Integrating Motion, Illumination, and Structure in Video Sequences with Applications in Illumination-Invariant Tracking

Yilei Xu, *Student Member*, *IEEE*, and Amit K. Roy-Chowdhury, *Member*, *IEEE*

Abstract—In this paper, we present a theory for combining the effects of motion, illumination, 3D structure, albedo, and camera parameters in a sequence of images obtained by a perspective camera. We show that the set of all Lambertian reflectance functions of a moving object, at any position, illuminated by arbitrarily distant light sources, lies "close" to a bilinear subspace consisting of nine illumination variables and six motion variables. This result implies that, given an arbitrary video sequence, it is possible to recover the 3D structure, motion, and illumination conditions simultaneously using the bilinear subspace formulation. The derivation builds upon existing work on linear subspace representations of reflectance by generalizing it to moving objects. Lighting can change slowly or suddenly, locally or globally, and can originate from a combination of point and extended sources. We experimentally compare the results of our theory with ground truth data and also provide results on real data by using video sequences of a 3D face and the entire human body with various combinations of motion and illumination directions. We also show results of our theory in estimating 3D motion and illumination model parameters from a video sequence.

Index Terms—Motion, illumination, reflectance, bilinear, 3D structure.

1 Introduction

The appearance of an image sequence depends upon the following physical entities: the 3D model of the object being imaged, its surface properties and texture, illumination condition, object motion, and camera parameters. Understanding the interaction of these variables from a video sequence is one of the key challenges in computer vision. However, this is not an easy problem, in general, because of the complex relationships between the various parameters. As a result, two of the fundamental cues for image formation, motion and illumination, have been studied more or less independently. In this paper, we present a theory that combines motion, illumination and 3D structure in a single framework, using a perspective projection camera model. We derive an analytical expression for the joint space of motion and illumination variables. This result implies that, when the 3D model of an object at one instance of time is known, the reflectance images at future time instances under varying illumination and motion can be calculated using the subspace (which is shown to be approximately bilinear) computed at this instance. Conversely, given an arbitrary video sequence, it is possible to recover the 3D structure, motion, and illumination conditions simultaneously using the bilinear subspace formulation. Our theory is valid under the assumption of continuous motion (i.e., small motion in the case of discrete frames) and is thus applicable to most video sequences.

Following the work of [39], we distinguish the variabilities in a image sequence as originating from three general sources: photometric, geometric, and object characteristic. Photometric variability is due to change of the illumination conditions. Geometric variability is due to the change of pose and relative spatial location of the object with respect to the view point. Object characteristic variability is due to the object itself, including nonrigid deformations and texture change. In this paper, we concentrate on unifying the geometric and photometric variabilities. Object variabilities, like nonrigidity, will be handled in future work.

1.1 Related Work

Traditionally, motion and illumination (i.e., the geometric and photometric issues) have been studied separately. One of the classical methods for 2D motion estimation on the image plane is optical flow [18]. It assumes that the intensity of a particular point does not change over time. Estimation of 3D motion and structure, usually referred to as the Structure from Motion (SfM) [5], [1], [44], [42], [6] problem, is another classical research area in computer vision. While largely constrained to the analysis of rigid objects, it has been recently extended to nonrigid objects under orthographic projection [45]. For reconstructing 3D structure from discrete views obtained over a wide baseline, stereo reconstruction algorithms (and multicamera generalizations) have been proposed [8], [15]. However, most SfM and stereo reconstructions algorithms do not take illumination variation into consideration. To understand the inaccuracies that arise in the solution of the 3D reconstruction problems, a number of strategies for statistical analysis of the errors and robust statistical algorithms have been developed [7], [51], [48], [9], [35], [37], [36], [31], [28]. A method for shape reconstruction of a moving object under arbitrary, unknown illumination, assuming motion is known, was presented in [41]. Recently,

- *The authors are with the Department of Electrical Engineering, University of California, Riverside, CA 92521. E-mail: {yxu, amitrc}@ee.ucr.edu.*

Manuscript received 28 June 2005; revised 17 Mar. 2006; accepted 22 Aug. 2006; published online 18 Jan. 2007.
Recommended for acceptance by I.A. Essa.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-0337-0605.
Digital Object Identifier no. 10.1109/TPAMI.2007.1047.

0162-8828/07/$25.00 © 2007 IEEE Published by the IEEE Computer Society

Zhang et al. [49] have proposed modeling the change of illumination in optical flow and combine it with structure from motion, photometric stereo, and multiview stereo in an optimization framework. In [21], the authors proposed a multiview stereo algorithm that can estimate the three-dimensional shape and reflectance parameters under fixed illumination. However, none of the above methods provide an explicit expression relating the image and the motion, structure, and illumination variables for video sequences.

In the study of illumination, Shape from Shading (SfS) [10], [17], [29] is one of the earliest and most widely known methods. It is based on the Lambertian reflectance law and relies on the illumination information in a *single* image to estimate the 3D structure in a scene. Shashua [39] and Moses [26] proposed that, ignoring the effect of shadows, the set of images under varying illumination lies in a 3D linear subspace and derived the representation of the space. Using this fact and under the condition that the object and camera are fixed, they showed that three images obtained under three independent lighting conditions is sufficient to reconstruct the image set without prior knowledge of illumination conditions. This is known as Photometric Stereo and requires the object and camera to be fixed. When an uniform ambient illumination component is considered, the subspace of the image becomes 4D. Belhumeur and Kriegman [3] showed that the set of images of an object under arbitrary illumination forms a convex cone in the space of all possible images. Furthermore, they also proved that, when attached shadows are considered, the subspace dimension grows to infinity. However, most of the energy is packed in a limited number of lower order harmonics, thereby leading to a low-dimensional subspace approximation. In [2] and [33], the authors independently derived that it is possible to use low order spherical harmonics to accurately approximate the reflectance images. Specifically, they analytically derived a 9D spherical harmonics based linear representation of the images produced by a Lambertian object with attached shadows. An overall framework for modeling reflected light as a convolution of incident illumination with the bidirectional reflectance distribution functions, along with applications, was presented in [34]. However, these methods focus primarily on the problem of object recognition and are restricted to the analysis of single images. Extending the work in [2] directly to video sequences would require repeating the processes for each image separately. However, this is inefficient since the images of a moving object illuminated from a given light source over a short time period would be related based on the motion of the object. We exploit this fact to derive the joint illumination and motion space of video sequences [47].

Motivation for integrating the effects of motion and lighting comes largely from the field of object recognition, where pose and illumination invariant recognition is still an open problem. In the FRVT 2004 evaluation report (http://www.frvt.org/), illumination and pose variations are cited as being two of the major problems facing face recognition algorithms. In the mid-1990s, Murase and Nayar [27] proposed a method for pose and illumination invariant object recognition where an object is represented as a manifold in an eigenspace parameterized by pose and illumination variables. The input image is projected to this eigenspace and the object is recognized based on the manifold it lies on. However, this method needs a large set of images to construct the object eigenspace. Recently, there have been a number of algorithms for illumination invariant face recognition, most of which are based on the fact that the image of an object under varying illumination lies in a lower dimensional linear subspace. Lee et al. [23] try 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. They also proposed a novel method to model and recognize human faces in video sequences in [24]. In [50], 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. 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. Gross et al. [12], [13] proposed using Eigen Light-Fields and Fisher Light-Fields to do the pose invariant face recognition. They use 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. Zhou and Chellappa [52] used photometric stereo methods for face recognition under varying illumination and pose. In spite of the existance of many methods, pose and illumination variations remain a challenging problem in object recogntion and it is important to understand the interplay of motion and illumination in the process of image sequence formation. Also, most of these methods usually deal with recognition across discrete poses and do not consider continuous video sequences.

1.2 Overview of the Paper

In this paper, we develop a theory to characterize the interaction of motion and illumination in generating image sequences of a 3D object. We show that the set of all Lambertian reflectance functions of a moving object with attached shadows at any position, illuminated by arbitrarily distant light sources, lies "close"[1] to a *bilinear subspace* consisting of (approximately) nine illumination variables and six motion variables. Our work generalizes the results in [2] to video sequences. We consider the case of continuous motion and represent variations in surface norms and albedo up to a first order approximation. The bilinear subspace formulation can be used to simultaneously estimate the motion, illumination, and structure from a video sequence. Using this result, we synthesize video sequences of a 3D face with various combinations of motion and illumination directions. We further demonstrate the application of this theory to estimate 3D motion and lighting from a video sequence of a moving face under unknown varying illumination.

The rest of the paper is organized as follows: Section 2 presents previous work on the Lambertian Reflectance Linear Subspace (LRLS) method for modeling illumination in an image. It also provides an intuitive motivation for our theoretical derivation. Section 3 presents the theoretical derivation of the bilinear space of motion and illumination variables, with some of the mathematical details in the Appendix. In Section 4, experimental analysis of the accuracy of the theory and image synthesis results are presented. Section 5 shows the application of this theory in 3D motion estimation and its results. Section 6 concludes the paper and highlights future work.

[1] The Lambertian reflectance function actually lies in a nonlinear space, which is approximately bilinear, as we show later in the paper.

2 Previous Work and Motivation

Before we derive our theoretical results, we first review some basic definitions and previous work. Lambertian surfaces reflect light in all directions. According to Lambert's cosine law, the brightness of a specific point on Lambertian surface is proportional to the inner product of the surface normal and the incidence direction, as well as the energy per unit area on the surface, i.e.,

$$I = A\rho cos\theta,$$

where I is the reflectance intensity, A is the incident ray intensity, ρ is the albedo of the surface point, and θ is the angle between the surface norm and the direction of the incident ray.

The authors in [2] have proved that, when the 3D model is fixed, the set of the reflectance images can be decomposed by an infinite series of spherical harmonics functions. However, as the lower order spherical harmonics capture more energy, it is possible to use only a few spherical harmonics to approximate the image under varying illumination conditions. In the paper, they proved that the image can be approximated by a linear combination of the first nine spherical harmonics, which accounts for 99.22 percent of the energy. That is, the image lies close to a 9D linear subspace. They also show that the reflectance intensity for an image pixel (x, y) can be approximately expressed as

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

where I is the reflectance intensity of the pixel, 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, and b_{ij} are the basis images. 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, Y_{ij} is the spherical harmonics function, and $\mathbf{n}$ is the unit norm vector at the reflection point (please refer to [2] for more detail). Thus, (1) relates the 3D structure of the object (in terms of surface normals), the illumination direction, and albedo to the generated image. For brevity, we will refer to the work in [2] as the Lambertian Reflectance Linear Subspace (LRLS) theory.

The LRLS theory, as described in [2], is suitable for the situation that the 3D model and position are fixed and only illumination changes. This is because the basis images do not change as long as the 3D model and its position are fixed. This works for still images, but, when we consider the situation that the rigid object is moving, the basis images, b_{ij}, change from frame to frame. That is to say, for different time instances, the frames are not in the same linear subspace. If we want to use the method in [2] directly to video sequences, the basis images would have to be calculated for each frame. This is not only time-consuming, but also inefficient because it does not take into account the fact that the images of the moving object would be related over a short period of time. In this paper, we show how to take into account the motion of an object so as to combine the effects of motion, illumination, and 3D structure in generating a sequence of images.

