A Compact Discriminative Representation for Efficient Image-set Classification with Application to Biometric Recognition

A Compact Discriminative Representation for Efficient Image-set Classification with Application to Biometric Recognition

Muhammad Uzair, Arif Mahmood, Ajmal Mian and Chris McDonald
Computer Science & Software Engineering
The University of Western Australia
35 Stirling Highway, Crawley, WA, Australia
{uzair,arifm,ajmal,chris}@csse.uwa.edu.au

Abstract

We present a simple yet compact and discriminative representation for image sets which can efficiently be used for image-set based object classification. For each image-set we compute a global covariance matrix which captures correlated variations in all image-set dimensions. Without loss of information, we compact the covariance matrix into a lower triangular matrix by using Cholesky decomposition. While preserving discrimination capability of the representation, we obtain further compression by applying Multiple Discriminant Analysis. As a result, we are able to represent image sets containing N samples each of dimensionality d by a single vector whose dimensionality is $<< Nd$. We apply the proposed representation to various biometric applications such as image-set based face recognition and person identification using image-sets of periocular regions. To show that our representation is generic, we also report results for image-set based object categorization. We observe improved accuracy and significant speedup over the current state-of-the-art techniques on standard datasets.

1. Introduction

Image-set based object classification has recently obtained significant attention from the research community [3, 6, 13, 14, 21, 22, 23, 25]. In image-set classification, the gallery consists of one or more sets for each class and each image-set contains multiple images of the same class complementing a wide range of rigid and non-rigid variations as well as illumination changes. In the case of faces, pose variations are relatively rigid while expression variations are non-rigid. The query set also contains an arbitrary number of images of the same subject and is assigned the label of the nearest gallery set by maximizing some similarity measure. The problem of image set classification may naturally arise in a wide range of biometric applications including video-based face recognition, surveillance, person re-identification in camera networks and classification based on long term observations [19].

Compared to recognizing an individual from a single mug-shot, set-to-set matching offers significantly more information. The image set contains useful data variability which can be efficiently modeled for more accurate recognition results [3, 6, 13, 14, 22]. Image-set based face recognition may be considered as a generalization of the video-based recognition, however, it may also be applied in situations where the images of a set may have large variations without any temporal relationship [13, 22].

Often image-set based classification is performed in two steps. The first step is image-set representation using a model that encodes the intra-image as well as inter-image variations within the set. The second step is to find the similarity between two image-set representations by defining a suitable distance measure. For a particular classification algorithm, accuracy, computational complexity and space complexity usually depend on both the specific set modeling approach and the distance metric used.

While image-set based face recognition provides an opportunity of better recognition, it poses many challenges as well. The main challenge is how to efficiently model an image-set in a compact representation without losing discriminative information. Existing algorithms which have relatively more accuracy exhibit more computational complexity [13], while simple and efficient algorithms such as nearest neighbor based classifiers exhibit reduced accuracy and robustness. In contrast to the existing algorithms, the image-set representation proposed in this paper is both compact and computationally efficient. In a wide range of experiments on four standard datasets, the proposed representation exhibited more accuracy than existing algorithms, with significant execution time speedup.

1.1. Overview of the Proposed Algorithm

An image x having d pixels may be considered as a point in a high dimensional Euclidean space $\mathcal{R}^d$, and an image-set as a point cloud in that space. Image-set based classifica-

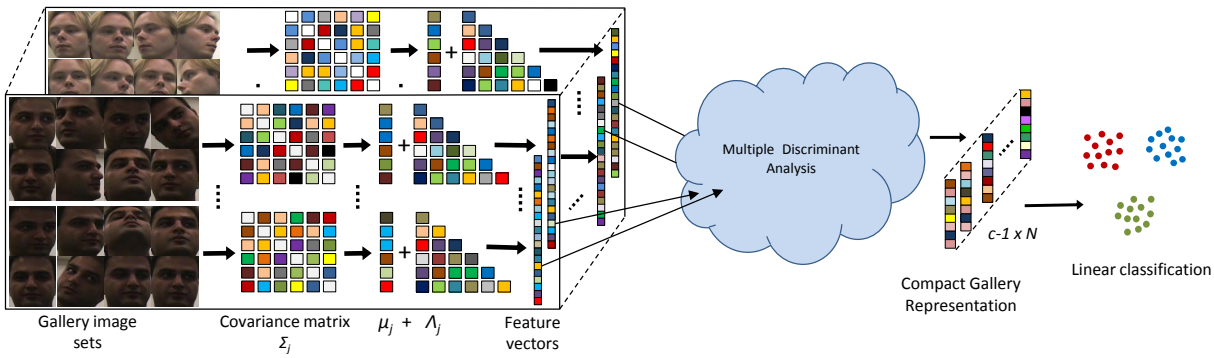

Figure 1. Overview of the proposed algorithm

tion is based on the hypothesis that the point-cloud of each subject has some unique characteristics which may be used to represent and uniquely identify that subject. Moghaddam and Pentland [17] assumed that the point-cloud has multivariate normal distribution $\mathcal{N}(\mu, \Sigma)$ with density given by

$$P(x|\mu, \Sigma) = \frac{\exp\left[-\frac{1}{2}(x-\mu)^t \Sigma^{-1}(x-\mu)\right]}{(2\pi)^{d/2}|\Sigma|^{1/2}}. \quad (1)$$

