Video-based Face Recognition via Joint Sparse Representation

Yi-Chen Chen, Vishal M. Patel, Sumit Shekhar, Rama Chellappa and P. Jonathon Phillips

Abstract— **In video-based face recognition, a key challenge is in exploiting the extra information available in a video; e.g., face, body, and motion identity cues. In addition, different video sequences of the same subject may contain variations in resolution, illumination, pose, and facial expressions. These variations contribute to the challenges in designing an effective video-based face-recognition algorithm. We propose a novel multivariate sparse representation method for video-to-video face recognition. Our method simultaneously takes into account correlations as well as coupling information among the video frames. Our method jointly represents all the video data by a sparse linear combination of training data. In addition, we modify our model so that it is robust in the presence of noise and occlusion. Furthermore, we kernelize the algorithm to handle the non-linearities present in video data. Numerous experiments using unconstrained video sequences show that our method is effective and performs significantly better than many state-of-the-art video-based face recognition algorithms in the literature.**

I. Introduction

Though face recognition research [1] has traditionally concentrated on recognition from still images, recently, video-based face recognition has also gained a lot of traction. Faces are essentially articulating three dimensional objects. For faces, cues from motion possesses useful information in the form of behavioral traits such as idiosyncratic head movements and gestures, which can potentially aid in recognition tasks. Humans efficiently fuse face, body, and motion when recognizing people in video [2]. From video sequence effective representations such as three dimensional face models or super-resolved frames can be estimated. These techniques have the potential to improve recognition results.

While the advantage of using motion information in face videos has been widely recognized, computational models for video-based face recognition have only recently gained attention. In this paper, we consider the problem of video-to-video face recognition where one is presented with a video sequence and the goal is to recognize the person in the video. A key challenge is exploiting the extra information available in a video. In addition, different video sequences of the same subject may contain variations in resolution, illumination, pose, and facial expressions. These variations contribute to the difficulties in designing an effective video-based face recognition algorithm.

Yi-Chen Chen, Vishal M. Patel, Sumit Shekhar and Rama Chellappa are with the Department of Electrical and Computer Engineering and the Center for Automation Research, UMIACS, University of Maryland, College Park, MD. {chenyc08, pvishalm, sshekha, rama}@umiacs.umd.edu

P. Jonathon Phillips is with National Institute of Standards and Technology, Gaithersburg, MD. jonathon.phillips@nist.gov. The identification of any commercial product or trade name does not imply endorsement or recommendation by NIST.

Numerous methods have been proposed to exploit the extra information available in video. Three proposed techniques include frame-based recognition algorithms and fusing the results [3], modeling the temporal correlations explicitly to recognize the human [4], and extract joint appearance and behavioral features from the sequences [5], [6]. A major drawback of the frame-based fusion approach is that it does not exploit the temporal information present in a video sequence.

It has been shown that in a generic video-face recognition algorithm, performance can be significantly improved by simultaneously performing recognition and tracking [5], [7], [8], [9], [10], [11]. A statistical method for video-based face recognition was recently presented in [12]. These methods use subspace-based models and tools from Riemannian geometry of the Grassmann manifold. Intrinsic and extrinsic statistics are derived for the maximum-likelihood classification applications. An image set classification method for the video-based face recognition problem was recently proposed in [13]. This method is based on a measure of between-set dissimilarity defined as the distance between sparse approximated nearest points of two image sets and uses a scalable accelerated proximal gradient method for optimization. A dictionary-based face recognition method from video was recently proposed in [14]. This method was shown to be robust to illumination and pose variations.

The method presented in [6] represents the appearance variations due to shape and illumination on faces by assuming that the shape-illumination manifold of all possible illuminations and poses is generic for faces. This in turn implies that the shape-illumination manifold can be estimated using a set of subjects independent of the test set. It was shown that the effects of face shape and illumination can be learnt using PCA from a small, unlabeled set of video sequences of faces acquired in randomly varying lighting conditions. Given a novel sequence, the learned model is used to decompose the face appearance manifold into albedo and shape-illumination manifolds, producing the classification decision using robust likelihood estimation.

In recent years, the theories of sparse representation and dictionary learning have emerged as powerful tools for efficiently processing of image and video data in non-traditional ways. This is due in part to the fact that signals and images of interest can be sparse in some properly designed dictionary. This has led to a resurgence of the principles of sparse representation and dictionary learning for biometrics recognition [15], [16], [17]. One of the main advantages of using sparse representations for biometrics recognition is that they tend to be to robust to noise and occlusion [15].

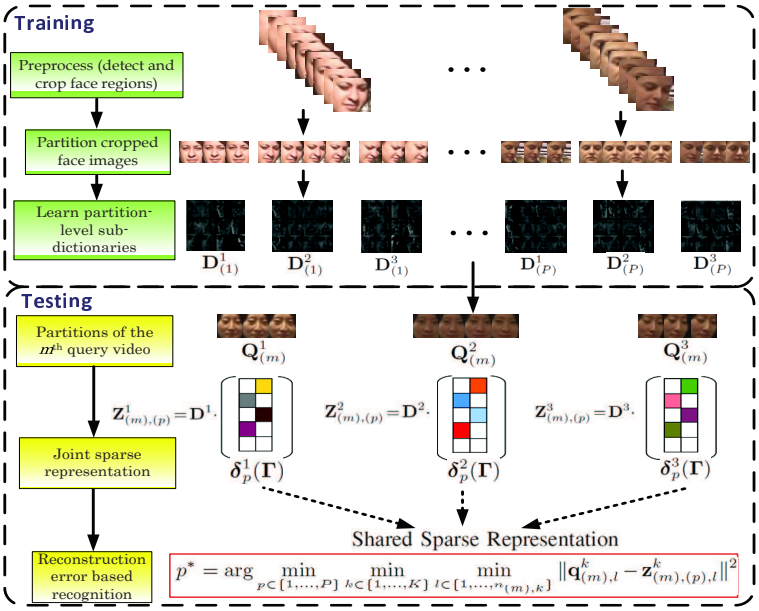


Fig. 1: Overview of the proposed approach.

