IEEE Access
The journal for rapid open access publishing

Received August 19, 2015, accepted September 14, 2015, date of current version October 14, 2015.

Digital Object Identifier 10.1109/ACCESS.2015.2485400

Dictionary-Based Face and Person Recognition From Unconstrained Video

YI-CHEN CHEN[1], (Student Member, IEEE), VISHAL M. PATEL[2], (Member, IEEE), P. JONATHON PHILLIPS[3], (Fellow, IEEE), AND RAMA CHELLAPPA[1], (Fellow, IEEE)

[1]Center for Automation Research, Department of Electrical and Computer Engineering, University of Maryland Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742, USA
[2]Department of Electrical and Computer Engineering, Rutgers University, Piscataway, NJ 08854, USA
[3]National Institute of Standards and Technology, Gaithersburg, MD 20899, USA

Corresponding author: V. M. Patel (vishal.m.patel@rutgers.edu)

This work was supported by the Cooperative Agreement through the National Institute of Standards and Technology under Grant 70NANB11H023. The work of V. M. Patel was supported by the Office of Naval Research under Grant N00014-12-1-0124. The work of P. J. Phillips was supported by the Federal Bureau of Investigation.

ABSTRACT To recognize people in unconstrained video, one has to explore the identity information in multiple frames and the accompanying dynamic signature. These identity cues include face, body, and motion. Our approach is based on video-dictionaries for face and body. Video-dictionaries are a generalization of sparse representation and dictionaries for still images. We design the video-dictionaries to implicitly encode temporal, pose, and illumination information. In addition, our video-dictionaries are learned for both face and body, which enables the algorithm to encode both identity cues. To increase the ability of our algorithm to learn nonlinearities, we further apply kernel methods for learning the dictionaries. We demonstrate our method on the Multiple Biometric Grand Challenge, Face and Ocular Challenge Series, Honda/UCSD, and UMD data sets that consist of unconstrained video sequences. Our experimental results on these four data sets compare favorably with those published in the literature. We show that fusing face and body identity cues can improve performance over face alone.

INDEX TERMS Video-based face recognition, person recognition, dictionary learning, kernel dictionary learning.

I. INTRODUCTION

Face recognition research has traditionally concentrated on recognition from still images [1]–[3]. Due to the widespread deployment of surveillance camera, face recognition from video has gained a lot of attention in recent years [4], [5]. In an unconstrained video, recognition purely from face ignores other useful information. In practice, one can enhance the recognition of people from video by fusing identity cues from the face and body and their motion [6].

There are a number of face recognition methods that rely on the temporal dynamics in face videos [4]. Temporal dynamics can be exploited to characterize how facial appearance and motions change together, represent idiosyncratic features of a person or improve recognition accuracy through a tracking scheme. While the advantage of using motion information in face videos has been widely recognized, computational models for video-based face recognition have only recently received attention [1], [4], [7]. In video-based face and person recognition, a key challenge is exploiting all available identity-cues in 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 challenges in designing an effective video-based face recognition algorithm.

Existing approaches to the video-based face recognition problem include multi-still face recognition [8], extracting joint appearance and behavioral features from video [9], or explicitly modeling the temporal correlations between faces in two videos [7]. It has been shown that for a generic video-face recognition algorithm, performance can be significantly improved by simultaneously performing recognition and tracking [5], [9].

There are a number of approaches for fusing face and gait for person recognition [10], [11]. However, there is a substantial difference between the composition of gait videos and unconstrained video. In the vast majority of gait videos, people are walking across the field of view and the complete body is visible [12]. In unconstrained videos, people can be performing any action and only a portion of a person's body could be visible.

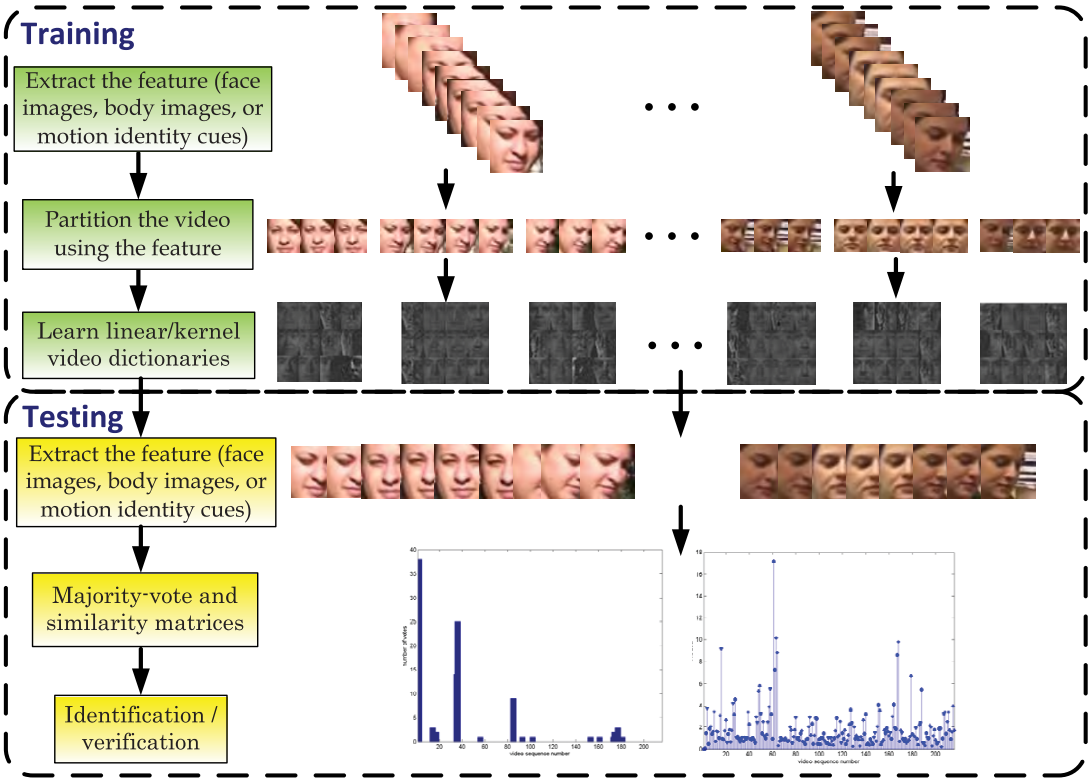

FIGURE 1. Overview of the proposed approach. The extracted feature can be face images, body images, or motion identity cues. For illustration purpose only, here we just show cropped face images as the feature.

To address the challenges in recognizing people from unconstrained video, we present a generative approach based on dictionary learning methods. Figure 1 shows an overview of our approach. While Figure 1 illustrates our approach for faces, we apply the same method for recognition using the body. From cropped images of faces or bodies extracted from a video sequence, we first partition the video sequence so that images with the same or close pose and illumination conditions are in one partition. This step removes the temporal redundancy while capturing variations due to changes in pose and illumination. For each partition, a sub-dictionary is learned where the representation error is minimized under a sparseness constraint. These partition-specific sub-dictionaries are combined to form a sequence-specific dictionary (i.e. a video dictionary). In the recognition phase, images of faces or bodies from a given query video sequence are projected onto the span of atoms in every sequence-specific dictionary. From the projection onto the atoms, the residuals are computed and combined to perform recognition or verification. The dictionary-based generative model is based on only the video sequence being processed. Thus, the method is scalable and incremental to increasing galleries. To handle the non-linearities present in the video data, we extend our original work in [13] by kernelizing the dictionary learning algorithm. In addition, we combine the face features with the upper body features to improve the recognition accuracy. We demonstrate the effectiveness of the proposed dictionary approach through comparisons with other recently proposed state-of-the-art methods, and with human performance on the Multiple Biometric Grand Challenge (MBGC) [14], [15], Face and Ocular Challenge Series (FOCS) [6], [16], Honda/UCSD [9], and UMD [17] datasets.

The key contributions of our work are[1]:

1. We introduce video-dictionaries for video-to-video face recognition. Through video partitioning, the learned dictionaries implicitly encode face pose and illumination information.
2. The dictionary learning algorithm is kernelized to handle non-linearities in the data samples.
3. The video dictionaries are further designed to encode the upper body features. The face features are combined with the upper body features to enhance the recognition accuracy.

The rest of the paper is organized as follows: In Section II we review some recent video-based face recognition methods. Section III describes the proposed dictionary-based video face recognition algorithm. Section IV describes the non-linear kernel dictionary learning. In section V, we present results on four challenging video datasets. Section VI concludes the paper with a summary and discussion.

[1]Preliminary version of this work appeared in [13]. Items **2**, and **3** are extensions to [13].

IEEE Access
The journal for rapid open access publishing

II. RELATED WORK

In this section, we review some of the recent video-based face recognition methods. In video-based face recognition, given a test video of a moving face, the first step is to track a set of facial features across all the frames of the video. Significant work has been done on face tracking using two-dimensional (2D) appearance-based models [18]–[20]. The 2D approaches, however, do not provide the three-dimensional (3D) configuration of the head, and are not robust to large changes in pose or viewpoint. To deal with this problem, several methods have been developed for 3D face tracking. Cascia *et al.* [21] proposed a cylindrical face model for face tracking. An extension of this work was proposed by Aggarwal *et al.* in [22] based on a particle filter for state estimation.

Temporal information in videos can be exploited for simultaneous tracking and recognition of faces without the need to perform these tasks in a sequential manner. One such method was proposed by Zhou *et al.* in [23]. Their tracking-and-recognition approach resolves uncertainties in tracking and recognition simultaneously in a unified probabilistic framework. Another method was proposed by Lee *et al.* [9], where a model of a subject is represented by a complex nonlinear appearance manifold. All frames in a video sequence are samples from an appearance manifold. To simplify the problem, the manifold is approximated by a collection of linear subspaces. Each subspace consists of nearby poses and is obtained by principle component analysis (PCA) of frames from training video sequences. If sufficient 3D view variations and illumination variations are available in the training set, this method is robust to large changes in appearance.

In a related work, Arandjelović [24] and Arandjelović and R. Cipolla [25] represent 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 learned using PCA from a small, unlabeled set of video sequences of faces acquired in randomly varying lighting conditions [5]. Given a novel sequence, the learned model is used to decompose the face appearance manifold into albedo and shape-illumination manifolds. Then a classification decision is made using robust likelihood estimation.

Sprechmann and Sapiro [26] proposed a framework for unsupervised clustering based on dictionary learning and sparse representation that can simultaneously learn a set of dictionaries. Each dictionary optimally represents the associated cluster in the sense that signals are best reconstructed in a sparse coding manner. As a result, they model the data as the of union of learned low dimensional subspaces. Unlike [26], our method is not specifically an unsupervised clustering method. It is mainly designed for video-based face recognition which uses clustering as one of its steps. Yang *et al.* [27] proposed to learn a set of metafaces from the training dataset and apply it to the sparse representation-based classification algorithm for face recognition. Another difference is that we propose a non-linear extension of our algorithm.

