Pose and Illumination Invariant Face Recognition Using Video Sequences

Amit K. Roy-Chowdhury and Yilei Xu

University of California, Riverside
{amitrc,yxu}@ee.ucr.edu

1 Abstract

Pose and illumination variations remain a persistent challenge in face recognition. In this paper, we present a framework for face recognition from *video sequences* that is robust to large changes in facial pose and lighting conditions. 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. The pose and illumination estimates in the probe and gallery sequences can then be compared for recognition applications. If similar parameters exist in both the probe and gallery, the similarity between the set of images can be directly computed. If the lighting and pose parameters in the probe and gallery are different, we will synthesize the images using the face model estimated from the training data corresponding to the conditions in the probe sequences. The method can handle situations where the pose and lighting conditions in the training and testing data are very different. We will show results on a video-based face recognition dataset that we have collected.

2 Introduction

Pose and illumination variations remain a persistent problem in face recognition, and has been documented in different studies [47, 30]. These two factors affect low-level tasks like face registration and tracking, which, in turn, reduce the final accuracy of the recognition algorithms. Also, it is often difficult to estimate illumination conditions accurately so as to factor them into the recognition strategies. Pose estimation problems are often made difficult by the fact that illumination is unknown. Therefore, it is extremely important to develop methods for face recognition that are robust to variations in pose and illumination.

It is believed by many that video-based systems hold promise in certain applications where motion can be used as a cue for face segmentation and tracking, and the presence of more data can increase recognition performance [47]. However, video-based face recognition systems have their own challenges such as low resolution of the face region, segmentation and tracking over time, 3D modeling, and developing measures for integrating information over the entire sequence. In this paper, we present a novel framework for video-based face tracking and recognition that is based on learning joint illumination and motion models from video, synthesizing novel views based on the learned parameters, and designing metrics that can compare two time sequences while being robust to outliers. We show experimentally that our method achieves high identification rates under extreme changes of pose and illumination.

2.1 Overview of the Approach

The underlying concept of this paper is a method for learning joint illumination and motion models of objects from video. The application focus is on video-based face recognition where the learned models are used to i) automatically and accurately track the face in the video, and ii) synthesize novel views under different pose and illumination conditions. We can handle a variety of lighting conditions, including the presence of multiple and extended light sources, which is natural in outdoor environments (where face recognition performance is still poor [47, 30, 31]). We can also handle gradual and sudden changes of lighting patterns over time. This is achieved using the spherical harmonics based representation of illumination [3, 33] and our previous work that integrates motion and illumination models for video analysis [43]. In [3, 33], the reflectance image was represented using a linear combination of spherical harmonics basis functions. For Lambertian objects, a ninth order expansion was deemed sufficient to capture most of the energy in the signal, while non-Lambertian objects required higher order coefficients. In [43, 44], we showed that the appearance of a moving object under arbitrary lighting could be represented as bilinear combination of 3D motion and the spherical harmonics coefficients for illumination.

This bilinear model of illumination and motion parameters allows us to develop an algorithm for tracking a moving object with arbitrary illumination variations. This is achieved by alternately projecting onto the appropriate motion and illumination bases of the bilinear space. In addition to the 3D motion estimates, we are also able to recover the illumination conditions as a function of time, which allows us to synthesize novel images under the same lighting conditions. The framework does not assume any model for the variation of the illumination conditions - lighting can change slowly or drastically and can originate from a combination of point and extended sources. The method *relies upon image differences and does not require computation of correspondences* *between images.* It leads to the development of an illumination invariant model based tracking algorithm that is initialized by registering the model (e.g., a generic face model) to the first frame of the sequence.

The recognition algorithm proceeds as follows. We assume that a 3D model of each face in the gallery is available. (We later show experimentally that an approximate 3D model with the correct texture is often good enough). Given a probe sequence, we track the face automatically in the video sequence under arbitrary pose and illumination conditions (as explained above). During the process, we also learn the illumination model parameters. The learned 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. 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. Robust distance measures are studied for this purpose.

Experimental evaluation is carried out on a database of 32 people that we collected for this purpose. One of the challenges in video-based face recognition is the lack of a good dataset, unlike in image-based approaches [47]. The dataset in [23] is small and consists mostly of pose variations. The dataset described in [28] has large pose variations under constant illumination, and illumination changes in natural environments but mostly in fixed frontal/profile poses (these are essentially for gait analysis). An ideal dataset for us would be similar to the CMU PIE dataset [37], but with video sequences instead of discrete poses. This is the reason why we collected our own data, which has large, simultaneous pose and illumination variations. We are presently enlarging this dataset and adding expression variations.

2.2 Relation to Previous Work

We divide our survey of the relevant literature into two broad parts. First we look at face recognition, especially the problem of pose and illumination variations. Next, we compare our joint illumination and motion models with other some approaches that deal with illumination variations in motion analysis.

Face Recognition

Due to want of space, we refer the reader to a recent review paper for existing work on face recognition [47]. A recently edited book [48] also deals with many of well-known approaches for face processing, modeling and recognition. For a comparison of the performance of various face recognition algorithms on standard databases, the reader can refer to [31, 30]. We will briefly review a few papers most directly related to this work.

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 [22], the authors propose to arrange physical lighting so that the acquired images of each object can be directly used as the basis vectors of the low-dimensional linear space. In [46], 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 morphable model based face recognition algorithm was proposed in [6], but they use the Phong illumination model, estimation of whose parameters can be more difficult than the spherical harmonics model in the presence of multiple and extended light sources. In [16], a method was proposed for using Locality Preserving Projections (LPP) to eliminate the unwanted variations resulting from changes in lighting, facial expression, and pose. The authors in [12, 13] proposed to use Eigen Light-Fields and Fisher Light-Fields to do pose invariant face recognition. They used generic training data and gallery images to estimate the Eigen/Fisher Light-Field of the subject's head, and then compare the probe image and gallery light-fields to match the face. In [49], the authors used photometric stereo methods for face recognition under varying illumination and pose. Their method requires iteration over all the poses in order to find the best match. Correlation filters have been proposed for illumination invariant face recognition from still images in [36]. A novel method for multilinear independent component analysis was proposed in [41] for pose and illumination invariant face recognition. All of these methods deal with recognition in a single image or across discrete poses and do not consider continuous video sequences. The authors in [23] deal with the issue of video-based face recognition, but concentrate mostly on pose variations. A method for video-based face verification using correlation filters was proposed in [42]. The advantage of using 3D models in face recognition has been highlighted in [8], but their focus is on 3D models obtained directly from the sensors and not estimated from video. This paper provides a method for *learning* the pose and illumination conditions from video, using a 3D model that can be estimated from images.