3 Theoretical Derivation

In order to deal with both illumination and motion, we divide the problem into two stages. In the first stage, the object's motion is considered and the change in its position from one time instance to the other is calculated. We refer to this change of position as the *coordinate change* of the object. Then, in the next stage, we consider the effect of the incident illumination ray, which is projected onto the object and reflected according to the Lambert's cosine law. We will use the results in [2], [33] for the second stage of the problem and incorporate the effect of the motion.

Lambert's cosine law relates the direction and intensity of the light ray incident at a point of a 3D object, the albedo at the point, and the surface normal to the reflectance intensity at an image pixel that corresponds to the 3D surface point. If the 3D object is moving, then different points on that object can correspond to the same image point, i.e., they lie on the same ray passing through the image point. Let $\mathbf{P}$ and $\mathbf{Q}$ be two such points on the object that project to the same image point. Direction of illumination remaining constant, we need to estimate the change in the surface normal and albedo from point $\mathbf{P}$ to point $\mathbf{Q}$ in order to compute the reflectance intensity at the pixel as generated by this point. Our derivation of the bilinear subspace depends upon estimating the change in surface norm and albedo, which in turn depends upon the motion of the object.

3.1 Problem Formulation

In our problem, we need to consider only the relative motion between the camera and the object. We assume a perspective projection model for the camera. We fix the origin of the frame of reference to the center of the projection of the camera, the z-axis to be the optical axis, and assume that it passes through the center of the image. Hence, the focal length, f, of the camera is the only intrinsic parameter we consider.[2]

For the moment, assume that, at time instance t_1, we know the 3D model of the object, its pose, and the illumination condition in terms of the coefficients $l_{ij}^{t_1}$, see (1). Without loss of generality, we also assume that the pixel (x, y) corresponds to the point $\mathbf{P_1}$ at t_1. Thus, from the LRLS theory, we have the reflectance intensity for the pixel (x, y) as:

$$I(x, y, t_1) = \sum_{i=0,1,2} \sum_{j=-i,-i+1\ldots i-1,i} l_{ij}^{t_1} b_{ij}(\mathbf{n}_{\mathbf{P_1}}). \tag{3}$$

Let us define the the motion of the object in the above reference frame as the translation $\mathbf{T} = [T_x \quad T_y \quad T_z]^T$ of the centroid of the object and the rotation $\mathbf{\Omega} = [\omega_x \quad \omega_y \quad \omega_z]^T$ about the centroid. At the new time instance t_2, the illumination can change and is represented in terms of the coefficients $l_{ij}^{t_2}$. We will now derive the relationship between $I(x, y, t_1)$, $I(x, y, t_2)$, $\mathbf{T}$, $\mathbf{\Omega}$, $l_{ij}^{t_1}$, and $l_{ij}^{t_2}$.

The overall derivation of the joint motion and illumination space will proceed as follows: We will first derive the new basis images, taking into consideration the motion of the object. We show that the new bases are approximately of the form $(A\mathbf{T} + B\mathbf{\Omega})$, where A and B are suitably defined functions, the precise form of which we will derive. Next, incorporating the lighting parameters (which can be

[2] Radial distortion is ignored.

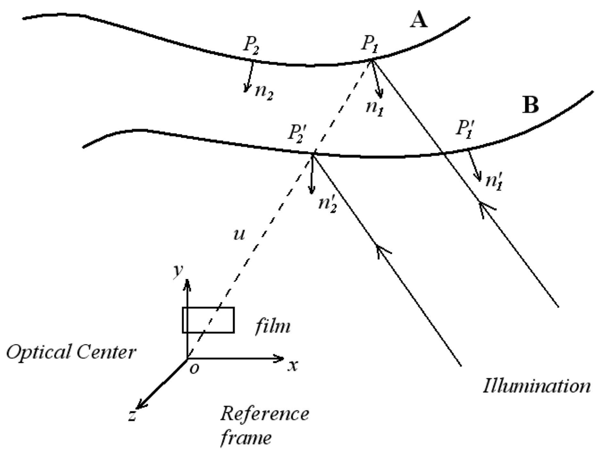


Fig. 1. Pictorial representation depicting imaging framework.

represented as a linear expansion using the LRLS theory), the joint motion and illumination space is shown to be bilinear.

3.2 Computation of the New Basis Image

Let A and B represent the same object before and after motion, respectively, as shown in Fig. 1. Consider the ray from the optical center to a particular pixel (x, y). We can find its intersection with the surface of the object by extending the ray. With respect to the camera, the direction of this ray does not change. Before the object's motion, the ray intersects with the surface at $\mathbf{P_1}$ (on A) and, after motion, it intersects at $\mathbf{P_2'}$ (on B). $\mathbf{P_1}$ (on A) moves to $\mathbf{P_1'}$ (on B) and $\mathbf{P_2}$ (on A) moves to $\mathbf{P_2'}$ (on B). Note that $\mathbf{P_2'}$ may not overlap with $\mathbf{P_1}$; they are just on the same projection ray. We will follow the convention of representing a point after motion with a prime ($'$).

We first define some notation required for our derivation. Let

$$\mathbf{J}_{\mathbf{P}_1} = \mathbf{J}\left(\frac{\partial \mathbf{n}_{\mathbf{P}_1}}{\partial \mathbf{P}}\right) \text{ and } \boldsymbol{\Delta} = \mathbf{P_2} - \mathbf{P_1} = \begin{pmatrix} \Delta x \\ \Delta y \\ \Delta z \end{pmatrix},$$

where $\mathbf{J}_{\mathbf{P}_1}$ is the Jacobian matrix of the norm, $\mathbf{n}_{\mathbf{P}_1}$, at point $\mathbf{P_1}$, with respect to $\mathbf{P} \triangleq (x, y, z)^T$, and $\boldsymbol{\Delta}$ is the difference in the coordinates of $\mathbf{P_2}$ and $\mathbf{P_1}$. Henceforth, we will refer to $\boldsymbol{\Delta}$ as the coordinate change.

From (1) and (2), we see that, when the illumination coefficients, l_{ij}, are known, only the norm and the albedo of the surface point of interest affect the reflection intensity at a particular pixel. The change in norm and albedo is obtained using the Jacobian matrix and gradient at the point of interest, as well as the coordinate change, which, in turn, is obtained from the motion information.

The norm changes from $\mathbf{P_1}$ to $\mathbf{P_2}$ and again from $\mathbf{P_2}$ to $\mathbf{P_2'}$. The first change is due to the fact that $\mathbf{P_2}$ is a different point on the surface, while the second change is due to the motion of the surface. Hence, the difference of $\mathbf{n}_{\mathbf{P}_1}$ and $\mathbf{n}_{\mathbf{P}_2'}$ is a function of the spatial (from $\mathbf{n}_{\mathbf{P}_1}$ to $\mathbf{n}_{\mathbf{P}_2}$) and temporal (from $\mathbf{n}_{\mathbf{P}_2}$ to $\mathbf{n}_{\mathbf{P}_2'}$) changes. Using the coordinate change $\boldsymbol{\Delta}$ and the Jacobian matrix of norm at $\mathbf{P_1}$, we are able to calculate the first order difference between $\mathbf{n}_{\mathbf{P}_1}$ and $\mathbf{n}_{\mathbf{P}_2}$. Using the motion information, we can obtain the difference between $\mathbf{n}_{\mathbf{P}_2}$ and $\mathbf{n}_{\mathbf{P}_2'}$. The albedo changes from $\mathbf{P_1}$ to $\mathbf{P_2}$, but is the same for $\mathbf{P_2}$ and $\mathbf{P_2'}$. Hence, the difference of $\rho_{\mathbf{P}_1}$ and $\rho_{\mathbf{P}_2'}$ is a function of spatial coordinates only, and can be obtained using the gradient of albedo. We can express the change in norm and albedo up to a first order approximation as

$$\Delta \mathbf{n} = \mathbf{n}_{\mathbf{P}_2'} - \mathbf{n}_{\mathbf{P}_1} = \mathbf{J}_{\mathbf{P}_1} \boldsymbol{\Delta} + \frac{\partial \mathbf{n}_{\mathbf{P}_2}}{\partial t} \Delta t \tag{4}$$

and

$$\Delta \rho = \rho_{\mathbf{P}_2'} - \rho_{\mathbf{P}_1} = \nabla \rho_{\mathbf{P}_1} \boldsymbol{\Delta}, \tag{5}$$

where $\nabla \rho_{\mathbf{P}_1}$ is the gradient of ρ at point $\mathbf{P_1}$. Thus, $\Delta \mathbf{n}$ and $\Delta \rho$ can be substituted into the expression for the basis images in (2), which can be rewritten as

$$\begin{aligned} b_{ij}(\mathbf{n}_{\mathbf{P}_2'}) &= (\rho_{\mathbf{P}_1} + \Delta\rho) r_i Y_{ij}(\mathbf{n}_{\mathbf{P}_1} + \Delta \mathbf{n}) \\ &= b_{ij}(\mathbf{n}_{\mathbf{P}_1}) + \nabla \rho_{\mathbf{P}_1} r_i Y_{ij}(\mathbf{n}_{\mathbf{P}_1}) \boldsymbol{\Delta} \\ &\quad + \rho_{\mathbf{P}_1} r_i \nabla Y_{ij}(\mathbf{n}_{\mathbf{P}_1}) \Delta \mathbf{n} + o(\boldsymbol{\Delta}). \end{aligned} \tag{6}$$

The last term is a higher order term and can be ignored when $\boldsymbol{\Delta}$ is small. Substituting $\Delta \mathbf{n}$ from (4), we see that the basis image is a linear function of $\boldsymbol{\Delta}$.

$$\begin{aligned} b_{ij}(\mathbf{n}_{\mathbf{P}_2'}) =& b_{ij}(\mathbf{n}_{\mathbf{P}_1}) \\ &+ \left(\nabla \rho_{\mathbf{P}_1} r_i Y_{ij}(\mathbf{n}_{\mathbf{P}_1}) \boldsymbol{\Delta} + \rho_{\mathbf{P}_1} r_i \nabla Y_{ij}(\mathbf{n}_{\mathbf{P}_1}) \mathbf{J}_{\mathbf{P}_1}\right) \boldsymbol{\Delta} \\ &+ \rho_{\mathbf{P}_1} r_i \nabla Y_{ij}(\mathbf{n}_{\mathbf{P}_1}) \frac{\partial \mathbf{n}_{\mathbf{P}_2}}{\partial t} \Delta t + o(\boldsymbol{\Delta}). \end{aligned} \tag{7}$$

$\frac{\partial \mathbf{n}_{\mathbf{P}_2}}{\partial t}$ is not a function of $\boldsymbol{\Delta}$, as we will show later in Section 3.4. We next show how to solve for $\boldsymbol{\Delta}$.

