# Dictionary-based Face Recognition from Video

Yi-Chen Chen[1], Vishal M. Patel[1], P. Jonathon Phillips[2], and Rama Chellappa[1]

[1]Department of Electrical and Computer Engineering
Center for Automation Research, University of Maryland, College Park, MD
[2]National Institute of Standards and Technology, Gaithersburg, MD
{chenyc08,pvishalm,rama}@umiacs.umd.edu jonathon.phillips@nist.gov

**Abstract.** The main challenge in recognizing faces in video is effectively exploiting the multiple frames of a face and the accompanying dynamic signature. One prominent method is based on extracting joint appearance and behavioral features. A second method models a person by temporal correlations of features in a video. Our approach introduces the concept of video-dictionaries for face recognition, which generalizes the work in sparse representation and dictionaries for faces in still images. Video-dictionaries are designed to implicitly encode temporal, pose, and illumination information. We demonstrate our method on the Face and Ocular Challenge Series (FOCS) Video Challenge, which consists of unconstrained video sequences. We show that our method is efficient and performs significantly better than many competitive video-based face recognition algorithms.

## 1 Introduction

Traditional face recognition algorithm recognize faces from still images [1], [2], [3]. 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 [4], [1]. In video-based face recognition, 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 challenges in designing an effective video-based face recognition algorithm.

Numerous methods have been proposed that approach the problem as a multi-still face recognition problem [5], or extract joint appearance and behavioral features from a video [6],[7], or explicitly model the temporal correlations between faces in two videos [4]. 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 [7].

To address the challenges of face recognition from unconstrained videos, we propose a generative approach based on dictionary learning methods, which is robust to changes in illumination and pose. One major advantage of our method

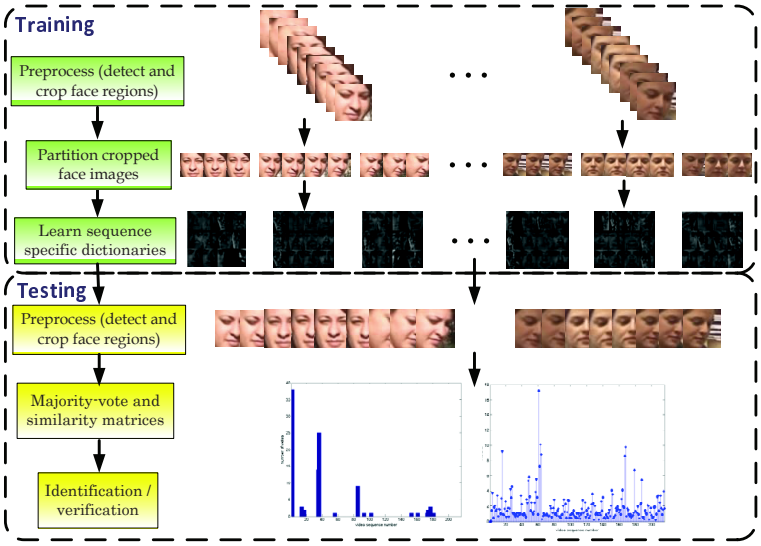



**Fig. 1.** Overview of the proposed approach

is that it is robust to some variations in video sequences. Figure 1 shows an overview of our approach. From cropped face images extracted from a video sequence, we first partition the video sequence so that frames with same pose and illumination 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. In the recognition phase, frames from a given query video sequence are projected onto the span of atoms in every sequence-specific dictionary. From the projection on to the atoms, the residuals are computed and combined to perform recognition or verification. 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 challenging Face and Ocular Challenge Series (FOCS) Video Challenge.

The rest of the paper is organized as follows: In Section 2 we review some recent video-based face recognition methods. Section 3 describes the proposed dictionary-based video face recognition algorithm. In section 4, we demonstrate results on three challenging video datasets. Section 5 concludes the paper with a summary and discussion.

