mu_{Y_i|\mathbf{X}_i}) = \mathbf{N}(y; \mu_{Y_i|\mathbf{X}_i}, \sigma^2)$, where $\mathbf{N}(a, b^2)$ is the normal density with mean a and variance b^2 . In this case the generalized linear model reduces to the ordinary multiple linear regression model, or ANOVA, for regressing Y on \mathbf{X} . The linear model here uses normalized image distance for Y . Distance is simply the negation of the normalized similarity score

Another form of generalized linear model arises from assuming that Y is a binary random variable. The well-known logistic regression model [6] is established by using $f(y; \mu_{Y_i|\mathbf{X}_i}) = \text{Bern}(y; \mu_{Y_i|\mathbf{X}_i})$, where $\text{Bern}(y; \pi) = \pi^y(1 - \pi)^{1-y}$ (with $0 \leq \pi \leq 1$) is the Bernoulli distribution². This assumption is paired with the canonical link $g(z) = \text{logit}(z) = \log(z/(1 - z))$.

We use this form of generalized linear model to estimate the probability of rank one recognition as predicted by the subject covariates \mathbf{X} . In this analysis, $Y_i = 1$ for the i th probe if and only if that probe is recognized at rank one. Otherwise $Y_i = 0$. The probability of rank one recognition is essentially synonymous with the expected recognition rate.

4.2 Results

The ANOVA results are summarized in Figure 1. For each algorithm, the bars indicate the change in normalized distance associated with the indicated predictor level, relative to the normalized distance associated with that algorithm when all predictors are at baseline levels. Thus, the figure compares algorithm-specific effects of each predictor, controlling for all other predictors. Bars whose lengths are statistically significantly greater than zero are indicated with a diamond.

To illustrate how one interprets an effect in Figure 1, consider the fourth predictor down: Eyes Not Open. This indicates the eyes factor being Not Open for both images. Relative to the baseline of having eyes open in both images, EBGm finds subjects with eyes Not Open in both images harder to recognize, but not significantly so. The linear model predicts an increase in normalized distance between a pair of images of the same subject of about 0.32 (= 1.98 - 1.66 from Table 1) for EBGm when eyes are

²There is another simple way to write this model that employs the Binomial distribution; see the references.

