Learning a Structured Dictionary for Video-based Face Recognition

Hongyu Xu, Jingjing Zheng, Azadeh Alavi and Rama Chellappa
Department of Electrical and Computer Engineering
and the Center for Automation Research, UMIACS
University of Maryland, College Park, MD, USA
{hyxu, zjngjng, azadeh, rama}@umiacs.umd.edu

Abstract

In this paper, we propose a structured dictionary learning framework for video-based face recognition. We discover the invariant structural information from different videos of each subject. Specifically, we employ dictionary learning and low-rank approximation to preserve the invariant structure of face images in videos. The learned dictionary is both discriminative and reconstructive. Thus, we not only minimize the reconstruction error of all the face images but also encourage a sub-dictionary to represent the corresponding subject from different videos. Moreover, by introducing the low-rank approximation, the proposed method is able to discover invariant structured information from different videos of the same subject. To this end, an efficient alternating algorithm is employed to learn our structured dictionary. Extensive experiments on three video-based face recognition databases show that our approach outperforms several state-of-the-art methods.

1. Introduction

Video-based face recognition has become a very popular topic of research in recent years [8, 9, 34, 15, 17, 19, 30, 31, 39]. Given a video sequence, the objective is to recognize the person in the video. It is often interchanged with image-set based face recognition [4, 20, 16, 10, 32, 33, 25, 26, 38, 7], when the image sets are sampled from videos. Compared with single image-based face recognition, a video provides more samples from frames containing the person of interest. However, it brings more challenges as videos are often acquired in unconstrained environments, under significant variations in poses, expressions, lighting conditions and backgrounds. These variations result in large intra-personal variations within a video sequence. Therefore, it is important to represent and model the same subject against these variations in videos.

Numerous methods have been proposed to exploit useful information contained in videos. Early approaches [2, 28, 20, 19, 22] addressed this problem through learning probabilistic models. This was then followed by computing the similarity between two videos to perform recognition. Later, more sophisticated statistical model-based approaches [4, 17, 16, 26, 32, 30, 31, 33] were proposed to learn discriminative and compact representations for each subject.

Recent works have shown that dictionary-based methods achieve impressive performance in various tasks, such as image-based face recognition, object and action recognition [1, 13, 11, 21, 27, 37, 40, 41, 42, 18, 35]. This is due to the fact that images could be well represented by an approximately learned dictionary and related sparse codes. However, there are only a few reported efforts on video-based face recognition [8, 39, 24]. Recently, [8] proposed to partition videos into several clusters and learned a separate sub-dictionary for each cluster. One limitation of this method is that the number of clusters needs to be predefined. [24] jointly learned a global projection matrix and a set of sub-dictionaries to encode the new features with discriminative sparse coefficients. However, this method suffers from high computational complexity. In addition, information useful for dictionary learning may be lost after projecting all the samples onto the same subspace. [39] learned a sub-dictionary along with a low-rank representation for each subject. However, the sub-dictionaries were independently learned and are not discriminative enough for classification.

To overcome the challenges discussed above, we propose a structured dictionary learning framework for video-based face recognition. The learned dictionary has the following three properties. First, it is reconstructive. We minimize the errors of all the face images when reconstructed from the dictionary, which encourages the learned dictionary to be reconstructive. Second, it is discriminative. For face images from each subject, we not only enforce the corresponding sub-dictionary to well represent them, but also enforce other sub-dictionaries not to be used for reconstruction. This will encourage different sub-dictionaries to en-

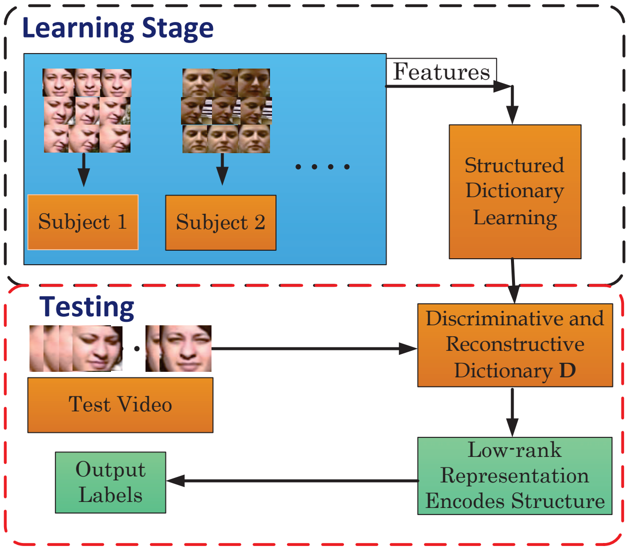

Learning Stage
Features
Subject 1
Subject 2
....
Structured Dictionary Learning
Testing
Discriminative and Reconstructive Dictionary **D**
Test Video
Low-rank Representation Encodes Structure
Output Labels

Figure 1. Overview of the proposed approach

code features from different subjects. Third, it is capable of discovering invariant structured information from different videos of the same subject. This is achieved by minimizing the rank of the representation matrix of face images from each subject. It is known that face images from one subject in different videos share some similar characteristics (*i.e.* consecutive pose change or similar facial appearance), which could be exploited to derive a low-dimensional subspace representation. Motivated by this underlying feature, we regularize the representation matrix of face images from the same subject in videos to produce a matrix of lower rank compared to the original data matrix. Figure 1 shows the overview of our approach.