Recently, Turaga *et al.* [28] presented a statistical method for video-based face recognition, which uses subspace-based models and tools from Riemannian geometry of the Grassmann manifold. Intrinsic and extrinsic statistics are derived for designing maximum-likelihood classification rules. An image set classification methods for video-based face recognition problem was recently proposed by Hu *et al.* [29]. This method is based on a measure of between-set dissimilarity. This dissimilarity is the distance between sparse approximated nearest points of two image sets and is found by a scalable accelerated proximal gradient method for optimization. Other image set-based face recognition methods include [30]–[35]. Recently, a three stage video-based face recognition algorithm was proposed in [36] that computes a discriminative video signature as an ordered list of still face images from a large dictionary. Some of the other recent face recognition and related algorithms include [37]–[45]. See [36] for a survey of recent video-based face recognition algorithms.

III. DICTIONARY-BASED VIDEO ALGORITHM

In this section, we present the details of our dictionary-based video face and person recognition algorithm. The details of our approach are described for face video dictionaries. The approach is exactly the same for learning dictionaries for bodies. We first describe how the video sequence is partitioned into sub-sequences in section III-A, and how we build sequence-specific dictionaries in section III-B. Identification and verification are described in sections III-C and III-D, respectively.

A. VIDEO SEQUENCE PARTITION

For each frame in a video sequence, we first detect and crop the face and body regions automatically using the Viola-Jones object detection framework [46]. We then partition all the cropped face images into K different partitions. We partition the cropped faces by a clustering algorithm that is inspired by a video summarization algorithm [47]. Let $S = \{\mathbf{f}_1, \ldots, \mathbf{f}_n\}$ be the set of all n cropped faces from a video sequence. The following steps summarize our video sequence partition approach.

One major difference between our method and [47] is that the overall cost $J(S) \triangleq \alpha \times err(S) + (1-\alpha) \times (D - div(S))$ used in [47], is now replaced with

$$M(S) \triangleq \frac{div(S)}{err(S)}, \tag{1}$$

where $err(S)$, $div(S)$ and D are the square error, diversity and an upper bound of diversity of summary $S(\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_K)$, respectively [47], where $\mathbf{s}_i$'s are representatives. The terms $err(S)$ and $div(S)$ are *square error* and *diversity*,

Algorithm 1 Video Sequence Partition Algorithm

Initialization of sets:

$S = \{\mathbf{f}_1, \ldots, \mathbf{f}_n\}$, $I = \{1, 2, \ldots, n\}$, $T = \phi$.

Procedure:

1. Find $(i^*, j^*) = \underset{i,j \in I, i \neq j}{\text{argmax}} \|\mathbf{f}_i - \mathbf{f}_j\|_2$.
2. Update of sets: $t_1 \leftarrow i^*$, $t_2 \leftarrow j^*$, $T \leftarrow T \bigcup \{t_1, t_2\}$, $I \leftarrow I \setminus \{i^*, j^*\}$.
3. Find $k^* = \underset{k \in I}{\text{argmax}} \prod_{l=1}^{|T|} \|\mathbf{f}_{t_l} - \mathbf{f}_k\|_2$.
4. Update of sets: $t_{|T|+1} \leftarrow k^*$, $T \leftarrow T \bigcup \{t_{|T|+1}\}$, $I \leftarrow I \setminus \{k^*\}$.
5. Repeat steps 3 and 4 until $|T| = K$.
6. Given $\{\mathbf{f}_{t_1}, \ldots, \mathbf{f}_{t_K}\}$, use the nearest neighbor criterion to partition S into K partitions, denoted by $S(\mathbf{f}_{t_1}, \ldots, \mathbf{f}_{t_K}) = \bigcup_{i=1}^{K} S_i$. $S(\mathbf{f}_{t_1}, \ldots, \mathbf{f}_{t_K})$ are the initial partitions which are followed by N iterations of updating described in step 7 and 8.
7. Randomly select s_i from S_i, $i = 1, 2, \ldots, K$, as representatives. Find the corresponding nearest neighbor partitions which are denoted by $S(s_1, s_2, \ldots s_K)$, and calculate the corresponding score $M(S(s_1, s_2, \ldots s_K))$.
8. Repeat step 7, and keep updating for $\{\mathbf{s}_1^*, \mathbf{s}_2^*, \ldots, \mathbf{s}_K^*\}$ which gives the highest score M, until the number of repeating iterations for step 7 reaches N. In other words, $\{\mathbf{s}_1^*, \mathbf{s}_2^*, \ldots, \mathbf{s}_K^*\} = \underset{\mathbf{s}_i \in S_i, i=1,2,\ldots,K, \text{ in } N \text{ iterations}}{\text{argmax}} M(S(\mathbf{s}_1, \mathbf{s}_2, \ldots \mathbf{s}_K))$.

Output:

K partitions, $S(\mathbf{s}_1^*, \mathbf{s}_2^*, \ldots, \mathbf{s}_K^*)$.

respectively [47]. They are defined as follows

$$err(S) \triangleq tr\left[\sum_{i=1}^{K}\sum_{\mathbf{s}\in S_i}(\mathbf{s}-\mathbf{s}_i)(\mathbf{s}-\mathbf{s}_i)^T\right], \tag{2}$$

and

$$div(S) \triangleq tr\left[\sum_{i=1}^{K}(\mathbf{s}_i-\bar{\mathbf{s}})(\mathbf{s}_i-\bar{\mathbf{s}})^T\right], \tag{3}$$

where $\bar{\mathbf{s}} = \frac{1}{K}\sum_{i=1}^{K}\mathbf{s}_i$ and $tr(\mathbf{A})$ denotes the trace of matrix $\mathbf{A}$. The *diversity* represents the scatter of representatives to their mean, while the *square error* represents the total summation of partition-specific scatters, over all K partitions. The maximization of $M(S)$ is achieved through maximizing the *diversity* while minimizing the *square error*. Using this score, there is no need to set the weighting factor α [47], and the original cost minimization problem becomes an equivalent score maximization problem.

The proposed video partition technique is summarized in Algorithm 1. In steps 1 and 3, K initial representatives are chosen so that they are separated as far apart as possible. The corresponding initial K partitions are then determined by the nearest neighbor criterion. For all subsequent iterations steps (7 and 8), K distinct representatives are chosen always from the predetermined K initial partitions, and are used to calculate the associated score. The representatives that give the maximum $M(S)$ among, say N iterations, are recorded as exemplars. The corresponding final partitions are obtained by the nearest neighbor criterion.

B. BUILDING SEQUENCE-SPECIFIC DICTIONARIES

By partitioning the original video sequence, we obtain K separate sequences each containing images with specific pose and/or lighting conditions. We find the best representation for each member in a given partition by learning a partition specific dictionary. A dictionary is learned with the minimum representation error under a sparseness constraint. Thus, there will be K sub-dictionaries built to represent a 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 position shifted face images, or by in-plane rotated face images. We assume that each partition contains B images.

Let $\mathbf{G}_{j,k}^i$ be the augmented gallery matrix of the kth partition of the jth video sequence of subject i. In

$$\mathbf{G}_{j,k}^i = [\mathbf{g}_{j,k,1}^i \mathbf{g}_{j,k,2}^i, \cdots, \mathbf{g}_{j,k,B}^i] \in \boldsymbol{R}^{L\times B}, \tag{4}$$

each column is a vectorized form of the corresponding cropped grayscale face image of size L. Given $\mathbf{G}_{j,k}^i$, a dictionary $\mathbf{D}_{j,k}^i \in \mathbb{R}^{L\times K_0}$ is learned such that the columns of $\mathbf{G}_{j,k}^i$ are best represented by linear combinations of K_0 atoms of $\mathbf{D}_{j,k}^i$. This can be done by solving the following optimization problem

$$(\hat{\mathbf{D}}_{j,k}^i, \hat{\Gamma}_{j,k}^i) = \underset{\mathbf{D}_{j,k}^i, \Gamma_{j,k}^i}{\text{argmin}} \|\mathbf{G}_{j,k}^i - \mathbf{D}_{j,k}^i \Gamma_{j,k}^i\|_F^2, \quad \text{subject to } \|\boldsymbol{\gamma}_l\|_0 \leq T_0, \quad \forall l, \tag{5}$$

where $\boldsymbol{\gamma}_l$ is the lth column of the coefficient matrix $\Gamma_{j,k}^i$ and T_0 is a sparsity parameter. The ℓ_0 sparsity measure $\|\cdot\|_0$ counts the number of nonzero elements in the representation and $\|\mathbf{G}\|_F$ is the Frobenius norm of the matrix $\mathbf{G}$ defined as $\|\mathbf{G}\|_F = \sqrt{\sum_i \sum_j |\mathbf{G}(i,j)|^2}$. Many approaches have been proposed in the literature for solving such optimization problems. In this paper, we adapt the K-SVD algorithm [48] for solving (5) due to its simplicity and fast convergence.[2] The K-SVD algorithm alternates between sparse-coding and dictionary update steps. In the sparse-coding step, the dictionary $\mathbf{D}_{j,k}^i$ is fixed and the representation vectors $\boldsymbol{\gamma}_l$ are found for each example $\mathbf{g}_{j,k,l}^i$. Then, the dictionary is updated atom-by-atom in an efficient way [48].

The video sequence-specific dictionary is constructed by concatenating partition-level sub-dictionaries.

[2]Here "K" in "K-SVD" equals number of atoms K_0 in a learned dictionary, not the number of partitions K of a video sequence.

In other words, the jth dictionary of subject i is

$$\mathbf{D}_j^i = [\mathbf{D}_{j,1}^i \mathbf{D}_{j,2}^i \dots \mathbf{D}_{j,k}^i]. \tag{6}$$

C. IDENTIFICATION

Let Q denote the total number of query video sequences. Given the mth query video sequence $\mathbf{Q}^{(m)}$, where $m = 1, 2, \dots, Q$, we can write $\mathbf{Q}^{(m)} = \bigcup_{k=1}^{K} \mathbf{Q}_k^{(m)}$. Partitions $\mathbf{Q}_k^{(m)}$ are expressed by $\mathbf{Q}_k^{(m)} = [\mathbf{q}_{k,1}^{(m)}\ \mathbf{q}_{k,2}^{(m)}\ \cdots\ \mathbf{q}_{k,n_k}^{(m)}]$, where $\mathbf{q}_{k,l}^{(m)}$ is the vectorized form of the lth of the total n_k cropped face images belonging to the kth partition. Assume that there are a total of P gallery video sequences. We can write the associated dictionaries $\mathbf{D}_{(p)}$ for $p = 1, 2, \dots, P$, where each $\mathbf{D}_{(p)}$ corresponds to $\mathbf{D}_j^i$ for some subject i and its jth video sequence. Image $\mathbf{q}_{k,l}^{(m)}$ votes for sequence $\hat{p}$ with the minimum residual. In other words,

$$\hat{p} = \underset{p}{\operatorname{argmin}} \|\mathbf{q}_{k,l}^{(m)} - \mathbf{D}_{(p)}\mathbf{D}_{(p)}^{\dagger}\mathbf{q}_{k,l}^{(m)}\|_2, \tag{7}$$

where $\mathbf{D}_{(p)}^{\dagger} = (\mathbf{D}_{(p)}^T\mathbf{D}_{(p)})^{-1}\mathbf{D}_{(p)}^T$ is the pseudoinverse of $\mathbf{D}_{(p)}$ and $\mathbf{D}_{(p)}\mathbf{D}_{(p)}^{\dagger}\mathbf{q}_{k,l}^{(m)}$ is the projection of $\mathbf{q}_{k,l}^{(m)}$ onto the span of atoms in $\mathbf{D}_{(p)}$.

