A Statistical Assessment of Subject Factors in the PCA Recognition of Human Faces*

Geof Givens
Statistics Department
Colorado State University
Fort Collins, CO

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

April 10, 2003†

Abstract

Some people's faces are easier to recognize than others, but it is not obvious what subject-specific factors make individual faces easy or difficult to recognize. This study considers 11 *factors that might make recognition easy or difficult for* 1,072 *human subjects in the FERET dataset. The specific factors are: race (white, Asian, African-American, or other), gender, age (young or old), glasses (present or absent), facial hair (present or absent), bangs (present or absent), mouth (closed or other), eyes (open or other), complexion (clear or other), makeup (present or absent), and expression (neutral or other). An ANOVA is used to determine the relationship between these subject covariates and the distance between pairs of images of the same subject in a standard Eigenfaces subspace. Some results are not terribly surprising. For example, the distance between pairs of images of the same subject increases for people who change their appearance, e.g., open and close their eyes, open and close their mouth or change expression. Thus changing appearance makes recognition harder. Other findings are surprising. Distance between pairs of images for subjects decreases for people who consistently wear glasses, so wearing glasses makes subjects more recognizable. Pairwise distance also decreases for people who are either Asian or African-American rather than white. A possible shortcoming of our analysis is that minority classifications such as African-Americans and wearers-of-glasses are underrepresented in training. Followup experiments with balanced training addresses this concern and corroborates the original findings. Another possible shortcoming of this analysis is the novel use of pairwise distance between images of a single person as the predictor of recognition difficulty. A separate experiment confirms that larger distances between pairs of subject images implies a larger recognition rank for that same pair of images, thus confirming that the subject is harder to recognize.*

*The work was funded in part by the Defense Advanced Research Projects Agency (DARPA) under contract DABT63-00-1-0007.

†This paper will be presented at the *Statistical Analysis in Computer Vision Workshop* being held in conjunction with CVPR 2003.

1 Introduction

Many algorithms have been proposed for human face recognition [13, 3, 14], spawning a new industry [6]. 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.

This paper presents a study of how factors associated with subjects affect recognition difficulty using a standard principal components analysis (PCA) based nearest neighbor classifier [8]. This study uses 2,144 images from the FERET data set [11, 7]: two images for each of 1,072 human subjects. While this is not all of the potential FERET subjects, it is large number that enables us to look for statistically significant relationships between ease of recognition and 11 subject covariates.

The covariates in our study are race (white, Asian, African-American, or other), gender, age (young or old), glasses (absent or present), facial hair (absent or present), bangs (absent or present), mouth (closed or other), eyes (open or other), complexion (clear or other), makeup (present or absent), and expression (neutral or other). These covariates were not collected at the time the FERET data were collected, and so it was necessary for us to reconstruct these as best we could by visual inspection of the images.

To carry out our analysis, a standard PCA classifier was trained on all 2,144 images and each image was projected into the resulting subspace. Distance between pairs of images of the same subject is used as the response variable. A linear model was then used to estimate the degree to which each covariate influenced distance.

Using the pairwise distance between images of the same

subject as the response variable has some advantages over other possible choices such as recognition rate or recognition rank. Recognition rate is a function over a set of images, and therefore cannot be tied clearly to any one subject. Recognition rank for a given subject is more appropriate, but it depends on the gallery of images being matched against, not just on the subject. There are also distributional properties of recognition rank that make it less suitable than pairwise distance for regression analysis. Recognition rank does play an important backup role, however, providing a way to test our implicit hypothesis that pairwise distance predicts ease of recognition.

The major conclusions of this study are summarized in Figure 1. It suggests that older subjects are easier to recognize than young subjects, that subjects who consistently wear glasses are easier to recognize than subjects without glasses, and that subjects whose eyes are always closed are easier to recognize than subjects whose eyes are always open. Not surprisingly, subjects who change across the pair of images, for example by changing their expression or opening or closing their eyes, are harder to recognize than subjects who are more consistent. Perhaps more surprisingly, white subjects are harder to recognize than Asian, African-American or other subjects, even when the system is trained with racially balanced data sets.

