Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2008, Article ID 469698, 13 pages
doi:10.1155/2008/469698

Research Article

Integrating Illumination, Motion, and Shape Models for Robust Face Recognition in Video

Yilei Xu, Amit Roy-Chowdhury, and Keyur Patel

Department of Electrical Engineering, University of California, Riverside, CA 92521, USA

Correspondence should be addressed to Amit Roy-Chowdhury, amitrc@ee.ucr.edu

Received 30 April 2007; Revised 1 October 2007; Accepted 25 December 2007

Recommended by N. Boulgouris

The use of video sequences for face recognition has been relatively less studied compared to image-based approaches. In this paper, we present an *analysis-by-synthesis* framework for face recognition from *video sequences* that is robust to large changes in facial pose and lighting conditions. This requires tracking the video sequence, as well as recognition algorithms that are able to integrate information over the entire video; we address both these problems. Our method is based on a recently obtained theoretical result that can integrate the effects of motion, lighting, and shape in generating an image using a perspective camera. This result can be used to estimate the pose and structure of the face and the illumination conditions for each frame in a video sequence in the presence of multiple point and extended light sources. We propose a new inverse compositional estimation approach for this purpose. We then synthesize images using the face model estimated from the training data corresponding to the conditions in the probe sequences. Similarity between the synthesized and the probe images is computed using suitable distance measurements. The method can handle situations where the pose and lighting conditions in the training and testing data are completely disjoint. We show detailed performance analysis results and recognition scores on a large video dataset.

Copyright © 2008 Yilei Xu et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

1. INTRODUCTION

It is believed by many that video-based facerecognition systems hold promise in certain applications where motion can be usedas a cue for face segmentation and tracking, and the presence of more data can increase recognition performance [1]. However, these systems have their own challenges. They require tracking the video sequence, as well as recognition algorithms that are able to integrate information over the entire video.

In this paper, we present a novel *analysis-by-synthesis* framework for *pose and illumination invariant, video-based face recognition* that is based on (i) learning joint illumination and motion models from video, (ii) synthesizing novel views based on the learned parameters, and (iii) designing measurements that can compare two time sequences while being robust to outliers. We can handle a variety of lighting conditions, including the presence of multiple point and extended light sources, which is natural in outdoor environments (where face recognition performance is still relatively poor [1–3]). We can also handle gradual and sudden changes of lighting patterns over time. The pose and illumination conditions in the gallery and probe can be *completely disjoint*. We show experimentally that our method achieves high identification rates under extreme changes of pose and illumination.

1.1. Previous work

The proposed approach touches upon aspects of face recognition, tracking and illumination modeling. We place our work in the context of only the most relevant ones.

A broad review of face recognition is available in [1]. Recently, there have been a number of algorithms for pose and/or illumination invariant face recognition, many of which are based on the fact that the image of an object under varying illumination lies in a lower-dimensional linear subspace. In [4], the authors proposed a 3D spherical harmonic basis morphable model (SHBMM) to implement a

face recognition system given one single image under arbitrary unknown lighting. Another 3D face morphable model-(3DMM-) based face recognition algorithm was proposed in [5], but they use the Phong illumination model, estimation of those parameters can be more difficult in the presence of multiple and extended light sources. The authors in [6] proposed to use Eigen light-fields and Fisher light-fields to do pose invariant face recognition. The authors in [7] introduced a probabilistic version of Fisher light-fields to handle the differences of face images due to within-individual variability. Another method of learning statistical dependency between image patches was proposed for pose invariant face recognition in [8]. Correlation filters, which analyze the image frequencies, have been proposed for illumination invariant face recognition from still images in [9]. A novel method for multilinear independent component analysis was proposed in [10] for pose and illumination invariant face recognition.

All of the above methods deal with recognition in a single image or across discrete poses and do not consider continuous video sequences. Video-based face recognition requires integrating the tracking, recognition modules, and exploitation of the spatiotemporal coherence in the data. The authors in [11] deal with the issue of video-based face recognition, but concentrate mostly on pose variations. Similarly, [12] used adaptive hidden Markov models for pose-varying video-based face recognition. The authors of [13] proposed to use a 3D model of the entire head for exploiting features like hairline and handled large pose variations in head tracking and video-based face recognition. However, the application domain is consumer video and requires recognition across a few individuals only. The authors in [14] proposed to perform face recognition by computing the Kullback-Leibler divergence between testing image sets and a learned manifold density. Another work in [15] learns manifolds of face variations for face recognition in video. A method for video-based face verification using correlation filters was proposed in [16], but the poses in the gallery and probe have to be similar.

Except [13] (which is not aimed at face recognition on large datasets), all the rest are 2D approaches, in contrast to our 3D model-based method. The advantage of using 3D models in face recognition has been highlighted in [17], but their focus is on acquiring 3D models directly from the sensors. The main reason for our use of 3D models is invariance to large pose changes and more accurate representation of lighting compared to 2D approaches. We do not need to learn models of appearance under different pose and illumination conditions. *This makes our recognition strategy independent of training data needed to learn such models, and allows the gallery and probe conditions to be completely disjoint.*

There are numerous methods for tracking objects in video in the presence of illumination changes [18–22]. However, most of them *compensate* for the illumination conditions of each frame in the video (as opposed to *recovering* the illumination conditions). In [23, 24], the authors independently derived a low order (9D) spherical harmonics-based linear representation to accurately approximate the reflectance images produced by a Lambertian object with attached shadows. In [24, 25], the authors discussed the advantage of this 3D model-based illumination representation compared to some image-based representations. Their methods work only for a single image of an object that is fixed relative to the camera, and do not account for changes in appearance due to motion. We proposed a framework in [26, 27] for integrating the spherical harmonics-based illumination model with the motion of the objects leading to a bilinear model of lighting and motion parameters. In this paper, we show how the theory can be used for video-based face recognition.

1.2. Overview of the approach

The underlying concept of this paper is a method for learning joint illumination and motion models of objects from video. We assume that a 3D model of each face in the gallery is available. For our experiments, the 3D model is estimated from images, but any 3D modeling algorithm, including directly acquiring the model through range sensors, can be used for this purpose. Given a probe sequence, we track the face automatically in the video sequence under arbitrary pose and illumination conditions using the bilinear model of the illumination and motion we developed before [27]. This is achieved by a new inverse compositional estimation approach leading to real-time performance [28]. The illumination invariant model-based tracking algorithm allows us not only to estimate the 3D motion, but also to *recover* the illumination conditions as a function of time. The learned illumination parameters are used to synthesize video sequences for each gallery under the motion and illumination conditions in the probe. The distance between the probe and synthesized sequences is then computed for each frame. Different distance measurements are explored for this purpose. Next, the synthesized sequence that is at a minimum distance from the probe sequence is computed and is declared to be the identity of the person.

Experimental evaluation is carried out on a database of 57 people that we collected for this purpose. We compare our approach against other image-based and video-based face recognition methods. One of the challenges in video-based face recognition is the lack of a good dataset, unlike in image-based approaches [1]. The dataset in [11] is small and consists mostly of pose variations. The dataset described in [29] has large pose variations under constant illumination, and illumination changes in (mostly) fixed frontal/profile poses (these are essentially for gait analysis). The XM2VTS dataset (http://www.ee.surrey.ac.uk/CVSSP/xm2vtsdb/) does not have any illumination variations, which is one of the main contributions of our work. An ideal dataset for us would be similar to the CMU PIE dataset [9], but with video sequences instead of discrete poses. This is the reason why we collected our own data, which has large, simultaneous pose, illumination, and expression variations. It is similar to the PIE dataset though the illumination change is random and uses pre-existing and natural indoor and outdoor lighting.

1.3. Contributions

The following are the main contributions of the paper.

(i) We propose an *analysis-by-synthesis* framework for video-based face recognition that can work with large pose and illumination changes that are normal in natural imagery.

(ii) We propose a novel, inverse compositional (IC) approach for estimating 3D pose, and lighting conditions in the video sequence. Unlike existing methods [30], our warping function involves a 2D → 3D → 2D transformation. Our method allows us to estimate the motion and lighting in real-time.

(iii) We propose different metrics to obtain the identity of the individual in a probe sequence by integrating over the entire video and compare their merits and demerits.

(iv) Our overall strategy does not require learning an appearance variation model, unlike many existing methods [10–12, 14–16]. Thus, the proposed strategy is not dependent on the quality of the learned appearance model and can handle situations where the pose and illumination conditions in the probe are completely independent of the gallery and training data.

(v) We perform a thorough evaluation of our method against well-known image-based approaches like Kernel PCA + LDA [31] and 3D model-based approaches like 3DMM [4, 5].

