Adaptive Representations for Video-based Face Recognition Across Pose

Yi-Chen Chen*, Vishal M. Patel*, Rama Chellappa * and P. Jonathon Phillips †

Abstract

In this paper, we address the problem of matching faces across changes in pose in unconstrained videos. We propose two methods based on 3D rotation and sparse representation that compensate for changes in pose. The first is Sparse Representation-based Alignment (SRA) that generates pose aligned features under a sparsity constraint. The mapping for the pose aligned features are learned from a reference set of face images which is independent of the videos used in the experiment. Thus, they generalize across data sets. The second is a Dictionary Rotation (DR) method that directly rotates video dictionary atoms in both their harmonic basis and 3D geometry to match the poses of the probe videos. We demonstrate the effectiveness of our approach over several state-of-the-art algorithms through extensive experiments on three challenging unconstrained video datasets: the video challenge of the Face and Ocular Challenge Series (FOCS), the Multiple Biometrics Grand Challenge (MBGC), and the Human ID datasets.

1. Introduction

Face recognition with its wide range of commercial and law enforcement applications has been one of the most active areas of research in the field of computer vision. Though face recognition research has traditionally concentrated on recognition from still images, video-based face recognition has gained considerable traction in recent years. Video is a rich source of information that can lead to potentially better representations by integrating multiple views of a face and their corresponding temporal signature.

Numerous approaches have been proposed to exploit the extra information available in video. Arandjelovic and Cipolla [2] represented the variations in shape and illumination by assuming that the shape-illumination manifold for faces is generic. Turaga *et al.* [14], [13] proposed statistical methods using subspace-based models and tools from Riemannian geometry of the Grassmann manifold. Hu *et al.* [8] proposed a method to measure the between-set dissimilarity defined as the distance between sparse approximated nearest points of two image sets. Cui *et al.* [5] proposed the use of a reference image set to align the two image sets. In their work, a video sequence of a person was randomly selected from the training set as the reference image set. The alignment between the reference set and an image set was formulated as an optimization problem and solved by quadratic programming. Chen *at al.* [4] partitioned the video sequence and built video dictionaries that capture changes in illumination and pose, and remove the temporal redundancy.

Though significant efforts have gone into understanding the different sources of variations affecting facial appearance, the accuracy of video-based face recognition algorithms in completely uncontrolled scenarios is still far from satisfactory. Pose and illumination variations still remain as one of the biggest challenges. Some of the above methods [2], [14], [13], [8], [5], [4], rely on the pose diversity contained in the gallery videos to handle pose variations. When there are pose differences between the videos, the robustness of these methods is limited.

Figure 1 shows two typical examples of face mismatching across pose. In Figure 1(a), the first face pair compares frontal and non-frontal images of subject A; the second pair compares frontal images of subjects A and B. In this case, the distance[1] shows a better match between the two frontal images than the true match across pose. Figure 1(b) gives another example where the distance shows a better match

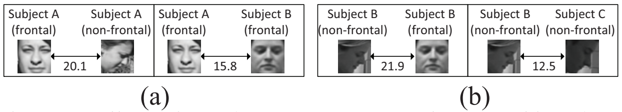


Figure 1. Illustration of common errors when matching faces across changes in pose. (a) The first face pair compares frontal and non-frontal images of subject A; the second pair compares frontal images of subjects A and B. (b) The first face pair compares non-frontal and frontal images of subject B; the second pair compares non-frontal images of subject B and C. In both cases, the distance shows a better match between the two in-pose images than the true match across pose.

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

† P. Jonathon Phillips is with National Institute of Standards and Technology, Gaithersburg, MD. jonathon.phillips@nist.gov

[1] Here, we take the ℓ_2-norm distance between two images.

between the two non-frontal images than the true match. Whenever the gallery videos contain only frontal poses and the probe videos contain only side-poses (and vice versa)[2], methods discussed above can result in recognition errors.

In this paper, we consider matching faces across very different poses between the probe and the gallery videos. Our reference sets are independent of the gallery and probe sets specified by the protocol. We propose two methods to compute pose aligned features based on 3-dimensional (3D) rotation and sparse representation. The first method is referred to as Sparse Representation-based Alignment (SRA) method. The pose aligned images obtained through this method are referred to as the SRA images. The second method is an adaptation of the SRA method that rotates the video dictionary atoms to align the pose prior to recognition. It is referred to as the Dictionary Rotation (DR) method.

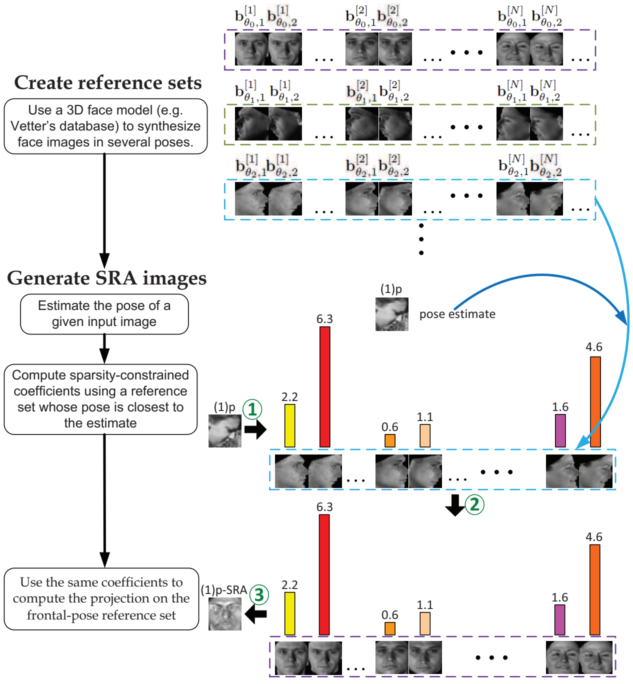

Create reference sets
Use a 3D face model (e.g. Vetter's database) to synthesize face images in several poses.
$b^{[1]}_{\theta_0,1}$ $b^{[1]}_{\theta_0,2}$ $b^{[2]}_{\theta_0,1}$ $b^{[2]}_{\theta_0,2}$ $b^{[N]}_{\theta_0,1}$ $b^{[N]}_{\theta_0,2}$
$b^{[1]}_{\theta_1,1}$ $b^{[1]}_{\theta_1,2}$ $b^{[2]}_{\theta_1,1}$ $b^{[2]}_{\theta_1,2}$ $b^{[N]}_{\theta_1,1}$ $b^{[N]}_{\theta_1,2}$
$b^{[1]}_{\theta_2,1}$ $b^{[1]}_{\theta_2,2}$ $b^{[2]}_{\theta_2,1}$ $b^{[2]}_{\theta_2,2}$ $b^{[N]}_{\theta_2,1}$ $b^{[N]}_{\theta_2,2}$
Generate SRA images
Estimate the pose of a given input image
Compute sparsity-constrained coefficients using a reference set whose pose is closest to the estimate
Use the same coefficients to compute the projection on the frontal-pose reference set
(1)p
pose estimate
(1)p ①
②
(1)p-SRA ③
2.2
6.3
0.6
1.1
1.6
4.6

Figure 2. Illustration of creating reference sets and generating the SRA images.