Modeling Illumination Variations in Video

Learning the parameters of the *joint* illumination and motion space is a novel contribution of this paper and we briefly review some related work. One of the well known approaches for 2D motion estimation is optical flow [18]. However, it involves the brightness constancy constraint, which is often violated in practice. Many researchers have tried overcoming this by introducing an illumination variation term within the standard optical flow formulation. In [29], the author coined the term "photometric motion" to define the intensity change of an image point due to object rotation, and applied it to solve for shape and reflectance. In [14], a parameterized function was proposed to describe the movement of the image points taking into account the illumination variation. In [27], the author combined the geometric and photometric effects for flow computation and highlighted the need for integrating the different variabilities in the process of image analysis. A method for shape reconstruction of a moving object under arbitrary, unknown illumination, assuming motion is known, was presented in [38]. Lighting changes were modeled by introducing illumination-specific parameters into the standard optical flow equations in [45]. Illumination invariant optical flow estimation was also the theme of [11], where an energy function was proposed to account for illumination changes and optimized using graph cuts. Another well-known approach for 2D motion estimation in monocular sequences is the Kanade-Lucas-Tomasi (KLT) tracker [40], which selects features that are optimal for tracking, and its extensions to handle illumination variations [19]. All of these approaches deal with 2D motion estimation that can handle only small changes in the pose of the object.

Our approach is illumination-invariant 3D motion estimation, *while simultaneously learning the parameters of the illumination model.* The 2D motion obtained by any of the above methods can be used along with the well-known structure from motion (SfM) methods [15] to compute 3D motion and structure. However, the accuracy of the 3D estimates will be limited by the accuracy of the 2D motion estimates in the case of lighting changes. As an alternative, model-based techniques have been used for direct 3D motion estimation from video [24]. Many 3D model based motion estimation algorithms rely on optical flow for the 2D motion and most existing methods are sensitive to lighting changes. The authors in [5] use probabilistic models and particle filters within a Bayesian framework to robustly track the human body, thus accounting for moderate illumination variations indirectly. A related work is [25], which uses SfM with photometric stereo to estimate surface structure. However, all the frames are needed a priori and an orthographic camera is assumed. Illumination invariant motion estimation is possible within the Active Appearance Model framework [10, 20], but the method requires training images under different illumination conditions. While these methods can handle illumination variations within the video sequence, they are not able to explicitly recover the illumination conditions of each frame in the video.

In [3] and [33], 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. This was an approximation of the infinite-dimensional convex cone representation derived in [4]. All of these 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 [43, 44] for integrating the spherical harmonics based illumination model with the motion of the objects leading to a bilinear model of lighting and motion parameters. This approach to illumination modeling takes into account the 3D shape of the object, which is in contrast to the 2D approaches for handling illumination variation, like gradient orientation histograms [9], scale-invariant feature transforms [26] and others [2, 39]. This is motivated by a number of reasons. Our final goal is to estimate the 3D motion and shape of the objects, in addition to the lighting conditions. Thus it makes sense to integrate the illumination models with the 3D shape models. Secondly, a number of authors have shown that 2D approaches to handle illumination variations have limited ability due to lack of knowledge of the underlying geometry of the object [1, 17, 32]. Thirdly, we not only want to achieve illumination invariance, but also learn the parameters of the illumination models from video sequences. The 3D approaches to illumination modeling allow this from video sequences of natural moving objects.

2.3 Organization of the Chapter

The rest of the paper is organized as follows. Section 3 presents a brief overview of the theoretical result describing the bilinear model of joint motion and illumination variables. Section 4 describes the algorithm for learning the parameters of the bilinear model. Section 5 describes our recognition algorithm. In Section 6 experimental results are presented. Section 7 concludes the paper and highlights future work.

3 Integrating Illumination and Motion Models in Video

The authors in [3, 33] proved that for a fixed Lambertian object, the set of reflectance images can be approximated by a linear combination of the first nine spherical harmonics, i.e,

$$I(x,y) = \sum_{i=0,1,2} \sum_{j=-i,-i+1 \ldots i-1,i} l_{ij} b_{ij}(\mathbf{n}), \tag{1}$$

where I is the reflectance intensity of the image pixel (x,y), i and j are the indicators for the linear subspace dimension in the spherical harmonics representation, l_{ij} is the illumination coefficient determined by the illumination direction, b_{ij} are the basis images, and $\mathbf{n}$ is the unit norm vector at the reflection point. The basis images can be represented in terms of the spherical harmonics as

$$b_{ij}(\mathbf{n}) = \rho r_i Y_{ij}(\mathbf{n}), i = 0,1,2; j = -i, \ldots, i, \tag{2}$$

where ρ is the albedo at the reflection point, r_i is constant for each spherical harmonics order, and Y_{ij} is the spherical harmonics function. For brevity, we will refer to the work in [3] as the Lambertian Reflectance Linear Subspace (LRLS) theory.

This result does not consider the relative motion between the object and the camera. In [43], it was shown that for moving objects it is possible to approximate the sequence of images by a bilinear subspace. We exploit this result for 3D motion estimation under arbitrarily varying illumination. We assume a perspective projection model for the camera, consider the focal length, f, of the camera as the only intrinsic parameter (can be relaxed), and assume the reference frame to be attached to the camera with the z-axis being along the optical axis. At time instance t_1, assume we know the 3D model of the object, its pose, and the illumination condition in terms of the coefficients $l_{ij}^{t_1}$. The ray from the optical center to the pixel (x, y) intersects with the surface at $\mathbf{P_1}$. Define the motion of the object in the above reference frame as the translation $\mathbf{T} = \left[T_x\ T_y\ T_z\right]^T$ of the centroid of the object and the rotation $\mathbf{\Omega} = \left[\omega_x\ \omega_y\ \omega_z\right]^T$ about the centroid. After the motion, $\mathbf{P_1}$ moves to $\mathbf{P_1}'$, and another point $\mathbf{P_2}$ moves to $\mathbf{P_2}'$. At the new time instance t_2, the direction of this ray does not change, and it intersects with the surface at $\mathbf{P_2}'$. The new illumination condition is represented in terms of the coefficients $l_{ij}^{t_2}$. This is represented pictorially in Figure 1.

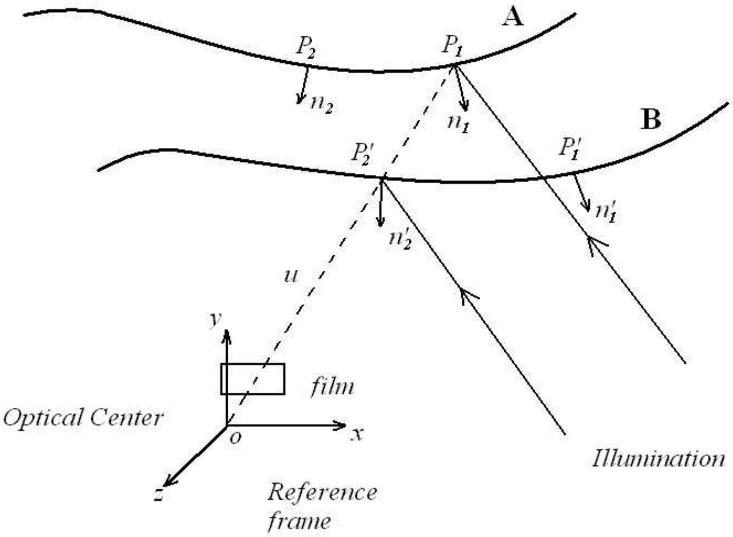


Fig. 1. Pictorial representation showing the motion of the object and its projection.

