Using A Generalized Linear Mixed Model to Study the Configuration Space of a PCA+LDA Human Face Recognition Algorithm *

Geof H. Givens
Statistics Department
Colorado State University
Fort Collins, CO, 80523

J. Ross Beveridge, Bruce A. Draper and David Bolme
Computer Science Department
Colorado State University
Fort Collins, CO, 80523

April 25, 2003

Abstract

A generalized linear mixed model is used to estimate how rank 1 recognition of human faces with a PCA+LDA algorithm is affected by the choice of distance metric, image size, PCA space dimensionality, supplemental training and inclusion of subjects in the training. Random effects for replicated training sets and for repeated measures on people were included in the model. Results indicate between people variation was a dominant source of variability, and that there was moderate correlation within people. Statistically significant effects and interactions were found for all configuration factors except image size. Changes to the PCA+LDA configuration only improved recognition for subjects who had images included in the training data. For subjects not included in training, no configuration changes were helpful. This study is instructive for what it reveals about PCA+LDA. It is also a model for how to conduct such studies. For example, by accounting for subject variation as a random effect and explicitly looking for interaction effects, we are able to discern effects that might otherwise have been masked by subject variation and interaction effects.

1 Introduction

Proper configuration or tuning is a critical factor in making effective use of most vision algorithms. Virtually all algorithms have a configuration space, and it is well understood performance varies with different configuration choices. There are a variety of approaches to configuration. Perhaps the most obvious and most common is semi-structured exploration: the developers of the algorithm run the algorithm, examine results, make changes, and repeat until they are satisfied. Somewhat more rigorous is iterative refinement with an objective performance standard as illustrated by the work of Bowyer [3].

It is less common to design experiments, collect large amounts of performance data, and then apply statistical tools to determine how different configuration choices influence performance. We adopt this approach here by applying sophisticated statistical models to understand the configuration of an interesting class of human face recognition algorithms. The statistical approach is a generalized linear mixed model. The face recognition algorithm uses principal components analysis (PCA) followed by linear discriminant analysis (LDA) and is based closely on the work of Zhao [12].

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

The use of the generalized linear mixed model in this context provides several very important features lacking from simpler approaches. First, it permits us to model what are called fixed effects that are attributable to various algorithm configuration choices. These include such things as what size images to use and how to train the algorithm. Second, it also permits modeling of random effects thought to influence performance but which are not under our direct control. A good example of a random effect is the specific identity of the human subjects.

Proper accounting for random effects due to subject identity is of particular importance in domains such as human face recognition. This is because some individuals are harder to recognize than others, while at the same time we are seldom afforded the luxury of choosing to only recognize the easy ones. Thus, a large source of variation in observed performance is attributable to a recorded variable that is not under our control. Were a simpler model used that did not explicitly capture and appropriately attribute this variance to subject identity, between subject variability could easily interfere with the detection of other effects of interest to us.

2 Image Pre-processing and the PCA+LDA Algorithm

How face image data is pre-processed is arguably as important as any other aspect of algorithm configuration. In the experiments presented below, the face imagery has already been pre-processed in essentially the same fashion in the FERET tests [8, 2]. Specifically, the following has been done to the images:

1. An integer to float conversion is done first.
2. Geometric normalization then aligns hand picked eye coordinates. By default, the resulting image is 130 by 150 pixels after normalization and eye centers are always at exactly the same position.
3. The aligned images are cropped using an elliptical mask centered on the face. This leaves the face visible but blanks out non-face portions of the image.
4. Histogram equalization is applied to the pixels in the face region.
5. Pixel normalization scales pixel values so each image has mean zero and standard deviation one.

Optional pre-processing used in the experimental design presented below includes:

- Add zero mean Gaussian noise with a specified standard deviation to each pixel.
- Reflect the image down the center of the face. This leaves the eyes in the same position, but obviously creates a mirror-image of the original face.
- Shift the image an amount left, right, up or down.
- Reduce the image size by a factor of two.

More will be said about these optional pre-processing steps in Section 3 as well as whether they alter how the face recognition algorithm performs.