Motivated by the success of sparse representation and dictionary learning in biometrics recognition, we propose a joint sparsity-based approach for unconstrained video-to-video face recognition. This method is based on the well known regularized regression method, multi-task multivariate Lasso [18], [19]. Our method simultaneously takes into account correlations as well as coupling information between frames of a video while enforcing joint sparsity within each frame's observation. We extend our model so that it is robust to both occlusion and noise. Furthermore, we kernelize the algorithm to enable it to handle the non-linearities present in video data. One of the main advantages of our method is that it does not require face tracking and is robust to changes in illumination and pose.

Figure 1 shows an overview of our approach. In the training stage, from cropped face images, we partition the pth video sequence, $p = 1, 2, ..., P$, so that frames with the same pose and illumination condition are in one partition. We then find the best representation for each member in these partitions by learning dictionaries under strict sparsity constraints. Each learned sub-dictionary $\mathbf{D}^k_{(p)}$ for $k = 1, 2, 3$, and $p = 1, 2, ..., P$, represents the pth video's kth face feature that is under a particular pose and/or illumination condition.

In the testing stage, the same partition step is applied on the mth query video sequence to acquire partitions, $\mathbf{Q}^k_{(m)}$, $k = 1, 2, 3$. Then, for each $\mathbf{Q}^k_{(m)}$, sub-dictionaries from all target videos are found and concatenated to form the dictionary $\mathbf{D}^k$. Using $\mathbf{D}^k, k = 1, 2, 3$, and a query sample, joint sparse representation $\mathbf{\Gamma} = [\mathbf{\Gamma}^1\ \mathbf{\Gamma}^2\ \mathbf{\Gamma}^3]$, is found to make decisions for recognition under the minimum class reconstruction error criterion.

A. Paper organization

This paper is organized as follows: Joint sparse representation-based recognition algorithm is detailed in Section II. This includes building partition-level sub-dictionaries for each video and fusion among these sub-dictionaries using joint sparse representation for identification and verification. Experimental results are presented in Section III and Section IV concludes the paper with a brief summary and discussion.

II. Proposed Approach

In this section, we present our joint sparsity-based method for video-to-video face recognition.

A. Building partition-level sub-dictionaries

For each frame in a video sequence, we first detect and crop the face regions. We then partition all the cropped face images into K different partitions using a variation of the algorithm [20] so that distinct partitions capture different pose and/or illumination conditions. To remove the temporal redundancy while capturing variations due to changes in pose and illumination, we construct a dictionary for each partition. A dictionary is learned with the minimum representation error under a strict sparseness constraint. Thus, there are K sub-dictionaries built to represent each video sequence. Due to changes in pose and lighting in a video sequence, the number of face images in a partition will vary. For partitions with very few images, before building the corresponding dictionary, we augment the partition by introducing synthesized face images. This is done by creating horizontally, vertically or diagonally shifted face images, or by in-plane rotation of faces.

Let $\mathbf{G}^k_{(p)}$ be the augmented gallery matrix of the kth partition of the pth video sequence. In augmented gallery

martix $\mathbf{G}_{(p)}^k = [\mathbf{g}_{(p),1}^k \mathbf{g}_{(p),2}^k ...]$, each column is a vectorized form of the corresponding cropped grayscale face image of size L. Given $\mathbf{G}_{(p)}^k$, a dictionary $\hat{\mathbf{D}}_{(p)}^k \in \mathbb{R}^{L \times \tilde{K}}$ is learned such that the columns of $\mathbf{G}_{(p)}^k$ are best represented by linear combinations of $\tilde{K}$ atoms of $\hat{\mathbf{D}}_{(p)}^k$. This can be done by minimizing the following representation error

$$(\hat{\mathbf{D}}_{(p)}^k, \hat{\mathbf{\Lambda}}_{(p)}^k) = \underset{\tilde{\mathbf{D}}_{(p)}^k, \mathbf{\Lambda}_{(p)}^k}{\operatorname{argmin}} \|\mathbf{G}_{(p)}^k - \tilde{\mathbf{D}}_{(p)}^k \mathbf{\Lambda}_{(p)}^k\|_F^2, \quad s.t. \ \|\boldsymbol{\lambda}_l\|_0 \leq T_0, \ \forall l, \qquad (1)$$

where $\boldsymbol{\lambda}_l$ is the lth column of coefficient matrix $\mathbf{\Lambda}_{(p)}^k$ and T_0 is a sparsity parameter. The ℓ_0 sparsity measure $\|\cdot\|_0$ counts the number of nonzero elements in the representation and $\|\mathbf{Y}\|_F$ is the Frobenius norm of the matrix $\mathbf{Y}$ defined as $\|\mathbf{Y}\|_F = \sqrt{\sum_{i,j} Y_{i,j}^2}$.

One of the most well-known algorithms for learning such dictionaries is the K-SVD[1] algorithm [21]. The K-SVD algorithm alternates between sparse-coding and dictionary update steps. In the sparse-coding step, the dictionary $\tilde{\mathbf{D}}_{(p)}^k$ is fixed and the representation vectors $\boldsymbol{\lambda}_l$ are found for each example $\mathbf{g}_{(p),i}^k$. Then, the dictionary is updated atom-by-atom in an efficient way [21]. Due to its simplicity and efficiency, we adapt the K-SVD algorithm to obtain $\hat{\mathbf{D}}_{(p)}^k$'s, the partition-specific dictionaries of the pth video.

B. Sparse representation for video-based face recognition (SRV)

Let Q denote the total number of query video sequences. Given the mth query video sequence $\mathbf{Q}_{(m)}$, where $m = 1, ..., Q$, we can write $\mathbf{Q}_{(m)} = \bigcup_{k=1}^K \mathbf{Q}_{(m)}^k$. Partitions $\mathbf{Q}_{(m)}^k$ are expressed by $\mathbf{Q}_{(m)}^k = [\mathbf{q}_{(m),1}^k \ \mathbf{q}_{(m),2}^k \ \cdots \ \mathbf{q}_{(m),n_{(m),k}}^k]$, where $\mathbf{q}_{(m),l}^k$ is the vectorized form of the lth of the total $n_{(m),k}$ cropped face images belonging to the kth partition.

