Face Recognition Based on Image Sets

Hakan Cevikalp
Eskisehir Osmangazi University
Meselik Kampusu, 26480, Eskisehir, Turkey
hakan.cevikalp@gmail.com

Bill Triggs
Laboratoire Jean Kuntzmann
B.P. 53, 38041 Grenoble Cedex 9, France
Bill.Triggs@imag.fr

Abstract

We introduce a novel method for face recognition from image sets. In our setting each test and training example is a set of images of an individual's face, not just a single image, so recognition decisions need to be based on comparisons of image sets. Methods for this have two main aspects: the models used to represent the individual image sets; and the similarity metric used to compare the models. Here, we represent images as points in a linear or affine feature space and characterize each image set by a convex geometric region (the affine or convex hull) spanned by its feature points. Set dissimilarity is measured by geometric distances (distances of closest approach) between convex models. To reduce the influence of outliers we use robust methods to discard input points that are far from the fitted model. The kernel trick allows the approach to be extended to implicit feature mappings, thus handling complex and nonlinear manifolds of face images. Experiments on two public face datasets show that our proposed methods outperform a number of existing state-of-the-art ones.

1. Introduction

Face recognition has traditionally been posed as the problem of identifying a face from a single image, and many methods assume that images are taken in controlled environments. However facial appearance changes dramatically under variations in pose, illumination, expression, *etc*., and images captured under controlled conditions may not suffice for reliable recognition under the more varied conditions that occur in real surveillance and video retrieval applications.

Recently there has been growing interest in face recognition from sets of images [11, 27, 9, 22, 10, 26, 2]. Here, rather than supplying a single query image, the user supplies a set of images of the same unknown individual. In general the gallery also contains a set of images for each known individual, so the system must recover the individual whose gallery set is the best match for the given query set. The query and gallery sets may contain large variations in pose, illumination, and scale. For example, even if the images were taken on the same occasion they may come from different viewpoints or from face tracking in surveillance video over several minutes.

Methods based on image sets are expected to give better performance than ones based on individual images both because they incorporate information about the variability of the individual's appearance and because they allow the decision process to be based on comparisons of the most similar pairs of query and gallery images – or on local models based on these. In many applications (*e.g*. surveillance systems incorporating tracking or correspondence between different cameras), image sets are also the most natural form of input to the system.

In this paper we develop a general geometric approach to set based face recognition that is particularly suited to cases where the sets contain a relatively wide range of images (*e.g*. ones collected over an extended period of time or under many different conditions). We represent each image as a feature vector in some linear or affine feature space, then for each image set we build a simple convex approximation to the region of feature space that is occupied by the set's feature vectors. Many convex models are possible but here we give results for the affine hull (affine subspace) and the convex hull of the set's feature space points.

To compare different sets, we use the geometric distances (distances of closest approach) between their convex models. This is a reasonable strategy because although sets from the same individual taken under different conditions are unlikely to overlap everywhere, they are more likely to lie close to one another at at least some points. Indeed, to the extent that it is permissible to synthesize new examples within each set by arbitrary linear or convex combinations of feature vectors, finding the distance between the sets corresponds to synthesizing the closest pair of examples, one from each set, then finding the distance between them. This

This work was supported in part by the Young Scientists Award Programme (TÜBA-GEBİP/2010-11) of the Turkish Academy of Sciences.

can be viewed as an approximate method of handling differences in lighting, viewpoint, scale, *etc*.

Classification methods based on convex models of sets of feature vectors have been used in a number of other contexts (*e.g*. [5,6]). Even though they only provide coarse geometric bounds for their underlying point sets – insensitive, *e.g*., to details of the distribution of the points within the convex set – this is still a reasonable strategy for classification with high-dimensional descriptors because in any case fine details of geometry or distribution can not be resolved with practical numbers of samples in high dimensions [14,16]. *C.f*. the successes of affine methods such as linear SVM in many high-dimensional vision problems. Secondly, like SVM, the kernel trick can be used to extend such methods to nonlinear ones capable of handling, *e.g*. nonlinear manifolds of facial appearances. Finally, to reduce the influence of outliers and noise, robust methods can be used to estimate the convex models.