2 The Subject Covariates

When the FERET images were collected, little or no data were recorded about the subjects themselves. Therefore, it has been necessary for us to estimate such data after the fact from the images themselves. Such post-hoc estimations are bound to be imperfect, but it is far more interesting to proceed with imperfect data than to do nothing. For this study, eleven covariates were selected for examination: age, race, gender, expression, skin appearance, glasses, facial hair, makeup, bangs, mouth and eyes. Each is further described below. All covariate values were estimated by a single viewer, so while there is clearly some degree of subjectivity in assigning values to covariates, we have at least avoided introducing further inconsistency by changing viewers.

Below is a list of the covariates and their values. Each item is a covariate and next to each is a list of the possible values. For each covariate a default value was designated. If the viewer found the evidence for one choice of covariate value versus another inconclusive, then the default value was selected. For example, if the viewer was uncertain about a subject's expression, the expression was assigned the default value neutral.

In some cases, our initial set of covariate values were consolidated into a smaller more manageable set. So, for example, initially age was divided into teens, twenties, thirties, etc. However, for all our analysis age was consolidated into only two categories, young versus old.

Age {Teen, 20, 30, 40, 50, 60+}. Consolidated to {Young [teen-30s] and Old [40s-60+]}. Default = Young. This covariate was one of the most difficult to judge.

Race {White, African-American, Asian, Other}. Default = White. This covariate was easier to judge than was age. If the image looked African-American or Asian, the corresponding category was selected. The "Other" category was used for Arab, Indian, Hispanic, mixed race, and any other apparent race that did not fit into the other three categories.

Gender {Male, Female}. Default = Male. This factor was easy to judge and is probably highly accurate.

Skin {Clear, Wrinkled, Freckled, Both, Other}. Consolidated to {Clear, Other}. Default = Clear. Skin was relatively easy to judge. Wrinkled was obvious. Freckled was more difficult to judge. If there was some doubt about this covariate the default value of clear was kept. For analysis, any image that was not rated as Clear was called Other.

Glasses {Yes, No}. Default = No. The category was easy to judge and should be accurate.

Facial_Hair {Yes, No}. Default = No. In many cases this category was easy to judge, for example, many of the men had mustaches or beards. However, there were a lot of men that had thin beards or were not clean shaven. In these cases, if there appeared to be hair visible then it was counted as facial hair. Otherwise, it was not.

Makeup {Yes, No}. Default = No. Like facial hair, this was also very difficult to judge. The general rule that was used for makeup was to only assign a Yes if it was obvious that a woman (or man) 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}. Default = 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 there was hair hardly visible around the edge of the image; these cases were assigned No.

Expression {Neutral, Other}. Default = Neutral. Neutral referred to a natural relaxed face. The other expressions were mostly smiles, but included any other distortion of the face.

Mouth {Open, Closed, Teeth, Other}. Consolidated to {Closed, Other}. Default = Closed. Closed was typically associated with a relaxed/neutral expression. When subjects had a mostly neutral expression with their mouth open they were assigned Open. In most cases Teeth referred to a smile. Other was used for indescribable expressions or closed mouth smiles. In the analysis, only the consolidated factor was used.

Eyes {Open, Closed, Other}. Consolidated to {Open, Other}. Default = Open. Open eyes were associated with relaxed open eyelids, with the person staring directly into the camera. Closed was also a relaxed expression, however with the eyelids closed. The Other rating was assigned to eyes that were half open, that looked somewhere other than directly at the camera, or that in some other way did not appear relaxed.

The system used to collect covariate data has a graphical user interface that displays a FERET image, both the original full image and the normalized version. The GUI also presents radio buttons for selecting covariate values. The person assigning the covariate values can rapidly step through images making appropriate selections. The results are logged to a simple ASCII file. The GUI is written in C++ using the Qt graphics API. We are happy to make this software available upon request.

3 Measuring Recognition Difficulty

The standard FERET evaluation protocol distinguishes between training, gallery and probe images. A PCA classifier uses the training data to determine a subspace in which a nearest neighbor classifier will match probe images to gallery images. Recognition rate is the fraction of the probe images that best match an image of the same subject, and thus recognition rate is defined over sets of probe and gallery images. Recognition rank is defined for a specific probe image and is the position of the first occurrence of an image of the same subject in the gallery images when those images are sorted by increasing distance relative to the probe image.