2. LEARNING JOINT ILLUMINATION AND MOTION MODELS FROM VIDEO

2.1. Bilinear model of the motion and illumination

In this section, we will briefly review the main results in [27] helping to lay the background and notation for this paper. It was proved that if the motion of the object (defined as the translation of the object centroid $\mathbf{\Delta T} \in \mathbb{R}^3$ and the rotation $\mathbf{\Delta\Omega} \in \mathbb{R}^3$ about the centroid in the camera frame) from time t_1 to new time instance $t_2 = t_1 + \delta t$ is small, then up to a first order approximation, the reflectance image $I(x, y)$ at t_2 can be expressed as

$$\begin{aligned} I_{t_2}(\mathbf{u}) &= \sum_{i=1}^{9} l^i b^i_{t_2}(\mathbf{u}),\ b^i_{t_2}(\mathbf{u}) \\ &= b^i_{t_1}(\mathbf{u}) + \mathbf{A}(\mathbf{u},\mathbf{n})\mathbf{\Delta}T + \mathbf{B}(\mathbf{u},\mathbf{n})\mathbf{\Delta\Omega}. \end{aligned} \tag{1}$$

In the above equations, $\mathbf{u}$ represents the image point projected from the 3D surface with surface normal $\mathbf{n}$ (see Figure 1), and $b^i_{t_1}(\mathbf{u})$ are the original basis images before motion. $\mathbf{A}$ and $\mathbf{B}$ contain the structure and camera intrinsic parameters, and are functions of $\mathbf{u}$ and the 3D surface normal $\mathbf{n}$. For each pixel $\mathbf{u}$, both $\mathbf{A}$ and $\mathbf{B}$ are $N_l \times 3$ matrices, where $N_l \approx 9$ for Lambertian objects with attached shadows. Please refer to [26] for the derivation of (1) and explicit expression for $\mathbf{A}$ and $\mathbf{B}$. From (1), we see that the new image spans a bilinear space of six motion and approximately nine illumination variables (for Lambertian objects with attached shadows). The basic result is valid for general illumination conditions, but requires consideration of higher order spherical harmonics.

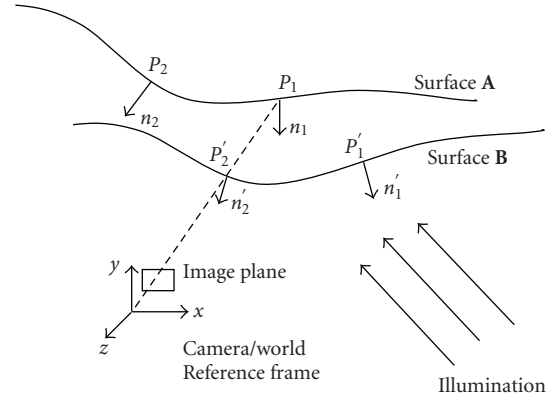


Figure 1: Pictorial representation showing the motion of the object and its projection (reproduced from [26]).

We can express the result in (1) succinctly using tensor notation as

$$\mathcal{I}_{t_2} = \left(\mathcal{B}_{t_1} + \mathcal{C}_{t_1} \times_2 \begin{pmatrix} \mathbf{\Delta T} \\ \mathbf{\Delta\Omega} \end{pmatrix}\right) \times_1 \mathbf{l}, \tag{2}$$

where $\times_n$ is called the *mode-n product* [32] and $\mathbf{l} \in \mathbb{R}^{N_l}$, is the vector of l_i components. The *mode-n product* of a tensor $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_n \times \cdots \times I_N}$ by a vector $\mathbf{V} \in \mathbb{R}^{1 \times I_n}$, denoted by $\mathcal{A} \times_n \mathbf{V}$, is the $I_1 \times I_2 \times \cdots \times 1 \times \cdots \times I_N$ tensor

$$(\mathcal{A} \times_n \mathbf{V})_{i_1 \cdots i_{n-1} 1 i_{n+1} \cdots i_N} = \sum_{i_n} a_{i_1 \cdots i_{n-1} i_n i_{n+1} \cdots i_N} v_{i_n}. \tag{3}$$

For each pixel (p, q) in the image, $\mathcal{C}_{klpq} = [\,\mathbf{A}\ \ \mathbf{B}\,]$ of size $N_l \times 6$. Thus for an image of size $M \times N$, $\mathcal{C}$ is $N_l \times 6 \times M \times N$. $\mathcal{B}_{t_1}$ is a subtensor of dimension $N_l \times 1 \times M \times N$, comprising the basis images $b^i_{t_1}(\mathbf{u})$, and $\mathcal{I}_{t_2}$ is a subtensor of dimension $1 \times 1 \times M \times N$, representing the image.

2.2. Pose and illumination estimation

Equation (2) provides us an expression relating the reflectance image $\mathcal{I}$ with the illumination coefficients $\mathbf{l}$ and motion variables $\mathbf{\Delta T}$, $\mathbf{\Delta\Omega}$. Letting $\mathbf{m} = \left(\begin{smallmatrix} \mathbf{\Delta T} \\ \mathbf{\Delta\Omega} \end{smallmatrix}\right)$, we have a method for estimating 3D motion and illumination as

$$(\hat{\mathbf{l}}_{t_2}, \hat{\mathbf{m}}_{t_2}) = \arg\min_{\mathbf{l},\mathbf{m}} \left|\left|\mathcal{I}_{t_2} - (\mathcal{B}_{t_1} + \mathcal{C}_{t_1} \times_2 \mathbf{m}) \times_1 \mathbf{l}\right|\right|^2 + \alpha \|\mathbf{m}\|^2, \tag{4}$$

where $\hat{x}$ denotes an estimate of x. Since the motion between consecutive frames is small, but illumination can change suddenly, we add a regularization term to the above cost function with the form of $\alpha\|\mathbf{m}\|^2$.

Since the image $\mathcal{I}_{t_2}$ lies approximately in a bilinear space of illumination and motion variables with the bases $\mathcal{B}_{t_1}$ and

$\mathcal{C}_{t_1}$ computed at the pose close to that of $\mathcal{I}_{t_2}$ (ignoring the regularization term for now), such a minimization problem can be achieved by alternately estimating the motion and illumination parameters with the bases $\mathcal{B}_{t_1}$ and $\mathcal{C}_{t_1}$ at the pose of the previous iteration. This process guarantees convergence to a local minimum. Assuming that we have tracked the sequence up to some frame for which we can estimate the motion (hence, pose) and illumination, we calculate the basis images, $b^i_{t_1}$, at the current pose and write it in tensor form $\mathcal{B}_{t_1}$. Similarly, we can also obtain $\mathcal{C}_{t_1}$ at the pose. (Assume an Nth-order tensor $\mathcal{A} \in \mathbf{C}^{I_1 \times I_2 \times \cdots \times I_N}$. The matrix unfolding $\mathbf{A}_{(n)} \in \mathbf{C}^{I_n \times (I_{n+1} I_{n+2} \cdots I_N I_1 I_2 \cdots I_{n-1})}$ contains the element $a_{i_1 i_2 \cdots i_N}$ at the position with row number i_n and column number equal to $(i_{n+1}-1) I_{n+2} I_{n+3} \cdots I_N I_1 I_2 \cdots I_{n-1} + (i_{n+2}-1) I_{n+3} I_{n+4} \cdots I_N I_1 I_2 \cdots I_{n-1} + \cdots + (i_N - 1) I_1 I_2 \cdots I_{n-1} + (i_1 - 1) I_2 I_3 \cdots I_{n-1} + \cdots + i_{n-1}$.) Unfolding $\mathcal{B}_{t_1}$ and the image $\mathcal{I}_{t_2}$ along the first dimension, [32] which is the illumination dimension, the image can be represented as

$$\mathcal{I}^T_{t_2(1)} = \mathcal{B}^T_{t_1(1)} \mathbf{l}. \tag{5}$$

This is a least squares problem, and the illumination $\mathbf{l}$ can be estimated as

$$\hat{\mathbf{l}} = \left(\mathcal{B}_{t_1(1)} \mathcal{B}^T_{t_1(1)}\right)^{-1} \mathcal{B}_{t_1(1)} \mathcal{I}^T_{t_2(1)}. \tag{6}$$

Keeping the illumination coefficients fixed, the bilinear space in (2) becomes a linear subspace, that is,

$$\mathcal{I}_{t_2} = \mathcal{B}_{t_1} \times_1 \mathbf{l} + \mathcal{G} \times_2 \mathbf{m}, \quad \text{where } \mathcal{G} = \mathcal{C}_{t_1} \times_1 \mathbf{l}, \tag{7}$$

and motion $\mathbf{m}$ can be estimated as

$$\hat{\mathbf{m}} = \left(\mathcal{G}_{(2)} \mathcal{G}^T_{(2)} + \alpha \mathbf{I}\right)^{-1} \mathcal{G}_{(2)} \left(\mathcal{I}_{t_2} - \mathcal{B}_{t_1} \times_1 \mathbf{l}\right)^T_{(2)}, \tag{8}$$

where $\mathbf{I}$ is an identity matrix of dimension 6×6.