The proposed SRA method consists of three steps (see Figure 2). In the first step, we obtain candidate reference sets for pose alignment from independent sources. The reference set does not contain videos in the gallery or probe sets specified in the protocol. Candidate reference sets can be other face datasets that contain images of many subjects in various poses, or generated from 3D face models through synthesizing face images. The second step is to generate the SRA images. Given a test image, we estimate its pose, compute the sparsity-constrained coefficient vector on the reference set for the estimated pose, and map the coefficient vector back onto the frontal-pose reference set to obtain the SRA image of the test image. Figure 2 illustrates the first two steps of our method. In the third step, we build the SRA video dictionaries and the base video dictionaries (DFRV [4])[3], and then effectively fuse both video dictionaries to construct the distance matrix for recognition. The SRA video dictionaries enable face recognition across changes in poses. Figures 3 (a) and (b) illustrate the training and the testing stages for building video dictionaries and constructing the distance matrix, respectively.

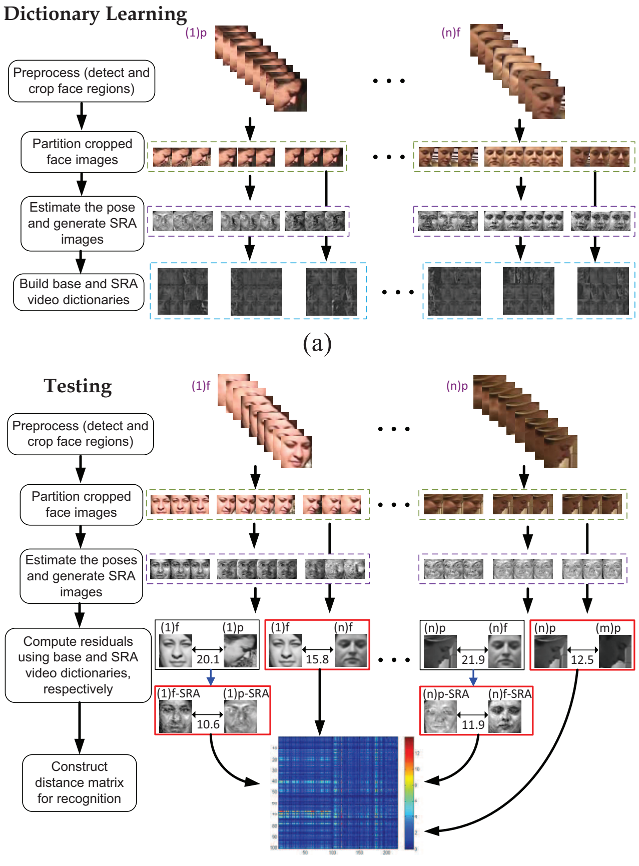

Dictionary Learning
Preprocess (detect and crop face regions)
Partition cropped face images
Estimate the pose and generate SRA images
Build base and SRA video dictionaries
(1)p
(n)f
(a)
Testing
Preprocess (detect and crop face regions)
Partition cropped face images
Estimate the poses and generate SRA images
Compute residuals using base and SRA video dictionaries, respectively
Construct distance matrix for recognition
(1)f
(n)p
(1)f (1)p 20.1
(1)f (n)f 15.8
(n)p (n)f 21.9
(n)p (m)p 12.5
(1)f-SRA (1)p-SRA 10.6
(n)p-SRA (n)f-SRA 11.9
(b)

Figure 3. Illustration of training and testing stages. (a) Training stage: build the base video dictionaries [4] and SRA video dictionaries. (b) Testing stage: compute residuals using both the base video dictionaries and SRA video dictionaries for recognition.

The rest of the paper is organized as follows. Section 2 details the proposed SRA and DR methods. Section 3 describe our approach to pose estimation from videos. We present experimental results and discussions in Section 4. Section 5 concludes the paper with a brief summary.

[2] We refer to gallery videos as enrolled videos for training, and probe videos as videos to be recognized for testing.

[3] We refer to the video dictionaries by DFRV [4] as "base video dictionaries".

2. Sparse Representation-based Alignment

The proposed SRA method computes the pose aligned feature as the re-projection of each face image under an arbitrary pose onto a fixed pose (e.g. frontal) reference set, and then measures the pairwise distances among these re-projections for recognition. The underlying assumption of our method is that whenever a face image under an arbitrary pose θ_1 is represented using a reference set under pose θ_1 weighed by a set of sparse coefficients, then the face image of the same subject under another pose θ_2 can be approximately represented by the re-projection using the *same* set of sparse coefficients on the reference set under pose θ_2.

Without loss of generality, let $\mathbf{y}_\theta$ be a d-dimensional vector representing an input face image under a non-frontal pose θ in its column-vectorized form, where θ_a, θ_e and θ_z stand for azimuth angle (wrt the y axis), elevation angle (wrt the x axis) and the rotation angle wrt the z axis, respectively. The input image can be a probe image for test, or a gallery image for training specified by the protocol.

The proposed SRA method consists of three steps. In what follows, we present details of these steps.