To summarize, we make the following contributions:

- We present a dictionary learning approach with both discriminative and reconstructive properties. The learned dictionary reveals that the structural information from video face images could be used for recognition directly.
- Our method learns a low-rank representation for video face images of the same subject, using an efficient alternating optimization algorithm.
- We demonstrate that the proposed algorithm outperforms the state-of-the-art methods on three benchmark databases for video-based face recognition.

Organization of the Paper: The rest of the paper is organized as follows: In Section 2, we review several video-based face recognition methods and related dictionary learning methods. In Section 3, we present our structured dictionary learning approach followed by an efficient optimization algorithm. We evaluate the proposed method for video-based face recognition on three benchmark databases in Section 4. Section 5 concludes the paper with a brief summary.

2. Related Work

Video-based Face Recognition: Existing video-based face recognition approaches [20, 19, 22, 4, 17, 16, 6, 26, 32, 30, 31, 33] can be categorized into two classes: parametric and non-parametric. Early approaches [20, 19, 22] computed the similarity between the query video and training videos based on probabilistic models. Such methods were based on the assumption that a strong statistical correlation existed between the training and testing videos. To overcome the drawback of the probabilistic approaches, non-parametric approaches [4, 17, 16, 26, 32, 30, 31, 33] represented the face images from videos as subspaces or manifolds. Linear/affine subspace-based methods [4, 20, 16, 36, 5, 7] modeled the video face images as a linear or affine subspace. Among them, [4, 36, 7, 16] used convex geometry to represent videos from one subject, yielding improved performance over the parametric approaches. However, to address the limitation of linear subspace models, more sophisticated nonlinear models have been extensively studied [14, 17, 30, 32, 6, 33]. To preserve the nonlinear structure, [14, 17, 32, 30] employed the concept of Grassmann manifolds, which is a special type of Riemannian manifold. [33] proposed more general discriminative analysis on Riemannian Manifold, which achieved encouraging results. A multi-kernel method combined with order statistics to perform classification was presented in [26]. Finally, deep learning approaches [15, 25] have achieved state-of-the-art performance.

Dictionary Learning: Dictionary learning [1, 13, 11, 21, 27, 37, 40, 41, 42, 18] has attracted great interest in subspace modeling for classification purpose. It overcomes the limitation of PCA subspaces by using non-orthogonal atoms (columns) in the dictionary to provide more flexibility to model the data. K-SVD [1] is one of the most common techniques to learn a dictionary. Several algorithms have been developed to make the dictionary more discriminative [40, 37, 27, 18, 41]. [18] proposed a Label Consistent K-SVD to learn a compact dictionary by incorporating the training labels. [41] presented a structured low-rank representation based on a dictionary to boost the classification performance. [37] integrated Fisher discrimination criterion with dictionary learning, which resulted in a more discriminative dictionary and sparse codes.

3. Proposed Approach

In this section, we detail the proposed structured dictionary learning framework. The dictionary learned by our method is both discriminative and reconstructive for video-based face recognition.

3.1. Problem Formulation

Assume that we have videos from P different subjects, and each video contains a sequence of face images. Let the data matrix $X = [X_1, ..., X_P] \in \mathbb{R}^{d\times N}$ denote face images from P different subjects from the given videos, where N is the total number of images. Each $X_i = [x_{i1}, x_{i2}, ..., x_{iN_i}] \in \mathbb{R}^{d\times N_i}, 1 \le i \le P$ be the features of face images from i-th subject identity, and each column is the feature vector extracted from one frame.

We aim to learn a dictionary $D \in \mathbb{R}^{d\times n}$ with both discriminative and reconstructive powers. The dictionary can be further decomposed into a set of sub-dictionaries as $D = [D_1, ..., D_P]$, where n is the number of atoms (columns) in the dictionary; and $D_i \in \mathbb{R}^{d\times n_i}$ is the i-th sub-dictionary corresponding to the i-th subject. We reconstruct the features of face images from each subject X_i using the dictionary D, and obtain the corresponding encoding coefficients $Z_i \in \mathbb{R}^{n\times N_i}$. We can write the coefficient matrix Z_i over the dictionary D as $Z_i = [Z_i^1, Z_i^2, ..., Z_i^P]^T$, where Z_i^j denotes the coefficients of X_i over the sub-dictionary D_j.

We propose to learn a structured dictionary with the following attributes: First, D should have small reconstruction errors for the training samples from all subjects. Second, each sub-dictionary D_i should represent face images only from the i-th subject, while different sub-dictionaries should be exclusive to each other. In order to achieve the above goal, the objective function for learning the dictionary D and representation coefficients Z is formulated as:

$$\begin{aligned} \min_{D,Z,E^1,E^2} & \sum_{i=1}^{P}(\|Z_i\|_* + \lambda_1\|E_i^1\|_1 + \lambda_2\|E_i^2\|_1) \\ s.t.\ \ & X_i = DZ_i + E_i^1, \\ & X_i = D_iZ_i^i + E_i^2, \qquad \forall\ i,\ \ 1 \le i \le P \end{aligned} \tag{1}$$

where $E_i^1 \in \mathbb{R}^{d\times N_i}$ and $E_i^2 \in \mathbb{R}^{d\times N_i}$ are the reconstruction errors of X_i using the dictionary D and sub-dictionary D_i respectively. The parameters λ_1 and λ_2 balance two types of reconstruction error terms. The objective function in (1) leads to a dictionary D with both discriminative and reconstructive powers at the same time, and has three terms:

1. The first term denotes the nuclear norm of Z_i, which is the low-rank approximation of representation Z_i. Minimization of this term enforces the representation Z_i of samples from the i-th subject to lie on the same low-dimensional subspace.