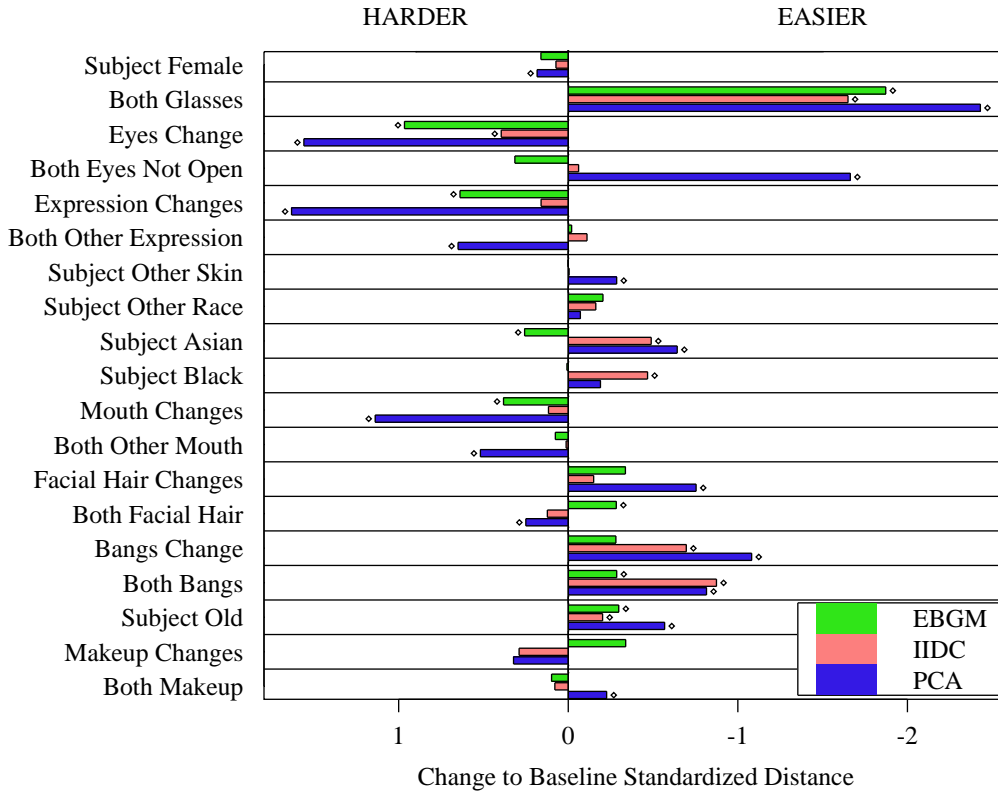


Figure 1: Summary of effects for the linear model.

Not Open. In contrast, the PCA algorithm sees a statistically significant drop in normalized distance of about 1.66 when eyes are Not Open in both images. Some effects are more universal: results for the third predictor down in Figure 1 show that all algorithms find recognition significantly harder when the eye state changes from one image to the other.

Table 1 shows the ANOVA results. The model is parameterized with treatment contrasts [2], which means that every parameter shown represents the deviation from the fit for PCA for young white clear-skinned males having all other covariates at baseline in both images. Thus, the p-values here test comparisons against PCA at baseline, whereas the significance diamonds in Figure 1 test within-algorithm effects. For example, consider the effects labeled Asian, IIDC:Asian, and EBGM:Asian. The estimated effect for Asian (-0.64) indicates that PCA distances are significantly decreased ($p < 0.0001$), which confirms the blue bar in Figure 1. The EBGM:Asian effect (0.90) shows that EBGM distances for Asians are significantly increased ($p < 0.0001$) compared to the Asian effect *for PCA*. The IIDC:Asian effect shows that Asian does not significantly increase IIDC distances compared to PCA, although Figure 1 shows that

Asians are significantly easier for IIDC than Whites.

The multiple R^2 for the ANOVA model is 0.657, and the partial R^2 for all the covariate predictors and interactions, given algorithm, is 0.344. In other words, after adjusting for algorithm, subject covariates and their interactions with algorithm account for 34.4% of the remaining variation in normalized distances.

The results for the generalized linear model are summarized in Table 2. An informal, iterative model search was used to arrive at the general linear model presented here. This model was arrived at after fitting the full model with all possible factors and two-way interactions, and then deleting terms based upon the the standard chi-squared (likelihood ratio) test for changes in deviance. In some cases, interactions could be deleted but main effects could not. We did not allow deletion of main effects in the presence of related significant interactions. After sequentially deleting all non-significant terms, deleted terms were individually reinserted in the model to test for significance. Additional cycles of addition/deletion of terms led to the final model given above. The (nonsequential) Δ Deviance entries indicate the change in deviance that would result from deleting these terms from the given model, and the p-value is