2.3. Inverse compositional (IC) pose and illumination estimation

The iteration involving alternate minimization over motion and illumination in the above approach is essentially a gradient descent method. In each iteration, as pose is updated, the gradients (i.e., the tensors $\mathcal{B}$ and $\mathcal{C}$) need to be recomputed, which is computationally expensive. The inverse compositional algorithm [30] works by moving these computational steps out of the iterative updating process.

Consider an input frame $I_{t_2}(\mathbf{u})$ at time instance t_2 with image coordinate $\mathbf{u}$. We introduce a warp operator $\mathbf{W}_\mathbf{p} : \mathbb{R}^2 \to \mathbb{R}^2$ such that, if the pose of $I_{t_2}(\mathbf{u})$ is $\mathbf{p}$, the pose of $I_{t_2}(\mathbf{W}_\mathbf{p}(\mathbf{u}, \mathbf{m}))$ is $\mathbf{p}+\mathbf{m}$ (see Figure 2). Basically, $\mathbf{W}_\mathbf{p}$ represents the displacement in the image plane due to a pose transformation of the 3D model. Denote the pose transformed image $I_{t_2}(\mathbf{W}_{\hat{\mathbf{p}}_{t_1}}(\mathbf{u}, \mathbf{m}))$ in tensor notation $\tilde{\mathcal{I}}^{\mathbf{W}_{\hat{\mathbf{p}}_{t_1}}(\mathbf{m})}_{t_2}$. Using this warp operator and ignoring the regularization term, we can restate the cost function (4) in the inverse compositional framework as

$$(\hat{\mathbf{l}}_{t_2}, \hat{\mathbf{m}}_{t_2}) = \arg \min_{\mathbf{l},\mathbf{m}} \left\| \tilde{\mathcal{I}}^{\mathbf{W}_{\hat{\mathbf{p}}_{t_1}}(-\mathbf{m})}_{t_2} - \mathcal{B}_{t_1} \times_1 \mathbf{l} \right\|^2. \tag{9}$$

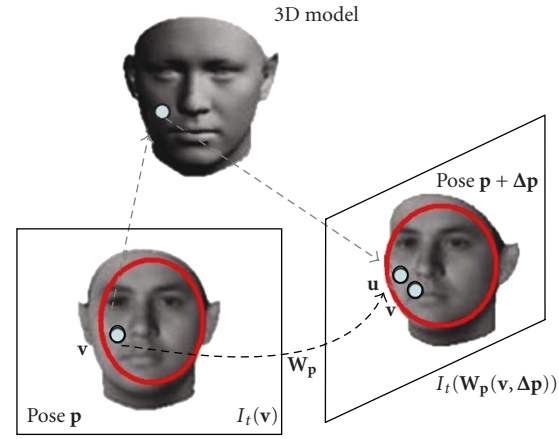


Figure 2: Illustration of the warping function $\mathbf{W}$. A point $\mathbf{v}$ in image plane is projected onto the surface of the 3D object model. After the pose transformation with $\Delta\mathbf{p}$, the point on the surface is back-projected onto the image plane at a new point $\mathbf{u}$. The warping function maps from $\mathbf{v} \in \mathbb{R}^2$ to $\mathbf{u} \in \mathbb{R}^2$. The red ellipses show the common part in both frames that the warping function $\mathbf{W}$ is defined upon.

This cost function can be minimized over $\mathbf{m}$ by iteratively solving for increments $\Delta\mathbf{m}$ in

$$\left\| \tilde{\mathcal{I}}^{\mathbf{W}_{\hat{\mathbf{p}}_{t_1}}(-\mathbf{m})}_{t_2} - (\mathcal{B}_{t_1} + \mathcal{C}_{t_1} \times_2 \Delta\mathbf{m}) \times_1 \mathbf{l} \right\|^2. \tag{10}$$

In each iteration, $\mathbf{m}$ is updated such that $\mathbf{W}_{\hat{\mathbf{p}}_{t_1}}(\mathbf{u}, -\mathbf{m}) \leftarrow \mathbf{W}_{\hat{\mathbf{p}}_{t_1}}(\mathbf{u}, -\mathbf{m}) \circ \mathbf{W}_{\hat{\mathbf{p}}_{t_1}}(\mathbf{u}, \Delta\mathbf{m})^{-1}$. (The compositional operator $\circ$ means the second warp is composed into the first warp, that is, $\mathbf{W}_{\hat{\mathbf{p}}_{t_1}}(\mathbf{u}, -\mathbf{m}) \equiv \mathbf{W}_{\hat{\mathbf{p}}_{t_1}}(\mathbf{W}_{\hat{\mathbf{p}}_{t_1}}(\mathbf{u}, \Delta\mathbf{m})^{-1}, -\mathbf{m}.)$) (The inverse of the warp $\mathbf{W}$ is defined to be the $\mathbb{R}^2 \to \mathbb{R}^2$ mapping such that if we denote the pose of $I_t(\mathbf{v})$ as $\mathbf{p}$, the pose of $I_t(\mathbf{W}_\mathbf{p}(\mathbf{W}_\mathbf{p}(\mathbf{v}, \Delta\mathbf{p}), \Delta\mathbf{p})^{-1})$ is $\mathbf{p}$ itself. As the warp $\mathbf{W}_\mathbf{p}(\mathbf{v}, \Delta\mathbf{p})$ transforms the pose from $\mathbf{p}$ to $\mathbf{p} + \Delta\mathbf{p}$, the inverse $\mathbf{W}_\mathbf{p}(\mathbf{v}, \Delta\mathbf{p})^{-1}$ should transform the pose from $\mathbf{p} + \Delta\mathbf{p}$ to $\mathbf{p}$, that is, $\mathbf{W}_\mathbf{p}(\mathbf{v}, \Delta\mathbf{p})^{-1} = \mathbf{W}_{\mathbf{p}+\Delta\mathbf{p}}(\mathbf{v}, -\Delta\mathbf{p})$. Thus $\{\mathbf{W}_\mathbf{p}\}$ is a group.) Using the additivity of pose transformation for small $\Delta\mathbf{m}$, $\mathbf{W}_{\hat{\mathbf{p}}_{t_1}}(\mathbf{W}_{\hat{\mathbf{p}}_{t_1}}(\mathbf{u}, \Delta\mathbf{m})^{-1}, -\mathbf{m}) = \mathbf{W}_{\hat{\mathbf{p}}_{t_1}}(\mathbf{W}_{\hat{\mathbf{p}}_{t_1}+\Delta\mathbf{m}}(\mathbf{u}, -\Delta\mathbf{m}), -\mathbf{m}) = \mathbf{W}_{\hat{\mathbf{p}}_{t_1}+\Delta\mathbf{m}}(\mathbf{u}, -\Delta\mathbf{m} - \mathbf{m}) \approx \mathbf{W}_{\hat{\mathbf{p}}_{t_1}}(\mathbf{u}, -\Delta\mathbf{m} - \mathbf{m})$. Thus, the above update is essentially $\mathbf{m} \leftarrow \mathbf{m} + \Delta\mathbf{m}$.

For the inverse compositional algorithm to be provably equivalent to the Lucas-Kanade algorithm up to a first order approximation of $\Delta\mathbf{m}$, the set of warps $\{\mathbf{W}_{\hat{\mathbf{p}}_{t_1}}\}$ must form a group, that is, every warp $\mathbf{W}_{\hat{\mathbf{p}}_{t_1}}$ must be invertible. If the change of pose is small enough, the visibility for most of the pixels will remain the same—thus $\mathbf{W}_{\hat{\mathbf{p}}_{t_1}}$ can be considered approximately invertible. However, if the pose change becomes too big, some portion of the object will become invisible after the pose transformation, and $\mathbf{W}_{\hat{\mathbf{p}}_{t_1}}$ will no longer be invertible. A detailed proof of convergence is available in [28].

We select a set of poses $\{\mathbf{p}_j\}$ with interval of 20 degrees in pan and tilt angles, and precompute the basis $\mathcal{B}$ and $\mathcal{C}$ at

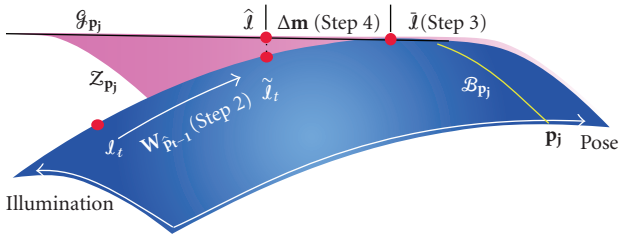


FIGURE 3: Pictorial representation of the inverse compositional tracking scheme. Starting with $\mathcal{I}_t$, we first warp it to $\tilde{\mathcal{I}}_t$ as in Step 2 below. This allows computation of the bases of the joint pose and illumination manifold at the cardinal pose p_j. Then, we search along the illumination dimension of this manifold to get the illumination estimate that best describes $\tilde{\mathcal{I}}_t$. This is Step 3. Then, in Step 4, $\tilde{\mathcal{I}}_t$ is projected onto the tangent plane of the manifold where the motion estimates was obtained.

these poses. We call these poses as cardinal poses. All frames that are close to a particular pose $\mathbf{p}_j$ will use the $\mathcal{B}$ and $\mathcal{C}$ at that pose, and the warp $\mathbf{W}_{\hat{\mathbf{p}}_{t_1}}$ should be performed to normalize the pose to $\mathbf{p}_j$. The pictorial representation of the inverse compositional tracking scheme is shown in Figure 3. While most of the existing inverse compositional methods move the expensive update steps out of the iterations for two-frame matching, we go even further and perform these expensive computations only once every few frames. This is by virtue of the fact that we estimate 3D motion.