2. The second term E_i^1 is the l_1 norm of the reconstruction error of X_i with respect to dictionary D. We encourage D to be reconstructive, by minimizing the reconstruction errors for samples from all different subjects.

3. The third term is the l_1 norm of the reconstruction error of X_i with respect to the i-th sub-dictionary D_i. By minimizing this reconstruction error term, we encourage the i-th sub-dictionary D_i to represent the samples from its own class, while discouraging the usage of sub-dictionaries $D_j (j \ne i)$ from other classes for reconstruction. This regularization will make the dictionary to be discriminative.

3.2. Optimization

In this section, we present an efficient algorithm to solve the optimization problem in (1). Our proposed algorithm uses the inexact ALM method to take advantage of its efficient convergence speed, for solving the low-rank related problems [3, 23].

In order to make the objective function separable, we first introduce auxiliary variables W_i to replace Z_i $(1 \le i \le P)$. Denote $W = \{W_1, ..., W_P\}$, then the function in (1) could be rewritten as:

$$\begin{aligned} \min_{D,Z,E^1,E^2,W} & \sum_{i=1}^{P}(\|W_i\|_* + \lambda_1\|E_i^1\|_1 + \lambda_2\|E_i^2\|_1) \\ s.t.\ \ & X_i = DZ_i + E_i^1, \\ & X_i = D_iZ_i^i + E_i^2 \\ & Z_i = W_i, \qquad \forall\ i,\ \ 1 \le i \le P \end{aligned} \tag{2}$$

The augmented Lagrangian function L of (2) is:

$$\begin{aligned} & L(D, Z, E^1, E^2, W, Y^1, Y^2, Y^3, \mu) \\ = & \sum_{i=1}^{P}(\|W_i\|_* + \lambda_1\|E_i^1\|_1 + \lambda_2\|E_i^2\|_1) \\ & + \sum_{i=1}^{P}(\langle Y_i^1, X_i - DZ_i - E_i^1\rangle + \langle Y_i^2, X_i - D_iZ_i^i - E_i^2\rangle \\ & + \langle Y_i^3, Z_i - W_i\rangle) + \frac{\mu}{2}\sum_{i=1}^{P}(\|X_i - DZ_i - E_i^1\|_F^2 \\ & + \|X_i - D_iZ_i^i - E_i^2\|_F^2 + \|Z_i - W_i\|_F^2) \end{aligned} \tag{3}$$

where $Y^1 = \{Y_1^1, ..., Y_P^1\}$, $Y^2 = \{Y_1^2, ..., Y_P^2\}$, $Y^3 = \{Y_1^3, ..., Y_P^3\}$ are all the multipliers, $\langle A, B\rangle = trace(A^TB)$ and μ is a positive scalar.

The optimization problem in (3) can be decomposed into two sub-problems and solved using the alternating method as in [41]. In the first sub-problem, the dictionary D is fixed and the optimal Z_i, E_i^1 and E_i^2 $(1 \le i \le P)$ are computed. In the second sub-problem, the Z_i, E_i^1 and E_i^2 $(1 \le i \le P)$ are fixed, and the dictionary D is updated. We alternate the steps of solving the two sub-problems until convergence.

Algorithm 1 First Sub-problem Optimization via Inexact ALM

```
Input: Training data $X = [X_1, ..., X_P]$, Dictionary $D$, parameter $\lambda_1, \lambda_2$
Output: $Z_i, E_i^1, E_i^2, Y_i^1, Y_i^2, Y_i^3$ $(1 \leq i \leq P)$
Initialize: $\forall i = 1, ..., P$, $Z_i = W_i = Y_i^3 = 0$, $E_i^1 = E_i^2 = Y_i^1 = Y_i^2 = 0$, $\mu = 10^{-6}$, $\mu_{max} = 10^7$, $\rho = 1.25$
for class $i = 1, ..., P$ do
    Update $Z_i$,$W_i$,$E_i^1$ and $E_i^2$
    while not converged do
        Fix the others and update $W_i$ according to (6)
        Fix the others and update $E_i^1$ according to (4)
        Fix the others and update $E_i^2$ according to (5)
        Fix the others and update $Z_i$ by
        $Z_i = \left[D^T D + (DM_i)^T (DM_i) + I\right]^{-1} [D^T(X_i - E_i^1) + (DM_i)^T(X_i - E_i^2) + W_i + \frac{1}{\mu}(D^T Y_i^1 + (DM_i)^T Y_i^2 - Y_i^3)]$
        Update Multipliers
        $Y_i^1 = Y_i^1 + \mu(X_i - DZ_i - E_i^1)$
        $Y_i^2 = Y_i^2 + \mu(X_i - D_i Z_i^i - E_i^2)$
        $Y_i^3 = Y_i^3 + \mu(Z_i - W_i)$
        Update $\mu$ by
        $\mu = \min(\rho\mu, \mu_{max})$.
        Check the convergence condition:
        $X_i - DZ_i - E_i^1 \to 0$
        $X_i - D_i Z_i^i - E_i^2 \to 0$
        $Z_i - W_i \to 0$
    end while
end for
```

