Minimum Risk Distance Measure for Object Recognition

Shyjan Mahamud Martial Hebert
Dept. of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Abstract

Recently, the optimal distance measure for a given object discrimination task under the nearest neighbor framework was derived [1]. For ease of implementation and efficiency considerations, the optimal distance measure was approximated by combining more elementary distance measures defined on simple feature spaces. In this paper, we address two important issues that arise in practice for such an approach: (a) What form should the elementary distance measure in each feature space take? We motivate the need to use optimal distance measures in simple feature spaces as the elementary distance measures; such distance measures have the desirable property that they are invariant to distance-respecting transformations. (b) How do we combine the elementary distance measures? We present the precise statistical assumptions under which a linear logistic model holds exactly. We benchmark our model with three other methods on a challenging face discrimination task and show that our approach is competitive with the state of the art.

1. Introduction

The nearest neighbor or exemplar-based framework for classification is widely used in vision for various classification tasks. The main appeal of the framework derives from the fact that it makes few assumptions about the objects to be classified. Let $S_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ be a training set of measurements x_i and corresponding class labels y_i. Let $d(\cdot, \cdot)$ be a given distance measure. On input query x, the 1-nearest neighbor (NN) rule assigns the class label y' corresponding to the nearest neighbor x' of x in S_n. The classification performance of the nearest neighbor rule obviously depends on the distance measure used.

A wide variety of distance measures have been used for the nearest neighbor rule in the literature. Here, we review some of the most relevant measures in the context of object recognition. Distance measures based on PCA or eigenspaces are perhaps the most popular [14, 8]. The underlying assumption when using PCA is that the measurement data can be explained (modulo noise) by a small dimensional linear subspace of X. This is a generative model that does not take into account how the different classes are distributed. Discriminative analysis, of which LDA [3, 16] is the most popular, on the other hand explicitly tries to find discriminative distance measures that separate the different classes from each other as much as possible. More recently, discriminative distance measures have also been derived from constructing Support Vector Machines [11] that maximize the margin between different classes. Bayesian approaches to classification estimates probability density models for each class and classifies an input query using the Bayes rule. For the two class case, the log-odds ratio can be considered to be a discriminative distance measure. For certain applications, for example face recognition [7], such an approach has been found to be competitive with the state of the art. However, in general, such approaches typically suffer from the need to specify an appropriate model for each class as well as estimating such models reliably from data. Thus in our work, we will work in the discriminative setting that makes as few modeling assumptions as possible about the objects of interest. Such an approach is necessitated especially for multi-class object recognition tasks where the objects of interest can be arbitrary.

Recently, a new criterion for directly finding discriminative distance measures for object recognition under the nearest neighbor framework was proposed [1]. The criterion is based on the optimal distance measure that minimizes the classification risk for the 1-nearest neighbor rule. In contrast to previous approaches, the new criterion allows us to combine discriminative feature spaces of different types (for example, color, texture, local shape) in a principled manner.

From practical considerations, the optimal distance measure was modeled by combining a set of simple elementary distance measures, each of which is defined on a simple feature space. In this paper, we address two important issues that arise in such a scheme: (a) What form should the elementary distance measure in each feature space take? We motivate the need to use optimal distance measures in simple feature spaces as the elementary distance measures; such distance measures have the desirable property that they are invariant to "distance-respecting" transformations. (b) How do we combine the elementary distance measures? We

motivate a linear combination model that can be shown to be exactly valid under certain statistical assumptions.

The rest of the paper is organized as follows: § 2 reviews the derivation given in [1] for the optimal distance measure that minimizes the risk for the 1-nearest neighbor rule. Estimation of this optimal distance measure directly from data leads to a general criterion for finding discriminative distance measures. Section 3 discusses the design and implementation of a model for the optimal distance measure motivated from practical considerations. We benchmark our approach against three other well-known approaches for a challenging face discrimination task in § 5.

2. Optimal 1-NN Distance Measure

Here we briefly review the results from [1]. As in the introduction, assume that we have a training set $S_n = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ where each tuple (x_i, y_i) is chosen i.i.d. from some unknown distribution over $X \times Y$ where X is the space of image measurements and Y is some discrete set of object class labels. We are also given a distance measure $d : X \times X \rightarrow \mathbb{R}$ between any two image measurements.