The authors in [43] proved that reflectance image at new time instance t_2 can be expressed as:

$$I(x, y, t_2) = \sum_{i=0,1,2} \sum_{j=-i,-i+1\ldots i-1,i} l_{ij}^{t_2} b_{ij}(\mathbf{n}_{\mathbf{P_2'}}), \qquad (3)$$

where

$$b_{ij}(\mathbf{n}_{\mathbf{P}_2'}) = b_{ij}(\mathbf{n}_{\mathbf{P}_1}) + \mathbf{AT} + \mathbf{B\Omega}. \tag{4}$$

In (3), $b_{ij}(\mathbf{n}_{\mathbf{P}_2'})$ and $l_{ij}^{t_2}$ are the basis images and illumination coefficients after motion. In (4), $b_{ij}(\mathbf{n}_{\mathbf{P}_1})$ are the original basis images before motion. **A** and **B** contain the structure and camera intrinsic parameters. Substituting (4) into (3), we see that the new image spans a bilinear space of six motion and approximately nine illumination variables (for Lambertian objects). The basic result is valid for general illumination conditions, but require consideration of higher order spherical harmonics.

When the illumination changes gradually, we can use the Talyor series to approximate the illumination coefficients as $l_{ij}^{t_2} = l_{ij}^{t_1} + \Delta l_{ij}$. Ignoring the higher order terms, the bilinear space now becomes a combination of two linear subspaces, as

$$\begin{aligned} I(x,y,t_2) = I(x,y,t_1) &+ \sum_{i=0,1,2} \sum_{j=-i,\dots,i} l_{ij}^{t_1}(\mathbf{AT} + \mathbf{B\Omega}) \\ &+ \sum_{i=0,1,2} \sum_{j=-i,\dots,i} \Delta l_{ij} b_{ij}(\mathbf{n}_{\mathbf{P}_1}). \end{aligned} \tag{5}$$

If the illumination does not change from t_1 to t_2 (often a valid assumption for a short interval of time), the new image at t_2 spans a linear space of the motion variables, since the third term in (5) is zero.

We can express the result in (3) succinctly using tensor notation as

$$\mathcal{I} = (\mathcal{B} + \mathcal{C} \times_2 \begin{pmatrix} \mathbf{T} \\ \mathbf{\Omega} \end{pmatrix}) \times_1 \mathbf{l}, \tag{6}$$

where $\times_n$ is called the *mode-n product* [41], and $\mathbf{l} \in \mathbf{R^9}$ is the vector of l_{ij} components. The *mode-n product* of a tensor $\mathcal{A} \in \mathbf{R}^{I_1 \times I_2 \times \ldots \times I_n \times \ldots \times I_N}$ by a vector $\mathbf{V} \in \mathbf{R}^{1 \times I_n}$, denoted by $\mathcal{A} \times_n \mathbf{V}$, is the $I_1 \times I_2 \times \ldots \times 1 \times \ldots \times I_N$ tensor

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

For each pixel (p,q) in the image, $\mathcal{C}_{klpq} = [A\ B]$ of size $N_l \times 6$, where N_l is the dimension of the illumination basis ($N_l \approx 9$ for Lambertian objects). Thus for an image of size $M \times N$, $\mathcal{C}$ is $N_l \times 6 \times M \times N$. $\mathcal{B}$ is a subtensor of dimension $N_l \times 1 \times M \times N$, comprising the basis images $b_{ij}(\mathbf{n}_{\mathbf{P}_1})$, and $\mathcal{I}$ is a subtensor of dimension $1 \times 1 \times M \times N$, representing the image. $\mathbf{l}$ is still the $N_l \times 1$ vector of the illumination coefficients.

These theoretical results can be used to synthesize video sequences of objects under different conditions of lighting and motion. This would rely on computing the basis images which are a function of the surface normal. In practice, the surface normals are computed by finding the intersection of the ray passing through a pixel with a 3D point, assuming that the 3D model is represented by a cloud of points. The normal is then calculated by considering neighboring points. If a mesh model of the object is used, the intersection of the ray with a triangular mesh is computed, and the normal to this mesh patch is calculated.

4 Learning Joint Illumination and Motion Models from Video

The joint illumination and motion space provides us with a novel method for 3D motion estimation under varying illumination. This is based on inverting the generative model for motion and illumination modeling. It can not only track the 3D motion under varying illumination, but also can estimate the illumination parameters.

Equation (3) provides us an expression relating the reflectance image I_{t2} with new illumination coefficients $l_{ij}^{t_2}$ and motion variables $\mathbf{m} = [\mathbf{T}, \boldsymbol{\Omega}]^T$, which lead to a method for estimating 3D motion and illumination as:

$$\begin{aligned}(\hat{\mathbf{l}}, \hat{\mathbf{T}}, \hat{\boldsymbol{\Omega}}) &= \arg\min_{\mathbf{l},\mathbf{T},\boldsymbol{\Omega}} \|I_{t2} - \sum_{i=0,1,2}\sum_{j=-i}^{i} l_{ij} b_{ij}(\mathbf{n}_{\mathbf{P}'_2})\|^2, \\ &= \arg\min_{\mathbf{l},\mathbf{T},\boldsymbol{\Omega}} \|\mathcal{I}_{t2} - (\mathcal{B}_{t1} + \mathcal{C}_{t1} \times_2 \begin{pmatrix}\mathbf{T}\\ \boldsymbol{\Omega}\end{pmatrix}) \times_1 \mathbf{l}\|^2, \qquad (7)\end{aligned}$$

where $\hat{x}$ denotes an estimate of x. The cost function is a square error norm, similar to the famous bundle-adjustment [15], but incorporates an illumination term. Motion and illumination estimates are obtained for each frame. Since the motion between consecutive frames is small, but illumination can change suddenly, we add a regularization term to the above cost function. It is of the form $\alpha||\mathbf{m}||^2$.