There are many ways to formalize the notion that a subject is hard to recognize. Given a specific gallery, one might equate difficulty with high recognition ranks: a subject with recognition rank 1 is easy, one with recognition rank 2 is harder, etc. There is some logic to this approach, but there are also problems. First, it not clear that difficulty is linear in recognition rank. Intuitively, the difference between ranks 1 and 10 carries far more weight than the difference between ranks 10 and 20. Second, since recognition rank is defined relative to a gallery, whether a subject is hard or easy to recognize becomes a global property of the entire set of images in the gallery: different galleries yield different ranks. While it is true that the performance of face recognition systems depends on the other subjects in the gallery, this dependency interferes with attempts to isolate the relative difficulty of specific subjects.

Here, we equate recognition difficulty with distance between pairs of images of the same subject. The assumption is a simple one: a nearest neighbor classifier is more likely to recognize a subject at rank 1 when the two images of the subject are close together. The advantage of this assumption is it yields a performance variable that depends solely upon the two images of the subject in question (and on training). To emphasize, this means the performance variable is not dependent upon distance between images of the subject and other subjects and is therefore not dependent upon a specific gallery set.

It is important to examine whether the assumption that rank 1 recognition is easier for images that are close together is valid in practice. One can imagine pathologies where many subjects might cluster very tightly in subspace, resulting in small distances between pairs of images of the same subject, but similarly small distances between pairs of images of different subjects. To test for such pathologies, we also look at rank-distance between pairs of images of the same subject. Rank-distance is closely associated with recognition rank, and is formally defined in Section 5.

4 Primary Experiment

We conducted an ANOVA to determine how subject covariates influence the distance between pairs of images of the same subject. In this section, the experimental design is described, followed by the model, and then the results. Subsequent sections will investigate specific questions/concerns associated with this primary experiment.

4.1 Primary Experiment Design

Several pilot studies were conducted using up to $2,974$ images of $1,120$ subjects prior to the primary experiment presented here. These pilot studies were important in enabling us to arrive at the final experiment design, and lead us to make decisions such as to limit ourselves to pairs of images for only $1,072$ subjects. Thus, only image pairs taken on the same day were included since increased time between images is known to make recognition harder. Also, several subjects were discarded because they wore glasses in one picture and not in another. The pilot studies confirmed that this makes recognition harder, and since it is not a surprising or terribly interesting result, these cases were removed.

Pilot studies also included multiple pairs of images for some subjects. This created several problems. It meant that some response values were correlated within subjects while

others were not. It also made subsequent balancing of the training data more difficult. Our primary experiment therefore includes only one pair of images per subject. More will be said about balanced training below.

The CSU standard PCA algorithm was used in this study. The code for this algorithm is part of the CSU Face Identification Evaluation System [5] and is available through our web site [2]. The PCA algorithm was trained using all $2,144$ images, and hence training was carried out using all the data subsequently used in our analysis. While it is typically a mistake to train on the test data when evaluating algorithms, it is appropriate when focusing specifically on questions of which subjects are close in subspace and which are not, and when one does not wish to complicate the question with whether zero, one or both of the subject images were in the training set. In keeping with common practice for a PCA classifier, the resulting subspace was truncated at 90% energy, resulting in 177 basis vectors. Subsequent to training, all $2,144$ images were projected into the subspace and the distance between all pairs of images was recorded.

The specific distance measure used, Mahalanobis metric, has been shown to perform best on the FERET data both in our own prior studies [15, 1] as well as in studies done by Moon and Phillips [11]. There has been some drift in the definition of this measure. For a pair of images A and B already projected into the PCA subspace, the distance measure was defined by Moon [11] as:

$$d_m(A,B) = -\sum_{i=1}^{k} \sqrt{\frac{\lambda_i}{\lambda_i+\alpha^2}} a_i b_i \quad \text{where} \quad \alpha = 0.25 \qquad (1)$$

where a_i and b_i are the ith components of the projected images and λ_i is the ith eigenvalue.

Yambor [15] found a simpler variant performed better and defined the measure as:

$$d_y(A,B) = -\sum_{i=1}^{k} \frac{1}{\sqrt{\lambda_i}} a_i b_i \qquad (2)$$

Most of the experiments performed by Yambor assumed a pre-processing step that normalized all images in PCA space to be of unit length. To generalize the measure intended by Yambor to images that may not be of unit length, the measure may be written as:

$$d(A,B) = -\sum_{i=1}^{k} \frac{1}{\sqrt{\lambda_i}} \frac{a_i}{|A|} \frac{b_i}{|B|} = -\frac{1}{|A||B|} \sum_{i=1}^{k} \frac{1}{\sqrt{\lambda_i}} a_i b_i \qquad (3)$$

This distance $d(A,B)$ is the Mahalanobis metric used here.