We exploit the joint sparsity of coefficients from different partitions to make a joint decision. We denote $\{\mathbf{Q}_{(m)}^k\}_{k=1}^K$ as a set of K partitions of the mth query video, with partition $\mathbf{Q}_{(m)}^k$ consisting of $n_{(m),k}$ face images, and $\mathbf{D}^k = [\mathbf{D}_{(1)}^k \ \mathbf{D}_{(2)}^k \ ... \ \mathbf{D}_{(P)}^k]$ as the concatenation of the kth sub-dictionaries from all target videos. Letting $\mathbf{\Gamma} = [\mathbf{\Gamma}^1 \ \mathbf{\Gamma}^2 \dots \mathbf{\Gamma}^K] \in I\!R^{d \times n_{(m)}}$ be the matrix formed by concatenating the coefficient matrices with $d = \sum_{j=1}^P d_j$ and $n_{(m)} = \sum_{k=1}^K n_{(m),k}$, we seek the row-sparse matrix $\mathbf{\Gamma}$ by solving the following ℓ_1/ℓ_q-regularized least square problem

$$\hat{\mathbf{\Gamma}} = \arg\min_{\mathbf{\Gamma}} \frac{1}{2} \sum_{k=1}^K \|\mathbf{Q}_{(m)}^k - \mathbf{D}^k \mathbf{\Gamma}^k\|_F^2 + \lambda \|\mathbf{\Gamma}\|_{1,q} \qquad (2)$$

where λ is a positive parameter and q is set greater than 1 to make the optimization problem convex. Here, $\|\mathbf{\Gamma}\|_{1,q}$ is a norm defined as $\|\mathbf{\Gamma}\|_{1,q} = \sum_{i=1}^d \|\boldsymbol{\gamma}^i\|_q$ where $\boldsymbol{\gamma}^i$'s are the row vectors of $\mathbf{\Gamma}$. Problem (2) can be solved using the classical Alternating Direction Method of Multipliers (ADMM) [22], [23], [24]. See [22], [23] for more details on ADMM. For our experiments, we choose $q = 2$.

[1]Here "K" in "K-SVD" equals number of atoms $\tilde{K}$ in a learned dictionary, not number of partitions K of a video sequence.

1) Identification: For identification, we use the knowledge of the correspondence $\boldsymbol{f}(\cdot)$ between subjects and sequences to assign the query video sequence $\mathbf{Q}_{(m)}$ to subject $i^* = \boldsymbol{f}(p^*)$, where p^* is the sequence-level decision. Once $\hat{\mathbf{\Gamma}}$ is obtained, p^* is declared as the one that produces the smallest approximation error.

$$p^* = \arg\min_p \min_{k \in \{1,...,K\}} \min_{l \in \{1,...,n_{(m),k}\}} \|\mathbf{q}_{(m),l}^k - \mathbf{D}^k \boldsymbol{\delta}_{p,l}^k(\hat{\mathbf{\Gamma}})\|^2, \qquad (3)$$

where $\boldsymbol{\delta}_{p,l}^k(\cdot)$ is the indicator function defined by keeping the coefficients corresponding to the kth partition from pth target video for the lth query image, and setting coefficients in all other rows and columns equal to zero.

2) Verification: For verification, given a query video sequence and any gallery video sequence, the goal is to correctly determine whether these two belong to the same subject. The well-known receiver operating characteristic (ROC) curve, which describes relations between false acceptance rates (FARs) and true acceptance rates (TARs), is used to evaluate the performance of verification algorithms. As the TAR increases, so does the FAR. Therefore, one would expect an ideal verification framework to have all TARs equal to 1 for any FARs. The ROC curves can be computed given a similarity matrix. In the proposed method, the residual between a query $\mathbf{Q}^{(m)}$ and the pth target video, is used to fill in the (m, p) entry of the similarity matrix. Denoting the residual by $\mathbf{R}^{(m,p)}$, we have

$$\mathbf{R}^{(m,p)} = \min_{k \in \{1,2,...,K\}} \mathbf{R}_k^{(m,p)}, \qquad (4)$$

where

$$\mathbf{R}_k^{(m,p)} \triangleq \min_{l \in \{1,...,n_{(m),k}\}} \|\mathbf{q}_{(m),l}^k - \mathbf{D}^k \boldsymbol{\delta}_{p,l}^k(\hat{\mathbf{\Gamma}})\|^2. \qquad (5)$$

In other words, we select the minimum residual among all $l \in \{1, 2, ..., n_{(m),k}\}$, and all $k \in \{1, 2, ..., K\}$, as the similarity between the query video sequence $\mathbf{Q}^{(m)}$ and the pth target video.

The SRV algorithm is summarized in Algorithm 1.

Algorithm 1: Sparse representation for video-based face recognition (SRV)

Input: Partition-level sub-dictionaries $\{\mathbf{D}^k\}_{i=k}^K$ and query videos $\{\mathbf{Q}_{(m)}^k\}_{k=1}^K$.

Procedure: Obtain $\hat{\mathbf{\Gamma}}$ by solving

$$\hat{\mathbf{\Gamma}} = \arg\min_{\mathbf{\Gamma}} \frac{1}{2} \sum_{k=1}^K \|\mathbf{Q}_{(m)}^k - \mathbf{D}^k \mathbf{\Gamma}^k\|_F^2 + \lambda_1 \|\mathbf{\Gamma}\|_{1,q},$$

Output:

(Identification) `video` $p^* = \arg\min_p \min_{k \in \{1,...,K\}} \min_{l \in \{1,...,n_{(m),k}\}} \|\mathbf{q}_{(m),l}^k - \mathbf{D}^k \boldsymbol{\delta}_{p,l}^k(\hat{\mathbf{\Gamma}})\|^2$, `subject` $i^* = \boldsymbol{f}(p^*)$.

(Verification) Use the similarity $\mathbf{R}^{(m,p)}$ computed by (4) and (5) to construct the similarity matrix, from which the ROC curves can be obtained.