1.1. Related Work

Existing classification methods using image sets differ in the ways in which they model the sets and compute distances between them. Fitzgibbon and Zisserman [10] (see also [3]) use image sets to recognize the principal characters in movies. They model faces detected in contiguous frames as affine subspaces in feature space, use Joint Manifold Distance (JMD) to measure distances between these, then apply a JMD-based clustering algorithm to discover the principal cast of the movie. Another approach [22,2] is to fit a parametric distribution function to each image set, then use Kullback-Leibler divergence to measure the similarity between the distributions. However as noted in [26], these methods must solve a difficult parameter estimation problem, they are not very robust when the test sets have only weak statistical relationships to the training ones, and large set sizes may be needed to approximate the distribution functions accurately.

Yamaguchi *et al*. [27] developed a mutual subspace based method in which image sets are modeled using linear subspaces and similarities between subspaces are measured using the canonical angles between them. Fukui and Yamaguchi [11] extended this approach to include a prior projection onto a more discriminative subspace. A basic limitation of these methods is that they incorporate only relatively weak information (linear subspace angles) about the locations of the samples in input space: for many feature sets, models based on affine subspaces are more discriminative than linear subspace based ones.

The above methods approximate image sets with linear or affine subspaces. There are also many methods that seek to build nonlinear approximations of the manifold of face appearances, typically embedding local linearity within a globally nonlinear model. This idea has been used widely in both descriptor dimensionality reduction and single-image face recognition [21,15,18]. Recently, [9,26] used approaches of this kind for image set based face recognition. Fan and Yeung [9] use hierarchical clustering to discover local structures, approximate each local structure with a linear (not affine) subspace, quantify similarities between subspaces using canonical angles, and finally measure similarities between face image sets by combining these local similarities using majority voting. Wang *et al*. [26] follow a similar approach, using nearest neighbor clustering to find the local structures forming the nonlinear manifold. They again use linear (not affine) subspaces and canonical angles, but also incorporate distances between the centers of the clusters into the similarity metric between local structures. Both of the above works were inspired by the nonlinear manifold modelling approach of Roweis and Saul [21], but they replace the locally affine / distance-based models used in [21] with locally linear / angle-based ones. For many feature sets, we believe that this reduces discrimination. Hadid and Pietikainen [13] also use local linearity to approximate nonlinear face manifolds. They reduce the dimensionality using Locally Linear Embedding, apply k-means clustering, represent the local patches using the resulting cluster centers as exemplars, and measure similarities between image sets by combining the pairwise distances between their exemplars.

In contrast to the above methods, we approximate each image set with a simple convex model – the affine or convex hull of the feature vectors of the set's images. Both approaches can be seen as enhancements of nearest neighbor classification [24,20,5] that attempt to reduce its sensitivity to random variations in sample placement by "filling in the gaps" around the examples. Although still based on the closest-point idea, they replace point-to-point or point-to-model comparisons with training-model to test-model ones. As we will see below, they have a number of attractive properties: the model for each individual can be fitted independently; computing distances between models is straightforward due to convexity; resistance to outliers can be incorporated by using robust fitting to estimate the convex models; and if desired they can be kernelized to produce more local nonlinear models. Moreover, as the experiments below show and despite their intrinsic simplicity, they provide more accurate classification than state-of-the-art methods that build nonlinear approximations to face manifolds by combining local linear models.

2. Method

Let the face image samples be $\mathbf{x}_{ci} \in \mathbb{R}^d$ where $c = 1, \ldots, C$ indexes the C image sets (individuals) and $i = 1, \ldots, n_c$ indexes the n_c samples of image set c. Our method approximates each gallery and test image set with a convex model – either an affine or a convex hull – then uses distances between such models to assign class labels. A test

individual is assigned to the gallery member whose image set's model is closest to the test individual's one.

Given convex sets H and H', the distance between them is the infimum of the distances between any point in H and any point in H' :

$$D(H, H') \;=\; \min_{\mathbf{x}\in H,\, \mathbf{y}\in H'} ||\mathbf{x}-\mathbf{y}||. \tag{1}$$

To implement this we need to introduce parametric forms for the points in H and H' and explicitly minimize the inter-point distance using mathematical programming.