Since the image I_{t2} lies approximately in a bilinear space of illumination and motion variables (ignoring the regularization term for now), such a minimization problem can be achieved by alternately estimating the motion and illumination parameters by projecting the video sequence onto the appropriate basis functions derived from the bilinear space. Assuming that we have tracked the sequence upto some frame for which we can estimate the motion (hence, pose) and illumination, we calculate the basis images, b_{ij}, at the current pose, and write it in tensor form $\mathcal{B}$. Unfolding[1] $\mathcal{B}$ and the image $\mathcal{I}$ along the first dimension [21], which is the illumination dimension, the image can be represented as:

$$\mathcal{I}_{(1)}^T = \mathcal{B}_{(1)}^T \mathbf{l}. \qquad (8)$$

[1] Assume an Nth-order tensor $\mathcal{A} \in \mathbf{C}^{I_1 \times I_2 \times \ldots \times I_N}$. The matrix unfolding $\mathbf{A}_{(n)} \in \mathbf{C}^{I_n \times (I_{n+1} I_{n+2} \ldots I_N I_1 I_2 \ldots I_{n-1})}$ contains the element $a_{i_1 i_2 \ldots i_N}$ at the position with row number i_n and column number equal to $(i_{n+1} - 1) I_{n+2} I_{n+3} \ldots I_N I_1 I_2 \ldots I_{n-1} + (i_{n+2} - 1) I_{n+3} I_{n+4} \ldots I_N I_1 I_2 \ldots I_{n-1} + \cdots + (i_N - 1) I_1 I_2 \ldots I_{n-1} + (i_1 - 1) I_2 I_3 \ldots I_{n-1} + \cdots + i_{n-1}$.

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

$$\hat{\mathbf{l}} = (\mathcal{B}_{(1)}\mathcal{B}_{(1)}^T)^{-1}\mathcal{B}_{(1)}\mathcal{I}_{(1)}^T. \tag{9}$$

Keeping the illumination coefficients fixed, the bilinear space in equations (3) and (4) becomes a linear subspace, i.e.,

$$\mathcal{I} = \mathcal{B} \times_1 \mathbf{l} + (\mathcal{C} \times_1 \mathbf{l}) \times_2 \begin{pmatrix} \mathbf{T} \\ \mathbf{\Omega} \end{pmatrix}. \tag{10}$$

Similarly, unfolding all the tensors along the second dimension, which is the motion dimension, and adding the effect of the regularization term, $\mathbf{T}$ and $\mathbf{\Omega}$ can be estimated as:

$$\begin{pmatrix} \hat{\mathbf{T}} \\ \hat{\mathbf{\Omega}} \end{pmatrix} = \left((\mathcal{C} \times_1 \mathbf{l})_{(2)}(\mathcal{C} \times_1 \mathbf{l})_{(2)}^T + \alpha\mathbf{I}\right)^{-1}(\mathcal{C} \times_1 \mathbf{l})_{(2)}(\mathcal{I} - \mathcal{B} \times_1 \mathbf{l})_{(2)}^T, \tag{11}$$

where $\mathbf{I}$ is an identity matrix of dimension 6×6. The above procedure for estimation of the motion should proceed in an iterative manner, since $\mathcal{B}$ and $\mathcal{C}$ are functions of the motion parameters. This should continue until the projection error $\|\mathcal{I} - \mathcal{B} \times_1 \hat{\mathbf{l}}\|^2$ does not decrease further. This process of alternate minimization leads to the local minimum of the cost function (which is quadratic in motion and illumination variables) at each time step. This can be repeated for each subsequent frame. We now describe the algorithm formally.

4.1 Algorithm

Consider a sequence of image frames I_t, $t = 0, ..., N-1$.

Initialization: Take one image of the object from the video sequence, register the 3D model onto this frame and map the texture onto the 3D model. Calculate the tensor of the basis images $\mathcal{B}_0$ at this pose. Use (9) to estimate the illumination coefficients. Now, assume that we know the motion and illumination estimates for frame t, i.e., $\mathbf{T}_t, \mathbf{\Omega}_t$ and $\mathbf{l}_t$.

- Step 1. Calculate the tensor form of the bilinear basis images $\mathcal{B}_t$ at the current pose using (4). Use (11) to estimate the new pose from the estimated motion.
- Step 2. Assume illumination does not change, i.e. $\hat{\mathbf{l}}_{t+1} = \hat{\mathbf{l}}_t$. Compute the motion $\mathbf{m}$ by minimizing the difference between an input frame and the rendered frame $\|\mathcal{I}_{t+1} - (\mathcal{B}_t + \mathcal{C}_t \times_2 \begin{pmatrix} \hat{\mathbf{T}}_{t+1} \\ \hat{\mathbf{\Omega}}_{t+1} \end{pmatrix}) \times_1 \hat{\mathbf{l}}_{t+1}\|^2$, and estimate the new pose.
- Step 3. Using the new pose estimate, re-estimate the illumination using (9). Repeat Steps 1 and 2 with the new estimated $\hat{\mathbf{l}}_{t+1}$ for that input frame, till the error is below an acceptable threshold.
- Step 4. Set t = t + 1. Repeat Steps 1, 2 and 3.
- Step 5. Continue till t = N - 1.

In many practical situations, the illumination changes slowly within a sequence (e.g., cloud covering the sun). In this case, we use the expression in (5) instead of (3,4) in the cost function (7) and estimate Δl_{ij}.

4.2 Handling Occlusions

The optimization function (7) yields the maximum likelihood estimate under the assumption of additive Gaussian noise to the image observations. However, in the presence of occlusion, the optimization function can be used only if we can work with the unoccluded pixels, which will have to be estimated a priori. A simple way to do this is to set a threshold and discard those pixels that have an intensity change (with respect to the previous frame) greater than the threshold. However, a simple threshold strategy may eliminate the pixels that are not occluded, but whose intensity changes because of the change in illumination conditions. Therefore, we propose the following modification to our algorithm to handle occlusion.

Assume that we are able to obtain the tracking and illumination estimates upto some instance t. Then, we can calculate the bilinear basis images at the current pose, and project the frame at the next time instance, $t+1$, onto the linear subspace of the basis images. This gives an estimate of the illumination coefficients for the frame. Using the basis images, we can synthesize the image with the newly estimated illumination coefficients $\mathbf{l_{t+1}}$. In order to do this, the motion between I_{t+1} and I_t is assumed to be the same as between I_t and I_{t-1} (i.e, uniform motion). If the difference between the synthesized image and the observed one is larger than some threshold for some pixels, we will discard these pixels. By doing this, we store a mask for the pixels which are occluded. Note that the synthesized image has the new illumination condition, and thus is not affected by the problem noted above. Using the unoccluded pixels and the algorithm described in Section 4.1, we re-estimate the 3D motion as well as the new illumination coefficients $\hat{\mathbf{l}}_{\mathbf{t+1}}$. For the image at time instance $t+2$, we will use the mask at time instance $t+1$ to estimate the illumination condition $\hat{\mathbf{l}}_{\mathbf{t+2}}$, then repeat what we have done for $t+1$ frame and update the mask. This method works provided the occlusion happens slowly(most practical cases). For sudden occlusion, a RANSAC approach [15], that works with random subsets of feature points, will be adopted.

5 Face Recognition From Video

The generative framework for integrating illumination and motion models described in Section 2 and the method for learning the model parameters as described in Section 4 set the stage for developing a novel face recognition algorithm that is particularly suited to handling video sequences. The method is able to handle arbitrary pose and illumination variations and can integrate information over an entire video sequence.