In addition to recording the Mahalanobis metric between the pair of images for each subject, the rank-distance is also recorded. For a given subject, this is done by selecting the first image of the subject as a probe, and then sorting the remaining 2,143 images by increasing distance. Rank-distance is then the position in this sorted list of the other image of the subject[1]. This rank-distance will be used below to test the strength of the relationship between Mahalanobis metric distance and rank 1 recognition.

4.2 ANOVA Model

The statistical modeling used in the primary experiment was an analysis of variance (ANOVA). The model is defined as follows:

$$Y_i = \text{distance between two images of subject } i \qquad (4)$$
$$X_{ij} = \text{subject covariate factor } j \text{ for subject } i \qquad (5)$$
$$\beta_j = \text{parameters quantifying factor } j\text{'s effect} \qquad (6)$$

and

$$Y_i = \beta_0 + X_{i1}\beta_1 + X_{i2}\beta_2 + \ldots + e_i \qquad (7)$$

where the e_i are assumed to be iid normal random variables with mean zero. Note that this model assumes purely additive effects with no interactions. We made this choice to ensure reliable parameter estimation with a simple model form. We did not have sufficient data to estimate interaction effects between covariates.

In this somewhat unconventional notation (adopted here to elucidate factor codings and effects), the components $X_{ij}\beta_j$ are products between a row vector X_{ij} and column vector β_j. This accommodates factors with more than two possible outcomes. In our case, this can either happen when a factor has more than two possible values, as with race, or when a factor may change from one image of the subject to another, as with Eyes. Table 1 shows our specific encoding for this analysis.

So, for example, with Eyes, the vector $[0,0]$ indicates the eyes are open in both images of the subject, $[1,0]$ indicates eyes are always closed, and $[0,1]$ indicates the eyes are open in one image an closed in another. Note that this encoding does not distinguish which image has the eyes open for the case where the images differ. This is intentional, since the case where there is a change between the images is of interest, but there is nothing special about the order of the images.

One can see in Table 1 that five factors are encoded with a 1x1 vector, five are encoded with a 2x1 vector, and one, Race, is encoded with a 3x1 vector. The model includes one parameter for each element in these vectors, and thus the entire model has 19 parameters to estimate: 18 for the factors plus the offset β_0. Note that β_0 is the regression parameter for our base-case in which all X_{ij} are zero. Reading off the zeroes in Table 1 shows that the base-case is a young, white, male with clear skin, no glasses, no facial hair, no makeup,

[1] We subtracted 1 so that the ideal outcome corresponded to a rank distance of 0.

no bangs, a neutral expression, a closed mouth and eyes open.

4.3 Primary Experiment Results

The ANOVA results are summarized in Figure 1. Base-case settings are indicated down the center of the diagram, with the degree and direction of effects noted. Effects are expressed as percent change from base-case, and rescaled in terms of similarity (1 minus distance) so that positive effects correspond to easier recognition. The threshold of a two-sided 95% confidence interval is shown as a thin vertical line. Solid bars indicate statistically significant changes in distance measure, whereas hollow bars indicate non-significant effects.

To illustrate, the pairwise distance between images of subjects always wearing glasses is reduced by 35% relative to the base-case of subjects not wearing glasses. In this case, the thin vertical line indicating statistical significance appears at 9%; thus the effect is highly significant and the bar is shaded solid. A reduction in distance suggests the subjects are more easily recognized, hence the bar for Glasses Always On is shown on the right side of the chart.

Conversely, consider the subjects whose eyes are open in one image and closed in another. In the base case, subjects have eyes open in both images. The effect for Eyes Open/Closed is the top solid bar on the side of Figure 1 corresponding to more difficult recognition, and it indicates a 12% increase in relative distance between pairs of subjects. Thus, not surprisingly, our study suggests that subjects are significantly harder to recognize if they close their eyes in one image but not the other. Perhaps more surprising, observe that subjects whose eyes are always closed are significantly easier to recognize than those whose eyes are always open.