2.0.1 The PCA+LDA Algorithm

The PCA+LDA human face recognition algorithm considered in this study is based upon the work of Zhao [12]. Prior to algorithm training, all images are pre-processed as just described. The PCA+LDA algorithm then takes a set of training images and builds a PCA subspace just as a standard Eigenfaces [7] algorithm would do. Next, all the training images are projected into the PCA subspace and are grouped according to subject identity. Each subject is treated as a distinct class and the LDA basis vectors are computed. In the experiments that follow, training is performed on 64 randomly selected subjects with 3 images each. Thus, the maximum possible dimensionality of the PCA space is 192, and one of the configuration choices involves how many of these to retain. Since there are 64 subjects, the LDA generates 63 Fisher basis vectors. The two linear projections are then composed, yielding a single projection from image space into the PCA+LDA space.

In the testing phase, the PCA+LDA algorithm reads the subspace basis vectors generated during training, projects test imagery into this subspace, and measures distance between all pairs of images. Two distance measures are considered in the experiments below. The first is the standard L2 norm or $\mathcal{L}_2$. The second is a heuristic distance measure defined by Zhao [12] and called here Soft L2 or $\mathcal{L}_{Soft}$. Thus, for two images A and B, these two distance measures are:

$$\mathcal{L}_2(A,B) = \sum_{i=1}^{N} (A_i - B_i)^2 \qquad \mathcal{L}_{Soft}(A,B) = \sum_{i=1}^{N} (\lambda_i)^{0.2} (A_i - B_i)^2 \tag{1}$$

where λ_i is the ith eigenvalue from the the solution of the generalized eigenvector problem associated with finding the fisher basis vectors [4].

3 Experimental Design

This study used $1,024$ images for 256 FERET subjects. The choice of subjects is based upon the data available: there are essentially only 256 FERET subjects with at least 4 frontal images. Before doing any algorithm training or testing, all $1,024$ images were pre-processed using the five steps described above.

The response variable for this experiment is the recognition rank for a given subject for a particular configuration of the algorithm. This lead us to select 1 of the 4 images for each of the subjects as the probe image and to then never vary this choice. Reviewing terminology briefly, a *probe image* is a new image of a person to be recognized and *gallery images* are stored images of known subjects against which the probe image is matched by a nearest neighbor classifier [8].

Recognition rank for a subject using a specific probe image is the placement of the subject's corresponding gallery image in the gallery when the gallery is sorted by increasing distance from the probe image. This rank is essentially measuring how well a nearest neighbor classifier will recognize a subject. Thus, if the recognition rank is 1 for a subject, this means a nearest neighbor classifier will correctly recognize the subject because the closest gallery image is an image of the same subject. A recognition rank of 2 says there is 1 gallery image of another subject that appears more similar to the subject than does the corresponding gallery image of that subject. A recognition rank much larger than 1 suggests an algorithm will do a bad job of recognizing that subject.

One of the more important aspects of how one configures an algorithm such as the PCA+LDA algorithm is through training. The training imagery was drawn from the remaining 3 images for each of the 256 subjects. In each test performed here, the PCA+LDA algorithm was trained on $\frac{1}{4}$ of the subjects. Thus, the algorithm was trained on 192 images, 3 images for each of the 64 randomly selected subjects. An additional 192 images were added to the training in some tests to see if supplementing the training with altered versions

of images improves performance. Three specific sources of supplemental imagery were used. Note in all three cases one supplemental image is generated for each of the original training images by doing one of the following three things:

1. Reflect each image horizontally about its mid-line.
2. Add independent mean zero Gaussian noise (with a standard deviation of 2% of the dynamic range of the image) to the grey-scale value of each pixel in each image.
3. Shift each image left by one pixel.

The entire training procedure using 64 randomly selected subjects was itself replicated 8 times in order to determine how much the exact choice of training subjects matters.

The change in the size of the training set from 192 images without supplemental training to 384 with supplemental training is particularly important if it affects the dimensionality of the PCA space used for discrimination. Therefore, this dimensionality was explicitly controlled by setting it to either 114 eigenvectors (60% of 191 or 30% of 383) or the number of eigenvectors that comprise 90% of the energy, denoted $E_{90\%}$ [5]. In other words, keep k eigenvectors where the sum of the k largest eigenvalues is equal to 90% of the sum of all the eigenvalues. Typical values of k using the $E_{90\%}$ rule to truncated the eigenspace ranged from 64 to 107.

To test how much effect the exact images chosen to be in the gallery has on algorithm performance, each probe was tested on 30 different gallery sets. These galleries were generated by randomly selecting 1 of the remaining 3 images per subject. The same 30 randomly generated gallery sets were used throughout the entire experiment.