2.1. Affine Hull Method

First consider the case where image sets are approximated by the affine hulls of their training samples, *i.e*., the smallest affine subspaces containing them:

$$H_c^{\text{aff}} \;=\; \left\{ \mathbf{x} = \sum_{k=1}^{n_c} \alpha_{ck}\mathbf{x}_{ck} \;\middle|\; \sum_{k=1}^{n_c} \alpha_{ck} = 1 \right\},\; c = 1, \ldots, C. \tag{2}$$

The affine model implicitly treats any affine combination of a person's face descriptor vectors as a valid face descriptor for him. This typically gives a rather loose approximation to the data, and one that is insensitive to the positions of the samples within the affine subspace.

To parametrize the affine hull, we can choose any point $\boldsymbol{\mu}_c$ on it as a reference (*e.g*. one of the samples $\mathbf{x}_{ck}$, or their mean $\frac{1}{n_c}\sum_{k=1}^{n_c}\mathbf{x}_{ck}$), and rewrite the hull as

$$H_c^{\text{aff}} \;=\; \left\{ \mathbf{x} = \boldsymbol{\mu}_c + \mathbf{U}_c\mathbf{v}_c \;\middle|\; \mathbf{v}_c \in \mathbb{R}^l \right\}. \tag{3}$$

Here, $\mathbf{U}_c$ is an orthonormal basis for the directions spanned by the affine subspace and $\mathbf{v}_c$ is a vector of free parameters that provides reduced coordinates for the points within the subspace, expressed with respect to the basis $\mathbf{U}_c$. Numerically, $\mathbf{U}_c$ is obtained by applying the thin Singular Value Decomposition (SVD) to $[\mathbf{x}_{c1} - \boldsymbol{\mu}_c, \ldots, \mathbf{x}_{cn_c} - \boldsymbol{\mu}_c]$. We discard directions corresponding to near-zero singular values in order to remove spurious noise dimensions within data. The effective dimension of $\mathbf{U}_c$ and the hull is the number of significantly non-zero singular values.

L2 norm based fitting procedures such as SVD are sensitive to outliers. Hulls can also be estimated using robust procedures such as the L1 norm based estimators of [8, 17]. However we used SVD in the experiments below because it proved adequate for the data sets studied there.

Given two non-intersecting affine hulls $\{\mathbf{U}_i\mathbf{v}_i + \boldsymbol{\mu}_i\}$ and $\{\mathbf{U}_j\mathbf{v}_j + \boldsymbol{\mu}_j\}$, the closest points on them can be found by solving the following optimization problem

$$\min_{\mathbf{v}_i, \mathbf{v}_j} ||(\mathbf{U}_i\mathbf{v}_i + \boldsymbol{\mu}_i) - (\mathbf{U}_j\mathbf{v}_j + \boldsymbol{\mu}_j)||^2. \tag{4}$$

Defining $\mathbf{U} \equiv \begin{pmatrix}\mathbf{U}_i & -\mathbf{U}_j\end{pmatrix}$ and $\mathbf{v} \equiv \begin{pmatrix}\mathbf{v}_i \\ \mathbf{v}_j\end{pmatrix}$, this can be written as a standard least squares problem

$$\min_{\mathbf{v}} \; ||\mathbf{U}\,\mathbf{v} - (\boldsymbol{\mu}_j - \boldsymbol{\mu}_i)||^2, \tag{5}$$

whose solution is $\mathbf{v} = (\mathbf{U}^\top\mathbf{U})^{-1}\mathbf{U}^\top(\boldsymbol{\mu}_j - \boldsymbol{\mu}_i)$. It follows that the distance between the hulls can be written as

$$D(H_i^{\text{aff}}, H_j^{\text{aff}}) = ||(\mathbf{I} - \mathbf{P})(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)|| \tag{6}$$

where $\mathbf{P} = \mathbf{U}(\mathbf{U}^\top\mathbf{U})^{-1}\mathbf{U}^\top$ is the orthogonal projection onto the joint span of the directions contained in the two subspaces and $\mathbf{I} - \mathbf{P}$ is the corresponding projection onto the orthogonal complement of this span. If the matrix $\mathbf{U}^\top\mathbf{U}$ is not invertible, $\mathbf{P}$ can be computed as $\tilde{\mathbf{U}}\,\tilde{\mathbf{U}}^\top$, where $\tilde{\mathbf{U}}$ is an orthonormal basis for $\mathbf{U}$ obtained using thin SVD.