The ANOVA yielded $R^2 = 0.39$, indicating that about 39% of the total variation in similarity can be explained by the subject covariates. When compared to baseline runs of the PCA algorithm, in which about 75% of subjects can be correctly recognized at rank 1, this R^2 is notable.

Subject Covariates

Factor	Values	X_{ij}
Age	Young	$[0]$
	old	$[1]$
Race	White	$[0,0,0]$
	Asian	$[1,0,0]$
	Black	$[0,1,0]$
	Other	$[0,0,1]$
Gender	Male	$[0]$
	Female	$[1]$
Skin	Clear	$[0]$
	Not Clear	$[1]$
Glasses	Always No	$[0]$
	Always Yes	$[1]$
Facial Hair	No	$[0]$
	Yes	$[1]$

Image Specific Covariates

Factor	Values		X_{ij}
Makeup	No	No	$[0,0]$
	Yes	Yes	$[1,0]$
	No	Yes	$[0,1]$
	Yes	No	$[0,1]$
Bangs	No	No	$[0,0]$
	Yes	Yes	$[1,0]$
	No	Yes	$[0,1]$
	Yes	No	$[0,1]$
Expression	Neutral	Neutral	$[0,0]$
	Other	Other	$[1,0]$
	Neutral	Other	$[0,1]$
	Other	Neutral	$[0,1]$
Mouth	Closed	Closed	$[0,0]$
	Open	Open	$[1,0]$
	Closed	Open	$[0,1]$
	Open	Closed	$[0,1]$
Eyes	Open	Open	$[0,0]$
	Closed	Closed	$[1,0]$
	Open	Closed	$[0,1]$
	Closed	Open	$[0,1]$

Table 1: Factor encoding used in the model, equation (7).

4.4 Primary Experiment Conclusions

Some aspects of Figure 1 are of particular interest. Starting with the most significant effect observed, glasses help face recognition. This is perhaps startling, since at least one author has suggested that synthetic removal of glasses is a critical pre-processing step required to improve face recognition [16]. It is tempting to invent post hoc explanations for why glasses improve performance, and one reasonable hypothesis is that glasses are distinguishing features being encoded in the PCA space. Note, it is implicit for our tests that a subject wearing glasses is wearing the *same* pair of glasses in each image. Further empirical work is needed to see if this result extends to other data sets and to better explain why the glasses effect is so pronounced.

Another somewhat startling outcome is the race effect. In our set of $1,072$ FERET subjects, 720 are white, 143 are Asian, 121 are African-American and 88 are other races. Relative to the majority of the subjects (which are white), Asians, African-Americans and others are all significantly easier to recognize. This is not what we expected going into this experiment. To the contrary, our expectation was that a PCA space trained primarily on white subjects would favor those subjects. Frul et. al. [10] have observed a similar