All the tests in this experiment were run using the standard $\mathcal{L}_2$ norm and the $\mathcal{L}_{Soft}$ measure proposed by Zhao [12] and defined above in equation 1. It was also brought to our attention in discussions with Zhao that he typically reduced the size of images prior to recognition. Thus, two different image sizes are tested in this experiment, the FERET standard (130×150) and small images with $\frac{1}{4}$ as many pixels (65×75).

One additional variable was observed in this experiment: whether a particular probe had been included in the training set used in each test. Since there were 256 probes and 64 subjects in the training set, only 25% of probes were included in any training set.

The experimental design was a balanced complete factorial; in other words we ran each choice in every combination. This results in $8 \times 30 \times 4 \times 2 \times 2 \times 2 = 7680$ attempts to recognize each probe image under different circumstances. A summary of the configuration and replication factors used in this experimental design is given in Table 1.

4 The Model

Let $R_{tpgusmdf}$ represent the recognition rank when the p^{th} probe image of the s^{th} size and f^{th} value of the training set inclusion flag is probed in the g^{th} gallery, using an algorithm with the m^{th} metric limiting the PCA discrimination space in the d^{th} manner, when the algorithm is trained on the t^{th} baseline training set with the u^{th} type of training set supplementation. Then let $Y_{tpusmdf}$ be the number of gallery sets for which $R_{tpgusmdf} = 1$. Note that $0 \leq Y_{tpusmdf} \leq 30$ for our experiment.

We modeled $Y_{tpusmdf}$ as a binomial random variable, namely $Y_{tpusmdf} \sim Bin(30, p_{tpusmdf})$. The probability of rank 1 recognition (namely $p_{tpusmdf}$) depends on various factors as follows:

$$\begin{aligned} \log \frac{p_{tpusmdf}}{1-p_{tpusmdf}} &= L + U_u + +M_m + D_d + zI_t(p) + \tau_t + \pi_p + (UM)_{um} + (UF)_{uf} + \\ &\quad (UD)_{ud} + (MF)_{mf} + (MD)_{md} + (DF)_{df} + (MDF)_{mdf} \end{aligned} \tag{2}$$

Factor	Index	Values
Training Set	t	$0,\ldots,7$
Gallery Set	g	$0,\ldots,29$
Probe Image	p	$0,\ldots,255$
Training Set Supplement	u	None, Shift, Reflection, or Noise
Metric	m	Standard or Soft L2,i.e. $\mathcal{L}_2$ or $\mathcal{L}_{Soft}$
PCA Space Dimension	d	114 or $E_{90\%}$
Image Size	s	Standard (130×150) or Small (65×75)
Training Set Inclusion Flag	f	No or Yes

Table 1: Summary of the configuration and replication factors used in the experimental design.

where

L	is the grand mean
U_u	is the effect of including the u^{th} supplement to the training set
M_m	is the effect of using the m^{th} distance metric
D_d	is the effect of the d^{th} strategy for limiting PCA dimensionality
F_f	is the effect of the f^{th} option for including a probe in the training set
$(UM)_{um},\ldots$	are 2- and 3-way interactions between main effects
τ_t	is the random effect of the t^{th} set of people used for training, and
π_p	is the random effect of the p^{th} probe image.

Commonly $\log(p/(1-p))$ is termed the logit transformation.

We are interested in the amount of variability in rank 1 recognition rate attributable to training set, but the 8 training sets used in our experiment are merely a random sample from the variety of sets that might have been used. Therefore, we fit $\tau_t \sim N\left(0,\sigma_\tau^2\right)$ as independent random effects. Using the same rationale, we modeled the π_p as mean zero normal random effects with constant variance (σ_π^2), but the covariance structure was more complex since results from probing the same probe image in different circumstances are probably more correlated than probing different probe images. We assumed a compound symmetric covariance structure, which allows correlation ρ between any pair of $Y_{tpusmdf}$ corresponding to the same probe image. Results for different probe images were modeled as independent. Thus, the model accounts for the possibilities that some training sets were more effective than others and that some probes were 'easier' than others. It also allows for correlation induced by repeated measurement of the same probes under different experimental conditions.

This model is a generalized linear mixed model [1], namely a mixed effects logistic regression with repeated measures on people. Such models can be fit routinely using restricted pseudo-likelihood or the MIVQEU0 method [9, 6, 10, 11].

