Simultaneous Feature and Dictionary Learning for Image Set Based Face Recognition

Jiwen Lu[1], Gang Wang[1,2], Weihong Deng[3], and Pierre Moulin[1,4]

[1] Advanced Digital Sciences Center, Singapore
[2] Nanyang Technological University, Singapore
[3] Beijing University of Posts and Telecommunications, Beijing, China
[4] University of Illinois at Urbana-Champaign, IL USA

Abstract. In this paper, we propose a simultaneous feature and dictionary learning (SFDL) method for image set based face recognition, where each training and testing example contains a face image set captured from different poses, illuminations, expressions and resolutions. While several feature learning and dictionary learning methods have been proposed for image set based face recognition in recent years, most of them learn the features and dictionaries separately, which may not be powerful enough because some discriminative information for dictionary learning may be compromised in the feature learning stage if they are applied sequentially, and vice versa. To address this, we propose a SFDL method to learn discriminative features and dictionaries simultaneously from raw face images so that discriminative information can be jointly exploited. Extensive experimental results on four widely used face datasets show that our method achieves better performance than state-of-the-art image set based face recognition methods.

Keywords: Face recognition, image set, feature learning, dictionary learning, simultaneous learning.

1 Introduction

Image set based face recognition has attracted increasing interest in computer vision in recent years [41,33,27,12,15,2,11,7,23,38,3,17,37,4,6,5,30,19,29]. Different from conventional image based face recognition systems where each training and testing example is a single face image, for image set based face recognition, each training and testing example contains a face image set captured from different poses, illuminations, expressions and resolutions. While more information can be provided to describe the person with image sets, image set based face recognition is still challenging because there are usually large intra-class variations within a set, especially when they are captured in unconstrained environments.

There has been a number of work on image set based face recognition over the past decade [41,33,27,12,15,2,11,7,23,38,3,17,37,4,6,5,30,19], and dictionary-based methods have achieved state-of-the-art performance [4,6,5] because the pose, illumination and expression information in face image sets can be implicitly

D. Fleet et al. (Eds.): ECCV 2014, Part I, LNCS 8689, pp. 265–280, 2014.

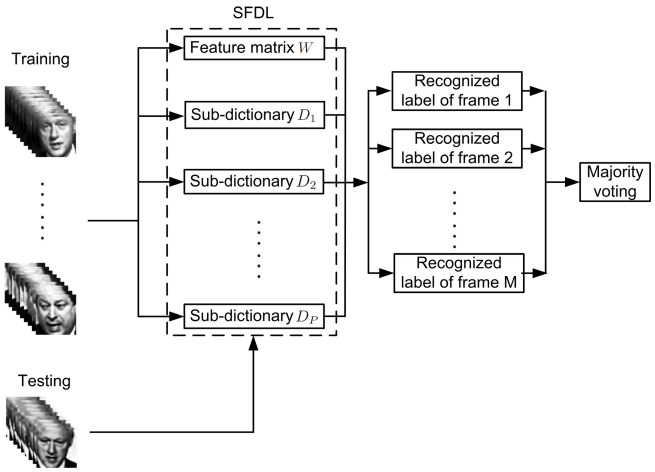

Fig. 1. The basic idea of our image set based face recognition approach, where discriminative features and dictionaries are learned simultaneously to encode the pose, illumination and expression information in face image sets, so that it is more robust to noise. In the training stage, we learn a feature projection matrix W and a structured dictionary $D = [D_1, D_2, \cdots, D_P]$ (one sub-dictionary per class) by using the proposed SFDL method, where P is the number of subjects in the training set. Given a testing face image set containing M image frames, we first apply the learned feature projection matrix W to project each sample into a feature and recognize its label by using the smallest reconstruction error corresponding to the associated sub-dictionary. Lastly, the majority voting strategy is used to classify the whole testing face image set.

encoded into the learned dictionaries. However, most existing dictionary-based image set based face recognition methods are unsupervised [4,6,5], which are not discriminative enough to classify face sets. Moreover, these methods learn dictionaries using the original raw pixels, which may contain some noisy components that are irrelevant to dictionary learning. Since face images usually lie on a low-dimensional manifold, it is desirable to seek the most discriminative features in a low-dimensional subspace and suppress the useless information to promote learning dictionaries for image sets.

In this paper, we propose a new simultaneous feature and dictionary learning (SFDL) method for image set based face recognition, where the basic idea is illustrated in Fig. 1. The goal of our method is to jointly learn a feature projection matrix and a structured dictionary, where each frame within a set is projected into a low-dimensional subspace and encoded with a discriminative coding coefficient, and face image sets from each person are represented by a sub-dictionary so that person-specific dictionaries can be learned to extract more discriminative information, simultaneously. Extensive experimental results on four widely used face datasets show that our method achieves better performance than state-of-the-art image set based face recognition methods.

