Jointly Learning Dictionaries and Subspace Structure for Video-based Face Recognition

Guangxiao Zhang[†], Ran He[§], Larry S. Davis[†]

[†]Institute for Advanced Computer Studies, University of Maryland, College Park, MD, USA 20742
[§] Institute of Automation, Chinese Academy of Sciences,
95 Zhongguancun East Road, P.O.Box 2728, Beijing, China 100190
{gxzhang, lsd}@umiacs.umd.edu, rhe@nlpr.ia.ac.cn

Abstract. In video-sharing websites and surveillance scenarios, there are often a large amount of face videos. This paper proposes a joint dictionary learning and subspace segmentation method for video-based face recognition (VFR). We assume that the face images from one subject video lie in a union of multiple linear subspaces, and there exists a global dictionary to represent these images and segment them to their corresponding subspaces. This assumption results in a "chicken and egg" problem, where subspace clustering and dictionary learning are mutually dependent. To solve this problem, we propose a joint optimization model that includes three parts. The first part seeks a low-rank representation for subspace segmentation; the second part encourages the dictionary to accurately represent the data while tolerating frame-wise corruption or outliers; and the third part is a regularization on the dictionary. An alternating minimization method is employed as an efficient solution to the proposed joint formulation. In each iteration, it alternately learns the subspace structure and the dictionary by improving the learning results. Experiments on three video-based face databases show that our approach consistently outperforms the state-of-the-art methods.

1 Introduction

Video-based face recognition (VFR) has become an active research topic in recent years in the computer vision community [1–7]. Compared to still-image face recognition, the task in video-to-video recognition is to efficiently exploit multiple frames, and to build a model robust against variations of the same subject appearing in different videos. This is challenging because faces detected from videos are usually acquired under non-ideal acquisition conditions in which illumination, pose, and facial expression variations dominate. Moreover, the cropped face images are often of low resolution, which makes many local feature methods inapplicable.

To solve the video-based face recognition problem, researchers have proposed numerous methods. Early frame-based attempts include fusing frame-based recognition results by voting [8], finding the minimal distance between two frames across videos [9], and matching the key frames with exemplars in the gallery [10]. Most of the recent approaches are based on either temporal models or image sets. Some researchers extract spatial-temporal representations from videos to enhance face recognition [11, 12]. Others discard the temporal information and treat the videos as image sets [1–3, 13–16]. The problem of VFR then becomes a more general image-set matching or classification problem. To solve this problem, many statistical models were proposed to describe the image sets as linear subspaces or manifolds. Under the linear subspace assumptions, methods such as [13], [14] (DCC) measure the distance or similarity between two subspaces by computing the angles between the principle components. Hu et al. [1] (SANP) find the minimal distance of the two nearest points, which can be sparsely approximated from the samples of their respective subspaces. Under the nonlinear manifold assumptions, [17] defines the distance between subspaces over Grassmann manifold, and then constructs the Sparse Approximated Nearest Subspaces (SANS) adaptively from the samples of the query image set. It approaches the nearest point to the reference point by minimizing the joint sparse representation error. Wang et al. [15] proposed Manifold-to-Manifold Distance learning (MMD), which partitions a manifold into several local linear models and integrates the pair-wise distances. They also extended MMD to a supervised version called Manifold Discriminant Analysis(MDA) [16]. Moreover, Wang et al. [2] represent image sets with their covariance matrices, and compute the distance between manifolds by mapping the covariance matrix from the Riemannian manifold to a Euclidean space (Cov+PLS). While those methods have received great success, Cui et al. [3] raised an uncertainty issue that commonly arises when partitioning a nonlinear manifold into local linear subspaces. They argue that face images with similar appearance can be clustered to different subspaces or clusters in different video sequences, making the distance between two manifolds ill-defined. They proposed to align the gallery set and the query set with respect to a pre-defined reference sequence [3] (ImgSetAlign). This image set alignment issue is also addressed in [18]. Given a well-aligned, high quality gallery, they combine three tasks in a unified framework: aligning faces geometrically, performing recognition, and selecting good quality frames (CAR). Another image set method proposed by Lu et. al [19] computes the holistic multiple order statistics features of the image sets, and performs multi-kernel metric learning. These methods have achieved the state of the art on several public face databases.

Most recently, sparse representation for videos has attracted attention. Although the advantage of dictionary learning for robust face recognition in still images has been widely recognized [20], dictionary learning for VFR is relatively new. Chen et al. [4,5] proposed a joint sparse representation method under a minimal reconstruction error criterion. It divides a video sequence into K partitions in order to capture different poses or illumination conditions. Next, partition-level sub-dictionaries are learned by minimizing the total reconstruction error within the partition. Experiments have demonstrated that the dictionary-based method also achieved the state of the art.