3.3. Computing Representation Z

Given the dictionary D, the augmented Lagrangian function of (3) could be decomposed as the summation of P different sub-functions, where each sub-function is only associated with one class label i $(1 \leq i \leq P)$. Therefore, all the variables Z_i, E_i^1, E_i^2 and W_i $(1 \leq i \leq P)$ in sub-functions could be updated in a class by class fashion. When updating class i, variables Z_i, E_i^1, E_i^2 and W_i could be obtained as follows:

$$\begin{aligned} E_i^1 &= \arg\min_{E_i^1} \lambda_1 \|E_i^1\|_1 + \langle Y_i^1, X_i - DZ_i - E_i^1 \rangle \\ &\qquad + \frac{\mu}{2} \|X_i - DZ_i - E_i^1\|_F^2 \\ &= \arg\min_{E_i^1} \|E_i^1\|_1 + \frac{\mu}{2\lambda_1} \|(X_i - DZ_i + \frac{Y_i^1}{\mu}) - E_i^1\|_F^2 \end{aligned} \tag{4}$$

Similar to E_i^1, E_i^2 is updated as:

$$E_i^2 = \arg\min_{E_i^2} \|E_i^2\|_1 + \frac{\mu}{2\lambda_2} \|(X_i - D_i Z_i^i + \frac{Y_i^2}{\mu}) - E_i^2\|_F^2 \tag{5}$$

Algorithm 2 Overall Learning Framework

```
Input: Training data $X = [X_1, ..., X_P] \in \mathbb{R}^{d \times N}$, dictionary size $n_0$, parameter $\lambda_1, \lambda_2$
Initialize: Sub-dictionary $D_i$ $(1 \leq i \leq P)$ by using k-SVD [1] algorithm, fix $\epsilon_d = 10^{-4}$
while not converged do
    Update $Z_i$,$W_i$,$E_i^1$ $E_i^2$ $(1 \leq i \leq P)$ class by class using Algorithm 1.
    Update Dictionary $D$ according to (9)
    Check the convergence conditions:
    $\|D^{new} - D^{old}\|_F^2 < \epsilon_d$
end while
Output: Structured dictionary $D$ and representation $Z$
```

W_i is updated as:

$$\begin{aligned} W_i &= \arg\min_{W_i} \|W_i\|_* + \langle Y_i^3, Z_i - W_i \rangle + \frac{\mu}{2} \|Z_i - W_i\|_F^2 \\ &= \arg\min_{W_i} \|W_i\|_* + \frac{\mu}{2} \|(Z_i + \frac{Y_i^3}{\mu}) - W_i\|_F^2 \end{aligned} \tag{6}$$

Specifically, (4), (5) and (6) can be solved by singular value thresholding operation as in [23].

Note that when updating Z_i with other variables fixed, Z_i^i is also the corresponding component in Z_i with respect to the i-th sub-dictionary D_i. Here, we construct a matrix M such that $D_i Z_i^i = DM_i Z_i$, $M_i = \text{diag}(0, ..., 0, I_{n_0}, 0, ..., 0) \in \mathbb{R}^{n \times n}$; where $I_{n_0} \in \mathbb{R}^{n_0 \times n_0}$ located between index $n_0(i-1)+1$ and $n_0 i$. Then we could rewrite (3) as:

$$\begin{aligned} Z_i = \arg\min_{Z_i} & \langle Y_i^1, X_i - DZ_i - E_i^1 \rangle + \langle Y_i^2, X_i - DM_i Z_i - E_i^2 \rangle \\ & + \langle Y_i^3, Z_i - W_i \rangle + \frac{\mu}{2} (\|X_i - DZ_i - E_i^1\|_F^2 \\ & + \|X_i - DM_i Z_i - E_i^2\|_F^2 + \|Z_i - W_i\|_F^2) \end{aligned} \tag{7}$$

The optimization problem in (7) is a quadratic form in the variable Z_i. Consequently we can derive the optimal Z_i by setting the first-order derivative with respect to Z_i to be zero.

The optimization procedure of the first sub-problem is illustrated in Algorithm 1.

3.4. Updating Dictionary D

With a fixed Z_i, E_i^1 and E_i^2 $(1 \leq i \leq P)$, D is the only variable in (3). Denote $A_i = M_i Z_i$, then we could rewrite $D_i Z_i^i = DA_i$, for $A_i = [A_i^1, A_i^2, ..., A_i^P]^T \in \mathbb{R}^{n \times N_i}$; where its component A_i^i corresponding to D_i is equal to Z_i^i, and other components $A_i^j (j \neq i)$ are all zeros. Then the optimization function of D is

$$\min_{D}\sum_{i=1}^{P}(\langle Y_i^1, X_i - DZ_i - E_i^1\rangle + \langle Y_i^2, X_i - DA_i - E_i^2\rangle) + \frac{\mu}{2}\sum_{i=1}^{P}(\|X_i - DZ_i - E_i^1\|_F^2 + \|X_i - DA_i - E_i^2\|_F^2) \tag{8}$$

The function in (8) is a quadratic form in variable D and the optimal solution is obtained as

$$D = \left[\frac{1}{\mu}(Y_i^1 Z_i^T + Y_i^2 A_i^T) + (X_i - E_i^1)Z_i^T + (X_i - E_i^2)A_i^T\right] \times \left[\sum_{i=1}^{P}(Z_i Z_i^T + A_i A_i^T)\right]^{-1} \tag{9}$$

The overall framework is summarized in Algorithm 2.