The distance of a test image y from the distribution was computed by using Mahalanobis distance measure:

$$\Delta(y, \mu, \Sigma) = (y-\mu)^t \Sigma^{-1}(y-\mu). \quad (2)$$

Shakhnarovich et al. [19] also assumed normal distribution of image-set cloud and computed set-to-set distance by using Kullback-Leibler (KL) divergence as a distance measure. KL divergence between two Gaussian distributions is given by

$$\Delta_{Kl}(P_j||P_i) = \frac{1}{2}\log\frac{|\Sigma_i|}{|\Sigma_j|} + \frac{1}{2}Tr(\Upsilon_{i||j}) - \frac{d}{2}, \quad (3)$$

where $\Upsilon_{i||j} = \Sigma_j \Sigma_i^{-1} + \Sigma_i^{-1}(\mu_j - \mu_i)(\mu_j - \mu_i)^t$. Note that KL divergence is not a metric, because it is not symmetric: $\Delta_{Kl}(P_j||P_i) \neq \Delta_{Kl}(P_i||P_j)$ and does not follow the triangular inequality.

Some other researchers such as Forstner and Moonen [8] directly compared the sample covariance matrices, Σ_i and Σ_j of different image-sets to evaluate the set-to-set distance. The non-singular symmetric positive definite matrices do not span the Euclidean space, rather these matrices are located on a high dimensional Riemannian manifold. Comparison between two such matrices in the Riemannian manifold can be done by computing their affine invariant distance $\Delta_\parallel$ [18]

$$\Delta_\parallel(\Sigma_i, \Sigma_j) = \sqrt{\sum_{p=1}^{d} \ln^2 \lambda_p(\Sigma_i, \Sigma_j)}, \quad (4)$$

where $\lambda_p(\Sigma_i, \Sigma_j)$ are the Eigenvalues computed by solving the polynomial given by the determinant $|\lambda\Sigma_i - \Sigma_j| = 0$. $\Delta_\parallel$ is invariant under affine transformations and inversions and is a valid Riemannian metric defined on the space $Sym^+(d, \mathbb{R})$ of real symmetric positive definite matrices. These matrices can also be compared by using the log Euclidean distance Δ_ℓ [4]

$$\Delta_\ell(\Sigma_i, \Sigma_j) = ||\log(\Sigma_i \Sigma_j^{-1})||_F, \quad (5)$$

where $||\cdot||_F$ is the matrix Frobenius norm. For a symmetric positive definite matrix Σ, its eigen-decomposition is given by $\Sigma = U\Lambda U^t$ and logarithm is defined as $\log\Sigma = U\log(\Lambda)U^t$. Both $\Delta_\parallel$ and Δ_ℓ are computationally expensive because of the fact that the covariance matrices have high dimensionality and matrix exponential and logarithm cannot be efficiently computed. Secondly, these measures only use one parameter of the distributions, Σ_i and Σ_j, completely ignore the second important parameters μ_i and μ_j, and there is no straight forward way to embed μ in these formulations.

Due to large dimensionality and small sample size, the empirical estimates of the covariance matrix does not necessarily remain positive definite, which is required to compute KL divergence, Mahalanobis distance, $\Delta_\parallel$ and Δ_ℓ. The solution is to regularize the sample covariance matrices. Different types of regularization of large covariance matrices have already been investigated in statistical literature. Furrer and Bengtsson [9] pointed out that positive definiteness of large covariance matrices can be preserved by point wise multiplication with a known positive definite matrix. In the proposed algorithm, instead of multiplication, we use addition of a positive definite matrix to the covariance matrix. This regularization process is motivated by the ridge regression [11]. In multiple linear regression: $Y = X\beta + \epsilon$, if columns of X are not linearly independent, then X^tX is rank deficit. Then instead of using least squares estimate $\beta = (X^tX)^{-1}X^tY$, a regularized estimate $\hat{\beta} = (X^tX + kI)^{-1}X^tY$, for $k \geq 0$, is used.

If an image space has dimensionality d, the Riemannian manifold will have dimensionality of the order of d^2, which is significantly larger. As an example, for an image-set with images of size 100×100 pixels, the covariance matrix is 10^8

dimensional. The covariance matrix, being symmetric, has $\frac{d(d+1)}{2}$ unique elements. Instead of using only these unique elements, we apply Cholesky decomposition [24] and get a lower triangular matrix which captures the full information present in the regularized covariance matrix. Our choice of Cholesky decomposition is also motivated by the target tracking and texture classification work of Hong et al. [12].

In order to enrich the second order statistic Σ with the first order statistic μ, we add the mean of the corresponding image set to each column of the lower triangular matrix. The mean enriched matrix is then rearranged in vector form to represent the image set (Figure 1). The dimensionality of the feature vectors is $d(d+1)/2$, which is still large. We achieve further dimensionality reduction by using Multiple Discriminant Analysis.