2.1. Obtain Reference Sets for Alignment

In the first step, we obtain the reference sets for pose alignment. Ideally, the reference sets are independent datasets from the protocol with face images from various subjects in different pose and illumination conditions. The poses of the reference sets should cover those in the probe and gallery videos. In practice, when these datasets are not available, or lacking of enough pose and/or subject variability, one alternative is to use a 3D face model (e.g. Vetter's database [3]) to synthesize face images in several poses with illumination changed accordingly [16]. The reference sets can then be built from the synthesized images. Let the resulting reference set from V subjects under a particular pose θ be denoted by

$$\mathbf{B}_\theta = [\mathbf{b}^{[1]}_{\theta,0} \dots \mathbf{b}^{[1]}_{\theta,U-1} | \cdots | \mathbf{b}^{[V]}_{\theta,0} \dots \mathbf{b}^{[V]}_{\theta,U-1}], \tag{1}$$

where $\mathbf{b}^{[v]}_{\theta,u}$ denotes the uth synthetically created variation of face image of the vth subject under pose θ in its column-vectorized form. The variations include slight changes in pose (including θ_a, θ_e and θ_z), illumination or spatial locations. These are created to account for variations among images that are non-ideally cropped from unconstrained videos, and for the pose errors due to non-ideal estimation.

In particular, there are in total only $U-1$ synthetic variations, appearing in the same sequence for all subjects and all poses. In other words, the uth synthetic variation applied to yield $\mathbf{b}^{[v]}_{\theta,u}$ is the same operation for all v and θ. This constraint is required to generate final aligned images in the frontal pose using the sparsity constraint coefficients, as discussed in section 2.2. For simplicity of notation, we use $\mathbf{B}_{\theta_0}$ to denote the reference set from V subjects under the frontal pose.

2.2. Generate SRA Images

In the second step, we generate SRA images using the reference sets presented in section 2.1. We present the motivation of using the sparse representation-based pose aligned feature as follows.

Under the assumption that $\mathbf{y}_\theta$ can be approximated by a sparse linear combination of vectors from $\mathbf{B}_\theta$, we compute the sparse coefficient vector $\hat{\boldsymbol{\gamma}}$ by solving the following optimization problem

$$\hat{\boldsymbol{\gamma}} = \underset{\boldsymbol{\gamma}}{\operatorname{argmin}} \|\boldsymbol{\gamma}\|_1 \;\text{ such that }\; \|\mathbf{y}_\theta - \mathbf{B}_\theta \boldsymbol{\gamma}\|_2^2 \le \varepsilon, \tag{2}$$

where $\|\cdot\|_1$ is the ℓ_1-norm. Let $\mathbf{y}_{\theta_0}$ denote $\mathbf{y}_\theta$'s frontal image, and $\hat{\boldsymbol{\gamma}}_0$ be the solution to (2) with $\mathbf{y}_\theta$ and $\mathbf{B}_\theta$ replaced by $\mathbf{y}_{\theta_0}$ and $\mathbf{B}_{\theta_0}$, respectively. We can relate $\mathbf{y}_\theta$ and $\mathbf{y}_{\theta_0}$ to $\mathbf{B}_\theta$ and $\mathbf{B}_{\theta_0}$ by

$$\mathbf{y}_\theta = \mathbf{B}_\theta \hat{\boldsymbol{\gamma}} + \mathbf{e}, \tag{3}$$

$$\mathbf{y}_{\theta_0} = \mathbf{B}_{\theta_0} \hat{\boldsymbol{\gamma}}_0 + \mathbf{e}_0, \tag{4}$$

where $\mathbf{e}$ and $\mathbf{e}_0$ are error terms. Now, consider the two re-projections: $\mathbf{B}_{\theta_0}\hat{\boldsymbol{\gamma}}_0$, and $\mathbf{B}_{\theta_0}\hat{\boldsymbol{\gamma}}$. In the following, we show the distance $\|\mathbf{B}_{\theta_0}\hat{\boldsymbol{\gamma}}_0 - \mathbf{B}_{\theta_0}\hat{\boldsymbol{\gamma}}\|_2$ can be made small if $\mathbf{B}_\theta\hat{\boldsymbol{\gamma}}$ and $\mathbf{B}_{\theta_0}\hat{\boldsymbol{\gamma}}_0$ can well approximate $\mathbf{y}_\theta$ and $\mathbf{y}_{\theta_0}$, respectively.

Rotating an input image $\mathbf{y}_\theta$ by δ according to the 3D face model can be approximated through the completion of the following two steps: (1) Perform δ-rotation on the harmonic basis of $\mathbf{y}_\theta$ [16]. (2) Apply spatial translation and interpolation according to the 3D δ-rotation matrix. In step (1), the harmonic basis of $\mathbf{y}_\theta$ is changed in accordance with the azimuth, elevation and z axis rotations [16]. We denote the resulting intermediate image vector by $\tilde{\mathbf{y}}_{\theta+\delta}$. In step (2), a spatial translation and interpolation operator $\mathcal{R}_\delta(\cdot)$ determined by the 3D rotation matrix, is applied on $\tilde{\mathbf{y}}_{\theta+\delta}$ to obtain the output image $\mathbf{y}_{\theta+\delta}$. It can be shown that

$$\mathbf{y}_{\theta+\delta} \approx \mathbf{B}_{\theta+\delta}\hat{\boldsymbol{\gamma}} + \mathcal{R}_\delta(\mathbf{e}), \tag{5}$$

$$\|\mathbf{B}_{\theta_0}\hat{\boldsymbol{\gamma}}_0 - \mathbf{B}_{\theta_0}\hat{\boldsymbol{\gamma}}\|_2 \approx \|\mathcal{R}_{-\theta}(\mathbf{e}) - \mathbf{e}_0\|_2 \le \|\mathcal{R}_{-\theta}(\mathbf{e})\|_2 + \|\mathbf{e}_0\|_2. \tag{6}$$

Due to the space limitations, we present more details on the harmonic basis rotation, as well as the derivations for (5) and (6) in the Supplementary Material.

Based on (6), $\|\mathbf{B}_{\theta_0}\hat{\boldsymbol{\gamma}}_0 - \mathbf{B}_{\theta_0}\hat{\boldsymbol{\gamma}}\|_2$ can be made small if the errors $\|\mathbf{e}\|_2$ and $\|\mathbf{e}_0\|_2$ are both small. Even if $\|\mathbf{e}\|_2$ and $\|\mathbf{e}_0\|_2$ cannot be ignored, since $\mathbf{e}_0$ is the reconstruction error of $\mathbf{y}_{\theta_0}$ under frontal pose θ_0, and $\mathbf{e}$ is the reconstruction error of $\mathbf{y}_\theta$ under pose θ, $\mathbf{e}_0$ and $\mathcal{R}_{-\theta}(\mathbf{e})$ should stay close to each other whenever θ is not large. In this case, $\|\mathbf{B}_{\theta_0}\hat{\boldsymbol{\gamma}}_0 - \mathbf{B}_{\theta_0}\hat{\boldsymbol{\gamma}}\|_2$ remains close to zero, and hence in general smaller than $\|\mathbf{y}_{\theta_0} - \mathbf{y}_\theta\|_2$, the distance between two original images under different poses.