C. Finding aligned sub-dictionaries for unconstrained videos

The formulation presented above is made under the assumption that $\mathbf{D}^k$ is a concatenation of sub-dictionaries that are aligned with $\mathbf{Q}^k_{(m)}$. In other words, if $\mathbf{Q}^k_{(m)}$ collects a subject's left side face images from the mth video, then $\mathbf{D}^k$ must also collect sub-dictionaries of left side faces from all target videos. In practical situations, unlike constrained videos, illumination and pose conditions in an unconstrained video vary. For example, some query videos contain left side face images only, while some target videos contain frontal face images only. In addition, no information among the partitions is given on which partition represents which specific pose and illumination condition. To overcome these difficulties before finding joint sparse representation, we find approximately aligned dictionaries $\mathbf{D}^k$ such that $\mathbf{D}^k_{(p)}, p = 1, 2, ..., P$ are obtained by:

$$\mathbf{D}^k_{(p)} = \arg \max_{\hat{\mathbf{D}}^u_{(p)}, u\in\{1,2,...,K\}} C_u, \tag{6}$$

where C_u is the number of votes for the uth sub-dictionary of the pth target video (i.e., $\hat{\mathbf{D}}^u_{(p)}$ in (1)) collected from each $\mathbf{q}^k_{(m),l}$ in $\mathbf{Q}^k_{(m)}$. In other words, the aligned sub-dictionaries are determined by the majority vote criterion. Each query image $\mathbf{q}^k_{(m),l}$ in the kth partition of the mth query video $\mathbf{Q}^k_{(m)}$ votes for $\hat{\mathbf{D}}^u_{(p)}$ such that it has the minimum reconstruction error from its projection on $\hat{\mathbf{D}}^v_{(p)}$:

$$u = \arg\min_v \|\mathbf{q}^k_{(m),l} - \hat{\mathbf{D}}^v_{(p)}\hat{\mathbf{D}}^{v\dagger}_{(p)}\mathbf{q}^k_{(m),l}\|^2, \tag{7}$$

where $\hat{\mathbf{D}}^{v\dagger}_{(p)}$ is the pseudo-inverse of $\hat{\mathbf{D}}^v_{(p)}$.

D. Kernel sparse representation for video-based face recognition (KSRV)

The class identities in different partitions may not be linearly separable. Hence, we also extend the joint sparse representation framework to the non-linear kernel space. The kernel function, $\kappa : \mathbb{R}^n \times \mathbb{R}^n$, is defined as the inner product

$$\kappa(\mathbf{d}_i, \mathbf{d}_j) = \langle \phi(\mathbf{d}_i), \phi(\mathbf{d}_j) \rangle$$

where, ϕ is an implicit mapping projecting the vector $\mathbf{d}$ into a higher dimensional space.

Considering the general case of K partitions of the mth query video with $\{\mathbf{Q}^k_{(m)}\}_{k=1}^K$ as a set of $n_{(m),k}$ observations, the feature space representation can be written as:

$$\mathbf{\Phi}(\mathbf{Q}^k_{(m)}) = [\phi(\mathbf{q}^k_{(m),1})\ \phi(\mathbf{q}^k_{(m),2})\ ...\ \phi(\mathbf{q}^k_{(m),n_{(m),k}})]$$

Similarly, the dictionary of training samples for the kth partition can be represented in feature space as

$$\mathbf{\Phi}(\mathbf{D}^k) = [\phi(\mathbf{D}^k_1), \phi(\mathbf{D}^k_2), \cdots, \phi(\mathbf{D}^k_P)]$$

As in joint linear space representation, we have:

$$\mathbf{\Phi}(\mathbf{Q}^k_{(m)}) = \mathbf{\Phi}(\mathbf{D}^k)\mathbf{\Gamma}^k$$

where, $\mathbf{\Gamma}^k$ is the coefficient matrix associated with partition k. Incorporating information from all the partitions, we solve the following optimization problem similar to the linear case:

$$\hat{\mathbf{\Gamma}} = \arg\min_{\mathbf{\Gamma}} \frac{1}{2}\sum_{k=1}^K \|\mathbf{\Phi}(\mathbf{Q}^k_{(m)}) - \mathbf{\Phi}(\mathbf{D}^k)\mathbf{\Gamma}^k\|_F^2 + \lambda\|\mathbf{\Gamma}\|_{1,q} \tag{8}$$

where, $\mathbf{\Gamma} = [\mathbf{\Gamma}^1, \mathbf{\Gamma}^2, \cdots, \mathbf{\Gamma}^K]$. It is clear that the information from all the partitions of a video are integrated via the shared sparsity pattern of the matrices $\{\mathbf{\Gamma}^k\}_{k=1}^K$. This can be reformulated in terms of kernel matrices as:

$$\begin{aligned}\hat{\mathbf{\Gamma}} = \arg\min_{\mathbf{\Gamma}} \frac{1}{2}\sum_{k=1}^K \big(&\text{trace}({\mathbf{\Gamma}^k}^T \mathbf{K}_{\mathbf{D}^k,\mathbf{D}^k}\mathbf{\Gamma}^k) \\ &-2\text{trace}(\mathbf{K}_{\mathbf{D}^k,\mathbf{Q}^k_{(m)}}\mathbf{\Gamma}^k)\big) + \lambda\|\mathbf{\Gamma}\|_{1,q}\end{aligned} \tag{9}$$

where, the kernel matrix $\mathbf{K}_{\mathbf{X},\mathbf{Y}}$ is defined as:

$$\mathbf{K}_{\mathbf{X},\mathbf{Y}}(i,j) = \kappa(\mathbf{x}_i, \mathbf{y}_j) = \langle \phi(\mathbf{x}_i), \phi(\mathbf{y}_j) \rangle, \tag{10}$$

with $\mathbf{x}_i$ and $\mathbf{y}_j$ being i^{th} and j^{th} columns of $\mathbf{X}$ and $\mathbf{Y}$ respectively. Similar to the linear case, problem (8) can be solved using the ADMM type of algorithm.