During testing, for each probe set we compute the feature vector and transform it by the MDA basis learned from the gallery sets. In the MDA space, any linear classifier such as SVM may be trained to separate one class from the others. We observe that a simple classifier, such as nearest neighbor, also yields very good accuracy since the learned representation is discriminative. The proposed algorithm is tested for image set based object classification on four standard datasets [15, 10, 16, 1]. The recognition rates and the execution time efficiency is compared with seven state of the art algorithms [14, 23, 21, 6, 13, 22]. The proposed algorithm has demonstrated significant improvements in the recognition rate as well as in the execution time.

1.2. Current Image-set Classification Techniques

Existing image-set classification techniques may be categorized into sample based (nearest neighbor) techniques and structure based set-to-set matching techniques. Sample based techniques measure the distance between nearest neighbor points of two image sets. For example, Cevikalp and Triggs [6] considered each image set as a convex geometric region in $\mathcal{R}^d$. Set dissimilarity was measured by the distance of closest approach between the regions represented by the affine (AHISD) or convex hulls (CHISD). For the case of affine hull, the minimum distance was computed using least squares while for the case of convex hull, an SVM was trained to separate the probe set from the gallery.

Instead of searching the nearest points with dense combinations of samples in the corresponding image sets, Hu et al. [13] proposed that each of the two points should be able to be approximated by a sparse combination from the samples of the respective sets. They argued that the sparse approximated nearest points (SANP) will lie close to some facet of the affine hull and hence, implicitly incorporate structural information of the sets as well. Wu et al. [25] performed distance metric learning on the geometric distance between the approximated convex hulls of each pair of query-gallery sets. A maximum-margin-based ranking algorithm is adopted to learn a good metric, making the closest distance between correct query-gallery pair smaller than that between incorrect ones. Sample based methods may be vulnerable to outliers. For example, if a query set contains a single outlier closer to a different gallery set, it may be misclassified based on that sample alone.

Structure based techniques model the underlying structure of an image set with one or more linear subspaces. Structural similarity between the sets is usually measured using subspace to subspace distance. Kim et al. [14] performed discriminative learning using canonical correlations between the structures of sets. A discriminant function was learned that maximized the within-class similarity and minimized the between-class canonical correlations. Wang et al. [23] proposed Manifold-Manifold Distance (MMD) which clustered each image set into multiple linear local models and represented each model by a linear subspace. The similarity between two sets was defined as the canonical correlation between the nearest local models. In addition, the nearest point distance was also combined with the structural similarity to calculate the final similarity between two sets. Wang and Chen [21] proposed Manifold Discriminant Analysis (MDA) that represented each image set by multiple local linear models. The local models were transformed by a linear discriminant function where different classes were better separable. The similarity between two sets was calculated as the pair-wise local model distances in the learned embedding space.

Wang et al. [22] modeled the structure of each image set directly by the covariance matrix. They mapped the covariance matrix of each image set from the Riemannian manifold to the Euclidean space by a kernel function based on Log Euclidean distance (5). The image sets were then classified according to a learned regression function calculated by the kernel partial least squares. The proposed algorithm is different from [22] because we use Cholesky decomposition to bring the covariance matrices to the Euclidean space. Also our representation is based on both μ and Σ, and we employ Multiple Discriminant Analysis for dimensionality reduction and improving discrimination and we use simple NN technique for set-to-set matching. In our experiments, we observe more accuracy and speedup than [22].

2. Proposed Algorithm

The proposed algorithm has two main steps in the training phase. The first step is to compactly represent an image set with a feature vector and the second step is to reduce the dimensionality of the feature vector while maintaining discrimination capability to get high speedup. During the test phase, the compact representations of the probe sets are very efficiently compared with the gallery set representations using a simple nearest neighbor technique.

2.1. Compact Image Set Representation

Let $G = \{X_j\}_{j=1}^{g} \in \mathcal{R}^{d \times N}$ be the gallery containing g image sets and N is the total number of images in the gallery: $N = \sum_{j=1}^{g} n_j$, where n_j is the number of images in the j^{th} image set. Let $X_j = \{x_j^i\}_{i=1}^{n_j} \in \mathcal{R}^{d \times n_j}$ be the j^{th} image set, where $x_j^i \in \mathcal{R}^d$ is a d dimensional feature vector obtained by lexicographic ordering of the pixel elements of the i^{th} image in the j^{th} set. Instead of pixel values, the vector x_j^i may also contain feature values such as LBP or Gabor features. The value of n_j may vary across image sets while the dimensionality of x_j^i will remain fixed.

The mean of the image set X_j is often used to capture the first order statistics

$$\mu_j = \frac{1}{n_j} \sum_{i=1}^{n_j} (x_j^i), \tag{6}$$

and the covariance Σ_j is used to capture the second order statistics of the image set

$$\Sigma_j = \frac{1}{n_j - 1} \sum_{i=1}^{n_j} (x_j^i - \mu_j)(x_j^i - \mu_j)^t. \tag{7}$$

If the number of images n_j in one image set are less than the dimensionality of the feature vector d then the matrix $\Sigma_j \in \mathcal{R}^{d \times d}$ will become rank deficient and it will be semi positive definite.