The dictionary-based method falls into the category of image-set methods. The basic assumption can be summarized as follows: all face images from one subject video lie in a union of multiple linear subspaces. In each of those sub-

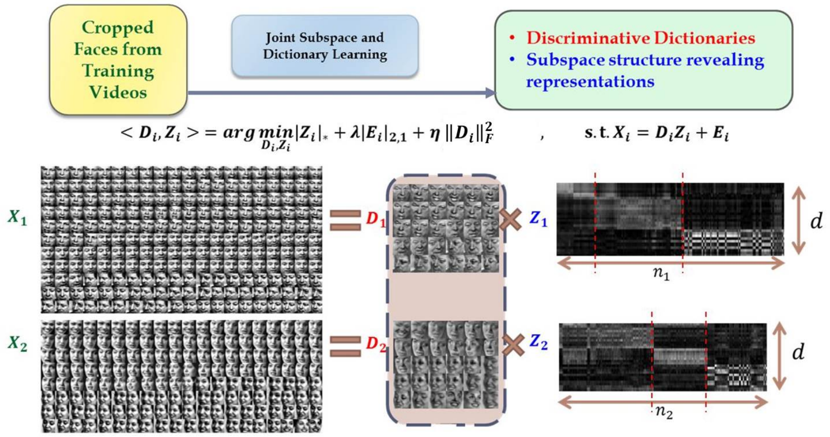

Cropped Faces from Training Videos
Joint Subspace and Dictionary Learning
• Discriminative Dictionaries
• Subspace structure revealing representations
$< D_i, Z_i > = \arg\min_{D_i, Z_i} |Z_i|_* + \lambda |E_i|_{2,1} + \eta \|D_i\|_F^2$, s.t. $X_i = D_i Z_i + E_i$
X_1
D_1
Z_1
d
n_1
X_2
D_2
Z_2
d
n_2

Fig. 1. An overview of our method. Sequences of cropped faces detected from videos are sent as inputs to our Jointly Learning Dictionary and Subspace Structure (JLDSS) algorithm. It learns class-specific dictionaries and the corresponding low-rank representations simultaneously. Examples of face sequences X_1, X_2, our dictionaries D_1, D_2, and the low-rank matrices Z_1, Z_2 are shown in the figure. The low-rank matrices can also be employed for video segmentation. Such examples are shown by the red dashed lines on Z_1 and Z_2.

spaces, there exists a sub-dictionary that can represent the data well. However, as in other manifold-partitioning-to-multiple-linear-subspaces methods, the clustering uncertainty issue recognized in [3] also exists here. Subspace clustering and dictionary learning are mutually beneficial and dependent on one another. To learn the sub-dictionaries, or perform any within subspace operations in general, subspace clustering, or equivalently "sequence partitioning" in [4] and [5], needs to be performed first. Yet it is impossible to define an "optimal" clustering result until the final reconstruction error is obtained. In other words, one needs to have the dictionaries in advance for reconstruction to make an appropriate choice of subspace clustering that captures the true characteristics of the video.

To address the above issue, we propose a joint learning framework that simultaneously learns a global dictionary and reveals the intrinsic subspace structure of the video. We assume that faces from one subject video lie in multiple linear subspaces, and there exists a global dictionary that can represent all the faces and reflects the subspace structure of the image set. The objective function of our model includes three parts. The first part forces the data to lie in multiple linear subspaces, in each of which one point can be represented by a set of bases called a dictionary that spans the same linear subspace; the second part encourages the dictionary to represent the data well with tolerance to outliers; the third part regularizes the learned dictionary. An alternating minimization method is employed as an efficient solution to the joint formulation. In each iteration, it alternately learns the subspace structure and the dictionary by improving the learning results.

The main contributions of our work are summarized as follows:

- We present a joint subspace and dictionary learning framework for VFR. Unlike the partition-then-learn framework, our approach implicitly learns the subspace segmentation along with a global dictionary simultaneously. The video-dictionaries are compact and compliant to the subspace structure of the data, meaning the more dynamic videos with larger variation are automatically assigned more dictionary atoms than the more static videos.
- Our model is robust to variation and frame-wise corruption. Since we model the same face under various poses and illumination as data points lying in multiple subspaces, it allows our model to handle data with large variation. Moreover, by minimizing the $l_{2,1}$ norm of the reconstruction error matrix, we essentially fuse the frame-wise results together while tolerating corruption and outliers, making our model robust.
- Experiments shows that our method not only achieves the best recognition performances on three standard databases, but also yields interpretable low-rank representations and more natural dictionaries.

1.1 Related Work

There are a couple of recent works focusing on subspace recovery [21–24]. One of the related models to our work is [21]. Liu et al. proposed a low-rank minimization framework to recover the subspace structures in the presence of noise, outliers, and corruption. The main interest of those works, however, is to analyze the subspace structures of a given set of observation points without considering how it generalizes to unseen data. For that reason, [21] first constructs a dictionary, and keeps it unchanged throughout the process. On the contrary, our model is designed for classification purpose and therefore the generative power is important. We *learn* a set of dictionaries that best represent the observation points. The recovery of the subspace structures in our model facilitates the dictionary learning and enhances the representation.