To make the sequence-level decision, we select p^* such that

$$p^* = \underset{p}{\operatorname{argmax}} \left(\sum_{k=1}^{K} C_{p,k} \right), \tag{8}$$

where $C_{p,k}$ is the total number of votes from partition k for sequence p. Finally, using the knowledge of the correspondence $\boldsymbol{m}(\cdot)$ between subjects and sequences, we assign the query video sequence $\mathbf{Q}^{(m)}$ to subject $i^* = \boldsymbol{m}(p^*)$.

D. 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 TARs all equal to 1 for any FARs. The ROC curves can be computed given a similarity matrix. In the proposed dictionary-based method, the residual between a query $\mathbf{Q}^{(m)}$ and a dictionary $\mathbf{D}_{(p)}$, is used to fill in the (m, p) entry of the similarity matrix. Denoting the residual by $\mathbf{R}^{(m,p)}$, we have

$$\mathbf{R}^{(m,p)} = \min_{k \in \{1,2,\dots,K\}} \mathbf{R}_k^{(m,p)}, \tag{9}$$

where

$$\mathbf{R}_k^{(m,p)} \triangleq \min_{l \in \{1,2,\dots,n_k\}} \|\mathbf{q}_{k,l}^{(m)} - \mathbf{D}_{(p)}\mathbf{D}_{(p)}^{\dagger}\mathbf{q}_{k,l}^{(m)}\|_2. \tag{10}$$

In other words, we select the minimum residual among all $l \in \{1, 2, \dots, n_k\}$, and all $k \in \{1, 2, \dots, K\}$, as the similarity between the query video sequence $\mathbf{Q}^{(m)}$ and dictionary $\mathbf{D}_{(p)}$. We denote the resulting dictionary-based face recognition algorithm as DFRV.

IV. NON-LINEAR KERNEL DICTIONARIES FOR VIDEO-BASED FACE RECOGNITION

The class identities in the face dataset may not be linearly separable. Hence, we also extend the DFRV framework to the kernel space. This essentially requires the dictionary learning model to be non-linear [49].

Let $\boldsymbol{\Phi} : \mathbb{R}^L \to \mathcal{H}$ be a non-linear mapping from L dimensional space into a dot product space $\mathcal{H}$. A non-linear dictionary can be trained in the feature space $\mathcal{H}$ by solving the following optimization problem

$$(\hat{\mathbf{A}}_{j,k}^i, \hat{\mathbf{\Gamma}}_{j,k}^i) = \arg \min_{\mathbf{A}_{j,k}^i, \mathbf{\Gamma}_{j,k}^i} \|\Phi(\mathbf{G}_{j,k}^i) - \boldsymbol{\Phi}(\mathbf{G}_{j,k}^i)\mathbf{A}_{j,k}^i\mathbf{\Gamma}_{j,k}^i\|_F^2, \quad \text{subject to } \|\boldsymbol{\gamma}_l\|_0 \le T_0, \quad \forall l, \tag{11}$$

where

$$\boldsymbol{\Phi}(\mathbf{G}_{j,k}^i) = [\boldsymbol{\Phi}(\mathbf{g}_{j,k,1}^i), \boldsymbol{\Phi}(\mathbf{g}_{j,k,2}^i), \cdots, \boldsymbol{\Phi}(\mathbf{g}_{j,k,B}^i)]. \tag{12}$$

Since the dictionary lies in the linear span of the samples $\Phi(\mathbf{G}_{j,k}^i)$, in (11) we have used the following model for the dictionary in the feature space,

$$\mathbf{D}_{j,k}^i = \boldsymbol{\Phi}(\mathbf{G}_{j,k}^i)\mathbf{A}_{j,k}^i, \tag{13}$$

where $\mathbf{A}_{j,k}^i \in \mathbb{R}^{B \times K_0}$ is a matrix with K_0 atoms [49]. This model provides adaptivity via modification of the matrix $\mathbf{A}_{j,k}^i$. Through some algebraic manipulations, the cost function in (11) can be rewritten as,

$$\|\boldsymbol{\Phi}(\mathbf{G}_{j,k}^i) - \boldsymbol{\Phi}(\mathbf{G}_{j,k}^i)\mathbf{A}_{j,k}^i\mathbf{\Gamma}_{j,k}^i\|_F^2 = \mathbf{tr}((\mathbf{I} - \mathbf{A}_{j,k}^i\mathbf{\Gamma}_{j,k}^i)^T\mathcal{K}(\mathbf{G}_{j,k}^i, \mathbf{G}_{j,k}^i)(\mathbf{I} - \mathbf{A}_{j,k}^i\mathbf{\Gamma}_{j,k}^i)), \tag{14}$$

where $\mathcal{K}(\mathbf{G}_{j,k}^i, \mathbf{G}_{j,k}^i)$ is a kernel matrix whose elements are computed from

$$\kappa(r, s) = \boldsymbol{\Phi}(\mathbf{g}_{j,k,r}^i)^T\boldsymbol{\Phi}(\mathbf{g}_{j,k,s}^i). \tag{15}$$

It is apparent that the objective function is feasible since it only involves a matrix of finite dimension $\mathcal{K} \in \mathbb{R}^{B \times B}$, instead of dealing with a possibly infinite dimensional dictionary.

An important property of this formulation is that the computation of $\mathcal{K}$ only requires dot products. Therefore, we are able to employ Mercer kernel functions to compute these dot products without carrying out the mapping $\boldsymbol{\Phi}$. Some commonly used kernels include polynomial kernels

$$\kappa(\mathbf{x}, \mathbf{y}) = (\langle \mathbf{x}, \mathbf{y} \rangle + c)^d \tag{16}$$

and Gaussian kernels

$$\kappa(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{\|\mathbf{x} - \mathbf{y}\|^2}{\sigma}\right), \tag{17}$$

where c, d and σ are the parameters.

Similar to the optimization of (5) using the linear K-SVD [48] algorithm, the optimization of (11) involves sparse coding and dictionary update steps in the feature space which results in the kernel K-SVD algorithm [49]. Details of the optimization can be found in [49].

A. FEATURE SPACE IDENTIFICATION

Let $\mathbf{A}_j^i = \text{diag}[\mathbf{A}_{j,1}^i, \mathbf{A}_{j,2}^i, \cdots, \mathbf{A}_{j,K}^i]$ denote the jth learned coefficient matrix of subject i. Assuming that there are a total of P gallery video sequences, we can write the associated coefficient matrices $\mathbf{A}_{(p)}$ for $p = 1, 2, \ldots, P$, where $\mathbf{A}_{(p)}$ equals $\mathbf{A}_j^i$ for some subject i and its jth video sequence. Accordingly, we use $\mathbf{G}_{(p)} \triangleq [\mathbf{g}_{(p),1} \cdots \mathbf{g}_{(p),K\times B}]$ to denote $\mathbf{G}_j^i = [\mathbf{G}_{j,1}^i, \mathbf{G}_{j,2}^i, \cdots, \mathbf{G}_{j,K}^i]$. We find the coefficient vectors, $\mathbf{x}_{k,l}^{(m)}$ with at most T_0 non-zero elements such that $\Phi(\mathbf{G}_{(p)})\mathbf{A}_{(p)}\mathbf{x}_{k,l}^{(m)}$ approximates $\mathbf{q}_{k,l}^{(m)}$ by minimizing the following problem

$$\min_{\mathbf{x}_{k,l}^{(m)}} \|\Phi(\mathbf{q}_{k,l}^{(m)}) - \Phi(\mathbf{G}_{(p)})\mathbf{A}_{(p)}\mathbf{x}_{k,l}^{(m)}\|_2 \quad \text{such that } \|\mathbf{x}_{k,l}^{(m)}\|_0 \le T_0. \tag{18}$$

The above problem can be solved by the Kernel Orthogonal Matching Pursuit (KOMP) algorithm [49].

Similar to (7), image $\mathbf{q}_{k,l}^{(m)}$ votes for sequence $\hat{p}$ such that

$$\begin{aligned}\hat{p} &= \underset{p}{\text{argmin}}\ r(\mathbf{q}_{k,l}^{(m)}, \mathbf{A}_{(p)}) \\ &= \underset{p}{\text{argmin}}\ \|\Phi(\mathbf{q}_{k,l}^{(m)}) - \Phi(\mathbf{G}_{(p)})\mathbf{A}_{(p)}\mathbf{x}_{k,l}^{(m)}\|_2^2 \\ &= \underset{p}{\text{argmin}}\ \mathcal{K}(\mathbf{q}_{k,l}^{(m)}, \mathbf{q}_{k,l}^{(m)}) - 2\mathcal{K}(\mathbf{q}_{k,l}^{(m)}, \mathbf{G}_{(p)})\mathbf{A}_{(p)}\mathbf{x}_{k,l}^{(m)} \\ &\quad + \mathbf{x}_{k,l}^{(m)T}\mathbf{A}_{(p)}^T\mathcal{K}(\mathbf{G}_{(p)}, \mathbf{G}_{(p)})\mathbf{A}_{(p)}\mathbf{x}_{k,l}^{(m)},\end{aligned} \tag{19}$$

where

$$\mathcal{K}(\mathbf{q}_{k,l}^{(m)}, \mathbf{G}_{(p)}) = [\kappa(\mathbf{q}_{k,l}^{(m)}, \mathbf{g}_{(p),1}), \kappa(\mathbf{q}_{k,l}^{(m)}, \mathbf{g}_{(p),2}), \cdots, \kappa(\mathbf{q}_{k,l}^{(m)}, \mathbf{g}_{(p),K\times B})]. \tag{20}$$

To make the sequence-level decision for identification, we select p^* by (8), with $C_{p,k}$ replaced by $\tilde{C}_{p,k}$, the total number of votes from the kth partition of the mth query video for the pth target video sequence according to (19).

B. FEATURE SPACE VERIFICATION

For verification using the kernel dictionaries, we construct the similarity matrix $\tilde{\mathbf{R}}^{(m,p)}$ by

$$\tilde{\mathbf{R}}^{(m,p)} = \min_{k\in\{1,2,\ldots,K\}} \tilde{\mathbf{R}}_k^{(m,p)}, \tag{21}$$

where $\tilde{\mathbf{R}}_k^{(m,p)}$ is the residual between $\mathbf{Q}_k^{(m)}$ and the kernel dictionary built from the pth target video sequence. It is computed by

$$\tilde{\mathbf{R}}_k^{(m,p)} = \min_{l\in\{1,2,\ldots,n_k\}} r(\mathbf{q}_{k,l}^{(m)}, \mathbf{A}_{(p)}). \tag{22}$$

We denote the resulting kernel DFRV algorithm as KDFRV. Both linear DFRV and non-linear KDFRV algorithms are summarized in Algorithm 2.