## 2 Related work

In this section, we review some of the recent video-based face recognition methods. In video 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. From the tracked features, one can extract a few key frames that can be used for matching with exemplars in the gallery. Significant work has been done on face tracking using 2D appearance-based models [8], [9], [10]. The 2D approaches; however, do not provide the 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.* [11] proposed a cylindrical face model for face tracking. An extension of this work was proposed by Aggarwal *et al.* in [12] that uses 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 [6]. 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.* [7], 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. This method is robust to large appearance changes if sufficient 3D view variations and illumination variations are available in the training set.

In a related work, Arandjelovic and Cipolla [13] 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 learnt 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, producing the classification decision using robust likelihood estimation.

Recently, Turaga *et al.* [14] presented a statistical method for video based face recognition. These methods use subspace-based models and tools from Riemannian geometry of the Grassmann manifold. Intrinsic and extrinsic statistics are derived for maximum-likelihood classification applications. An image set classification methods for video-based face recognition problem was recently proposed by Hu *et al.* [15]. This method is based on a measure of between-set dissimilarity that is the distance between sparse approximated nearest points of two image sets and uses a scalable accelerated proximal gradient method for optimization.

## 3 Proposed Approach

In this section, we present the details of our proposed dictionary-based video face recognition algorithm. We describe how the video sequence is partitioned into sub-sequences in section 3.1, and how we build sequence-specific dictionaries in

section 3.2. Identification and verification are described in sections 3.3 and 3.4, respectively.

### 3.1 Video Sequence Partition

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. We partition the cropped faces by a $k$-means clustering type of algorithm that is inspired by a video summarization algorithm [16]. Let $S = \{f_1, ..., f_n\}$ 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

**Algorithm 1**: Video sequence partition algorithm.

**Initialization of sets:**

$S = \{f_1, ..., f_n\}, I = \{1, 2, ..., n\}, T = \phi$.

**Procedure:**

1. Find $(i^*, j^*) = \underset{i,j \in I, i \neq j}{\operatorname{argmax}} \|f_i - f_j\|_2$.
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 $k^* = \underset{k \in I}{\operatorname{argmax}} \prod_{l=1}^{|T|} \|f_{t_l} - f_k\|_2$.
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 $\{f_{t_1}, ..., f_{t_K}\}$, use the nearest neighbor criterion to partition $S$ into $K$ partitions, denoted by $S(f_{t_1}, ..., f_{t_K}) = \bigcup_{i=1}^{K} S_i$. $S(f_{t_1}, ..., f_{t_K})$ is the initial partitions which are followed by $N$ iterations of updating described in step 7 and 8.
7. Randomly select $s_i$ from $S_i$, $i = 1, 2, ..., K$, as representatives. Find the corresponding nearest neighbor partitions which are denoted by $S(s_1, s_2, ...s_K)$, and calculate the corresponding score $M(S(s_1, s_2, ...s_K))$.
8. Repeat step 7, and keep updating for $\{s_1^*, s_2^*, ..., s_K^*\}$ which gives the highest score $M$, until the number of repeating iterations for step 7 reaches $N$. In other words,
$\{s_1^*, s_2^*, ..., s_K^*\} = \underset{s_i \in S_i, i=1,2,...,K, \text{ in } N \text{ iterations}}{\operatorname{argmax}} M(S(s_1, s_2, ...s_K))$.

**Output:**

$K$ partitions, $S(s_1^*, s_2^*, ..., s_K^*)$.

and [16] is that the overall cost $J(S) \triangleq \alpha \times err(S) + (1-\alpha) \times (D - div(S))$, in [16] is now replaced with $M(S) \triangleq \frac{div(S)}{err(S)}$, where $err(S)$, $div(S)$ and $D$ are the square error, diversity and an upper bound of diversity of summary $S$, respectively [16]. Using this score, there is no need to set the weighting factor $\alpha$, and the original cost minimization problem becomes an equivalent score maximization problem. The other major difference is that we initialize the partitions deterministically (steps 1 to 6 above). As seen in steps 1 and 3, these $K$ initial representatives

are chosen so they are separated as far apart as possible. The corresponding initial $K$ partitions are then determined by the nearest neighbor criterion. Under the assumption that there exist $K$ exemplars, we expect each of the $K$ initial partitions determined by finding the nearest neighbor among $K$ initial representatives, to contain exactly one exemplar. 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. As long as each of the $K$ exemplars fall in a distinct initial partition, they can be found after sufficient number of iterations. 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.