2 Preliminaries

2.1 Subspace learning via low-rank minimization

Suppose we have a set of corrupted data points (in columns), $X = [x_1, x_2, ..., x_n]$, drawn from a union of multiple subspaces $\mathcal{S}_1, ..., \mathcal{S}_k$. We wish to decompose the data matrix as the sum of a clean, self-expressive, and low-rank matrix plus a matrix of noise or outliers. This can be achieved by solving:

$$Z^* = \arg\min_Z \operatorname{rank}(Z) + \gamma \|E\|_l, \quad \text{s.t. } X = DZ + E \tag{1}$$

where D is a pre-defined dictionary that linearly spans the entire data space, and Z is the representation with respect to D. The optimal solution is then used for estimating the lowest-rank recovery of the corrupted data DZ^{*1}. With replacement of the rank function with the nuclear norm, the problem becomes a convex optimization and can be solved by the Augmented Lagrange Multiplier (ALM) algorithm, also known as an alternating direction method:

$$Z^* = \arg\min_{Z} \|Z\|_* + \gamma \|E\|_l, \quad \text{s.t. } X = DZ + E \tag{2}$$

Depending on the error types in different applications, one can choose:

- $l = 0$ to model element-wise sparse error. As minimizing the l_0 norm is NP hard, the l_1 norm is often employed as a good relaxation, which is defined by $\|E\|_1 := \sum_{i,j} |[E]_{i,j}|$.
- $l = 2$ to model Gaussian noise (white noise). $\|E\|_2 := \sqrt{\sum_{i,j} |E_{i,j}|^2}$.
- $l = 2,1$ to model sample-wise sparse error. This is suitable when outliers and corruption exist. $\|E\|_{2,1} := \sum_i \|[E]_{:,i}\|_2$.

In most of the literature, the dictionary D in (2) is pre-defined. In particular, by setting $D = X$, one essentially assumes that any data point (column of X) can be represented by a linear combination (with the coefficients given by columns of Z) of all the other points in the same subspace. Columns of Z thereby are considered as new representations of the original points [21].

Low-rank minimization and subspace structure recovery have been successfully used in applications such as data clustering, image denoising, saliency detection, and recognition and classification. In particular, for recognition and classification where the dataset contains multiple subjects, samples of one subject are considered to be drawn from the same linear subspace, while samples of different subjects are drawn from different linear subspaces. However, for video-based face recognition, it is beneficial to consider a nonlinear subspace or multiple linear subspaces for one subject because of large appearance variations [4,5,14,15].

2.2 Dictionary learning for sparse representation

Suppose we have the original training data $X = [X_1, X_2, ..., X_c]$, where X_i is the data from the i^{th} class. We wish to learn a set of bases D_i, called "dictionaries", such that the projection of X_i to the bases is "sparse", i.e.

$$\min_{D_i, Z_i} \|X_i - D_i Z_i\|_F^2, \quad \text{s.t. } \|z_j\| \leq T_0, \forall j = 1, ..., n_i \tag{3}$$

where $Z_i = [z_1, z_2, ..., z_{n_i}]$ is the sparse representation of the original data X_i with respect to D_i. T_0 is the sparsity constant which specifies the maximum number of nonzero elements.

The standard solution to (3) alternates between sparse coding and dictionary learning. There are many off-the-shelf algorithms to find the sparse codes for a given dictionary, such as Orthogonal Matching Pursuit (OMP), coordinate

[1] Here and for the rest of the paper, a variable with a superscript * denotes the optimal solution. One should not confuse the notation with the symbol of Hermitian transpose.

descent, first-order/proximal methods, etc. Conversely, given the sparse representation, one can derive the optimal dictionary by finding the least-square-based closed form solution, or adopt the popular K-SVD algorithm [25] for its computational efficiency.

3 Joint discriminative dictionary learning and subspace structure recovery for videos

3.1 Problem formulation

Suppose we have a set of cropped faces from the training videos (gallery) for N people. Denote the face sequence of person i as $X_i = [x_1^{[i]}, ..., x_{n_i}^{[i]}]$, where $x_j^{[i]}$ is a column feature vector that describes the j^{th} face in the sequence for person i. Due to facial expression, pose, and illumination changes, we assume that $x_j^{[i]}, j = 1, ..., n_i$ are noisy data points drawn from a union of an unknown number of subspaces. The objective is to learn a dictionary D_i that: (1) is good for reconstruction; (2) yields a new representation Z_i, which has low rank and reveals the multiple subspace structure of the "clean" data. This can be formulated as the following optimization problem:

$$\begin{aligned} < D_i, Z_i >= \arg \min_{D_i, Z_i} \|Z_i\|_* + \lambda \|E_i\|_{2,1} + \eta \|D_i\|_F^2, \\ \text{s.t. } X_i = D_i Z_i + E_i, \quad \text{for all } i. \end{aligned} \tag{4}$$

The first term is the low-rank requirement. The second one, which is the $l_{2,1}$ norm of the reconstruction error, encourages accurate reconstruction while tolerating sample-specific corruption such as occlusion and outliers. The choice of the trade-off parameter λ depends on the nature of the data. For example, if the person's face in a video appears to be fairly still (with small changes in pose and expression) and switches to another still pose very quickly, then that means that the data points are lying in the subspaces with few outliers, therefore the low-rank constraint should be relaxed and the dictionary should aim to achieve better reconstruction. Conversely, if the person's facial expression or pose changes gradually over time with no obvious cutoff, then the low-rank constraint should be emphasized more so it allows the dictionary to capture key features. The third term is the regularization.

