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 possess 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.

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 robust to noise and occlusion [15].
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 $p$th 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 $D_k^{(p)}$ for $k = 1, 2, 3$, and $p = 1, 2, ..., P$, represents the $p$th video’s $k$th face feature that is under a particular pose and/or illumination condition.

In the testing stage, the same partition step is applied on the $m$th query video sequence to acquire partitions, $Q_k^{(m)}$, $k = 1, 2, 3$. Then, for each $Q_k^{(m)}$, sub-dictionaries from all target videos are found and concatenated to form the dictionary $D_k^{T}$. Using $D_k^{T}$, $k = 1, 2, 3$, and a query sample, joint sparse representation $\Gamma = [\Gamma^1 \; \Gamma^2 \; \Gamma^3]$, 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 $G_k^{(p)}$ be the augmented gallery matrix of the $k$th partition of the $p$th video sequence. In augmented gallery
matrix $G_{(p)}^k = [g_{(p),1}^k, g_{(p),2}^k, ...]$, each column is a vectorized form of the corresponding cropped grayscale face image of size $L$. Given $G_{(p)}^k$, a dictionary $D_{(p)}^k \in \mathbb{R}^{L \times K}$ is learned such that the columns of $G_{(p)}^k$, are best represented by linear combinations of $K$ atoms of $D_{(p)}^k$. This can be done by minimizing the following representation error
\[
(D_{(p)}^k, \Lambda_{(p)}^k) = \arg \min_{D_{(p)}^k, \Lambda_{(p)}^k} \|G_{(p)}^k - D_{(p)}^k \Lambda_{(p)}^k\|^2_F, \quad s.t. \quad \|\lambda_i\|_0 \leq T_0, \quad \forall i, \tag{1}
\]
where $\lambda_i$ is the $i$th column of coefficient matrix $\Lambda_{(p)}^k$ and $T_0$ is a sparsity parameter. The $\ell_0$ sparsity measure $\|\|_0$ counts the number of nonzero elements in the representation and $\|Y\|_F$ is the Frobenius norm of the matrix $Y$ defined as $\|Y\|_F = \sqrt{\sum_{i,j} Y_{i,j}^2}$.

One of the most well-known algorithms for learning such dictionaries is the K-SVD algorithm [21]. The K-SVD algorithm alternates between sparse-coding and dictionary update steps. In the sparse-coding step, the dictionary $D_{(p)}^k$ is fixed and the representation vectors $\lambda_i$ are found for each example $g_{(p),i}$. 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 $D_{(p)}^k$’s, the partition-specific dictionaries of the $p$th video.

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

Let $Q$ denote the total number of query video sequences. Given the $m$th query video sequence $Q_{(m)}$, where $m = 1,...,Q$, we can write $Q_{(m)} = \bigcup_{k=1}^K Q_{(m),k}^k$. Partitions $Q_{(m),k}^k$ are expressed by $Q_{(m),k}^k = [Q_{(m),k,1}^k, Q_{(m),k,2}^k, ..., Q_{(m),n(m),k}^k]$, where $Q_{(m),k,\ell}$ is the vectorized form of the $\ell$th of the total $n(m),k$ cropped face images belonging to the $k$th partition.

We exploit the joint sparsity of coefficients from different partitions to make a joint decision. We denote $(Q_{(m),k})_{k=1}^K$ as a set of $K$ partitions of the $m$th query video, with partition $Q_{(m),k}^k$ consisting of $n(m),k$ face images, and $D^k = \{D_{(1)}^k, D_{(2)}^k, ..., D_{(p)}^k\}$ as the concatenation of the $k$th sub-dictionaries from all target videos. Letting $\Gamma = [\Gamma^1, \Gamma^2, ..., \Gamma^K] \in \mathbb{R}^{d \times n(m)}$ be the matrix formed by concatenating the coefficient matrices with $d = \sum_{j=1}^d d_j$ and $n(m) = \sum_{k=1}^K n(m),k$, we seek the row-sparse matrix $\Gamma$ by solving the following $\ell_1/\ell_q$-regularized least square problem
\[
\Gamma = \arg \min_{\Gamma} \frac{1}{2} \sum_{k=1}^K \|Q_{(m),k}^k - D^k \Gamma^k\|^2_F + \lambda \|\Gamma\|_{1,q}, \tag{2}
\]
where $\lambda$ is a positive parameter and $q$ is set greater than 1 to make the optimization problem convex. Here, $\|\Gamma\|_{1,q}$ is a norm defined as $\|\Gamma\|_{1,q} = \sum_{\ell=1}^d \gamma^\ell \|\gamma^\ell\|_q$ where $\gamma^\ell$‘s are the row vectors of $\Gamma$. 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) Identification: For identification, we use the knowledge of the correspondence $f(\cdot)$ between subjects and sequences to assign the query video sequence $Q_{(m)}$ to subject $i^* = f(p^*)$, where $p^*$ is the sequence-level decision. Once $\hat{\Gamma}$ is obtained, $p^*$ is declared as the one that produces the smallest approximation error.
\[
p^* = \arg \min_p \min_{k \in \{1,...,K\}} \min_{\ell \in \{1,...,n(m),k\}} \|q_{(m),\ell}^k - D^k \delta_{p^*,\ell}(\hat{\Gamma})\|^2, \tag{3}
\]
where $\delta^p_{\ell}(\cdot)$ is the indicator function defined by keeping the coefficients corresponding to the $k$th partition from $p$th target video for the $\ell$th 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 $Q_{(m)}$ and the $p$th target video, is used to fill in the $(m,p)$ entry of the similarity matrix. Denoting the residual by $R_{(m,p)}^{(m,p)}$, we have
\[
R_{(m,p)}^{(m,p)} = \min_{k \in \{1,...,K\}} R_{k}^{(m,p)}, \tag{4}
\]
where
\[
R_{k}^{(m,p)} = \min_{\ell \in \{1,...,n(m),k\}} \|q_{(m),\ell}^k - D^k \delta_{p^*,\ell}(\hat{\Gamma})\|^2. \tag{5}
\]
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 $Q_{(m)}$ and the $p$th 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 $\{D^k\}_{k=1}^K$ and query videos $Q_{(m),k}^k_{k=1}^K$.