Algorithm 2 Video-Based Face Recognition (DFRV & KDFRV)

Training:

1. Given a sequence - the jth video of subject i, extract all the frames from it. Detect and crop face regions to form a set S_j^i.
2. Separate S_j^i into K partitions. Augment each partition by adding artificial images and obtain the resulting augmented gallery matrix from the kth partition, $\mathbf{G}_{j,k}^i, \forall k = 1, 2, \ldots, K$.
3. Use (5) for DFRV (and (11) for KDFRV) to learn the partition-specific sub-dictionary $\mathbf{D}_{j,k}^i, \forall k = 1, 2, \ldots, K$. Construct the sequence-specific dictionary $\mathbf{D}_j^i$ as in (6).

Testing:

1. Partition the mth query video sequence $\mathbf{Q}^{(m)} = \bigcup_{k=1}^K \mathbf{Q}_k^{(m)}$, where $\mathbf{Q}_k^{(m)} = [\mathbf{q}_{k,1}^{(m)}\ \mathbf{q}_{k,2}^{(m)}\ \cdots\ \mathbf{q}_{k,n_k}^{(m)}]$.
2. (Identification) Use (7) for DFRV (and (19) for KDFRV) to determine the vote from $\mathbf{q}_{k,l}^{(m)}, \forall k, l$. Then, use (8) and subject-sequence correspondence $\boldsymbol{m}(\cdot)$ to make the final decision.
3. (Verification) Find the similarity matrix between $\mathbf{Q}^{(m)}$ and $\mathbf{D}_{(p)}$ by (9) for DFRV (and (21) for KDFRV). The ROC curve can be obtained from the similarity matrix.

V. EXPERIMENTAL RESULTS

To illustrate the effectiveness of our method, we present experimental results on four publicly available datasets for video-based face recognition: the Multiple Biometric Grand Challenge (MBGC) [14], [15], the Face and Ocular Challenge Series (FOCS) [6], [16], the Honda/UCSD [9], and the UMD Comcast10 [17] datasets. For MBGC and FOCS videos, we use the upper body information in addition to faces for recognizing humans. All cropped face and upper body images were resized to $L = 20 \times 20$ pixels. Kernel parameters, the number of partitions per video K and the number of atoms per sub-dictionary K_0 are selected through 5-fold cross-validation. We summarize in Table 1, K and K_0 used in our experiments on the four datasets. The Gaussian kernel with parameter $\sigma = 32$ was used for kernel dictionaries. To train a video sequence-specific dictionary with three partitions per class, on average our method takes about 0.54 seconds on a desktop PC with processor Intel(R) Core(TM) i5-3470 CPU @ 3.20GHz and 8.00 GB RAM using Matlab.

TABLE 1. Summary of number of partitions per video (K) and number of atoms per sub-dictionary (K_0) in our experiments.

datasets	MBGC	FOCS	Honda/UCSD	UMD Comcast10
K	3	5	3	3
K_0	40	25	14	5

We compare the performance of our method with that of several state-of-the-art video-based face recognition

methods, including the Wrapped Gaussian Common Pole (WGCP) method [28], [50], and an image set-based Sparse Approximated Nearest Points (SANP) method [29]. When reporting the experimental results on face and upper body parts using the DFRV-based methods, we use the following naming convention:

- DFRV-f: DFRV on face images
- DFRV-b: DFRV on upper body images
- DFRV-bf: Score-level fusion of DFRV on both face and upper body images
- KDFRV-f: KDFRV on face images.

A. 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 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. The top row of Figure 2(a) shows 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 bottom row of Figure 2(a) shows example frames from four different activity sequences, where each subject reads from a paper, and the sequences consists of non-frontal views of the subject. There exist several challenging conditions including frontal and profile faces in shadow, and the profile faces sometimes being heavily covered by one's hair.

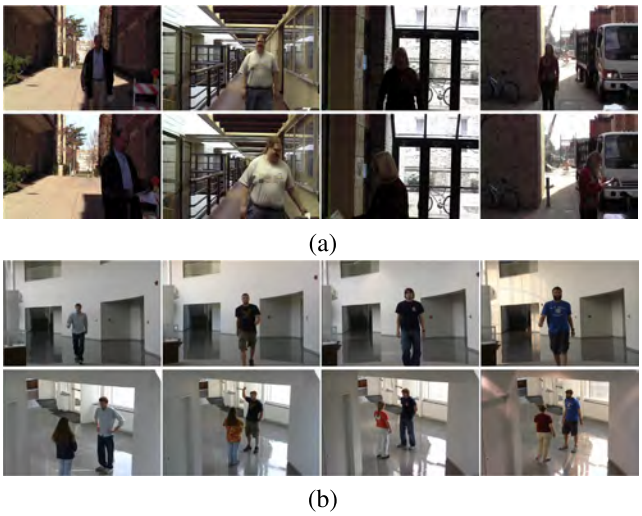

FIGURE 2. Examples of MBGC and UT-Dallas video sequences. (a) MBGC walking (top row) and activity (bottom row) sequences. (b) UT-Dallas walking (top row) and activity (bottom row) sequences.

Figure 3 shows an example of the output from the video partitioning stage. For results in Figure 3, the number of partitions is set equal to $K = 3$. Results are presented for 2 subjects for both walking and activity sequences.[3] For subject faces from walking videos shown in Figure 3(a), the corresponding cropped upper body images from activity videos are shown in Figure 3(b).[4] Each row shows up to 30 partitioned cropped face (or upper body) images from the same video sequence. The red lines distinguish between different subjects. It can be seen that each partition from a video sequence encodes a particular pose and/or illumination condition, and different partitions represent different conditions.

1) IDENTIFICATION RESULTS ON THE MBGC DATASET

Following the experiment design in [28], we conducted a leave-one-out identification experiment on 3 subsets of the cropped face and upper body images from walking videos. These 3 subsets are $S2$ (subjects which have at least two video sequences: 144 subjects, 397 videos), $S3$ (subjects which have at least three video sequences: 55 subjects, 219 videos) and $S4$ (subjects which have at least four video sequences: 54 subjects, 216 videos).

Table 2 lists the percentages of correct identifications for this experiment. The proposed DFRV-based methods (DFRV-f, KDFRV-f, DFRV-b and DFRV-bf) outperform the other state-of-the-art methods [28], [29] and [50]. For most subjects in this dataset, videos of the same subject wearing the same clothes and performed similar activities, were recorded in the same day. As different subjects possess different body appearance, compared to DFRV-f, the use of body information in DFRV-b and DFRV-bf enhance the discriminative identification rate. Comparing DFRV-f and KDFRV-f, we observe that kernel dictionaries obtained higher average identification rate on this dataset. This may be the case due to the fact that kernel dictionaries are able to capture the non-linearities in data. Hence, with the proper choice of kernel and parameters, the performance obtained using the kernel dictionaries is in general better than that given by the linear dictionaries.

We further compared our method on face images with a baseline method where the dictionary learning stage in our DFRV method is omitted and the cropped images in each partition are directly used as dictionaries. This method is denoted as "no DL". As shown in Table 2, omitting the dictionary learning stage results in the poor performance compared to the DFRV-f method. This baseline, however, remains better than SANP [29] as it keeps video partitioning that accounts for the pose and illumination variations.

In the second set of experiments, we selected videos associated for those 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" (SD as probe; HD as gallery) and "HD vs SD" (HD as probe; SD as gallery) experiments. Correct identification rates are shown in Table 3.

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

[4] As lower body parts are not available for some videos, in our work only face and upper body images were used for recognition.

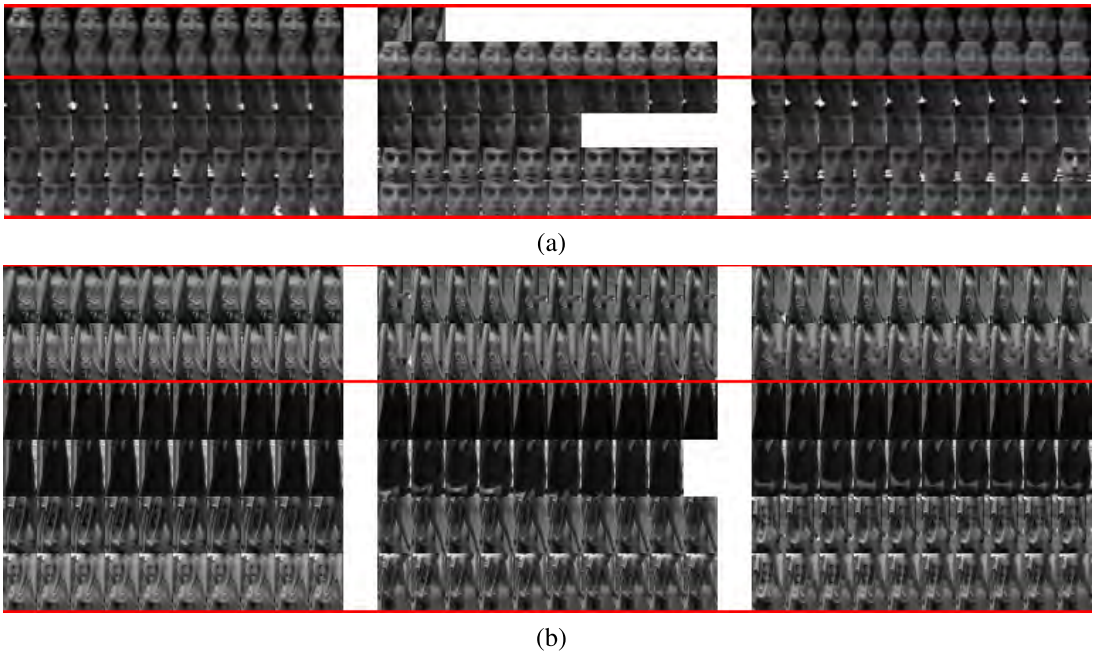

FIGURE 3. Partition results of example face and upper body images from MBGC videos: (a) Face images from walking videos. (b) The corresponding upper body images from activity videos. Red lines separate different subjects. A subject has at least two video sequences. Face (or upper body) images from a video sequence are shown in a row, and are further divided into three partitions. Each partition shows up to 10 face (or upper body) images. A partition represents a particular pose and illumination condition.

TABLE 2. Identification rates (%) of leave-one-out testing experiments on the MBGC walking videos. The proposed DFRV-based methods outperform statistical methods and the SANP method, recently proposed in [28] and [29], respectively.

MBGC walking videos	Procrustes Metric [29], [51]	Kernel Density [29], [51]	WGCP [29]	SANP [30]	Baseline (no DL)	DFRV-f	KDFRV-f	DFRV-b	DFRV-bf
$S2$	43.79	39.74	63.79	83.88	78.09	85.64	84.89	94.71	**95.97**
$S3$	53.88	50.22	74.88	84.02	77.63	88.13	89.50	94.98	**95.89**
$S4$	53.70	50.46	75	84.26	77.78	88.43	89.81	95.37	**96.30**
Average	50.46	46.81	71.22	84.05	77.83	87.40	88.07	95.02	**96.05**

TABLE 3. 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. The DFRV-bf method achieves the best identification rates.

MBGC walking videos	Procrustes Metric [29], [51]	Kernel Density [29], [51]	WGCP [29]	SANP [30]	Baseline (no DL)	DFRV-f	KDFRV-f	DFRV-b	DFRV-bf
SD vs HD	61.31	55.78	30.15	41.71	77.39	86.93	89.45	95.48	**96.48**
HD vs SD	68.69	56.06	30.30	45.96	85.35	91.41	89.90	**95.96**	**95.96**
Average	65	55.92	30.23	43.84	81.37	89.17	89.68	95.72	**96.22**

The DFRV-based methods outperformed the other methods significantly. The WGCP [28] 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. On the other hand, the SANP [29] 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. The proposed DFRV-based methods overcome this limitation by video partitioning and effectively combining different partition-level sub-dictionaries.