3.2 Optimization

For the rest of this section, the class index i is dropped for convenience. Following the standard procedures of the Augmented Lagrangian Multiplier method, we introduce auxiliary variables J, Y_1, Y_2 and μ. The optimization problem above becomes:

$$\min_{D,Z,E,J} \|J\|_* + \lambda \|E\|_{2,1} + \eta \|D\|_F^2, \text{s.t.} X = DZ + E, Z = J \tag{5}$$

Algorithm 1 Video-based face recognition by JLDSS

```
Input: $X_1, X_2, ..., X_C$, $Y$, $\lambda$, and $\eta$
Output: Recognition $p$
Initialization:
  Initialize $D_i$ by finding the first $d$ principle components on columns of $X_i$.
Training:
for i=1:C do
  Learn $D_i$ and $Z_i$ by Algorithm 2 (JLDSS), given $X_i$, $D_i$, $\lambda$, and $\eta$.
end for
  $D = [D_1|D_2|...|D_C]$
Testing:
  Find $Z_y$ in (12) by Algorithm 2 without updating $D$, given $Y$, $D$, and $\lambda$.
  Recognize $p$ given by equation (13).
```

with the Lagrangian function given by

$$\mathcal{L}(D, Z, J, E, Y_1, Y_2, \mu) = \|J\|_* + \lambda\|E\|_{2,1} + \eta\|D\|_F^2 + \langle Y_1, X - DZ - E\rangle + \langle Y_2, Z - J\rangle + \frac{\mu}{2}\left(\|X - DZ - E\|_F^2 + \|Z - J\|_F^2\right) \tag{6}$$

where $\langle A, B\rangle = trace(A^T B)$. This problem can be optimized in an alternating way described as follows. In each iteration, it first solves for Z with D fixed, and then solves for D with Z fixed. Repeat until the convergence is achieved.

Solve for Z With D fixed, (5) becomes a typical low-rank minimization problem with auxiliary variable J:

$$\min_{Z,J,E} \|J\|_* + \lambda\|E\|_{2,1}, \text{s.t.} X = DZ + E, Z = J \tag{7}$$

with solutions given by,

$$J^* = \arg\min_J \frac{1}{\mu}\|J\|_* + \frac{1}{2}\|J - (Z + Y_2/\mu)\|_F^2 \tag{8}$$

$$Z^* = (I + D^T D)^{-1}\left(D^T(X - E) + J + (D^T Y_1 - Y_2)/\mu\right) \tag{9}$$

$$E^* = \arg\min \frac{\lambda}{\mu}\|E\|_{2,1} + \frac{1}{2}\|E - (X - DZ + Y_1/\mu)\|_F^2 \tag{10}$$

Details of derivation are provided in the supplemented material.

Solve for D Once we have updated Z, J, and E, the Lagrangian function (6) becomes a quadratic function of D. Finding the solution to $\nabla_D \mathcal{L}(D; Z, J, E, Y_1, Y_2, \mu) = 0$ is equivalent to:

$$\min_D \left\{\eta\|D_i\|_F^2 + \langle Y_1, X - DZ - E\rangle + \frac{\mu}{2}\|X - DZ - E\|_F^2\right\} \tag{11}$$

which has a closed form solution: $(D^*)^T = \left(\frac{2\eta}{\mu} I + ZZ^T\right)^{-1} Z\left((X - E) + \frac{Y_1}{\mu}\right)^T$. See the supplemented material for derivation.

The advantage of our method is that for each class, it seeks a solution for D and Z simultaneously without conducting explicit subspace clustering. Yet if one wants to, one can easily find the subspace structure by performing spectral clustering on Z [21]. The typical matrices are displayed in Figure 2(d) and 2(e). It clearly shows that the video contains 3 distinct poses or illumination conditions.

We describe the complete algorithm in Algorithm 1.