On input measurement $x \in X$, the 1-*nearest neighbor rule* reports the class label y' associated with the training image $x' \in S_n$ that is closest to x according to the distance measure d. The n-sample NN mis-classification risk $R(n)$ is defined as:

$$R(n) \equiv \mathrm{E}_{(x,y),S_n}[1_{[y \neq y']}] \quad (1)$$

where $1_{[y \neq y']}$ is an indicator function for the event $y \neq y'$. Note that the risk is averaged over all inputs x *as well as* all training sets of size n.

Conditioning on input x, the risk can be re-written as follows:

$$\begin{aligned} R(n) &\equiv \mathrm{E}_{x,X_n}[r(x, x')] \\ r(x, x') &\equiv \mathrm{E}_{y,y'}[1_{[y \neq y']} | x, x'] \\ &= p(y \neq y' | x, x') \end{aligned} \quad (2)$$

where $r(x, x')$ is the conditional risk of assigning input x with the class label corresponding to x', and X_n is the set of all training measurements x_i from S_n. For any given training set size of n, the risk $R(n)$ depends only on the distance measure d used for the nearest neighbor search. Thus, it is natural to ask for the distance measure that minimizes the risk.

Since the conditional risk $r(x_i, x_j) = p(y_i \neq y_j | x_i, x_j)$ is itself a measure defined over any two input measurements $x_i, x_j \in X$, we can consider using it as a candidate distance measure. Under this distance measure, two images are "closer" to each other if they are both likely to come from the same class. We can easily show that this distance measure minimizes the NN risk:

Property 1 (Optimality) *The distance measure* $d(x_i, x_j) \equiv p(y_i \neq y_j | x_i, x_j)$ *minimizes the risk* $R(n)$ *for any* n.

See [2] for a proof.

The optimal distance measure is not a metric distance measure. Of the metric axioms, it somewhat surprisingly satisfies triangle inequality and is of course symmetric since the loss function is symmetric. However, it does not in general satisfy self-similarity, i.e., $d(x, x') = 0$ iff $x = x'$ is not satisfied. In fact, for most real applications one expects some uncertainty, however small, as to which class a measurement belongs to. It can be shown that due to this uncertainty, self-similarity is always violated. See [2] for further details.

Our strategy for finding a discriminative distance measure will be based on modeling and estimating this optimal distance measure from task-dependent training data. Implicitly, such a strategy finds a distance measure that minimizes the NN risk.

An approach that is similar in spirit has been explored before in the literature starting with [13]. However, the risk minimized there is the asymptotic risk R^M when using any *metric* distance measure. Using the non-metric optimal distance measure on the other hand always gives a risk that is better than R^M. The risk can even approach the Bayes optimal risk depending on the task, even when the risk for any metric distance is strictly worse than the Bayes optimal.

3. Modeling the Optimal Distance

Under the i.i.d. assumption the optimal distance measure $p(y_i \neq y_j | x_i, x_j)$ can be expressed in terms of generative models $p(x|y)$ for each class as follows:

$$p(y_i \neq y_j \mid x_i, x_j) = \sum_y p(y|x_i)(1 - p(y|x_j)) \quad (3)$$

Thus one approach [4] is to first estimate a generative model $p(x|y)$ for each class from training data and then construct the optimal distance measure using the expression above. The disadvantage of such an approach is the need to estimate reliably the generative models from data. If we can indeed estimate generative models reliably from data, then we should get better classification performance using the Bayes' decision rule directly. In practice, it is more likely that estimating generative models from data may not be reliable since a good model may require the estimation of many parameters, even though most of them may be irrelevant to the task of discriminating one object from another. Moreover, for a multi-class object discrimination task, formulating a generative model is likely to be difficult in practice for an arbitrary collection of objects of interest.

