Incorporating invariants in Mahalanobis distance based classifiers: Application to Face Recognition

Andrew M. Fraser
Portland State University and Los Alamos National Laboratory
Nicolas W. Hengartner, Kevin R. Vixie, and Brendt E. Wohlberg
Los Alamos National Laboratory
Los Alamos, NM 87545
USA

***Abstract*— We present a technique for combining prior knowledge about transformations that should be ignored with a covariance matrix estimated from training data to make an improved Mahalanobis distance classifier. Modern classification problems often involve objects represented by high-dimensional vectors or images (for example, sampled speech or human faces). The complex statistical structure of these representations is often difficult to infer from the relatively limited training data sets that are available in practice. Thus, we wish to efficiently utilize any available *a priori* information, such as transformations of the representations with respect to which the associated objects are known to retain the same classification (for example, spatial shifts of an image of a handwritten digit do not alter the identity of the digit). These transformations, which are often relatively simple in the space of the underlying objects, are usually nonlinear in the space of the object representation, making their inclusion within the framework of a standard statistical classifier difficult. Motivated by prior work of Simard et al., we have constructed a new classifier which combines statistical information from training data and linear approximations to known invariance transformations. When tested on a face recognition task, performance was found to exceed by a significant margin that of the best algorithm in a reference software distribution.**

I. Introduction

The task of identifying objects and features from image data is central in many active research fields. In this paper we address the inherent problem that a single object may give rise to many possible images, depending on factors such as the lighting conditions, the pose of the object, and its location and orientation relative to the camera. Classification should be invariant with respect to changes in such parameters, but recent empirical studies [1] have shown that the variation in the images produced from these sources for a single object are often of the same order of magnitude as the variation between different objects.

Inspired by the work of Simard et al. [2] [3], we think of each object as generating a low dimensional manifold in image space by a group of transformations corresponding to changes in position, orientation, lighting, etc. If the functional form the transformation group is known, we could in principle calculate the entire manifold associated with a given object from a single image of it. Classification based on the entire manifold, instead of a single point leads to procedures that will be invariant to changes in instances from that group of transformations. The procedures we describe here approximate such a classification of equivalence classes of images. They are quite general and we expect them to be useful in the many contexts outside of face recognition and image processing where the problem of transformations to which classification should be invariant occur. For example, they provide a framework for classifying near field sonar signals by incorporating Doppler effects in an invariant manner. Although the procedures are general, in the remainder of the paper, we will use the terms *faces* or *objects* and *image classification* for concreteness.

Of course there are difficulties. The set of manifolds in image space from all possible objects does not fill the image space, and thus does not properly partition it into equivalence classes of images. Since the manifolds are highly nonlinear, finding the manifold to which a new point belongs is computationally expensive. For noisy data, the computational problem is further compounded with the uncertainty in the assigned manifold.

To address these problems, we use tangents to the manifolds at selected points in image space. Using first and second derivatives of the transformations, our procedures provide substantial improvements to current image classification methods.

II. Combining Within Class Covariances and Linear Approximations to Invariances

Here we outline our approach. For a more detailed development, see [4]. We start with the standard Mahalanobis distance classifier

$$\hat{k}(Y) = \underset{k}{\operatorname{argmin}} (Y-\mu_k)^T C_w^{-1} (Y-\mu_k),$$

where C_w is the within class covariance for all of the classes, μ_k is the mean for class k, and Y is the image to be classified. We incorporate the known invariances while retaining this classifier structure by augmenting the within class covariance C_w to obtain class specific covariances, C_k for each class k. We design the augmentations to allow excursions in directions tangent to the manifold generated by the transformations to which the classifier should be invariant. We have sketched a geometrical view of our approach in Fig. 1.