2 Related Work

Image Set Based Face Recognition: Over the past recent years, we have witnessed a considerable interest in developing new methods for image set based face recognition [33,27,12,15,2,11,23,38,36,3,17,8,4,6,5,30]. These methods can be mainly categorized into two classes: parametric and non-parametric. Parametric methods first model each face image set as a distribution function and then compute the divergence between two distributions as the similarity of two face image sets. The key shortcoming of these methods is that if there are not strong correlations between two face image sets, the estimated model cannot well characterize the sets and may fail to measure their similarity. Non-parametric methods usually represent each face image set as a single or mixture of linear subspaces, and then use the subspace distance to measure the similarity of face image sets. Representative subspace distance methods include principal angle [16], affine/convext hull similarity [3], and nearest points distance [17,18,44]. However, these methods are generally sensitive to outliers and occlusions. To address this, Chen *et al.* [4,6,5] presented a dictionary-based approach for image set based face recognition by building one dictionary for each face image set and using these dictionaries to measure the similarity of face image sets. While reasonably good recognition rates can be obtained, their approach is generative and the dictionaries are learned from the original raw pixels, which may contain some noisy and irrelevant components.

Dictionary Learning: There have been extensive work on dictionary learning in the literature [1,32,45,20,34,42,31,24,14,28,39,46,9,10]. Dictionary learning aims to seek a collection of atoms for sparse representation of the input samples, where each data is linearly represented by a small number of atoms. Existing dictionary learning methods can be mainly classified into two categories: unsupervised [1] and supervised [42,24]. In recent years, dictionary learning has been extensively used in face recognition and also shown good performance [42,24]. However, most existing dictionary learning methods have been developed for image based face recognition and little progress has been made for image set based face recognition. More recently, Chen *et al.* [4] presented a discretionary learning method for video-based face recognition, where each face video is first clustered into several clusters and then the dictionary is learned for each cluster. However, their method is unsupervised, which may not be discriminative enough for classification.

3 Proposed Approach

Fig. 1 shows the basic idea of our proposed approach, and the following subsections present the details of the proposed approach.

3.1 SFDL

Generally, there are two key components in an image set based face recognition approach [36,37,30]: image set representation and image set matching. Previous

work [36] has shown that feature learning is an effective tool for image set representation because it can extract discriminative information from face image sets. Recent study [4] has also shown that dictionary learning is a promising solution to image set matching because face images with varying poses, illuminations and expressions within a set can be encoded as dictionaries so that the noise can be effectively alleviated and better matching performance can be obtained. However, most previous image set based face recognition methods learned features and dictionaries separately, which may not be powerful enough because some discriminative information for dictionary learning may be compromised in the feature learning stage, and vice versa. That is because the objective of feature learning is usually inconsistent to that of dictionary learning because feature learning is essentially a feature selection problem while dictionary learning is intrinsically a clustering problem. Hence, it is suboptimal to apply feature learning and dictionary learning for image set based face recognition. To address this shortcoming, we propose a SFDL method to learn discriminative features and dictionaries simultaneously in the following.

Let $X = [X_1, X_2, \cdots, X_P]$ be the training set of face image sets from P different subjects. Assume there are N images in total in the training set by concatenating all the frames from image sets, we rewrite X as $X = [x_1, x_2, \cdots x_N] \in R^{d \times N}$, where x_i is a d-dimensional vector of the cropped face image. To extract more discriminative and robust information from the training set, SFDL aims to simultaneously learn a feature projection matrix and a discriminative structured dictionary to project each image frame in all image sets into a low-dimensional subspace, under which each image frame is encoded by a discriminative coding coefficient. To achieve this, we formulate the following optimization problem:

$$\begin{aligned}\min_{W,D,A} J &= J_1 + \lambda_1 J_2 + \lambda_2 J_3 \\ &= \sum_{i=1}^{N} (\|Wx_i - D\alpha_i\|_2^2 + \eta_1 \|\alpha_i\|_1) + \sum_{p=1}^{P} \sum_{i=1}^{N_p} \|Wx_{ip} - D_p \alpha_{ip}^p\|_2^2 \\ &+ \lambda_1 \sum_{i=1}^{N} (\|W^T W x_i - x_i\|_2^2 + \eta_2 \sum_{j=1}^{k} h(W_j x_i)) \\ &+ \lambda_2 \sum_{i=1}^{N} \sum_{j=1}^{N} \|\alpha_i - \alpha_j\|_2^2 S_{ij} \end{aligned} \tag{1}$$

where W is the feature projection matrix, $D = [D_1, D_2, \cdots, D_P]$ is the structured dictionary, D_p is the sub-dictionary for the pth class in D, x_{ip} is the ith raw pixel sample from the pth class, N_p is the number of samples in the pth class, $A = [\alpha_1, \alpha_2, \cdots, \alpha_N]$ is the sparse representation of the training samples in X, α_i is the coefficient vector of x_i, α_{ip}^p is the representation coefficient vector of α_i from the pth class, h is a nonlinear convex function which is defined as a smooth l_1 penalty: $h(\cdot) = \log(\cosh(\cdot))$ [25], λ_1, λ_2, η_1 and η_2 are four parameters to balance the importance of different terms, S is an affinity matrix to measure the

similarity of the sparse codes α_i and α_j according to their label and appearance information, which is defined as follows:

$$S_{ij} = \begin{cases} 1, & \text{if } x_i \in N_{k_1}(x_j) \text{ or } x_j \in N_{k_1}(x_i) \\ & \text{and } l(x_i) = l(x_j) \\ -1, & \text{if } x_i \in N_{k_2}(x_j) \text{ or } x_j \in N_{k_2}(x_i) \\ & \text{and } l(x_i) \neq l(x_j) \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

where $N_k(x_i)$ and $l(x_i)$ denote the k-nearest neighbors and the label of x_i, respectively.