3.3 Computation of Coordinate Change $\boldsymbol{\Delta}$

Since $\mathbf{P_2'}$ and $\mathbf{P_1}$ are on the same ray, we can represent the difference between them using a unit vector $\mathbf{u}$ under the perspective camera model, i.e.,

$$\mathbf{P_2'} - \mathbf{P_1} = k\mathbf{u}, \tag{8}$$

where

$$\mathbf{u} = \frac{1}{\sqrt{x^2 + y^2 + f^2}} \begin{pmatrix} x \\ y \\ f \end{pmatrix} \tag{9}$$

and k is a scalar. Since the motion of the object is considered as a pure rotation with respect to its centroid and a pure translation of the centroid, the new coordinate of $\mathbf{P_2}$ can be expressed as

$$\mathbf{P_2'} = \mathbf{R}(\mathbf{P_2} - \mathbf{T_0}) + \mathbf{T_0} + \mathbf{T}, \tag{10}$$

where $\mathbf{R}$ is the Rodrigues rotation matrix obtained from the rotation $\boldsymbol{\Omega}$ with respect to the centroid and $\mathbf{T_0}$ is the position of the centroid of the object. Substituting it into (7), we get

$$k\mathbf{u} = \mathbf{R}(\mathbf{P_2} - \mathbf{T_0}) + \mathbf{T_0} + \mathbf{T} - \mathbf{P_1}. \tag{11}$$

Under the assumption of small motion, we have an additional constraint. We may consider the new point $\mathbf{P_2}$ to be on the tangent plane that passes through the original intersection point $\mathbf{P_1}$, i.e.,

$$\mathbf{n}_{\mathbf{P}_1}^T (\mathbf{P_1} - \mathbf{P_2}) = 0. \tag{12}$$

Using (11) and (12) and, after some algebraic manipulation (see Appendix A), we can show that

$$\begin{aligned}\mathbf{\Delta} &= (\mathbf{R}^{-1} - \mathbf{I})(\mathbf{P_1} - \mathbf{T_0}) - \mathbf{R}^{-1}\mathbf{T} \\ &\quad - \mathbf{R}^{-1}\frac{\mathbf{n}_{\mathbf{P_1}}^T\big((\mathbf{R}^{-1} - \mathbf{I})(\mathbf{P_1} - \mathbf{T_0}) - \mathbf{R}^{-1}\mathbf{T}\big)}{\mathbf{n}_{\mathbf{P_1}}^T\mathbf{R}^{-1}\mathbf{u}}\mathbf{u}.\end{aligned} \tag{13}$$

The coordinate change, $\mathbf{\Delta}$, obtained in (13) captures the effect of the motion. However, as it is a nonlinear function of the object motion variables $\mathbf{T}$ and $\mathbf{\Omega}$, its complex form makes it difficult to analyze. Henceforth, we will denote this as $\mathbf{\Delta}_{nl}$.

Since the motion is small, we can simplify the above equation using certain approximations that neglect terms with small magnitude with respect to terms with large magnitudes. This will allow us to interpret the joint effect of motion and illumination analytically, while sacrificing little in terms of accuracy. Using a series of mathematical calculations, we can obtain $\mathbf{\Delta}$ as a linear function of the motion variables (see Appendix B) as:

$$\begin{aligned}\mathbf{\Delta} &\cong \hat{\mathbf{P}}\mathbf{\Omega} + \mathbf{T} - \frac{1}{\mathbf{u}^T\mathbf{n}_{\mathbf{P_1}}}\mathbf{u}\mathbf{n}_{\mathbf{P_1}}^T\hat{\mathbf{P}}\mathbf{\Omega} - \frac{1}{\mathbf{u}^T\mathbf{n}_{\mathbf{P_1}}}\mathbf{u}\mathbf{n}_{\mathbf{P_1}}^T\mathbf{T} \\ &= \left(\mathbf{I} - \frac{1}{\mathbf{n}_{\mathbf{P_1}}^T\mathbf{u}}\mathbf{u}\mathbf{n}_{\mathbf{P_1}}^T\right)(\hat{\mathbf{P}}\mathbf{\Omega} - \mathbf{T}) \\ &\triangleq \mathbf{C}(\hat{\mathbf{P}}\mathbf{\Omega} - \mathbf{T}),\end{aligned} \tag{14}$$

where $\hat{\mathbf{P}} = (\mathbf{P_1} - \mathbf{T_0})^\wedge$.[3]

We will refer to this as $\mathbf{\Delta}_l$. Henceforth, when we use $\mathbf{\Delta}$, we will refer to $\mathbf{\Delta}_l$; when required to be specific, we will mention $\mathbf{\Delta}_l$ or $\mathbf{\Delta}_{nl}$.

3.4 Temporal Change of Norm

In order to obtain the change of norm $\Delta\mathbf{n}$, we still need to compute the effect of temporal change on the right-hand side (RHS) of (4). Using the assumption of small motion, we can compute:

$$\begin{aligned}\frac{\partial \mathbf{n}_{\mathbf{P_2}}}{\partial t}\Delta &= \frac{\partial(\mathbf{n}_{\mathbf{P}_1} + \mathbf{J}_{\mathbf{P}_1}\mathbf{\Delta})}{\partial t} = \mathbf{\Omega} \times (\mathbf{n}_{\mathbf{P}_1} + \mathbf{J}_{\mathbf{P}_1}\mathbf{\Delta}) \\ &= \mathbf{\Omega} \times \mathbf{n}_{\mathbf{P}_1} + o(\mathbf{\Omega}\mathbf{T}) \cong (-\mathbf{n}_{\mathbf{P}_1})^\wedge\mathbf{\Omega} \\ &\triangleq -\hat{\mathbf{N}}\mathbf{\Omega}.\end{aligned} \tag{15}$$

As $\mathbf{\Delta}$ is a linear function of the motion variables $\mathbf{\Omega}$ and $\mathbf{T}$, the cross product of $\mathbf{\Omega}$ and $\mathbf{J}_{\mathbf{P}_1}\mathbf{\Delta}$ is a second order term and can be ignored when the motion is small. Thus, the temporal change is not a function of $\mathbf{\Delta}$, a fact that was used in (7).

3.5 Bilinear Space of Motion and Illumination

Substituting (14) and (15) into (4), we get a linear expression for $\Delta\mathbf{n}$ as a function of motion variables, i.e.,

$$\Delta\mathbf{n} = (\mathbf{J}_{\mathbf{P}_1}\mathbf{C}\hat{\mathbf{P}} - \hat{\mathbf{N}})\mathbf{\Omega} - \mathbf{J}_{\mathbf{P}_1}\mathbf{C}\mathbf{T}. \tag{16}$$

So far, we have expressed the coordinate and norm change as linear expressions of the motion variables. Substituting (14) and (16) into (1) and (7), which contain the illumination variables, we have

$$I(x, y, t_2) = \sum_{i=0,1,2}\ \sum_{j=-i,-i+1\ldots i-1,i} l_{ij}^{t_2} b_{ij}(\mathbf{n}_{\mathbf{P}'_2}), \tag{17}$$

where

$$b_{ij}(\mathbf{n}_{\mathbf{P}'_2}) = b_{ij}(\mathbf{n}_{\mathbf{P}_1}) + \mathbf{A}\mathbf{T} + \mathbf{B}\mathbf{\Omega}, \tag{18}$$

$$\mathbf{A} = -r_i\big(\nabla\rho_{\mathbf{P}_1}Y_{ij}(\mathbf{n}_{\mathbf{P}_1}) + \rho_{\mathbf{P}_1}\nabla Y_{ij}(\mathbf{n}_{\mathbf{P}_1})\mathbf{J}_{\mathbf{P}_1}\big)\mathbf{C}, \tag{19}$$

and

$$\mathbf{B} = -\mathbf{A}\hat{\mathbf{P}} - r_i\rho_{\mathbf{P}_1}\nabla Y_{ij}(\mathbf{n}_{\mathbf{P}_1})\hat{\mathbf{N}}. \tag{20}$$

In (18), $b_{ij}(\mathbf{n}_{\mathbf{P}'_2})$ are the basis images after motion. The first term, $b_{ij}(\mathbf{n}_{\mathbf{P}_1})$, are the original basis images before motion. They are only determined by the object model and do not change with the variation of illumination. The illumination change is reflected in the change of the coefficients from $l_{ij}^{t_1}$ to $l_{ij}^{t_2}$. The effect of the motion is reflected in $\mathbf{A}\mathbf{T} + \mathbf{B}\mathbf{\Omega}$, where the first term describes the effect of translation, and the second term describes the effect of rotation. Substituting (18) into (17), we see that the new image spans a bilinear space of the motion variables and illumination variables.

When the illumination changes gradually, we may 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, defined by the motion and illumination variables.

$$\begin{aligned}I(x, y, t_2) &= I(x, y, t_1) + \sum_{i=0,1,2}\ \sum_{j=-i,\ldots,i} l_{ij}^{t_1}(\mathbf{A}\mathbf{T} + \mathbf{B}\mathbf{\Omega}) \\ &\quad + \sum_{i=0,1,2}\ \sum_{j=-i,\ldots,i} \Delta l_{ij} b_{ij}(\mathbf{n}_{\mathbf{P}_1}).\end{aligned} \tag{21}$$

If the illumination does not change from t_1 to t_2 (often a valid assumption for a short interval of time), we see that the new image at t_2 spans a linear space of the motion variables since the third term in (21) is zero.

3.6 Tensor Notation

We can express the above result succinctly using tensor notation as

$$\mathcal{I} = \left(\mathcal{B} + \mathcal{C} \times_2 \begin{pmatrix}\mathbf{T} \\ \mathbf{\Omega}\end{pmatrix}\right) \times_1 \mathbf{l}, \tag{22}$$

where $\times_n$ is called the *mode-n product* [22] and $\mathbf{l} \in \mathbb{R}^{\mathbf{9}}$ is the vector of l_{ij} components. The *mode-n product* of a tensor $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_n \times \ldots \times I_N}$ by a vector $\mathbf{V} \in \mathbb{R}^{1 \times I_n}$, denoted by $\mathcal{A} \times_n \mathbf{V}$, is the $I_1 \times I_2 \times \ldots \times 1 \times \ldots \times I_N$ 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, $\mathcal{C}_{klpq} = [\,\mathbf{A} \quad \mathbf{B}\,]$ of size 9×6. Thus, for an image of size $M \times N$, $\mathcal{C}$ is $9 \times 6 \times M \times N$. $\mathcal{B}$ is a subtensor of dimension $9 \times 1 \times M \times N$, comprised of 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. l is still the 9×1 vector of the illumination coefficients.