Denote the transformations with respect to which invariance is desired by $\tau(Y,\theta)$, where $Y \in \mathcal{Y}$ and $\theta \in \Theta$ are the image and transform parameters respectively. The second order Taylor series for the transformation is

$$\tau(Y,\theta) = \tau(Y,0) + V\theta + \theta^T H \theta + R,$$

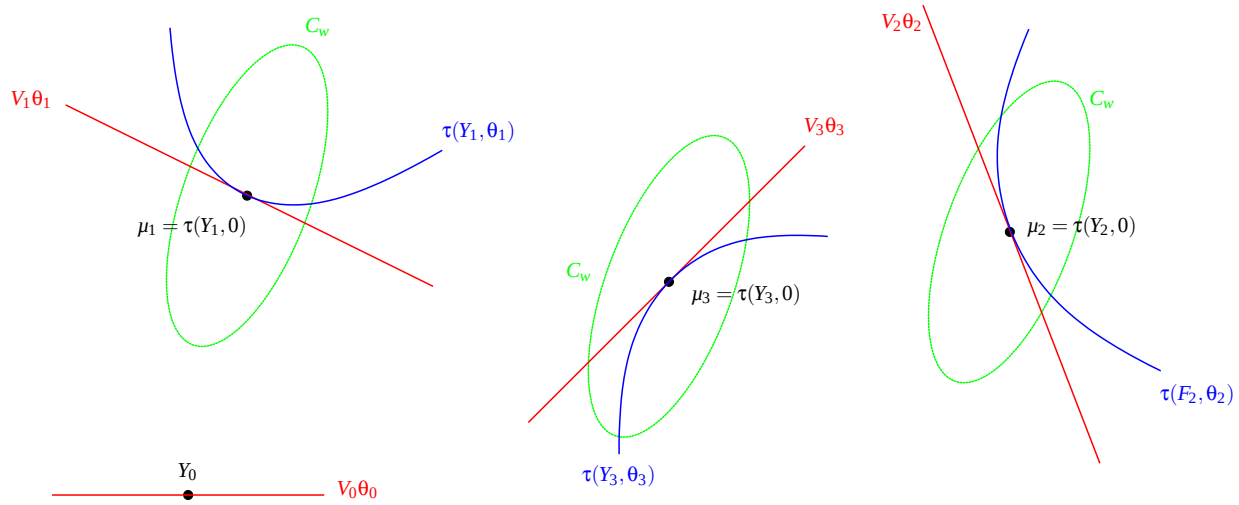

Fig. 1. A geometrical view of classification with augmented covariance matrices: The dots represent the centers μ_k about which approximations are made, the curves represent the true invariant manifolds, the straight lines represent tangents to the manifolds, and the ellipses represent the pooled within class covariance C_w estimated from the data. A new observation Y is assigned to a class $k \in \{1,2,3\}$ using $\hat{k}(Y) = \text{argmin}_k (Y-\mu_k)^T C_k^{-1} (Y-\mu_k)$. The novel aspect is our calculation of $C_k = C_w + \alpha \tilde{C}_k$ where α is a parameter corresponding to a Lagrange multiplier, and $\tilde{C}_k$ is a function of the tangent and curvature of the manifold (from the first and second derivatives respectively) with weighting of directions according to relevance estimated by diagonalizing C_w.

where R is the remainder,

$$(V_k)_i = \left.\frac{\partial \tau(Y_k,\theta)}{\partial \theta_i}\right|_{\theta=0}, \text{ and } (H_k)_{i,j} = \left.\frac{\partial^2 \tau(Y,\theta)}{\partial \theta_i \partial \theta_j}\right|_{\theta=0}.$$

We define

$$C_k = C_w + \alpha V_k C_{\theta,k} V_k^T, \tag{1}$$

Where $C_{\theta,k}$ is a $\dim(\Theta) \times \dim(\Theta)$ matrix. We require that $C_{\theta,k}$ be non-negative definite. Consequently $V_k C_{\theta,k} V_k^T$ is also non-negative definite. When C_k^{-1} is used as a metric, the effect of the term $V_k C_{\theta,k} V_k^T$ is to discount displacement components in the subspace spanned by V_k, and the degree of the discount is controlled by $C_{\theta,k}$. We developed [4] our treatment of $C_{\theta,k}$ by thinking of θ as having a Gaussian distribution and calculating expected values with respect to its distribution. Here we present some of that treatment, minimizing the probabilistic interpretation. Roughly, $C_{\theta,k}$ characterizes the costs of excursions of θ. We choose $C_{\theta,k}$ to balance the conflicting goals

Big: We want to allow θ to be large so that we can classify images with large displacements in the invariant directions.