2) VERIFICATION RESULTS ON THE MBGC DATASET

Figure 4(a) and (b) show the corresponding ROC curves for "SD vs HD" and "HD vs SD" verification

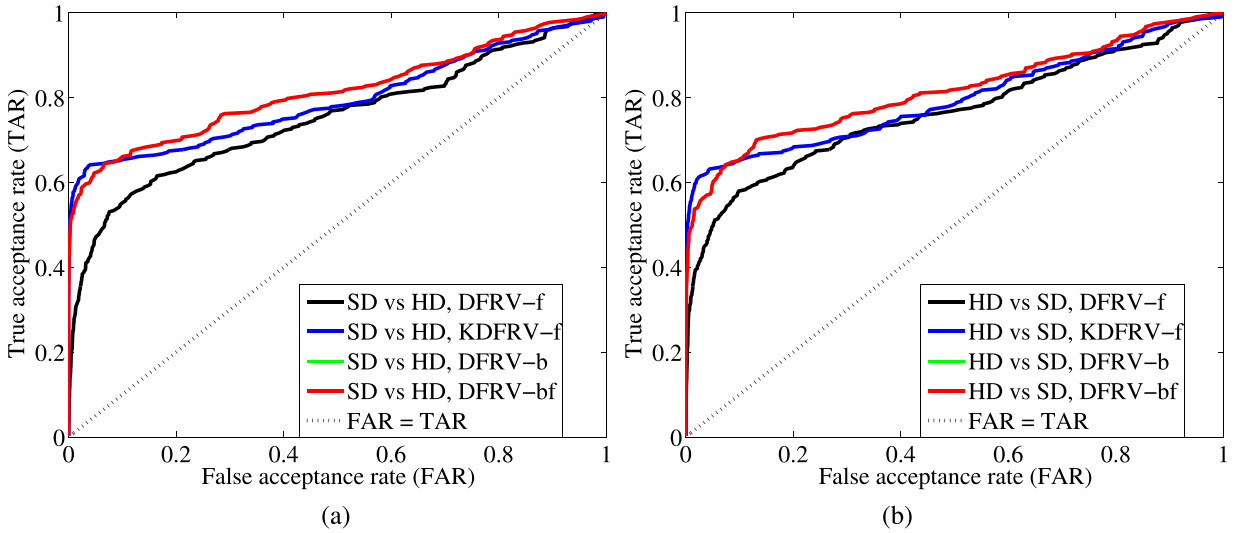

FIGURE 4. ROC curves of DFRV-based methods on the MBGC walking videos: (a) SD vs HD. (b) HD vs SD. There is no difference between DFRV-b and DFRV-bf curves. Both DFRV-b and DFRV-bf obtained better verification performances than DFRV-f.

experiments, respectively. As shown in both figures, one could hardly see the difference between DFRV-b (body only, in color green) and DFRV-bf (body and face, in color red) curves as the body feature dominates the overall performance. In addition, both DFRV-b and DFRV-bf obtained better verification performances than the DFRV-f method. For both identification and verification, the HD test samples had better performances than the SD test samples.

We further examine the effect on the performance of varying the number of video sequences per person in the gallery. We divide the videos into two groups beforehand either as probe, or as gallery. For the 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). Results presented above show that the WGCP method in this setting does not perform so well. We observe that the WGCP method is able to give satisfactory performance only when there are enough video sequences for training, which allows one to obtain more discriminative metrics for different subjects.

In the MBGC [14] protocol, verifications are specified by two sets: target and query. The protocol requires the algorithm to match each target sequence with all query sequences. We performed three verification experiments: walking vs walking (WW), activity vs walking (AW), activity vs activity (AA). Figure 5(a) shows the ROC curves. We observe that DFRV-f gives better ROC curve than WGCP for almost all FARs, in WW experiments. In AW and AA experiments; however, all curves are pretty close to random performance. These two experiments are very challenging. According to the MBGC website [15], for the AW and AA experiments, no results have been reported that are better than random.

Figures 5(b)(c)(d) show the comparisons between DFRV-f and DFRV-bf in WW, AW and AA experiments, respectively. As the MBGC verification protocol is designed to exclude matching videos of the same subject recorded in the same day, the body feature no longer contributes as much as it does in the identification experiments. Therefore, the gain obtained from the DFRV-bf is limited. A slightly larger improvement of DFRV-bf over DFRV-f can be observed in AA experiments (Figure 5(d)) only.

TABLE 4. Score level fusion summary of MBGC version 1 (Notre Dame) experiments.

MBGC v1 experiments	score-level fusion
$S2$, $S3$, $S4$ identification	$0.55\mathbf{C}_b+0.45\mathbf{C}_f$
HD(SD) vs SD(HD) identification	$0.5\mathbf{C}_b+0.5\mathbf{C}_f$
HD(SD) vs SD(HD) verification	$0.95\left(\frac{\mathbf{R}_b-\text{MED}(\mathbf{R}_b)}{\text{MAD}(\mathbf{R}_b)}\right)+0.05\left(\frac{\mathbf{R}_f-\text{MED}(\mathbf{R}_f)}{\text{MAD}(\mathbf{R}_f)}\right)$
'walking vs walking' verification	$0.95\left(\frac{\mathbf{R}_b-\text{MED}(\mathbf{R}_b)}{\text{MAD}(\mathbf{R}_b)}\right)+0.05\left(\frac{\mathbf{R}_f-\text{MED}(\mathbf{R}_f)}{\text{MAD}(\mathbf{R}_f)}\right)$
'walking vs activity' verification	$0.95\left(\frac{\mathbf{R}_b-\text{MED}(\mathbf{R}_b)}{\text{MAD}(\mathbf{R}_b)}\right)+0.05\left(\frac{\mathbf{R}_f-\text{MED}(\mathbf{R}_f)}{\text{MAD}(\mathbf{R}_f)}\right)$
'activity vs activity' verification	$0.95\left(\frac{\mathbf{R}_b-\text{MED}(\mathbf{R}_b)}{\text{MAD}(\mathbf{R}_b)}\right)+0.05\left(\frac{\mathbf{R}_f-\text{MED}(\mathbf{R}_f)}{\text{MAD}(\mathbf{R}_f)}\right)$

We regard face and body as distinct biometric modalities. Table 4 summarizes results of the score level fusion of face and body similarity scores. The vote and distance scores for faces are denoted by $\mathbf{C}_f$ and $\mathbf{R}_f$, respectively; the vote and distance scores for upper bodies are denoted by $\mathbf{C}_b$ and $\mathbf{R}_b$, respectively. We linearly combine the face and body similarity scores after normalization. We experimented with median and median absolute deviation (MAD) normalization. Normalizing by median and MAD is robust to outliers [51].

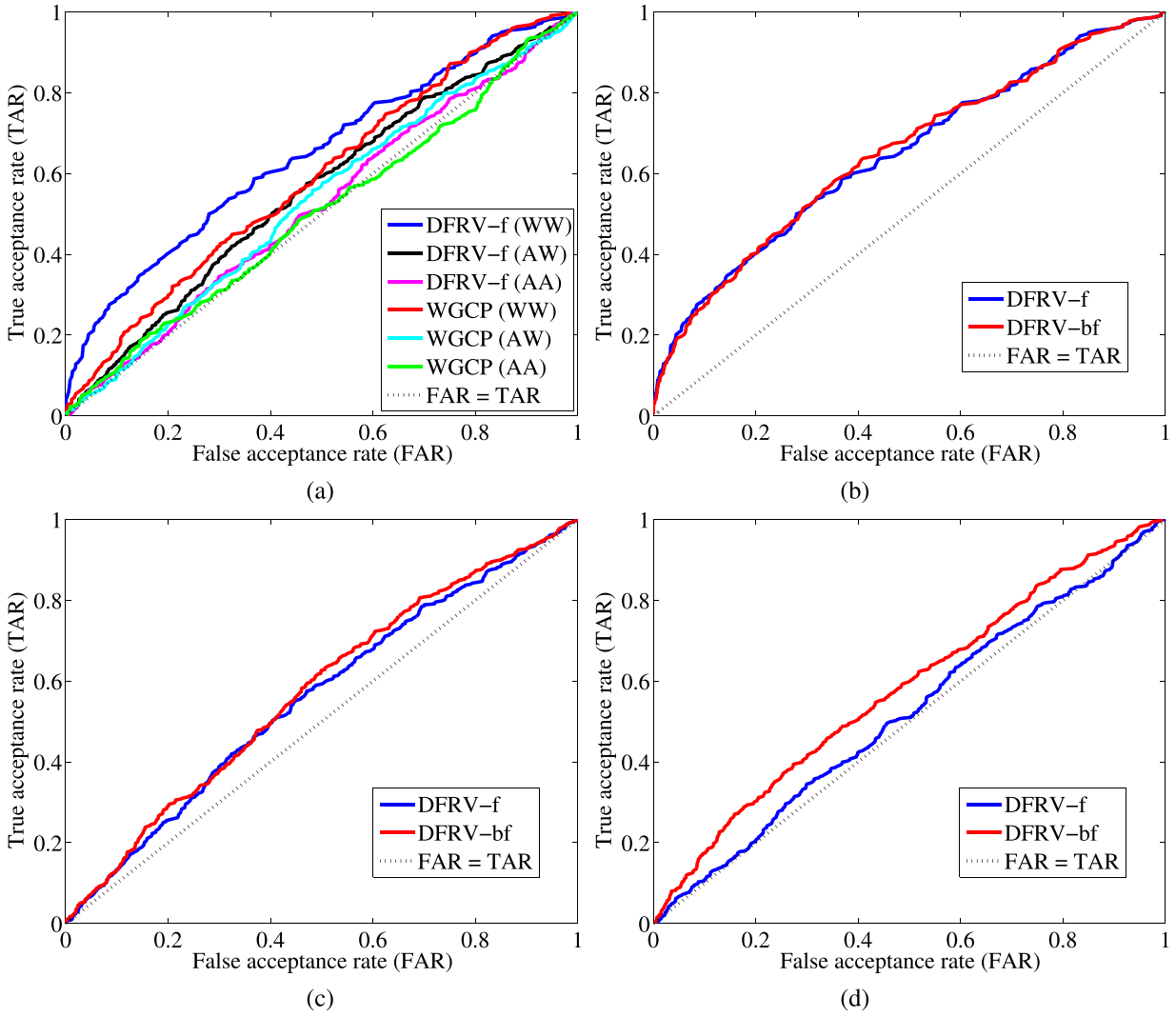

FIGURE 5. ROC curves of the MBGC experiments on walking and activity videos: (a) Comparing DFRV-f with WGCP in WW, AW and AA experiments. The proposed DFRV-f method gives better ROC curves than WGCP in WW experiments. Both curves are close to the random guess in the challenging AW and AA experiments. (b) Comparing DFRV-f and DFRV-bf in WW experiments. (c) Comparing DFRV-f and DFRV-bf in AW experiments. (d) Comparing DFRV-f and DFRV-bf in AA experiments, where a better improvement of DFRV-bf over DFRV-f is obtained.