The first term J_1 in Eq. (1) is to ensure that for each face sample x_i from the pth class in the low-dimensional feature subspace, it is not only well reconstructed by the whole dictionary D, but also the sub-dictionary D_p of the pth class. The second term J_2 in Eq. (1) is to ensure that the feature projection matrix W can preserve the energy of each x_i as much as possible and each column in W is to be as sparse as possible. The third term J_3 in Eq. (1) is to ensure that the difference of the sparse codes of two face images is minimized if they are from the same class and look similar, and the difference of the sparse codes of two face images is maximized if they are from different classes and also look similar, such that discriminative information can be discovered when learning sparse representation coefficients.

We rewrite A as $A = [A_1, A_2, \cdots, A_P]$, where A_p denotes the sub-matrix from the pth class containing the coding coefficients of X_p over D. Let A_p^p be the coding coefficient of X_p over the sub-dictionary D_p. Then, J_1 in Eq. (1) can be re-written as follows:

$$\begin{aligned} J_1 &= \sum_{p=1}^{P} (\|WX_p - DA_p\|_F^2 + \|WX_p - D_p A_p^p\|_F^2) + \eta_1 \|A\|_1 \\ &= \sum_{p=1}^{P} G_p(W, X_p, D, A_p) + \eta_1 \|A\|_1 \end{aligned} \tag{3}$$

where

$$G_p(W, X_p, D, A_p) \triangleq \|WX_p - DA_p\|_F^2 + \|WX_p - D_p A_p^p\|_F^2 \tag{4}$$

We can also simplify J_2 and J_3 in Eq. (1) as follows:

$$J_2 = \|W^T WX - X\|_2^2 + \eta_2 H(WX) \tag{5}$$

$$J_3 = tr(A^T CA) - tr(A^T SA) = tr(A^T LA) \tag{6}$$

where $H(Z)$ is the sums of the outputs of the nonlinear convex function h which is applied on all elements in the matrix Z, $C = diag\{c_1, c_2, \cdots, c_N\}$ is a diagonal matrix whose diagonal elements are the sums of the row elements of S, and $L = C - S$.

Combining Eqs. (4)-(6) into Eq. (1), we have the following SFDL model:

$$\min_{W,D,A} J = \sum_{p=1}^{P} G_p(W, X_p, D, A_p) + \eta_1 \|A\|_1 + \lambda_1(\|W^T W X - X\|_2^2 + \eta_2 H(WX)) + \lambda_2 tr(A^T L A) \quad (7)$$

While the objective function in Eq. (7) is not convex for W, D and A simultaneously, it is convex to one of them when the other two are fixed. Following the work in [20], [42], [14], [24], we iteratively optimize W, D and A by using the following three-stage method.

Step 1: Learn W with fixed D and A: when D and A are fixed, Eq. (7) can be rewritten as

$$\min_{W} J = \sum_{p=1}^{P}(\|WX_p - DA_p\|_F^2 + \|WX_p - D_p A_p^p\|_F^2) + \lambda_1(\|W^T W X - X\|_2^2 + \eta_2 H(WX)) \quad (8)$$

Eq. (8) is an unconstrained optimization problem and many existing fast unconstrained optimizers can be applied to solve this problem. In our implementations, we use the conjugate gradient decent method in [25] to get W.

Step 2: Learn A with fixed W and D: when W and D are fixed, Eq. (7) can be rewritten as

$$\min_{A} J = \sum_{p=1}^{P}(\|Y_p - DA_p\|_F^2 + \|Y_p - D_p A_p^p\|_F^2) + \eta_1 \|A\|_1 + \lambda_2 tr(A^T L A) \quad (9)$$

where $Y_p = WX_p$ is the projection of X_p in the feature space. We compute A_p sequentially by fixing the other coefficient matrices A_q ($q \neq p$, and $1 \leq q \leq P$). Then, Eq. (9) can be simplified as

$$\min_{A_p} J = \|Y_p - DA_p\|_F^2 + \|Y_p - D_p A_p^p\|_F^2 + \eta_1 \|A_p\|_1 + \lambda_2 tr(A_p^T L A_p) \quad (10)$$

Following the work in [26], we optimize each α_{ip} in A_p alternatively. To obtain each α_{ip}, we fix the encoding coefficients α_{jp} ($j \neq i$) for other samples, and rewrite Eq. (10) as

$$\min_{\alpha_{ip}} J = \|Y_p - D\alpha_{ip}\|_F^2 + \|Y_p - D_p \alpha_{ip}^p\|_F^2 + \eta_1 \|\alpha_{ip}\|_1 + \lambda_2 F(\alpha_{ip}) \quad (11)$$

where

$$F(\alpha_{ip}) = \lambda_2(\alpha_{ip}^T (A_p L_i) + (A_p L_i)^T \alpha_{ip} - \alpha_{ip}^T L_{ii} \alpha_{ip}) \quad (12)$$

L_i is the ith column of L, and L_{ii} is the entry in the ith row and ith column of L. We apply the feature sign search algorithm [26] to solve α_{ip}.