From this motivation, we define the SRA image of $\mathbf{y}_\theta$, denoted by $\mathbf{y}_{\theta,\mathrm{SRA}}$, as the re-projection on the frontal reference set $\mathbf{B}_{\theta_0}$ using $\hat{\boldsymbol{\gamma}}$ in (2). It is a synthesized face image in the frontal pose:

$$\mathbf{y}_{\theta,\mathrm{SRA}} = \mathbf{B}_{\theta_0}\hat{\boldsymbol{\gamma}}. \tag{7}$$

Similarly, the SRA image of $\mathbf{y}_{\theta_0}$, denoted by $\mathbf{y}_{\theta_0,\mathrm{SRA}}$, is obtained by replacing $\hat{\boldsymbol{\gamma}}$ with $\hat{\boldsymbol{\gamma}}_0$ in (7). The top part of Figure 2 illustrates example reference sets in different poses created from the Vetter's 3D face model [3]. We assume that images in the frontal pose are initially available. Using the frontal images, we create synthetic variations and then images in different poses $\{\theta_l\}_{l=1}^{L}$, from which reference sets are constructed. Algorithm 1 describes the details for generating the reference sets, $\{\mathbf{B}_{\theta_l}\}_{l=1}^{L}$. The bottom part of Figure 2 illustrates how the SRA feature of an input image is computed. With the pose estimate of the input image, we select the reference set whose pose is closest to the estimate among the all available poses. The coefficient vector is computed with the selected reference set using (2) and then mapped back to $\mathbf{B}_{\theta_0}$ in (7), where the projection onto $\mathbf{B}_{\theta_0}$ is computed as the output SRA image.

```
Algorithm 1: Generate reference sets for poses $\{\theta_l\}_{l=1}^{L}$.
Input: Properly cropped frontal face images from $V$ subjects
    $\{\mathbf{b}_{\theta_0,0}^{[v]}\}_{v=1}^{V}$, a set of possible poses $\{\theta_l\}_{l=1}^{L}$, and Vetter's
    3D face model [3].
Algorithm:
1. Apply predefined $(U-1)$ synthetic variations on each $\mathbf{b}_{\theta_0,0}^{[v]}$ to
obtain $\{\mathbf{b}_{\theta_0,u}^{[v]}\}_{u=1}^{U-1}$, $\forall v \in \{1,...,V\}$. Form $\mathbf{B}_{\theta_0}$ by
concatenating $\mathbf{b}_{\theta_0,u}^{[v]}$'s.
2. Estimate the basis harmonics [17]. Repeat 3 and 4
$\forall\, v \in \{1,...,V\},\ u \in \{1,...,U-1\},\ l \in \{1,...,L\}$:
3. Let $\delta_l = \theta_l - \theta_0$. Given $\mathbf{b}_{\theta_0,u}^{[v]}$, rotate the basis harmonics and
compute the intermediate image $\tilde{\mathbf{b}}_{\theta_l,u}^{[v]}$, where the illumination is
changed accordingly with rotation $\delta_l$ [16].
4. Compute the 3D rotation matrix $\mathbf{R}_{\delta_l}$. Obtain the final rotated
image $\mathbf{b}_{\theta_l,u}^{[v]}$ for each pixel using either direct mapping from the
corresponding source pixel, or interpolation from neighboring
pixels.
5. Collect $\{\mathbf{b}_{\theta_l,u}^{[v]}\}_{l=1}^{L}$'s and obtain $\{\mathbf{B}_{\theta_l}\}_{l=1}^{L}$.
Output: $\mathbf{B}_{\theta_0}$ and $\{\mathbf{B}_{\theta_l}\}_{l=1}^{L}$
```

2.3. Building video dictionaries and Computing Distances

In this section, we describe how the video dictionaries are built and used to compute distances. We refer to the video dictionaries proposed in [4] as the base video dictionaries, and the video dictionaries built using our SRA images as the SRA video dictionaries. Our approach extends the DFRV method [4] to effectively combine both base and SRA video dictionaries in such a way that the base video dictionaries are used only when there is small difference in pose across the the probe and gallery videos, otherwise, the SRA video dictionaries are used to account for the large pose difference across the the probe and gallery videos.

In the DFRV method [4], given the gth video sequence in the training stage, cropped face images extracted from all frames form a set denoted by $S^{(g)}$, and uses the video partition algorithm [4] to separate $S^{(p)}$ into K partitions. Let $\mathbf{G}_k^{(g)}$ denote the resulting gallery matrix from the kth partition, $\forall k = 1, ..., K$. In our method, to further obtain the SRA images of $\mathbf{G}_k^{(g)}$, we assume that all images belonging to partition $\mathbf{G}_k^{(g)}$ are in close poses. Let $\hat{\theta}$ be the estimated pose of $\mathbf{G}_k^{(g)}$. Among all available $\mathbf{B}_\theta$'s, we choose $\mathbf{B}_{\bar{\theta}}$ such that $\bar{\theta}$ is the closest pose to $\hat{\theta}$ among the other poses in the reference sets. For each column in $\mathbf{G}_k^{(g)}$, we use (2) and (7) (with $\mathbf{B}_\theta$ replaced by $\mathbf{B}_{\bar{\theta}}$ accordingly) to compute its SRA image, and concatenate the columns of SRA images to form $\mathbf{G}_{k,\mathrm{SRA}}^{(g)}$. Next, from $\mathbf{G}_k^{(g)}$ and $\mathbf{G}_{k,\mathrm{SRA}}^{(g)}$, we use the K-SVD algorithm [1] to learn the partition-level sub-dictionaries $\mathbf{D}_{(g),k}$, $\mathbf{D}_{(g),k,\mathrm{SRA}}$, $\forall k = 1, ..., K$. Then the base video dictionaries $\mathbf{D}_{(g)}$ [4], and SRA video dictionaries $\mathbf{D}_{(g),\mathrm{SRA}}$ are constructed by concatenating the corresponding sub-dictionaries.