B. FOCS UT-DALLAS VIDEO

The video challenge of Face and Ocular Challenge Series (FOCS) [6] is designed to match "frontal vs frontal", "frontal vs non-frontal", and "non-frontal vs non-frontal" video sequences. In this section we present our experimental results on the UT Dallas video sequences contained in the FOCS video challenge. The performance of the DFRV-f algorithm on the UT Dallas dataset shows the strength of our approach on a difficult data set. In addition, it allows us to directly compare the performance of the DFRV-f algorithm to humans [6].

The FOCS UT Dallas dataset contains 510 walking (frontal face) and 506 activity (non-frontal face) video sequences recorded from 295 subjects with frame size 720×480 pixels. The top row of Figure 2(b) shows key frames from four different walking sequences of one subject. The sequences were acquired on different days. In the walking sequences, the subject is originally positioned far away from the video camera, walks towards it with a frontal pose, and finally turns away from the video camera with profile face. The bottom row of figure 2(b) shows key frames of four different activity sequences of the same subject. In these sequences, the subject stands and talks with another person with a non-frontal face view to the video camera. The sequences contain normal head motions that occur during a conversation; e.g., the head turning up to 90 degrees, hand raising and/or pointing somewhere.

1) IDENTIFICATION RESULTS ON THE FOCS DATASET

We conducted the same leave-one-out tests on 3 subsets: $S2$ (189 subjects, 404 videos), $S3$ (19 subjects, 64 videos), and $S4$ (6 subjects, 25 videos) from the UT-Dallas walking videos. For body images, in order to capture both shape and temporal information in a low resolution scenario, we took

FIGURE 6. Sequential upper body differences in grayscale: the grayscale differences between a reference upper body frame and its subsequent frames in a cycle period of $L = 18$ frames. For each subject, the corresponding upper body differences computed from a reference frame are shown in a row as a motion cue of that reference frame. Here there are three rows shown for three different subjects. This feature captures both the shape and its temporal movement information, while not requiring either silhouette extraction or background subtraction.

TABLE 5. Identification rates (%) of leave-one-out testing experiments on the FOCS UT-Dallas walking videos. The DFRV-bf method performs the best.

UT-Dallas walking videos	Procrustes Metric [29], [51]	Kernel Density [29], [51]	WGCP [29]	SANP [30]	Baseline (no DL)	DFRV-f	KDFRV-f	DFRV-b	DFRV-bf
$S2$	38.12	40.84	53.22	48.27	45.05	59.90	46.53	20.30	**61.14**
$S3$	60.94	64.06	70.31	60.94	67.19	78.13	71.88	42.19	**79.69**
$S4$	64	64	76.00	68.00	76.00	80.00	76.00	60.00	**84.00**
Average	54.35	54.97	66.51	59.07	62.75	72.68	64.80	40.83	**74.94**

the grayscale differences between a reference upper body frame and all of its subsequent frames in a cycle period (L subsequent frames). Then we resized the resulting concatenated sequential differences as a motion cue of that reference frame. Figure 6 shows for the three example subjects their sequential upper body differences (in grayscale) over $L = 18$ frames, where each row captures a subject's upper body shape and information on its temporal movements. This method does not require silhouette extraction or background subtraction. Table 5 shows the identification results. Among all the compared methods, the DFRV-bf method achieved the best identification rates. Among methods other than DFRV-bf and DFRV-b (i.e., methods using face only), the KDFRV-f method, however, did not obtain better identification performance than DFRV-f and WGCP. Perhaps the Gaussian kernel with $\sigma = 32$ is not the best kernel for this dataset. Multiple Kernel Learning (MKL) methods can be adapted to optimally learn the kernel weights [52], [53]. However, this tremendously increases the complexity of the learning algorithm. The optimization of the choice of kernel and its parameters is one of our future research directions.

2) VERIFICATION RESULTS ON THE FOCS DATASET

Like MBGC, FOCS specifies a verification protocol: **1A** (walking vs walking), **2A** (activity vs walking), and **3A** (activity vs activity). In these experiments, 481 walking videos and 477 activity videos are chosen as query videos. The size of target sets ranges from 109 to 135 video sequences. Figure 7 shows ROC curves of verification experiments. In Figure 7(a), we compare the proposed algorithm with WGCP [28]. In all three experiments, the DFRV-f algorithm is superior to the WGCP algorithm.

O'Toole *et al.* [6] evaluated the accuracy of humans recognizing people in the UT Dallas data set. Human performance was reported for both static and dynamic presentations of faces and bodies. This included humans viewing the original sequence and for sequences edited to contain only the head. Since the DFRV-f algorithm only encodes face information, it is reasonable to compare the DFRV-f with human performance on the original sequences and the edited face only sequences. In Figure 7(b)(c)(d) we compare the performance of the DFRV-f algorithm and humans for experiments **1A**, **2A**, and **3A**. In Figures 7(b) and (d), we observe that the performance of the DFRV-f algorithm is very close to humans on the face only matching task. Experiments **1A** and **3A** are within pose matching tasks; whereas, **2A** is cross pose. Reported performance is better than random; however, not near human level of performance.

In Figures 8(a)(b)(c), we compare DFRV-f and DFRV-bf in 1A, 2A and 3A experiments, respectively. As shown, there is not much difference between the two methods. In fact, unlike MBGC, a subject with different cloth and facial appearances (as shown in Figure 2(b)) was recorded in different days. The body feature becomes much less discriminative and DFRV-b no longer gives satisfactory identification results. Therefore, for this challenging dataset, as the face feature dominates the performance, both DFRV-f and DFRV-bf obtained similar identification and verification results. The score level fusion between face and body for DFRV-bf is summarized in Table 6, where scores of the face feature weigh more as the face features are more discriminative on this dataset.

C. HONDA/UCSD DATASET

The third set of experiments is conducted on the Honda/UCSD Dataset [9]. The Honda Dataset consists of 59 video sequences from 20 distinct subjects. We follow the same procedure used in [29]. The experiments are done in three cases of the maximum set length (available number of cropped-face images per video sequence) as defined in [29]: 50, 100 and full length frames. Table 7 shows identification rates of our methods and other state-of-the-art methods. Both DFRV-f and KDFRV-f obtained the highest average

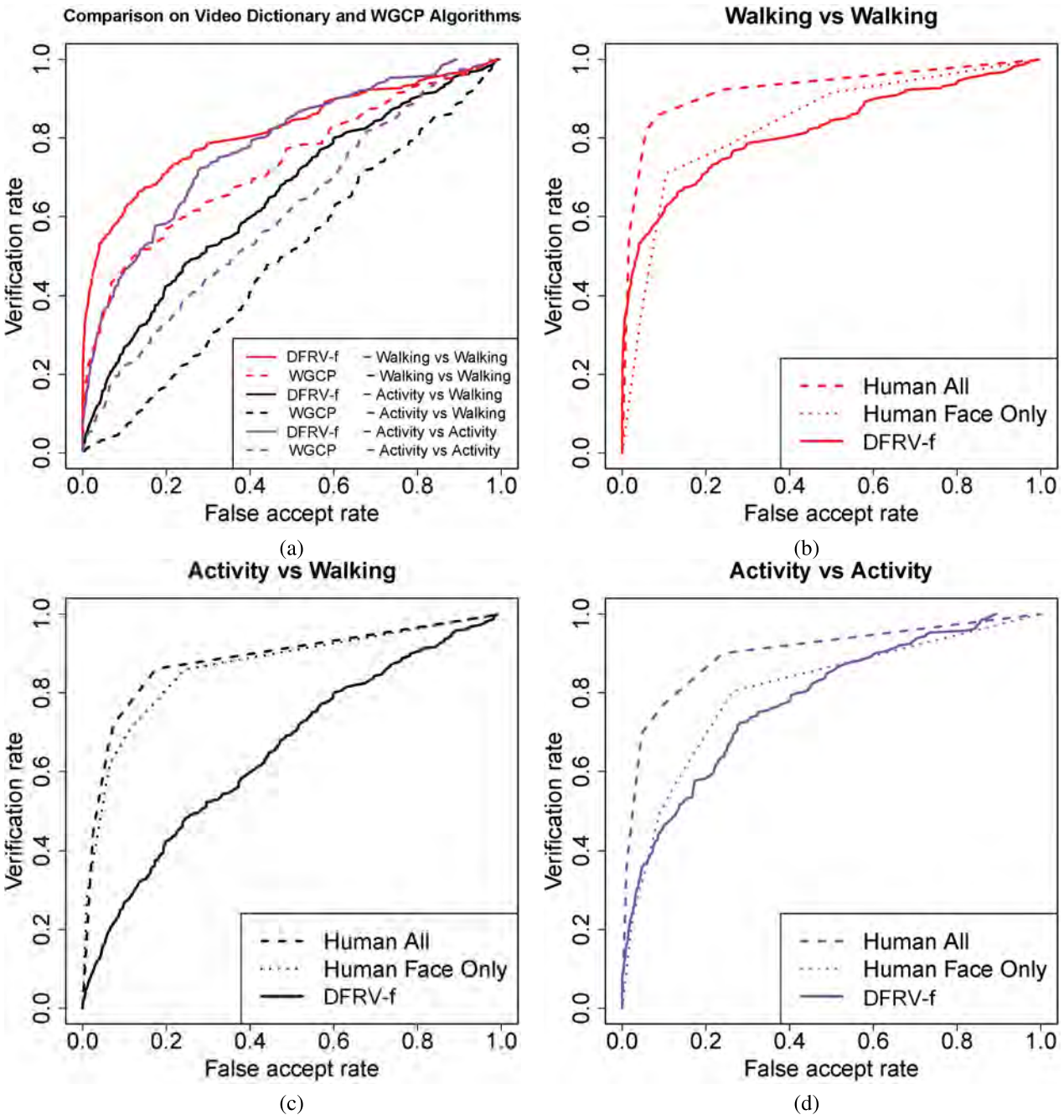

FIGURE 7. ROC curves of FOCS experiments on UT-Dallas videos: (a) comparison between DFRV-f and WGCP [28]; (b)(c)(d) comparison between DFRV-f and human perception [6]: (b) walking vs walking (c) activity vs walking (d) activity vs activity. Compared to WGCP, our DFRV-f method gives better ROC curves, which also stay very close to those of face-only human perception in (b)(d) cases.

identification rates. They ranked the second and tied with the MDA method [54] for the full length case.