2.2. Reduced Affine Hull Method

The above formulation fails if several gallery hulls intersect the given test one, because the test class will have distance zero to each of the corresponding gallery classes. This can occur for several reasons. Firstly, if there are outliers (incorrect or very poor images) in any of the image sets, the corresponding affine hulls may be over-large. The solution in this case is to use a more robust hull fitting procedure or some other form of outlier removal. Secondly, if the feature set is too weak, the features may not suffice to linearly separate the candidates. In this case one can either use more discriminative features, or possibly kernelize the method to make the corresponding decision rules more local and nonlinear. Thirdly, if the affine hulls overlap but the underlying image sets do not, affine approximations may be too loose to give good discrimination and it may be preferable to use a tighter convex approximation. The convex hull of the samples is the tightest convex model containing the samples, but unless the number of samples is exponential in their effective dimension it is typically a significant underestimate of the region spanned by the class. Here we develop a parametric family that includes both affine and convex hulls and many models intermediate between them. The approach is based on constraining the coefficients that can be used to form affine combinations – *c.f*. the reduced convex hulls of [4].

To produce our reduced affine hulls we introduce lower and upper bounds L, U on the allowable α coefficients in (2) to control the looseness of the convex approximation:

$$H_c^{\text{raff}} \;=\; \left\{ \mathbf{x} = \sum_{k=1}^{n_c} \alpha_{ck}\mathbf{x}_{ck} \;\middle|\; \sum_{k=1}^{n_c} \alpha_{ck} = 1, L \le \alpha_{ck} \le U \right\} \tag{7}$$

In the full affine case the bounds are inactive, $(L, U) = (-\infty, \infty)$. In the convex hull case, $L = 0$ and $U \ge 1$ is irrelevant. If $L = 0$ and $U < 1$, several samples need to be

active to ensure $\sum_k \alpha_{ck} = 1$, giving a convex approximation that lies strictly inside the convex hull of the samples. Similarly, if $-\infty < L < 0, U \geq 1$, the region is larger than the convex hull, but smaller than the affine one.

We can write the points of H_c^{raff} more compactly in the form $\{\mathbf{x} = \mathbf{X}_c\, \boldsymbol{\alpha}_c\}$ where $\mathbf{X}_c$ is a matrix whose columns are the feature vectors of set c and $\boldsymbol{\alpha}_c$ is a vector containing the corresponding α_{ck} coefficients. H_c^{raff} is convex because any convex sum of its points, *i.e.* of $\boldsymbol{\alpha}_c$ vectors satisfying the sum 1 and L, U constraints, still satisfies these constraints. For simplicity we apply the same L, U constraints to each α_{ck} coefficient, although this is not strictly necessary.

Given two such reduced affine hulls, the distance between them can be found by solving the following constrained convex optimization problem

$$\begin{aligned} (\boldsymbol{\alpha}_i^*, \boldsymbol{\alpha}_j^*) \;&=\; \underset{\boldsymbol{\alpha}_i,\, \boldsymbol{\alpha}_j}{\arg\min}\; ||\mathbf{X}_i\, \boldsymbol{\alpha}_i - \mathbf{X}_j\, \boldsymbol{\alpha}_j||^2 \\ \sum_{k=1}^{n_i} \alpha_{ik} = 1 = \sum_{k'=1}^{n_j} \alpha_{jk'}, &\quad L \leq \alpha_{ik}, \alpha_{jk'} \leq U \end{aligned} \tag{8}$$

and taking $D(H_i^{\text{raff}}, H_j^{\text{raff}}) = ||\mathbf{X}_i\, \boldsymbol{\alpha}_i^* - \mathbf{X}_j\, \boldsymbol{\alpha}_j^*||$. As before we can write this as a constrained least squares problem $\min \|\mathbf{X}\, \boldsymbol{\alpha}\|^2$ in terms of $\mathbf{X} = \begin{pmatrix} \mathbf{X}_i & -\mathbf{X}_j \end{pmatrix}$ and $\boldsymbol{\alpha} = \left(\begin{smallmatrix} \boldsymbol{\alpha}_i \\ \boldsymbol{\alpha}_j \end{smallmatrix}\right)$, but the constraints are now nonstandard.