Algorithm 2 JLDSS: Jointly Learning Dictionary and Subspace Structure

```
Input: $X$, $D_0$, $\lambda$, and $\eta$
Output: $D$ and $Z$
Initialization:
  $Z = J = 0, D = D_0, E = 0, Y_1 = 0, Y_2 = 0, \mu = 10^{-6}, \mu_{max} = 10^6, \rho = 1.1$, and $tol = 10^{-6}$
while not converge do
  Update $J \leftarrow J^*$, where
      $$J^* = \arg\min_J \frac{1}{\mu}\|J\|_* + \frac{1}{2}\|J - (Z + Y_2/\mu)\|_F^2$$
  Update $Z \leftarrow Z^*$, where
      $$Z^* = (I + D^T D)^{-1}\left(D^T(X - E) + J + (D^T Y_1 - Y_2)/\mu\right)$$
  Update $E \leftarrow E^*$, where
      $$E^* = \arg\min \frac{\lambda}{\mu}\|E\|_{2,1} + \frac{1}{2}\|E - (X - DZ + Y_1/\mu)\|_F^2$$
  (For recognition, skip this step) Update the dictionary $D \leftarrow D^*$, where
      $$(D^*)^T = \left(\frac{2\eta}{\mu}I + ZZ^T\right)^{-1} Z\left((X - E) + \frac{Y_1}{\mu}\right)^T$$
  Update the parameter $\mu \leftarrow \min(\rho\mu, \mu_{max})$
  Update the multipliers
      $$Y_1 \leftarrow Y_1 + \mu(X - DZ - E), \quad Y_2 \leftarrow Y_2 + \mu(Z - J)$$
  Check the convergence conditions: $\|X - DZ - E\|_\infty < tol$ and $\|Z - J\|_\infty < tol$.
end while
```

3.3 Recognition

Once we obtain the class-specific dictionaries $D_1, D_2, ..., D_C$, the global dictionary is the concatenation, i.e. $D = [D_1|D_2|...|D_C]$. Denote the test sequence (query) of a face as Y. We assume all the faces belong to a single subject to be recognized. The low-rank representation is given by:

$$Z_y = \arg\min_{Z_y} \|Z\|_* + \lambda \|E_y\|_{2,1}, \quad \text{s.t. } Y = DZ_y + E_y \tag{12}$$

Suppose we have d dictionary atoms for each class-specific dictionary. The first d rows of Z_y correspond to the dictionary of the 1st class; the second d rows, or the $(d+1)$-th to the $2d$-th rows, of Z_y correspond to the dictionary of the 2nd class; and so on. Denote the k-th d rows of Z_y as $Z_{y,k}$. Choose the subject $p*$ with the best reconstruction given by D_k and $Z_{y,k}$ as our recognition decision:

$$p* = \arg\min_{k \in 1,\dots,C} \|Y - D_k Z_{y,k}\|_{2,1} \tag{13}$$

4 Experiments

We evaluated the proposed method on three data sets for video-based face recognition: Honda/UCSD video database [12], the CMU Motion of Body (MoBo) database [26], and the more challenging YouTube Celebrities Face Tracking and Recognition dataset [11].

4.1 Comparison methods

The methods we compare ours against include:

- A linear subspace method: discriminative canonical correlations (DCC) [14];
- A nonlinear manifold method: manifold discriminant analysis (MDA) [15];
- An affine subspace method: sparse approximated nearest point (SANP) [1];
- A covariance-on-manifold method: covariance discriminative learning (Cov+PLS) [2];
- A manifold alignment method: image sets alignment (ImgSetsAlign) [3];
- A dictionary based method: sparse representation for video (SRV) and its kernelized version KSRV [5]. The higher recognition rates between the two versions are adopted for comparison, which we denote as (K)SRV.

We compare our method especially to the dictionary-based method (K)SRV to show the effect of learning subspace structure with the dictionary without performing video partitioning. The recognition rates for other competing methods are cited directly from their papers, except for (K)SRV in Honda/UCSD, because [5] had a slightly different setting.

4.2 Experimental Set-up

For all experiments, we extracted face sequences by a cascaded face detector [27], and resized them to 20*20 gray images (30*30 for YouTube Celebrities database). The feature vectors are simply the vectorized faces with histogram equalization for reducing lighting effects. We also doubled the size of the gallery by adding the mirror-symmetric faces. This avoids the tendency of assigning unknown profile faces to the one with similar poses in the gallery.

Honda/UCSD Database There are 59 videos for 20 people in a wide range of different poses in Honda/UCSD database. Each person has at least 2 videos. We randomly selected 1 video as training and tested on the rest, and repeated for 10 times. To follow the procedures in [1], we tested four cases of maximum set length: 50, 100, and full length. Note that for 50/100 maximum length, we tested on the first 50/100 frame as an standard setting, as well as on randomly selected 50/100 frames as in [5] for fair comparison. With the randomly chosen frames, we achieved 100% accuracy for both 50 frames case and 100 frames case. The average recognition rates over 10 trials under the standard setting are reported in Table 1. Our rates are obtained by setting the dictionary size $d = 10$. Performance is not sensitive to the choices of λ and η. We outperform all other methods in all settings.

CMU MoBo Database The CMU MoBo contains 96 sequences of 24 subjects, each of which has 4 sequences (roughly 300 frames each) captured in different walking situations. We performed 10-fold cross validation where 1 video was randomly chosen as training and the remaining 3 for testing. The average recognition rate is shown in Table 1. For our method, we set $d = 20, \lambda = 0.1, \text{and } \eta = 0.01$. For (K)SRV, we set the number of partitions $K = 3$. The dictionary size for each subject is $d = 7 * 3$ (7 for each partition), which is comparable to 20 in our method. Again we achieved the best performance among all.