Our approach instead will be to model the optimal distance *directly* in terms of more elementary distance measures defined on simple to construct feature spaces. Our rationale for such an approach is based on the fact that it is relatively easy and efficient to construct various simple feature spaces from an input image. For example, in the face discrimination task that we consider in § 5, the feature spaces will be various single-dimensional linear projections. More generally, simple discriminative feature spaces can be based on color, texture, local shape properties. From a practical point of view, it would thus be advantageous to approximate the optimal distance measure by combining the discriminative information from such simple-to-construct feature spaces.

Having decided on such a scheme, two issues arise: (a) What are appropriate elementary distance measures? and (b) How do we combine the elementary distance measures for approximating the optimal distance measure.

3.1. Choosing Elementary Distance Measures

Formally, let $\mathcal{C} = \{d_1, d_2, \ldots, d_N\}$ be a possibly large collection of elementary distance measures, each of which is defined on some simple feature space. For run-time considerations, we wish to select $K \ll N$ elementary distance measures $d_k \in \mathcal{C}$ from this collection that best approximate the optimal distance measure $p(y_i \neq y_j | x_i, x_j)$.

What is a good choice for the elementary distance measures? The question is analogous to the issue of choosing an appropriate kernel (which is essentially the inverse of a distance measure) for Support Vector Machines [15].

One desirable property that any sound selection scheme for elementary distance measures should possess is invariance to distance "respecting" transformations. If s and s' are two distance scores under some distance measure d, then f is a distance respecting transformation of d if f does not change distance relationships:

$$s < s' \implies f(s) < f(s')$$

In other words, f has to be some strictly increasing function. Clearly, selecting two elementary distance measures related by some distance respecting transformation should be considered redundant, since intuitively, such transformations cannot give us any new information about the elementary distance measure being transformed.

We now show that one way for ensuring such invariance to such distance respecting transformations is to choose the *optimal* elementary distance measure $p(y_i \neq y_j \mid d)$ given an elementary distance d rather than choosing d itself directly. Intuitively, such a choice induces a partition of the space of all possible elementary distance measures in a feature space into equivalence classes. Two distance measures belong to the same equivalence class if they are related by some distance respecting transformation. We show below that the optimal elementary distance measure $p(y_i \neq y_j \mid d)$ is the same for every distance measure d from a given equivalence class. Different equivalence classes have different optimal elementary distance measures associated with them and the task is to select the best one over all equivalence classes. Thus, potentially, rather than having to search over the space of all elementary distance measures, we only need to search over the space of optimal elementary distance measures.

Property 2 (Invariance) *The optimal distance measure $p(y_i \neq y_j \mid d)$ given an elementary distance d is invariant to distance respecting transformations.*

Proof. Let f be a distance respecting transformation, i.e., f is a strictly increasing function. Let $d' = f(d)$. We then have from Bayes rule:

$$p(y_i \neq y_j \mid d') = \frac{p(d' \mid y_i \neq y_j) p(y_i \neq y_j)}{p(d')}$$

We now use the following identities that relates probability densities when transformed by any strictly increasing function f [9]:

$$p(d' \mid y_i \neq y_j) = \frac{p(d \mid y_i \neq y_j)}{f'(d)}$$

$$p(d') = \frac{p(d)}{f'(d)}$$

where f' is the derivative w.r.t. d. Substituting in the above equation, we get:

$$\begin{aligned} p(y_i \neq y_j \mid d') &= \frac{p(d \mid y_i \neq y_j) p(y_i \neq y_j)}{p(d)} \\ &= p(y_i \neq y_j \mid d) \end{aligned}$$

□

In practice, we can construct the optimal elementary distance measure $d^* \equiv p(y_i \neq y_j \mid d)$ given some elementary distance measure d and training data as follows. From the training data, estimate the intra-class distribution of the distance scores $p(d \mid y_i = y_j)$ as well as the extra-class distribution $p(d \mid y_i \neq y_j)$ and the priors $p(y_i = y_j)$, $p(y_i \neq y_j)$. Then the optimal elementary distance measure d^* is given by using Bayes rule. In our work, we estimate the intra- and extra-class distribution of the distance scores using a Parzen window estimator [6]. This is practical since these distribution are over one-dimensional quantities. We sample the values of the Parzen window estimator at regular fine intervals and store it in a table at training time for fast access at run-time. See Figure 1.