[3] We define the skew symmetric matrix of a vector

$$\mathbf{X} = \begin{pmatrix}x_1 \\ x_2 \\ x_3\end{pmatrix}$$

as

$$\mathbf{X}^\wedge = \hat{\mathbf{X}} = \begin{pmatrix}0 & -x_3 & x_2 \\ x_3 & 0 & -x_1 \\ -x_2 & x_1 & 0\end{pmatrix}.$$

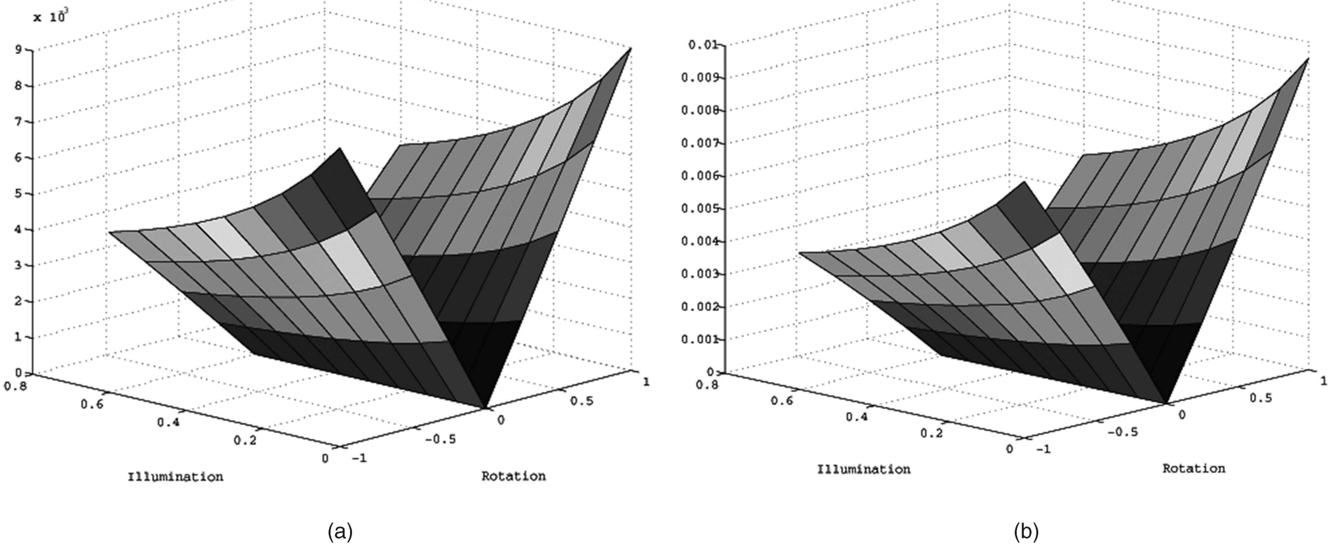


Fig. 2. (a) Shows the normalized error for one particular pixel between true intensity and the estimate obtained with linear approximation, $\mathbf{\Delta}_l$, in (14). (b) Shows the normalized error for the same pixel between the true intensity and the estimate with nonlinear approximation, $\mathbf{\Delta}_{nl}$, in (13). The error is plotted for a typical variation range for l_{11} and rotation of the object. The form of the error surface is similar for the other motion and illumination changes.

In [46], the authors assumed the face image lies in a multilinear space parameterized by factors like illumination, viewpoint, identity, expression, etc., and then used Multilinear Independent Components Analysis to recognize faces. Our result provides a theoretical underpinning for this assumed model.

3.7 Discussion on the Theoretical Result

Physical Interpretation. This bilinear space result integrates the effects of illumination and motion in generating an image from a 3D object using a perspective camera. When the object does not move, the second and third motion terms of the basis image $b_{ij}(\mathbf{n}_{\mathbf{P}'_2})$ are zero and the result is the same as the one in [2], a 9D Lambertian Reflectance Linear Subspace. When the illumination remains the same, the reflectance image spans a linear subspace of motion variables. When the illumination and motion variables all change, the image space is "close to" bilinear. Thus, the joint illumination and motion space for a sequence of images is bilinear with (approximately) nine illumination variables and six motion variables. The shape of the object is encoded in the $\mathbf{A}$ and $\mathbf{B}$ matrices and in $b_{ij}(\mathbf{n}_{\mathbf{P}_1})$. The camera intrinsic parameters are implicitly present in $\mathbf{\Delta}$ (thus, in $\mathbf{A}$ and $\mathbf{B}$) through $\mathbf{u}$. Therefore, (17) and (18) integrate the motion, illumination, 3D structure, albedo, and camera intrinsic parameters into one single framework.

Effect of Linearization. Fig. 2 shows the effect of approximating the nonlinear function $\mathbf{\Delta}_{nl}$ in (13) with a linear approximation $\mathbf{\Delta}_l$. We plot the difference in the image intensity (| true intensity - estimated intensity using $\mathbf{\Delta}_{nl}$| and | true intensity - estimated intensity using $\mathbf{\Delta}_l$|) at a particular point as a function of $l_{11}^{t_1}$ and ω_y. The rotation range is defined as in Section 4 and the illumination changes in a typical range. We take the intensity obtained from the LRLS method as the true value. The difference in Fig. 2a is computed between the true value and the intensity obtained with the linear expression of $\mathbf{\Delta}_l$ (using (14)) and normalized with regard to the true one. The difference in Fig. 2b is computed between the true value and the intensity obtained with the nonlinear expression of $\mathbf{\Delta}_{nl}$ (using (13)) and also normalized with regard to the true one. As can be seen, there is no perceptible difference between the bilinear and nonlinear image spaces.

Generalizations of the theory. Even though the above result is derived using previous work on the LRLS theory, the basic result (i.e., the joint motion and illumination space is bilinear with the bases of this space determined by the surface normals and camera intrinsic parameters) is valid in more general circumstances. If we can write the image appearance as a linear dot product of lighting coefficients and basis images and, if the basis images change linearly with the 3D rigid motion parameters, the joint motion and illumination space will be bilinear. This could be achieved using higher order coefficients in the spherical harmonics representation of illumination or a different set of basis functions [43], [32]. However, for other basis functions, the precise form of the expression would have to be rederived, while using higher order spherical harmonics coefficients would require imposing additional constraints to enforce nonnegativity of the lighting function (see [19] for details). Also, for glossy surfaces, the gradient of the albedo can have high frequency components which can affect the parameter estimates in scene understanding applications.

Effect of scale changes. To understand this, we consider that the motion is purely in the direction of the optical axis, i.e., zooming effect. Irrespective of how the objects move, (11), is satisfied. Thus, even when the object moves toward the camera, the intersection points of a ray with the object surface at two consecutive time instances should still be close to each other, provided the motion is small. Therefore, $\mathbf{P_2}$ can still be considered to be on the tangent plane passing through $\mathbf{P_1}$. So, (11) and (12) are satisfied and the coordinate change $\mathbf{\Delta}$, which completely determines the change of norm and albedo, can be calculated accurately. We will show some images of this case in Section 4.

The motion of a plane. When the plane moves with pure translation, there is no difference between $\mathbf{n}_{\mathbf{P}_2}$ and $\mathbf{n}_{\mathbf{P}'_2}$ (in

Fig. 1); thus, the change of norm is completely due to the spatial component in (4). When the object plane moves with pure rotation confined to the image plane, the rotation axis is parallel to the norm on the object plane; thus, $\mathbf{n}_{\mathbf{P}_2}$ and $\mathbf{n}_{\mathbf{P}'_2}$ are the same, so the change of norm again has only the spatial component. When the object plane purely rotates but the rotation is not confined to the image plane, there will be both spatial and temporal change of the norm. So, the change of norm, which determines the reflectance intensity, can be described by the theory. The albedo change is the same as in the main theory (see (5)).

Pixels for which the unit vector $\mathbf{u}$ is perpendicular to the norm. In this case, (11) is still satisfied; however, because the ray is now coplanar with the tangent plane passing through $\mathbf{P}_1$, there are an infinite number of solutions for (12). In implementing the theory, this affects all points for which the angle between the unit vector $\mathbf{u}$ and the norm $\mathbf{n}_{\mathbf{P}}$ are very close to 90°, making the denominators in (13), (14), (24), (25), (26), (28), (30), (31), (32), (34) very small. In this case, the two constraints ((11), (12)) for calculating the coordinate change $\boldsymbol{\Delta}$ become only one and it is not possible to compute $\boldsymbol{\Delta}$. However, this happens only at a very few points near the object's edge (e.g., near the edge of a face) and is not a serious impediment to the application of the theory in practical problems. In the implementation, the value of the pixels where this happens is replaced with values from nearby pixels. This is not a shortcoming of our theory since it is not possible to view a point if the viewing direction and the surface normal are perpendicular (there is no light reflected along the viewing direction).

Applications of the theory. The theoretical result has important applications in image sysnthesis, illumination invariant tracking, 3D modeling, object recognition, video compression, and others. All of these methods 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 Experimental Analysis

In this section, we experimentally analyze the theoretical results obtained above. Specifically, we show the accuracy of an image obtained by the above model with a true image. We also show the results of synthesized images.

The above derivation is based on the small motion assumption, which is used in three places. First, it is used to obtain (12) by making the tangent plane approximation. Next, the small motion assumption is used to obtain the linear approximation of (14). The third place where it is used is the first order approximation of the norm and albedo. We show that the effect of this assumption on the resultant video sequences is very small.

For the sake of brevity, we show the effect of the translation along and rotation about the y-axis on the change of albedo, norm and $\boldsymbol{\Delta}$. We also compare the synthesized images with those obtained with LRLS theory. The results are similar for other combinations of motion. For the experimental error analysis in Fig. 3, the translation is normalized with respect to the width of the face and the unit of the rotation is degree. In addition, the pose is fixed as the front view and the illumination is fixed from the front of the face. In this experiment, we calculated the errors in a typical motion range. We assume the largest distance the face can move along the positive and negative directions of y axis in one second is half of the width of the face. We also assume that the largest angle the face rotates in one second is 30°. Using the convention of 30 frames per second, we can get that the maximum translation between the consequent frames is $\frac{0.5}{30} = 0.0167$ of the width of the face (henceforth referred to as the 0.0167 normalized translation unit) and the maximum rotation between consequent frames is $\frac{30^\circ}{30} = 1^\circ$. So, we calculate the error in the range of -0.020 normalized translation units to +0.020 normalized translation units and the rotation from -1.00° to $+1.00^\circ$. In addition, because of the discontinuity effects at the extremities of the face, there may be a few points with large errors. To avoid the bias caused by these points, we represent the total error using the median of the errors of all the points.

Fig. 3a depicts the difference between $\boldsymbol{\Delta}_{nl}$ and $\boldsymbol{\Delta}_{l}$ normalized with regard to $\boldsymbol{\Delta}_{nl}$. Within the typical motion range defined above, the largest relative error of the linear solution (with regard to the nonlinear solution) is about 5 percent. Next, in Figs. 3b and 3c, we compute the error introduced by the first order approximation of $\mathbf{n}_{\mathbf{P}'_2}$ and $\rho_{\mathbf{P}'_2}$ (see (4) and (5)). We compute the normalized error as the difference of these variables obtained using our theory and those obtained using the LRLS theory in [2] (see Section 2) and normalized with regard to the LRLS results obtained for each image separately. Fig. 3d gives the normalized error of the image obtained with the bilinear approximation of (17) and (18). We see that the maximum error in all of the above cases is about 5 percent. Typically, the motion between the consecutive frames is much smaller than the extremities of the above range; hence, the difference in practice is about $2 \sim 3$ percent. Moreover, if we consider only rotation, the error at the extremities of the above range is $1 \sim 2$ percent. (See Figs. 2a and 2b.) Fig. 3e computes the normalized error of the images obtained with the nonlinear expression of coordinate change in (13). The normalized error of the images obtained with linear coordinate change (14) in Fig. 3d and nonlinear coordinate change (13) in Fig. 3e are very similar, which validates the approximations in the linearization part of the derivation.

Finally, we synthesized a video sequence of a rotating face with our theory and LRLS theory respectively. The image resolutions is 240×320 pixels for all the images and the generic face model is rotating along y axis from -30° to $+30^\circ$. Illumination changes in the same way as pose, and always comes from the front of the face. Fig. 3f gives the normalized error of the video sequence synthesized with our theory. The maximum error is about 5 percent, though, as we show next, there is no perceptible difference in image quality.[4] Moreover, the computational complexity for generating a sequence of images using our theory is much lower than that using the LRLS theory. The time taken to compute the first frame is the same in both cases; for subsequent frames, LRLS has to repeat the same procedure, while our approach uses the bilinear

[4] The periodicity appears because we do the reinitialization for every 20 frames.