In the testing stage, we partition the pth probe video sequence denoted by $\mathbf{Q}^{(p)} = \bigcup_{k=1}^{K} \mathbf{Q}_k^{(p)}$, where $\mathbf{Q}_k^{(p)} = [\mathbf{q}_{k,1}^{(p)} \ldots \mathbf{q}_{k,n_k}^{(p)}]$ as in [4], and then use (2) and (7) to compute the SRA partition $\mathbf{Q}_{k,\mathrm{SRA}}^{(p)}$, $\forall k = 1, ..., K$. These partitions are collected as $\mathbf{Q}_{\mathrm{SRA}}^{(p)}$.

Let $\mathbf{R}$ be the distance matrix with entry $\mathbf{R}^{(p,g)}$ denoting the residual between the pth probe video and the gth gallery video. Our method to compute $\mathbf{R}^{(p,g)}$ requires using SRA images and SRA video dictionaries only when a gallery video dictionary and partitions of a probe video appear in very different poses. In particular, when poses of $\mathbf{Q}_k^{(p)}$ and $\mathbf{D}_{(g)}$ are close, the corresponding $\mathbf{R}^{(p,g)}$ remains computed from $\mathbf{Q}_k^{(p)}$ and base $\mathbf{D}_{(g)}$ [4]. On the other hand, when their poses are very different, $\mathbf{R}^{(p,g)}$ is computed using their $\mathbf{Q}_{\mathrm{SRA}}^{(p)}$ and $\mathbf{D}_{(g),\mathrm{SRA}}$. Therefore,

$$\mathbf{R}^{(p,g)} = \min_{k\in\{1,\ldots,K\}} \mathbf{R}_k^{(p,g)}, \tag{8}$$

where $\mathbf{R}_k^{(p,g)} =$

$$\begin{cases} \min\limits_{l} \|\mathbf{q}_{k,l}^{(p)} - \mathbf{D}_{(g)}\mathbf{D}_{(g)}^{\dagger}\mathbf{q}_{k,l}^{(p)}\|_2, & \text{if } \eta(\mathbf{Q}_k^{(p)}, \mathbf{D}_{(g)}) = 1, \\ \min\limits_{l} \|\mathbf{q}_{k,l,\mathrm{SRA}}^{(p)} - \mathbf{D}_{(g),\mathrm{SRA}}\mathbf{D}_{(g),\mathrm{SRA}}^{\dagger}\mathbf{q}_{k,l,\mathrm{SRA}}^{(p)}\|_2, & \text{else.} \end{cases}$$

In (8), $\mathbf{D}^\dagger$ denotes the pseudo-inverse of $\mathbf{D}$, and $\eta(\mathbf{Q}_k^{(p)}, \mathbf{D}_{(g)})$ is an indicator function such that $\eta = 1$ if $\mathbf{Q}_k^{(p)}$ and $\mathbf{D}_{(g)}$ are in close poses, and $\eta = 0$ otherwise. Figures 3(a) and (b) are illustrations of building base video dictionaries and SRA video dictionaries, and constructing the distance matrix, respectively. Algorithm 2 summarizes our SRA method.

Algorithm 2: The SRA algorithm.

```
Training:
  1. Given a sequence - the $g$th video, extract all the frames from it.
     Detect and crop face regions to form a set $S^{(g)}$.
  2. Separate $S^{(g)}$ into $K$ partitions. Augment each partition by
     adding synthetic images and obtain the resulting augmented
     gallery matrix from the $k$th partition, $\mathbf{G}_k^{(g)}, \forall k = 1, ..., K$.
  3. For each column in $\mathbf{G}_k^{(g)}$, use (2) and (7) to compute its SRA
     image. The resulting $\mathbf{G}_{k,\mathrm{SRA}}^{(g)}$ is formed by concatenating
     columns of the corresponding SRA images.
  4. From $\mathbf{G}_k^{(g)}$ and $\mathbf{G}_{k,\mathrm{SRA}}^{(g)}$, use the K-SVD algorithm to learn
     the corresponding partition-level sub-dictionaries $\mathbf{D}_{(p),k}$,
     $\mathbf{D}_{(p),k,\mathrm{SRA}}, \forall k = 1, ..., K$, and video dictionaries $\mathbf{D}_{(g)}$,
     $\mathbf{D}_{(g),\mathrm{SRA}}$.
Testing:
  1. Partition the $p$th probe video sequence $\mathbf{Q}^{(p)} = \bigcup_{k=1}^{K} \mathbf{Q}_k^{(p)}$,
     where $\mathbf{Q}_k^{(p)} = [\mathbf{q}_{k,1}^{(p)} \mathbf{q}_{k,2}^{(p)} ... \mathbf{q}_{k,n_k}^{(p)}]$ as in [4].
  2. Use (2) and (7) to compute the SRA images of $\mathbf{Q}_k^{(m)}$, denoted
     by $\mathbf{Q}_{k,\mathrm{SRA}}^{(p)}$. Then obtain the corresponding $\mathbf{Q}_{\mathrm{SRA}}^{(p)}$.
  3. Using $\mathbf{D}_{(g)}$, $\mathbf{D}_{(g),\mathrm{SRA}}$, $\mathbf{Q}^{(p)}$ and $\mathbf{Q}_{\mathrm{SRA}}^{(p)}$, construct the
     distance matrix $\mathbf{R}^{(p,g)}$ by (8).
```

2.4. Dictionary Rotation

The second method for pose alignment is an extension of the SRA algorithm, designed by rotating the video dictionary atoms in both their harmonic basis and 3D geometry. In other words, it performs 3D rotation on atoms of video dictionaries to match the pose prior to recognition. We refer to this method as Dictionary Rotation (DR). We first obtain the pose estimate for the kth partition of the pth probe video $\mathbf{Q}_k^{(p)}$, and then use steps $2 \sim 4$ of Algorithm 1 to rotate each column of $\mathbf{D}_{(g)}$ to the pose estimate[4]. Let the resulting video dictionary be denoted by $\mathbf{D}_{(g),DR}^{(p),k}$. The same steps are repeated for all K partitions of $\mathbf{Q}^{(p)}$. Next, we use (8) to find $\mathbf{R}^{(p,g)}$, with $\mathbf{R}_k^{(p,g)}$ replaced by

$$\mathbf{R}_k^{(p,g)} = \min_l \|\mathbf{q}_{k,l}^{(p)} - \mathbf{D}_{(g),DR}^{(p),k} \left(\mathbf{D}_{(g),DR}^{(p),k}\right)^{\dagger} \mathbf{q}_{k,l}^{(p)}\|_2. \quad (9)$$