Interestingly, although our motivation for using optimal elementary distance measures is purely for the sake of invariance, we also find that for the experiments on face discrimination that we report in § 5, the recognition performance when using optimal elementary distance measures is also empirically superior compared with using the elementary distance measures directly.

3.2. Linear Combination Model

Next, given a set of K one-dimensional optimal elementary distance measures $\mathbf{d} = [d_1^*, \ldots, d_k^*]$, we want to approximate the optimal distance measure by combining the elementary distance measures. In other words, we want to model and estimate $p(y_i \neq y_j \mid \mathbf{d})$. Here we motivate the following linear logistic model for the optimal distance measure that we use in practice:

$$\log \frac{p(y_i \neq y_j \mid \mathbf{d})}{p(y_i = y_j \mid \mathbf{d})} = \alpha_0 + \sum_k^K \alpha_k \log \frac{p(y_i \neq y_j \mid d_k^*)}{p(y_i = y_j \mid d_k^*)} \tag{4}$$

Below, we develop the statistical assumptions and practical considerations under which such a model holds exactly.

We can model the optimal distance measure by modeling the intra-class and extra-class distributions of the set of one-dimensional optimal elementary distance measures d_k^*, since Bayes rule gives us:

$$\frac{p(y_i \neq y_j \mid \mathbf{d})}{p(y_i = y_j \mid \mathbf{d})} = \frac{p(\mathbf{d} \mid y_i \neq y_j)}{p(\mathbf{d} \mid y_i = y_j)} \frac{p(y_i \neq y_j)}{p(y_i = y_j)}$$

Modeling the intra- and extra-class distributions $p(\mathbf{d} \mid \cdot)$ directly is inconvenient since each d_k^* takes values in $[0, 1]$. We will instead model the distribution of a distance-respecting transform of each d_k^*, namely the logit transform, which transforms the range $[0, 1]$ into the whole real line which is more convenient to model. The logit transform is given by

$$\text{logit}(x) \equiv \log \frac{x}{1-x}, \qquad x \in (0, 1)$$

Redefine $\mathbf{d} = [\text{logit}(d_1^*), \ldots, \text{logit}(d_K^*)]$ to be the vector of logit transforms of the 1-dimensional optimal distance measures. Then we assume that the intra- and extra-class distribution over $\mathbf{d}$ can be modeled well by two Gaussians:

$$\mathbf{d}_{\text{intra}} \sim \mathcal{N}(\mu_{\text{intra}}, \Sigma_{\text{intra}}), \quad \mathbf{d}_{\text{extra}} \sim \mathcal{N}(\mu_{\text{extra}}, \Sigma_{\text{extra}})$$

The linear logistic model (4) for the optimal distance measure can now be motivated by assuming $\Sigma_{\text{intra}} = \Sigma_{\text{extra}} = \Sigma$, resulting in the following parameters for the model:

$$\alpha_0 = \mu_{\text{intra}}^T \Sigma^{-1} \mu_{\text{intra}} - \mu_{\text{extra}}^T \Sigma^{-1} \mu_{\text{extra}} + \log \frac{p(y_i \neq y_j)}{p(y_i = y_j)}$$

$$[\alpha_1, \ldots, \alpha_K]^T = \Sigma^{-1}(\mu_{\text{extra}} - \mu_{\text{intra}})$$

The above modeling assumption is the same as that used for deriving linear discriminants and is primarily motivated by the need to make the estimation of the model parameters computationally tractable, since we need to estimate only O(K) parameters in (4) with the assumption as opposed to estimating O(K^2) parameters without the assumption.

4. Estimation

Having specified a model for the optimal distance measure, in this section we describe a simple maximum likelihood framework for estimating such a model from data.