We propose to decompose Σ_j into lower triangular matrices by applying Cholesky decomposition. However, Cholesky decomposition will yield a unique lower triangular matrix only if the covariance matrix is positive definite. In order to ensure a unique decomposition, we have to ensure that all eigenvalues are large (positive), which we obtain by introducing a regularization term:

$$\hat{\Sigma}_j = \Sigma_j + \frac{\lambda_{\Sigma_j}}{\gamma} I, \tag{8}$$

where $\lambda_{\Sigma_j} = \sum_{i=1}^{d} \lambda_j$ is the sum of all eigenvalues of Σ_j, I is an identity matrix of the same size as that of Σ_j, and $\gamma > 1$ is a positive constant. To make the analysis process simple, in all of our experiments we use a fixed value of $\gamma = 1000$. Note that this type of regularization is also similar to the one used by Moghaddam and Pentland [17], in which they applied eigen-decomposition on $\Sigma_j = UDU^t$, and in the diagonal matrix D, replaced the k smallest eigenvalues by their average $\rho = \frac{1}{k} \sum_{i=1}^{k} \lambda_i$. Since SVD is computationally expensive and efficiency is an important criterion of our approach, we avoid SVD and add a fraction of the sum of all eigenvalues at the leading diagonal of Σ_j, which is equivalent to adding the values to the diagonal of D. Thus we get more efficiency and also get rid of the user defined parameter k. The value of λ_{Σ_j} is computed as $\lambda_{\Sigma_j} = \text{trace}(\Sigma_j)$.

By applying Cholesky decomposition on $\hat{\Sigma}_j$ we get

$$\hat{\Sigma}_j = \Lambda_j \times \Lambda_j^t, \tag{9}$$

where Λ_j is the lower triangular matrix with positive diagonal entries. In the lower triangular matrix Λ_j we add first order statistics of the image set

$$\hat{\Lambda}_j = \Lambda_j + \mu_j 1^{1 \times d}, \tag{10}$$

where $1^{1 \times d}$ is a row of ones. The feature vector f_j is obtained by applying a function $\psi()$ on $\hat{\Lambda}_j$.

$$f_j = \psi(\hat{\Lambda}_j) = \begin{cases} \hat{\Lambda}_j(p,q) \text{ if } p \geq q \\ 0 \text{ otherwise .} \end{cases} \tag{11}$$

We rearrange the non zero entries of f_j in a vector form, $f_j \in \mathcal{R}^{\frac{d(d+1)}{2}}$. Since f_j globally represents an image set X_j therefore, the distance between two images sets X_a and X_b can be efficiently computed by computing the distance between the corresponding feature vectors f_a and f_b in the Euclidean space. We observe that this distance computation is more efficient than the existing image set models such as manifold set representations [23, 21] or affine and convex hull based image set representations [6, 13].

2.2. Multiple Discriminant Analysis

In the last section we explained how we represent an image set with a single feature vector. This feature vector is still of very high dimensionality. In this section, we show how we gain further computational and memory efficiency. We reduce the dimensionality of the feature vectors while maintaining the discrimination.

Let $G_p = \{f_{ij}\}_{i=1,j=1}^{n_j,c} \in \mathcal{R}^{\frac{d(d+1)}{2} \times g}$ be the compact gallery representation learned in the previous section, where c are the subject classes (or object categories) contained in the gallery and $n_j \geq 1$ are the number of image sets in each class, $g = \sum_{j=1}^{c} n_j$, and f_{ij} represents the i^{th} feature vector in the j^{th} class. We intend to reduce the dimensionality of the f_{ij} from $\frac{d(d+1)}{2}$ to $c-1$, by using Multiple discriminant analysis [7].

Multiple discriminant analysis is a generalization of Fisher Linear Discriminant (FLD) and requires $c-1$ discriminant functions to be learned for a classification problem with c classes. Therefore, we project the feature vectors f_{ij} from $d(d+1)/2$ dimensional space to a $c-1$ dimensional space: $\hat{f}_{ij} = W^t f_{ij}$ or in terms of the gallery matrix: $\hat{G}_p = W^t G_p$, where W is an ortho-normal transformation matrix of size $\frac{d(d+1)}{2} \times (c-1)$ and the size of the transformed gallery matrix $\hat{G}_p$ is $(c-1) \times g$. Traditionally W is learned such that the between-class scatter of $\hat{G}_p$ is maximized and within-class scatter is minimized. Let S_j be the

within-class scatter for each of the j^{th} class

$$S_j = \sum_{i=1}^{n_j} (f_{ij} - \mu_j)(f_{ij} - \mu_j)^t, \tag{12}$$

where μ_j is the mean of the feature vectors within the j^{th} class: $\mu_j = \frac{1}{n_j} \sum_{i=1}^{n_j} f_{ij}$. Since the outer product $(f_{ij} - \mu_j)(f_{ij} - \mu_j)^t$ has rank one, the rank of S_j will be bounded by n_j $\mathbb{R}(S_j) \leq n_j$. The overall within-class scatter is given by the summation of all class-scatter matrices

$$S_w = \sum_{j=1}^{c} S_j, \tag{13}$$

and the rank of $S_w \in \mathcal{R}^{\frac{d(d+1)}{2} \times \frac{d(d+1)}{2}}$ is upper bounded by the sum of ranks of the individual class scatter matrices and thus by the number of image-sets in the gallery $\mathbb{R}(S_w) \leq g$, which shows that S_w is rank deficient.