3.5. Video-based Recognition

Once the discriminative and reconstructive dictionary D is learned, we predict the label of a given query video Y by computing the following terms:

$$Z = \arg\min_{Z} \|Z\|_* + \lambda_1\|E\|_1 \ s.t.\ Y = DZ + E \tag{10}$$

where $Y = [y_1, ..., y_{N_y}] \in \mathbb{R}^{d\times N_y}$, N_y is the total number of face images. Note that during the training stage, D is learned such that each sub-dictionary D_i represents the i-th class, while different sub-dictionaries are exclusive to each other. Therefore, we assign the label p^* with the smallest reconstruction error as:

$$p^* = \arg\min_{p\in 1,...,P}\sum_{k=1}^{N_y}\|y_k - D_p z_k^p\|_2 \tag{11}$$

where y_k is the k-th face image vector in the query video and z_k^p is the sparse coefficient of y_k corresponding to the p-th sub-dictionary D_p.

4. Experiments

In this section, we present experimental results for video-based face recognition on three benchmark database, Honda/UCSD [22], CMU Mobo [12] and YouTube Celebrities [19] databases. We will first introduce three databases and their experimental settings. This is then followed by discussion of the proposed approach.

4.1. Database and Settings

Honda/UCSD [22]: There are in total 59 video sequences of 20 different subjects, where each subject has 2 or 3 video sequences. The video is acquired under large variations in expressions and head poses. Following the protocol in [22, 25, 24, 31], we select one sequence from each subject for training and test on the remaining sequences. We also evaluate our method with different lengths of training frames as in [16, 38, 8] by selecting 50 and 100 frames from each training video. The face detector presented in [29] was used to detect the faces. Faces were resized to 20 × 20 after histogram equalization to remove moderate illumination effect.

Figure 2. Examples of YouTube Celebrities (YTC) database [19]

CMU Mobo [12]: It contains 96 video sequences of 24 subjects. Each subject has 4 video sequences captured in different walking situations. Face images were encoded using Local Binary Pattern (LBP) feature as in [16]. Following the standard protocol as in [4, 20], we randomly selected one video from each subject to train while testing on the rest of all video sequences. This was repeated ten times.

Youtube Celebrities [19]: Youtube Celebrities Video is a widely used challenging database, which contains 1910 video clips of 47 subjects collected from YouTube. Some exemplar video frames are given in Figure 2. Each face is resized to 20 × 20 after using the face detector in [29] and pre-processed by histogram equalization as in [30, 31, 24, 26]. Intensity features are extracted for each face image. We conduct 10-fold cross validation experiments. For each subject, we randomly select 3 video clips for training and 6 for testing in each of the 10 folds. This setting ensures that both training and test data covered the whole video clips of each subject, which is the same with the protocol in [10, 30, 31, 33, 17] and similar to [26, 25].

We set all the sub-dictionaries to have the same number of atoms (columns), *i.e.* $n_i = n_0$. For Honda/UCSD and CMU Mobo databases, we run ten different trails under the standard settings and report the average recognition rate. The parameters λ_1,λ_2 have been empirically set to be 0.1 and 1 respectively. For a fair comparison with other dictionary learning approaches, the dictionary size n_0 is set at 10 for the Honda/UCSD database and at 20 for the CMU Mobo database. For the YTC databese, we employ 10-fold cross validation and report the average recognition rate. Our rates are reported by settings n_0 = 40, $\lambda_1 = 0.02$ and $\lambda_2 = 0.1$.

Methods	DCC [20]	MMD [32]	MDA [30]	AHISD [4]	CHISD [4]	SANP [16]	DFRV [8]	CDL [31]	JDSSL [39]	JRNP [36]	Ours
50 Frames	76.9	69.3	74.4	87.2	82.1	84.6	89.7	87.2	87.2	92.3	**93.6**
100 Frames	84.6	87.2	94.8	84.6	84.6	92.3	97.4	94.3	97.4	**100.0**	**100.0**
Full Length	94.9	97.1	97.4	89.7	92.3	94.8	97.4	**100.0**	**100.0**	**100.0**	**100.0**
Year	2006	2008	2009	2010	2010	2011	2012	2012	2014	2015	

Table 1. Video-based face recognition results for the Honda/UCSD database [22] using different number of frames in each image set for training. Rank-1 recognition accuracy results are presented.

Methods	DCC [20]	MMD [32]	MDA [30]	AHISD [4]	CHISD [4]	SANP [16]	DFRV [8]	CDL [31]	JDSSL [39]	Ours
Accuracy	88.9	92.5	94.4	92.9	96.5	96.1	95.2	94.1	96.3	**98.2**

Table 2. Video-based face recognition results for the CMU Mobo database [12]. Rank-1 recognition accuracy results are presented.

Methods	DCC [20]	MMD [32]	MDA [30]	AHISD [4]	CHISD [4]	SANP [16]	CDL [31]	JDSSL [39]	PML [17]	DARG [33]	Ours
Accuracy	66.8	65.3	67.0	63.7	66.5	65.0	70.1	70.1	70.4	72.5	**72.8**

Table 3. Video-based face recognition results for the YTC database [19]. Rank-1 recognition accuracy results are presented.