### 3.2 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. 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 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 (by certain degrees with respect to Z axis) face images.

Let $\mathbf{G}^i_{j,k}$ be the augmented gallery matrix of the $k$th partition of the $j$th video sequence of subject $i$. In $\mathbf{G}^i_{j,k} = [\mathbf{g}^i_{j,k,1}\mathbf{g}^i_{j,k,2}...]$, each column is a vectorized form of the corresponding cropped grayscale face image of size $L$. Given $\mathbf{G}^i_{j,k}$, a dictionary $\mathbf{D}^i_{j,k} \in \mathbb{R}^{L\times\tilde{K}}$ is learned such that the columns of $\mathbf{G}^i_{j,k}$ are best represented by linear combinations of $\tilde{K}$ atoms of $\mathbf{D}^i_{j,k}$. This can be done by minimizing the following representation error

$$(\hat{\mathbf{D}}^i_{j,k}, \hat{\mathbf{\Gamma}}^i_{j,k}) = \underset{\mathbf{D}^i_{j,k},\mathbf{\Gamma}^i_{j,k}}{\operatorname{argmin}} \|\mathbf{G}^i_{j,k} - \mathbf{D}^i_{j,k}\mathbf{\Gamma}^i_{j,k}\|^2_F \quad s.t. \quad \|\boldsymbol{\gamma}_l\|_0 \leq T_0, \ \ \forall l, \tag{1}$$

where $\boldsymbol{\gamma}_l$ is the $l$th column of coefficient matrix $\mathbf{\Gamma}^i_{j,k}$ and $T_0$ is a sparsity parameter. The $\ell_0$ sparsity measure $\|\cdot\|_0$ counts the number of nonzero elements in the representation and $\|\cdot\|_F$ denotes the Frobenius norm. One of the simplest algorithms for finding such a dictionary is the K-SVD algorithm [17][1] that we use to obtain $\mathbf{D}^i_{j,k}$. The video sequence-specific dictionary is constructed by concatenating partition-level sub-dictionaries. In other words, the $j$th dictionary of subject $i$ is

$$\mathbf{D}^i_j = [\mathbf{D}^i_{j,1}\ \mathbf{D}^i_{j,2}\ ...\ \mathbf{D}^i_{j,k}]. \tag{2}$$

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

### 3.3 Identification

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

$$\hat{p} = \underset{p}{\operatorname{argmin}} \|\mathbf{q}_{k,l}^{(m)} - \mathbf{D}_{(p)}\mathbf{D}_{(p)}^{\dagger}\mathbf{q}_{k,l}^{(m)}\|_2, \tag{3}$$

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

$$p^* = \underset{p}{\operatorname{argmax}} \left( \sum_{k=1}^{K} w_k C_{p,k} \right), \tag{4}$$

where $C_{p,k}$ is the total number of votes from partition $k$ for sequence $p$, and $w_k$ is the weight associated with partition $\mathbf{Q}_k^{(m)}$. 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 $i^* = \boldsymbol{m}(p^*)$.

### 3.4 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

$$\mathbf{R}^{(m,p)} = \min_{k \in \{1,2,...,K\}} \mathbf{R}_k^{(m,p)}, \tag{5}$$

where

$$\mathbf{R}_k^{(m,p)} \triangleq \min_{l \in \{1,2,...,n_k\}} \|\mathbf{q}_{k,l}^{(m)} - \mathbf{D}_{(p)}\mathbf{D}_{(p)}^{\dagger}\mathbf{q}_{k,l}^{(m)}\|_2. \tag{6}$$

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

**Algorithm 2**: Video-based Face Recognition (DFRV)

**Training:**

1. Given a sequence - the $j$th 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 $k$th partition, $\mathbf{G}_{j,k}^i, \forall k = 1, 2, ..., K$.
3. Use the K-SVD algorithm to learn the partition-specific sub-dictionary $\mathbf{D}_{j,k}^i, \forall k = 1, 2, ..., K$. Construct the sequence-specific dictionary $\mathbf{D}_j^i$ as in (2).