The between class scatter matrix is defined as

$$S_b = \sum_{j=1}^{c} n_j(\mu_j - \mu)(\mu_j - \mu)^t, \tag{14}$$

where μ is the average of all feature vectors in the gallery and μ_j is the class specific mean. Since S_w is a summation of c rank-one matrices, it can have only c non-zero eigenvalues.

The between class scatter of the transformed feature vectors is $\hat{S}_b = W^T S_b W$ and within class scatter is $\hat{S}_w = W^t S_w W$. Traditionally W is learned such that the ratio of $\hat{S}_b$ to the $\hat{S}_w$ is, in some way, maximized. Since the determinant of a matrix is the product of its eigenvalues which represents its scatter, therefore, often the ratio of the determinants of both scatter matrices is maximized

$$W_{opt} \equiv \arg\max_{W} \frac{|W^t S_b W|}{|W^t S_w W|}. \tag{15}$$

W_{opt} may be considered as the set of generalized eigenvectors of S_b and S_w corresponding to the $c-1$ largest eigenvalues:

$$S_b W = \Lambda S_w W, \tag{16}$$

where Λ is a diagonal matrix containing eigenvalues. For a non-singular S_w, W may be computed as eigenvectors of $S_w^{-1} S_b$.

In our case, S_w is rank deficient which means the traditional solution cannot be applied. Assuming each image-set will result in an independent feature vector, the null space of S_w, has $\frac{d(d-1)}{2} - g$ dimensions. Minimization of within class scatter, $W^t S_w W$ can easily find a W within the null space of S_w, resulting in $S_w W = \mathbf{0}$ and $|W^t S_w W| = 0$, which will result in $|\hat{S}_b|/|\hat{S}_w| \rightarrow \infty$, without considering the maximization of $|\hat{S}_b|$. This may be considered as a degenerated case of MDA. One may think of applying some type of regularization on S_w to make it positive definite. However, we chose to reduce the dimensionality of the feature space by using PCA such that S_w will become full rank [5].

The gallery matrix G_p has $\frac{d(d+1)}{2}$ rows which are significantly larger than the number of columns. Since the row and column ranks are always equal, the number of linearly independent rows in G_p are bounded by g. In order to discard additional rows, we use Principal Component Analysis (PCA). PCA basis is learned such that the total scatter $\hat{S}_w + \hat{S}_b$ of the $\hat{G}_p$ is maximized. An ortho-normal transformation matrix Ψ may be computed such that

$$\Psi_{opt} \equiv \arg\max_{\Psi} |\Psi^T (S_w + S_b) \Psi|, \tag{17}$$

or $(S_w + S_b)\Psi = \Lambda\Psi$, which shows that Ψ is the matrix of eigenvectors of $S_w + S_b$.

Transformation of S_w and S_b with Ψ will be $\Psi^t S_w \Psi$ and $\Psi^t S_b \Psi$. The size of both transformed scatter matrices is $g \times g$, and the rank of $\Psi^t S_w \Psi$ is g. Using these reduced size scatter matrices, the transformation matrix Φ may now be found

$$\Phi_{opt} \equiv \arg\max_{\Phi} \frac{|\Phi^t \Psi^t S_b \Psi \Phi|}{|\Phi^t \Psi^t S_w \Psi \Phi|}, \tag{18}$$

or in terms of generalized eigenvalue problem

$$\Psi^t S_b \Psi \Phi = \Lambda \Psi^T S_w \Psi \Phi. \tag{19}$$

For a non-singular $\Psi^T S_w \Psi$, Φ may be computed as eigenvectors of $(\Psi^t S_w \Psi)^{-1}(\Psi^t S_b \Psi)$. In order to obtain a compact and discriminative representation of the gallery, we project G_p on $W = \Psi\Phi$: $\hat{G}_p = (\Psi\Phi)^T G_p$. The final compact and discriminative gallery representation $\hat{G}_p \in \mathcal{R}^{c-1 \times g}$ is used for estimating the label of the probe image set.

The compact representation of the probe image set is also projected on the $\Psi\Phi$: $\hat{f}_p = (\Psi\Phi)^t f_p$, and the distance is computed from each of the feature vectors in $\hat{G}_p$

$$L_p \equiv \min_{1 \leq j \leq c} \big(\min_{1 \leq i \leq n_j} (||\hat{f}_{ij} - \hat{f}_p||_n) \big), \tag{20}$$

where L_p is the predicted label of the probe image set and $||\cdot||_n$ represents n^{th} norm distance between the two vectors. In our experiments discussed in the following section, we have used L_2 norm in (20).

3. Experimental Evaluation

To evaluate the proposed algorithm, we have performed extensive experimentation on four standard datasets capturing a wide range of operating conditions. Holistic face recognition experiments are performed on the Honda/UCSD [15]

and CMU Mobo [10] datasets. Experiment on periocular biometric recognition are performed on periocular images generated from the MBGC NIR video v2 dataset [1] and object categorization on the ETH-80 dataset [16]. The proposed algorithm is compared with seven state-of-the-art algorithms for recognition rate performance and for execution time complexity. Experimental setup details are given in the following subsection.