4.2. Results and Analysis

Comparison with State-of-the-art Methods: In this section, we compare our results with several state-of-the-art listed next: Discriminant Canonical Correlation analysis (DCC) [20], Manifold-to-Manifold Distance (MMD) [32], Manifold Discriminative Analysis (MDA) [30], Covariance Discriminative Learning (CDL) [31], the linear version of Affine Hull-based Image Set Distance (AHISD) [4], Convex Hull-based Image Set Distance (CHISD) [4] and Sparse Approximated Nearest Points (SANP) [16], Joint Regularized Nearest Points (JRNP) [36], Dictionary-based Face Recognition from Video (DFRV) [8], Joint Dictionary and Subspace Learning (JDSSL) [39]. All the competing methods were implemented using the code provided by the authors except for JDSSL and JRNP. The parameters were tuned based on the settings reported in their papers. We implement the JDSSL following the algorithm in [39] and cite the results directly reported in JRNP [36] as a fair comparison for the Honda/UCSD database[1].

Honda/UCSD: The average recognition rates using 50, 100 and full length of training frames on Honda/UCSD are reported in Table 1. It is seen that most state-of-the-art methods achieve 100% rank-1 accuracy using full length of frames for training. When the number of training frames is reduced to 50 and 100, the performance of other comparing methods degrades. However, when the frame length is 100, our method could still achieve 100% accuracy as reported in [36], which demonstrates that our dictionary is able to preserve the subspace structure even with a small number of training samples. In particular, our method consistently outperforms all other dictionary-based approaches [8, 39]. This is because the learned dictionary by the proposed method is not only reconstructive and discriminative, but also can encourage the discriminative coefficients to be of low rank. Overall, our method achieves the best performance under all three settings.

CMU Mobo: We repeated 10 trials by different randomly selected training and testing image sets. The average recognition rates of the proposed method along with other methods are reported in Table 2. As shown in Table 2, our method achieves very high performance of 98.2% and outperforms all other methods.

YouTube Celebrities: We used the cropped face samples of size 20 × 20 for consistency with Honda/UCSD database and reported results using 10-fold validation. These are the proposed settings used in [10, 30, 31]. We also compared with other state-of-the-art methods in [17] and [33]. Table 3 summarizes the average recognition rates of different methods.

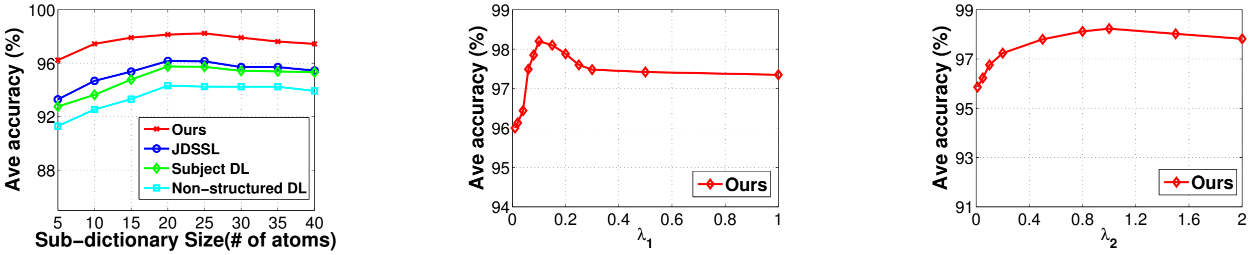
It is noted that the performance of all the methods on YTC degenerates significantly compared with Honda/UCSD and CMU Mobo. This is due to the large diversity and variations in appearance of each subject. Moreover, the high compressed rate, which results in low quality and resolution of the images, makes the recognition problem more difficult. It can be seen that our method outperforms the dictionary-based approach [39] by 2.7%, which demonstrates the effectiveness of the dictionary. In addition, our method achieves state-of-the-art performance compared to [33, 17][2].

Comparison with Different Dictionary Learning Approach: We further compare the proposed method with two different dictionary leaning strategies to further illustrate the effectiveness of our method.

1. Subject-specific Dictionary Learning (Subject DL): Instead of learning a global structured dictionary, we simply learn each sub-dictionary $D_i, i = \{1, ..., P\}$ independently by setting $\lambda_1 = 0$. Then we concate-

[1] Results of JRNP [36] on CMU Mobo [12] and YTC [19] databases have not been reported because the experimental settings we used are different from the ones in [36].

[2] Note that results from recent approaches [25, 36, 24, 15] have not been reported here since they employed different protocols from the settings in this paper.



(a) dictionary size

(b) parameter λ_1

(c) parameter λ_2

Figure 3. The effects of dictionary size n_0, parameters λ_1 and λ_2 on CMU Mobo database [12].

Methods	Honda/UCSD	CMU Mobo	YTC
Subject DL	98.4	95.8	69.5
Non-structured DL	95.9	94.3	67.7
Our method	**100.0**	**98.2**	**72.8**

Table 4. Average recognition rates of different dictionary learning approaches on Honda/UCSD, CMU Mobo and YTC databases. Rank-1 recognition accuracy are presented.

nate all the sub-dictionaries D_i together to construct D.

2. Non-structured Dictionary Learning (Non-structured DL): We only consider two terms of reconstruction errors using D and D_i and remove the nuclear term $\|Z_i\|_*$ in (1) without encouraging the representations to be low-rank. Then we perform recognition directly using (11).