Let $y_{ij} \equiv 2 1_{[y_i \neq y_j]} - 1$. As before, let $S = \{(x_1, y_1), \ldots, (x_N, y_N)\}$ be the training set of image measurements and corresponding class labels. Let $\mathbf{d} = \{d_1^*, \ldots, d_K^*\}$ be a particular selection of optimal elementary distance measures from $\mathcal{C}$ (see § 3.1). Maximizing the log-likelihood of the optimal distance measure model (4), given the training set, can be shown [2] to be equivalent to minimizing the following cost function:

$$J_{\mathbf{d}}(\boldsymbol{\alpha}) \equiv \sum_{i,j}^N \log \left(1 + e^{-\sum_k \alpha_k y_{ij} \text{ logit } (d_k^*(x_i, x_j))}\right) \tag{5}$$

where, for compactness of notation, we have assumed the inclusion of a constant distance measure $d_0 \equiv 1$. This cost function is convex [2] and can be optimized using standard iterative techniques like Newton's method [12].

Finally, the best choice for $\mathbf{d}$ is the one that maximizes the likelihood or equivalently minimizes $J_{\mathbf{d}}$ over all choices of K optimal elementary distance measures from the collection $\mathcal{C}$. The brute-force search over all choices is clearly unfeasible when K is large. Instead, we adopt a simple greedy strategy in which at each iteration k we choose the best $d_k^* \in \mathcal{C}$ that along with the distance measures $\{d_1^*, \ldots, d_{k-1}^*\}$ chosen in the previous iterations minimizes the cost function.

5. Experiments

In this section, we benchmark our approach against other well-known algorithms on a challenging face recognition task. At the outset, we would like to emphasize that our method is not limited to face recognition. In fact, in contrast to other approaches, our approach allows for the combination of elementary distance measures from disparate feature spaces like color, texture and local shape in a principled manner, see [2]. Here, we choose a face recognition task primarily for benchmarking purposes, since the task has been well-studied in the literature resulting in a large number of algorithms developed for the task as well as well-known standard data-sets.

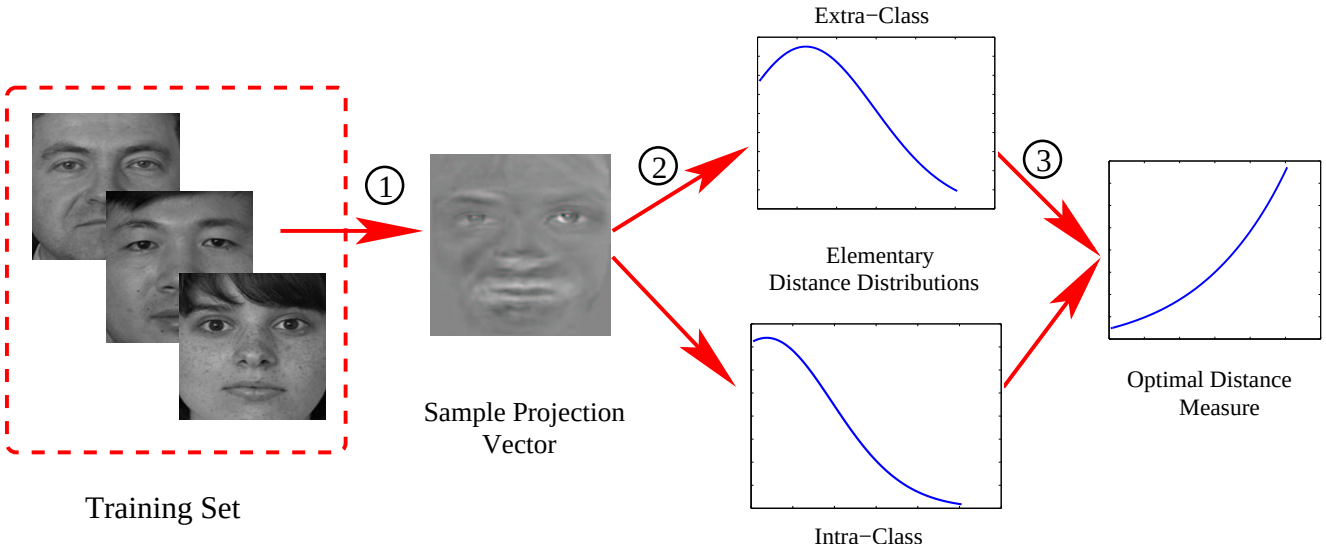

Figure 1: Computing the optimal elementary distance measure: (1) Each training image is projected onto some projection vector. (2) The intra- and extra-class distributions for the elementary distance measure (in our work, the magnitude of the difference between two images along the projection vector) over the training set are estimated using a Parzen window estimator. (3) The optimal elementary distance measure is computed from the intra- and extra-class distributions using Bayes rule. The result is sampled and stored in a table for efficient access at run-time. See text for further details.