D. UMD COMCAST10 DATASET

The UMD Comcast10 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(s) 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 9(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 9(b) shows example frames from four different walking sequences, in each of

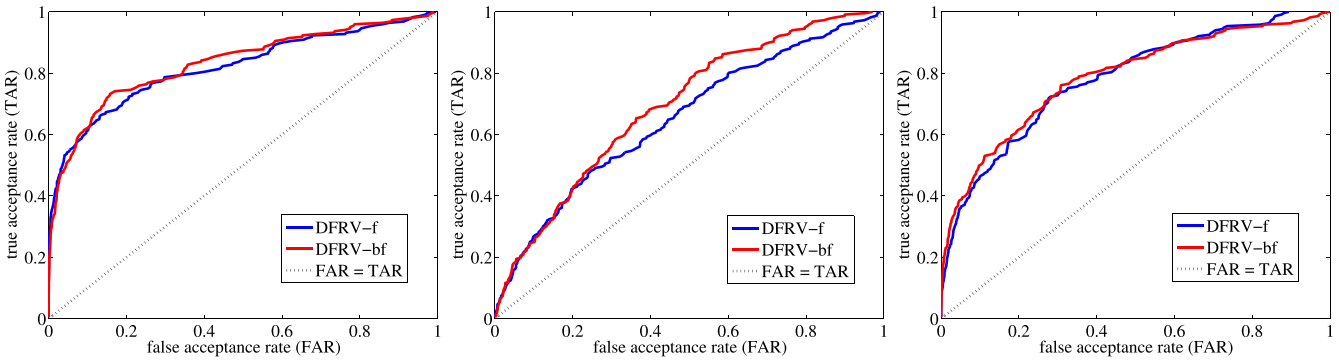

FIGURE 8. ROC curves of DFRV-f and DFRV-bf on the UT-Dallas videos. (a) walking vs walking. (b) activity vs walking. (c) activity vs activity. DFRV-bf obtained higher detection rates than DFRV-f (for FARs > 0.3) in the activity vs activity experiment.

TABLE 6. Score level fusion summary of FOCS (UT-Dallas) experiments.

FOCS (UT-Dallas) experiments	score-level fusion
$S2$, $S3$, $S4$ identification	$0.4\mathbf{C}_b + 0.6\mathbf{C}_f$
'walking vs walking' verification	$0.15\left(\frac{\mathbf{R}_b - \text{MED}(\mathbf{R}_b)}{\text{MAD}(\mathbf{R}_b)}\right) + 0.85\left(\frac{\mathbf{R}_f - \text{MED}(\mathbf{R}_f)}{\text{MAD}(\mathbf{R}_f)}\right)$
'activity vs walking' verification	$0.15\left(\frac{\mathbf{R}_b - \text{MED}(\mathbf{R}_b)}{\text{MAD}(\mathbf{R}_b)}\right) + 0.85\left(\frac{\mathbf{R}_f - \text{MED}(\mathbf{R}_f)}{\text{MAD}(\mathbf{R}_f)}\right)$
'activity vs activity' verification	$0.15\left(\frac{\mathbf{R}_b - \text{MED}(\mathbf{R}_b)}{\text{MAD}(\mathbf{R}_b)}\right) + 0.85\left(\frac{\mathbf{R}_f - \text{MED}(\mathbf{R}_f)}{\text{MAD}(\mathbf{R}_f)}\right)$

FIGURE 9. Example frames of UMD Comcast10 videos. (a) standing sequences. (b) walking sequences. (c) Frames with blurred subjects due to the moving camera. Faces in standing sequences were sometimes non-frontal or partially occluded, while faces in walking sequences were frontal for most of the time. Camera's movement raises the difficulty of face tracking and recognition.

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. In fact quite often it switched back and forth, which created more challenging conditions in these unconstraint video sequences. Figure 9(c) shows example frames with blurred subjects due to the movement of the camera.

Following the experiment design in [28], we conducted a leave-one-out identification experiment on 3 subsets of the cropped face images from walking videos performed. 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 four sequences: 7 subjects, 51 sequences). Note that for these particular segmented sequences, the three sets $S3$, $S4$ and $S5$ are identical. Table 8 lists the percentages of correct identifications for this experiment. Both KDFRV-f and DFRV-f outperformed the other compared methods. In particular, KDFRV-f achieved 100% identification rates on $S3 \sim S6$ video subsets.

Figure 10(a) shows the verification performances in $S2$, $S3$ and $S6$ experiments through ROC curves. From this figure, ROC curves of $S2$, $S3$ and $S6$ under either the DFRV-f method or the WGCP method, are indistinguishable. Both DFRV-f and KDFRV-f give better ROC curves than the WGCP method. Figure 10(b) shows the ROC curves for "Standing vs Walking" (standing sequences as probe;

TABLE 7. Identification rates (%) on Honda/UCSD Dataset. Both DFRV-f and KDFRV-f obtained the highest average identification rates.

Set length	DCC [33]	MMD [31]	MDA [55]	AHISD [34]	CHISD [34]	SANP [30]	Baseline (no DL)	DFRV-f	KDFRV-f
50 frames	76.92	69.23	74.36	87.18	82.05	84.62	87.18	89.74	**92.31**
100 frames	84.62	87.18	94.87	84.62	84.62	92.31	87.18	**97.44**	94.87
full length	94.87	94.87	97.44	89.74	92.31	**100**	92.31	97.44	97.44
Average	85.47	83.76	88.89	87.18	86.33	92.31	88.89	**94.87**	**94.87**

TABLE 8. Identification rates (%) of leave-one-out testing experiments on the UMD Comcast10 dataset. The KDFRV-f method outperforms DFRV-f and other compared methods.

UMD Comcast10 videos	Procrustes Metric [29], [51]	Kernel Density [29], [51]	WGCP [29]	SANP [30]	Baseline (no DL)	DFRV-f	KDFRV-f
$S2$	82.80	81.72	82.97	92.47	91.40	94.62	**95.70**
$S3$, $S4$, $S5$	84.62	83.52	83.52	93.41	92.31	96.70	**100**
$S6$	98.04	96.08	88.23	98.04	92.31	96.70	**100**
Average	88.49	87.11	84.91	94.64	92.01	96.00	**98.57**

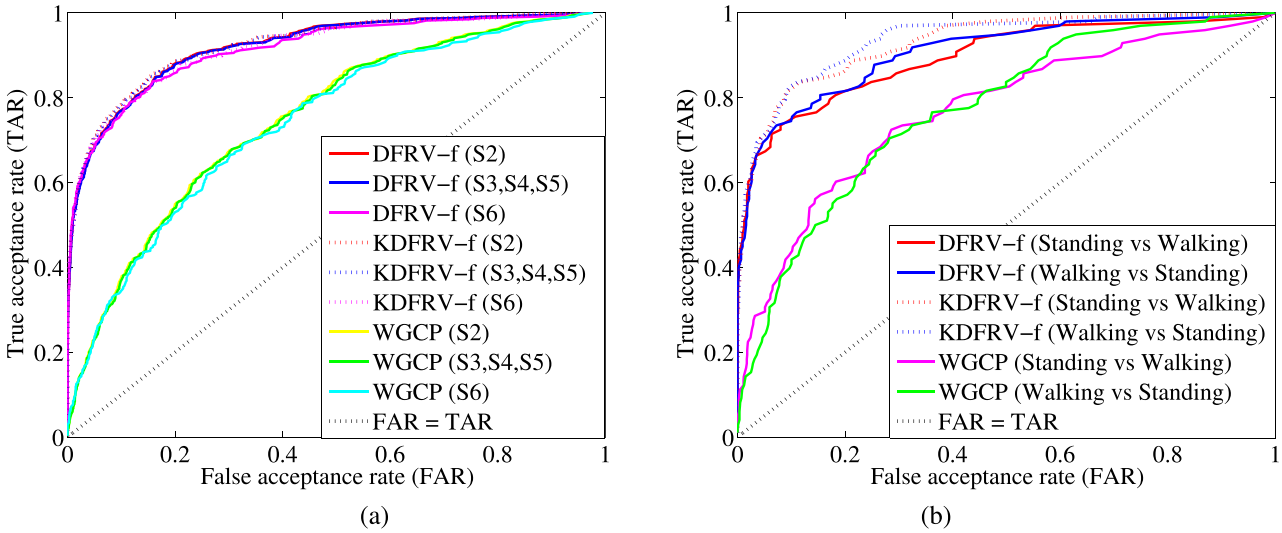

FIGURE 10. (a) ROC curves of the S2, S3, and S6 leave-one-out identification experiments on the UMD Comcast10 dataset. (b) ROC curves of standing vs walking, and walking vs standing verification experiments on the Comcast10 sequences. Both DFRV-f and KDFRV-f obtained better ROC curves than the WGCP method.

walking sequences as gallery) and "Walking vs Standing" (walking sequences as probe; standing sequences as gallery) experiments. For DFRV-f and KDFRV-f, the ROC curve of the "Walking vs Standing" experiment is slightly better than the other, which can be explained by the fact that there are more frames with frontal faces available in a walking sequence than a standing sequence. Similar to the identification results, KDFRV-f performs slightly better than the DFRV-f method. However, they both outperform the WGCP method.

VI. CONCLUSIONS

We presented a video dictionary-based family of algorithms for unconstrained video-to-video human identification and verification. To enhance the discriminative recognition, we extended our original work in [13] to handle the nonlinearities in video data by learning kernel dictionaries. Moreover, we used upper body features to improve the recognition accuracy. We further demonstrated the effectiveness of our dictionary-based approach by experimentally measuring the performance gain of our method over a baseline that omits dictionary learning. Finally, extensive experiments on four unconstrained video datasets show that our approach performs better than many well known video-based face recognition methods in the literature.

Several future directions of inquiry are possible considering our new approach. Many complicated fusion techniques go beyond a simple score level fusion such as decision and feature level fusion. It is very likely that utilization of such fusion techniques will provide an even greater performance when facial features are combined with body features for person recognition.

ACKNOWLEDGMENT

The identification of any commercial product or trade name does not imply endorsement or recommendation by NIST or the University of Maryland.

REFERENCES

[1] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *ACM Comput. Surv.*, vol. 35, no. 4, pp. 399–458, Dec. 2003.

[2] V. M. Patel, T. Wu, S. Biswas, P. J. Phillips, and R. Chellappa, "Dictionary-based face recognition under variable lighting and pose," *IEEE Trans. Inf. Forensics Security*, vol. 7, no. 3, pp. 954–965, Jun. 2012.

[3] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, Feb. 2009.

[4] J. R. Barr, K. W. Bowyer, P. J. Flynn, and S. Biswas, "Face recognition from video: A review," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 26, no. 5, p. 1266002, Aug. 2012.

[5] S. Z. Li and A. Jain, Eds., *Handbook of Face Recognition*. New York, NY, USA: Springer-Verlag, 2011.

[6] A. J. O'Toole *et al.*, "Recognizing people from dynamic and static faces and bodies: Dissecting identity with a fusion approach," *Vis. Res.*, vol. 51, no. 1, pp. 74–83, Jan. 2011.

[7] M. Tistarelli, S. Z. Li, and R. Chellappa, Eds., *Handbook of Remote Biometrics: For Surveillance and Security*. New York, NY, USA: Springer-Verlag, 2009.

[8] A. A. Ross, K. Nandakumar, and A. K. Jain, *Handbook of Multibiometrics*. New York, NY, USA: Springer-Verlag, 2006.

[9] K.-C. Lee, J. Ho, M.-H. Yang, and D. Kriegman, "Visual tracking and recognition using probabilistic appearance manifolds," *Comput. Vis. Image Understand.*, vol. 99, no. 3, pp. 303–331, Sep. 2005.

[10] A. Kale, A. K. Roychowdhury, and R. Chellappa, "Fusion of gait and face for human identification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, vol. 5. May 2004, pp. V-901–V-904.