3.1. Comparative Techniques and Setup

The seven classification techniques studied in this paper include Discriminant Canonical Correlation Analysis (DCC) [14], Manifold-to-Manifold Distance [23], Manifold Discriminant Analysis (MDA) [21], Affine Hull based Image Set Distance (AHISD), Convex Hull based Image Set Distance (CHISD) [6], Sparse Approximated Nearest Points (SANP) [13] and Covariance Discriminative Learning (CDL) [22]. Brief details are given in Section 1.2.

We have used the implementations made available by the original authors, except for MDA and CDL techniques. For MDA, Hu [13] implementation is used, while we have our own implementation of CDL. To allow comparison with the previous methods we followed the same protocol as used by [6, 13, 21, 22, 23]. We performed ten-fold cross validation experiments for all the datasets by randomly selecting gallery/probe combinations in each fold. We report average recognition rates of different methods for ten folds. The important parameters of different methods are carefully optimized. For DCC, the dimension of the embedding space is set to 100. The subspace dimension is set to 10 which preserves 90% energy and the corresponding 10 maximum canonical correlations are used to calculate set similarity.

For MMD and MDA, the parameters are configured according to [23, 21]. The ratio between Euclidean distance and geodesic distance is optimized for different datasets (i.e. 2.0 for Honda, 5.0 for Mobo and 2.0 for ETH-80 dataset). The maximum canonical correlation is used in defining MMD. For MDA, the number of between-class NN local models and the dimension of MDA embedding space are tuned to achieve the best results for each dataset as specified in [21]. The number of connected nearest neighbors for computing geodesic distance in both MMD and MDA is set to 12. For CHISD, we set the error penalty parameter to be the same as in [6] i.e. $C = 100$ for gray-scale pixel values and $C = 50$ for LBP in linear SVM. There is no parameter setting for AHISD and CDL.

For Honda, MoBo and MBGC datasets, each subject has one image set as the gallery and the rest of the images sets for probes. For ETH-80, each category has 5 objects for gallery and the other 5 objects for probes. Whenever the gallery contains only one image set for a particular class, we randomly partition that image set into two non-overlapping sub-sets and use each subset as a full image set.

Figure 2. Sample images from HONDA/UCSD and CMU Mobo Data sets.

The execution times of different algorithms are also compared. Average execution times over ten folds are reported on a Pentium 3.4GHz CPU with 8GB RAM. The memory requirements of the proposed algorithm are also analyzed and found smaller than the existing algorithms. For test, only the compact gallery representation along with Multiple Discriminant basis are required to be stored. The compact gallery for Honda dataset is shown in Figure 6.

3.2. Dataset Details

The Honda/UCSD dataset contains 59 video sequences of 20 different subjects. Each video contains approximately 300 to 500 frames covering large variations in head pose and facial expression. The CMU MoBo dataset contains 96 video sequences of 24 different subjects. Each subject has 4 sequences captured in different walking situations and each sequence has about 300 frames. For both databases, the faces in every frame of the video sequences are automatically detected by applying [20], cropped and converted to gray scale. Almost 30% frames on which face detection failed were dropped. For Honda dataset, we use 20x20 face images after histogram equalization. For CMU Mobo dataset, the grayscale face images are resized to 40x40 as in [6]. The Local Binary Patterns (LBP) [2] provided by [6] are used as the features of individual images. Sample face images from both datasets are shown in Fig. 2.

For periocular biometric recognition we have performed experiments on periocular images extracted from the NIR face videos of the MBGC portal challenge dataset version 2 [1]. Sample periocular images are shown in Fig. 3. This dataset consists of 114 different subjects, each subject has 1 to 11 image sets and each image set contains 5 to 25 images. The experiment is repeated ten fold and for each fold, the

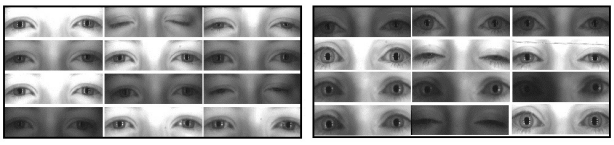

Figure 3. Sample periocular image sets for two subjects from the MBGC dataset

Figure 4. Eight object categories in the ETH-80 dataset.

gallery is constructed by randomly selecting one image set for each of the 114 subjects. The remaining 336 image sets are used as probes.

ETH-80 dataset [16] contains images of 8 object categories and each category has 10 objects. Each object has 41 images of different views which form an image set. We use 20×20 intensity images for the task of classifying an image set of an object into a known category.

3.3. Results and Discussion

In the face biometric experiments, the proposed algorithm has achieved 100% accuracy on the Honda dataset and 96.76% accuracy on the CMU Mobo dataset (Table 1). On the Honda dataset, SANP and CDL have also obtained 100% accuracy, while other algorithms have remained relatively less accurate. On the CMU Mobo, SANP has obtained 97% while CDL has obtained 95.83% accuracy (Table 1). Although SANP has shown good recognition performance, SANP is 117.63 times slower than our algorithm (Table 3). As compared to SANP, CDL is quite fast, however it has also remained 2.78 times slower than our algorithm. In the CDL algorithm, computational time is taken by the kernel computation both in the training and the testing steps. Training time of CDL is 10 times slower than our algorithm. In the speedup experiments, DCC has performed significantly faster than all algorithms, however it has shown less accuracy, especially on the CMU Mobo dataset. The proposed algorithm has exhibited very high accuracy and also high computational efficiency for the face biometric.