**Testing:**

1. Partition the $m$th query video sequence $\mathbf{Q}^{(m)} = \bigcup_{k=1}^{K} \mathbf{Q}_k^{(m)}$, where $\mathbf{Q}_k^{(m)} = [\mathbf{q}_{k,1}^{(m)} \ \mathbf{q}_{k,2}^{(m)} \ ... \ \mathbf{q}_{k,n_k}^{(m)}]$.
2. (Identification) Use (3) to determine the vote from $\mathbf{q}_{k,l}^{(m)}, \forall k, l$. Then, use (4) and subject-sequence correspondence $\boldsymbol{m}(\cdot)$ to make the final decision.
3. (Verification) Find the similarity $\mathbf{R}^{(m,p)}$ between $\mathbf{Q}^{(m)}$ and $\mathbf{D}_{(p)}$ by (5) and (6). Use $\mathbf{R}^{(m,p)}$ to construct the similarity matrix, from which the ROC curves can be obtained.

## 4 Experimental Results

To illustrate the effectiveness of our method, we present experimental results on three publicly available datasets for video-based face recognition: the Multiple Biomertic Grand Challenge (MBGC) [18],[19], the Face and Ocular Challenge Series (FOCS) [20],[21], and the Honda/UCSD datasets [7].

### 4.1 MBGC Video version 1

(a) (b)

**Fig. 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.

The MBGC Video version 1 dataset (Notre Dame dataset) contains 399 walking (frontal-face) and 371 activity (profile-face) video sequences recorded of 146 subjects. Both types of sequences were collected in standard definition (SD) format (720 × 480 pixels) and high definition (HD) format (1440 × 1080 pixels). The 399 walking sequences consist of 201 sequences in SD and 198 in HD. For the 371 walking video sequences, 185 are in SD and 186 are in HD. 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 in these videos. The challenging conditions include frontal and profile faces in shadow, and profile faces sometimes being heavily covered by one's hair.

Figure 3 shows an example of output from the video partitioning stage. For results in Figure 3, the number of partitions is $K = 3$. Results are presented for 4 subjects for walking sequences.Each row shows up to 30 partitioned cropped face 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.

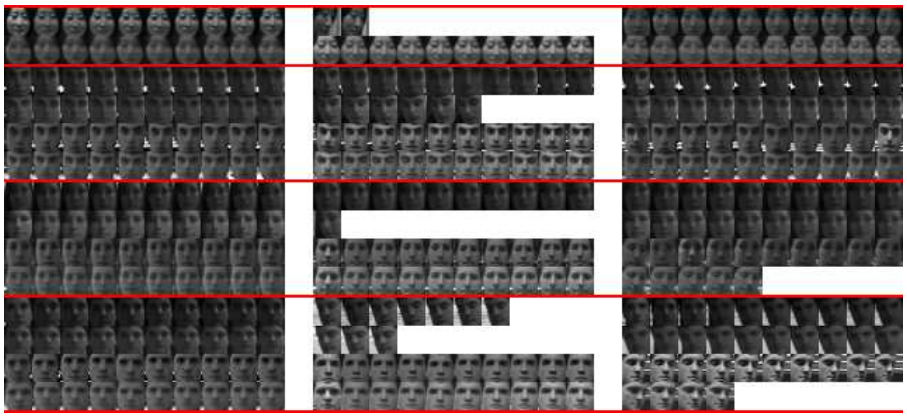


**Fig. 3.** Partition results of MBGC walking sequences (4 subjects only). Red lines separate different subjects. A subject has at least two video sequences. Face images from a video sequence are shown in a row, and are further divided into three partitions. Each partition shows up to 10 face images. A partition represents a particular pose and illumination condition.