Input: Training set $X = [X_1, X_2, \cdots, X_P]$, affinity matrix S, parameters λ_1, λ_2, η_1 and η_2, iteration number T, convergence error ϵ.
Output: Feature weighting matrix W, dictionary D, and coding coefficient matrix A.
Step 1 (Initialization):
1.1: Initialize each column d_p^i in D_p as a random vector with unit l2-norm.
1.2: Initialize each column in A as a random vector.
Step 2 (Local optimization):
For $t = 1, 2, \cdots, T$, repeat
2.1. Solve W^t with fixed D^{t-1} and A^{t-1} via Eq. (8).
2.2. Solve A^t with fixed W^t and D^{t-1} via Eq. (11).
2.3. Solve D^t with fixed W^t and A^t via Eq. (14).
2.3. If $|D^t - D^{t-1}| < \epsilon$ or $|W^t - W^{t-1}| < \epsilon$ and $t > 2$, go to Step 3.
Step 3 (Output):
Output $W = W^t$, $D = D^t$, and $A = A^t$.

Algorithm 1. SFDL

Step 3: Learn D with fixed W and A: when W and A are fixed, Eq. (7) can be rewritten as

$$\min_D J = \sum_{p=1}^{P} (\|Y_p - DA_p\|_F^2 + \|Y_p - D_p A_p^p\|_F^2) \tag{13}$$

We update D_p sequentially by fixing the other sub-dictionaries D_q ($q \neq p$, and $1 \leq q \leq P$). Then, Eq. (13) can be reduced as

$$\min_{D_p} J = \|Y_p - D_p A_p\|_F^2 + \|Y_p - D_p A_p^p\|_F^2 \tag{14}$$

We restrict that each column d_p^i in D_p is a unit vector, where $1 \leq i \leq K_p$, K_p is the number of atoms in D_p. Eq. (14) is a quadratic programming problem and can be solved by using the algorithm in [43], which updates D_p atom by atom.

We repeat the above three steps until the algorithm is convergent. The proposed SFDL algorithm is summarized in **Algorithm 1**.

3.2 Identification

Given a testing face video $X^q = [x_1^q, x_2^q, \cdots, x_M^q]$, where x_j^q is the jth ($1 \leq j \leq M$) frame of this video and M is the number of image frames in this video, we first apply the learned feature projection matrix W to project each frame x_j^q in this video into a feature and recognize its label by using the smallest reconstruction error corresponding to each sub-dictionary D_p ($q \leq p \leq P$), which is computed as follows:

$$p' = \arg\min_p \|W x_j^q - D_p D_p^\dagger x_j^q\|_2 \tag{15}$$

where $D_p^\dagger = (D_p^T D_p)^{-1} D_p^T$ is the pseudoinverse of D_p.

Then, we adopt the majority voting strategy to classify the whole testing face video:

$$p^* = \arg\max_p Z_p \tag{16}$$

where Z_p is the total number of votes from the pth class.

3.3 Verification

Different from face identification, the goal of video face verification is to determine whether a pair of given face videos belongs to the same person or not. Assume $X^a = [x_1^a, x_2^a, \cdots, x_{M_1}^a]$ and $X^b = [x_1^b, x_2^b, \cdots, x_{M_2}^b]$ be the given testing face video pair, x_i^a and x_j^b are the ith and jth frames of these two videos, M_1 and M_2 are the number of image frames in these two videos, $1 \leq i \leq M_1$, $1 \leq j \leq M_2$, we first apply the learned feature projection matrix W to project each x_i^a and x_j^b in these two videos into a low-dimensional feature and recognize their labels by using the smallest reconstruction error corresponding to each sub-dictionary D_p $(1 \leq p \leq P)$ as defined in Eq. (15). Then, we compute the number of votes from each class for these two videos by counting the labels of all frames in each video and get two voting vectors $H^a = [h_1^a, h_2^a, \cdots, h_P^a]$ and $H^b = [h_1^b, h_2^b, \cdots, h_P^b]$, where h_p^a and h_p^b denote the total voting number of votes from the pth class of X^a and X^b, respectively. Lastly, the intersection metric is applied to measure the similarity of the normalized H^a and H^b as follows:

$$s(H^a, H^b) = \sum_{p=1}^{P} \min(\bar{h_p^a}, \bar{h_p^b}) \tag{17}$$

where $\bar{h_p^a} = \frac{1}{M_1} h_p^a$ and $\bar{h_p^a} = \frac{1}{M_2} h_p^b$.

4 Experimental Results

We evaluate our proposed approach on four publicly available video face databases including the Honda [27], Mobo [13], YouTube Celebrities (YTC) [22] and YouTube Face (YTF) [40] datasets. The Honda, MoBo, and YTC datasets are used to show the effectiveness of our approach for face classification with image sets, and the YTF dataset is selected to show the effectiveness of our approach to face verification with image sets.

4.1 Datasets

The Honda dataset [27] contains 59 face videos of 20 subjects, where there are large pose and expression variations and the average length of these videos are approximately 400 frames.

There are 96 videos from 24 subjects in the MoBo dataset [13]. For each subject, four videos corresponding to different walking patterns on a treadmill

such as slow, fast, inclined and carrying a ball were captured and each video corresponds to one walking pattern. For each video, there are around 300 frames covering pose and expression variations.

The YTC dataset [22] contains 1910 video sequences of 47 celebrities (actors, actresses and politicians) which are collected from YouTube. Most videos are low resolution which leads to noisy and low-quality image frames. The number of frames for these videos varied from 8 to 400.

The YTF dataset [40] contains 3425 videos of 1596 subjects which are also downloaded from YouTube. The average length of each video clip is about 180 frames. There are large variations in pose, illumination, and expression, and resolution in these videos.