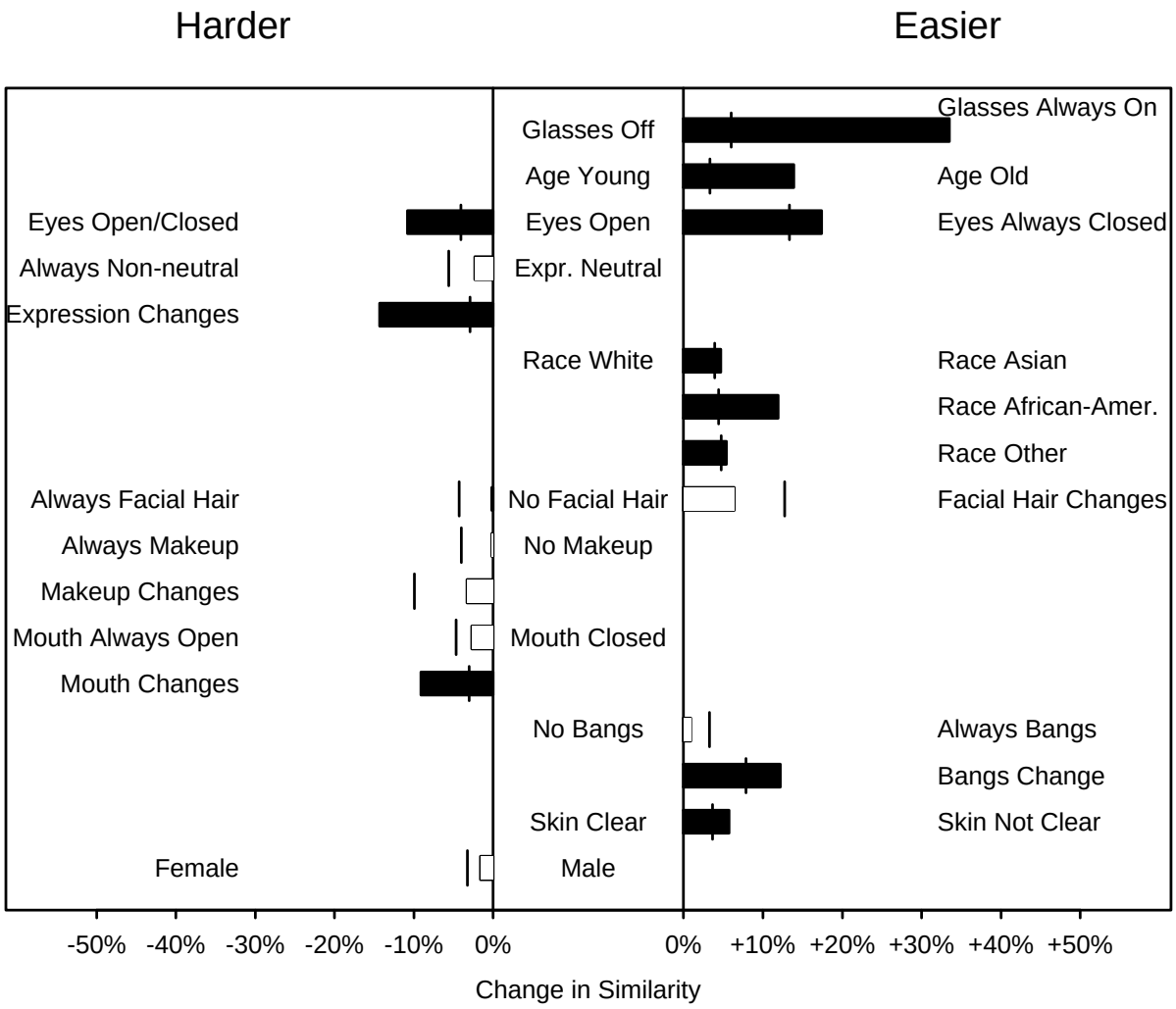

Figure 1: Results of ANOVA for primary FERET subject covariate study. See section 4.3 for explanation of this graph.

result for a smaller subset of the FERET data and looking only at the distinction between White and Asian.

Another result from Figure 1 worth noting is the lack of a significant gender effect. Of the $1,072$ subjects used in this study, 624 were male and 448 are female. Many researchers engaged in face recognition work have at one time or another been part of informal discussions of whether men or woman our more easily recognized, and it is intriguing how often researchers have an opinion on this question.

There is little prior formal evaluation of gender. One important exception is the work of Gross et. al. [12]. They report recognition rates of 87.6 for males and 93.7 for females using $1,119$ subjects. However, direct comparison is not appropriate. Gross et. al. used a different data set, a different algorithm (FaceIt [4].), and a weaker analytic technique; comparing recognition rates over whole galleries partitioned only by gender. The difference in analytic technique alone could explain the difference. We are currently working on analyzing our data using the simpler approach employed by Gross et. al. in order to determine for certain if our failure to observe a significant gender effect might be direct consequence of our doing a more complete covariate analysis that controls for other factors potentially confounded with gender in the simpler analysis.

Finally, we note a variety of other significant results are shown in Figure 1, including the result that old people were significantly easier to recognize than young ones. Perhaps this confirms the idea that a face gains character with age.

Looking at the race effect, one might imagine obtaining the result shown in Figure 1 without it directly translating to

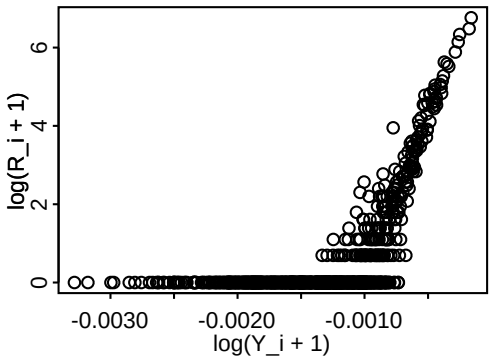

Figure 2: Log-log relationship between rank distance and distance.

improved recognition performance. Essentially, one might argue that Asians, African-Americans and others races are clustering more tightly due to low representation in the PCA training set. Indeed, one might raise this concern with any of the results we've found where the "easier" category represents a minority of the total subjects. In Section 6 we will report on a series of follow up experiments that test and solidly refute this criticism.