*Procedure:* Obtain $\Gamma$ by solving
\[
\hat{\Gamma} = \arg \min_{\Gamma} \frac{1}{2} \sum_{k=1}^K \|Q_{(m),k}^k - D^k \Gamma^k\|^2_F + \lambda \|\Gamma\|_{1,q},
\]
Output:

(Identification) $p^* = \arg \min_p \min_{k \in \{1,...,K\}} \min_{\ell \in \{1,...,n(m),k\}} \|q_{(m),\ell}^k - D^k \delta_{p^*,\ell}(\hat{\Gamma})\|^2$.

(Verification) Use the similarity $R_{(m,p)}^{(m,p)}$ 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 $D_k$ is a concatenation of sub-dictionaries that are aligned with $Q_{(m)}^k$. In other words, if $Q_{(m)}^k$ collects a subject’s left side face images from the $m$th video, then $D_k$ 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 $D_k$ such that $D_{(p)}^k, p = 1, 2, \ldots, P$ are obtained by:

$$D_{(p)}^k = \arg \max_{D_{(p)}^k, u \in \{1, \ldots, K\}} C_u,$$

where $C_u$ is the number of votes for the $u$th sub-dictionary of the $p$th target video (i.e., $D_{(p)}^k$ in (1)) collected from each $Q_{(m),l}^k$ in $Q_{(m)}^k$. In other words, the aligned sub-dictionaries are determined by the majority vote criterion. Each query image $Q_{(m),l}^k$ in the $k$th partition of the $m$th query video $Q_{(m)}^k$ votes for $D_{(p)}^k$ such that it has the minimum reconstruction error from its projection on $D_{(p)}^k$:

$$u = \arg \min_{u} \|q_{(m),l}^k - D_{(p)}^k D_{(p)}^{\dagger} q_{(m),l}^k\|^2,$$

where $D_{(p)}^{\dagger}$ is the pseudo-inverse of $D_{(p)}$.

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

$$\kappa(d_i, d_j) = \langle \phi(d_i), \phi(d_j) \rangle$$

where, $\phi$ is an implicit mapping projecting the vector $d$ into a higher dimensional space.

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

$$\Phi(Q_{(m),l}^k) = [\phi(q_{(m),l,1}^k), \phi(q_{(m),l,2}^k) \ldots \phi(q_{(m),l,n_{(m),k}}^k)]$$

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

$$\Phi(D_k^t) = [\phi(D_{k,1}^t), \phi(D_{k,2}^t), \ldots, \phi(D_{k,P}^t)]$$

As in joint linear space representation, we have:

$$\Phi(Q_{(m),l}^k) = \Phi(D_k^t)\Gamma_k$$

where, $\Gamma_k$ 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:

$$\hat{\Gamma} = \arg \min_{\Gamma} \frac{1}{2} \sum_{k=1}^K \|\Phi(Q_{(m),l}^k) - \Phi(D_k^t)\Gamma_k\|^2 + \lambda \|\Gamma\|_{1,q}$$

(8)

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

$$\hat{\Gamma} = \arg \min_{\Gamma} \frac{1}{2} \sum_{k=1}^K (\text{trace}(\Gamma_k^T K_{D_k^t} \hat{\Gamma} \Gamma_k) - 2\text{trace}(K_{D_k^t} Q_{(m),l}^k \hat{\Gamma})) + \lambda \|\Gamma\|_{1,q}$$