The individual examples (feature vectors) $\mathbf{x}_{ck}$ only appear in the quadratic term of (8) so it is easy to kernelize the method by rewriting the quadratic in terms of dot products $\mathbf{x}_{ck}^\top \mathbf{x}_{c'k'}$ and replacing these with kernel evaluations $k(\mathbf{x}_{ck}, \mathbf{x}_{c'k'})$. In the general case there is no reason to expect sparsity so all of the gallery and test points need to be retained in their respective models (although for each given pair of convex models, the corresponding closest point solution *is* usually sparse). However the size of the computation typically remains modest because each class (individual) is fitted separately.

2.3. Convex Hull Approximation

Taking L=0, $U \geq 1$ in (7) approximates the examples with their convex hull (the smallest convex set containing them). As mentioned above, this is much tighter than the affine approximation, but – particularly for small numbers of samples in high dimensions – it can seriously underestimate the true extent of the underlying class, which sometimes leads to false rejections of candidates.

Distances between convex hulls can be found using (8) with L=0 and no U constraint. This problem is closely related to the classical hard margin SVM, which finds a separating hyperplane between the two convex hulls based on exactly the same pair of closest points, but scales its solution differently. Thus – at the cost of SVM training for small problems at run time – one can also find convex hull distances by training an SVM that separates the given test set from the given gallery one, and taking the inter-hull distance to be $2/\|\mathbf{w}\|$ where $\mathbf{w}$ is the SVM weight vector.

Similarly, to handle outliers we can produce an even more restrictive inner approximation by setting $U < 1$, and the resulting problem can be related to the classical soft-margin SVM and the ν-SVM [4].

3. Experiments

We tested[1] the linear and kernelized versions of the proposed methods, AHISD (Affine Hull based Image Set Distance) and CHISD (Convex Hull based Image Set Distance), on two public face recognition data sets: Honda/UCSD [19] and CMU MoBo [12]. These contain several video sequences each from a number of different individuals. Image sets for training and test were constructed by detecting faces in each video sequence using a Viola-Jones face detector [25]. To allow comparison with the literature we followed the simple protocol of [26]: the detected face images were histogram equalized but no further preprocessing such as alignment or background removal was performed on them, and the image features were simple pixel (gray level) values. For CMU MoBo we also tested a Local Binary Pattern (LBP) [1] feature set. For the linear AHISD method, the best separating hyperplane is determined by using affine subspace estimation formulation, and subspace dimensions are set by retaining enough leading eigenvectors to account for 98% of the overall energy in the eigen-decomposition. For the nonlinear AHISD method, we set the bounds as $-L = U = \tau$, where the value of τ is chosen between 1 and 5. The upper bound of nonlinear CHISD method is set to $U = 0.7$ for the Honda/UCSD database, but we used SVM algorithm to compute the distances between convex hulls for CMU MoBo because of speed issues. We set the error penalty term of SVM to $C = 100$ for gray values and to $C = 50$ for LBP features. For all kernelized methods we used the Gaussian kernels.

We compared the proposed linear methods to the Mutual Subspace Method (MSM) [11,27] and the kernelized ones to manifold learning methods that use patchwise local representations [9,13]. Representative examples and linear subspaces were used to model local patches as in [9,13], but instead of Locally Linear Embedding and Isomap based clustering, we used Spectral Clustering to determine samples forming the local patches. In addition to testing locally constant (exemplar) and locally linear subspace (LS) patchwise models, we also tested locally affine (AH) ones to illustrate that the latter are often superior. We used a Gaussian kernel based similarity function for edge weighting during Spectral Clustering and set the number of local patches to 6 for each manifold learning algorithm. To combine the decisions of the local models regarding the label of

[1] For software see *http://www2.ogu.edu.tr/∼mlcv/softwares.html*.

Figure 1. Some detected face images from videos of two subjects from the Honda/UCSD data set.

Figure 2. Some detected face images from videos of two subjects from the MoBo data set.

the test manifold, the majority voting scheme of [9, 13] was used.

3.1. Experiments on the Honda/UCSD Data Set

The Honda/UCSD data set was collected for video-based face recognition. It consists of 59 video sequences involving 20 individuals. Each sequence contains approximately 300–500 frames. Twenty sequences were set aside for training, leaving the remaining 39 for testing. The detected faces were resized to 40×40 gray-scale images and histogram equalized, and the resulting pixel values were used as feature vectors. Some examples are shown in Fig. 1.