In our method, the gallery is represented by a 3D model of the face. The model can be built from a single image [7], a video sequence [35] or obtained directly from 3D sensors [8]. In our experiments, the face model will be estimated from video. Given a probe sequence, we will estimate the motion and illumination conditions using the algorithms described in Section 4. 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. Given the rendered images from the 3D models in the gallery and the probe images, we will design robust metrics for comparing these two sequences. A feature of these metrics will be their ability to integrate the identity over all the frames, ignoring some frames that may have the wrong identity. Since 3D shape modeling is done for the gallery sequences only, we avoid the issues of high computational complexity of 3D modeling algorithms in real-time.

One of the challenges faced is to design suitable metrics capable of comparing two video sequences. This metric should be general enough to be applicable to most videos and robust to outliers. Let $P(f_i), i = 1, ..., N$ be N frames from the probe sequence. Let $SG_j(f_i), i = 1, ..., N$ be the frames of the synthesized sequence for galley j, where $j = 1, ..., M$ 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 and the synthesis algorithm. Let d_{ij} be the distance between the i^{th} frames of P and G_j. We now compare two distance measures that can be used for obtaining the identity of the probe sequence.

$$\begin{aligned} &1.\ ID = \arg\min_j \min_i d_{ij} \\ &2.\ ID = \arg\min_j \max_i d_{ij} \end{aligned} \tag{12}$$

The first alternative computes the distance between the frames in the probe and each synthesized sequence that are the most similar and chooses the identity as the individual with the smallest distance in the gallery. This can be looked upon as obtaining the identity of the probe from one image of it that is most similar to the gallery. The second distance measure can be interpreted as minimizing the maximum separation between the probe and synthesized gallery images. Both of these measures suffer from a lack of robustness, which can be critical for their performance since the correctness of the synthesized images depend upon the accuracy of the illumination and motion parameter estimates. For this purpose, we replace the max by the f^{th} percentile and the min (in the inner distance computation of 1 in (12)) by the $(1-f)^{th}$ percentile. In our experiments, we choose f to be 0.8 and use the first option.

A third possible option is to assign a weight to each image of each synthesized gallery that is inversely proportional to its distance from the corresponding probe image, sum all the weights and choose the gallery with largest weight as the identity. The problem with this method is that the recognition accuracy depends upon the choice of the weighting function, which in turn can vary with the probe and gallery sequences.

One point that still needs to be addressed is on how do we compute d_{ij}. Recall that a generic face model is used to track the face in the probe video and the estimated illumination and motion parameters are used to synthesize the videos for each person in the gallery using their 3D model. This sets up a mapping between the pixels in the synthesized images with the probe images through the 3D models. Also, the number of synthesized images is the same as the number of images in the probe, thus obviating any synchronization issues. Thus d_{ij} can be computed directly as the squared difference between the synthesized and probe image frames.

We now describe formally the video-based face recognition algorithm. Using the above notation, let $P(f_i), i = 1, ..., N$ be N frames from the probe sequence. Let $G_1, ..., G_M$ be the 3D models for each of M galleries.

- Step 1. Register a 3D generic face model to the first frame of the probe sequence. Estimate the illumination and motion model parameters for each frame of the probe sequence using the method described in Section 4.
- Step 2. Using the estimated illumination and motion parameters, synthesize, for each gallery, a video sequence using the generative model of (4). Denote these as $SG_j(f_i), i = 1, ..., N$ and $j = 1, ..., M$.
- Step 3. Compute d_{ij} in (12).
- Step 4. Obtain the identity using a suitable distance measure from (12), modifying it for robustness as necessary (see discussion above).

6 Experimental Results

Since the tracking and synthesis algorithms are the foundation for the recognition strategy, we first present results on these two aspects highlighting the accuracy of the methods in a controlled environment. We then describe our face video database and the results of the recognition algorithms.

6.1 Tracking and Synthesis Results

We synthesized a video sequence of a face with known motion and lighting. A generic 3D model was registered to the first frame of the sequence manually and tracked using the algorithm described in Section 4. Figures 2, 3 and 4 show the results of our tracking algorithm on this sequence. The images in Fig. 2 are synthesized from a 3D model, and thus the motion and illumination are known. The face is rotating along y axis from $-30°$ to $+30°$, and the illumination is changing such that the light always comes from the front of the face. The resolution of the image is 240 by 320. Figures 3 and 4 show plots of the estimated motion and illumination against the true values.

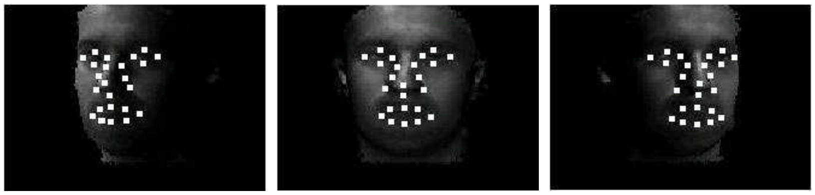


Fig. 2. The back projection of the mesh vertices of the 3D face model using the estimated 3D motion onto some input frames. Face is rotating about the y axis, and illumination is changing in the same way as pose.

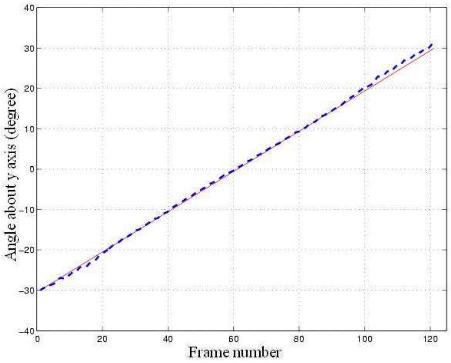
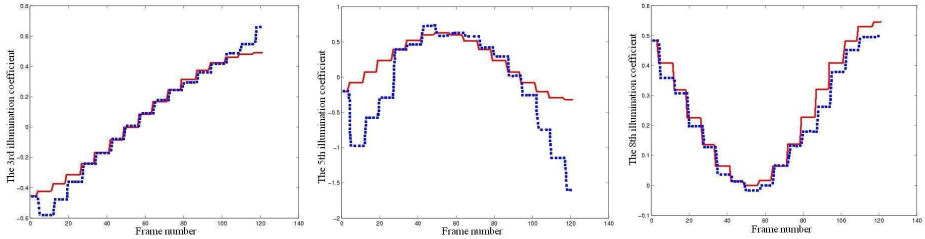


Fig. 3. The solid line shows the true pose (represented by the angle of face about y axis) and the broken line is the estimated pose.



(a) (b) (c)

Fig. 4. (a), (b), (c) are the estimates of the 3rd, 5th, and 8th illumination coefficients respectively. The solid line shows the true illumination coefficients using the LRLS method, and the dotted line shows the estimated illumination coefficients.

We also show results of the synthesis algorithm on a real-life video sequence. Frames from a synthesized video sequence using learned motion and illumination parameters are shown in Figure 5. Motion and illumination are learned from the frames in the first and second row respectively, and images in the third row are synthesized with the motion and illumination parameters learned from the corresponding frames in the same column. The reader can visually compare the synthesized images for accuracy of pose and illumination estimates.