Following the experiment design in [14], 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: 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 1 lists the percentages of correct identifications for this experiment. The DFRV method outperforms the statistical-pattern recognition methods reported in [14],[22] and the Sparse Approximated Nearest Points (SANP) method [15].

| MBGC walking videos | Procrustes Metric [14],[22] | Kernel Density [14],[22] | WGCP [14] | SANP [15] | DFRV |
|---|---|---|---|---|---|
| $S2$ | 43.79 | 39.74 | 63.79 | 83.88 | **85.64** |
| $S3$ | 53.88 | 50.22 | 74.88 | 84.02 | **88.13** |
| $S4$ | 53.70 | 50.46 | 75 | 84.26 | **88.43** |
| Average | 50.46 | 46.81 | 71.22 | 84.05 | **87.40** |

**Table 1.** Identification rates of leave-one-out testing experiments on the MBGC walking videos. Our DFRV method outperforms statistical methods and the SANP method, recently proposed in [14] and [15], respectively.

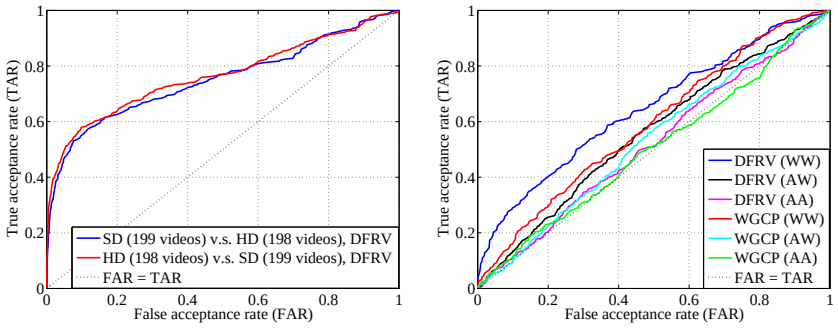
In the second set of experiments, we selected videos associated for those subjects that are in at least two videos (i.e., $S2$). We divide all these videos into SD and HD videos, to conduct "SD v.s. HD" (SD as probe; HD as gallery) and "HD v.s. SD" (HD as probe; SD as gallery) experiments. Correct identification rates are shown in Table 2. The DFRV method performed the best and it outperformed the other methods significantly. Figure 4(a) shows the corresponding

| MBGC walking videos | Procrustes Metric [14],[22] | Kernel Density [14],[22] | WGCP [14] | SANP [15] | DFRV |
|---|---|---|---|---|---|
| SD v.s. HD | 61.31 | 55.78 | 30.15 | 41.71 | **86.93** |
| HD v.s. SD | 68.69 | 56.06 | 30.30 | 45.96 | **91.41** |
| Average | 65 | 55.92 | 30.23 | 43.84 | **89.17** |

**Table 2.** Identification rates of "SD v.s. HD" and "HD v.s. 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 method achieves the best identification rates.

ROC curves for verification experiments. For both identification and verification, HD probes had better performances than SD probes. In this experiment, we examine the effect on performance of varying the number video sequences per person in the gallery. We divide the videos into two groups beforehand either as probe, or as gallery. 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). Results presented above show that WGCP in this setting does not perform so well. We observe that WGCP is able to give satisfactory performance only when there are enough video sequences for training, to obtain more discriminative metrics for different subjects.

In the MBGC [18] protocol, verifications are specified by two sets: target and query. The protocol requires the algorithm to match each target sequence




(a) (b)

**Fig. 4.** (a) ROC curves of SD v.s. HD and HD v.s. SD verification testing experiments on the MBGC frontal (walking) videos using DFRV. (b) ROC curves of the MBGC experiments on frontal (walking) and profile (activity) videos. The proposed DFRV 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.

with all query sequences. We performed three verification experiments: walking v.s. walking (WW), activity v.s. walking (AW), activity v.s. activity (AA). Figure 4(b) shows the ROC curves. We observe that DFRV gives better ROC curve than WGCP for most 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[19], for the AW and AA experiments, no results have been reported that are better than random.

### 4.2 FOCS UT-Dallas Video

The video challenge of Face and Ocular Challenge Series (FOCS) [20] is designed to match "frontal v.s. frontal", "frontal v.s. non-frontal", and "non-frontal v.s. 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 algorithm on the UT Dallas dataset shows the strength of our approach on a second hard data set. In addition, it allows us to directly compare the performance of the DFRV algorithm to humans [23].