2.4. The IC pose and illumination estimation algorithm

Consider a sequence of image frames $\mathcal{I}_t$, $t = 0, \ldots, N-1$. In keeping with standard notation used in tracking, we assume $\delta t = 1$, and consider two frames at t and $t-1$.

Assume that we know the pose and illumination estimates for frame $t-1$, that is, $\hat{\mathbf{p}}_{t-1}$ and $\hat{\mathbf{l}}_{t-1}$.

Step 1. For the new input frame $\mathcal{I}_t$, find the closest $\mathbf{p_j}$ to the pose estimates at $t-1$, that is, $\hat{\mathbf{p}}_{t-1}$. Set $\hat{\mathbf{m}}_t$ to be 0.

Step 2. Apply the pose transformation operator $\mathbf{W}_{\hat{\mathbf{p}}_{t-1}}$ to get the pose normalized version of the frame $\tilde{\mathcal{I}}_t^{\mathbf{W}_{\hat{\mathbf{p}}_{t-1}}(\mathbf{p_j}-\hat{\mathbf{p}}_{t-1}-\hat{\mathbf{m}}_t)}$s, that is, $I(\mathbf{W}_{\hat{\mathbf{p}}_{t-1}}(\mathbf{u}, \mathbf{p_j} - \hat{\mathbf{p}}_{t-1} - \hat{\mathbf{m}}_t), t)$.

Step 3. Use

$$\hat{\mathbf{l}}_t = \left(\mathcal{B}_{\mathbf{p}_{j|t-1}} \mathcal{B}_{\mathbf{p}_{j|t-1}}^{\mathrm{T}}\right)^{-1} \mathcal{B}_{\mathbf{p}_{j|t-1}}^{\mathrm{T}} \tilde{\mathcal{I}}_{t(1)}^{\mathbf{W}_{\hat{\mathbf{p}}_{t-1}}(\mathbf{p_j}-\hat{\mathbf{p}}_{t-1}-\hat{\mathbf{m}}_t)} \quad (11)$$

to estimate $\hat{\mathbf{l}}_t$ of the pose normalized image $\tilde{\mathcal{I}}_t^{\mathbf{W}_{\hat{\mathbf{p}}_{t-1}}(\mathbf{p_j}-\hat{\mathbf{p}}_{t-1}-\hat{\mathbf{m}}_t)}$.

Step 4. With the estimated $\hat{\mathbf{l}}_t$ from Step 3, use

$$\Delta\hat{\mathbf{m}} = \left[\mathcal{G}_{\mathbf{p_j}} \mathcal{G}_{\mathbf{p_j}}^{\mathrm{T}}\right]^{-1} \mathcal{G}_{\mathbf{p_j}} \left(\tilde{\mathcal{I}}_t^{\mathbf{W}_{\hat{\mathbf{p}}_{t-1}}(\mathbf{p_j}-\hat{\mathbf{p}}_{t-1}-\hat{\mathbf{m}}_t)} - \mathcal{B}_{\mathbf{p}_{j|t-1}} \times_1 \hat{\mathbf{l}}_t\right) \quad (12)$$

to estimate the motion increment $\Delta\mathbf{m}$, where

$$\mathcal{G}_{\mathbf{p_j}} = \mathcal{C}_{\mathbf{p}_{j|t-1}} \times_1 \hat{\mathbf{l}}_t. \quad (13)$$

Update $\hat{\mathbf{m}}_t$ with $\hat{\mathbf{m}}_t \leftarrow \hat{\mathbf{m}}_t + \Delta\mathbf{m}$.

Step 5. Repeat Steps 2, 3, and 4 for that input frame till the difference error ε between the pose normalized image $\tilde{\mathcal{I}}_t^{\mathbf{W}_{\hat{\mathbf{p}}_{t-1}}(\mathbf{p_j}-\hat{\mathbf{p}}_{t-1}-\hat{\mathbf{m}}_t)}$ and the rendered image $(\mathcal{B}_{\mathbf{p}_{j|t-1}} + \mathcal{C}_{\mathbf{p}_{j|t-1}} \times_2 \hat{\mathbf{m}}_t) \times_1 \hat{\mathbf{l}}_t$ can be reduced below an acceptable threshold. This gives $\hat{\mathbf{l}}_t$ and $\hat{\mathbf{m}}_t$ of (4).

Step 6. Set $t = t + 1$. Repeat Steps 1, 2, 3, 4, and 5. Continue till $t = N - 1$.

3. FACE RECOGNITION FROM VIDEO

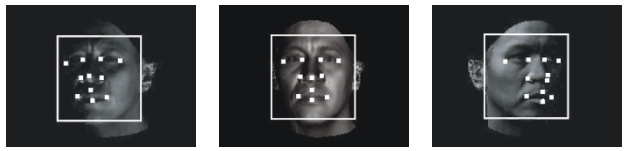
We now explain the face recognition algorithm and analyze the importance of different measurements for integrating the recognition performance over a video sequence. In our method, the gallery is represented by a textured 3D model of the face. The model can be built from a single image [33], a video sequence [34] or obtained directly from 3D sensors [17]. In our experiments, the face model will be estimated from a gallery video sequence for each individual. Face texture is obtained by normalizing the illumination of the first frame in the gallery sequence to an ambient condition, and mapping it onto the 3D model. Given a probe sequence, we will estimate the motion and illumination conditions using the algorithms described in Section 2.2. Note that the tracking does not require a person-specific 3D model—a generic face model is usually sufficient. Given the motion and illumination estimates, we will then render images from the 3D models in the gallery. The rendered images can then be compared with the images in the probe sequence. For this purpose, we will design robust measurements for comparing these two sequences. A feature of these measurements will be their ability to integrate the identity over all the frames, ignoring some frames that may have the wrong identity.

Let I_i, $i = 0, \ldots, N-1$ be the ith frame from the probe sequence. Let $S_{i,j}$, $i = 0, \ldots, N-1$ be the frames of the synthesized sequence for individual j, where $j = 1, \ldots, M$ and M is the total number of individuals in the gallery. Note that the number of frames in the two sequences to be compared will always be the same in our method. By design, each corresponding frame in the two sequences will be under the same pose and illumination conditions, dictated by the accuracy of the estimates of these parameters from the probes sequences. Let d_{ij} be the Euclidean distance between the ith frames I_i and $S_{i,j}$. We now compare three distance measures that can be used for obtaining the identity of the probe sequence:

$$(1) \quad \mathrm{ID} = \arg\min_j \min_i d_{ij}, \quad (14)$$

$$(2) \quad \mathrm{ID} = \arg\min_j \max_i d_{ij}, \quad (15)$$

$$(3) \quad \mathrm{ID} = \arg\min_j \frac{1}{N} \sum_i d_{ij}. \quad (16)$$



(a) (b) (c)

Figure 4: The back projection of the feature points on the generated 3D face model using the estimated 3D motion onto some input frames.

The first alternative computes the distance between the frames in the probe sequence and each synthesized sequence that are the most similar and chooses the identity as the individual with the smallest distance. The second distance measure can be interpreted as minimizing the maximum separation between the frames in the probe sequence and synthesized sequences. Both of these measures suffer from a lack of robustness, which can be critical for their performance since the correctness of the frames in the synthesized sequences depends upon the accuracy of the illumination and motion parameter estimates. For this purpose, we replace the max by the fth percentile and the min (in the inner distance computation of 1) by the $(1-f)$th percentile. In our experiments, we choose f to be 0.8.

The third option (16) chooses the identity as the minimum mean distance between the frames in the probe sequence and each synthesized sequence. Under the assumptions of Gaussian noise and uncorrelatedness between frames, this can be interpreted as choosing the identity with the maximum a-posterior probability given the probe sequence.

As the images in the synthesized sequences are pose and illumination normalized to the ones in the probe sequence, d_{ij} can be computed directly using the Euclidean distance. Other distance measurements, like [14, 35], can be considered in situations where the pose and illumination estimates may not be reliable or in the presence of occlusion and clutter. We will look into such issues in our future work.

3.1. Video-based face recognition algorithm

Using the above notation, let I_i, $i = 0, \ldots, N-1$ be N frames from the probe sequence. Let $G_1, \ldots, G_M$ be the 3D models with texture for each of M galleries.