Prior to computing the distance, the pose alignment is done by directly rotating dictionary atoms to the estimated pose from each partition of a given probe video. The underlying motivation of this method is based on the fact that if a probe image $\mathbf{q}_\theta$ is represented as a linear combination of video dictionary atoms plus an error term

$$\mathbf{q}_\theta = \mathbf{D}_{(g)}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad (10)$$

then from (3) and (5), the δ-rotated copy of $\mathbf{q}_\theta$ is

$$\mathbf{q}_{\theta+\delta} \approx \mathcal{R}_\delta(\check{\mathbf{D}}_{(g)})\boldsymbol{\beta} + \mathcal{R}_\delta(\boldsymbol{\epsilon}), \quad (11)$$

where $\check{\mathbf{D}}_{(g)}$ is $\mathbf{D}_{(g)}$ with δ-rotated harmonic basis, and $\mathcal{R}_\delta(\check{\mathbf{D}}_{(g)})$ is the δ-rotated $\mathbf{D}_{(g)}$. Each column of $\mathcal{R}_\delta(\check{\mathbf{D}}_{(g)})$ is δ-rotated version of the corresponding column of $\mathbf{D}_{(g)}$. Therefore, δ-rotating an image can be approximated by a linear combination of the corresponding δ-rotated dictionary atoms weighed by the same coefficient vector.

3. Pose Estimation

Various geometric-based approaches have been proposed in the literature for pose estimation using facial landmarks, such as the location of the eyes, nose, and mouth [6], [7], [15], [9], [16]. Unlike constrained still images, face images extracted from unconstrained videos may suffer from low resolution or bad illumination. This makes automatic detection of landmarks much more difficult. We present a semi-automatic method for estimating poses in videos. First, we select face images of V_1 out of V subjects from the reference set with various poses. For each face image, we manually locate T landmarks. Let $\mathbf{l}_{t,\theta}^v$ be the resulting two dimensional vector representing the spatial location the tth landmark of subject v under pose θ. Let s_k^* be the exemplar of the kth partition obtained through the video sequence partition algorithm presented in [4], with the corresponding vector of the tth landmark denoted by $\mathbf{l}_t(s_k^*)$. For the given test video, we assume that all the images in a partition have approximately the same pose. Due to the fact that a video usually contains a large quantity of images, instead of locating landmarks for all images, we manually locate landmarks on the K exemplar images only. The pose estimate of each exemplar is used to represent the pose of the corresponding partition. Using nearest neighbor criterion, we select the pose with landmark vectors $\{\mathbf{l}_{t,\theta}^v\}_{v=1,t=1}^{V_1,T}$ that gives the minimum average distance to $\{\mathbf{l}_t(s_k^*)\}_{t=1}^T$ as the pose estimate $\hat{\theta}$ of the partition. In other words,

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \frac{\sum_{v=1}^{V_1} \sum_{t=1}^{T} \|\mathbf{l}_{t,\theta}^v - \mathbf{l}_t(s_k^*)\|_2}{V_1 T}. \quad (12)$$

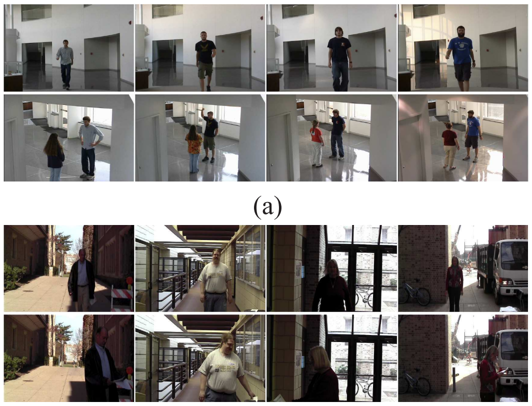
For images from unconstrained videos, however, sometimes even manually locating the landmarks is impossible due to the image's extremely poor resolution and illumination. In this case, we simply examine the face image and roughly estimate the pose directly.

4. Experimental results

We evaluate our methods on three challenging datasets: the video challenge of the Face and Ocular Challenge Series (FOCS) [10], the Multiple Biometrics Grand Challenge (MBGC) [12], and the Human ID [11] datasets. For FOCS

[4] Here, the frontal pose θ_0 in steps $2 \sim 4$ of Algorithm 1 is replaced by a general pose θ.

and MBGC datasets, we created the reference sets using the 3D face model [3] from 100 subjects in the Vetter's database. For the Human ID dataset, as it contains facial moving mug shot videos with face poses in θ_a ranging from $-90^\circ \sim 90^\circ$, we collected frames directly from these videos as reference sets. There is no overlap between subjects whose videos are used as reference sets and subjects whose videos are used as probe and gallery videos for testing.


(a)

(b)

Figure 4. Examples of UT-Dallas and MBGC video sequences. (a) UT-Dallas walking (top row) and activity (bottom row) sequences. (b) MBGC walking (top row) and activity (bottom row) sequences.

4.1. FOCS UT-Dallas Video

The video challenge of Face and Ocular Challenge Series (FOCS) [10] contains an experiment on data collected at the University of Texas at Dallas (UT-Dallas). This dataset consists of 510 walking (frontal face) and 506 activity (non-frontal face) video sequences recorded from 295 subjects with frame size 720 × 480 pixels. This dataset allows for the design of experiments that match "frontal vs. frontal", "frontal vs. non-frontal", and "non-frontal vs. non-frontal" video sequences. The top row of Figure 4(a) shows key frames from four different walking sequences of one subject. The videos were recorded on different days. In the walking sequences shown on the top row, 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 a profile face. The bottom row of figure 4(a) shows key frames of four different activity sequences of the same subject. In these sequences, the subject stands and talks with another person whose back is to the video camera. The sequences contain normal head motions turning up to $\theta_a = \pm 90^\circ$ during a conversation.

We resized the faces to 20 × 20 pixels and conducted leave-one-out tests on 3 subsets: $S2$ (294 subjects, 1014 videos), $S4$ (183 subjects, 782 videos), and $S6$ (19 subjects, 126 videos)[5]. Unlike DFRV [4] where only walking videos were chosen for identification tests, we conduct experiments across *both* walking and activity videos[6]. Table I shows identification rates. Table I shows our SRA and DR methods performed better than state-of-the-art algorithms including SANP [8] and DFRV [4]. The SRA approach was better than the DR method on 2 of 3 cases, and tied on 1 case. Statistics-based approaches including PM, KD and WG [14], [13], no longer give satisfactory results. The 'no DL' is a baseline method that represents each video partition directly as a set of basis atoms without dictionary learning.