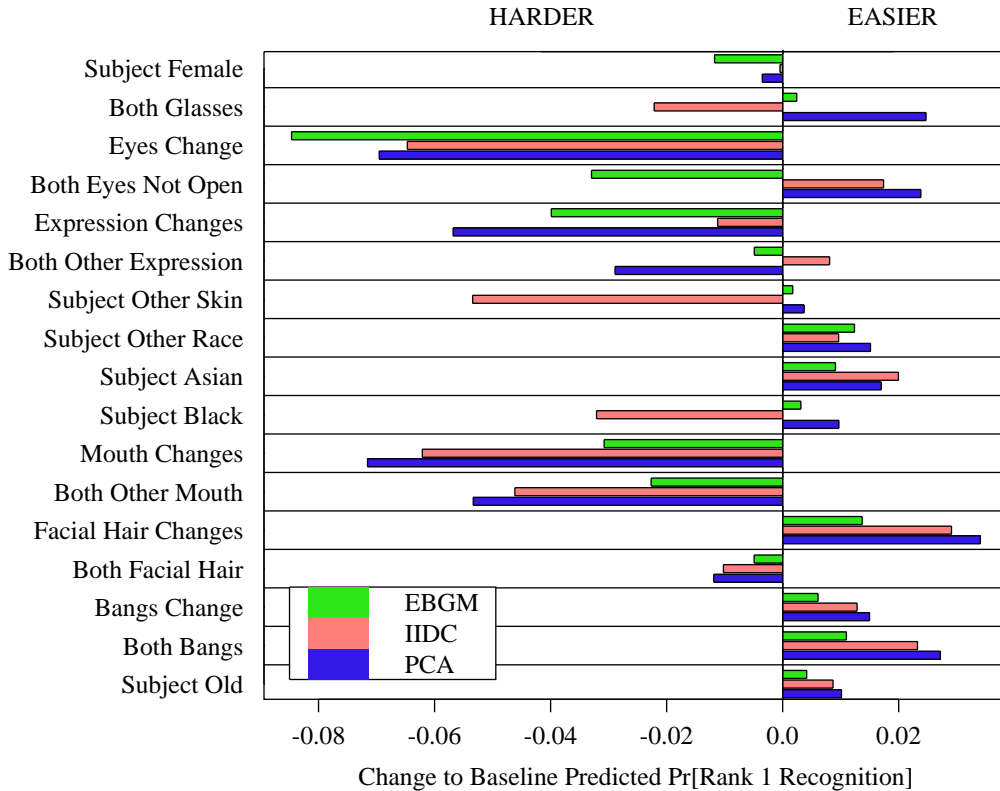


Figure 2: Summary of effects the generalized linear model.

obtained from the corresponding chi-squared test. All interactions and main effects not included in the table above would have non-significant p-values if they were tested for addition to the model given above.

The effects fitted in the generalized linear model are shown in Figure 2, which uses the same organizational and inferential structure as Table 1 except that the response is now the estimated probability of rank one recognition and the significance levels of individual bars are not shown. Figure 2 is organized so that predictors with significant algorithm interactions appear at the top and those with no algorithm interaction appear at the bottom.

To illustrate how to interpret Figure 2, consider the second covariate down: Glasses. For PCA, the probability of recognizing a subject with eyes Not Open in both images, relative to the baseline of a subject with eyes Open in both images, is about 0.02 greater. In contrast, there is a drop in expected probability of rank one recognition of about 0.03 for IIDC. So between PCA and IIDC, the difference is about 0.05, meaning that PCA will recognize 5 more subjects out of 100 correctly at rank one when both algorithms are faced with an Eyes Not Open image pair instead of Eyes Open.

There are only 5 (out of 51 possible) cases where the

linear and generalized linear models indicate significant and opposite effects. These cases suggest something interesting is going on relative to how images are locally distributed, in that normalized similarity scores and rank one recognition probability are not positively correlated.

Both the linear and generalized linear model indicate many significant covariate effects and interactions between algorithms and covariates. While there is a wealth of interesting results, let us draw special attention to several:

Age: Older subjects are consistently easier to recognize.

Gender The linear model indicates women are harder to recognize than men only for PCA ($p=0.0324$). The other two algorithms have no significant dependence upon gender. For the generalized linear model, the story is more complex. If either the algorithm gender interaction or the main gender effect is removed separately, neither appears significant. However, if both are taken together, both become significant and EBGM appears to have more difficulty recognizing women.

Gender as a covariate is of particular interest because others have studied it using a simple data partition approach [17, 15]. In other words, they have divided their test data into two sets, men and women, and noted higher

Table 1: ANOVA results for the linear model. ‘B’=‘both images’, ‘O’=‘Other’, ‘Ch’=‘changes from one image to the other’, and ‘.’ indicates an interaction.