For face videos in the Honda, Mobo and YTC datasets, each image frame is first automatically detected by applying the face detector method proposed in [35] and then resized to a 30×30 intensity image. For the YTF dataset, each image frame was cropped into 30×30 according to the provided eye coordinates. Hence, each video is represented as an image set. For each image frame in all these four datasets, we only perform histogram equalization to remove the illumination effect.

4.2 Experimental Settings

To make a fair comparison with state-of-the-art image set based face recognition methods, we follow the same protocol used in [38], [36], [3], [17], [37], [40]. On the Honda, MoBo, and YTC datasets, we conduct experiments 10 times by randomly selecting training and testing sets, compute and compare the average identification rate. For both the Honda and MoBo datasets, we randomly select one face video per person to construct the training set and the remaining videos as the testing set. For the YTC dataset, we equally divide the whole dataset into five folds (with minimal overlapping), and each fold contains 9 videos for each person. For each fold, we randomly select 3 face videos for each person for training and use the remain 6 for testing. For the YTF dataset, we follow the standard evaluation protocol and evaluate our approach by using 5000 video pairs which were randomly selected in [40], where half of them are from the same person and the remaining half are from different persons. These pairs are equally divided into 10 folds and each fold contains 250 intra-personal pairs and 250 inter-personal pairs. We also use the 10-fold cross validation strategy in our experiments [40]. Specifically, we use 6 folds from the 9 folds in the training set to train the SFDL model and the rest 3 to learn a discriminative distance metric by using the method in [21].

In our implementations, the feature dimension of W and the parameters λ_1, λ_2, η_1 and η_2 of our SFDL method were empirically specified as 200, 1, 1, 0.05, and 0.2, respectively, and the number of atoms per person (K_p) for different datasets are summarized in Table 1.

Table 1. Summary of number of atoms per person (K_p) for different face datasets in our experiments

Dataset	Honda	MoBo	YTC	YTF
K_p	20	25	35	40

4.3 Results and Analysis

Comparison with Existing State-of-the-Art Image Set Based Face Recognition Methods: We compare our approach with ten state-of-the-art image set based face recognition methods, including Mutual Subspace Method (MSM) [41], Discriminant Canonical Correlation analysis (DCC) [23], Manifold-to-Manifold Distance (MMD) [38], Manifold Discriminant Analysis (MDA) [36], Affine Hull based Image Set Distance (AHISD) [3], Convex Hull based Image Set Distance (CHISD) [3], Sparse Approximated Nearest Point (SANP) [17], Covariance Discriminative Learning (CDL) [37], Dictionary-based Face Recognition from Video (DFRV) [4], and Local Multi-Kernel Metric Learning (LMKML) [30].

The standard implementations of all the other compared methods were provided by the original authors except the CDL and DFRV methods because their codes have not been publicly available. We carefully implemented their methods by following their settings in [37] and [4]. We tuned the parameters of different methods as follows: For MSM and DCC, we performed PCA to learn a linear subspace for each face image set where each subspace dimension was set as 10 to preserve 90% of the energy to compute the similarity of two image sets. For MMD and MDA, the parameters were configured according to [38] and [36], respectively. Specifically, the maximum canonical correlation was used to compute MMD, and the number of connected nearest neighbors for computing geodesic distance in both MMD and MDA was fixed as 12. No parameter is required in AHISD. For CHISD and SANP, we follow the same parameter settings as those in [3] and [17]. For CDL, the KLDA was employed for discriminative learning and the regularization parameter was set the same as that in [37]. For DFRV, we followed the parameter settings in [4]. For the DCC, CDL and LMKML methods, if there is a single video from each class in the Honda, MoBo and YTF datasets, we randomly and equally divided each video clip into two image sets to model the within-class variation.

Table 2 tabulates the average recognition rates of different image set based face recognition methods on these four datasets. We see that our approach performs better than the other ten compared image set based face recognition methods on the Honda, MoBo, and YTF datasets, and achieves comparable results on the YTC dataset. Compared with the existing unsupervised image set based face recognition methods such as MSM, DCC, MMD, AHISD, CHISD, SANP, and DFRV, our SFDL can exploit more discriminative information in the learned feature projection matrix and dictionary. Compared with the existing supervised image set based face recognition methods such as MDA, CDL, and LMKML, our SFDL can project each image frame into a discriminative feature subspace

Table 2. Average recognition rates (%) of different image set based face recognition methods on different video face datasets

Method	Honda	MoBo	YTC	YTF	Year
MSM [41]	92.5	85.5	61.5	62.5	1998
DCC [23]	94.9	88.1	64.8	70.8	2007
MMD [38]	94.9	91.7	66.7	65.0	2008
MDA [36]	97.4	94.4	68.1	72.5	2009
AHISD [3]	89.5	94.1	66.5	66.5	2010
CHISD [3]	92.5	95.8	67.4	66.3	2010
SANP [17]	93.6	96.1	68.3	63.7	2011
CDL [37]	97.4	87.5	69.7	74.5	2012
DFRV [4]	97.4	94.4	74.5	78.6	2012
LMKML [30]	98.5	96.3	78.2	77.8	2013
SFDL	**100.0**	**96.7**	**76.7**	**80.2**	

Table 3. Average recognition rates (%) of different feature and dictionary learning strategies on different face datasets

Method	Honda	MoBo	YTC	YTF
Structured IFDL	98.3	94.1	74.3	78.5
Structured SFDL	**100.0**	**96.7**	**76.7**	**80.2**

and encode it with a class-specific dictionary, so that more person-specific information can be extracted.