Figure 5 compares ROC curves of DFRV [4] and our SRA method. As shown, while there is no difference between both methods under "W vs W" (walking vs walking[7]) and "A vs A" (activity vs activity) verification protocols, the proposed SRA method outperforms DFRV under "A vs W" (activity vs walking) and "W vs A" (walking vs activity) protocols[8]. This is explained by the fact that the SRA method takes the same distances as DFRV [4] when matching in-pose videos ("W vs W" and "A vs A"), while it uses pose aligned feature (i.e. SRA image) to measure the distance between videos across different poses ("A vs W" and "W vs A"). Based on Table I and Figure 5, the proposed SRA outperforms other methods through the use of pose aligned feature in matching out-of-pose videos.

4.2. MBGC Video version 1

The MBGC Video version 1 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 top row of Figure 4(b) 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 4(b) shows example frames from four different activity sequences, where each subject reads from a paper. The activity sequences consists of non-frontal views of the subject.

Each cropped face image was resized to 20 × 20 pixels for the experiment. We conducted leave-one-out tests on 3 subsets: $S2$ (145 subjects, 769 videos), $S5$ (55 subjects, 426 videos), and $S8$ (48 subjects, 384 videos). Similar to section 4.1, the identification experiments are performed across *both* walking and activity videos. Table II shows identification results. As shown, the proposed SRA and DR obtained

[5] We refer to Sn as subjects that have at least n video sequences.

[6] According to (8)(9), the proposed methods can achieve the same performance as DFRV [4] when matching in-pose (walking) videos. We chose not to repeat identification experiments on only walking videos, because in-pose video matching is not the focus of this paper.

[7] This means walking videos as probe, and walking videos as gallery. "A vs A", "A vs W" and "W vs A" can be explained in the same manner.

[8] This is true when the false acceptance rate (FAR) is less than 0.5, which covers the upper limit of FAR for most applications.

UT-Dallas all (W & A) videos	PM [14],[13]	KD [14],[13]	WGCP [14]	SANP [8]	baseline (no DL)	DFRV [4]	DR	SRA
$S2$	17.46	14.89	8.48	25.54	22.98	23.67	**28.40**	**28.40**
$S4$	24.30	20.33	11.64	33.38	29.80	31.59	36.70	**38.36**
$S6$	47.62	43.65	30.16	51.59	50.79	55.56	59.52	**62.70**
Average	29.79	26.29	16.76	36.84	34.52	36.94	41.54	**43.15**

Table I. Identification rates of leave-one-out testing experiments on the FOCS UT-Dallas (both walking and activity) videos.

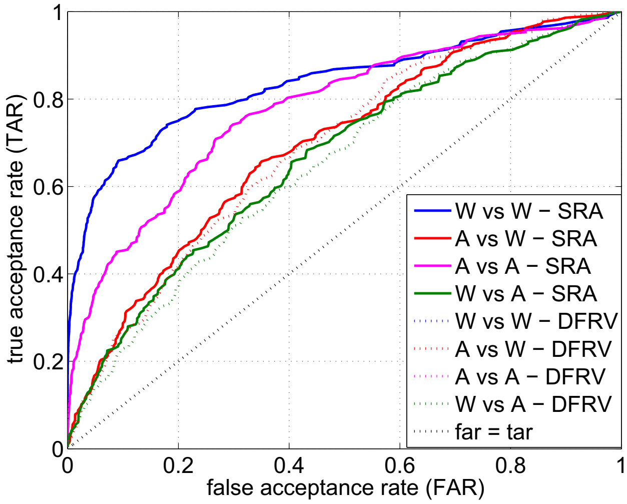


Figure 5. ROC curves of the DFRV [4] and our SRA methods on the FOCS UT-Dallas videos. Our SRA method takes the same distances as DFRV [4] when matching in-pose videos ("W vs W" and "A vs A"), and uses the pose aligned feature to measure distances between out-of-pose videos ("A vs W" and "W vs A"). As shown, it outperforms DFRV in out-of-pose scenarios.

improved identification rates over comparable algorithms. In addition, the MBGC dataset contains videos in both HD and SD formats for the same subject recorded in the same day, while videos of each subject in the FOCS dataset were recorded on different days, during which the subjects may have changed style in their hair, facial hair, expression, pose and illumination. This explains the overall much higher recognition rates on the MBGC dataset than those on the FOCS dataset.

4.3. Human ID database

The Human ID database [11] contains videos of human faces and people, which is useful for testing algorithms for face and person recognition. For each selected subject, there are videos of moving facial mug shots, facial speech, dynamic facial expressions, walking on the same day, and walking on a different day. A complete set of videos is available for 284 subjects. We selected videos of 30 out of 284 subjects from the Human ID database for our experiments. The face region was properly cropped and resized to 30 × 24 pixels. The first three rows of Figures 6 show cropped face images of one subject from its moving facial mug shot, facial speech and dynamic facial expression videos, respectively. The last row of Figures 6 shows the walking video frames of the same subject recorded on the same day (left) as the first three videos, and on a different day (right). The facial mug shot video contains poses from the left side pose to the right side pose (θ_a from $-90° \sim 90°$ wrt the y axis), incremented in a step of $22.5°$. Each gallery video is a trimmed facial moving mug shot video that contains face images with poses in θ_a ranging from about $0° \sim 90°$, while probe videos of the same subject contains facial speech, expression, walking (on the same day) and walking (on a different day) videos. Cropped face images from the probe videos are almost always frontal.

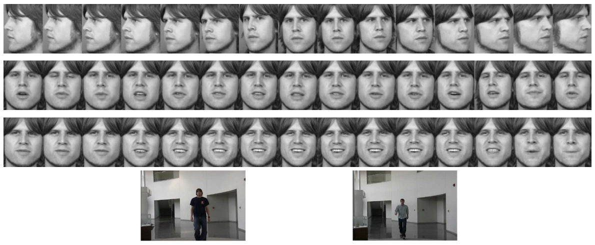


Figure 6. Example frames from the Human ID database. Videos include: moving facial mug shots (1st row), facial speech (2nd row), dynamic facial expression (3rd row), walking on the same day (4th row left), and walking on a different day (4th row right).