Step 1. Register a 3D generic face model to the first frame of the probe sequence. This is achieved using the method in [36]. Estimate the illumination and motion model parameters for each frame of the probe sequence using the method described in Section 2.4

Step 2. Using the estimated illumination and motion parameters, synthesize, for each gallery, a video sequence using the generative model of (1). Denote these as $S_{i,j}$, $i = 1, \ldots, N$ and $j = 1, \ldots, M$.

Step 3. Compute d_{ij} as above.

Step 4. Obtain the identity using a suitable distance measure as in (14) or (15) or (16).

4. EXPERIMENTAL RESULTS

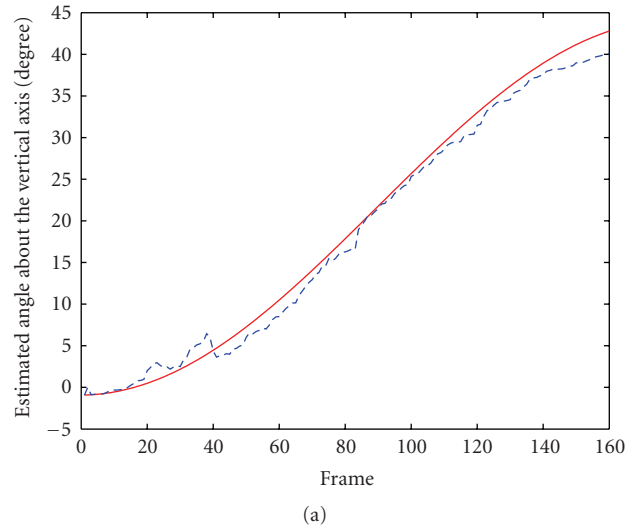
4.1. Accuracy of tracking and illumination estimation

We will first show some results on the accuracy of tracking and illumination estimation with known ground truth. This is because of the critical importance of this step in our proposed recognition scheme. We use the 3DMM [33] to generate a face. The generated face model is rotated along the vertical axis at some specific angular velocity, and the illumination is changing both in direction (from right-bottom corner to the left-top corner) and in brightness (from dark to bright to dark). In Figure 4, the images show the back projection of some feature points on the 3D model onto the input frames using the estimated motion under three different illumination conditions. In Figure 5, (a) shows the comparison between the estimated motion (in blue) and the ground truth (in red). The maximum error in pose estimates is 2.53° and the average error is 0.67°. Figure 5(b) shows the norm of the error between the ground truth illumination coefficients and the estimated ones, normalized with the ground truth. The maximum error is 4.93% and the average is 4.1%.

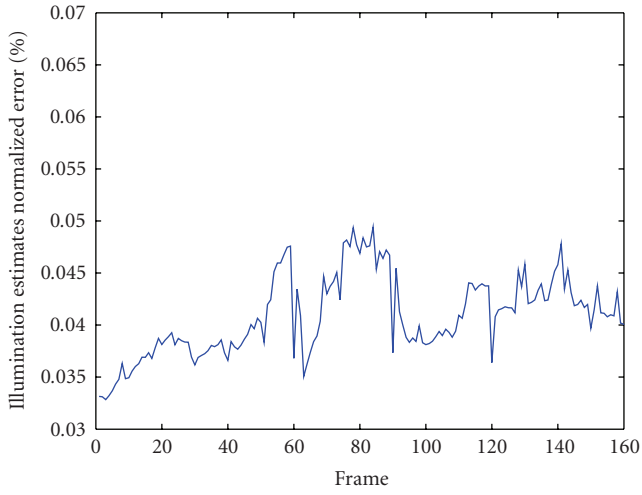
The results on tracking and synthesis on two of the probe sequences in our database (described next) are shown in Figure 6. The inverse compositional tracking algorithm can track about 20 frames per second on a standard PC using a MATLAB implementation. Real-time tracking could be achieved through better software and hardware optimization.

4.2. Face database and experimental setup

Our database consists of videos of 57 people. Each person was asked to move his/her head as they wished (mostly rotate their head from left to right, and then from down to up), and the illumination was changed randomly. The illumination consisted of ceiling lights, lights from the back of the head and sunlight from a window on the left side of the face. Random combinations of these were turned on and off and the window was controlled using dark blinds. There was no control over how the subject moves his/her head or on facial expression. Sample frames of these video sequences are shown in Figure 7. The images are scale normalized and centered. Some of the subjects had expression changes also, for example, the last row of the Figure 7. The average size of the face was about 70×70 with the minimum size being 50×50. Videos are captured with uniform background. We recorded 2 to 3 sessions of video sequences for each individual. All the video sessions are recorded within one week. The first session is used as the gallery for constructing the 3D textured model of the head, while the remaining are used for testing. We used a simplified version of the method in [34] for this purpose. We would like to emphasize that any other 3D modeling algorithm would also have worked. Texture is obtained by normalizing the illumination



(a)



(b)

Figure 5: (a) 3D estimates (blue) and ground truth (red) of pose against frames. (b) The normalized error of the illumination estimates versus frame numbers.

Figure 6: Original images, tracking and synthesis results are shown in three successive rows for two of the probe sequences.

Figure 7: Sample frames from the video sequences collected for our database (best viewed on a monitor).

of the first frame in each gallery sequence to an ambient illumination condition and mapping onto the 3D model.

As can be seen from Figure 7, the pose and illumination vary randomly in the video. For each subject, we designed three experiments by choosing different probe sequences.

Experiment A

A video was used as the probe sequence with the average pose of the face in the video being about 15° from frontal.

Experiment B

A video was used as the probe sequence with the average pose of the face in the video being about 30° from frontal.

Experiment C

A video was used as the probe sequence with the average pose of the face in the video being about 45° from frontal.

Each probe sequence has about 20 frames around the average pose. The variation of pose in each sequence was less than 15°, so as to keep pose in the experiments disjoint. The probe sequences are about 5 seconds each. This is because we wanted to separate the probes based on pose of the head (every 15 degrees) and it does not take the subject more than 5 seconds to move 15 degrees when continuously rotating the head. To show the benefit of video-based methods over image-based approaches, we designed three new experiments: D, E, and F by taking random single images from A, B, and C, respectively.

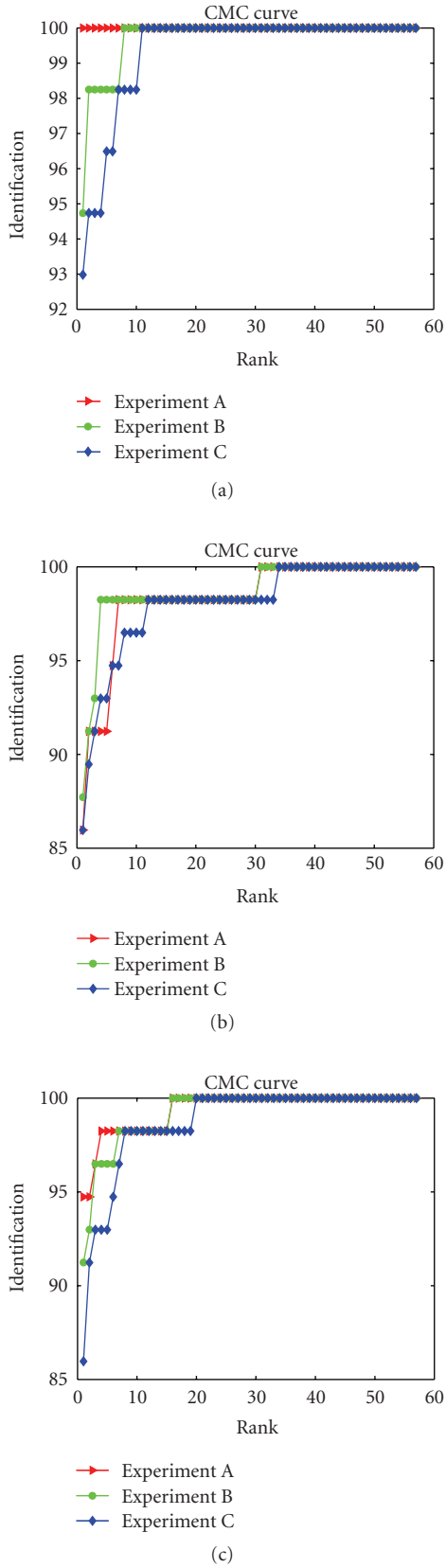


Figure 8: CMC curve for video-based face recognition experiments A to C; (a) with distance measurement 1 in (14), (b) with distance measurement 2 in (15), and (c) with distance measurement 3 in (16).