Small: We want $\theta^T H \theta \in \mathcal{Y}$ to be small so that the truncated Taylor series will be a good approximation.

We search for a resolution of these conflicting goals in terms of a norm on θ and the covariance $C_{\theta,k}$. For the remainder of this section let us consider a single individual k and drop the extra subscript, i.e., we will denote the covariance of θ for this individual by C_θ.

If, for a particular image component d, the Hessian H_d has both a positive eigenvalue λ_1 and a negative eigenvalue λ_2, then the quadratic term $\theta^T H \theta$ is zero along a direction e_0 which is a linear combination of the corresponding eigenvectors, i.e. $(\gamma e_0)^T H_d (\gamma e_0) = 0 \ \forall \gamma$. We suspect that higher order terms will contribute to significant errors when $\gamma \geq \min\left(|\lambda_1|^{\frac{1}{2}}, |\lambda_2|^{\frac{1}{2}}\right)$, so we eliminate the canceling effect by replacing H_d with its *positive square root*, i.e. if an eigenvalue λ of H_d is negative, replace it with $-\lambda$. This suggests the following *mean root square* norm

$$|\theta|_{H_{\text{mrs}}} \equiv \sqrt{\sum_{d=1}^{N} \theta^T \sqrt{H_d H_d}\, \theta}. \tag{2}$$

Consider the following objection to the norm in Eqn. (2). If there is an image component d which is unimportant for recognition and for which H_d is large, e.g. a sharp boundary in the background, then requiring $|\theta|_{H_{\text{mrs}}}$ to be small might prevent parameter excursions that would only disrupt the background. To address this objection, we use the eigenvalues of the pooled within class covariance matrix C_w to quantify the importance of the components. If there is a large within class variance in the direction of component d, we will not curtail particular parameter excursions just because they cause errors in component d.

We develop our formula for C_θ in terms of the eigendecomposition

$$C_w = \sum_d e_d \lambda_d e_d^T$$

as follows. Break the $\dim(\Theta) \times \dim(\mathcal{Y}) \times \dim(\Theta)$ tensor H into components

$$H_d \equiv e_d^T H. \tag{3}$$

Then for each component, define the $\dim(\Theta) \times \dim(\Theta)$ matrix

$$H_d^+ \equiv \sqrt{(H_d)^T H_d}, \tag{4}$$

and take the average to get

$$\bar{H} \equiv \sum_d H_d^+ |\lambda_d|^{-\frac{1}{2}}. \tag{5}$$

Define the norm

$$|\theta|_{\bar{H}} \equiv \sqrt{\theta^T \bar{H} \theta}.$$

Given H and C_w, one can calculate $\bar{H}$ using Equations (3), (4), and (5). Then by using the determinant $|C_\theta|$ to quantify goal **Big:** (allow θ to be large) and using $\mathbb{E}\,|\theta|^2_{\bar{H}}$ to quantify goal **Small:** (keep $\theta^T H \theta \in \mathcal{Y}$ small), we get the constrained optimization problem:

Maximize the determinant $|C_\theta|$
Subject to

$$\mathbb{E}\,|\theta|^2_{\bar{H}} \leq \gamma, \tag{6}$$

where γ is a constant.

The solution to the problem is

$$C_\theta = \alpha (\bar{H})^{-1}, \tag{7}$$

where α, which is a function of γ, is a constant that balances the competing goals.

To verify that Eqn. (7) indeed solves the optimization problem, note:

$$\begin{aligned}\mathbb{E}\,|\theta|^2_{\bar{H}} &= \mathbb{E}\left(\sum_{k,l} \theta_k \bar{H}_{k,l} \theta_l\right) \\ &= \sum_{k,l} \bar{H}_{k,l} \mathbb{E}\,(\theta_k \theta_l) \\ &= \mathrm{Tr}\,(\bar{H} C_\theta)\,.\end{aligned}$$

In the coordinates that diagonalize $\bar{H}$, Eqn. (6) only constrains the diagonal entries of C_θ. Of the symmetric positive definite matrices with specific diagonal entries, the matrix that has the largest determinant is simply diagonal. So C_θ and $\bar{H}$ must be simultaneously diagonalizable, and the problem reduces to

$$\text{Maximize: } \prod_{l=1}^{\dim(\Theta)} \sigma_l$$

$$\text{Subject to: } \sum_{l=1}^{\dim(\Theta)} \sigma_l h_l = \gamma.$$

The Lagrange multipliers method yields Eqn. (7).