Table III shows recognition rates on the four types of probe videos. The proposed SRA obtained the best average results. The SRA* column in Table III gives recognition rates when the reference set is constructed using frames of facial moving mug shot videos from another set of 30 subjects (i.e. other than gallery/probe videos). We observe SRA*, however, did not improve the performance compared to SRA. This can be explained by the fact that images used to construct the reference sets were chosen from a fix set of indices for all 30 selected subjects. When recorded, however, the timing of head turning may vary among the different subjects. Therefore, unlike Vetter's face reference sets, poses and their variations were in fact not aligned across different subjects. The resulting projection error may make the SRA* distances even greater than the original out-of-pose distances for the same subject. Hence, this experiment highlights the importance of the choice of reference sets.

MBGC v1 all	PM	KD	WGCP	SANP	baseline	DFRV	DR	SRA
(W & A) videos	[14],[13]	[14],[13]	[14]	[8]	(no DL)	[4]		
$S2$	41.48	31.86	14.17	68.79	62.55	69.70	80.88	**82.18**
$S5$	43.90	35.68	17.84	69.25	63.38	70.66	80.28	**81.22**
$S8$	44.53	35.94	18.49	71.09	64.32	71.35	81.51	**82.55**
Average	43.30	34.49	16.83	69.71	63.42	70.57	80.89	**81.98**

Table II. Identification rates of leave-one-out testing experiments on the MBGC v1 (both walking and activity) videos.

Human ID video types	PM	KD	WGCP	SANP	baseline	DFRV	DR	SRA*	SRA
	[14],[13]	[14],[13]	[14]	[8]	(no DL)	[4]			
Facial speech	40.00	33.33	20.00	43.33	36.67	63.33	**73.33**	63.33	63.33
Facial expression	33.33	13.33	10.00	36.67	26.67	**56.67**	53.33	53.33	**56.67**
Walking (same day)	3.33	3.33	6.67	**20.00**	16.67	20.00	13.33	20.00	**30.00**
Walking (different day)	10.00	6.67	6.67	6.67	10.00	13.33	13.33	10.00	**23.33**
Average	21.67	14.17	10.84	26.67	22.50	38.33	38.33	36.67	**43.33**

Table III. Identification rates of matching 4 types of probe videos with the moving facial mug shot gallery videos on the Human ID database.

5. Conclusions

We have extended the existing unconstrained video-to-video face recognition frameworks to the one that explicitly addresses the challenge of matching probe and gallery videos in different poses. Our approaches include a sparse representation-based alignment method that generates pose aligned features through pre-designed reference sets under a sparsity constraint, and a dictionary rotation method that directly rotates gallery video dictionary atoms in both their harmonic basis and 3D geometry to match the poses of the probe videos. Through extensive experiments on three challenging unconstrained video datasets across poses, illuminations and facial changes, the proposed SRA and DR have been shown to achieve better recognition performances than several state-of-the-art methods. Our future research directions include applying effective fusion methods and extracting class discriminative feature from pose aligned features.

Acknowledgment

This work was partially supported by Army Research Office MURI Grant W911NF-09-1-0383.

References

[1] M. Aharon, M. Elad, and B. A. K-SVD: An algorithm for desigining overcomplete dictionaries for sparse representation. *IEEE Transactions on Signal Processing*, Nov. 2006. 4

[2] O. Arandjelovic and R. Cipolla. Face recognition from video using the generic shape-illumination manifold. *European Conference on Computer Vision*, 3954:27–40, 2006. 1

[3] V. Blanz and T. Vetter. Face recognition based on fitting a 3D morphable model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(9):1063–1074, September 2003. 3, 4, 6

[4] Y.-C. Chen, V. M. Patel, P. J. Phillips, and R. Chellappa. Dictionary-based face recognition from video. *European Conference on Computer Vision*, October 2012. 1, 2, 4, 5, 6, 7, 8

[5] Z. Cui, S. Shan, H. Zhang, S. Lao, and X. Chen. Image sets alignment for video-based face recognition. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2626–2633, June 2012. 1

[6] A. Gee and R. Cipolla. Determining the gaze of faces in images. *Image and Vision Computing*, 12(10):639–647, 1994. 5

[7] T. Horprasert, Y. Yacoob, and L. Davis. Computing 3-D head orientation from a monocular image sequence. *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 242–247, 1996. 5

[8] Y. Hu, A. S. Mian, and R. Owens. Sparse approximated nearest points for image set classification. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 27–40, 2011. 1, 6, 7, 8

[9] K. Okada and C. v. d. Malsburg. Face recognition and pose estimation with parametric linear subspaces. *Applied Pattern Recognition, Studies in Computational Intelligence*, 91:49–74, 2008. 5

[10] A. J. O'Toole, J. Harms, S. L. Snow, D. R. Hurst, M. R. Pappas, J. H. Ayyad, and H. Abdi. Recognizing people from dynamic and static faces and bodies: Dissecting identity with a fusion approach. *Vision Research*, 51(1):74–83, 2005. 5, 6

[11] A. J. O'Toole, J. Harms, S. L. Snow, D. R. Hurst, M. R. Pappas, J. H. Ayyad, and H. Abdi. A video database of moving faces and people. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(5):812–816, May 2005. 5, 7

[12] P. J. Phillips, P. J. Flynn, J. R. Beveridge, W. T. Scruggs, A. J. O'Toole, D. Bolme, K. W. Bowyer, B. A. Draper, G. H. Givens, Y. M. Lui, H. Sahibzada, J. A. Scallan III, and S. Weimer. Overview of the multiple biometrics grand challenge. *International Conference on Biometrics*, 2009. 5

[13] P. K. Turaga, A. Veeraraghavan, and R. Chellappa. Statistical analysis on Stiefel and Grassmann manifolds with applications in computer vision. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008. 1, 6, 7, 8

[14] P. K. Turaga, A. Veeraraghavan, A. Srivastava, and R. Chellappa. Statistical computations on Grassmann and Stiefel manifolds for image and video-based recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(11):2273–2286, Nov. 2011. 1, 6, 7, 8

[15] J.-G. Wang and E. Sung. EM enhancement of 3D head pose estimated by point at infinity. *Image and Vision Computing*, 25(12):1864–1874, 2007. 5

[16] Z. Yue, W. Zhao, and R. Chellappa. Pose-encoded spherical harmonics for face recognition and synthesis using a single image. *EURASIP Journal on Advances in Signal Processing*, pages 65:1–65:18, January 2008. 3, 4, 5

[17] L. Zhang and D. Samaras. Face recognition from a single training image under arbitrary unknown lighting using spherical harmonics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(3):351–363, March 2006. 4