In the object categorization experiment on ETH-80 dataset, our algorithm again outperformed all existing algorithms by exhibiting 91.25% average accuracy (Table 1). CDL was the next nearest by obtaining 89.20% and the DCC algorithm was the third most accurate by achieving 87.50% accuracy. On this dataset, SANP has obtained only 72.10% accuracy, which indicates that SANP may not be a good scheme for generic object classification. In contrast, our algorithm has delivered best performance for holistic face recognition, periocular recognition and object categorization as well.

In the periocular region biometric experiments on the

Table 1. Average recognition rates (%) of different algorithms on three datasets in ten fold experiments.

Algorithm	Honda	MoBo	ETH-80
DCC	94.87 ± 1.32	91.53 ± 2.51	90.95±5.32
MMD	94.87 ± 1.16	89.72 ± 1.68	85.72±8.29
MDA	97.44 ± 0.91	94.98 ± 2.22	80.50±6.81
AHISD	89.74 ± 1.85	94.58 ± 2.12	74.76±3.31
CHISD	92.31 ± 2.12	96.52 ± 1.71	71.00±3.93
SANP	100.00 ± 0.00	**97.00 ± 0.63**	72.43±4.98
CDL	100.00 ± 0.00	95.83 ± 2.51	89.20±6.82
Proposed	**100.00 ± 0.00**	96.76 ± 0.60	**91.25±3.91**

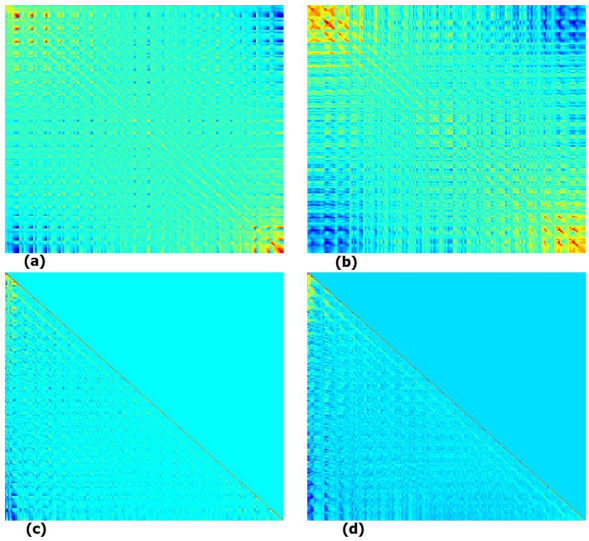

Figure 5. (a) & (b) Covariance matrices of image sets of different subjects in Honda/UCSD data set. (c) & (d) Lower triangular matrices for the same image sets. High values along diagonals are due to the regularization term in (8). Please note that the value of γ is same for both gallery and test sets. Therefore, its effect is canceled when norm is computed in (20).

MBGC dataset, our algorithm outperformed all others by achieving 92.57% accuracy (Table 2). SANP obtained 90.06% accuracy and CHISD obtained 88.99%, and CDL obtained only 64.37% accuracy which was significantly lower than the proposed algorithm. While CDL performed good on Honda/UCSD, CMU Mobo and ETH-80 datasets, its performance deteriorated with change of image acquisition sensors.

4. Conclusion

A computationally efficient yet highly accurate image set classification algorithm is presented. A novel image set representation which was based on covariance matrices and means was proposed. The covariance matrices were decomposed by using Cholesky decomposition and the means were added to the lower triangular matrices. Dimensionality of the resulting feature vectors was reduced using Multiple Discriminant Analysis. The proposed algorithm was compared with seven state of the art algorithms on four stan-

Table 2. Identification rates (%) of 10-fold experiments on the MBGC dataset

Fold	1	2	3	4	5	6	7	8	9	10	Mean	STD
MMD	85.71	82.14	81.25	82.44	82.14	79.19	81.55	80.95	80.36	79.76	81.549	1.72
MDA	90.77	87.58	87.39	86.61	87.69	86.61	87.58	87.5	87.1	87.1	87.593	1.12
AHISD	92.45	89.37	89.07	87.81	88.88	87.69	88.29	87.13	88.5	87.8	88.699	1.41
CHISD	93.45	90.67	90.48	87.99	88.88	87.69	87.99	87.1	88.5	87.1	88.985	1.90
DCC	74.37	68.97	67.37	67.23	61.37	64.03	60.83	62.32	64.3	58.25	64.904	4.46
SANP	93.15	89.88	89.68	88.8	88.1	88.8	87.34	88.1	88.61	88.1	89.056	1.54
CDL	61.37	67.37	61.37	67.37	61.37	67.37	67.37	61.37	67.37	61.37	64.370	3.00
Proposed	**95.83**	**92.86**	**93.75**	**91.67**	**94.05**	**89.88**	**93.15**	**91.96**	**90.93**	**91.67**	**92.575**	1.63