YouTube Celebrities The YouTube Celebrities contains 1910 video clips of 47 human subjects from YouTube. Roughly 41 clips were segmented from 3 unique videos for each person. This dataset is challenging because it contains a lot of noise and facial variations (see Figure 2(a)). Following the standard setup, we selected 3 training clips, 1 from each unique video, and 6 test clips, 2 from each unique video, per person. The performance of all methods is summarized in Table 1. Our rates are obtained by setting $d = 30, \lambda = 0.1, \text{and } \eta = 0.001$.

To the best of our knowledge, the top performance levels on this dataset are reported as 80.75% in [7], 78.9% in [6], and 74.6% in [3]. However, their experiments employed different protocols from the standard one. [7] not only benefits from its own tracker, which gives 92% success rate versus 80% using the standard tracker, but also takes advantage of sophisticated features including LBP, HOG and Gabor wavelets. Other methods only use 30*30 vectorized faces as features. [6] only tested on the videos of the first 29 celebrities out of 47, which makes the task easier than the standard one. Recognition rates are of course higher when a smaller number of subjects are included. In addition to the 3 training clips for each subject in the gallery, [3] uses one more sequence from any subject as a reference for alignment. This gives their method an advantage of seeing more faces in the gallery. We also noticed that [3] reported higher recognition rates than the literature for the competing methods that we also used for comparison: DCC: 0.673 in [3] vs 0.648 in [2], and MDA: 0.676 [3] vs 0.653 [16], suggesting a systematic bias might exist. Under the standard settings, our

Table 1. Recognition rates on three databases. We cited the recognition rates of the competing methods from the literature except for (K)SRV. The highest rate in each experiment is highlighted in bold font. In the last row, the number with the superscript * was achieved by employing a different protocol than the standard one.

Dataset		DCC [14]	MDA [16]	SANP [1]	Cov+ PLS [2]	ImgSets Align [3]	(K)SRV [5]	Our method
Honda/ UCSD [12]	50 frames	0.769	0.744	0.846	-	-	0.846±0.02	**0.872**±0.01
	100 frames	0.846	0.949	0.923	-	-	0.964±0.02	**0.974**±0.01
	full length	0.949	0.974	**1.000**	**1.000**	0.989	0.974±0.01	**1.000**
	Average	0.856	0.889	0.923	-	-	0.931±0.01	**0.949**±0.01
CMU MoBo [26]		0.903	0.947	0.900	0.941	0.950	0.952±0.03	**0.968**±0.02
YouTube [11]		0.648	0.653	0.684	0.701	**0.746***	0.684±0.03	0.723±0.03

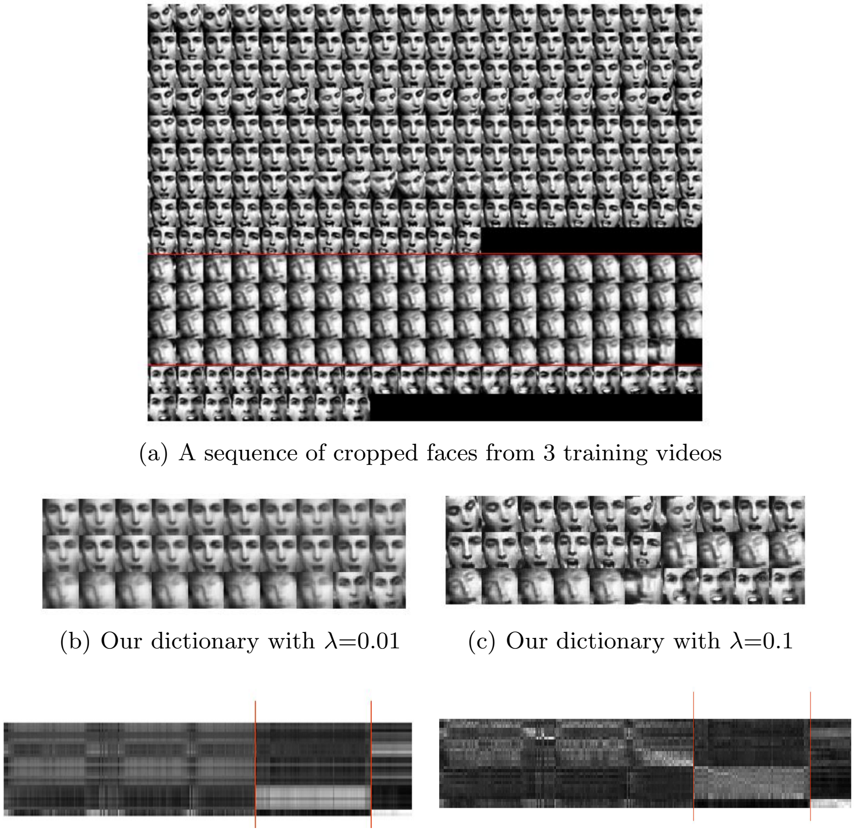
performance is the best, and it could be further improved with better tracking and advanced features.