One might also question whether the results of our initial experiment are flawed due to the reliance on pairwise distance to predict recognition performance. As already suggested in our introduction, this is a potentially valid concern, and one that can be addressed by studying how pairwise distance between images of the same subject relates to rank-distance. This analysis of rank-distance is presented in the following section, and again the criticism is found to be empirically unsubstantiated.

5 Relating Distance to Rank Distance

Figure 2 shows the log-log relationship between the distance (Y_i) between images used as our response variable, and the rank distance (R_i) described above which relates directly to recognition rank. Clearly there is a very strong relationship between these two variables. This is reassuring because it suggests that inferences about subject covariates based on distance are likely to hold for recognition rank, too.

To further confirm this relationship, we fit a logistic regression [9] to rank 1 recognition: the response variable was $Z_i = 1$ if $R_i = 0$, and $Z_i = 0$ if $R_i > 0$. The predictor variable was distance, Y_i. The model can be summarized as

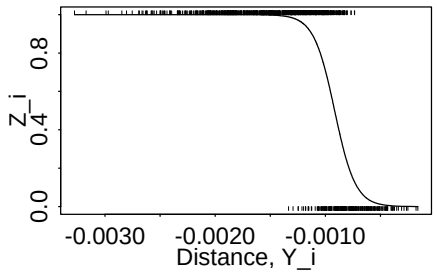

Figure 3: Results of logistic regression of rank distance on Mahalanobis metric distance.

$Z_i|Y_i \sim \text{Bernoulli}(p_i)$ where $\log(p_i/(1-p_i)) = \beta_0 + \beta_1 Y_i$.

Figure 3 shows the results of this model. The individual Z_i are shown as hash marks. The smooth line is the fitted curve, which shows that the probability of a rank 1 match decreases sharply with increasing Mahalanobis metric distance. In fact, the estimate of β_1 is -10585.3 with standard error 809.5, and the negative relationship between these variables is strongly significant.

The same analysis was carried out individually for individual groups of subjects, grouped by race, glasses, age, skin, and gender. The same conclusion was found in every case. These results effectively refute the potential criticism of our primary study that our chosen response variable is uninformative about recognition performance of the algorithm.

6 Balance Experiments

Many of the results from our primary experiment could be explained by arguing that groups of subjects underrepresented in training appear closer in PCA subspace because PCA is proportionally under-representing the portion of the space in which these groups lie. Consequently, they appear more tightly clustered than do the majority groupings. If this hypothesis is true, it would be of considerable concern, since we are drawing the conclusion that Asians are easy to recognize based upon the decreased average distance between pairs of images of Asians relative to whites.

To test this hypothesis we repeated the primary experiment with a training data set that was carefully balanced, so that groups of interest are equally represented. We ran a total of six additional experiments, balancing training on various groupings of one or more variables—race, age, skin or glasses—while controlling for others. It was not possible to balance over the eyes factor due to an insufficient number of subjects in some categories.

Table 2 summarizes the experiments we conducted. For

To Test Race *Glasses = Off, Eyes = Open, Skin = Clear*					
		Balance Age		Total	PCA
Test	**Compare**	**Young**	**Old**	**Images**	**Dim.**
Asian	Asian	89	6	380	78
	White	89	6		
Black	Black	78	16	376	102
	White	78	16		
Other	Other	62	6	272	75
	White	62	6		

To Test Age *Glasses = Off, Eyes = Open, Race = White*					
		Balance Skin		Total	PCA
Test	**Compare**	**Clear**	**Other**	**Images**	**Dim.**
Age	Old	131	57	752	130
	Young	131	57		

To Test Skin *Glasses = Off, Eyes = Open, Race = White*					
		Balance Age		Total	PCA
Test	**Compare**	**Old**	**Young**	**Images**	**Dim.**
Skin	Clear	72	57	516	117
	Other	72	57		