Predictor	Est.	S.E.	t	p
Intercept	-8.44	0.08	-107.76	< 0.0001
IIDC	5.48	0.11	49.46	< 0.0001
EBGM	3.54	0.11	31.98	< 0.0001
Old	-0.57	0.08	-7.09	< 0.0001
Female	0.18	0.09	2.14	0.0324
Afr.-American	-0.19	0.11	-1.76	0.0790
Asian	-0.64	0.10	-6.43	< 0.0001
O Race	-0.07	0.12	-0.59	0.5534
O Skin	-0.29	0.09	-3.08	0.0021
B Bangs	-0.82	0.08	-9.74	< 0.0001
Bangs Ch	-1.08	0.19	-5.63	< 0.0001
B O Expression	0.65	0.15	4.39	< 0.0001
Expression Ch	1.63	0.08	19.94	< 0.0001
B Eyes Not Open	-1.66	0.32	-5.22	< 0.0001
Eyes Ch	1.56	0.11	13.79	< 0.0001
B Facial Hair	0.25	0.10	2.40	0.0164
Facial Hair Ch	-0.75	0.32	-2.34	0.0191
B Glasses	-2.43	0.13	-18.14	< 0.0001
B Makeup	-0.23	0.11	-2.02	0.0439
Makeup Ch	0.32	0.26	1.23	0.2179
B O Mouth	0.52	0.12	4.20	< 0.0001
Mouth Ch	1.14	0.08	13.69	< 0.0001
IIDC : Old	0.37	0.11	3.22	0.0013
EBGM : Old	0.27	0.11	2.39	0.0171
IIDC : Female	-0.11	0.12	-0.92	0.3602
EBGM : Female	-0.02	0.12	-0.19	0.8526
IIDC : Afr.-Amer.	-0.28	0.15	-1.82	0.0693
EBGM : Afr. Amer.	0.20	0.15	1.29	0.1956
IIDC : Asian	0.15	0.14	1.09	0.2778
EBGM : Asian	0.90	0.14	6.36	< 0.0001
IIDC : O Race	-0.09	0.17	-0.53	0.5930
EBGM : O Race	-0.13	0.17	-0.78	0.4339
IIDC : O Skin	0.28	0.13	2.14	0.0327
EBGM : O Skin	0.29	0.13	2.19	0.0286
IIDC : B Bangs	-0.06	0.12	-0.50	0.6201
EBGM : B Bangs	0.53	0.12	4.47	< 0.0001
IIDC : Bangs Ch	0.39	0.27	1.42	0.1557
EBGM : Bangs Ch	0.80	0.27	2.95	0.0032
IIDC : B O Expr.	-0.76	0.21	-3.64	0.0003
EBGM : B O Expr.	-0.67	0.21	-3.20	0.0014
IIDC : Expr. Ch	-1.47	0.12	-12.72	< 0.0001
EBGM : Expr. Ch	-0.99	0.12	-8.58	< 0.0001
IIDC : B O Eyes	1.60	0.45	3.55	0.0004
EBGM : B O Eyes	1.98	0.45	4.39	< 0.0001
IIDC : Eyes Ch	-1.16	0.16	-7.28	< 0.0001
EBGM : Eyes Ch	-0.59	0.16	-3.71	0.0002
IIDC : B Fac. Hair	-0.13	0.15	-0.86	0.3889
EBGM : B Fac. Hair	-0.53	0.15	-3.62	0.0003
IIDC : Fac. Hair Ch	0.60	0.45	1.33	0.1841
EBGM : Fac. Hair Ch	0.42	0.45	0.92	0.3593
IIDC : B Glasses	0.78	0.19	4.12	< 0.0001
EBGM : B Glasses	0.56	0.19	2.95	0.0032
IIDC : B Makeup	0.31	0.16	1.92	0.0545
EBGM : B Makeup	0.32	0.16	2.04	0.0409
IIDC : Makeup Ch	-0.03	0.37	-0.09	0.9294
EBGM : Makeup Ch	-0.66	0.37	-1.79	0.0742
IIDC : B O Mouth	-0.51	0.17	-2.90	0.0037
EBGM : B O Mouth	-0.44	0.17	-2.54	0.0113
IIDC : Mouth Ch	-1.02	0.12	-8.69	< 0.0001
EBGM : Mouth Ch	-0.76	0.12	-6.44	< 0.0001