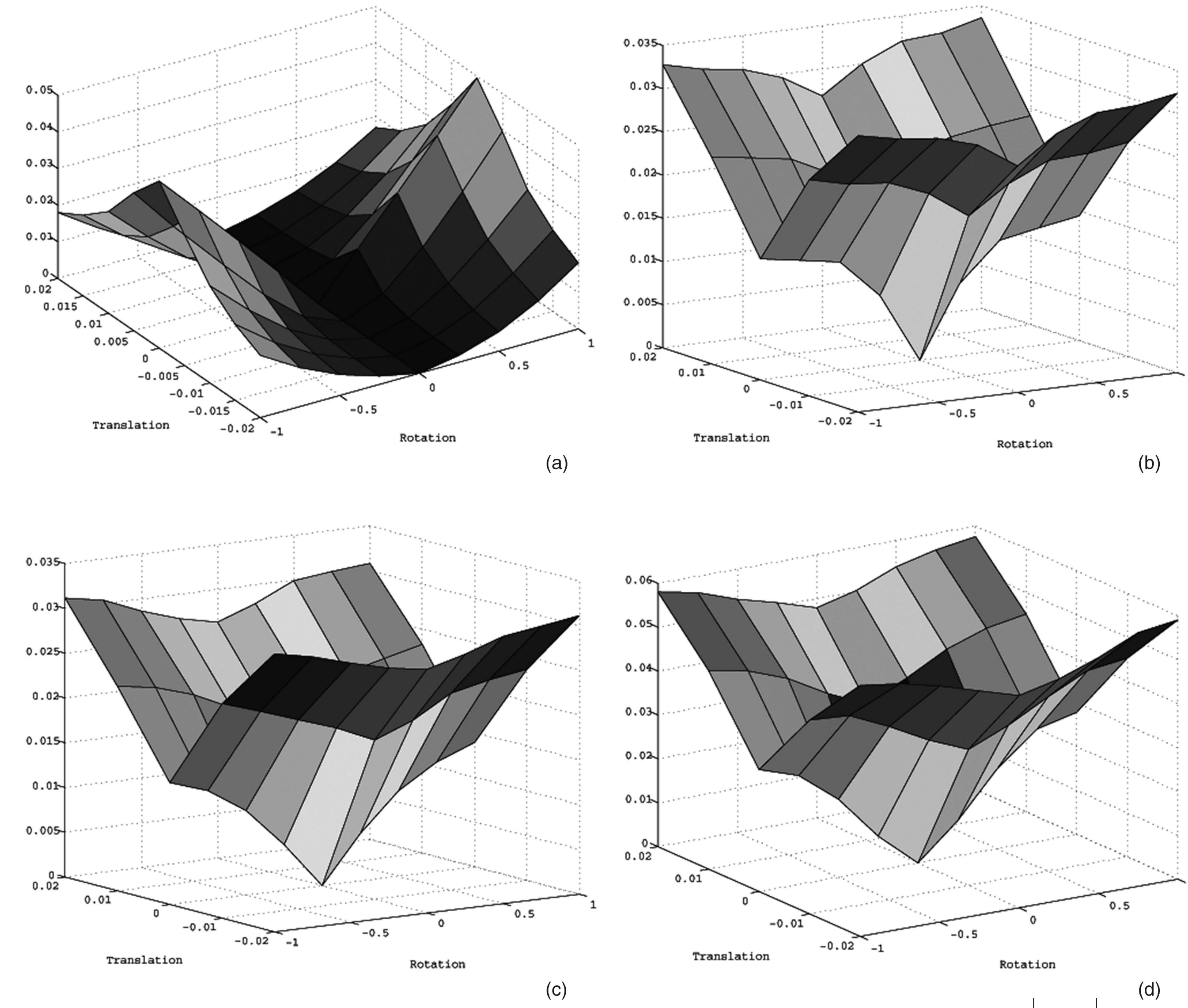
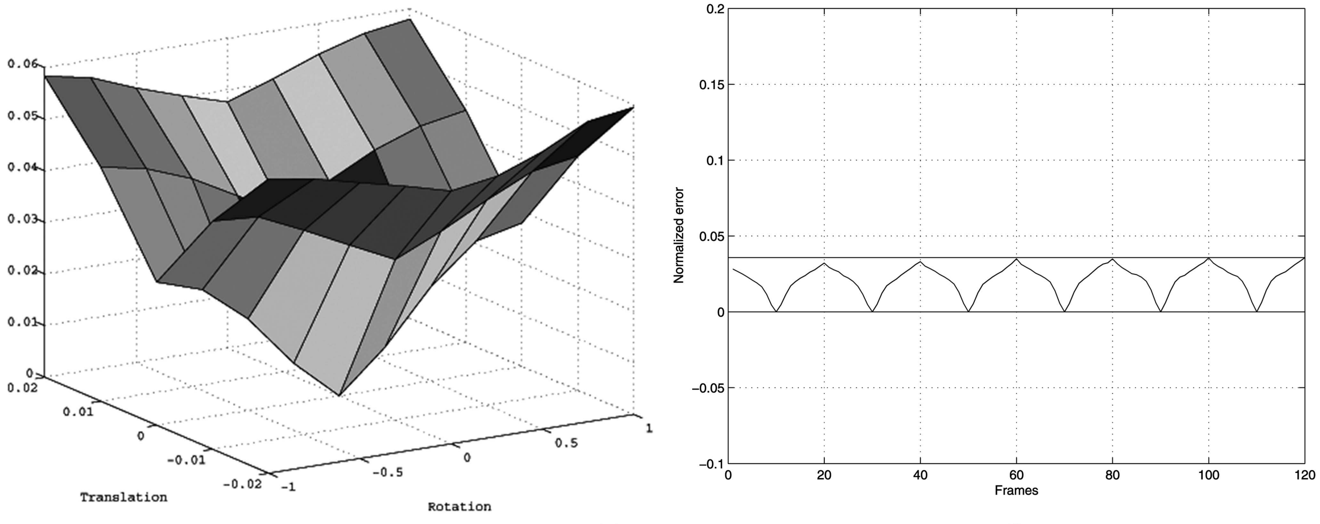


Fig. 3. (a) Normalized difference of the linear and nonlinear coordinate change, $\frac{|\boldsymbol{\Delta}_l-\boldsymbol{\Delta}_{nl}|}{|\boldsymbol{\Delta}_{nl}|}$. (b) Normalized difference of the norm, $\frac{\left|\mathbf{n}_{\mathbf{P}_2'}-\bar{\mathbf{n}}_{\mathbf{P}_2'}\right|}{\left|\bar{\mathbf{n}}_{\mathbf{P}_2'}\right|}$, where $\mathbf{n}_{\mathbf{P}_2'}$ is the first order approximation of the norm at point $\mathbf{P}_2'$ with linearized coordinate change and $\bar{\mathbf{n}}_{\mathbf{P}_2'}$ is the true value of the norm at point $\mathbf{P}_2'$. (c) Normalized difference of the albedo change, $\frac{\left|\rho_{\mathbf{P}_2}-\bar{\rho}_{\mathbf{P}_2}\right|}{\left|\bar{\rho}_{\mathbf{P}_2'}\right|}$, where $\rho_{\mathbf{P}_2}$ is the first order approximation of the albedo at point $\mathbf{P}_2'$ with linearized coordinate change and $\bar{\rho}_{\mathbf{P}_2'}$ is the true value of the albedo at point $\mathbf{P}_2'$. (d) Normalized difference of the synthesized image, $\frac{\left|I(.,.,t_2)^l-I(.,.)^{LRLS}\right|}{\left|I(.,.)^{LRLS}\right|}$, where $I(.,.,t_2)^l$ is the image generated by linear coordinate change $\boldsymbol{\Delta}_l$ and $I(.,.)^{LRLS}$ is the image obtained with the LRLS theory.

space of (17) and (18), the computation of which is very fast. In our implementation (which has not been optimized for efficiency), generating each frame using LRLS theory will take 15 $\sim$ 20 seconds, while generating 20 frames with the bilinear space takes almost the same time.

Next, we applied our theory to a 3D face to synthesize image sequences for different combinations of motion and illumination directions using the bilinear space theory. In Fig. 4, the pose of the 3D face is fixed and illumination is rotating with respect to the face. From (17) and (18), $\mathbf{T}$ and $\boldsymbol{\Omega}$ are zero, basis images b_{ij} remain the same, and only l_{ij} change. Thus, all of the images lie in a linear subspace of l_{ij}. The results obtained here are the same as using the LRLS theory. In Fig. 5, illumination is fixed but the face is rotating about y-axis from $-25°$ to $+25°$, thus l_{ij} are fixed and b_{ij} is a linear function of $\mathbf{T}$ and $\boldsymbol{\Omega}$; thus, $I(x, y, t_2)$ lies in a linear subspace of the motion variables. For comparison, we also show the results using the LRLS theory repeated for each pose. There is no perceptible difference between the images synthesized by the two methods, while the time taken for synthesizing is drastically less than that for LRLS theory. (see the last sentence of the previous paragraph).

In Fig. 6, we show synthesis results using the data in [4]. The face is moving and the illumination always comes from the front of the face, thus b_{ij} is a linear function of $\mathbf{T}$ and $\boldsymbol{\Omega}$, and $I(x, y, t_2)$ is the combination of b_{ij} with varying coefficients l_{ij}. Thus, $I(x, y, t_2)$ lies in a bilinear space of the illumination and motion variables. We also show



(e) (f)

Fig. 3 (continued). (e) Normalized difference of the synthesized image, $\frac{|I(.,.,t_2)^{nl} - I(.,.)^{LRLS}|}{|I(.,.)^{LRLS}|}$, where $I(.,.,t_2)^{nl}$ is the image generated by the nonlinear coordinate change Δ_{nl}. (f) Normalized error as in (d), plotted as a function of time for a video sequence.

synthesized images when the face is moving toward the camera. In Fig. 7, we use the human body model to synthesize image sequences under varying illumination. The human body is rotating about the vertical axis and the illumination is changing in the same way.

5 Application to 3D Motion Estimation

In this section, we demonstrate the application of the theory developed earlier in the paper to the problem of 3D motion estimation. We also recover the illumination model parameters. The input is a video sequence captured under arbitrary conditions of motion and illumination. We would like to emphasize that this section provides an example of an application of the theory in tracking an object under variable illumination. It is not possible to include more generalized tracking scenarios (e.g., occlusion, clutter, multiple objects) within the constraints of this paper. They will be the focus of future work.

Estimating 3D motion from video has been one of the most extensively studied problems in computer vision. One of the

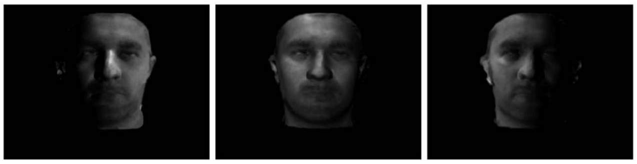


Fig. 4. Reflectance images under fixed pose and rotating illumination. All these images lie in a linear subspace of illumination variables.