5 Results

We originally fit a larger model including the main effect for image size and all its two-way interactions. A major finding of our study was that image size did not significantly affect rank 1 recognition rate, either overall or in any interaction. Through a sequence of model fitting steps, we removed all image size terms from the original model (and added the three-way *MDF* interaction) to obtain the final fitted model given in (2).

On the logit scale, the random training sets accounted for virtually no variability in rank 1 recognition

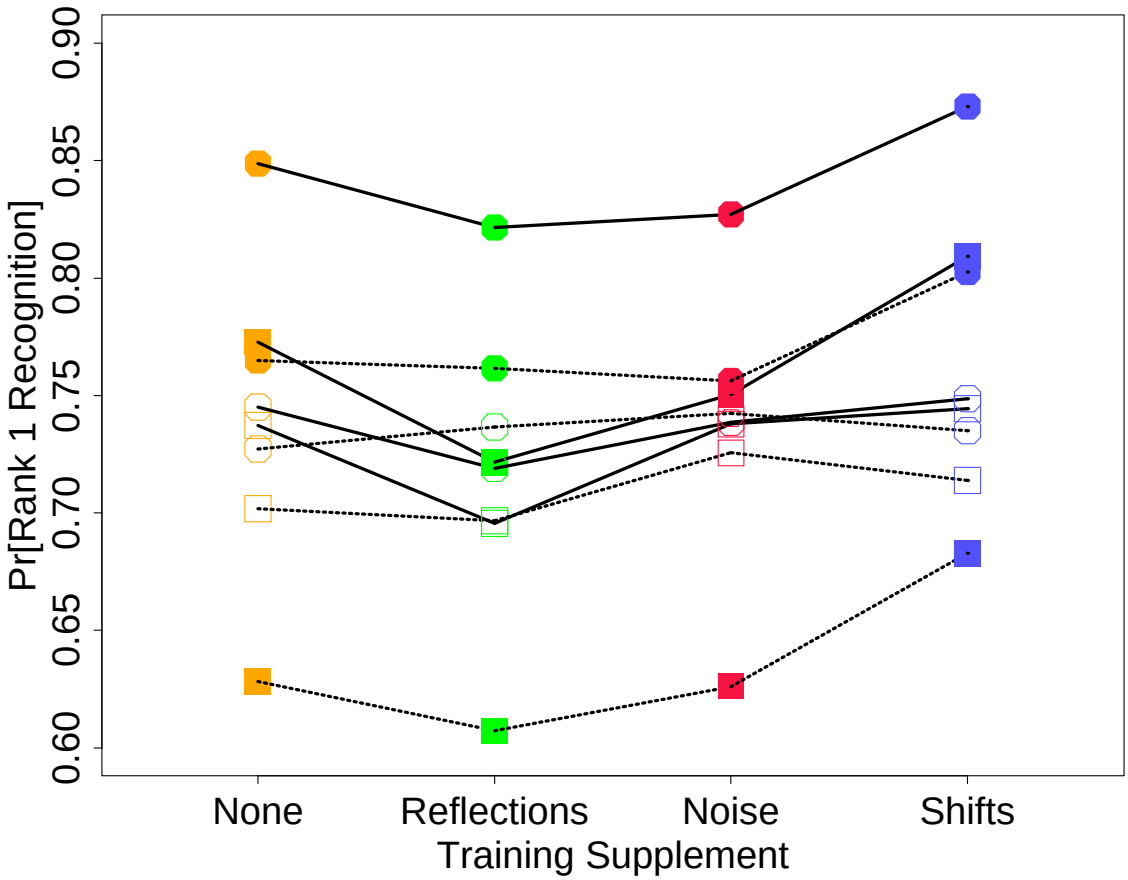

Figure 1: Effect of training set supplementation on predicted probability of rank 1 recognition. See the text for codings for color, shape, fill, and line type.

rate. Any randomly selected training set of this size is as good as any other. Also on the logit scale, the correlation between probes of the same person under different conditions was about 0.70. Our findings strongly support the conclusion that some people are harder to recognize at rank 1 than others, and that this variation accounts for the majority of the total variability in rank 1 recognition rate on the logit scale.

The best tested configuration of the PCA+LDA algorithm used the $\mathcal{L}_{Soft}$ distance metric with PCA dimensionality set at 60% of the number of eigenvectors and supplementing the training set with shifted images. Estimated probability of rank 1 recognition for this configuration was .87 for probes included in the training set and .75 for those not. In comparison, the estimated probabilities were .77 and .74, respectively, for the 'baseline' PCA+LDA configuration that use the standard $\mathcal{L}_2$ distance metric and no training supplementation.