Table 2: Summary of generalized linear model results.

	df	Δ Deviance	p
Intercept	1	<i>Note 1</i>	
Algorithm	2	<i>Note 2</i>	
Age	1	5.73	0.0167
Bangs	2	63.99	< 0.0001
Facial Hair	2	11.12	0.0039
Mouth	2	76.50	< 0.0001
Race & Alg. : Race	9	46.48	< 0.0001
Skin & Alg. : Skin	3	24.00	< 0.0001
Expr. & Alg. : Expr.	6	54.64	< 0.0001
Eyes & Alg. : Eyes	6	131.87	< 0.0001
Glasses & Alg. : Glasses	3	8.15	0.0430
Gender & Alg. : Gender	3	9.55	0.0228

Note 1 The null model deviance is 4,266.9 on 6,425 df. The model using all terms given above has residual deviance of 3,676.9 on 6,386 df—highly significant.

Note 2 The factor indicating algorithm has many significant interactions in this model and is highly significant. In a table organized to show subject covariate effects, an analogous test for algorithm would be distracting.

recognition rates for men. However, these studies do not control for other covariates such as facial hair or makeup.

Glasses: The linear model suggests all algorithms benefit from a subject wearing glasses: presumably the same glasses. However, here we see a rare reversal between the linear and generalized linear model for IIDC, where the probability of rank one recognition drops by about 0.02 for subjects wearing glasses vs. subjects not wearing glasses. Indeed, for the generalized linear model, the only algorithm to benefit from glasses is PCA.

Eyes Not Open: Subjects who have their eyes closed in both images are, according to both models, easier to recognize for PCA and harder for EBGM. A plausible explanation is that PCA has no means of discounting or ignoring strong variations associated with pupils and the whites of the eyes. Consequently, subjects with closed eyes in both images are more easily recognized. Conversely, EBGM is a landmark algorithm that suffers when the eyes cannot be reliably located.

Race: Race has a modest but statistically significant effect for most algorithms, and not always the same effect for each algorithm. There is a trend toward non-white subjects being easier to recognize, but it is not universal across races and algorithms. It is true for all non-white subjects using PCA, and we’ve observed this same effect in prior studies.

Bangs and Facial Hair: There is a clear trend suggesting bangs make recognition easier, and that having bangs or facial hair in one image and not another, makes recognition easier. This is unexpected, given that bangs and facial hair

are sources of interpersonal variation. These results suggest a need for further study.

An obvious hypothesis to discount many of our results is that subjects with covariate factor levels that are under-represented in the dataset will be easier to recognize due to their relative uniqueness in the gallery. In earlier work with these same data, we have specifically tested this hypothesis with a series of carefully designed experiments that balance representation of the factors in question [5]. In every case, the hypothesis that modeled recognition impacts could be explained by data imbalance was soundly refuted.

5. Conclusion

We have presented two studies relating subject and image covariates to the performance of face recognition algorithms. One study uses an ANOVA to model the effects of covariates on the similarity between two images of the same subject. The second study uses a generalized linear model to measure the effects of the same covariates on rank one recognition rates. Both methods control for other covariates, thereby eliminating variable confounding or surrogacy. The generalized linear model has the advantage that it directly predicts recognition rates, but the disadvantage that only a limited portion of the data is directly informative about recognition failure (because of high recognition rates). Roughly speaking, the ANOVA has the advantage that it uses data more efficiently, but the disadvantage that its response variable is imperfectly related to recognition performance. Most of the time the two models agree, increasing our confidence in their conclusions.