Summary: Given a new image Y, we estimate its class with

$$\hat{k}(Y) = \underset{k}{\mathrm{argmin}} (Y - \mu_k)^T C_k^{-1} (Y - \mu_k),$$

where $C_k = C_w + \alpha V_k C_{\theta,k} V_k^T$. We have derived the parameters of this classifier by synthesizing statistics from training data with analytic knowledge about transformations we wish to ignore.

III. Face Recognition Results

We tested our techniques by applying them to a face recognition task and found that they reduce the error rate by more than 20% (from an error rate of 26.7% to an error rate of 20.6%). We used an analytic expression for transformations in image space and developed procedures for evaluating first and second derivatives of the transformations. The transformations have the following five degrees of freedom:

- Horizontal translation
- Vertical translation
- Horizontal scaling
- Vertical scaling
- Rotation

To implement the test, we relied on the FERET data set [5] and a source code package from Beveridge et al. [6], [7] at CSU for evaluating face recognition algorithms.

Version 4.0 (October 2002) of the CSU package contains source code that implements 13 different face recognition algorithms, scripts for applying those algorithms to images from the FERET data set, and source code for Monte Carlo studies of the distribution of the performance of the recognition algorithms. Following Turk and Pentland [8], all of the CSU algorithms use principal component analysis as a first step. Those with the best recognition rates also follow Zhao *et al.* [9] and use a discriminant analysis. For each algorithm tested, the CSU evaluation procedure reports a distribution of performance levels. The specific task is defined in terms of a single *probe* image and a *gallery* of N_G images. The images in the gallery are photographs of N_G distinct individuals. The gallery contains a single *target* image, which is another photograph of the individual represented in the probe image. Using distances reported by the algorithm under test, the evaluation procedure sorts the gallery into a list, placing the target image as close to the top as it can. The algorithm scores a success at rank n if the target is in the first n entries of the sorted list. The CSU evaluation procedure randomly selects $N_G \times 10,000$ gallery-probe pairs and reports the distribution of successful recognition rates as a function of rank.

Restricting the test data set to those images in the FERET data that satisfy the following criteria:

- Coordinates of the eyes have been measured and are part of the FERET data.
- There are at least four images of each individual.
- The photographs of each individual were taken on at least two separate occasions.

yields a set of 640 images consisting of 160 individuals with 4 images of each individual. Thus we use $N_G = 160$. Of the remaining images for which eye coordinates are given, we used a training set of 591 images consisting of 3 images per individual for 197 individuals. The testing and training images were uniformly preprocessed by code from the CSU package. In [6] the authors describe the preprocessing as,

> "All our FERET imagery has been preprocessed using code originally developed at NIST and used in the FERET evaluations. We have taken this code and converted it ...
> Spatial normalization rotates, translates and scales the images so that the eyes are placed at fixed points in the imagery based on a ground truth file of eye coordinates supplied with the FERET data. The images are cropped to a standard size, 150 by 130 pixels. The NIST code also masks out pixels not lying within an oval shaped face region and scales the pixel data range of each image within the face

region. In the source imagery, grey level values are integers in the range 0 to 255. These pixel values are first histogram equalized and then shifted and scaled such that the mean value of all pixels in the face region is zero and the standard deviation is one."

Each recognition algorithm calculates subspaces and fits parameters using the preprocessed training images and knowledge of the identity of the individuals in the images. Then, using those parameters, each algorithm constructs a matrix consisting of the distances between each pair of images in the testing set of 640 images. Thus, in the training phase, one can calculate the mean image, μ_k, of an individual, but in the testing phase, the algorithm has no information about the identity of the individuals in the images.

We developed three recognition algorithms: the first consists of the general techniques of Section II combined with minor modifications to fit the test task. We developed the second two algorithms after observing that the CSU algorithms based on angular distance perform best (see Fig. 2). In Section II we supposed that we would have several examples of each class, making an estimate of each class mean μ_k plausible, but for the task defined by the CSU evaluation procedure, we must simply provide 640×640 interimage distances.