Simultaneous vs. Individual Feature and Dictionary Learning: The feature learning and dictionary learning can also be learned in an individual manner. To show the effect of SFDL, we compare our SFDL method with the individual feature and dictionary learning (IFDL) method. IFDL means the feature projection matrix and the structured dictionaries are learned from the training set separately. Table 3 tabulates the average recognition rates of these two methods. We can observe that our simultaneous method can achieve higher recognition rate than the individual method, which shows that jointly learning the feature subspace and dictionary is better because some useful information for dictionary learning may be lost in the feature learning phase in the individual method.

Structured vs. Shared SFDL: To demonstrate the advantage of the structured dictionary in our SFDL, we also compare it with a shared SFDL method which learns a common dictionary in SFDL rather than a structured dictionary. Table 4 tabulates the average recognition rates of these two types of SFDL methods. We can observe that the structured SFDL achieves higher recognition rate than the shared SFDL method. This is because the structured SFDL can characterize more class-specific information than the shared SFDL.

Parameter Analysis: We first evaluate the effect of the feature dimension of the learned feature projection matrix of our SFDL on the recognition performance. Fig. 2 shows the recognition accuracy of our SFDL versus different

Table 4. Average recognition rates (%) of the structured and shared dictionary learning methods on different face datasets

Method	Honda	MoBo	YTC	YTF
Shared SFDL	98.3	95.3	74.7	78.9
Structured SFDL	**100.0**	**96.7**	**76.7**	**80.2**

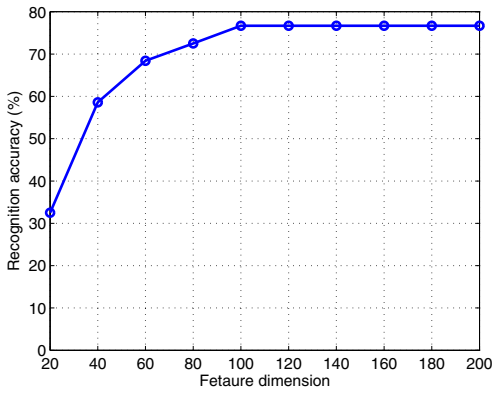

Fig. 2. Average recognition rate (%) of our SFDL versus different feature dimension of the learned feature projection matrix on the YTC dataset

feature dimensions on the YTC dataset. We can see that our proposed SFDL can achieve stable performance when the feature dimension reaches 100.

We also investigate the performance of our SFDL versus different number of iterations. Fig. 3 shows the recognition accuracy of our SFDL over different number of iterations on the YTC dataset. We see that our proposed SFDL can achieve stable performance in several iterations.

Robustness Analysis: We first test the robustness of our proposed approach versus different amount of noisy data in face videos. We follow the settings in [3], [37], [30] and conducted three experiments where the training and/or testing face image sets were corrupted by adding one image from each of the other classes. The original data and three noisy scenarios are called as "original", "NTR" (only training videos have noisy data), "NTE" (only testing videos have noisy data), and "NTT" (both training and testing videos have noisy data), respectively. Table 5 records the recognition accuracy of different image set based face recognition methods with different amounts of noisy data on the YTC dataset.

We also evaluate the performance of our approach when face videos contain varying number of image frames. We randomly selected N frames from each face image set (both training and testing) and used them for recognition. If there are less than N image frames for one face image set, all image frames within this image set were used for recognition. Fig. 4 shows the performance of different methods on the YTC dataset with varying image frames. From Table 5 and

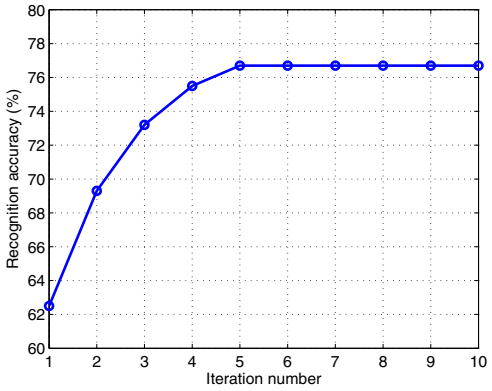

Fig. 3. Average recognition rate (%) of our approach versus different number of iterations on the YTC dataset

Table 5. Average recognition rates (%) of different image set based face recognition methods with different amounts of noisy data on the YTC dataset

Method	Original	NTR	NTE	NTT
MSM [41]	62.8	59.7	45.3	52.2
DCC [23]	64.8	58.7	49.9	54.2
MMD [38]	66.7	62.5	46.4	55.4
MDA [36]	68.1	65.8	52.5	53.4
AHISD [3]	66.5	62.5	44.5	35.6
CHISD [3]	67.4	66.8	42.5	38.5
SANP [17]	68.3	67.2	47.5	39.4
CDL [37]	69.7	68.4	54.5	58.4
DFRV [4]	74.5	71.1	60.8	62.1
LMKML [30]	78.2	76.1	64.5	66.1
SFDL	**76.7**	**76.3**	**64.8**	**67.2**

Fig. 4, we observe that our approach demonstrates strong robustness with some slight performance drop than the other compared methods. That is because we use dictionaries to represent each face image set and such dictionary-based methods are robust to noise and the number of samples in face image set. Hence, the effects of the noisy samples and varying data size can be alleviated in our proposed approach.