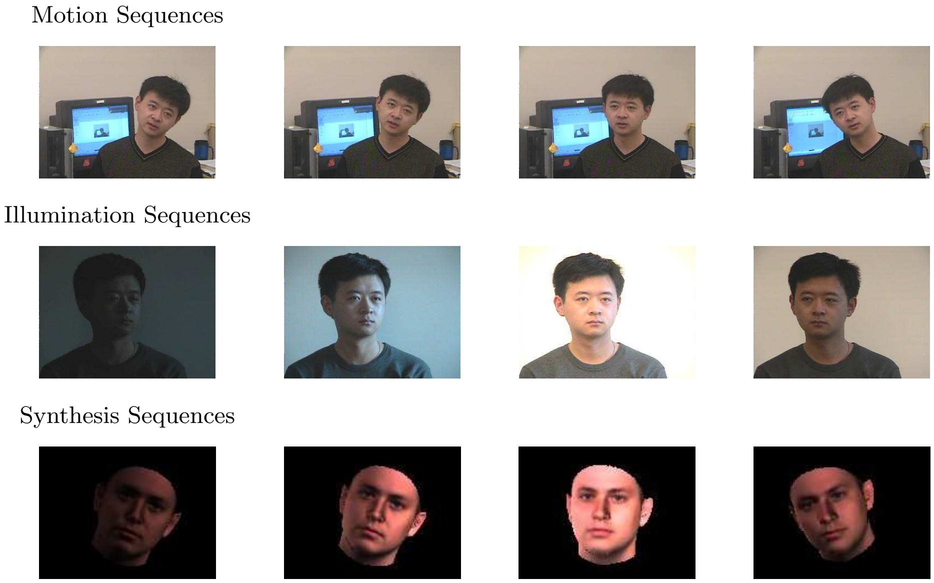


Fig. 5. An example of video synthesis with learned motion and illumination models. Motion and illumination are learned from the frames in the first and second row respectively, and images in the third row are synthesized with the motion and illumination parameters learned from the corresponding frames in the same column.

6.2 Face Recognition Results

Face Database

Our database consists of videos of 32 people. Each person was asked to move his/her head as they wished 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. An example of some of the images in the video database is shown in Figure 6. The resolution of the face varied depending on the person and the movement. A statistical analysis showed that the average size was about 70 x 70, with the minimum size being 50 x 50. Each sequence was divided into two parts - gallery and probe. The frames in Figure 6 are arranged in the same order as in the original video, with the first column representing a frame from

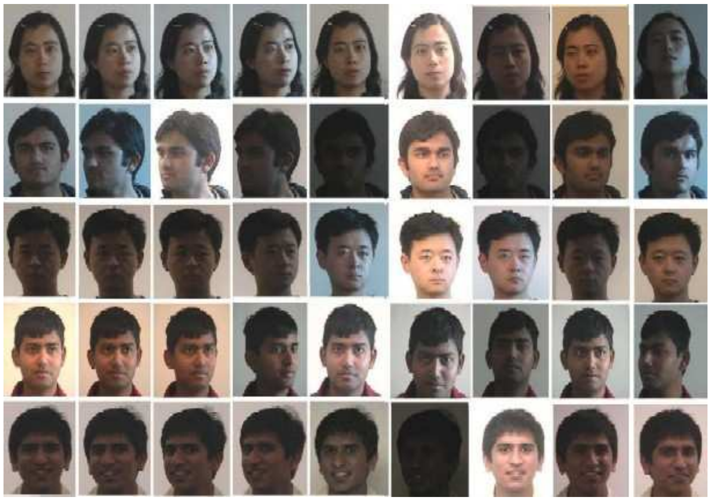


Fig. 6. Sample frames from the video sequences collected for our database.

the gallery, the third column representing the image in Expt. 1 (see below), and the fifth column representing the image in Expt. 3 (see below).

A 3D model of each face was constructed from the gallery sequence. In the set of experiments shown, a generic model was registered to one approximately frontal image in the gallery manually by choosing seven points on the face. Thereafter the texture of the face was mapped onto the model. The shape was not changed from the generic model. We would like to emphasize that any other 3D modeling algorithm would also have worked and we plan to integrate our previous work in [34] with this system.

From the portion of each sequence designated as probe, we designed five experiments by choosing different parts of it, as described below.

- Expt. 1: A single image, some examples of which are shown in the third column of Figure 6, was used as the probe.
- Expt. 2: A video sequence starting with the frame in Expt. 1 was used as the probe. Examples of these frames can be seen from the third column and beyond in Figure 6.
- Expt. 3: A single image, some examples of which are shown in the fifth column of Figure 6, was used as the probe.
- Expt. 4: A video sequence starting with the frame in Expt. 3 was used as the probe. Examples of which can be seen from the third column and beyond in Figure 6.
- Expt. 5: A video sequence that has a portion with frontal face and illumina-