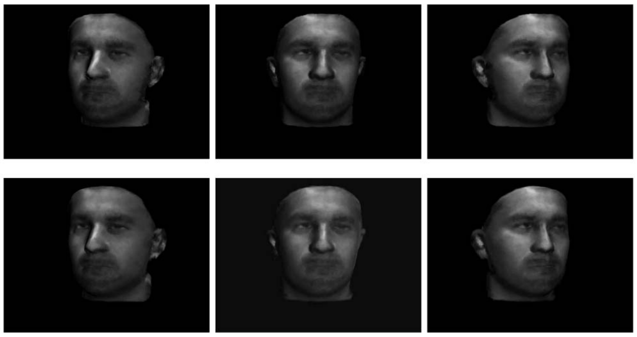
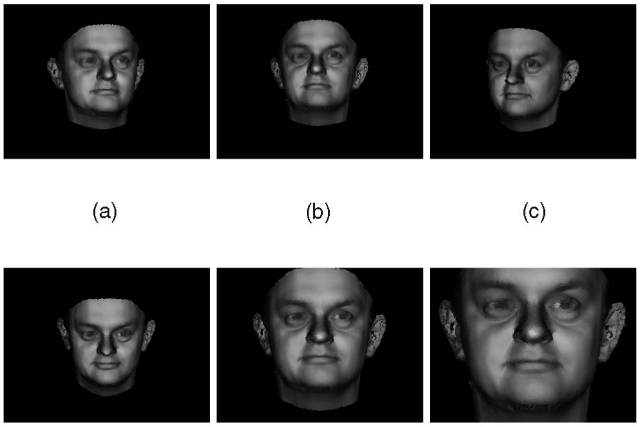


Fig. 5. Reflectance images of the face rotating along the vertical axis under fixed illumination. The images in the upper row are generated by our theory and the images in the lower row are generated by the LRLS theory repeated for each pose.



(a) (b) (c)

(d) (e) (f)

Fig. 6. Reflectance images of the moving face with changing illumination directions. Illumination changes in the same way as pose and always comes from the front of the face. We also show synthesized images when the face moves toward the camera. The results are obtained using the data from the Three-Dimensional Morphable Face Model described in [4].

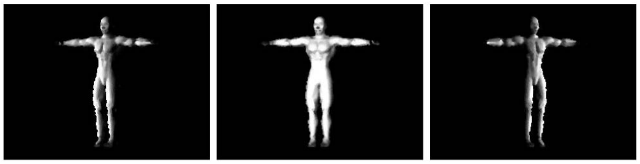


Fig. 7. Reflectance images of a rotating human body under varying illumination generated by the bilinear space.

well-known approaches is to use optical flow and the SfM formulation to reconstruct 3D motion [48], [9]. However, optical flow involves the brightness constancy constraint, which is violated when the illumination is static, but the object moves relative to the direction of illumination. Pentland [30] 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. Zhang et al. [49] modeled lighting changes by introducing illumination-specific parameters into the standard optical flow equation. Another well-known approach for 2D motion estimation in monocular sequences is the Kanade-Lucas-Tomasi (KLT) tracker [40], [20], which selects features that are optimal for tracking. Hager and Belhumeur [14] proposed using a parameterized function to describe the movement of the image points, taking into account illumination variation by modifying the brightness constancy constraint. Freedman and Turek in [11] proposed graph algorithms for illumination invariant tracking. All of these methods deal with estimation of 2D motion. Illumination invariant 3D tracking is considered within the Active Appearance Model (AAM) framework in [38], but the method requires training images. A review of 3D model-based motion estimation algorithms is available in [25], but most of them do not have explicit illumination models.

Our theory, which provides an explicit expression for all of the images that an object can produce under arbitrary motion and illumination conditions, allows us to develop a theoretical framework for estimating the 3D motion of an object under changing illumination conditions. Also, we are able to recover the illumination conditions as a function of time. This is based on inverting the generative model for motion and illumination modeling. We can handle the cases of gradual and sudden change of complicated lighting patterns that include combinations of point and extended sources. We assume that an approximate 3D model of the object that we are trying to track is available (e.g., a generic model of a face) and the tracking algorithm is initialized by registering it to the first frame of the sequence.

5.1 Illumination Invariant 3D Motion Estimation

Equation (17) provides us an expression relating the reflectance image I_{t2} with illumination coefficients $l_{ij}^{t_2}$ and motion variables $\mathbf{T}$, $\mathbf{\Omega}$, which leads to a method for estimating 3D motion and illumination as:

$$\begin{aligned}(\hat{\mathbf{l}}, \hat{\mathbf{T}}, \hat{\mathbf{\Omega}}) &= \arg\min_{\mathbf{l},\mathbf{T},\mathbf{\Omega}} \| I_{t2} - \sum_{i=0,1,2} \sum_{j=-i}^{i} l_{ij}^{t_2} b_{ij}(\mathbf{n}_{\mathbf{P}'_2}) \|^2 \\ &= \arg\min_{\mathbf{l},\mathbf{T},\mathbf{\Omega}} \left\| \mathcal{I}_{t2} - \left(\mathcal{B}_{t1} + \mathcal{C}_{t1} \times_2 \begin{pmatrix} \mathbf{T}_{t2} \\ \mathbf{\Omega}_{t2} \end{pmatrix} \right) \times_1 \mathbf{l}_{t2} \right\|^2,\end{aligned} \tag{23}$$

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 and motion and illumination estimates are obtained for each frame. Since the image I_{t2} lies approximately in a bilinear space of illumination and motion variables, 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[5] $\mathcal{B}$ and the image $\mathcal{I}$ along the first dimension [22], which is the illumination dimension, the image can be represented as:

$$\mathcal{I}_{(1)}^T = \mathcal{B}_{(1)}^T \mathbf{l}. \tag{24}$$

This is a least square problem and the illumination l can be estimated as:

$$\hat{\mathbf{l}} = (\mathcal{B}_{(1)} \mathcal{B}_{(1)}^T)^{-1} \mathcal{B}_{(1)} \mathcal{I}_{(1)}^T. \tag{25}$$

Keeping the illumination coefficients fixed, the bilinear space in (17) and (18) becomes a linear subspace, i.e.,

$$\mathcal{I} = \mathcal{B} \times_1 \mathbf{l} + (\mathcal{C} \times_1 \mathbf{l}) \times_2 \begin{pmatrix} \mathbf{T} \\ \mathbf{\Omega} \end{pmatrix}. \tag{26}$$

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

$$\begin{aligned}\begin{pmatrix} \hat{\mathbf{T}} \\ \hat{\mathbf{\Omega}} \end{pmatrix} &= \left((\mathcal{C} \times_1 \mathbf{l})_{(2)} (\mathcal{C} \times_1 \mathbf{l})_{(2)}^T \right)^{-1} (\mathcal{C} \times_1 \mathbf{l})_{(2)} \\ &\quad \times (\mathcal{I} - \mathcal{B} \times_1 \mathbf{l})_{(2)}^T.\end{aligned} \tag{27}$$

This can be repeated for each subsequent frame. 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. We now describe the tracking algorithm formally.

Tracking Algorithm

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

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. Use LRLS method [2] to calculate the tensor of the basis images $\mathcal{B}_0$ at this pose. Use (25) 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 (18). Use (27) to estimate the new pose from the estimated motion.

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

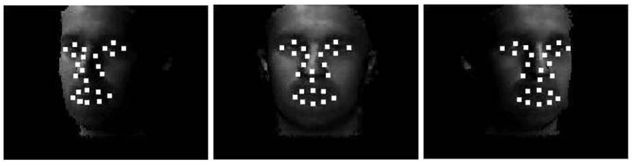


Fig. 8. The back projection of the mesh vertices of the 3D face model using the estimated 3D motion onto some input frames. The face is rotating about the y axis and illumination is changing in the same way as pose.

Step 2. Assume illumination does not change, i.e., $\hat{\mathbf{l}}_{t+1} = \hat{\mathbf{l}}_t$. If the MSE between an input frame and the rendered frame,

$$\|\mathcal{I}_{t+1} - \left(\mathcal{B}_t + \mathcal{C}_t \times_2 \begin{pmatrix} \hat{\mathbf{T}}_{t+1} \\ \hat{\mathbf{\Omega}}_{t+1} \end{pmatrix}\right) \times_1 \hat{\mathbf{l}}_{t+1}\|^2,$$

is above a certain threshold (in the experimental section, we will discuss our strategy for choosing a appropriate threshold), repeat Step 1 for $\mathcal{I}_{t+1}$, till the MSE falls below an acceptable threshold.

Step 3. If the MSE is still larger than some threshold, reestimate the illumination using (25). Repeat Steps 1 and 2 with the new estimated $\hat{\mathbf{l}}_{t+1}$ for that input frame.

Step 4. Set $\mathrm{t} = \mathrm{t} + 1$. Repeat Steps 1, 2, and 3.

Step 5. Continue till $\mathrm{t} = \mathrm{N} - 1$.

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 (21) instead of (17) and (18) in the cost function (23) and estimate Δl_{ij}.

Figs. 8, 9, and 10 show the results of our tracking algorithm on a controlled data sequence. The images in Fig. 8 are synthesized from a 3D model and, thus, the motion and illumination are known. The face is rotating along y axis from -30° to $+30^\circ$ and the illumination is changing in the same way as pose. The resolution of the image is 240×320. Figs. 9 and 10 show plots of the estimated motion and illumination against the true values.

Finally, we show the results of the tracking on two real video sequences in Fig. 11, one of a face moving arbitrarily under varying illumination and the other of a person walking in a corridor where the light changes significantly. The image resolution is 240×320. Here, we map the texture of the person onto a generic 3D face model in order to perform the tracking. This is done by registering a few control points on the 3D model to the image of the face in the front view. We see from the results that the tracking is quite accurate and can handle different kinds of motion, including changes of scale.

6 Conclusions

In this paper, we have shown that the joint space of motion and illumination variables lies "close" to a bilinear subspace consisting of (approximately) nine illumination variables and six motion variables. The main novelty of our work is to formulate the combined effects of motion and illumination in the reflectance image. A detailed derivation of the bilinear space from fundamentals is presented.

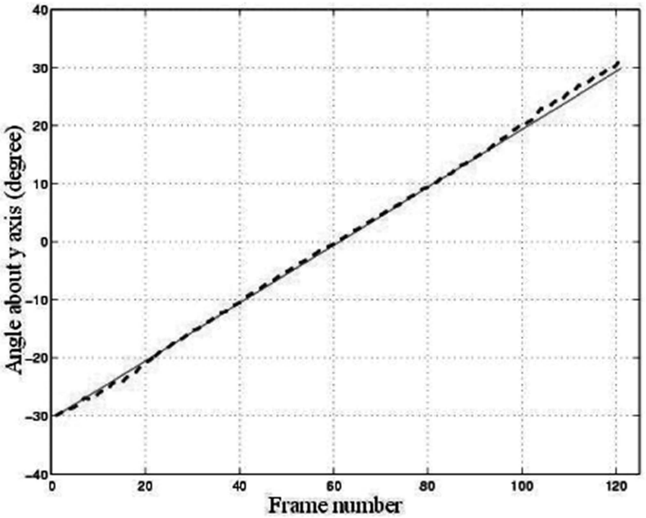


Fig. 9. The solid line shows the true pose (represented by the angle of face about y axis) and the broken line is the estimated pose.

Experimental analysis of the theory, synthesized results of face images under varying motion and illumination and application in 3D motion estimation under varying illumination are presented. Future work will involve 3D modeling from a monocular video sequence under varying illumination, object recognition across illumination, and pose variations, and accounting for specularities in the image (which are not modeled by Lambert's law). We also to plan to extend the theory for the analysis of deformable objects in video sequences.