The most obvious method for fitting our classification approach within this distance-based framework is to define the distance between image Y_k and Y_l as the Mahalanobis distance

$$d_0(Y_k, Y_l) = (Y_k - Y_l)^T C_k^{-1} (Y_k - Y_l).$$

Note, however, that this distance is not symmetric, since the augmented covariance is only relevant to one of the two images. Consequently, the symmetrized distance

$$d_0'(Y_k, Y_l) = \frac{d_0(Y_k, Y_l) + d_0(Y_l, Y_k)}{2}$$

is used for the distance matrix. After observing that of the CSU algorithms, those based on angular distance perform best (see Fig. 2), we developed two additional algorithms. The "Mahalanobis Angle" distance is

$$d_1(Y_k, Y_l) = \frac{Y_k^T C_k^{-1} Y_l}{\sqrt{Y_k^T C_k^{-1} Y_k}\sqrt{Y_l^T C_k^{-1} Y_l}},$$

with symmetrized version

$$d_1'(Y_k, Y_l) = \frac{d_1(Y_k, Y_l) + d_1(Y_l, Y_k)}{2}.$$

Instead of symmetrizing $d_1(Y_k, Y_l)$, we also define the symmetric distance

$$d_2'(Y_k, Y_l) = \frac{Y_k^T A_{kl}^{-1} Y_l}{\sqrt{Y_k^T A_{kl}^{-1} Y_k}\sqrt{Y_l^T A_{kl}^{-1} Y_l}},$$

where

$$A_{kl} = (C_k + C_l)^{-1}.$$

Evaluating each of the first two distances on the test set of 640 images takes about 30 minutes on a 2.2 GHz Pentium III. We found that the second distance performed better than the first. Because we estimated that evaluating the third distance would take about 160 hours, we instead implemented a hybrid, constructed by computing $d_1'(Y_k, Y_l)$ and then computing $d_2'(Y_k, Y_l)$ only for those distance below some threshold (further detail may be found in [4]).

Each of our algorithms operates in a subspace learned from the training data and uses an estimated covariance,

$$C_k = C_w + \alpha V_k \bar{H}_k^{-1} V_k^T,$$

associated with each image Y_k. We list the key ideas here:

- Use the training data (which includes image identities) to calculate raw within-class sample covariances, C_w'. Regularize the raw covariances as follows: (1) Do an eigenvalue-eigenvector decomposition to find $C_w' = Q\Lambda' Q^T$. (2) Sum the eigenvalues, $S = \sum_i \lambda_i'$. (3) Set $C_w = C_w' + \delta S \mathbb{I}$, which has no eigenvalues less than δS.
- Conceptually convolve the test image with a Gaussian kernel that has mean zero and variance

$$\begin{bmatrix} \left(\frac{h-1}{8}\right)^2 & 0 \\ 0 & \left(\frac{h-1}{8}\right)^2 \end{bmatrix},$$

  where h is an adjustable parameter in the code that must be an odd integer. Change variables to transfer differentiation from the image to the kernel. Evaluate the matrices V_k and $\bar{H}_k$ by convolving (using FFT methods) differentiated kernels with the image.

Thus α, δ, and h are three adjustable parameters in the estimate of C_k. We investigated the dependence of the performance on these parameters [4], and chose the values $\alpha = 100$, $h = 11$, and $\delta = 0.0003$. Our experiments indicated that the classification performance was not sensitive to small changes in these choices.

Results are displayed in Fig. 2 and Fig. 3. Each of our algorithms performs better than all of the algorithms in the CSU package.

IV. Conclusions

We have presented techniques for constructing classifiers that combine statistical information from training data with tangent approximations to known transformations, and we demonstrated the techniques by applying them to a face recognition task. The techniques we created are a significant step forward from the work of Simard et al. due to the careful use of the curvature term for the control of the approximation errors implicit in the procedure. For the face recognition task we used a five parameter group of invariant transformations consisting of rotation, shifts, and scalings. On the face test case, a classifier based on our techniques has an error rate more than 20% lower than that of the best algorithm in a reference software distribution.