Computational Time: Lastly, we report the computational time of different methods using the YTC dataset. Our hardware configuration is a 2.8-GHz CPU and a 24GB RAM. Table 6 shows the computational time for different methods under the Matlab paltform. It is to be noted that training time is only required for some discriminative learning and dictionary learning methods such as DCC, MDA, CDL, DFRV, LMKML and our SFDL. We see that the computational

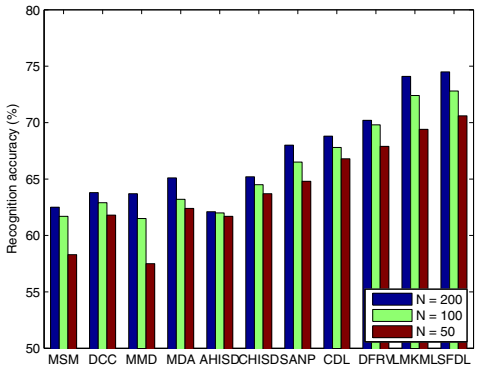

Fig. 4. Average recognition rates (%) of different image set based face recognition methods with different number of image frames on the YTC dataset

Table 6. Computation time (seconds) of different image set based face recognition methods on the YTC dataset for the training and testing phases per face video

Method	MSM	DCC	MMD	MDA	AHISD	CHISD	SANP	CDL	DFRV	LMKML	SFDL
Training	N.A.	98.6	N.A.	185.3	N.A.	N.A.	N.A.	68.5	8656.5	4232.8	7532.5
Testing	2.7	2.5	3.5	3.2	8.7	6.7	48.6	12.8	5.4	210.6	6.5

time of our SFDL is generally larger than many other compared methods for training and is comparable to them for testing.

5 Conclusion and Future Work

In this paper, we propose a new simultaneous feature and dictionary learning (SFDL) method for image set based face recognition. By jointly learning the feature projection matrix and the structured dictionary, our approach extracts more discriminative information for image set based face representation. Experimental results on four widely used face datasets have shown the superiority of our approach over the state-of-the-art image set based face recognition methods in terms of accuracy and robustness. How to design more efficient optimization methods to improve the speed of our SFDL method appears to be an interesting future work.

Acknowledgement. This work is supported by the research grant for the Human Cyber Security Systems (HCSS) Program at the Advanced Digital Sciences Center from the Agency for Science, Technology and Research of Singapore, and the research grant of MOE Tier 1 RG84/12, MOE Tier 2 ARC28/14 and SERC 1321202099.

References

1. Aharon, M., Elad, M., Bruckstein, A.: Ksvd: An algorithm for designing overcomplete dictionaries for sparse representation. TSP 54(11), 4311–4322 (2006)
2. Arandjelovic, O., Shakhnarovich, G., Fisher, J., Cipolla, R., Darrell, T.: Face recognition with image sets using manifold density divergence. In: CVPR, pp. 581–588 (2005)
3. Cevikalp, H., Triggs, B.: Face recognition based on image sets. In: CVPR, pp. 2567–2573 (2010)
4. Chen, Y.-C., Patel, V.M., Phillips, P.J., Chellappa, R.: Dictionary-based face recognition from video. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part VI. LNCS, vol. 7577, pp. 766–779. Springer, Heidelberg (2012)
5. Chen, Y.C., Patel, V.M., Phillips, P.J., Chellappa, R.: Dictionary-based face recognition from video. Tech. rep., University of Maryland (2013)
6. Chen, Y.C., Patel, V.M., Shekhar, S., Chellappa, R., Phillips, P.J.: Video-based face recognition via joint sparse representation. In: FG, pp. 1–8 (2013)
7. Chin, T.J., Schindler, K., Suter, D.: Incremental kernel svd for face recognition with image sets. In: FG, pp. 461–466 (2006)
8. Cui, Z., Shan, S., Zhang, H., Lao, S., Chen, X.: Image sets alignment for video-based face recognition. In: CVPR, pp. 2626–2633 (2012)
9. Deng, W., Hu, J., Guo, J.: Extended src: Undersampled face recognition via intraclass variant dictionary. PAMI 34(9), 1864–1870 (2012)
10. Deng, W., Hu, J., Lu, J., Guo, J.: Transform-invariant pca: A unified approach to fully automatic face alignment, representation, and recognition. PAMI 36(6), 1275–1284 (2014)
11. Fan, W., Yeung, D.: Locally linear models on face appearance manifolds with application to dual-subspace based classification. In: CVPR, pp. 1384–1390 (2006)
12. Fitzgibbon, A., Zisserman, A.: Joint manifold distance: a new approach to appearance based clustering. In: CVPR, pp. 26–33 (2003)
13. Gross, R., Shi, J.: The cmu motion of body (mobo) database. Tech. rep., Carnegie Mellon University (2001)
14. Guo, H., Jiang, Z., Davis, L.S.: Discriminative dictionary learning with pairwise constraints. In: Lee, K.M., Matsushita, Y., Rehg, J.M., Hu, Z. (eds.) ACCV 2012, Part I. LNCS, vol. 7724, pp. 328–342. Springer, Heidelberg (2013)
15. Hadid, A., Pietikainen, M.: From still image to video-based face recognition: an experimental analysis. In: FG, pp. 813–818 (2004)
16. Hotelling, H.: Relations between two sets of variates. Biometrika 28(3/4), 321–377 (1936)
17. Hu, Y., Mian, A., Owens, R.: Sparse approximated nearest points for image set classification. In: CVPR, pp. 121–128 (2011)
18. Hu, Y., Mian, A.S., Owens, R.: Face recognition using sparse approximated nearest points between image sets. PAMI 34(10), 1992–2004 (2012)
19. Huang, L., Lu, J., Tan, Y.P.: Co-learned multi-view spectral clustering for face recognition based on image sets. IEEE Signal Processing Letters 21(7), 875–879 (2014)
20. Jiang, Z., Lin, Z., Davis, L.S.: Learning a discriminative dictionary for sparse coding via label consistent k-svd. In: CVPR, pp. 1697–1704 (2011)
21. Kan, M., Shan, S., Xu, D., Chen, X.: Side-information based linear discriminant analysis for face recognition. In: BMVC, pp. 1–12 (2011)
22. Kim, M., Kumar, S., Pavlovic, V., Rowley, H.: Face tracking and recognition with visual constraints in real-world videos. In: CVPR, pp. 1–8 (2008)