4.3. Recognition results

We plot the cumulative match characteristic (CMC) [1, 2] for experiments: A, B, and C with measurement 1 (14), measurement 2 (15), and measurement 3 (16) in Figure 8. In experiment A, where pose is 15°away from frontal, all the videos with large and arbitrary variations of illumination are recognized correctly. In experiment B, we achieve about 95% recognition rate, while for experiment C it is 93% using the distance measure (14). Irrespective of the illumination changes, the recognition rate decreases consistently with large difference in pose from frontal (which is the gallery), a trend that has been reported by other authors [4, 5]. *Note that the pose and illumination conditions in the probe and gallery sets can be completely disjoint.*

4.4. Performance analysis

Performance with changing average pose

Figures 8(a), 8(b), and 8(c) show the recognition rate with the measurements in (14), (15), and (16). Measurement 1 in (14) gives the best result. This is consistent with our expectation, as (14) is not affected by the few frames in which the motion and illumination estimation error is relatively high. The recognition result is affected mostly by registration error which increases with nonfrontal pose (i.e., A→B→C). On the other hand, measurement 2 in (15) is mostly affected by the errors in the motion and illumination estimation and registration, and thus the recognition rate in Figure 8(b) is lower than that of Figure 8(a). Ideally, measurement 3 should give the best recognition rate as this is the MAP estimation. However, the assumptions of Gaussianity and uncorrelatedness may not be valid. This affects the recognition rate for measurement 3, causing it perform worse than measurement 1 (14) but better than measurement 2 (15). We also found that small errors in 3D shape estimation have negligible impact on the motion and illumination estimates and the overall recognition result.

Effect of registration and tracking errors

There are two major error sources: registration and motion/illumination estimation. The error in registration may affect the motion and illumination estimation accuracy in subsequent frames, while robust motion and illumination estimation may regain tracking back after some time, if the registration errors are small.

In Figures 9(a), 9(b), and 9(c), we show the plots of error curves under three different cases. Figure 9(a) is the ideal case, in which the registration is accurate and the error in motion and illumination estimation is consistently small through the whole sequence. The distance d_{ik} from the probe sequence I_i with the true identity k to the synthesized sequence with the correct model $S_{i,k}$, will always be smaller than d_{ij}, $j = 1, \ldots, k-1, k+1, \ldots, M$. In this case, all the measurements 1, 2, and 3 in (14), (15) or (16) will work. In the case shown in Figure 9(b), the registration is correct but the error in the motion and illumination estimation accumulates. Finally, the drift error causes d_{ik},

the distance from the probe sequence to the synthesized sequence with the correct model (shown in bold red) to be higher than some other distance d_{ij}, $j \neq k$ (shown in green). In this case, measurement 2 in (15) will be wrong but measurements 1 and 3 in (14) or (16) still work. In Figure 9(c), the registration is not accurate (the error d_{ik} at the first frame is significantly higher than in (a) and (b)), but the motion and illumination estimation is able to regain tracking after a number of frames where the error decreases. Under this case, both measurements 1 and 2 in (14) and (15) will not work, as it is not any individual frame that reveals the true identity, but the behavior of the error over the collection of all frames. Measurement 3 in (16) computes the overall distance by taking every frame into consideration, thus it works in such cases. This shows the importance of using different distance measurements based on the application scenario. Also, the effect of obtaining the identity by integrating over time is seen.

4.5. Comparison with other approaches

The area of video-based face recognition is less standardized than image-based approaches. There is no standard dataset on which both image and video-based methods have been tried, thus we do the comparison on our own dataset. This dataset can be used for such comparison by other researchers in the future.

Comparison with 3DMM-based approaches

3DMM has achieved a significant impact in the face biometrics area, and obtained impressive results in pose and illumination varying face recognition. It is similar to our proposed approach in the sense that both methods are 3D approaches, estimate the pose, illumination, and do synthesis for recognition. However, 3DMM [5] method uses the Phong illumination model, thus it cannot model extended light sources (like the sky) accurately. To overcome this, Samaras and Zhang [4] proposed the 3D shperical harmonics basis morphable model (SHBMM) that integrates the spherical harmonics illumination representation into the 3DMM. Also, 3DMM and SHBMM methods have been applied to single images only. Although it is possible to repeatedly apply 3DMM or SHBMM approach to each frame in the video sequence, it is inefficient. Registration of the 3D model to each frame will be needed, which requires a lot of computation and manual work. None of the existing 3DMM approaches integrate tracking and recognition. Our proposed method, which integrates 3D motion into SHBMM, is a unified approach for modeling lighting and motion in a face video sequence.

Using our dataset, we now compare our proposed approach against the SHBMM method of [4], which was shown, give better results than 3DMM in [5]. We will also compare our results with the published results of SHBMM method [4] in the later part of this section.

Recall that we designed three new experiments: D, E, and F by taking random single images from A, B, and C, respectively. In Figure 10, we plot the CMC curve with measurement 1 in (14) (which has the best performance for experiments: A, B, and C) for the experiments: D, E, and F and compare them with the ones of the experiments: A, B, and C. The image-based approach recognition was

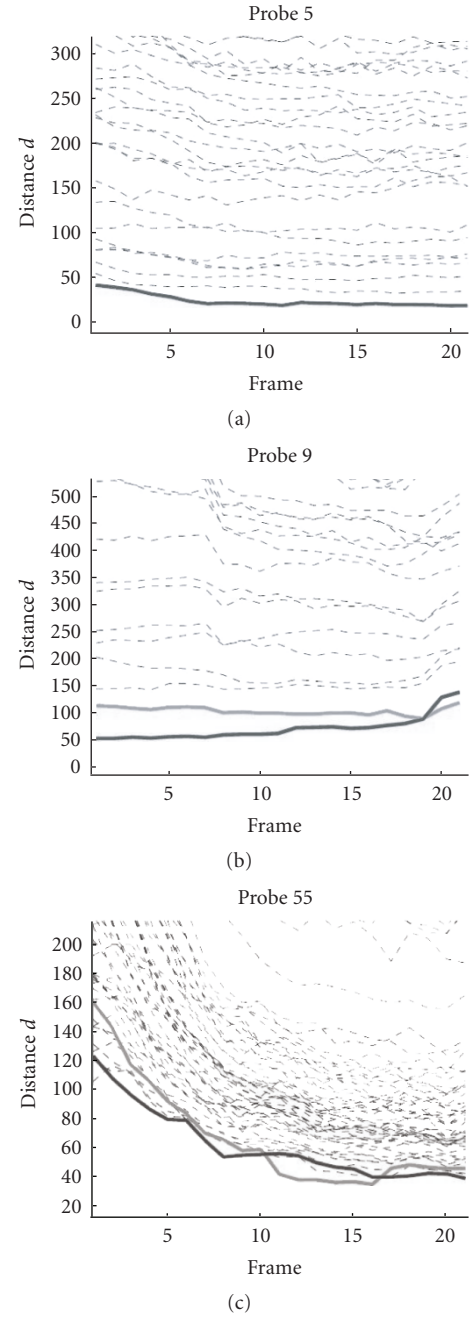


Figure 9: The plots of error curves under three different cases: (a) both registration and motion/illumination estimation are correct, (b) registration is correct but motion/illumination estimation has drift error, and (c) registration is inaccurate, but robust motion/illumination estimation can regain tracking after a number of frames. The black, bold curve shows the distance of the probe sequence with the synthesized sequence of the correct identity, while both the gray bold and dotted curves show the distance with the synthesized sequences using the incorrect identity.

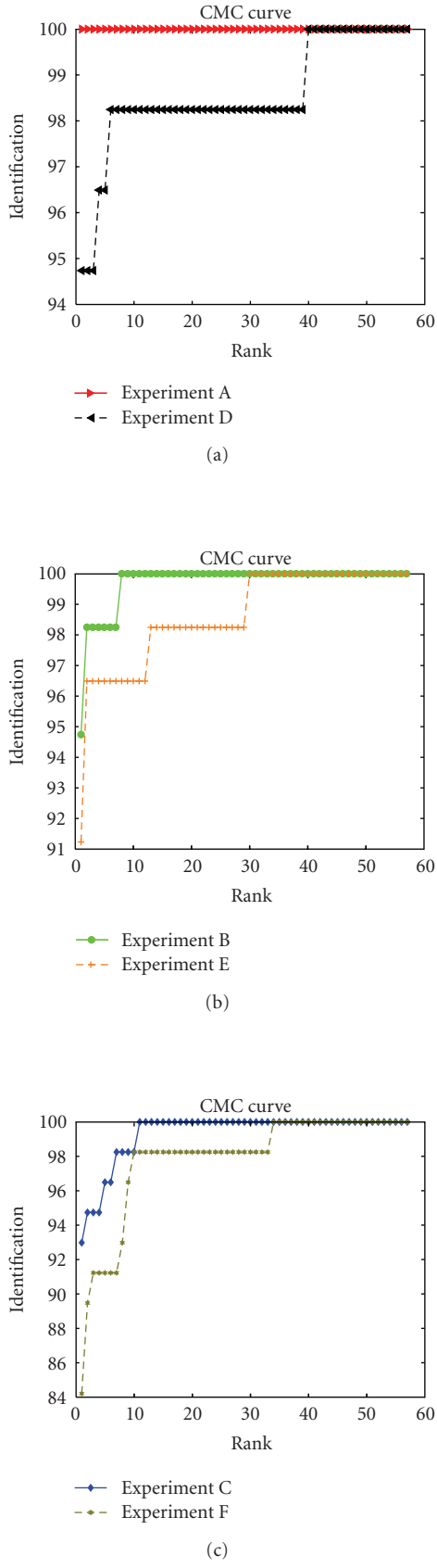


Figure 10: Comparison between the CMC curves for the video-based face experiments A to C with distance measurement 1 against SHBMM method of [4].

achieved by integrating spherical harmonics illumination model with the 3DMM (which is essentially the idea in SHBMM [4]) on our data. For this comparison, we randomly chose images from the probe sequences of experiments: A, B, and C and computed the recognition performance over multiple such random sets. Thus the experiments D, E, and F average the image-based performance over different conditions. By analyzing the plots in Figure 10, we see that the recognition performance with the video-based approach is consistently higher than the image-based one, both in rank 1 performance as well as the area under the CMC curve. This trend is magnified as the average facial pose becomes more nonfrontal. Also, we expect that registration errors, in general, will affect image-based methods more than video-based methods (since robust tracking may be able to overcome some of the registration errors, as shown in Section 4.4).