The improvement we obtained is surprising because our techniques handle rotation, shifts, and scalings, but we also preprocessed the FERET data with a program from CSU that centers, rotates, and scales each image based on measured eye coordinates. While our techniques may compensate for

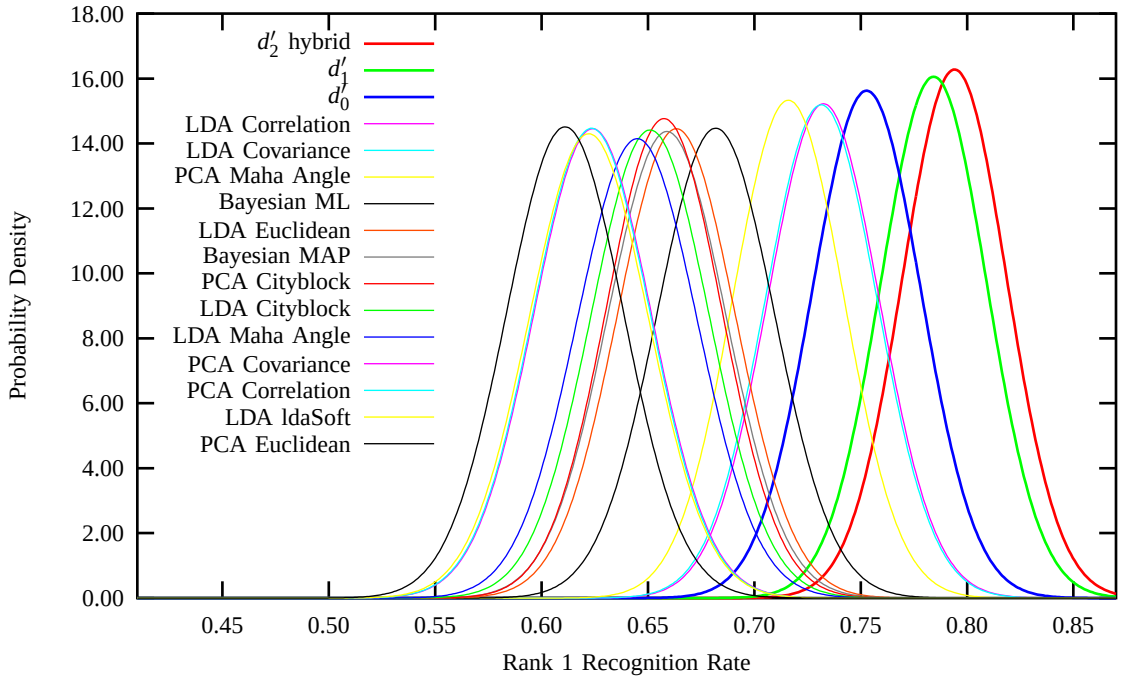

Fig. 2. Approximate distributions for the rank one recognition performance of the algorithms. For each algorithm, a Gaussian is plotted with a mean and variance estimated by a Monte-Carlo study. Note that the key lists the algorithms in order of decreasing mean of the distributions; the first three are the algorithms described in Section III, and the remainder are those implemented in the CSU software distribution.