Table 3. Execution times on Honda dataset

Algorithm	Training Time	Testing Time
DCC	0.91s	0.30s
MMD	184.57s	38.10s
MDA	10.55s	33.00s
AHISD	N/A	9.10s
CHISD	N/A	110.10s
SANP	N/A	482.30s
CDL	3.50s	11.40s
Proposed	0.35s	4.10s

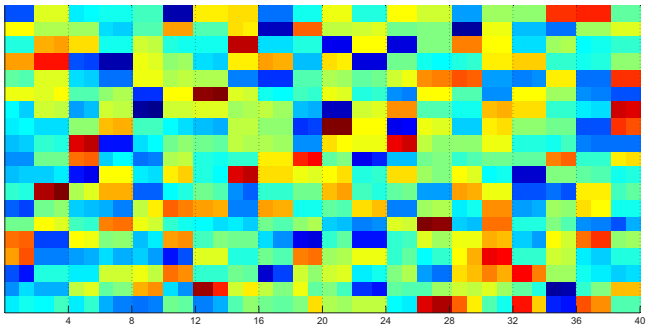

Figure 6. Graphical representation of the compact feature vectors of the gallery image sets of Honda/UCSD data by different colors. The size of each vector is 19 because there are 20 classes in the gallery. The number of vectors is 40 because for each class we have two vectors. Please note that the vectors representing different classes are significantly different while vectors belonging to same class have relatively closer values.

dard datasets. The experimental results demonstrate that the proposed algorithm was significantly faster than the current most accurate algorithms while exhibiting similar or better recognition rates.

5. Acknowledgements

This research was supported by ARC grants DP1096801 and DP110102399.

References

[1] Multiple Biometric Grand Challenge (MBGC) dataset http://face.nist.gov/mbgc/.

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

[3] O. Arandjelovic, G. Shakhnarovich, J. Fisher, R. Cipolla, and T. Darrell. Face recognition with image sets using manifold density divergence. In *CVPR*, June 2005.

[4] V. Arsigny, P. Fillard, X. Pennec, and N. Ayache. Geometric means in a novel vector space structure on symmetric positive-definite matrices. *SIAM J. Matrix Analysis Applications*, 29(1):328–347, 2006.

[5] P. Belhumeur, J. Hespanha, and D. Kriegman. Eigenfaces vs. fisherfaces: recognition using class specific linear projection. *IEEE PAMI*, 19(7):711 –720, July 1997.

[6] H. Cevikalp and B. Triggs. Face recognition based on image sets. In *CVPR*, June 2010.

[7] R. Duda, P. Hart, and D. Stork. *Pattern Classification*. New York: John Wiley & Sons, 2001.

[8] W. Frstner and B. Moonen. A metric for covariance matrices. Technical report, Stuttgart University, 1999.

[9] R. Furrer and T. Bengtsson. Estimation of high-dimensional prior and posterior covariance matrices in kalman filter variants. *J. Multivariate Analysis*, 98(2):227 – 255, 2007.

[10] R. Gross and J. Shi. The cmu motion of body database. Technical Report CMU-RI-TR-01-18, Robotics Institute, 2001.

[11] A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):pp. 55–67, 1970.

[12] X. Hong, H. Chang, S. Shan, X. Chen, and W. Gao. Sigma set: A small second order statistical region descriptor. In *CVPR*, June 2009.

[13] Y. Hu, A. Mian, and R. Owens. Face recognition using sparse approximated nearest points between image sets. *IEEE Transactions on PAMI*, 2012.

[14] T.-K. Kim, J. Kittler, and R. Cipolla. Discriminative learning and recognition of image set classes using canonical correlations. *IEEE Trans. on PAMI*, 29(6):1005 –1018, June 2007.

[15] K.-C. Lee, J. Ho, M.-H. Yang, and D. Kriegman. Video-based face recognition using probabilistic appearance manifolds. In *CVPR*, June 2003.

[16] B. Leibe and B. Schiele. Analyzing appearance and contour based methods for object categorization. In *CVPR*, 2003.

[17] B. Moghaddam and A. Pentland. Probabilistic visual learning for object representation. *IEEE Trans. PAMI*, 19(7):696 –710, 1997.

[18] X. Pennec, P. Fillard, and N. Ayache. A riemannian framework for tensor computing. *IJCV*, 66:41–66, 2006.

[19] G. Shakhnarovich, J. W. Fisher, and T. Darrell. Face recognition from long-term observations. In *Proc. ECCV*, pages 851–868, 2002.

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

[21] R. Wang and X. Chen. Manifold discriminant analysis. In *CVPR*, pages 429 –436, June 2009.

[22] R. Wang, H. Guo, L. Davis, and Q. Dai. Covariance discriminative learning: A natural and efficient approach to image set classification. In *CVPR*, June 2012.

[23] R. Wang, S. Shan, X. Chen, and W. Gao. Manifold-manifold distance with application to face recognition based on image set. In *CVPR*, June 2008.

[24] W. B. Wu and M. Pourahmadi. Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika*, 90(4):831–844, 2003.

[25] Y. Wu, M. Minoh, M. Mukunoki, and S. Lao. Set based discriminative ranking for recognition. In *ECCV*, 2012.