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 $720 \times 480$ 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 showing the 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. 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. Table 3 shows identification results. The DFRV algorithm has the best identification rates among all the compared algorithms.

| UT-Dallas walking videos | Procrustes Metric [14],[22] | Kernel Density [14],[22] | WGCP [14] | SANP [15] | DFRV |
|---|---|---|---|---|---|
| $S2$ | 38.12 | 40.84 | 53.22 | 48.27 | **59.90** |
| $S3$ | 60.94 | 64.06 | 70.31 | 60.94 | **78.13** |
| $S4$ | 64 | 64 | 76 | 68.00 | **80.00** |
| Average | 54.35 | 54.97 | 66.51 | 59.07 | **72.68** |

**Table 3.** Identification rates of leave-one-out testing experiments on the FOCS UT-Dallas walking videos. The DFRV method performs the best.

Like MBGC, FOCS specifies a verification protocol: **1A** (walking v.s. walking), **2A** (activity v.s. walking), and **3A** (activity v.s. 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 5 shows ROC curves of verification experiments. In Figure 5(a), we compare the proposed algorithm with WGCP [14]. In all three experiments, the DFRV algorithm is superior to the WGCP algorithm.

O'Toole *et al.* [23] evaluated the accuracy of humans recognizing people in the UT Dallas data set. Human performance was reported for both through static and dynamic presentations of faces and bodies. Performance in [23] was reported for humans viewing the original sequence and for sequences edited to contain only the head. Since the DFRV algorithm only encodes face information, it is reasonable to compare the DFRV with human performance on the original sequences and the edited face only sequences. In Figure 5(b)(c)(d) we compare the performance of the DFRV algorithm and humans for experiments 1A, 2A, and 3A. In Figures 5(b)(d), the performance of the DFRV 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. This suggests that recognition across pose in a video is a good direction for future research.

### 4.3 Honda/UCSD Dataset

The third set of experiments is conducted on the Honda/UCSD Dataset [7]. The Honda Dataset consists of 59 video sequences from 20 distinct subjects. We follow the same experiment procedure in [15]. The experiments are done in three cases of the maximum set length (available number of cropped-face images per video sequence) as defined in [15]: 50, 100 and Full Length frames. Image resolution is $20 \times 20$ pixels. Table 4 shows identification rates of the DFRV and

other methods [15]. The DFRV method ranks first except for the full length case, where the DFRV ranks second, tied with the MDA method [24].