Appendix A
Derivation of (13)

Equation (13) is the nonlinear solution of $\mathbf{\Delta}$ from (11) and (12) in Section 3. From (11), we have

$$\mathbf{P_2} = \mathbf{R}^{-1}(k\mathbf{u} + \mathbf{P_1} - \mathbf{T_0} - \mathbf{T}) + \mathbf{T_0}. \tag{28}$$

Substituting it into (12), we can solve for k as

$$k = -\frac{\mathbf{n}_{\mathbf{P}_1}^T((\mathbf{R}^{-1} - \mathbf{I})(\mathbf{P_1} - \mathbf{T_0}) - \mathbf{R}^{-1}\mathbf{T})}{\mathbf{n}_{\mathbf{P}_1}^T \mathbf{R}^{-1}\mathbf{u}}. \tag{29}$$

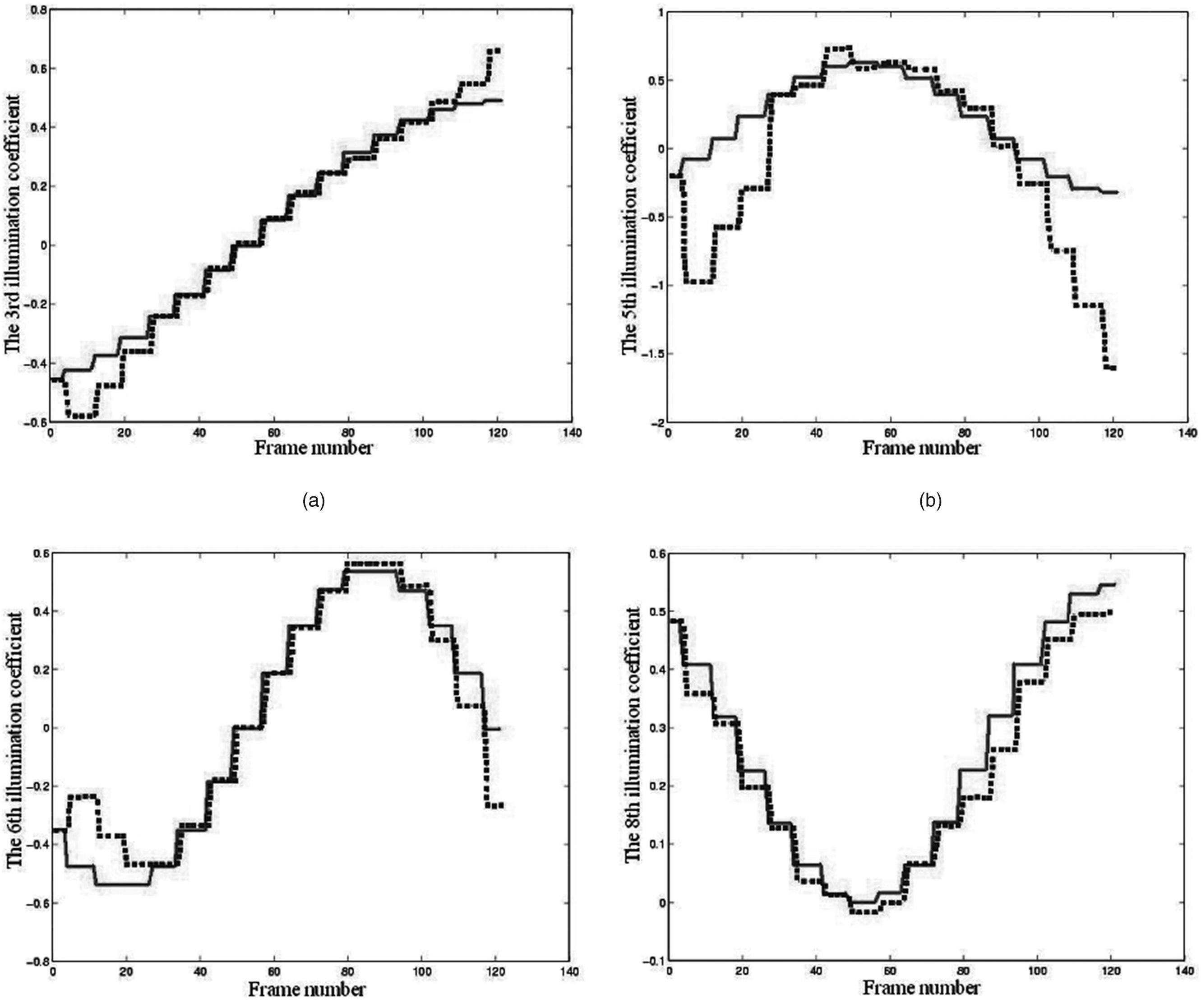
Substituting back into (11), $\mathbf{P_2}$ can be expressed as

$$\begin{aligned} \mathbf{P_2} = -\mathbf{R}^{-1} & \frac{\mathbf{n}_{\mathbf{P}_1}^T((\mathbf{R}^{-1} - \mathbf{I})(\mathbf{P_1} - \mathbf{T_0}) - \mathbf{R}^{-1}\mathbf{T})}{\mathbf{n}_{\mathbf{P}_1}^T \mathbf{R}^{-1}\mathbf{u}}\mathbf{u} \\ & + (\mathbf{R}^{-1} - \mathbf{I})(\mathbf{P_1} - \mathbf{T_0}) - \mathbf{R}^{-1}\mathbf{T} + \mathbf{P_1}. \end{aligned} \tag{30}$$

Thus, the coordinate difference between $\mathbf{P_2}$ and $\mathbf{P_1}$

$$\begin{aligned} \mathbf{\Delta} = & (\mathbf{R}^{-1} - \mathbf{I})(\mathbf{P_1} - \mathbf{T_0}) - \mathbf{R}^{-1}\mathbf{T} \\ & - \mathbf{R}^{-1}\frac{\mathbf{n}_{\mathbf{P}_1}^T(\mathbf{R}^{-1} - \mathbf{I})(\mathbf{P_1} - \mathbf{T_0})}{\mathbf{n}_{\mathbf{P}_1}^T \mathbf{R}^{-1}\mathbf{u}}\mathbf{u} \\ & + \mathbf{R}^{-1}\frac{\mathbf{n}_{\mathbf{P}_1}^T \mathbf{R}^{-1}\mathbf{T}}{\mathbf{n}_{\mathbf{P}_1}^T \mathbf{R}^{-1}\mathbf{u}}\mathbf{u}, \end{aligned} \tag{31}$$

from which (13) follows.



(a)

(b)

(c)

(d)

Fig. 10. (a), (b), (c), and (d) are the estimates of the third, fifth, sixth, and eighth illumination coefficients, respectively. The solid line shows the true illumination coefficients using the LRLS method and the dotted line shows the estimated illumination coefficients. More detailed analysis of the illumination estimation process will be provided in the future.

Appendix B

Derivation of (14)

When the motion is small, the inverse of the Rodrigues Rotation Matrix $\mathbf{R}^{-1}$ can be obtained from $-\mathbf{\Omega}$ as

$$\mathbf{R}^{-1} \cong \begin{pmatrix} 1 & \omega_z & -\omega_y \\ -\omega_z & 1 & \omega_x \\ \omega_y & -\omega_x & 1 \end{pmatrix}.$$

So, the first term in the RHS of (31) can be rewritten as

$$\begin{aligned}
&(\mathbf{R}^{-1} - \mathbf{I})(\mathbf{P_1} - \mathbf{T_0}) \\
\cong &\begin{pmatrix} 0 & -P_{w1z} + T_{0z} & P_{w1y} - T_{0y} \\ P_{w1z} - T_{0z} & 0 & -P_{w1x} + T_{0x} \\ -P_{w1y} + T_{0y} & P_{w1x} - T_{0x} & 0 \end{pmatrix} \begin{pmatrix} \omega_x \\ \omega_y \\ \omega_z \end{pmatrix} \\
\triangleq &\ \hat{\mathbf{P}}\mathbf{\Omega}.
\end{aligned} \tag{32}$$

For the third term in (31), we have

$$\begin{aligned}
&\mathbf{R}^{-1} \frac{\mathbf{n}^T_{\mathbf{P_1}}(\mathbf{R}^{-1} - \mathbf{I})(\mathbf{P_1} - \mathbf{T_0})}{\mathbf{n}^T_{\mathbf{P_1}}\mathbf{R}^{-1}\mathbf{u}} \mathbf{u} \cong \mathbf{R}^{-1} \frac{\mathbf{n}^T_{\mathbf{P}_{bf1}} \hat{\mathbf{P}}\mathbf{\Omega}}{\mathbf{n}^T_{\mathbf{P_1}}\mathbf{R}^{-1}\mathbf{u}} \mathbf{u} \\
= &\ \mathbf{R}^{-1} \frac{1}{\mathbf{n}^T_{\mathbf{P_1}}\mathbf{R}^{-1}\mathbf{u}} \left(\mathbf{n}^T_{\mathbf{P_1}} \hat{\mathbf{P}}\mathbf{\Omega}\right)\mathbf{u} = \mathbf{R}^{-1} \frac{1}{\mathbf{n}^T_{\mathbf{P_1}}\mathbf{R}^{-1}\mathbf{u}} \mathbf{u}\mathbf{n}^T_{\mathbf{P_1}} \hat{\mathbf{P}}\mathbf{\Omega} \\
= &\ \frac{1}{\mathbf{n}^T_{\mathbf{P_1}}\mathbf{R}^{-1}\mathbf{u}} \mathbf{R}^{-1}\mathbf{u}\mathbf{n}^T_{\mathbf{P_1}} \hat{\mathbf{P}}\mathbf{\Omega}.
\end{aligned} \tag{33}$$

Since $\mathbf{u}$ is a unit vector, each of its components is each less than or equal to 1. However, due to the small motion assumption, the elements of $\mathbf{\Omega}$ are far less than 1. Thus,

$$R^{-1}\mathbf{u} = \begin{pmatrix} 1 & \omega_z & -\omega_y \\ -\omega_z & 1 & \omega_x \\ \omega_y & -\omega_x & 1 \end{pmatrix} \begin{pmatrix} u_x \\ u_y \\ u_z \end{pmatrix} \cong \mathbf{u}. \tag{34}$$

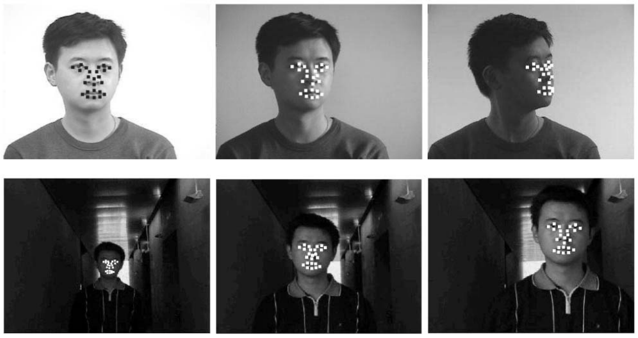


Fig. 11. The back projection of the mesh vertices of the 3D face model using the estimated 3D motion onto some input frames obtained under varying illumination. The top row is a video sequence of a face moving arbitrarily, while the second is of a person walking in a corridor.