23. Kim, T., Kittler, J., Cipolla, R.: Discriminative learning and recognition of image set classes using canonical correlations. PAMI 29(6), 1005–1018 (2007)
24. Kong, S., Wang, D.: A dictionary learning approach for classification: Separating the particularity and the commonality. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part I. LNCS, vol. 7572, pp. 186–199. Springer, Heidelberg (2012)
25. Le, Q.V., Karpenko, A., Ngiam, J., Ng, A.: Ica with reconstruction cost for efficient overcomplete feature learning. In: NIPS, pp. 1017–1025 (2011)
26. Lee, H., Battle, A., Raina, R., Ng, A.: Efficient sparse coding algorithms. In: NIPS, pp. 801–808 (2006)
27. Lee, K., Ho, J., Yang, M., Kriegman, D.: Video-based face recognition using probabilistic appearance manifolds. In: CVPR, pp. 313–320 (2003)
28. Lin, T., Liu, S., Zha, H.: Incoherent dictionary learning for sparse representation. In: ICPR, pp. 1237–1240 (2012)
29. Lu, J., Tan, Y.P., Wang, G., Yang, G.: Image-to-set face recognition using locality repulsion projections and sparse reconstruction-based similarity measure. TCSVT 23(6), 1070–1080 (2013)
30. Lu, J., Wang, G., Moulin, P.: Image set classification using multiple order statistics features and localized multi-kernel metric learning. In: ICCV, pp. 1–8 (2013)
31. Ma, L., Wang, C., Xiao, B., Zhou, W.: Sparse representation for face recognition based on discriminative low-rank dictionary learning. In: CVPR, pp. 2586–2593 (2012)
32. Mairal, J., Bach, F., Ponce, J., Sapiro, G., Zisserman, A.: Discriminative learned dictionaries for local image analysis. In: CVPR, pp. 1–8 (2008)
33. Shakhnarovich, G., Fisher III, J.W., Darrell, T.: Face recognition from long-term observations. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002, Part III. LNCS, vol. 2352, pp. 851–865. Springer, Heidelberg (2002)
34. Tosic, I., Frossard, P.: Dictionary learning. IEEE Signal Processing Magazine 28(2), 27–38 (2011)
35. Viola, P., Jones, M.: Robust real-time face detection. IJCV 57(2), 137–154 (2004)
36. Wang, R., Chen, X.: Manifold Discriminant Analysis. In: CVPR, pp. 1–8 (2009)
37. Wang, R., Guo, H., Davis, L., Dai, Q.: Covariance discriminative learning: A natural and efficient approach to image set classification. In: CVPR, pp. 2496–2503 (2012)
38. Wang, R., Shan, S., Chen, X., Gao, W.: Manifold-manifold distance with application to face recognition based on image set. In: CVPR, pp. 1–8 (2008)
39. Wang, X., Wang, B., Bai, X., Liu, W., Tu, Z.: Max-margin multiple-instance dictionary learning. In: ICML, pp. 846–854 (2013)
40. Wolf, L., Hassner, T., Maoz, I.: Face recognition in unconstrained videos with matched background similarity. In: CVPR, pp. 529–534 (2011)
41. Yamaguchi, O., Fukui, K., Maeda, K.: Face recognition using temporal image sequence. In: FG, pp. 318–323 (1998)
42. Yang, M., Zhang, L., Feng, X., Zhang, D.: Fisher discrimination dictionary learning for sparse representation. In: ICCV, pp. 543–550 (2011)
43. Yang, M., Zhang, L., Yang, J., Zhang, D.: Metaface learning for sparse representation based face recognition. In: ICIP, pp. 1601–1604 (2010)
44. Yang, M., Zhu, P., Van Gool, L., Zhang, L.: Face recognition based on regularized nearest points between image sets. In: FG, pp. 1–7 (2013)
45. Zhang, Q., Li, B.: Discriminative k-svd for dictionary learning in face recognition. In: CVPR, pp. 2691–2698 (2010)
46. Zuo, Z., Wang, G.: Learning discriminative hierarchical features for object recognition. IEEE Signal Processing Letters 21(9), 1159–1163 (2014)