1) Identification: Once $\hat{\mathbf{\Gamma}}$ is obtained, we assign $\mathbf{Q}_{(m)}$ to subject $i^* = \boldsymbol{f}(p^*)$, where p^* is obtained as follows.

$$\begin{aligned}p^* &= \arg\min_p \min_k \min_{l\in\{1,...,n_{(m),k}\}} \|\phi(\mathbf{q}^k_{(m),l}) - \mathbf{\Phi}(\mathbf{D}^k_{(p)})\boldsymbol{\delta}^k_{p,l}(\hat{\mathbf{\Gamma}})\|^2 \\ &= \arg\min_p \min_k \min_{l\in\{1,...,n_{(m),k}\}} \Big\{\text{trace}(\mathbf{K}_{\mathbf{Q}^k_{(m)}\mathbf{Q}^k_{(m)}}) \\ &\quad -2\ \text{trace}(\boldsymbol{\delta}^k_{p,l}(\hat{\mathbf{\Gamma}})^T \mathbf{K}_{\mathbf{D}^k_{(p)}}\mathbf{K}_{\mathbf{Q}^k_{(m)}}\boldsymbol{\delta}^k_{p,l}(\hat{\mathbf{\Gamma}})) \\ &\quad +\text{trace}(\boldsymbol{\delta}^k_{p,l}(\hat{\mathbf{\Gamma}})^T \mathbf{K}_{\mathbf{D}^k_{(p)}\mathbf{D}^k_{(p)}}\boldsymbol{\delta}^k_{p,l}(\hat{\mathbf{\Gamma}}))\Big\}.\end{aligned} \tag{11}$$

2) Verification: Similar to the linear case in II-B.1, we use (4) to construct the similarity $\mathbf{R}^{(m,p)}$, with $\mathbf{R}^{(m,p)}_k$ in (5) replaced with

$$\mathbf{R}^{(m,p)}_k \triangleq \min_{l\in\{1,...,n_{(m),k}\}} \|\phi(\mathbf{q}^k_{(m),l}) - \mathbf{\Phi}(\mathbf{D}^k_{(p)})\boldsymbol{\delta}^k_{p,l}(\hat{\mathbf{\Gamma}})\|^2. \tag{12}$$

The KSRV algorithm is summarized in Algorithm 2.

Algorithm 2: Kernel sparse representation for video-based face recognition (KSRV)

Input: Partition-level sub-dictionaries $\{\mathbf{D}^k\}_{i=k}^K$ and query videos $\{\mathbf{Q}_{(m),k}\}_{k=1}^K$.

Procedure: Obtain $\hat{\mathbf{\Gamma}}$ by solving

$$\hat{\mathbf{\Gamma}} = \arg\min_{\mathbf{\Gamma}} \frac{1}{2}\sum_{k=1}^K \|\mathbf{\Phi}(\mathbf{Q}^k_{(m)}) - \mathbf{\Phi}(\mathbf{D}^k)\mathbf{\Gamma}^k\|_F^2 + \lambda\|\mathbf{\Gamma}\|_{1,q} \tag{13}$$

Output:
(Identification) `video` $p^* = \arg\min_p \min_k \min_{l\in\{1,...,n_{(m),k}\}} \|\phi(\mathbf{q}^k_{(m),l}) - \mathbf{\Phi}(\mathbf{D}^k_{(p)})\boldsymbol{\delta}^k_{p,l}(\hat{\mathbf{\Gamma}})\|^2$, `subject` $i^* = \boldsymbol{f}(p^*)$.
(Verification) Use the similarity $\mathbf{R}^{(m,p)}$ computed by (4) and (12) to construct the similarity matrix, from which the ROC curves can be obtained.

III. Experimental Results

To illustrate the effectiveness of our method, we present experimental results on three datasets for video-based face recognition: the UMD dataset [25], the Multiple Biometric Grand Challenge (MBGC) dataset [26],[27], and the Honda/UCSD dataset [5].

A. UMD video

The UMD dataset contains 12 videos recorded of a group of 16 subjects. The videos were collected in a high definition format (1920×1088 pixels). They contain sequences of subjects standing without walking toward the camera, which we refer to as standing sequences, and sequence of each subject walking toward the camera, which we refer to as walking sequences. After segmenting the videos according to subjects and sequence types, we obtained 93 sequences in total: 70 standing sequences and 23 walking sequences. Figure 2(a) shows example frames from four different standing sequences, where most subjects are standing in a group. As some subjects were having conversations and others were looking elsewhere, their faces were sometimes non-frontal or partially occluded. Figure 2(b) shows example frames from four different walking sequences, in each of which a single subject was walking toward the camera, with a frontal face for most of the time. However, the walking subject's head sometimes turned to the right or left showing a profile face. Furthermore, for both types of sequences, the camera was not always static. Figure 2(c) shows example frames with blurred subjects due to the camera motion.

(a)

(b)

(c)

Fig. 2: Example frames from the UMD dataset. (a) Standing sequences. (b) Walking sequences. (c) Frames with blurred subjects due to the camera motion. Faces in standing sequences were sometimes non-frontal or partially occluded, while faces in walking sequences were frontal most of the time. Camera movements raise the additional difficulty for face tracking and recognition.

Figure 3 shows an example of output from the video partitioning stage. For results in Figure 3, the number of partitions is $K = 3$. Results are presented for 8 subjects for walking sequences[2]. Each row shows up to 30 partitioned cropped face images from the same video sequence. We use blue lines to distinguish among different subjects. It can be seen that each partition from a video sequence encodes a particular pose, illumination or blur condition, and different partitions represent different conditions. We can see the partition results are not ideal for all frames, as some frames suffer from misalignment or camera movements.

Following the experimental setup of [12], we conducted a leave-one-out identification experiment on 3 subsets of cropped face images from the walking videos. These 3 subsets are $S2$ (subjects which have at least two video sequences: 16 subjects, 93 sequences), $S3$ (subjects which have at least three sequences: 15 subjects, 91 sequences) and $S6$ (subjects which have at least six sequences: 7 subjects, 51 sequences)[3]. Table I lists the percentages of correct identifications for this experiment. The proposed sparsity-based methods, SRV and KSRV obtained average identification rates better than other compared methods. The methods compared in Table I include the wrapped Gaussian common pole (WGCP) method and other statistical methods reported in [12],[28], as well as the sparse approximated nearest points (SANP) method [13].

UMD videos	PM [12],[28]	KD [12],[28]	WGCP [12]	SANP [13]	SRV	KSRV
$S2$	82.80	81.72	82.97	92.47	92.47	**93.55**
$S3, S4, S5$	84.62	83.52	83.52	93.41	94.51	**94.51**
$S6$	**98.04**	96.08	88.23	**98.04**	**98.04**	**98.04**
Average	88.49	87.11	84.91	94.64	95.01	**95.37**

TABLE I: Identification rates of leave-one-out testing experiments on UMD videos. Both SRV and KSRV outperform the other compared methods.