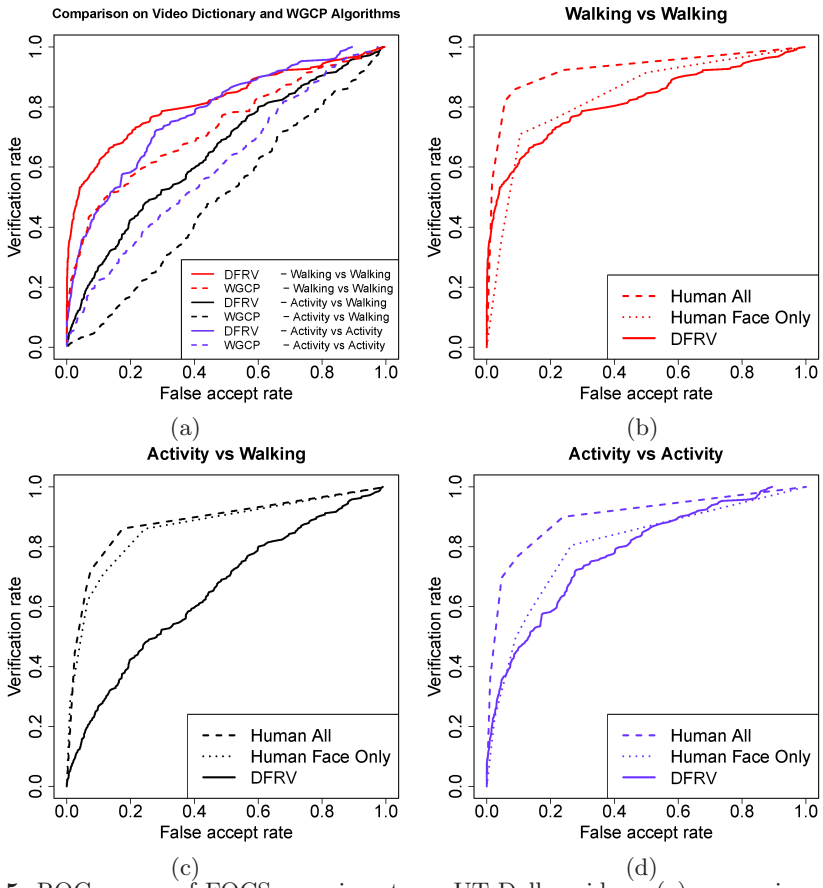



**Fig. 5.** ROC curves of FOCS experiments on UT-Dallas videos: (a) comparison with WGCP [14]; (b)(c)(d) comparison with human perception [23]: (b) walking v.s. walking (c) activity v.s. walking (d) activity v.s. activity. Compared to WGCP, the proposed DFRV method gives better ROC curves, which also stay very close to those of face-only human perception in (b)(d) cases.

## 5 Conclusions and Future work

We have demonstrated the effectiveness of the proposed dictionary approach for video-based face identification and verification. Our experiments show that our method performs better than many competitive video-based face recognition methods. Our experimental results are on three different datasets. It was observed that when viewing the face and body in motion, humans can achieve

| Set length | DCC [25] | MMD [26] | MDA [24] | AHISD [27] | CHISD [27] | SANP [15] | DFRV |
|---|---|---|---|---|---|---|---|
| 50 frames | 76.92 | 69.23 | 74.36 | 87.18 | 82.05 | 84.62 | **89.74** |
| 100 frames | 84.62 | 87.18 | 94.87 | 84.62 | 84.62 | 92.31 | **97.44** |
| Full Length | 94.87 | 94.87 | 97.44 | 89.74 | 92.31 | **100** | 97.44 |
| Average | 85.47 | 83.76 | 88.89 | 87.18 | 86.33 | 92.31 | **94.87** |

**Table 4.** Identification rates on Honda/UCSD Dataset. The proposed DFRV method ranks first except for the full length case where it ranks the second.

the best identification performances [23]. As as shown in Figures 5(b)(c)(d), an important future research direction is developing algorithms that effectively fuse both face and body information for recognition from video.

**Acknowledgement**

This paper was partially supported by a Cooperative Agreement from the National Institute of Standards and Technology under the Grant 70NANB11H023.

## References

1. Zhao, W., Chellappa, R., Phillips, J., Rosenfeld, A.: Face recognition: A literature survey. ACM Computing Surveys (Dec. 2003) 399–458
2. Phillips, P.J.: Matching pursuit filters applied to face identification. IEEE Transactions on Image Processing **7**(8) (Aug. 1998) 1150–1164
3. Patel, V.M., Wu, T., Biswas, S., Philips, P.J., Chellappa, R.: Dictionary-based face recognition under variable lighting and pose. IEEE Transactions on Information Forensics and Security **7**(3) (June 2012) 954–965
4. Tistarelli, M., Li, S.Z., Chellappa, R.: Handbook of Remote Biometrics: For Surveillance and Security. Springer (2009)
5. Ross, A., Nandakumar, K., Jain, A.K.: Handbook of Multibiometrics. Springer (2006)
6. Zhou, S., Krueger, V., Chellappa, R.: Probabilistic recognition of human faces from video. Computer Vision and Image Understanding, Special Issue on Face Recognition **91** (July 2003) 214–245
7. Lee, K.C., Ho, J., Yang, M.H., Kriegman, D.: Visual tracking and recognition using probabilistic appearance manifolds. Computer Vision and Image Understanding **99** (2005) 303–331
8. Hager, G., Belhumeur, P.: Efficient region tracking with parametric models of geometry and illumination. IEEE Transactions on Pattern Analysis and Machine Intelligence **20**(10) (Oct. 1998) 1025–1039
9. Lanitis, A., Taylor, C., Cootes, T.: Automatic interpretation and coding of face images using flexible models. IEEE Transactions on Pattern Analysis and Machine Intelligence **19**(7) (July 1997) 743 –756
10. Zhou, S.K., Chellappa, R., Moghaddam, B.: Visual tracking and recognition using appearance-adaptive models in particle filters. IEEE Transactions on Image Processing **13**(11) (Nov. 2004) 1491–1506