5.1. Face Recognition Task

We chose a subset of frontal face images from the FERET [10] database that had varying expressions and some illumination changes. Specifically, we chose a subset corresponding to 200 subjects, each of which had 2 images with varying expression and illumination. This was divided into a training and test set with 100 subjects in each set. None of the subjects in the test set were represented in the training set. The test set was further divided into a gallery set of 100 images and a probe set of 100 images. During testing of each algorithm, the closest match for each probe image from the gallery was found. Figure 2 shows a sample of the selected images.

Following [11], the images were pre-processed as follows. Each of the images were aligned using a similarity transform (rotation, translation and scale) such that the locations of the eyes, whose positions were provided in the FERET database, fell on pre-specified pixel locations. Next, the images were cropped with a common mask to exclude background and hair. The non-masked pixels were then histogram-equalized and the resulting pixels were further processed to have zero mean and unit variance. Finally, all the images were scaled to have a size of 150×200. Figure 3 shows an image before and after pre-processing.

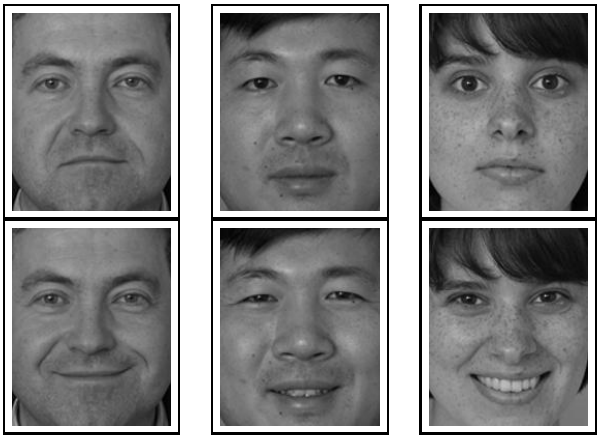

Figure 2: Sample images from the FERET database.

Figure 3: A face image before and after pre-processing.

5.2. Methods

Our Approach. For our approach, we need to specify the collection $\mathcal{C}$ of elementary distance measures from which a

subset of K distance measures are chosen for approximating the optimal distance measure (see § 3). In order to make a fair comparison with the other approaches, the elementary distance measures we consider are distances between two images along various one-dimensional linear projections. Formally, if x is a measurement vector of pixels (after pre-processing as described above) for an image, then an elementary distance measure that we consider takes the form $d_v(x_i, x_j) = |v^T(x_i - x_j)|$, where v is a projection vector. The collection of elementary distances that we consider then corresponds to a collection of densely sampled projection vectors v. Assuming the training set is representative of the testing examples that will be seen, we only consider projection directions within the "face"-space spanned by the training images, since the orthogonal space does not contain any discriminative information. For this experiment, we sample 1000 projection vectors in this face-space.

For each of the sampled directions, optimal elementary distance measures are constructed as detailed in § 3.1 using Parzen window estimators. We use a Gaussian window for the estimators with a variance that is $1/10$th of the maximum distance between any two images along any of the sampled projection directions. Given K, the greedy scheme detailed in § 4 is used to approximate the optimal distance measure by the best K optimal elementary distance measures.

PCA. The principal component vectors of the training images are extracted. We use the Mahalanobis distance in PCA space to compensate for the natural scales of each principal component direction. Specifically, the distance measure used is:

$$d_{\mathrm{PCA}}(x_i, x_j) = \sum_k^K \frac{1}{\lambda_k} |u_k^T(x_i - x_j)|^2$$

where $u_k, k = 1, \ldots, K$ correspond to the K principal component directions corresponding to the top eigenvalues λ_k.