(9)

where, the kernel matrix $K_{X,Y}$ is defined as:

$$K_{X,Y}(i,j) = \kappa(x_i, y_j) = \langle \phi(x_i), \phi(y_j) \rangle,$$

(10)

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

1) Identification: Once $\hat{\Gamma}$ is obtained, we assign $Q_{(m),l}^k$ to subject $i^\star = f(p^\star)$, where $p^\star$ is obtained as follows:

$$p^\star = \arg \min_p \min_k \min_{\{q_{(m),l}^k\}} \|\phi(q_{(m),l}^k) - \Phi(D_k^t)\delta_{p,l}(\hat{\Gamma})\|^2$$

$$= \arg \min_p \min_k \min_{\{q_{(m),l}^k\}} \{\text{trace}(K_{Q_{(m),l}^k} Q_{(m),l}^k) - \text{trace}(\delta_{p,l}(\hat{\Gamma})^TK_{D_k^t} \delta_{p,l}(\hat{\Gamma})) + \text{trace}(\delta_{p,l}(\hat{\Gamma})^TK_{D_k^t} D_k^t \delta_{p,l}(\hat{\Gamma}))\}.$$ 

(11)

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

$$R_{(m,p)} = \min_{l \in \{1, \ldots, n_{(m),k}\}} \|\phi(q_{(m),l}^k) - \Phi(D_k^t)\delta_{p,l}(\hat{\Gamma})\|^2.$$ 

(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 $\{D_k\}_{k=1}^K$ and query videos $\{Q_{(m),l}\}_{k=1}^K$.

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

$$\hat{\Gamma} = \arg \min_{\Gamma} \frac{1}{2} \sum_{k=1}^K \|\Phi(Q_{(m),l}^k) - \Phi(D_k^t)\Gamma_k\|^2 + \lambda \|\Gamma\|_{1,q}$$

(13)

**Output:**

- **(Identification) video $p^\star = \arg \min_p \min_k \min_{\{q_{(m),l}^k\}} \|\phi(q_{(m),l}^k) - \Phi(D_k^t)\delta_{p,l}(\hat{\Gamma})\|^2,$** subject $i^\star = f(p^\star)$.

- **(Verification) Use the similarity $R_{(m,p)}$ 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 \times 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.

![Example frames from the UMD dataset. (a) Standing sequences. (b) Walking sequences. (c) Frames with blurred subjects due to the camera motion.](image)

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.

![Table II: Identification rates of “Standing vs. Walking” and “Walking vs. Standing” experiments on the UMD videos.](table)

![Table I: Identification rates of leave-one-out testing experiments on UMD videos. Both SRV and KSRV outperform the other compared methods.](table)

![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-based methods tied with each other and they are comparable with other methods.](image)

2For the illustration purpose only, here we show results of 8 subjects only.
based methods obtained better ROC curves than the WGCP method.

**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 \((720 \times 480\) pixels) and high definition (HD) format \((1440 \times 1080\) 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 \(S_2\) (144 subjects, 397 videos), \(S_3\) (55 subjects, 219 videos) and \(S_4\) (54 subjects, 216 videos). Table III lists the percentages of correct identifications for this experiment. Our sparsity-based methods outperform the other compared methods.

<table>
<thead>
<tr>
<th>MBGC walking</th>
<th>PM [12],[28]</th>
<th>KD [12],[28]</th>
<th>WGCP [12]</th>
<th>SANP [13]</th>
<th>SRV</th>
<th>KSRV</th>
</tr>
</thead>
<tbody>
<tr>
<td>(S_2)</td>
<td>43.79</td>
<td>39.74</td>
<td>63.79</td>
<td>83.88</td>
<td>86.65</td>
<td>86.65</td>
</tr>
<tr>
<td>(S_3)</td>
<td>53.88</td>
<td>50.22</td>
<td>74.88</td>
<td>84.02</td>
<td>87.67</td>
<td>88.58</td>
</tr>
<tr>
<td>(S_4)</td>
<td>53.70</td>
<td>50.46</td>
<td>75.00</td>
<td>84.26</td>
<td>87.96</td>
<td>88.89</td>
</tr>
<tr>
<td>Average</td>
<td>50.46</td>
<td>46.81</td>
<td>71.22</td>
<td>84.05</td>
<td>87.43</td>
<td>88.04</td>
</tr>
</tbody>
</table>

**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., \(S_2\)). We divide all these videos into SD and HD videos, to conduct “SD vs. HD”
(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.

\[
\begin{array}{ccccccc}
\text{MBGC walking} & \text{PM} & \text{KD} & \text{WGCP} & \text{SANP} & \text{SRV} & \text{KSRV} \\
\text{SD vs. HD} & 61.31 & 55.78 & 30.15 & 41.71 & 91.96 & 91.46 \\
\text{HD vs. SD} & 68.69 & 56.06 & 30.30 & 45.96 & 90.40 & 91.41 \\
\text{Average} & 65 & 55.92 & 30.25 & 43.84 & 91.18 & 91.44 \\
\end{array}
\]

TABLE IV: Identification rates of “SD vs. HD” and “HD vs. SD” experiments on the MBGC walking video subset \(S^2\) (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.