In the next set of experiments with the UMD dataset, we conduct "S vs. W" (i.e., "Standing vs. Walking" - standing sequences as probe and walking sequences as gallery) and "W vs. S" (i.e., "Walking vs. Standing" - walking sequences as probe and standing sequences as gallery) experiments. Correct identification rates are shown in Table II. Our sparsity-based methods tied with each other and they are comparable with other methods.

UMD videos	PM [12],[28]	KD [12],[28]	WGCP [12]	SANP [13]	SRV	KSRV
S vs. W	65.71	51.43	30.00	**87.14**	84.29	84.29
W vs. S	73.91	65.22	43.48	**91.30**	**91.30**	**91.30**
Average	69.81	58.33	36.74	**89.22**	87.80	87.80

TABLE II: Identification rates of "Standing vs. Walking" and "Walking vs. Standing" experiments on the UMD videos.

Figure 4(a) shows the ROCs for the verification experiments using $S2$, $S3$ and $S6$ on the UMD dataset. Figure 4(b) shows the ROC curves for "Walking vs. Standing" and "Standing vs. Walking" experiments. The proposed sparsity-

[2]For the illustration purpose only, here we show results of 8 subjects only.

[3]For the UMD video sequences, the three sets $S3$, $S4$ and $S5$ are identical.

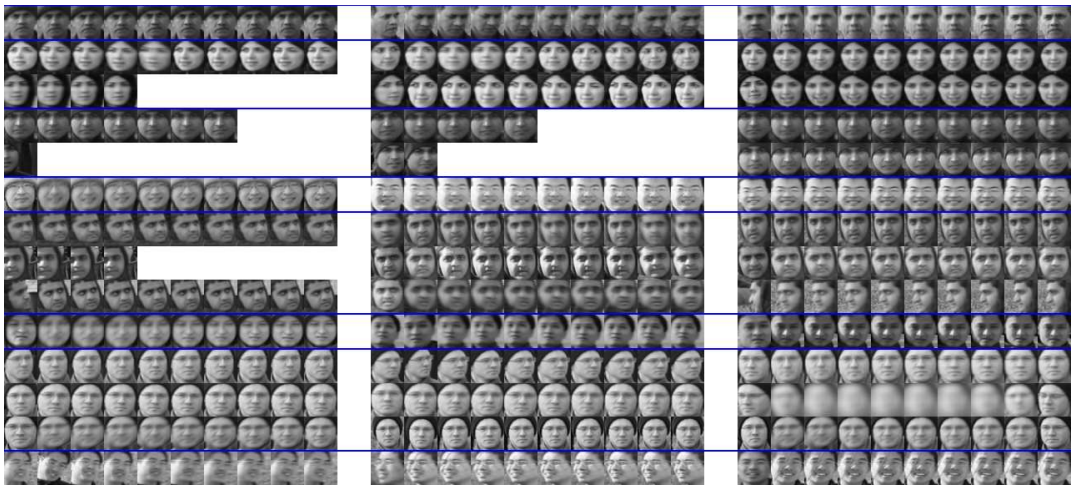
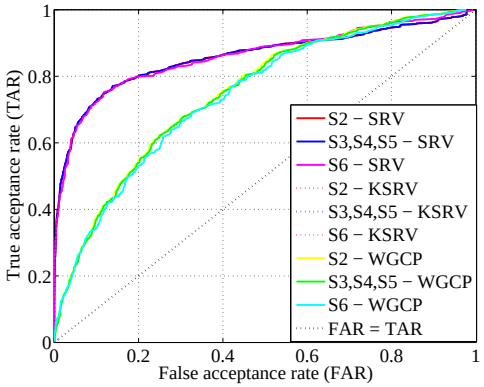
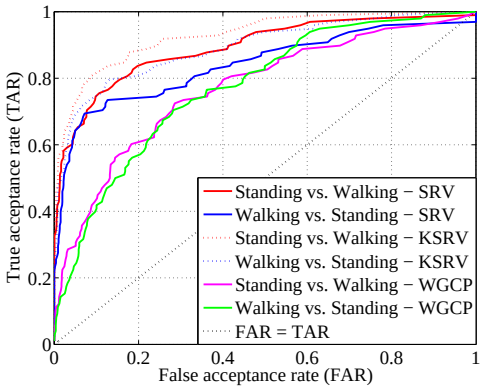


Fig. 3: Partition results of the UMD walking videos. Blue lines separate different subjects. Face images from a video sequence are shown in a row, and are further divided into three partitions. Each partition shows up to 10 face images. A partition represents a particular pose, illumination or blur condition.

based methods obtained better ROC curves than the WGCP method.



(a)



(b)

Fig. 4: (a) The ROC curves of the S2, S3, and S6 leave-one-out verification experiments on UMD videos. (b) The ROC curves of standing vs. walking, and walking vs. standing verification experiments on UMD videos. The sparsity-based methods obtained better ROC curves than the WGCP method for most FARs.

B. MBGC Video version 1

The MBGC Video version 1 dataset (Notre Dame dataset) contains 399 walking (frontal-face) and 371 activity (profile-face) video sequences recorded of 146 subjects. Both types of sequences were collected in standard definition (SD) format (720×480 pixels) and high definition (HD) format (1440×1080 pixels). The 399 walking sequences consist of 201 sequences in SD and 198 in HD. For the 371 walking video sequences, 185 are in SD and 186 are in HD. Figure 5 shows four example frames from four different walking sequences, where each subject walks toward the video camera with a frontal pose for most of the time and turns to the left or right showing the profile face at the end. The challenging conditions in these videos include frontal and non-frontal faces in shadow.

We conducted leave-one-out identification experiments on 3 subsets of the cropped face images from the walking videos. These 3 subsets are $S2$ (144 subjects, 397 videos), $S3$ (55 subjects, 219 videos) and $S4$ (54 subjects, 216 videos). Table III lists the percentages of correct identifications for this experiment. Our sparsity-based methods outperform the other compared methods.