It is interesting to compare these results against the results in [4], for image-based recognition. The size of the databases in both cases is close (though ours is slightly smaller). Our recognition rate with a video sequence at average 15 degrees facial pose (with a range of 15 degrees about the average) is 100%, while the average recognition rate for approximately 20 degrees (called side view) in [4] is 92.4%. For the experiments B and C, [4] does not have comparable cases and goes directly to profile pose (90 degrees), which we do not have. Our recognition rate at 45° average pose is 93%. In [4], the quoted rates at 20° is 92% and at 90° is 55%. Thus the trend of our video-based recognition results are significantly higher than image-based approaches that deal with both pose and illumination variations.

We would like to emphasize that the above paragraph shows a comparison of recognition rates on two different datasets. While this may not seem completely fair, we are constrained by the lack of a standard dataset on which to compare image- and video-based methods. We have shown a comparison on our dataset using our implementation in Figure 9. The objective of the above paragraph is just to point out some trends with published results on other datasets that do not have video—these should be taken as very definitive statements.

Comparison with 2D approaches

In addition to comparing with 3DMM-based methods, we also do the comparison against traditional 2D methods. We choose the Kernel PCA [31] based approaches as it has performed quite well in many applications. We downloaded the Kernel PCA code from http://asi.insa-rouen.fr/arakotom/toolbox/index.html, and implemented the Kernel PCA with the LDA in MATLAB. In the training phase, we applied KPCA using the polynomial kernel and decrease the dimension of the training samples to 56. Then multiclass LDA is used for separating between different people. For each individual, we use the same images that we used for constructing the 3D shape in our proposed 3D approach as the training set. With this KPCA/LDA approach, we tested

the recognition performance using single frames and the whole video sequences.

When we have a single frame as probe, we use k-Nearest Neighbor for the recognition, while in the case of video sequence, we compute the distance from every frame in the probe sequence to the centroid of the training samples in each class, take the summation over time, and then rank the distance of the sequence to each class. Here, we show the results of recognition with the described 2D approach using single frames and video sequences about 15 degrees (comparable to experiments: A and D), 30 degrees (comparable to experiments: B and E), and 45 degrees (comparable to experiments: C and F) in Figure 11. For the comparison, we also show the results of our approach with video sequences in experiments: A, B, and C. Note that testing frames and sequences are the same as those used in experiments: A/B/C and D/E/F. Since 2D approaches cannot model the pose and illumination variation well, the recognition results are much worse compared to 3D approaches under arbitrary pose and illumination variation. However, we can still see the advantage of integrating the video sequences in Figure 11.

Comparison with 2D illumination methods

The major disadvantage of the 2D illumination methods is that they cannot handle local illumination conditions (lighting coming from some specific direction such that only part of the object is illuminated). In Figure 12, we show the comparison in removing local illumination effects between the spherical harmonics illumination model against the local histogram equalization method. In the three images in Figure 12(a), the top one is the original frame with illumination coming from the left side of the face. The left image in the second row is local histogram equalized, and the right one is resynthesized with the spherical harmonics illumination model with some predefined ambient illumination. In the local histogram equalized image, although the right side of the face is enhanced compared with the original one, the illumination direction can still be clearly perceived. But in the one synthesized with the spherical harmonics illumination model, the direction of illumination is almost completely removed, and no illumination direction information is retained. In Figure 12(b), we show the plot of the error curves of the probe sequence (an image of which is shown in Figure 12(a)) with the local histogram equalization method, while in Figure 12(c) we show the error curves with the method we proposed. It is clear that 3D illumination methods can achieve better results under local illumination conditions.

5. CONCLUSIONS

In this paper, we have proposed an *analysis-by-synthesis* method for video-based face recognition that relies upon a novel theoretical framework for integrating illumination motion and shape models for describing the appearance of a video sequence. We started with a brief exposition of this theoretical result, followed by methods for learning

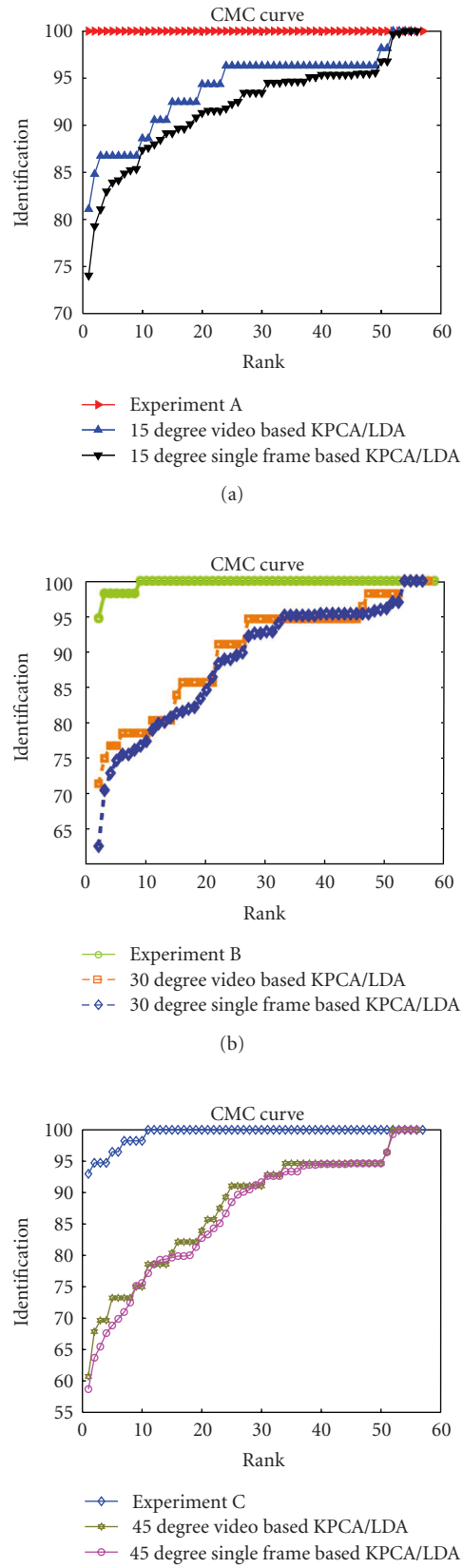


Figure 11: Comparison between the CMC curves for the video-based face experiments A to C with distance measurement 1 in (14) against KPCA+LDA-based 2D approaches.