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.
and can perform significantly better than many state-of-the-
for the query video. Then joint sparse representation is
unconstrained video-based face recognition. In the training

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

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>50 frames</td>
<td>76.92</td>
<td>69.23</td>
<td>74.36</td>
<td>87.18</td>
</tr>
<tr>
<td>100 frames</td>
<td>84.62</td>
<td>87.18</td>
<td>94.87</td>
<td>84.62</td>
</tr>
<tr>
<td>full length</td>
<td>94.87</td>
<td>94.87</td>
<td>97.94</td>
<td>89.74</td>
</tr>
<tr>
<td>Average</td>
<td>85.47</td>
<td>83.76</td>
<td>88.89</td>
<td>87.18</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>50 frames</td>
<td>82.05</td>
<td>84.62</td>
<td>94.87</td>
<td>94.87</td>
</tr>
<tr>
<td>100 frames</td>
<td>84.62</td>
<td>92.31</td>
<td>97.44</td>
<td>97.44</td>
</tr>
<tr>
<td>full length</td>
<td>92.31</td>
<td>100</td>
<td>97.44</td>
<td>97.44</td>
</tr>
<tr>
<td>Average</td>
<td>86.33</td>
<td>92.31</td>
<td>96.58</td>
<td>96.58</td>
</tr>
</tbody>
</table>

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 recog-
nition: A literature survey,” ACM Computing Surveys, pp. 399–458,

R. Barwick, and J. Dunlop, “Recognizing people from dynamic and
static faces and bodies: Dissecting identity with a fusion approach,”


and recognition using probabilistic appearance manifolds,” Computer

the generic shape-illumination manifold,” European Conference on

models of geometry and illumination,” IEEE Transactions on Pattern
Analysis and Machine Intelligence, vol. 20, no. 10, pp. 1025–1039,

coding of face images using flexible models,” IEEE Transactions on
Pattern Analysis and Machine Intelligence, vol. 19, no. 7, pp. 743–756,
July 1997.

recognition using appearance-adaptive models in particle filters,” IEEE
Transactions on Image Processing, vol. 13, no. 11, pp. 1491–1506,

under varying illumination: an approach based on registration of
texture-mapped 3d models,” IEEE Transactions on Pattern Analysis

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 analysis,” IEEE Conference on Pattern Analysis and

for image set classification,” IEEE Conference on Computer Vision and
Pattern Recognition, pp. 72–80, 2011.

based face recognition from video,” European Conference on Compu-

recognition via sparse representation,” Pattern Analysis and Machine
2009.

iris recognition using random projections and sparse representations,”
Pattern Analysis and Machine Intelligence, IEEE Transactions on,

[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,
nos. 3, pp. 954–965, June 2012.

grouped variables,” Journal of the Royal Statistical Society: Series B,

regression,” Journal of the Royal Statistical Society: Series B, vol. 70,

diverse aspects of videos,” IEEE Transactions on Multimedia, vol. 12,

overcomplete dictionaries for sparse representation,” IEEE Transac-

[22] J. Yang and Y. Zhang, “Alternating direction algorithms for ll prob-
lems in compressive sensing,” SIAM Journal on Scientific Computing,

lagrangian approach to the constrained optimization formulation of
imaging inverse problems,” IEEE Transactions on Image Processing,

sparsity-based robust multimodal biometrics recognition,” European
Conference on Computer Vision Workshop on Information Fusion in
Computer Vision for Concept Recognition, 2012.

problems, prospects, and progress,” Pattern Recognition Letters,

O’Toole, D. Bolme, K. W. Bowyer, B. A. Draper, G. H. Givens, Y. M.
multiple biometrics grand challenge,” International Conference on

[27] National Institute of Standards and Technology, “Multiple
biometric grand challenge (MBGC).”

on stiefel and grassmann manifolds with applications in computer vi-
sion,” IEEE Conference on Computer Vision and Pattern Recogni-
tion, pp. 1–8, 2008.

and recognition of image set classes using canonical correlations,”
IEEE Transactions on Pattern Analysis and Machine Intelligence,

for face recognition based on image sets,” IEEE Conference on Compu-
ter Vision and Pattern Recognition, pp. 1–8, 2008.

[31] R. Wang and X. Chen, “Manifold discriminant analysis,” IEEE Confer-
ence on Computer Vision and Pattern Recognition, pp. 429–436,
2009.

IEEE Conference on Computer Vision and Pattern Recognition,