11. La Cascia, M., Sclaroff, S., Athitsos, V.: Fast, reliable head tracking under varying illumination: an approach based on registration of texture-mapped 3d models. IEEE Transactions on Pattern Analysis and Machine Intelligence **22**(4) (Apr. 2000) 322–336
12. Aggarwal, G., Veeraraghavan, A., Chellappa, R.: 3d facial pose tracking in uncalibrated videos. International Conference on Pattern Recognition and Machine Intelligence (2005)
13. Arandjelovic, O., Cipolla, R.: Face recognition from video using the generic shape-illumination manifold. European Conference on Computer Vision **3954** (2006) 27–40
14. Turaga, P.K., Veeraraghavan, A., Srivastava, A., Chellappa, R.: Statistical computations on grassmann and stiefel manifolds for image and video-based recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **33**(11) (Nov. 2011) 2273–2286
15. Hu, Y., Mian, A.S., Owens, R.: Sparse approximated nearest points for image set classification. IEEE Conference on Computer Vision and Pattern Recognition (2011) 27–40
16. Shroff, N., Turaga, P., Chellappa, R.: Video précis: Highlighting diverse aspects of videos. IEEE Transactions on Multimedia **12**(8) (Dec. 2010) 853–868
17. Aharon, M., Elad, M., A., B.: K-svd: An algorithm for desigining overcomplete dictionaries for sparse representation. IEEE Transactions on Signal Processing (Nov. 2006)
18. Phillips, P.J., Flynn, P.J., Beveridge, J.R., Scruggs, W.T., O'Toole, A.J., Bolme, D., Bowyer, K.W., Draper, B.A., Givens, G.H., Lui, Y.M., Sahibzada, H., Scallan III, J.A., Weimer, S.: Overview of the multiple biometrics grand challenge. International Conference on Biometrics (2009)
19. Information Technology Laboratory, National Institute of Standards and Technology: Multiple biomertic grand challenge, http : //www.nist.gov/itl/iad/ig/mbgc.cfm
20. O'Toole, A.J., Harms, J., Snow, S.L., Hurst, D.R., Pappas, M.R., Ayyad, J.H., Abdi, H.: Recognizing people from dynamic and static faces and bodies: Dissecting identity with a fusion approach. Vision Research **51**(1) (2005) 74–83
21. Information Technology Laboratory, National Institute of Standards and Technology: Face and ocular challenge series, http : //www.nist.gov/itl/iad/ig/focs.cfm
22. Turaga, P.K., Veeraraghavan, A., Chellappa, R.: Statistical analysis on stiefel and grassmann manifolds with applications in computer vision. IEEE Conference on Computer Vision and Pattern Recognition (2008) 1–8
23. O'Toole, A.J., Phillips, P.J., Weimer, S., Roark, D.A., Ayyad, J., Barwick, R., Dunlop, J.: Recognizing people from dynamic and static faces and bodies: Dissecting identity with a fusion approach. Vision Research **51**(1) (2011) 74–83
24. Wang, R., Chen, X.: Manifold dsicriminant analysis. IEEE Conference on Computer Vision and Pattern Recognition (2009) 429–436
25. Kim, M.K., Arandjelovic, O., Cipolla, R.: Discrimitive learning and recognition of image set classes using canonical correlations. IEEE Transactions on Pattern Analysis and Machine Intelligence **29**(6) (June 2007) 1005–1018
26. Wang, R., Shan, S., Chen, X., Gao, W.: Manifold-manifold distance with application to face recognition based on image set. IEEE Conference on Computer Vision and Pattern Recognition (2008) 1–8
27. Cevikalp, H., Triggs, B.: Face recognition based on image sets. IEEE Conference on Computer Vision and Pattern Recognition (2010) 2567–2573