Table 4 shows the average recognition rates of three different dictionary learning strategies. Our method consistently outperforms Subject DL and Non-structured DL on all three databases. Compared to Subject DL, the dictionary learned in our method is both discriminative and reconstructive. As it is designed to have small reconstruction errors for all the samples. Second, each sub-dictionary could well represent the corresponding subject while different sub-dictionaries would be exclusive to each other. In contrast, Subject DL only learns sub-dictionary for representing the corresponding subject. Moreover, Non-structured DL only focuses on reconstruction error of the samples. However, our method encourages face images from the same subject to have similar representation by enforcing them to lie in a low-dimensional subspace, which leads to independency across different subjects.

Parameter Sensitivity: In order to evaluate the effects of dictionary size n_0 and hyper-parameters λ_1, λ_2 on our method, we run different choices of parameters on the CMU Mobo database and plot the results in Figure 3.

Firstly, in Figure 3(a), we compare our method with JDSSL [39] and two different learning strategies (Subject DL and Non-structured DL) under the same number of sub-dictionary atoms for a fair comparison. It is seen that our approach outperforms [39] and the other two dictionary learning algorithms, by a large margin for all the different number of atoms. This is because we learn more discriminative and reconstructive dictionaries to preserve the structure of the samples from videos, while [39] only learned each sub-dictionary to encode the samples from the corresponding subject. We can also observe that increasing the size of sub-dictionary from 5 to 25 can result in improving the recognition performance. All the methods achieve the best performance when $n_0 = 25$. It is also interesting to note that when the size of sub-dictionary is 40, the performance degenerates slightly for all the methods. With a large sized dictionary, some redundant atoms in sub-dictionaries may be learned without being useful for recognition, thus affecting the partition-based decision to be made.

We also evaluate our approach with varying values of parameters λ_1 and λ_2 as shown in Figure 3(b)(c). It is observed that the performance is more sensitive to the choice of λ_1, which is associated with the reconstruction error when using dictionary D for reconstruction. This is because our method learns a discriminative and reconstructive global dictionary instead of concatenating the sub-dictionaries together, which are learned class by class.

5. Conclusion

In this paper, we presented a novel structured dictionary learning framework for video-based face recognition. We encouraged our sub-dictionaries to better represent the corresponding subject face images, while also preserving the subspace structure by enforcing the representation to be low-rank. This framework learned a dictionary with both discriminative and reconstructive properties for recognition purposes. Moreover, we proposed an efficient alternating optimization algorithm that converges reasonable faster. Finally, we extensively evaluated our approach on three benchmark databases for video-based face recognition. The experimental results clearly demonstrate the superior performance over the state-of-the-art.

6. Acknowledgement

This research is based upon work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA R&D Contract No. 2014-14071600012. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

References

[1] M. Aharon, M. Elad, and A. Bruckstein. K-SVD : An algorithm for designing of overcomplete dictionaries for sparse representation. *IEEE Transactions on Signal Processing*, 54(11):4311–4322, 2006. 1, 2, 4

[2] O. Arandjelovic, G. Shakhnarovich, J. Fisher, R. Cipolla, and T. Darrell. Face recognition with image sets using manifold density divergence. In *CVPR*, pages 581–588, 2005. 1

[3] E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *J. ACM*, 58(3):11, 2011. 3

[4] H. Cevikalp and B. Triggs. Face recognition based on image sets. In *CVPR*, pages 2567–2573, 2010. 1, 2, 5, 6

[5] L. Chen. Dual linear regression based classification for face cluster recognition. In *CVPR*, pages 2673–2680, 2014. 2

[6] S. Chen, C. Sanderson, M. T. Harandi, and B. C. Lovell. Improved image set classification via joint sparse approximated nearest subspaces. In *CVPR*, pages 452–459, 2013. 2

[7] S. Chen, A. Wiliem, C. Sanderson, and B. C. Lovell. Matching image sets via adaptive multi convex hull. In *WACV*, pages 1074–1081, 2014. 1, 2

[8] Y. Chen, V. M. Patel, P. J. Phillips, and R. Chellappa. Dictionary-based face recognition from video. In *ECCV*, pages 766–779, 2012. 1, 5, 6

[9] Y. Chen, V. M. Patel, S. Shekhar, R. Chellappa, and P. J. Phillips. Video-based face recognition via joint sparse representation. In *International Conference on Automatic Face and Gesture Recognition, FG*, pages 1–8, 2013. 1

[10] Z. Cui, S. Shan, H. Zhang, S. Lao, and X. Chen. Image sets alignment for video-based face recognition. In *CVPR*, pages 2626–2633, 2012. 1, 5, 6

[11] W. Deng, J. Hu, and J. Guo. Extended src: Undersampled face recognition via intraclass variant dictionary. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(9):1864–1870, 2012. 1, 2

[12] R. Gross and J. Shi. The cmu motion of body (mobo) database. Technical Report CMU-RI-TR-01-18, Robotics Institute, June 2001. 5, 6, 7

[13] H. Guo, Z. Jiang, and L. S. Davis. Discriminative dictionary learning with pairwise constraints. In *ACCV*, pages 328–342, 2012. 1, 2

[14] M. T. Harandi, C. Sanderson, S. A. Shirazi, and B. C. Lovell. Graph embedding discriminant analysis on grassmannian manifolds for improved image set matching. In *CVPR*, pages 2705–2712, 2011. 2