To Test Glasses *Eyes = Open, Race = White, Skin = Clear*					
		Balance Age		Total	PCA
Test	**Compare**	**Old**	**Young**	**Images**	**Dim.**
Glasses	Off	14	16	120	50
	On	14	16		

Table 2: Subject counts, image counts and PCA dimensions for training in the balanced training experiments.

example, to confirm whether Asians were really easier to recognize, we balanced Asians and Whites across age. In other words, we chose equal numbers of young Asians and Whites, and equal numbers of old Asians and Whites. We controlled for three variables (glasses, eyes, and skin) by limiting consideration to only subjects with open eyes, clear skin, and no glasses. This balancing and controlling defines a subset of the subjects used in the primary analysis, and the sample sizes shown in Table 2 reflect the limited numbers of some types of subjects in the dataset. Our balancing and controlling strategy was designed to provide the largest possible sample sizes in the individual cells in Table 2.

Continuing the example of comparing Asians and Whites, we then trained PCA on the 380 images listed in row one of Table 2. Thus, Asians and Whites were equally represented in the training, and balanced or controlled on other important covariates. We then projected all 2,144 images into PCA space, computed the distances between images, and conducted ANOVA modeling as before. An analogous balancing and controlling process was conducted for the other factors listed in Table 2.

Of the several ANOVA models we fit to the data from these six experiments, we report here results from full models identical to (7). In these models, it is important to examine only the effect of the particular variable for which balancing is being conducted. The experiments confirmed that, adjusting for other factors, Asians are easier than whites (p-value $= 0.0104$), African-Americans are easier than whites (p-value $= 0.0064$), other race members are easier than whites (p-value $= 0.0249$), old people are easier than young people (p-value < 0.0001), other skin people are easier to recognize than clear skin people (p-value $= 0.0122$), and subjects with glasses are easier to recognize than subjects without glasses (p-value $= 0.0005$).

Thus, whether PCA is trained on an imbalanced collection of diverse people (as is likely in many real applications), or on a carefully balanced collection of people, our results confirm that some people are easier to recognize than others, and that subject-specific covariates can explain a significant portion of this variation.

References

[1] J. Ross Beveridge, Kai She, Bruce Draper, and Geof H. Givens. A nonparametric statistical comparison of principal component and linear discriminant subspaces for face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 535 – 542, December 2001.

[2] Ross Beveridge. Evaluation of face recognition algorithms web site. `http:\\cs.colostate.edu\evalfacerec`.

[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] Identix Corporation. Faceit system homepage. www.identix.com, 2002.

[5] M. Teixeira D. Bolme, R. Beveridge and B. Draper. The csu face identification evaluation system: Its purpose, features and structure. In *Proceedings of the Third International Conference on Vision Systems*, page (to appear), Graz, Austria, April 2003.

[6] Duane M. Blackburn, Mike Bone and P. Jonathon Phillips. Facial Recognition Vendor Test 2000. http://www.dodcounterdrug.com/facialrecognition/frvt2000/frvt2000.htm, DOD, DARPA and NIJ, 2000.

[7] FERET Database. http://www.itl.nist.gov/iad/humanid/feret/. NIST, 2001.

[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, New York, NY, 1989.

[10] Nicholas Furl, P. Jonathon Phillips and Alice J. O'Toole. Face recognition algorithms and the other-race effect: computational mechanisms for a developmental contact hypothesis. *Cognitive Science*, 26:797 – 815, 2002.

[11] 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.

[12] Jeffrey F. Cohn Ralph Gross, 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.

[13] 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.

[14] W. Zhao, R. Chellappa, A. Rosenfeld, and J. Phillips. Face Recognition: A Literature Survey. Technical Report CS-TR4167R, Univ. of Maryland, 2000. Revised 2002.

[15] Bruce A. Draper Wendy S. Yambor and J. Ross Beveridge. Analyzing pca-based face recognition algorithms: Eigenvector selection and distance measures. In H. Christensen and J. Phillips, editors, *Empirical Evaluation Methods in Computer Vision*. World Scientific Press, Singapore, 2002.

[16] Robert Mainani Zhong Jing and Jiankang Wu. Glasses detection and extraction by deformable contour. In *International Conference on Pattern Recognition Proceedings*, Spain, 2000.