LDA. We implemented the well-motivated "soft-LDA" scheme in [16]. First, the training images are projected onto the most significant principal components (in our case, the ones that capture at least 99% of the signal energy). The intra-class S_{intra} and extra-class S_{extra} scatter matrices are then created in PCA space. LDA components are found by maximizing the Fisher ratio or equivalently solving the eigen-problem $S_{\mathrm{extra}}L = S_{\mathrm{intra}}L\Lambda$ where L is the matrix of column vectors of the LDA components. To avoid singularities, it is crucial to regularize S_{intra} by adding a small quantity along the diagonal. Without this regularization, LDA performs very poorly. For robustness, a distance measure that used "soft"-weights based on the eigenvalues is suggested by [16]:

$$d_{\mathrm{LDA}}(x_i, x_j) = \sum_k^K \lambda_k^\alpha |l_k^T(x_i - x_j)|^2$$

where l_k are the LDA components and λ_k the corresponding eigenvalues. We use $\alpha = 0.2$ in our experiments.

Bayesian Classifier. We implemented the scheme proposed in [7] as follows. Intra-class and extra-class image differences are each modeled by a Gaussian density based on a decomposition of the image difference space in each case into the space spanned by the first K principal components of the corresponding class of image differences from the training set, as well as a residue space. The distance measure in this case is given by the log-odds ratio between the extra-class and intra-class densities. See [7] for details.

5.3. Results

At run-time, all of the methods above have two steps: (a) projecting the input image onto K vectors. In our approach, the vectors are the projection vectors along which elementary distances are measured. For PCA, they are the principal components. For LDA, they are the linear discriminant directions. For the Bayesian approach, the vectors are the PCA components required for the intra- and extra-class density models. In our experiments, we divide K equally between the two density models for this approach. (b) finding the closest match from the training set using a distance measure. The second step is roughly comparable for all the methods, since each distance measure can be efficiently computed in each case. The first step is proportional to K. In practice, the choice for K should be based on a trade-off between the desired running time performance and the desired recognition performance.

Figure 4 shows the recognition performance on the FERET test set as K varies for each method. PCA performs poorly on this task, showing that the task is challenging. As can be seen, our approach is competitive in performance with the other state of the art approaches.

Figure 5 compares the performance of our model when using optimal elementary distance measures (labeled "invariant" in the figure, same as the one labeled "Our Model" in Figure 4), which is invariant to distance-respecting transformations, and when using elementary distance measures directly. Although our motivation for using optimal elementary distance measures was purely for the sake of invariance (see § 3.1), we can see from the figure that using optimal elementary distance measures is also empirically superior compared with using the elementary distance measures directly.

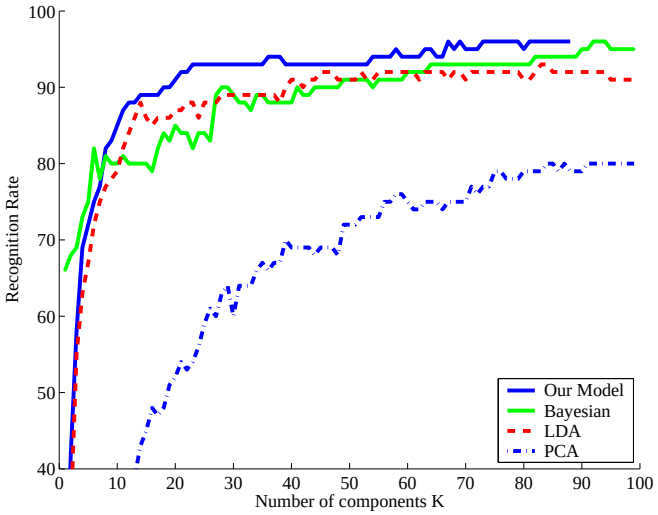

Figure 4: Recognition performance for the various methods as K increases.