MBGC walking	PM [12],[28]	KD [12],[28]	WGCP [12]	SANP [13]	SRV	KSRV
$S2$	43.79	39.74	63.79	83.88	**86.65**	**86.65**
$S3$	53.88	50.22	74.88	84.02	87.67	**88.58**
$S4$	53.70	50.46	75	84.26	87.96	**88.89**
Average	50.46	46.81	71.22	84.05	87.43	**88.04**

TABLE III: Identification rates of leave-one-out testing experiments on the MBGC walking videos. Our sparsity-based methods obtained the best results.

In the second set of experiments, we selected videos of subjects that are in at least two videos (i.e., $S2$). We divide all these videos into SD and HD videos, to conduct "SD vs. HD"

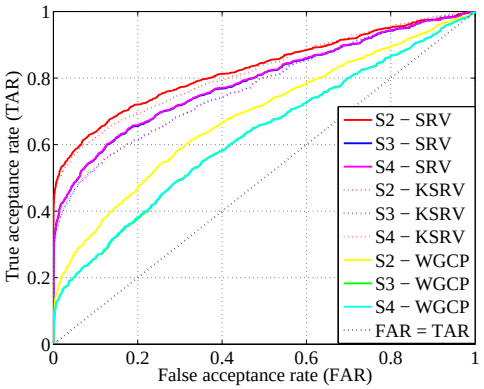
Fig. 5: Examples of MBGC walking video sequences.

(SD as probe; HD as gallery) and "HD vs. SD" (HD as probe; SD as gallery) experiments. In this setting, we examine the effect of varying the number video sequences per person in the gallery. We divide the videos into two groups: gallery and probe. For most subjects (89 out of 144), this setting allows only one video per subject for training, unlike the previous leave-one-out test in which there are always at least two training video sequences per subject (the subject whose video is currently used as probe is excluded). Correct identification rates are shown in Table IV. Our fusion methods significantly outperformed other methods. The WGCP [12] method finds projections of training samples on a Grassmann manifold on its tangent plane and uses them to learn a pre-assumed Gaussian model. While the geodesic distance of any point on the manifold to the pole (i.e., the tangent point of the manifold and the corresponding tangent plane) is maintained, this property does not always apply to the geodesic distance between any pair of points on the manifold. Also, the pre-assumed Gaussian model may not be appropriate to model the training samples. The SANP [13] method is based on image set classification. The major limitation of this method is that it relies on the unseen appearances of a set to be modeled by affine combinations of samples. While this may be true for some variations in facial illumination, it does not hold for the extreme variations especially in the presence of shadows, pose and expression variations. Our method overcomes this limitation by learning and fusing across different partition specific dictionaries.

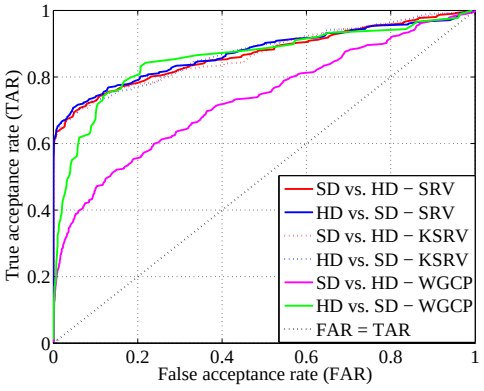
MBGC walking	PM [12],[28]	KD [12],[28]	WGCP [12]	SANP [13]	SRV	KSRV
SD vs. HD	61.31	55.78	30.15	41.71	**91.96**	91.46
HD vs. SD	68.69	56.06	30.30	45.96	90.40	**91.41**
Average	65	55.92	30.23	43.84	91.18	**91.44**

TABLE IV: Identification rates of "SD vs. HD" and "HD vs. SD" experiments on the MBGC walking video subset $S2$ (the subset that contains subjects who have at least two video sequences). In this experiment, most subjects (89 out of 144) have only one video per subject available for training. Our sparsity-based fusion methods, SRV and KSRV achieve the best identification rates.

Figures 6(a) and (b) show the corresponding ROC curves for the verification experiments. The SRV and KSRV methods have similar performances in both figures. They give better ROC curves than the WGCP method in Figure 6(a) for all FARs, and in Figure 6(b) for low FARs.



(a)



(b)

Fig. 6: (a) The ROC curves for S2, S3 and S4 verification experiments on the MBGC walking videos. (b) The ROC curves of "SD vs. HD" and "HD vs. SD" experiments on the MBGC walking video subset $S2$. The proposed sparsity-based methods give better ROC curves than the WGCP method shown in (a), and in (b) for low FARs.

C. Honda/UCSD Dataset

For the final set of experiments, we use the Honda/UCSD Dataset [5]. The Honda Dataset consists of 59 video sequences from 20 distinct subjects. We follow the same experimental procedure presented in [13]. The experiments are done in three cases of the maximum set length (available number of cropped-face images per video sequence) as defined in [13]: 50, 100 and full length frames. Image resolution is 20×20 pixels. Table V shows the identification rates of the proposed sparsity-based methods and other six state-of-the-art methods [29],[30],[31],[32],[13]. We observe that both fusion methods obtained the highest average identification rates, tieing with each other.

Set length	DCC [29]	MMD [30]	MDA [31]	AHISD [32]
50 frames	76.92	69.23	74.36	87.18
100 frames	84.62	87.18	94.87	84.62
full length	94.87	94.87	97.44	89.74
Average	85.47	83.76	88.89	87.18
Set length	CHISD [32]	SANP [13]	SRV	KSRV
50 frames	82.05	84.62	**94.87**	**94.87**
100 frames	84.62	92.31	**97.44**	**97.44**
full length	92.31	**100**	97.44	97.44
Average	86.33	92.31	**96.58**	**96.58**

TABLE V: Identification rates on the Honda/UCSD Dataset. The proposed sparsity-based methods obtained the highest average identification rates.

IV. Conclusions

We proposed an effective joint sparsity-based approach for unconstrained video-based face recognition. In the training stage, we partition the face images extracted from a given video. Each partition captures different pose and illumination conditions and is encoded in different video sub-dictionaries. Each sub-dictionary encodes a face in a particular viewing condition. In the testing stage, the same partition is found for the query video. Then joint sparse representation is found to make decisions for recognition under the minimum class reconstruction error criterion. Various experiments on publicly available data sets show that our method is efficient and can perform significantly better than many state-of-the-art face recognition algorithms in the literature.