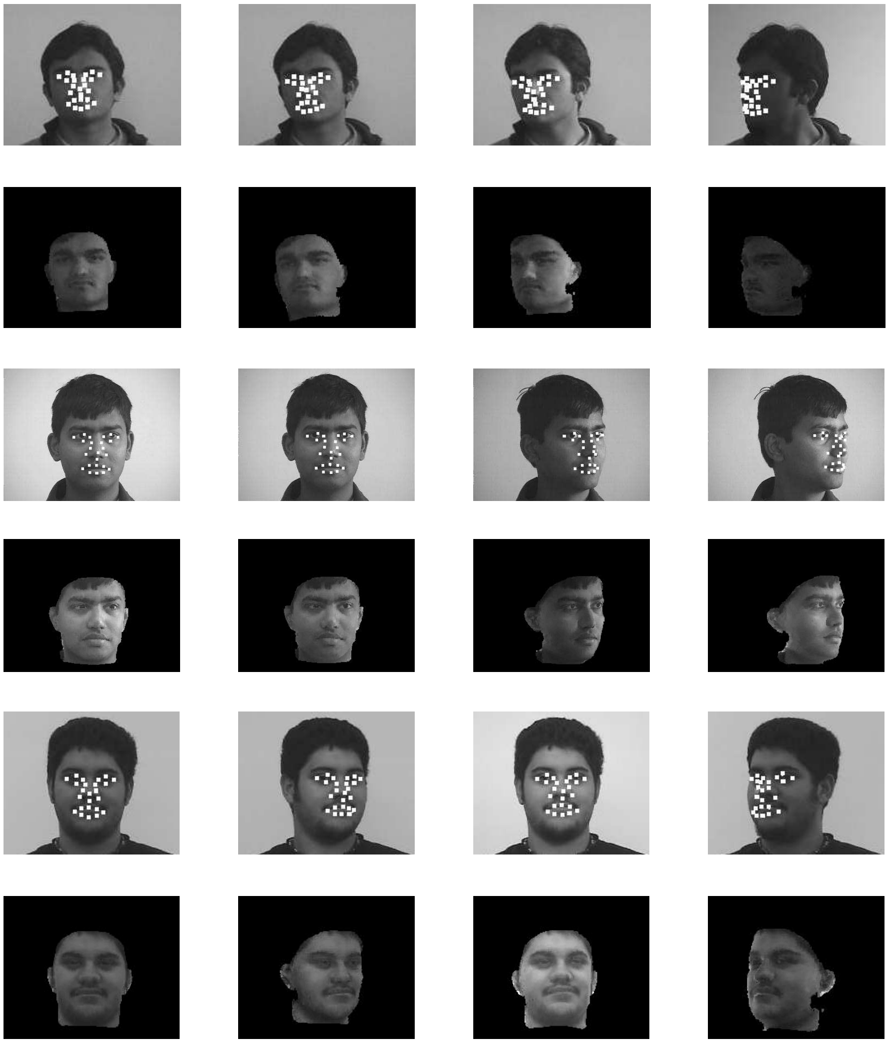


Fig. 7. Tracking and synthesis results are shown in alternating rows for three of the probes.

tion similar to the gallery was used as the probe. This is achieved by considering the probe sequence to start immediately after the gallery sequence ends in our collected data.

As can be seen from Figure 6, the pose and illumination varies randomly in the video. The reason for choosing the experiments in this way are the following: i) to study the advantage video provides over image-based recognition,

ii) how sensitive recognition rates are with respect to the actual frames in the video (hence the change in the starting frame in Expt. 4 compared to Expt. 2), and iii) how recognition rates are affected if there is a small portion of the video in the probe very similar to the gallery, even though the other frames may not be.

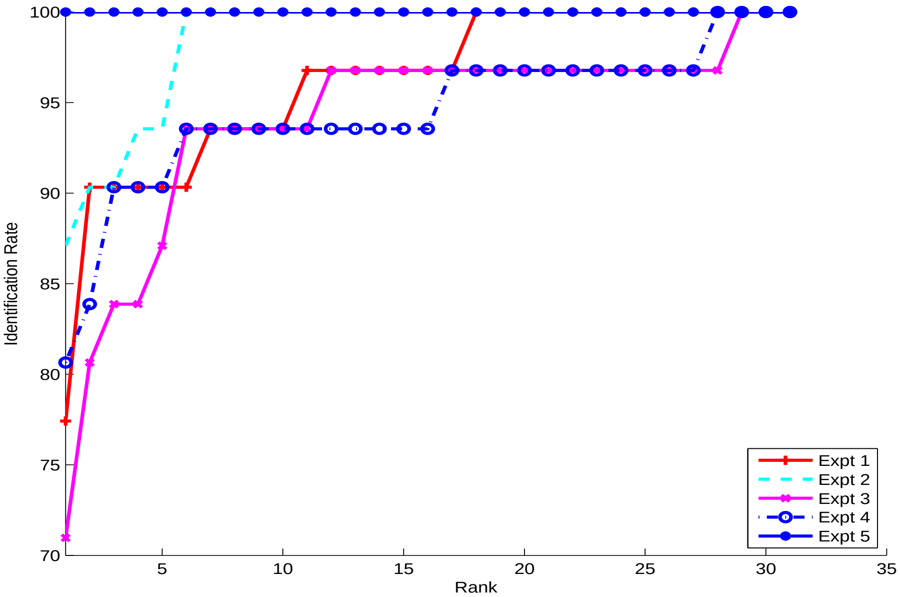


Fig. 8. CMC curve for video-based face recognition experiments.

The results on tracking and synthesis on three of the probes are shown in Figure 6. We plot the Cumulative Match Characteristic (CMC) [47, 30] for all the experiments in Figure 8. The following are the main conclusions that we can draw from our experiments.

• Our proposed algorithm gives relatively high performance (about 90% on the average for Expts. 1, 3 and 5 that deal with video sequences) on videos with large and arbitrary variations of pose and illumination.

• There is a significant change increase in performance in considering a video sequence compared to a single image, as evidenced by the improvements between Expts. 1 and 2, and between Expts. 3 and 4. Between Expts. 1 and 2 there is a 10% increase in the Rank 1 identification rate, as well as a significant increase in the slope of the CMC curve. Between Expts. 3 and 4, there is again a 10% increase in the identification rate. However, the recognition rates between Expts. 2 and 4 are different, demonstrating the sensitivity of the algorithm to the actual frames in the sequence (which is to be expected).

• When a part of the video sequence has overlap with the gallery (even one frame), our system gives a 100% recognition rate (Expt. 5).

All these experiments demonstrate the effectiveness of video-based face recognition methods over still image-based approaches. However, the recognition rate is affected significantly by the actual conditions under which the video was captured.

7 Conclusions

In this paper, we have proposed a method for video-based face recognition that relies upon a novel theoretical framework for integrating illumination and motion models for describing the appearance of a video sequence. We started with a brief exposition of this theoretical result, followed by methods for learning the model parameters. Then, we described our recognition algorithm that relies on synthesis of video sequences under the conditions of the probe. Finally, we demonstrated the effectiveness of the method on video databases with large and arbitrary variations in pose and illumination. In future, we will work on improving the tracking and synthesis algorithms (which we believe will improve recognition performance), performing thorough experimentation to understand the effect of the different variabilities, and analyzing performance on larger datasets.

8 Acknowledgments

The authors would like to acknowledge Keyur Patel and Manas Desai for designing a video-based face recognition system that was very helpful for experimenting with the proposed algorithms.

References

1. Y. Adini, Y. Moses, and S. Ullman. Face recognition: The problem of compensating for changes in illumination direction. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19(7):721–732, July 1997.
2. R. Alferez and Y.F. Wang. Geometric and illumination invariants for object recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 21(6):505–536, June 1999.
3. R. Basri and D.W. Jacobs. Lambertian Reflectance and Linear Subspaces. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 25(2):218–233, February 2003.
4. P. Belhumeur and D. Kriegman. What Is the Set of Images of an Object Under All Possible Lighting Conditions? In *Computer Vision and Pattern Recognition*, 1996.