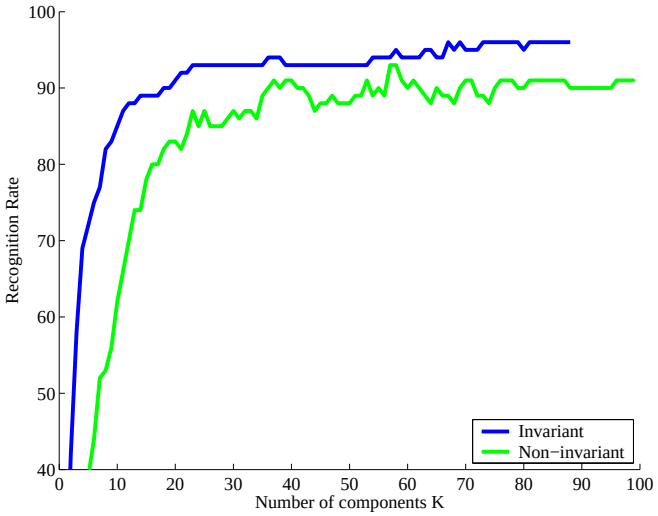

Figure 5: Performance comparison of our model when using optimal elementary distance measures (labeled "invariant"), which is invariant to distance-respecting transformations, and when using elementary distance measures directly. See text.

6. Discussion

We have addressed two issues that arise when modeling the optimal distance measure as a combination of more elementary distance measures. First, we motivated the need for using elementary distance measures that are invariant to distance-respecting transformations. Second, we presented a simple linear combination model that can be justified exactly under certain assumptions. Finally, we have shown that our approach is competitive with the best methods for a face discrimination task.

In contrast to the other approaches compared in this paper, our approach is more widely applicable, since it can also combine multiple modalities (color, texture, shape) in a principled manner within the same framework. We simply need to define appropriate elementary distance measures in each of these feature spaces. This is currently ongoing work.

In § 5, we made the observation that using the invariant elementary distance measures leads to better empirical performance when compared with using the elementary distance measures directly. In future, we would like to investigate both theoretically and empirically whether this observation holds more generally across more tasks.

References

[1] Same Author. In *Conference paper*, 2002.

[2] Same Author. PhD thesis, 2002.

[3] P.N. Belhumeur, J.P. Hespanha, and D.J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19:711–720, 1997.

[4] Enrico Blanzieri and Francesco Ricci. A minimum risk metric for nearest neighbor classification. In *Proc. 16th International Conf. on Machine Learning*, pages 22–31. Morgan Kaufmann, San Francisco, CA, 1999.

[5] T.M. Cover and P.E. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13:21–27, January 1967.

[6] R.O. Duda, P.E. Hart, and D.G. Stork. *Pattern Classification*. Wiley, 2001.

[7] B. Moghaddam and A. Pentland. Beyond eigenfaces: Probabilistic matching for face recognition. In *Intl. Conf. on Automatic Face and Gensture Recogntition*, April 1998.

[8] H. Murase and S.K. Nayar. Detection of 3d objects in cluttered scenes using hierarchical eigenspace. *PRL*, 18(4):375–384, April 1997.

[9] A. Papoulis. *Probability, Random Variables, and Stochastic Processes*. McGraw–Hill, 1965.

[10] P. Phillips, H. Moon, P. Rauss, and S. Rizvi. The feret evaluation methodology for face-recognition algorithms. In *CVPR*, pages 137–143, 1997.

[11] P. J. Phillips. Support vector machines applied to face recognition. In *Neural Information Processing Systems 11*, pages 803–809, 1999.

[12] W.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery. *Numerical Recipes in C*. Cambridge University Press, 1992.

[13] R. Short and K. Fukanaga. The optimal distance measure for nearest neighbor classification. *IT*, 27:622–627, 1981.

[14] M. Turk and A.P. Pentland. Eigenfaces for recognition. *CogNeuro*, 3(1):71–96, 1991.

[15] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1999.

[16] W. Zhao, R. Chellappa, and P. Phillips. Subspace linear discriminant analysis for face recognition. Technical Report Technical Report CAR-TR-914, Center for Automation Research, University of Maryland, College Park, 1999.