Linear Methods	Clean	Noisy G.	Noisy T.	Noisy G+T.
Linear AHISD	97.4	**97.4**	**92.3**	87.2
Linear CHISD	94.9	92.3	**92.3**	82.1
MSM	97.4	**97.4**	87.2	76.9
Nonlinear Methods				
Kernel AHISD	97.4	**97.4**	**92.3**	**92.3**
Kernel CHISD	**100**	**97.4**	**92.3**	82.1
Spec Clus + Exemp.	94.9	89.7	84.6	79.5
Spec Clus + LS	97.4	**97.4**	89.7	79.5
Spec Clus + AH	97.4	94.9	**92.3**	82.1

Table 1. Classification Rates (%) on the Honda/UCSD data set, respectively for the clean data, the data with noisy gallery sets but clean test ones, the data with clean gallery sets and noisy test ones, and the data with noise in both gallery and test sets.

The results are summarized in Table 1. For outlier free image sets (first column), classification is relatively easy and all of the methods tested yielded high recognition rates. To demonstrate the different methods' resistance to outliers, we ran three more experiments in which the training and/or the test sets were systematically corrupted by adding one image from all other classes.

Among the linear methods tested, our AHISD one performed the best in all cases. MSM does well on clean image sets but its performance drops significantly for the corrupted ones, especially when both the training and the test sets are corrupted. Among the nonlinear methods tested, our kernelized AHISD and CHISD ones outperformed the manifold learning ones in most of the cases tested. The kernelized methods also outperform their linear counterparts, especially on the corrupted image sets. Among the manifold learning methods, the one based on exemplars yields the worst accuracies but there is not a clear winner between the locally linear and locally affine subspace based ones. Overall our proposed methods seem to be the best performers, winning in most of the cases tested, particularly on the more corrupted data sets.

3.2. Experiments on the MoBo Data Set

The MoBo (Motion of Body) data set contains 96 image sequences of 24 individuals walking on a treadmill. The images were collected from multiple cameras under four different walking situations: slow walking, fast walking, incline walking, and carrying a ball. Thus, there are 4 image sets for each individual. Each image set includes both frontal and profile views of the subject's faces. Some examples of the detected faces are shown in Fig. 2. As before, the detected faces were converted to 40×40 gray-scale images and histogram equalized, with the resulting pixel values used as features. We also tested a Local Binary Pattern feature set in which each 40×40 image is partitioned into 25 8×8-pixel squares, with a uniform LBP histogram using circular (8,1) neighborhoods being extracted from each square and the resulting histograms being concatenated to produce the final feature vector.

We randomly selected one image set from each class (individual) for the gallery and used the remaining 3 for testing. This was repeated 10 times and we report averages and standard deviations of the resulting classification rate over the 10 runs. Fig. 3 shows the classification rates of each run for the gray level features, and the overall results are shown in Table 2. The asterisks indicate performance differences

Linear Methods	**Gray Level**	**LBP**
Linear AHISD	$92.7^* \pm 3.3$	$94.6^* \pm 2.3$
Linear CHISD	94.2 ± 2.7	**98.1** ± 0.9
MSM	$92.0^* \pm 3.0$	$92.4^* \pm 1.9$
Nonlinear Methods		
Kernel AHISD	93.8 ± 2.8	97.6 ± 1.8
Kernel CHISD	**95.3** ± 2.2	**98.0** ± 1.1
Spec. Clus. + Exemplar	$85.5^* \pm 4.4$	$91.6^* \pm 3.0$
Spec. Clus. + LS	$88.2^* \pm 4.5$	$93.0^* \pm 2.8$
Spec. Clus. + AH	$89.5^* \pm 5.0$	$92.8^* \pm 2.2$

Table 2. Mean classification rates (%) and their standard deviations across the 10 trials on the MoBo data set, for gray level pixel features and LBP features.

that are statistically significant at the 5% level between the given method and the best method for that feature set (indicated in bold).

For the gray level features, our kernelized CHISD method either matches or outperforms all of the others tested in each trial, and overall it significantly outperforms the others. Our linear CHISD method is second best, followed by kernelized AHISD. Among the manifold learning methods, the one based on exemplar images performs the worst, as before, while the locally affine method outperforms the locally linear one. Our methods are significantly more consistent than the manifold learning ones across the different trials. Similar conclusions hold for the LBP features, with the linear and kernelized CHISD methods leading the table and exemplar based manifold learning trailing it as before. Replacing gray level features with LBP ones improves the performance for all methods tested, and these improvements are significant most of the time. Overall, our methods significantly outperform the existing state-of-the-art.