[15] M. Hayat, M. Bennamoun, and S. An. Learning non-linear reconstruction models for image set classification. In *CVPR*, pages 1915–1922, 2014. 1, 2, 6

[16] Y. Hu, A. S. Mian, and R. A. Owens. Sparse approximated nearest points for image set classification. In *CVPR*, pages 121–128, 2011. 1, 2, 5, 6

[17] Z. Huang, R. Wang, S. Shan, and X. Chen. Projection metric learning on grassmann manifold with application to video based face recognition. In *CVPR*, pages 140–149, 2015. 1, 2, 5, 6

[18] Z. Jiang, Z. Lin, and L. S. Davis. Learning a discriminative dictionary for sparse coding via label consistent K-SVD. In *CVPR*, pages 1697–1704, 2011. 1, 2

[19] M. Kim, S. Kumar, V. Pavlovic, and H. A. Rowley. Face tracking and recognition with visual constraints in real-world videos. In *CVPR*, 2008. 1, 2, 5, 6

[20] T. Kim, J. Kittler, and R. Cipolla. Learning discriminative canonical correlations for object recognition with image sets. In *ECCV*, pages 251–262, 2006. 1, 2, 5, 6

[21] S. Kong and D. Wang. A dictionary learning approach for classification: Separating the particularity and the commonality. In *ECCV*, volume 7572, pages 186–199, 2012. 1, 2

[22] K. Lee, J. Ho, M. Yang, and D. J. Kriegman. Video-based face recognition using probabilistic appearance manifolds. In *CVPR*, pages 313–320, 2003. 1, 2, 5, 6

[23] Z. Lin, R. Liu, and Z. Su. Linearized alternating direction method with adaptive penalty for low-rank representation. In *NIPS*, pages 612–620, 2011. 3, 4

[24] J. Lu, G. Wang, W. Deng, and P. Moulin. Simultaneous feature and dictionary learning for image set based face recognition. In *ECCV*, pages 265–280, 2014. 1, 5, 6

[25] J. Lu, G. Wang, W. Deng, P. Moulin, and J. Zhou. Multi-manifold deep metric learning for image set classification. In *CVPR*, pages 1137–1145, 2015. 1, 2, 5, 6

[26] J. Lu, G. Wang, and P. Moulin. Image set classification using holistic multiple order statistics features and localized multi-kernel metric learning. In *ICCV*, pages 329–336, 2013. 1, 2, 5

[27] L. Ma, C. Wang, B. Xiao, and W. Zhou. Sparse representation for face recognition based on discriminative low-rank dictionary learning. In *CVPR*, pages 2586–2593, 2012. 1, 2

[28] G. Shakhnarovich, J. W. F. III, and T. Darrell. Face recognition from long-term observations. In *ECCV*, pages 851–868, 2002. 1

[29] P. A. Viola and M. J. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, 2004. 5

[30] R. Wang and X. Chen. Manifold discriminant analysis. In *CVPR*, pages 429–436, 2009. 1, 2, 5, 6

[31] R. Wang, H. Guo, L. S. Davis, and Q. Dai. Covariance discriminative learning: A natural and efficient approach to image set classification. In *CVPR*, pages 2496–2503, 2012. 1, 2, 5, 6

[32] R. Wang, S. Shan, X. Chen, and W. Gao. Manifold-manifold distance with application to face recognition based on image set. In *CVPR*, 2008. 1, 2, 6

[33] W. Wang, R. Wang, Z. Huang, S. Shan, and X. Chen. Discriminant analysis on Riemannian manifold of Gaussian distributions for face recognition with image sets. In *CVPR*, pages 2048–2057, 2015. 1, 2, 5, 6

[34] L. Wolf, T. Hassner, and I. Maoz. Face recognition in unconstrained videos with matched background similarity. In *CVPR*, pages 529–534, 2011. 1

[35] H. Xu, J. Zheng, and R. Chellappa. Bridging the domain shift by domain adaptive dictionary learning. In *Proceedings of the British Machine Vision Conference, BMVC*, 2015. 1

[36] M. Yang, W. Liu, and L. Shen. Joint regularized nearest points for image set based face recognition. In *International Conference Automatic Face and Gesture Recognition, FG*, pages 1–7, 2015. 2, 6

[37] M. Yang, L. Zhang, X. Feng, and D. Zhang. Fisher discrimination dictionary learning for sparse representation. In *ICCV*, pages 543–550, 2011. 1, 2

[38] M. Yang, P. Zhu, L. J. V. Gool, and L. Zhang. Face recognition based on regularized nearest points between image sets. In *International Conference on Automatic Face and Gesture Recognition, FG*, pages 1–7, 2013. 1, 5

[39] G. Zhang, R. He, and L. S. Davis. Jointly learning dictionaries and subspace structure for video-based face recognition. In *ACCV*, pages 97–111, 2014. 1, 6, 7

[40] Q. Zhang and B. Li. Discriminative K-SVD for dictionary learning in face recognition. In *CVPR*, pages 2691–2698, 2010. 1, 2

[41] Y. Zhang, Z. Jiang, and L. S. Davis. Learning structured low-rank representations for image classification. In *CVPR*, pages 676–683, 2013. 1, 2, 3

[42] J. Zheng and Z. Jiang. Learning view-invariant sparse representations for cross-view action recognition. In *ICCV*, pages 3176–3183, 2013. 1, 2