References

[1] W. Zhao, R. Chellappa, J. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *ACM Computing Surveys*, pp. 399–458, Dec. 2003.

[2] A. J. O'Toole, P. J. Phillips, S. Weimer, D. A. Roark, J. Ayyad, R. Barwick, and J. Dunlop, "Recognizing people from dynamic and static faces and bodies: Dissecting identity with a fusion approach," *Vision Research*, vol. 51, no. 1, pp. 74–83, 2011.

[3] A. Ross, K. Nandakumar, and A. K. Jain, *Handbook of Multibiometrics*. Springer, 2006.

[4] M. Tistarelli, S. Z. Li, and R. Chellappa, *Handbook of Remote Biometrics: For Surveillance and Security*. Springer, 2009.

[5] K.-C. Lee, J. Ho, M.-H. Yang, and D. Kriegman, "Visual tracking and recognition using probabilistic appearance manifolds," *Computer Vision and Image Understanding*, vol. 99, pp. 303–331, 2005.

[6] O. Arandjelovic and R. Cipolla, "Face recognition from video using the generic shape-illumination manifold," *European Conference on Computer Vision*, vol. 3954, pp. 27–40, 2006.

[7] G. Hager and P. Belhumeur, "Efficient region tracking with parametric models of geometry and illumination," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 10, pp. 1025–1039, Oct. 1998.

[8] A. Lanitis, C. Taylor, and T. Cootes, "Automatic interpretation and coding of face images using flexible models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 743 –756, July 1997.

[9] S. K. Zhou, R. Chellappa, and B. Moghaddam, "Visual tracking and recognition using appearance-adaptive models in particle filters," *IEEE Transactions on Image Processing*, vol. 13, no. 11, pp. 1491–1506, Nov. 2004.

[10] M. La Cascia, S. Sclaroff, and V. Athitsos, "Fast, reliable head tracking under varying illumination: an approach based on registration of texture-mapped 3d models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 4, pp. 322–336, Apr. 2000.

[11] G. Aggarwal, A. Veeraraghavan, and R. Chellappa, "3d facial pose tracking in uncalibrated videos," *International Conference on Pattern Recognition and Machine Intelligence*, 2005.

[12] P. K. Turaga, A. Veeraraghavan, A. Srivastava, and R. Chellappa, "Statistical computations on grassmann and stiefel manifolds for image and video-based recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 11, pp. 2273–2286, Nov. 2011.

[13] Y. Hu, A. S. Mian, and R. Owens, "Sparse approximated nearest points for image set classification," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 27–40, 2011.

[14] Y.-C. Chen, V. M. Patel, P. J. Phillips, and R. Chellappa, "Dictionary-based face recognition from video," *European Conference on Computer Vision*, 2012.

[15] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 2, pp. 210 –227, Feb. 2009.

[16] J. Pillai, V. Patel, R. Chellappa, and N. Ratha, "Secure and robust iris recognition using random projections and sparse representations," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, no. 9, pp. 1877 –1893, Sept. 2011.

[17] V. M. Patel, T. Wu, S. Biswas, P. J. Philips, and R. Chellappa, "Dictionary-based face recognition under variable lighting and pose," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 3, pp. 954–965, June 2012.

[18] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B*, vol. 68, pp. 49–67, Feb. 2006.

[19] L. Meier, S. V. D. Geer, and P. Bhlmann, "The group lasso for logistic regression," *Journal of the Royal Statistical Society: Series B*, vol. 70, pp. 53–71, Feb. 2008.

[20] N. Shroff, P. Turaga, and R. Chellappa, "Video précis: Highlighting diverse aspects of videos," *IEEE Transactions on Multimedia*, vol. 12, no. 8, pp. 853–868, Dec. 2010.

[21] M. Aharon, M. Elad, and B. A., "K-svd: An algorithm for desigining overcomplete dictionaries for sparse representation," *IEEE Transactions on Signal Processing*, Nov. 2006.

[22] J. Yang and Y. Zhang, "Alternating direction algorithms for l1 problems in compressive sensing," *SIAM Journal on Scientific Computing*, vol. 33, pp. 250–278, 2011.

[23] M. Afonso, J. Bioucas-Dias, and M. Figueiredo, "An augmented lagrangian approach to the constrained optimization formulation of imaging inverse problems," *IEEE Transactions on Image Processing*, vol. 20, pp. 681–695, March 2011.

[24] S. Shekhar, V. M. Patel, N. M. Nasrabadi, and R. Chellappa, "Joint sparsity-based robust multimodal biometrics recognition," *European Conference on Computer Vision Workshop on Information Fusion in Computer Vision for Concept Recognition*, 2012.

[25] R. Chellappa, J. Ni, and V. M. Patel, "Remote identification of faces: problems, prospects, and progress," *Pattern Recognition Letters*, vol. 33, no. 15, pp. 1849–1859, Oct. 2012.

[26] P. J. Phillips, P. J. Flynn, J. R. Beveridge, W. T. Scruggs, A. J. O'Toole, D. Bolme, K. W. Bowyer, B. A. Draper, G. H. Givens, Y. M. Lui, H. Sahibzada, J. A. Scallan III, and S. Weimer, "Overview of the multiple biometrics grand challenge," *International Conference on Biometrics*, 2009.

[27] National Institute of Standards and Technology, "Multiple biomertic grand challenge (MBGC)."

[28] P. K. Turaga, A. Veeraraghavan, and R. Chellappa, "Statistical analysis on stiefel and grassmann manifolds with applications in computer vision," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2008.

[29] M. K. Kim, O. Arandjelovic, and R. Cipolla, "Discrimitive learning and recognition of image set classes using canonical correlations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 6, pp. 1005–1018, June 2007.

[30] R. Wang, S. Shan, X. Chen, and W. Gao, "Manifold-manifold distance with application to face recognition based on image set," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2008.

[31] R. Wang and X. Chen, "Manifold dsicriminant analysis," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 429–436, 2009.

[32] H. Cevikalp and B. Triggs, "Face recognition based on image sets," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2567–2573, 2010.