4. Discussion and Conclusions

In this work we developed methods for face recognition from sets of images (rather than from individual images). Our methods characterize each image set (individual) from the gallery and the test set in terms of a convex region in feature space – the affine hull or the convex hull of the feature vectors of its images. Recognition is performed by finding the gallery region (individual) that is closest to the given test region (individual) in the sense of minimum distance between convex sets. The methods can be made resistant to outliers by using robust fitting procedures, and they can easily be kernelized because they are based on Euclidean geometry in feature space. Each class is handled separately so the size of the resulting kernel matrices remains modest.

In experiments on two publicly available face video data sets, we tested our linear and kernelized methods against one (MSM) based on fitting global linear subspaces to the image sets and using canonical angles between subspaces as a similarity measure, and against several others designed to model nonlinear face manifolds [9, 26, 13] by fitting patchwise constant (exemplar), patchwise linear or patchwise affine models to the samples. Our methods performed best overall. Both MSM and the manifold models had lower overall performance and were less consistent over trials and more sensitive to outliers in the data. In part this variability is due to the nonconvex optimization problem that must be solved for the manifold based methods, whereas our methods lead to convex problems. On the data sets tested, the accuracy of our linear methods was only slightly worse than that of the corresponding kernelized ones, although the latter were also slightly stabler on the whole.

Our methods are not limited to face images. They can also be used in other visual recognition problems where each example is represented by a set of images, and more generally in machine learning problems where the classes and test examples are represented by sets of feature vectors. One machine learning use of this kind is to supplement each input example with a set of virtual examples generated using known invariances of the problem. For example, in hand written digit recognition, virtual examples can be created by applying small spatial transformations, changes in thickness of the pen strokes, *etc*., to the input data [23, 7]. In such cases, the problem becomes one of set matching. Traditional approaches such as DeCoste and Schölkopf's kernel jittering use pairwise distances between the generated examples for matching. However as demonstrated in our experiments, if the number of such exemplars is limited, methods that interpolate a dense set (convex model) between the exemplars often do better than ones based on the exemplars alone.

References

[1] T. Ahonen, A. Hadid, , and M. Pietikainen. Face description with local binary patterns: Application to face recognition. *IEEE Transactions on on PAMI*, 28(12):2037–2041, 2006.

[2] O. Arandjelovic, G. Shakhnarovich, J. Fisher, R. Cipolla, and T. Darrell. Face recognition with image sets using manifold density divergence. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2005.

[3] O. Arandjelovic and A. Zisserman. Automatic face recognition for film character retrieval in feature-length films. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2005.

[4] K. P. Bennett and E. J. Bredensteiner. Duality and geometry in svm classifiers. In *ICML*, 2000.

[5] H. Cevikalp, B. Triggs, and R. Polikar. Nearest hyperdisk methods for high-dimensional classification. In *International Conference on Machine Learning*, 2008.

[6] C. Cortes and V. Vapnik. Support vector networks. *Machine Learning*, 20:273–297, 1995.