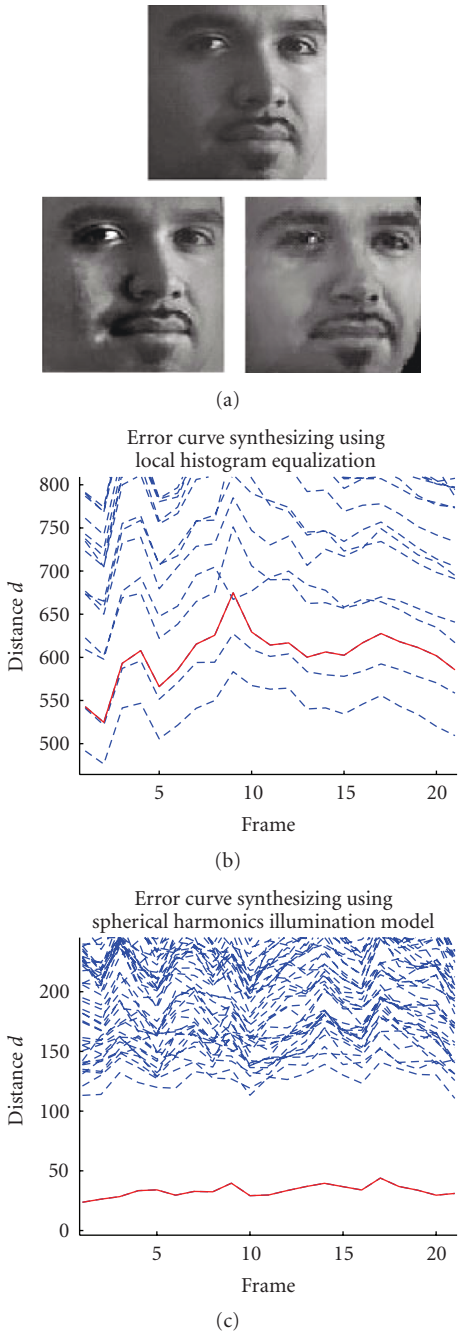


FIGURE 12: The comparison over local illumination effects between the spherical harmonics illumination model and the local histogram equalization method. (a) Top: original image; bottom left: local histogram equalized image; bottom right: synthesis with spherical harmonics illumination model in a predefined ambient illumination. (b) Plots of the error curves using the local histogram equalization. (c) Plots of the error curves using the proposed method. The bold curve is for the face with the correct identity.

the model parameters. Then, we described our recognition algorithm that relies on synthesis of video sequences under the conditions of the probe. We collected a face video database consisting of 57 people with large and arbitrary variation in pose and illumination and demonstrated the effectiveness of the method on this new database. A detailed analysis of performance is also carried out. Future work on video-based face recognition will require experimentation on large datasets, design of suitable metrics, and tight integration of the tracking and recognition phases.

ACKNOWLEDGMENT

Y. Xu and A. Roy-Chowdhury were supported by NSF Grant IIS-0712253.

REFERENCES

[1] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, Face recognition: a literature survey, *ACM Computing Surveys*, vol. 35, no. 4, pp. 399458, 2003.

[2] P. J. Phillips, P. J. Grother, R. J. Micheals, D. M. Blackburn, E. Tabassi, and J. M. Bone, Face recognition vendor test 2002: evaluation report, *Tech. Rep. NISTIR 6965*, National Institute of Standards and Technology, Gaithersburgh, Md, USA, 2003, http://www.frvt.org/.

[3] P. J. Phillips, P. J. Flynn, T. Scruggs, et al., Overview of the face recognition grand challenge, in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '05)*, vol. 1, pp. 947954, San Diego, Calif, USA.

[4] 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*, vol. 28, no. 3, pp. 351363, 2006.

[5] V. Blanz, P. Grother, P. J. Phillips, and T. Vetter, Face recognition based on frontal views generated from non-frontal images, in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '05)*, vol. 2, pp. 454461, San Diego, Calif, USA.

[6] I. Matthews, R. Gross, and S. Baker, Appearance-based face recognition and light-fields, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 4, pp. 449465, 2004.

[7] S. Lucey and T. Chen, Learning patch dependencies for improved pose mismatched face verification, in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '06)*, vol. 1, pp. 909915, New York, NY, USA.

[8] S. J. D. Prince and J. H. Elder, Probabilistic linear discriminant analysis for inferences about identity, in *Proceedings of the IEEE 11th International Conference on Computer Vision (ICCV '07)*, pp. 18, Rio de Janeiro, Brazil, October 2007.

[9] T. Sim, S. Baker, and M. Bsat, The CMU pose, illumination, and expression database, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 12, pp. 16151618, 2003.

[10] M. A. O. Vasilescu and D. Terzopoulos, Multilinear independent components analysis, in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '05)*, vol. 1, pp. 547553, San Diego, Calif, USA.

[11] K.-C. Lee, J. Ho, M.-H. Yang, and D. Kriegman, Video-based face recognition using probabilistic appearance manifolds, in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '03)*, vol. 1, pp. 313320, Madison, Wis, USA.

[12] X. Liu and T. Chen, Video-based face recognition using adaptive hidden Markov models, in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '03)*, vol. 1, pp. 340345, Madison, Wis, USA.

[13] M. Everingham and A. Zisserman, Identifying individuals in video by combinig 'generative' and discriminative head models, in *Proceedings of the 10th IEEE International Conference on Computer Vision (ICCV '05)*, vol. 2, pp. 11031110, Beijing, China, October 2005.

[14] O. Arandjelović, G. Shakhnarovich, J. Fisher, R. Cipolla, and T. Darrell, Face recognition with image sets using manifold density divergence, in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '05)*, vol. 1, pp. 581588, San Diego, Calif, USA.

[15] O. Arandjelovic and R. Cipolla, An illumination invariant face recognition system for access control using video, in *Proceedings of the British Machine Vision Conference (BMVC '04)*, pp. 537546, Kingston, Canada, September 2004.

[16] C. Xie, B. V. K. Vijaya Kumar, S. Palanivel, and B. Yegnanarayana, A still-to-video face verification system using advanced correlation filters, in *Proceedings of the 1st International Conference on Biometric Authentication (ICBA '04)*, vol. 3072, pp. 102108, Hong Kong.

[17] K. W. Bowyer and K. Chang, A survey of 3D and multimodal 3D+2D face recognition, in *Face Processing: Advanced Modeling and Methods*, Academic Press, New York, NY, USA, 2005.

[18] Y.-H. Kim, A. M. Martínez, and A. C. Kak, Robust motion estimation under varying illumination, *Image and Vision Computing*, vol. 23, no. 4, pp. 365375, 2005.

[19] G. D. Hager and P. N. Belhumeur, Efficient region tracking with parametric models of geometry and illumination, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 10, pp. 10251039, 1998.

[20] H. Jin, P. Favaro, and S. Soatto, Real-time feature tracking and outlier rejection with changes in illumination, in *Proceedings of the 8th International Conference on Computer Vision (ICCV '01)*, vol. 1, pp. 684689, Vancouver, BC, USA.

[21] S. Koterba, S. Baker, I. Matthews, et al., Multi-view AAM fitting and camera calibration, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV ' 05)*, vol. 1, pp. 511518, Beijing, China.

[22] P. Eisert and B. Girod, Illumination compensated motion estimation for analysis sythesis coding, in *Proceedings of the 3D Image Analysis and Synthesis*, pp. 6166, Erlangen, Germany, November 1996.

[23] R. Basri and D. W. Jacobs, Lambertian reflectance and linear subspaces, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 2, pp. 218233, 2003.

[24] R. Ramamoorthi, Modeling illumination variation with spherical harmonics, in *Face Processing: Advanced Modeling and Methods*, Academic Press, New York, NY, USA, 2005.

[25] J. Ho and D. Kriegman, On the effect of illumination and face recognition, in *Face Processing: Advanced Modeling and Methods*, Academic Press, New York, NY, USA, 2005.

[26] Y. Xu and A. Roy-Chowdhury, Integrating the effects of motion, illumination and structure in video sequences, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV '05)*, vol. 2, pp. 16751682, Beijing, China.

[27] Y. Xu and A. Roy-Chowdhury, Integrating motion, illumination, and structure in video sequences with applications in illumination-invariant tracking, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 5, pp. 793806, 2007.

[28] Y. Xu and A. Roy-Chowdhury, Inverse compositional estimation of 3D pose and lighting in dynamic scenes, *IEEE Transactions on Pattern Analysis and Machine Intelligence*. In press.

[29] A. J. O'Toole, J. Harms, S. L. Snow, et al., A video database of moving faces and people, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 5, pp. 812816, 2005.

[30] S. Baker and I. Matthews, Lucas-Kanade 20 years on: a unifying framework, *International Journal of Computer Vision*, vol. 56, no. 3, pp. 221255, 2004.

[31] B. Schölkopf, A. Smola, and K.-R. Müller, Nonlinear component analysis as a Kernel Eigenvalue problem, *Neural Computation*, vol. 10, no. 5, pp. 12991319, 1998.

[32] L. D. Lathauwer, B. D. Moor, and J. Vandewalle, A multillinear singular value decomposition, *Journal on Matrix Analysis and Applications*, vol. 21, no. 4, pp. 12531278, 2000.

[33] V. Blanz and T. Vetter, Face recognition based on fitting a 3D morphable model, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 9, pp. 10631074, 2003.

[34] A. K. Roy Chowdhury and R. Chellappa, Face reconstruction from monocular video using uncertainty analysis and a generic model, *Computer Vision and Image Understanding*, vol. 91, no. 1-2, pp. 188213, 2003.

[35] G. Shakhnarovich, J. W. Fisher, and T. Darrell, Face recognition from long-term observations, in *Proceedings of the 7th European Conference on Computer Vision (ECCV '02)*, vol. 235 of *Lecture Notes In Computer Science*, pp. 851868, Copenhagen, Denmark, May 2002.

[36] Y. Xu and A. Roy-Chowdhury, Pose and illumination invariant registration and tracking for video-based face recognition, in *Proceedings of the IEEE Computer Society Workshop on Biometrics, in Association with CVPR*, New York, NY, USA.