[11] X. Zhou and B. Bhanu, "Integrating face and gait for human recognition at a distance in video," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 37, no. 5, pp. 1119–1137, Oct. 2007.

[12] S. Sarkar, P. J. Phillips, Z. Liu, I. R. Vega, P. Grother, and K. W. Bowyer, "The humanID gait challenge problem: Data sets, performance, and analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 2, pp. 162–177, Feb. 2005.

[13] Y.-C. Chen, V. M. Patel, P. J. Phillips, and R. Chellappa, "Dictionary-based face recognition from video," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2012, pp. 766–779.

[14] P. J. Phillips *et al.*, "Overview of the multiple biometrics grand challenge," in *Proc. Int. Conf. Biometrics*, 2009, pp. 705–714.

[15] National Institute of Standards and Technology. *Multiple Biometric Grand Challenge (MBGC)*. [Online]. Available: http://www.nist.gov/itl/iad/ig/mbgc.cfm, accessed May 15, 2015.

[16] National Institute of Standards and Technology. *Face and Ocular Challenge Series (FOCS)*. [Online]. Available: http://www.nist.gov/itl/iad/ig/focs.cfm, accessed May 15, 2015.

[17] R. Chellappa, J. Ni, and V. M. Patel, "Remote identification of faces: Problems, prospects, and progress," *Pattern Recognit. Lett.*, vol. 33, no. 14, pp. 1849–1859, Oct. 2012.

[18] G. D. Hager and P. N. Belhumeur, "Efficient region tracking with parametric models of geometry and illumination," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 10, pp. 1025–1039, Oct. 1998.

[19] A. Lanitis, C. J. Taylor, and T. F. Cootes, "Automatic interpretation and coding of face images using flexible models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 743–756, Jul. 1997.

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

[21] 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 Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 4, pp. 322–336, Apr. 2000.

[22] G. Aggarwal, A. Veeraraghavan, and R. Chellappa, "3D facial pose tracking in uncalibrated videos," in *Proc. Int. Conf. Pattern Recognit. Mach. Intell.*, 2005, pp. 515–520.

[23] S. Zhou, V. Krueger, and R. Chellappa, "Probabilistic recognition of human faces from video," *Comput. Vis. Image Understand.*, vol. 91, nos. 1–2, pp. 214–245, Jul./Aug. 2003.

[24] O. Arandjelović, "Computationally efficient application of the generic shape-illumination invariant to face recognition from video," *Pattern Recognit.*, vol. 45, no. 1, pp. 92–103, Jan. 2012.

[25] O. Arandjelović and R. Cipolla, "Achieving robust face recognition from video by combining a weak photometric model and a learnt generic face invariant," *Pattern Recognit.*, vol. 46, no. 1, pp. 9–23, Jan. 2013.

[26] P. Sprechmann and G. Sapiro, "Dictionary learning and sparse coding for unsupervised clustering," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Mar. 2010, pp. 2042–2045.

[27] M. Yang, L. Zhang, J. Yang, and D. Zhang, "Metaface learning for sparse representation based face recognition," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2010, pp. 1601–1604.

[28] P. Turaga, A. Veeraraghavan, A. Srivastava, and R. Chellappa, "Statistical computations on Grassmann and Stiefel manifolds for image and video-based recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 11, pp. 2273–2286, Nov. 2011.

[29] Y. Hu, A. S. Mian, and R. Owens, "Sparse approximated nearest points for image set classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 121–128.

[30] R. Wang, S. Shan, X. Chen, and W. Gao, "Manifold-manifold distance with application to face recognition based on image set," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[31] J. Lu, G. Wang, and P. Moulin, "Image set classification using holistic multiple order statistics features and localized multi-kernel metric learning," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 329–336.

[32] T.-K. Kim, J. Kittler, and R. Cipolla, "Discriminative learning and recognition of image set classes using canonical correlations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 1005–1018, Jun. 2007.

[33] H. Cevikalp and B. Triggs, "Face recognition based on image sets," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 2567–2573.

[34] W. Fan and D.-Y. Yeung, "Locally linear models on face appearance manifolds with application to dual-subspace based classification," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2. 2006, pp. 1384–1390.

[35] A. Hadid and M. Pietikainen, "From still image to video-based face recognition: An experimental analysis," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, May 2004, pp. 813–818.

[36] H. S. Bhatt, R. Singh, and M. Vatsa, "On recognizing faces in videos using clustering-based re-ranking and fusion," *IEEE Trans. Inf. Forensics Security*, vol. 9, no. 7, pp. 1056–1068, Jul. 2014.

[37] L. An, M. Kafai, and B. Bhanu, "Dynamic Bayesian network for unconstrained face recognition in surveillance camera networks," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 3, no. 2, pp. 155–164, Jun. 2013.

[38] S. Liao, A. K. Jain, and S. Z. Li, "Partial face recognition: Alignment-free approach," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 5, pp. 1193–1205, May 2013.

[39] L. Best-Rowden, H. Han, C. Otto, B. F. Klare, and A. K. Jain, "Unconstrained face recognition: Identifying a person of interest from a media collection," *IEEE Trans. Inf. Forensics Security*, vol. 9, no. 12, pp. 2144–2157, Dec. 2014.

[40] M. Du, A. C. Sankaranarayanan, and R. Chellappa, "Robust face recognition from multi-view videos," *IEEE Trans. Image Process.*, vol. 23, no. 3, pp. 1105–1117, Mar. 2014.

[41] M. Kafai, L. An, and B. Bhanu, "Reference face graph for face recognition," *IEEE Trans. Inf. Forensics Security*, vol. 9, no. 12, pp. 2132–2143, Dec. 2014.

[42] X. Xing, K. Wang, T. Yan, and Z. Lv, "Complete canonical correlation analysis with application to multi-view gait recognition," *Pattern Recognit.*, 2015.

[43] G. Goswami, R. Bhardwaj, R. Singh, and M. Vatsa, "MDLFace: Memorability augmented deep learning for video face recognition," in *Proc. IEEE Int. Joint Conf. Biometrics*, Sep./Oct. 2014, pp. 1–7.

[44] Z. Huang, R. Wang, S. Shan, and X. Chen, "Face recognition on large-scale video in the wild with hybrid Euclidean-and-Riemannian metric learning," *Pattern Recognit.*, vol. 48, no. 10, pp. 3113–3124, Oct. 2015.

[45] M. De-la-Torre, E. Granger, P. V. W. Radtke, R. Sabourin, and D. O. Gorodnichy, "Partially-supervised learning from facial trajectories for face recognition in video surveillance," *Inf. Fusion*, vol. 24, pp. 31–53, Jul. 2015.

[46] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, May 2004.

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

[48] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, Nov. 2006.

[49] H. Van Nguyen, V. M. Patel, N. M. Nasrabadi, and R. Chellappa, "Design of non-linear kernel dictionaries for object recognition," *IEEE Trans. Image Process.*, vol. 22, no. 12, pp. 5123–5135, Dec. 2013.

[50] P. Turaga, A. Veeraraghavan, and R. Chellappa, "Statistical analysis on Stiefel and Grassmann manifolds with applications in computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[51] A. Jain, K. Nandakumar, and A. Ross, "Score normalization in multimodal biometric systems," *Pattern Recognit.*, vol. 38, no. 12, pp. 2270–2285, Dec. 2005.

[52] J. J. Thiagarajan, K. N. Ramamurthy, and A. Spanias, "Multiple kernel sparse representations for supervised and unsupervised learning," *IEEE Trans. Image Process.*, vol. 23, no. 7, pp. 2905–2915, Jul. 2014.

[53] A. Shrivastava, V. M. Patel, and R. Chellappa, "Multiple kernel learning for sparse representation-based classification," *IEEE Trans. Image Process.*, vol. 23, no. 7, pp. 3013–3024, Jul. 2014.

[54] R. Wang and X. Chen, "Manifold discriminant analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 429–436.

YI-CHEN CHEN received the B.S. degree in electrical engineering from National Tsing Hua University, Hsinchu, Taiwan, the M.S. degree in communication engineering from National Taiwan University, Taipei, Taiwan, and the Ph.D. degree in electrical engineering from the University of Maryland, College Park, MD, USA.

His research interests are in computer vision and pattern recognition, with a primary focus on face recognition, salient view selection, clustering, content-based image retrieval, and communications.

VISHAL M. PATEL (M'01) received the B.S. (Hons.) degree in electrical engineering and applied mathematics and the M.S. degree in applied mathematics from North Carolina State University, Raleigh, NC, USA, in 2004 and 2005, respectively, and the Ph.D. degree in electrical engineering from the University of Maryland, College Park, MD, USA, in 2010. He was a Research Faculty Member with the Institute for Advanced Computer Studies, University of Maryland. He is currently an Assistant Professor with the Department of Electrical and Computer Engineering, Rutgers University. His current research interests include signal processing, computer vision, and pattern recognition with applications in biometrics and imaging. He is a member of Eta Kappa Nu, Pi Mu Epsilon, and Phi Beta Kappa. He was a recipient of the ORAU Post-Doctoral Fellowship in 2010.

P. JONATHON PHILLIPS (M'96–SM'06–F'10) received the Ph.D. degree in operations research from Rutgers University. He was with the U.S. Army Research Laboratory. His previous efforts include the Iris Challenge Evaluations, the Face Recognition Vendor Test (FRVT) 2006, the Face Recognition Grand Challenge, and FERET. From 2000 to 2004, he was assigned to the Defense Advanced Projects Agency as a Program Manager for the Human Identification at a Distance program. He was the Test Director for the FRVT 2002. He is currently a Leading Researcher in the fields of computer vision, biometrics, face recognition, and human identification. He is with the National Institute of Standards and Technology, where he is involved in designing grand challenges for advancing face recognition and visual biometric technology and science. His work has been reported in print media of record, including *The New York Times* and *The Economist*. He is a fellow of IAPR. For his work on the FRVT 2002, he received the Department of Commerce Gold Medal. He won the inaugural Mark Everingham Prize. From 2004 to 2008, he was an Associate Editor of the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE and a Guest Editor of the Special Issue of the PROCEEDINGS OF THE IEEE on Biometrics. In an Essential Science Indicators analysis of face recognition publication over the past decade, his work ranks at second by total citations and first by cites per paper.

RAMA CHELLAPPA (F'92) is currently a Minta Martin Professor of Engineering and the Chair of the Electrical and Computer Engineering Department with the University of Maryland (UMD). He holds four patents. He is a fellow of the International Association of Pattern Recognition (IAPR), OSA, AAAS, ACM, and AAAI. He received the K.S. Fu Prize from IAPR. He is a recipient of the Society, Technical Achievement, and Meritorious Service Awards from the IEEE Signal Processing Society and four IBM Faculty Development Awards. He also received the Technical Achievement and Meritorious Service Awards from the IEEE Computer Society. At UMD, he received college and university level recognitions for research, teaching, innovation, and mentoring of undergraduate students. In 2010, he was recognized as an Outstanding ECE by Purdue University. He served as the Editor-in-Chief of the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE. He is a Golden Core Member of the IEEE Computer Society, and served as a Distinguished Lecturer of the IEEE Signal Processing Society and the President of the IEEE Biometrics Council.