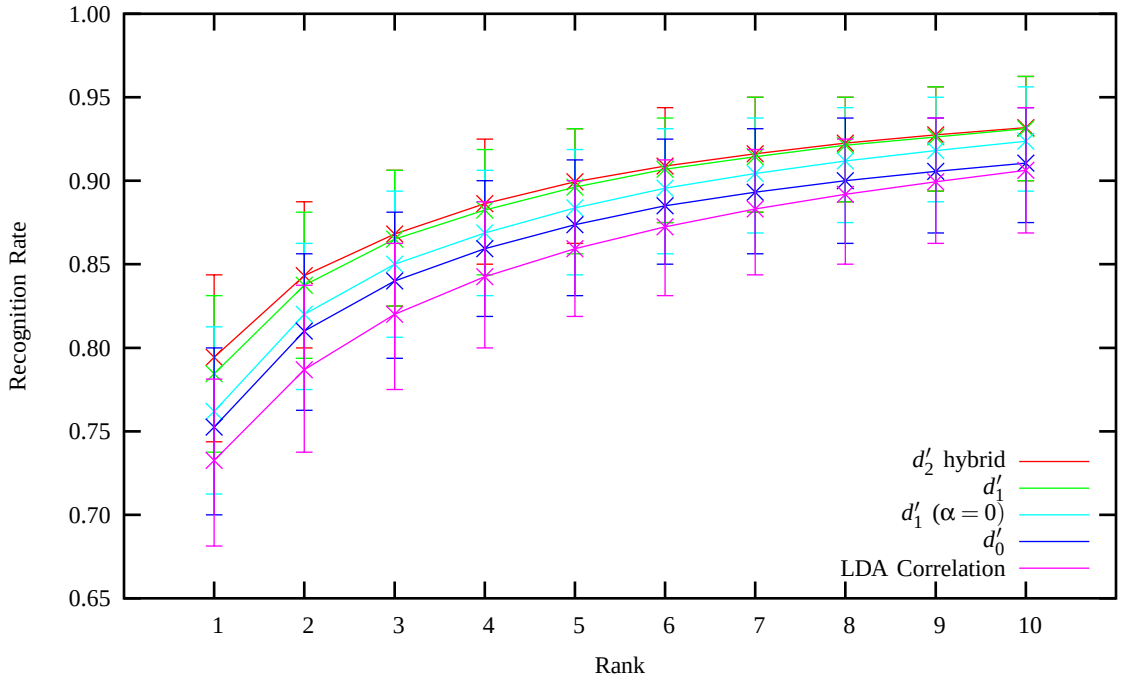

Fig. 3. The mean recognition rate and 95% confidence intervals as a function of rank for the following algorithms: d'_2 hybrid (the hybrid of d'_1 and d'_2), d'_1 (the symmetrized Mahalanobis Angle with tangent augmentation), d'_1 ($\alpha = 0$) (the symmetrized Mahalanobis Angle with no tangent augmentation, illustrating the benefit obtained from the regularization of C'_w), d'_1 (the symmetrized Mahalanobis distance), and LDA Correlation (the best performing algorithm in the CSU distribution).

errors in the measured eye coordinates or weaknesses in the preprocessing algorithms, we suspect that much of the improvement is due to similarities between the transformations we handle and differences between images. For example, a smile is probably something like a dilation in the horizontal direction.

V. Acknowledgment

This work was supported by a LANL 2002 Homeland defense LDRD-ER (PI K. Vixie) and a LANL 2003 LDRD-DR (PI J. Kamm).

References

[1] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 643–660, June 2001.

[2] P. Y. Simard, Y. A. L. Cun, J. S. Denker, and B. Victorri, "Transformation invariance in pattern recognition - tangent distance and tangent propagation," in *Neural Networks: Tricks of the Trade*, G. B. Orr and K.-R. Muller, Eds. Springer, 1998, ch. 12.

[3] P. Y. Simard, Y. A. Cun, J. S. Denker, and B. Victorri, "Transformation invariance in pattern recognition: Tangent distance and propagation," *International Journal of Imaging Systems and Technology*, vol. 11, no. 3, pp. 181–197, 2000.

[4] A. Fraser, N. Hengartner, K. Vixie, and B. Wohlberg, "Classification modulo invariance, with application to face recognition," *Journal of Computational and Graphical Statistics*, 2003, invited paper, in preparation.

[5] P. J. Phillips, H. Moon, P. J. Rauss, and S. Rizvi, "The feret evaluation methodology for face recognition algorithms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 10, Oct. 2000, available as report NISTR 6264.

[6] J. R. Beveridge, K. She, B. Draper, and G. 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*, 2001. [Online]. Available: http://www.cs.colostate.edu/evalfacerec/index.html

[7] R. Beveridge, "Evaluation of face recognition algorithms web site." http://www.cs.colostate.edu/evalfacerec/, Oct. 2002.

[8] M. Turk and A. Pentland, "Face recognition using eigenfaces," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Maui, HI, USA, 1991.

[9] W. Zhao, R. Chellappa, and A. Krishnaswamy, "Discriminant analysis of principal components for face recognition," in *Face Recognition: From Theory to Applications*, Wechsler, Phillips, Bruce, Fogelman-Soulie, and Huang, Eds., 1998, pp. 73–85.