The main focus of our study is the estimation of the main effects and interactions included in (2). These allow us to quantify how changes to the PCA+LDA configuration affect estimated probability of rank 1

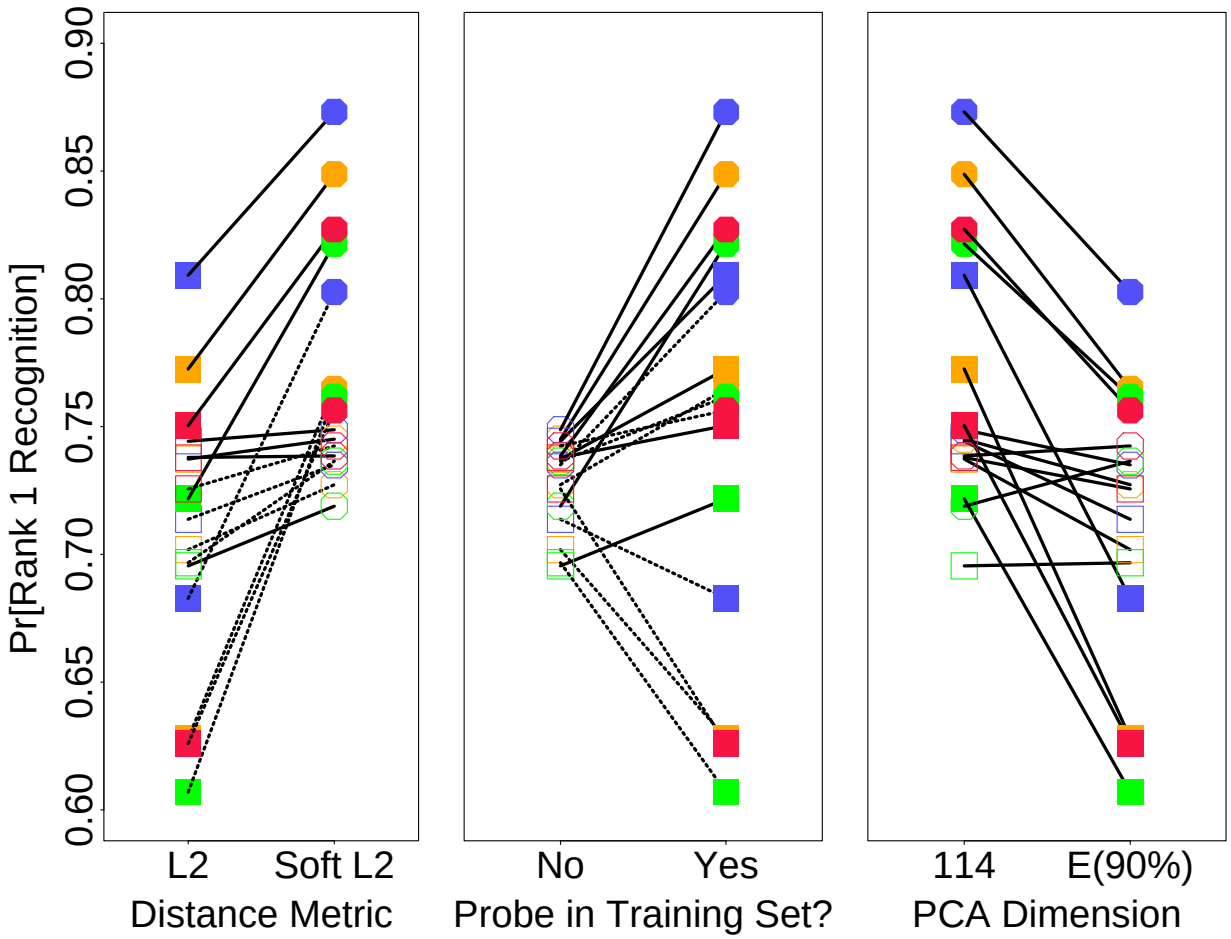

Figure 2: Effects of distance metric, PCA dimensionality, and inclusion of the probe in the training set on predicted probability of rank 1 recognition. See the text for codings for color, shape, fill, and line type. There is no line type variation in the right panel because PCA dimensionality is labeled on the horizontal axis.

recognition. Table 2 shows tests of the statistical significance of the various fixed effects. Figures 1 and 2 show the various estimated effects, as described in more detail in the rest of this section. The threshold for statistical significance in these figures is substantially smaller than the magnitude of difference that would be scientifically important in the context of human identification problems. Therefore it suffices to examine these figures for interesting effects with the understanding that any interesting effect would be statistically significant according to a simple F-test. The significance threshold is so small because the large sample size provides a lot of discriminatory power.