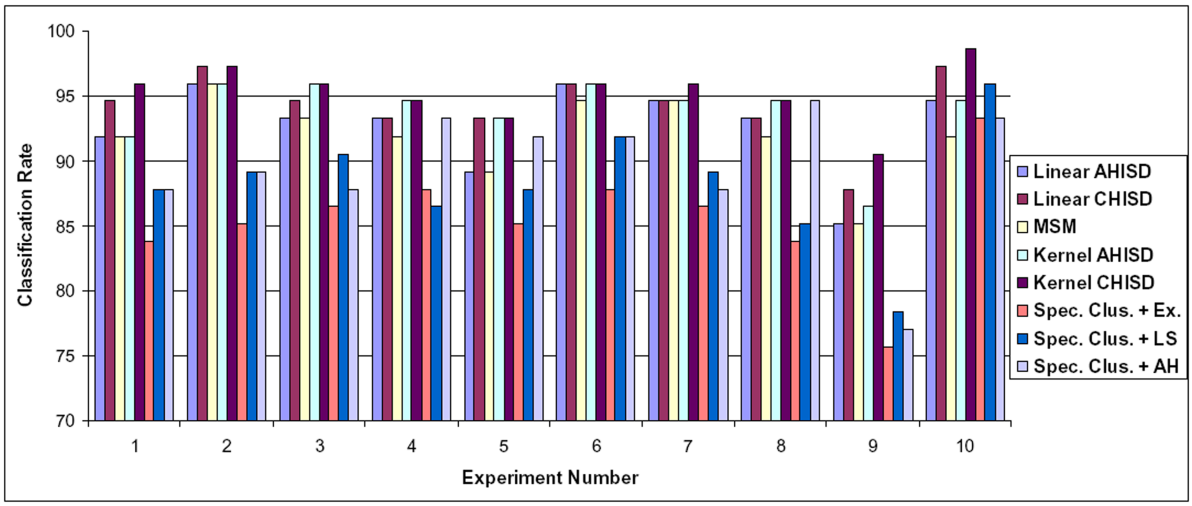

Figure 3. Classification rates of tested methods for each trial, for gray level features on the MoBo data set.

[7] D. DeCoste and B. Schölkopf. Training invariant support vector machines. *Machine Learning*, 46:161–190, 2002.

[8] C. Ding, D. Zhou, X. He, and H. Zha. R1-pca: rotational invariant l1-norm principal component analysis for robust subspace factorization. In *International Conference on Machine Learning*, 2006.

[9] W. Fan and D.-Y. Yeung. Locally linear models on face appearance manifolds with application to dual-subspace based classification. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2006.

[10] A. W. Fitzgibbon and A. Zisserman. Joint manifold distance: a new approach to appearance based clustering. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2003.

[11] K. Fukui and O. Yamaguchi. Face recognition using multi-viewpoint patterns for robot vision. In *International Symposium of Robotics Research*, pages 192–201, 2003.

[12] R. Gross and J. Shi. The cmu motion of body (mobo) database. Technical report, Robotics Institute, Carnegie Mellon University, 2001.

[13] A. Hadid and M. Pietikainen. From still image to video-based face recognition: an experimental analysis. In *International Conference on Automatic Face and Gesture Recognition*, 2004.

[14] P. Hall, J. S. Marron, , and A. Neeman. Geometric representation of high dimension, low sample size data. *Journal of the Royal Statistical Society Series B*, 67(3):427–444, 2005.

[15] G. E. Hinton, P. Dayan, and M. Revow. Modeling the manifold of images of handwritten digits. *IEEE Transactions on Neural Networks*, 18:65–74, 1997.

[16] L. O. Jimenez and D. A. Landgrebe. Supervised classification in high-dimensional space: geometrical, statistical, and asymptotical properties of multivariate data. *IEEE Transactions on Systems, Man, and Cybernetics-Part C: Applications and Reviews*, 28(1):39–54, 1998.

[17] Q. Ke and T. Kanade. Robust l1 norm factorization in the presence of outliers and missing data by alternative convex programming. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2005.

[18] T. Kim and J. Kittler. Locally linear discriminant analysis for multimodally distributed classes for face recognition with a single model image. *IEEE Transactions on PAMI*, 27:318–327, 2005.

[19] K. C. Lee, J. Mo, M. H. Yang, and D. Kriegman. Video-based face recognition using probabilistic appearance manifolds. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2003.

[20] G. I. Nalbantov, P. J. F. Groenen, and J. C. Bioch. Nearest convex hull classification. Technical report, Econometric Institute and Erasmus Research Institute of Management, 2007.

[21] S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290:2319–2323, 2000.

[22] G. Shakhnarovich, J. W. Fisher, and T. Darrell. Face recognition from long-term observations. In *IEEE European Conference on Computer Vision*, pages 851–868, 2002.

[23] P. Simard, Y. L. Cun, J. Denker, and B. Victorri. Transformation invariance in pattern recognition - tangent distance and tangent propagation. *Lecture Notes in Computer Science*, 1524:239–274, 1998.

[24] P. Vincent and Y. Bengio. K-local hyperplane and convex distance nearest neighbor algorithms. In *NIPS*, 2001.

[25] P. Viola and M. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57:137–154, 2004.

[26] R. Wang, S. Shan, X. Chen, and W. Gao. Manifold-manifold distance with application to face recognition based on image sets. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2008.

[27] O. Yamaguchi, K. Fukui, and K.-I. Maeda. Face recognition using temporal image sequence. In *International Symposium of Robotics Research*, pages 318–323, 1998.