4.3 Analysis and Discussions

As seen in Table 1, we consistently outperformed other competing methods in all datasets under all settings, especially compared with the other dictionary-based method [5]. We take the video clips for one subject from YouTube Celebrities as an example to further demonstrate the benefit of jointly learning a dictionary and subspace structure.

The Effect of λ The comparison between Figure 2(b) and 2(c) shows the effect of the choice of λ, which is the trade-off between low-rank and reconstruction accuracy. When λ is small, we assume the data is more uniform, thus the dictionary atoms from the same subspace look very similar to each other in Figure2(b).The faces which look different from the dictionary atoms are considered as outliers. When λ is large, we assume the data contains large variations and therefore put more emphasis on the reconstruction accuracy. As a result, the dictionary contains faces with more variety as shown in Figure 2(c). The corresponding low-rank representations also reflect the impact of choosing different λ, where the columns of the matrix in Figure 2(d) look quite similar to each other while the columns in Figure 2(e) are much more diverse. However, the subspace segmentation results are the same.

Low-rank Matrix Interpretation Figure 2(d), 2(e) show a typical low-rank representation of original faces from training videos constructed by our method, which has a block diagonal structure indicating the subspace structure of the data. The brightness indicates the value of each entry, where the darkest entries are zeros. Looking at the columns, one can easily construct a similarity matrix and apply spectral clustering to it if explicit segmentation is desired. Furthermore, since each row of the matrix corresponds to the coefficients of a particular



(a) A sequence of cropped faces from 3 training videos

(b) Our dictionary with λ=0.01

(c) Our dictionary with λ=0.1

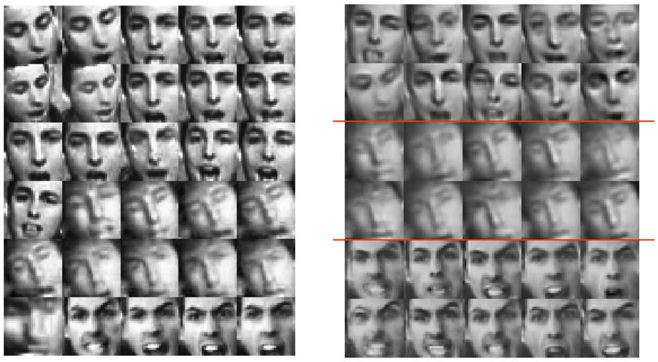
(d) The low-rank matrix with λ=0.01

(e) The low-rank matrix with λ=0.1

Fig. 2. An example of the training faces from YouTube Celebrities database (a), the dictionaries (b) (c), and the low-rank representations (d) (e).

dictionary atom, the row structure also reflects the structure of the dictionary. A skinny and tall block in the low-rank matrix suggests a relatively short clip with large variation, so that it requires many dictionary atoms to represent it, whereas a fat and short block indicates a long clip with little variation so that only a small number of dictionary atoms are needed.

Dictionary Comparison Figure 2(a) shows the sequence of cropped faces from our training clips, where the red line shows the true partition of the sequence that is used in our implementation of [5]. The dictionaries learned by both methods with $d = 30$ are shown in Figure 3. It clearly shows the limitations of the partition-level dictionaries. First, [5] assigns the same number of dictionary atoms regardless of the length and variation present in each clip. The first clip obviously contains much larger variation than the second clip. As a result, the first 10 dictionary faces learned by SRV [5] are blurry, indicating the dictionary



(a) Our dictionary with λ=0.1 (b) Dictionary learned by SRV

Fig. 3. Dictionary comparison of our method and SRV [5] using the true video partitions. The red dashed lines separate the partition-level sub-dictionaries.

is not big enough to capture the variation, while the second 10 dictionary faces are more or less uniform, indicating 10 atoms are more than necessary. Increasing the size of dictionary or the number of partitions might help, but with bigger dictionaries the computational cost will increase dramatically, especially in the testing stage when partition-level decisions need to be made. In addition, SRV suffers from the unknown length of each partition. In such situations where the shortest partition contains fewer frames than the size of sub-dictionary, artificial frames need to be inserted to obtain an augmented partition. In contrast, our method enjoys the flexibility of no explicit partitioning, so that the dictionary reflects the distribution of the training data.

5 Conclusion and Future Work

We introduced a novel joint learning framework for video-based face recognition. We modeled the set of faces as a union of multiple subspaces, and attempted to find a global dictionary that reveals the subspace structure. To achieve this goal, we proposed an objective function that encourages low-rank representation and reduces reconstruction error. We explained how our optimization problem can be solved with an alternating minimization algorithm. Finally, we conducted experiments on three data sets which resulted in the state-of-the-art performance. Future work to achieve better VFR includes running the face tracker and identifying the faces online, incorporate alignment with recognition, and developing a more effective down-sampling method that resizes the tracked face images to smaller size but preserves discriminative information.

Acknowledgement. This work was supported by the Army Research Office MURI Grant W911NF-09-1-0383. We also thank Dr. Ruiping Wang for sharing the processed data.