Figures 1 and 2 employ a consistent labeling to illustrate the results. Color indicates training supplement (orange for none, green for reflections, red for noise, and blue for shifts). Shape indicates distance metric (square for standard $\mathcal{L}_2$ and circle for $\mathcal{L}_{Soft}$). Fill indicates whether a probe is included in the training set

Effect	Label	(Num.) d.f.	F value	p-value
Training Supplement	U	3	139.1	$<.0001$
Distance Metric	M	1	1087.8	$<.0001$
PCA Dimensionality	D	1	936.6	$<.0001$
Training Inclusion Flag	F	1	333.8	$<.0001$
$U \times M$ Interaction	UM	3	14.2	$<.0001$
$U \times F$ Interaction	UF	3	67.0	$<.0001$
$U \times D$ Interaction	UD	3	41.5	$<.0001$
$M \times F$ Interaction	MF	1	717.6	$<.0001$
$M \times D$ Interaction	MD	1	59.1	$<.0001$
$D \times F$ Interaction	DF	1	703.7	$<.0001$
$M \times D \times F$ Interaction	MDF	1	5.04	.0248

Table 2: Table of significance of model effects and interactions based on approximate F-tests.

(solid fill for yes and hollow for no). Finally, PCA dimensionality is indicated by line type (solid lines for 114 and dashed for $E_{90\%}$).

Figure 1 shows the effect of training set supplementation on predicted probability of rank 1 recognition. Overall, adding shifted images was best and adding reflected images was actually worse than no supplementation. This refutes a hypothesis we originally held that reflections were effectively equivalent to shifts due to the rough horizontal symmetry of human faces. This figure also illustrates the interaction between the training set inclusion variable and supplementation type: the improvement or interference provided by training set supplementation was substantial only for recognition of probes that were included in the training set.

The left panel of Figure 2 shows the effect of distance metric on predicted probability of rank 1 recognition. Overall, the soft L2, $\mathcal{L}_{Soft}$, metric was superior, but this difference was mostly limited to instances when the probe was included in the training set. When training did not include the probe, very little was gained by switching from $\mathcal{L}_2$ to $\mathcal{L}_{Soft}$. This is the significant interaction.

The effect of including the probe in the training set is further shown in the center panel of Figure 2. Somewhat surprisingly, the overall effect of training set inclusion on predicted probability of rank 1 recognition was only slightly positive. There was an important three-way interaction with distance metric and PCA dimensionality which can be described as follows. The effect of training set inclusion was always positive when the PCA dimensionality was limited to 114. When the PCA dimensionality was set at $E_{90\%}$, the effect of training set inclusion depended on the choice of metric: positive effect for $\mathcal{L}_{Soft}$ and negative for standard $\mathcal{L}_2$. A definitive explanation for why inclusion of the probe in the training set actually impeded recognition when using the $\mathcal{L}_{Soft}$ with $E_{90\%}$ PCA eigenvectors eludes us.

The right panel of Figure 2 shows the effect of PCA dimensionality on predicted probability of rank 1 recognition. Overall, the use of $E_{90\%}$ PCA eigenvectors was inferior to fixing 114 eigenvectors. Since $E_{90\%}$ was virtually always less than 114, it is perhaps not surprising that this configuration was inferior. Again a significant interaction can be seen: the negative effect was mostly confined to probes that were included in the training set.

An important broader conclusion can be distilled from the details above. Variations in distance metric, PCA dimensionality, and training set supplementation exemplify the many sorts of strategies employed by researchers to improve the performance of the basic PCA+LDA algorithm. Such strategies appear to span a conceptual dimension with opposing approaches at each extreme. Direct manipulation of the training lies

at one extreme (eg. training supplementation) and changes in discriminatory strategy of the algorithm lies at the other extreme (eg. changing the distance metric). Both ends of the spectrum are motivated by the hope that such changes will improve recognition for probes not used in training. Our results firmly establish that none of these strategies succeed in this goal. On the other hand, our results also show that either end of the strategic spectrum can yield improved results on probes that are included in the training set. Changes to the algorithm provide about twice the improvement as changes to the training.