5. M. Black and A. Jepson. EigenTracking: Robust matching and tracking of articulated objects using a view-based representation. In *Proc. ECCV*, pages 329–342, 1996.
6. V. Blanz, P. Grother, P. Phillips, and T. Vetter. Face Recognition Based on Frontal Views Generated From Non-Frontal Images. In *Computer Vision and Pattern Recognition*, 2005.
7. V. Blanz and T. Vetter. Face recognition based on fitting a 3D morphable model. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 25(9):1063–1074, September 2003.
8. K.W. Bowyer and Chang. A survey of 3D and Multimodal 3D+2D Face Recognition. In *Face Processing: Advanced Modeling and Methods*. Academic Press, 2005.
9. H.F. Chen, P.N. Belhumeur, and D.W. Jacobs. In search of illumination invariants. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pages I: 254–261, 2000.
10. T.F. Cootes, G.J. Edwards, and C.J. Taylor. Active Appearance Models. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 23(6):681–685, June 2001.
11. D. Freedman and M. Turek. Illumination-Invariant Tracking via Graph Cuts. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, 2005.
12. R. Gross, I. Matthews, and S. Baker. Eigen Light-Fields and Face Recognition Across Pose. In *Proc. of the IEEE International Conference on Automatic Face and Gesture Recognition*, 2002.
13. R. Gross, I. Matthews, and S. Baker. Fisher Light-Fields for Face Recognition Across Pose and Illumination. In *Proc. of the German Symposium on Pattern Recognition*, 2002.
14. G. D. Hager and P.N. Belhumeur. Efficient Region Tracking With Parametric Models of Geometry and Illumination. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 20(10):1025–1039, 1998.
15. R.I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision.* Cambridge University Press, 2000.
16. X. He, S. Yan, Y. Hu, P. Niyogi, and H.J. Zhang. Face Recognition Using Laplacianfaces. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 27(5), 2005.
17. J. Ho and D Kriegman. On the Effect of Illumination and Face Recognition. In *Face Processing: Advanced Modeling and Methods*. Academic Press, 2005.
18. B.K.P. Horn and B.G. Schunck. Determining Optical Flow. *Artificial Intelligence*, 17:185–203, 1981.
19. H. Jin, P. Favaro, and S. Soatto. Real-time feature tracking and outlier rejection with changes in illumination. In *IEEE Intl. Conf. on Computer Vision*, 2001.
20. S. Koterba, S. Baker, I. Matthews, C. Hu, H. Xiao, J. Cohn, and T. Kanade. Multi-view aam fitting and camera calibration. In *IEEE Intl. Conf. on Computer Vision*, 2005.
21. L. D. Lathauwer, B. D. Moor, and J. Vandewalle. A Multillinear Singular Value Decomposition. *SIAM J. Matrix Anal. Appl.*, 21(4):1253–1278, 2000.
22. K. Lee, J. Ho, and D.J. Kriegman. Acquiring linear subspaces for face recognition under variable lighting. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 27(5):684–698, May 2005.
23. K.C. Lee, J. Ho, M.H. Yang, and D.J. Kriegman. Video-based face recognition using probabilistic appearance manifolds. In *Computer Vision and Pattern Recognition*, pages I: 313–320, 2003.

24. V. Lepetit and P. Fua. *Monocular Model-Based 3D Tracking of Rigid Objects.* Now Publishers Inc., 2005.
25. J. Lim, J. Ho, M.H. Yang, and D.J. Kriegman. Passive photometric stereo from motion. *Proc. of IEEE International Conference on Computer Vision*, pages II: 1635–1642, 2005.
26. D.G. Lowe. Object recognition from local scale-invariant features. In *Proc. of IEEE International Conference on Computer Vision*, pages 1150–1157, 1999.
27. S. Negahdaripour. Revised Definition of Optical Flow: Integration of Radiometric and Geometric Cues for Dynamic Scene Analysis. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 20(9):961–979, September 1998.
28. A. O'Toole et al. A video database of moving faces and people. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, pages 812–816, May 2005.
29. A. Pentland. Photometric Motion. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 13(9):879–890, 1991.
30. 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. Technical Report NISTIR 6965, http://www.frvt.org, 2003.
31. P. J. Phillips et al. Overview of the face recognition grand challenge. In *Computer Vision and Pattern Recognition*, 2005.
32. R. Ramamoorthi. Modeling Illumination Variation With Spherical Harmonics. In *Face Processing: Advanced Modeling and Methods.* Academic Press, 2005.
33. R. Ramamoorthi and P. Hanrahan. On the relationship between radiance and irradiance: determining the illumination from images of a convex Lambertian object. *Journal of the Optical Society of America A*, 18(10), Oct 2001.
34. A. Roy-Chowdhury and R. Chellappa. Face Reconstruction From Monocular Video Using Uncertainty Analysis and a Generic Model. *Computer Vision and Image Understanding*, 91(1-2):188–213, July-August 2003.
35. A. Roy-Chowdhury, R. Chellappa, and R. Gupta. 3D Face Modeling From Monocular Video Sequences. In *Face Processing: Advanced Modeling and Methods.* Academic Press, 2005.
36. M. Savvides, B.V.K. Vijaya Kumar, and P.K. Khosla. Corefaces - robust shift invariant pca based correlation filter for illumination tolerant face recognition. In *Computer Vision and Pattern Recognition*, 2004.
37. T. Sim, S. Baker, and M. Bsat. The cmu pose, illumination and expression database. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 25:1615–1618, December 2003.
38. D. Simakov, D. Frolova, and R. Basri. Dense shape reconstruction of a moving object under arbitrary, unknown lighting. In *IEEE Intl. Conf. on Computer Vision*, 2003.
39. D.A. Slater and G. Healey. The illumination-invariant matching of deterministic local structure in color images. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19(10):1146–1151, October 1997.
40. C. Tomasi and J. Shi. Good Features to Track. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pages 593–600, 1994.
41. M.A.O. Vasilescu and D. Terzopoulos. Multilinear Independent Components Analysis. In *Computer Vision and Pattern Recognition*, 2005.
42. C. Xie, B.V.K. Vijaya Kumar, S. Palanivel, and B. Yegnanarayana. A still-to-video face verification system using advanced correlation filters. In *First International Conference on Biometric Authentication*, 2004.

43. Y. Xu and A. Roy-Chowdhury. Integrating the Effects of Motion, Illumination and Structure in Video Sequences. In *Proc. of IEEE International Conference on Computer Vision*, 2005.
44. Y. Xu and A. Roy-Chowdhury. Integrating Motion, Illumination and Structure in Video Sequences, With Applications in Illumination-Invariant Tracking. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2006 (Accepted).
45. L. Zhang, B. Curless, A. Hertzmann, and S.M. Seitz. Shape and Motion under Varying Illumination: Unifying Structure from Motion, Photometric Stereo, and Multi-view Stereo. In *Proc. of IEEE International Conference on Computer Vision*, 2003.
46. L. Zhang, S. Wang, and D. Samaras. Face Synthesis and Recognition from a Single Image under Arbitrary Unknown Lighting using a Spherical Harmonic Bais Morphable Model. In *Computer Vision and Pattern Recognition*, 2005.
47. W. Zhao, R. Chellappa, P.J. Phillips, and A. Rosenfeld. Face Recognition: A Literature Survey. *ACM Transactions*, 2003.
48. W. Zhao and R. Chellappa (Eds.). *Face Processing: Advanced Modeling and Methods*. Academic Press, 2005.
49. S. Zhou, R. Chellappa, and D. Jacobs. Characterization of human faces under illumination variations using rank, integrability, and symmetry constraints. In *European Conference on Computer Vision*, 2004.

Index