References

1. Hu, Y., Mian, A., Owens, R.: Sparse approximated nearest points for image classification (2011) Proceedings of EEE Conference on Computer Vision and Pattern Recognition.
2. Wang, R., Guo, H., Davis, L., Dai, Q.: Covariance discriminative learning: A natural and efficient approach to image set classification (2012) Proceedings of IEEE Conference on Computer Vision and Pattern Recognition.
3. Cui, Z., Zhang, H., Lao, S., Chen, X.: Image sets alignment for video-based face recognition (2012) Proceedings of IEEE Conference on Computer Vision and Pattern Recognition.
4. Chen, Y.C., Patel, V., Phillips, P., Chellappa, R.: Dictionary-based face recognition from video (2012) Proceedings of European Conference of Computer Vision.
5. Chen, Y.C., Patel, V., Shekhar, S., Chellappa, R., Phillips, P.: Video-based face recognition via joint sparse representation (2013) Proceedings of IEEE Conference on Automatic Face and Gesture Recognition.
6. Yang, M., Zhu, P., Zhang, L.: Face recognition based on regularized points between image sets (2013) Proceedings of IEEE Conference on Automatic Face and Gesture Recognition.
7. Ortiz, E., Wright, A., Shah, M.: Face recognition in movie trailers via mean sequence spars representation-based classification (2013) Proceedings of IEEE Conference on Computer Vision and Pattern Recognition.
8. Shakhnarovich, G., Fisher, J., Darrell, T.: Face recognition from long-term observations (2002) Proceedings of European Conference on Computer Vision.
9. Satoh, S.: Conparative evaluation on face sequence matching for content-based video access (2000) Proceedings of IEEE Automatic Face and Gesture Recognition.
10. Kreger, V., Zhou, S.: Exemplar-based face recognition from video (2002) Proceedings of European Conference on Computer Vision.
11. Kim, M., Kumar, S., Pavlovic, V., Rowley, H.: Face tracking and recognition with visual constraints in real-world videos (2008) Proceedings of IEEE Conference on Computer Vision and Pattern Recognition.
12. Lee, K., Ho, J., Yang, M., Kriegman, D.: Visual tracking and recognition using probabilistic appearance manifolds (2005) Proceedings of Computer Vision and Image Understanding.
13. Cevikalp, H., Triggs, B.: Face recognition based on image sets (2010) Proceedings of IEEE Conference on Computer Vision and Pattern Recognition.
14. Kim, T., Arandjelovic, O., Cipolla, R.: Discriminative learning and recognition of image set classes using canonical correlations. IEEE Transactions on Pattern Analysis and Machine Intelligence **29** (2007) 1005–1018
15. Wang, R., Shan, S., Chen, X., Gao, W.: Manifold-manifold distance with application to face recognition based on image set (2008) Proceedings of IEEE Conference on Computer Vision and Pattern Recognition.
16. Wang, R., Chen, X.: Manifold discrimininant analysis (2009) Proceedings of IEEE Conference on Computer Vision and Pattern Recognition.
17. Chen, S., Sanderson, C., Harandi, M.T., Lovell, B.: Improved image set classification via joint sparse approximated nearest subspaces (2013) Proceedings of IEEE Conference on Computer Vision and Pattern Recognition.
18. Huang, Z., Shan, S., Wang, R., Chen, X.: Coupling alignments with recognition for still-to-video face recognition (2013) IEEE International Conference on Computer Vision.

19. Lu, J., Wang, G., Moulin, P.: Image set classification using holistic multiple order statistics features and localized multi-kernel metric learning (2013) IEEE International Conference on Computer Vision.
20. Wright, J., Yang, A., Ganesh, A., Sastry, S., Ma, Y.: Robust face recognition via sparse representation. **31** (2009) 210–227 IEEE Transactions on Pattern Analysis and Machine Intelligence.
21. Liu, G., Lin, Z., Yu, Y.: Robust subspace segmentation by low-rank representation (2010) International Conference on Machine Learning.
22. Elhamifar, E., Vidal, R.: Sparse subspace clustering: Algorithm, theory, and applications. IEEE Transactions on Pattern Analysis and Machine Intelligence **35** (2013) 2765–2781
23. Favaro, P., Vidal, R., Ravichandran, A.: A closed form solution to robust subspace estimation and clustering (2011) Proceedings of IEEE Conference on Computer Vision and Pattern Recognition.
24. He, R., Sun, Z., Tan, T., Zheng, W.S.: Recovery of corrupted low-rank matrices via half-quadratic based non convex minimization (2011) Proceedings of IEEE Conference on Computer Vision and Pattern Recognition.
25. Aharon, M., M., E., Bruckstein, A.: K-svd: An algorithm for designing overcomplete dictionaries for sparse representation. IEEE Transactions on Signal Processing **54** (2006) 4311–4322
26. Gross, R., Shi, J.: The cmu motion of body (mobo) database. Technical Report CMU-RI-TR-01-18, Robotics Institute, Pittsburgh, PA (2001)
27. Viola, P., Jones, M.: Robust real-time face detection. International Journal of Computer Vision **57** (2004) 137–154