6 Discussion

Since all the tested variations in the baseline configuration of the PCA+LDA algorithm only improved rank 1 recognition for probes that were included in the training set, our results raise several important questions. First, the earlier work of Zhao indicated the soft-L2 measure would improve performance for subjects not in the training set. The discrepancy between our findings and Zhao's suggests more work may be needed before this matter is fully understood. Second, and more broadly, how might one best configure a PCA+LDA algorithm to enhance performance for such subjects.

Another avenue of future work might consider alternate and more generous definitions of recognition success. For example, do our results hold when a 'success' is defined as ranking the correct gallery image at rank 10 or higher? Although our statistical methodology could be applied equally well to such a problem, the FERRET data are not very suitable for such an analysis. About 47% of the 65,536 attempts at recognition under different configurations and training resulted in a rank 1 match for all 30 galleries; the overall frequency of rank 1 matches was 76%. If the response variable in our analysis were changed to, say, rank 10 recognition, the outcomes would be skewed even more toward successful matches and the reliability of the inference would be decreased, all else being unchanged.

The randomness inherent in our outcome—rank 1 recognition—is generated by the repeated testing of probes under 30 different galleries. In a real application, similar randomness applies to probe images as well. The advantage of the experimental design used here is that the binomial sample size parameter can be set quite large (we chose 30). Consequently, the $Y_{tpusmdf}$ are very informative about the recognition rate parameter, $p_{tpusmdf}$. An alternative experimental design would have been to fix a single gallery and to test different probe images of the same person. In the FERRET data, however, there are only a small number of frontal images for most subjects. Thus, the observed recognition rate for each individual subject is less informative. When we have access to a large data set with many replicates per subject, we plan to run such an analysis varying probe presentation.

To close, the results presented here represent one of the largest tuning studies of a significant face recognition algorithm on record. As an exercise in advancing the state of experiment design, it is great success. For example, the existence of strong interactions between whether a subject is in the training set and whether the algorithm does better or worse is expected, but characterizing this dependence in clear quantitative terms is significant. As to what this study tells us about our ability to significantly alter, in particular improve, algorithm performance, the results are somewhat disappointing. Clearly, there is a need for additional study of the PCA+LDA algorithm in order to truly demarcate those factors of algorithm configuration that may, in future, prove to significantly improve performance. At a minimum, such studies should adhere to the standards established here.

Acknowledgments

This work supported by the Defense Advanced Research Projects Agency under contract DABT63-00-1-0007.

References

[1] N. E. Breslow and D. G. Clayton. Approximate inference in generalized linear mixed models. *J. Amer. Statist. Assoc.*, 8:9–25, 1993.

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

[3] M. Heath, S. Sarkar, T. Sanocki, and K. Bowyer. A robust visual method for assessing the relative performance of edge detection algorithms. *PAMI*, 19(12):1338–1359, December 1997.

[4] J. Ross Beveridge. The Geometry of LDA and PCA Classifiers Illustrated with 3D Examples. Technical Report CS-01-101, Computer Science, Colorado State University, 2001.

[5] M. Kirby. *Dimensionality Reduction and Pattern Analysis: An Empirical Approach*. Wiley, 2000.

[6] L. R. LaMotte. Quadratic estimation of variance components. *Biometrics*, 29:311–330, 1973.

[7] 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.

[8] P. Phillips, H. Moon, S. Rizvi, and P. Rauss. The FERET Evaluation Methodology for Face-Recognition Algorithms. *T-PAMI*, 22(10):1090–1104, October 2000.

[9] C. R. Rao. Estimation of variance and covariance components in linear models. *J. Amer. Statist. Assoc.*, 67:112–115, 1972.

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

[11] R. Wolfinger, R. Tobias, and J. Sall. Computing gaussian likelihoods and their derivatives for general linear mixed models. *SIAM Journal of Scientific Computing*, 15:1294–1310, 1994.

[12] W. Zhao, R. Chellappa, and A. Krishnaswamy. Discriminant analysis of principal components for face recognition. In *In Wechsler, Philips, Bruce, Fogelman-Soulie, and Huang, editors, Face Recognition: From Theory to Applications*, pages 73–85, 1998.