Both statistical models suggest that older subjects are easier to recognize than younger subjects, regardless of which face recognition algorithm is used. Subjects who close their eyes are easier to recognize using PCA, but harder to recognize with EBGm. Race also plays a part: there is a general trend toward non-white subjects being easier to recognize than white subjects, but the trend is not universal across algorithms or other races. Occasionally the ANOVA and generalized linear model disagree. For example, the ANOVA suggests that all algorithms benefit from having a subject wear (the same pair of) glasses in both images. According to the generalized linear model, however, only PCA improves under these conditions. Further empirical studies with more subjects, more replication, more algorithms, more precise and varied codings of covariates, and completely independent algorithm training will further elucidate important subject factors that affect recognition.

References

[1] *Omitted for Blind Review.*

- [2] J. M. Chambers and T. J. Hastie (Eds.), *Statistical Models in S*, Chapman and Hall, New York, 1992.
- [3] R. Chellappa, C.L. Wilson, and S. Sirohey. Human and machine recognition of faces: A survey. *Proceedings of the IEEE*, 83(5):705–740, May 1995.
- [4] FERET Database. <http://www.itl.nist.gov/iad/humanid/feret/>. NIST, 2001.
- [5] *Omitted for Blind Review.*
- [6] D. W. Hosmer and S. Lemeshow. *Applied Logistic Regression*. Wiley, New York, NY, 2000.
- [7] *Omitted for Blind Review.*
- [8] M. A. Turk and A. P. Pentland. Face Recognition Using Eigenfaces. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–591, June 1991.
- [9] P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Chapman and Hall, London, 1983. i-xiii+216pp.
- [10] B. Moghaddam, C. Nastar, and A. Pentland. A bayesian similarity measure for direct image matching. *ICPR*, B:350–358, 1996.
- [11] B. Moghaddam and A.P. Pentland. Probabilistic visual learning for object representation. *PAMI*, 19(7):696–710, July 1997.
- [12] H. Moon and J. Phillips. Analysis of pca-based face recognition algorithms. In K. Boyer and J. Phillips, editors, *Empirical Evaluation Techniques in Computer Vision*. IEEE Computer Society Press, 1998.
- [13] J. Neter, W. Wasserman, and M. H. Kutner. *Applied Linear Statistical Models*. Irwin, Boston, 1990. i-xvi+1181pp.
- [14] Kazunori Okada, Johannes Steffens, Thomas Maurer, Hai Hong, Egor Elagin, Hartmut Neven, and Christoph von der Malsburg. The Bochum/USC Face Recognition System And How it Fared in the FERET Phase III test. In H. Wechsler, P. J. Phillips, V. Bruce, F. Fogeman Soulié, and T. S. Huang, editors, *Face Recognition: From Theory to Applications*, pages 186–205. Springer-Verlag, 1998.
- [15] P. Jonathon Phillips, Patrick Grother, Ross J. Micheals, Duane M. Blackburn, Elham Tabassi and Mike Bone. Face recognition vendor test 2002. Technical report, NIST, March 2002.
- [16] P.J. Phillips, H.J. Moon, S.A. Rizvi, and P.J. Rauss. The FERET Evaluation Methodology for Face-Recognition Algorithms. *T-PAMI*, 22(10):1090–1104, October 2000.
- [17] Jeffrey F. Cohn, Ralph Gross, and Jianbo Shi. Quo vadis face recognition?: The current state of the art in face recognition. Technical Report TR-01-17, Carnegie Mellon University, June 2001.
- [18] *Omitted for Blind Review.*

- [19] D. Valentin, H. Abdi, A.J. O'Toole, and G.W. Cottrell. Connectionist models of face processing: A survey. *Pattern Recognition*, 27(9):1209–1230, September 1994.
- [20] W. Zhao, R. Chellappa, A. Rosenfeld, and J. Phillips. Face Recognition: A Literature Survey. Technical Report CS-TR4167R, Univ. of Maryland, 2000. Revised 2002.