Substituting back into (33), we have

$$\mathbf{R}^{-1}\frac{\mathbf{n}_{\mathbf{P}_1}^T(\mathbf{R}^{-1}-\mathbf{I})(\mathbf{P}_1-\mathbf{T}_0)}{\mathbf{n}_{\mathbf{P}_1}^T\mathbf{R}^{-1}\mathbf{u}}\mathbf{u}\cong\frac{1}{\mathbf{n}_{\mathbf{P}_1}^T\mathbf{u}}\mathbf{u}\mathbf{n}_{\mathbf{P}_1}^T\hat{\mathbf{P}}\mathbf{\Omega}. \tag{35}$$

Using similar reasoning for the fourth term on the RHS of (31), we have

$$\begin{aligned}\mathbf{R}^{-1}\frac{\mathbf{n}_{\mathbf{P}_1}^T\mathbf{R}^{-1}\mathbf{T}}{\mathbf{n}_{\mathbf{P}_1}^T\mathbf{R}^{-1}\mathbf{u}}\mathbf{u}&=\frac{\mathbf{n}_{\mathbf{P}_1}^T\mathbf{R}^{-1}\mathbf{T}}{\mathbf{n}_{\mathbf{P}_1}^T\mathbf{R}^{-1}\mathbf{u}\mathbf{R}^{-1}\mathbf{u}}\cong\frac{\mathbf{n}_{\mathbf{P}_1}^T\mathbf{R}^{-1}\mathbf{T}}{\mathbf{n}_{\mathbf{P}_1}^T\mathbf{u}}\mathbf{u}\\&=\frac{1}{\mathbf{n}_{\mathbf{P}_1}^T\mathbf{u}}(\mathbf{n}_{\mathbf{P}_1}^T\mathbf{R}^{-1}\mathbf{T})\mathbf{u}=\frac{1}{\mathbf{n}_{\mathbf{P}_1}^T\mathbf{u}}\mathbf{u}\mathbf{n}_{\mathbf{P}_1}^T\mathbf{R}^{-1}\mathbf{T}.\end{aligned} \tag{36}$$

Consider

$$R^{-1}\mathbf{T}=\begin{pmatrix}1&\omega_z&-\omega_y\\-\omega_z&1&\omega_x\\\omega_y&-\omega_x&1\end{pmatrix}\begin{pmatrix}T_x\\T_y\\T_z\end{pmatrix}\cong\mathbf{T}. \tag{37}$$

Substituting back into (31), we get

$$\mathbf{R}^{-1}\frac{\mathbf{n}_{\mathbf{P}_1}^T\mathbf{R}^{-1}\mathbf{T}}{\mathbf{n}_{\mathbf{P}_1}^T\mathbf{R}^{-1}\mathbf{u}}\mathbf{u}=\frac{1}{\mathbf{n}_{\mathbf{P}_1}^T\mathbf{u}}\mathbf{u}\mathbf{n}_{\mathbf{P}_1}^T\mathbf{T}. \tag{38}$$

Substituting (27), (30), (32), and (33) back into (31), we have

$$\begin{aligned}\mathbf{\Delta}&\cong\hat{\mathbf{P}}\mathbf{\Omega}-\frac{1}{\mathbf{n}_{\mathbf{P}_1}^T\mathbf{u}}\mathbf{u}\mathbf{n}_{\mathbf{P}_1}^T\hat{\mathbf{P}}\mathbf{\Omega}-\mathbf{T}-\frac{1}{\mathbf{n}_{\mathbf{P}_1}^T\mathbf{u}}\mathbf{u}\mathbf{n}_{\mathbf{P}_1}^T\mathbf{T}\\&=\left(\mathbf{I}-\frac{1}{\mathbf{n}_{\mathbf{P}_1}^T\mathbf{u}}\mathbf{u}\mathbf{n}_{\mathbf{P}_1}^T\right)(\hat{\mathbf{P}}\mathbf{\Omega}-\mathbf{T}).\end{aligned} \tag{39}$$

ACKNOWLEDGMENTS

The authors would like to acknowledge support from the Bourns College of Engineering (University of California, Riverside) Initial Complement Fund.

REFERENCES

[1] A. Azarbayejani and A.P. Pentland, "Recursive Estimation of Motion, Structure, and Focal Length," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 17, no. 6, pp. 562-575, June 1995.

[2] R. Basri and D.W. Jacobs, "Lambertian Reflectances and Linear Subspaces," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 25, no. 2, pp. 218-233, Feb. 2003.

[3] P. Belhumeur and D. Kriegman, "What Is the Set of Images of an Object under All Possible Lighting Conditions?" *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 1996.

[4] V. Blanz and T. Vetter, "Face Recognition Based on Fitting a 3D Morphable Model," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 25, no. 9, pp. 1063-1074, Sept. 2003.

[5] T.J. Broida, S. Chandrashekhar, and R. Chellappa, "Recursive 3-D Motion Estimation from a Monocular Image Sequence," *IEEE Trans. Aerospace and Electronic Systems,* vol. 26, no. 4, pp. 639-656, July 1990.

[6] A. Chiuso, P. Favaro, H. Jin, and S. Soatto, "Structure from Motion Causally Integrated over Time," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 24, no. 4, pp. 523-535, Apr. 2002.

[7] K. Daniilidis and H. Nagel, "Analytic Results on Error Sensitivity of Motion Estimation from Two Views," *Image and Vision Computing,* vol. 8, no. 4, pp. 297-303, 1990.

[8] O. Faugeras, *Three-Dimensional Computer Vision.* MIT Press, 2002.

[9] C. Fermuller and Y. Aloimonos, *Foundations of Image Understanding.* Kluwer, 2001.

[10] R.T. Frankot and R. Challappa, "A Method for Enforcing Integrability in Shape from Shading Algorithms," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 10, no. 4, pp. 439-451, July 1988.

[11] D. Freedman and M. Turek, "Illumination-Invariant Tracking via Graph Cuts," *Proc. Conf. Computer Vision and Pattern Recognition,* 2005.

[12] R. Gross, I. Matthews, and S. Baker, "Eigen Light-Fields and Face Recognition across Pose," *Proc. IEEE Int'l Conf. Automatic Face and Gesture Recognition,* May 2002.

[13] R. Gross, I. Matthews, and S. Baker, "Fisher Light-Fields for Face Recognition across Pose and Illumination," *Proc. German Symp. Pattern Recognition,* Sept. 2002.

[14] G.D. Hager and P.N. Belhumeur, "Efficient Region Tracking with Parametric Models of Geometry and Illumination," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 20, no. 10, pp. 1025-1039, Oct. 1998.

[15] R.I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision.* Cambridge Univ. Press, 2000.

[16] X. He, S. Yan, Y. Hu, P. Niyogi, and H.J. Zhang, "Face Recognition Using Laplacianfaces," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 27, no. 5, pp. 328-340, May 2005.

[17] B.K.P. Horn and M.J. Brooks, "The Variational Approach to Shape from Shading," *Computer Vision Graphics and Image Processing,* vol. 33, no. 2, pp. 174-208, 1986.

[18] B.K.P. Horn and B.G. Schunck, "Determining Optical Flow," *Artificial Intelligence,* vol. 17, pp. 185-203, 1981.

[19] D. Jacobs and S. Shirdhonkar, "Non-Negative Lighting and Specular Object Recognition," *Proc. Int'l Conf. Computer Vision,* 2005.

[20] H. Jin, P. Favaro, and S. Soatto, "Real-Time Feature Tracking and Outlier Rejection with Changes in Illumination," *Proc. IEEE Int'l Conf. Computer Vision,* 2001.

[21] H. Jin, S. Soatto, and A.J. Yezzi, "Multi-View Stereo Reconstruction of Dense Shape and Complex Appearance," *Int'l J. Computer Vision,* vol. 63, no. 3, pp. 175-189, 2005.

[22] L.D. Lathauwer, B.D. Moor, and J. Vandewalle, "A Multilinear Singular Value Decomposition," *SIAM J. Matrix Analysis and Applications,* vol. 21, no. 4, pp. 1253-1278, 2000.

[23] K. Lee, J. Ho, and D.J. Kriegman, "Acquiring Linear Subspaces for Face Recognition under Variable Lighting," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 27, no. 5, pp. 684-698, May 2005.

[24] K. Lee, J. Ho, M. Yang, and D. Kriegman, "Video-Based Face Recognition Using Probabilistic Appearance Manifolds," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2003.

[25] V. Lepetit and P. Fua, *Monocular Model-Based 3D Tracking of Rigid Objects.* Now Publishers Inc., 2005.

[26] Y. Moses, "Face Recognition: Generalization to Nobel Images," PhD thesis, Weizmann Inst. of Sciences, 1993.

[27] H. Murase and S.K. Nayar, "Visual Learning and Recognition of 3D Objects from Appearance," *Int'l J. Computer Vision,* vol. 14, no. 1, pp. 5-24, 1995.

[28] J. Oliensis, "A Critique of Structure from Motion Algorithms," *Computer Vision and Image Understanding,* vol. 80, no. 2, pp. 172-214, 2000.

[29] J. Oliensis and P. Dupuis, "Direct Method for Reconstructing Shape from Shading," *Proc. SPIE Conf. 1570 Geometric Methods in Computer Vision,* 1991.

[30] A. Pentland, "Photometric Motion," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 13, no. 9, pp. 879-890, Sept. 1991.

[31] G. Qian, R. Chellappa, and Q. Zheng, "Robust Structure from Motion Estimation Using Inertial Data," *J. Optical Soc. Am. A,* vol. 18, no. 12, pp. 2982-2997, 2001.

[32] R. Ramamoorthi, R. Ng, and P. Hanrahan, "Wavelet Triple Product Integrals for All-Frequency Relighting," *Proc. ACM SIGGRAPH,* pp. 475-485, 2004.

[33] R. Ramamoorthi and P. Hanrahan, "On the Relationship between Radiance and Irradiance: Determining the Illumination from Images of a Convex Lambertian Object," *J. Optical Soc. Am. A,* vol. 18, no. 10, Oct. 2001.

[34] R. Ramamoorthi and P. Hanrahan, "A Signal Processing Framework for Reflection," *ACM Trans. Graphics,* pp. 1004-1042, Oct. 2004.

[35] A.K. Roy-Chowdhury and R. Chellappa, "Stochastic Approximation and Rate Distortion Analysis for Robust Structure and Motion Estimation," *Int'l J. Computer Vision,* vol. 55, no. 1, pp. 27-53, 2003.

[36] A.K. Roy-Chowdhury and R. Chellappa, "An Information Theoretic Criterion for Evaluating the Quality of 3D Reconstructions from Video," *IEEE Trans. Image Processing,* pp. 960-973, July 2004.

[37] A.K. Roy-Chowdhury and R. Chellappa, "Statistical Bias in 3D Reconstruction from a Monocular Video," *IEEE Trans. Image Processing,* pp. 1057-1062, Aug. 2005.

[38] I. Matthews, C. Hu, J. Xiao, J. Cohn, S. Koterba, S. Baker, and T. Kanade, "Multi-View AAM Fitting and Camera Calibration," *Proc. Int'l Conf. Computer Vision,* Oct. 2005.

[39] A. Shashua, "On Photometric Issues in 3D Visual Recognition from a Single 2D Image," *Int'l J. Computer Vision,* vol. 21, nos. 1-2, pp. 99-122, 1997.

[40] J. Shi and C. Tomasi, "Good Features to Track," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 1994.

[41] D. Simakov, D. Frolova, and R. Basri, "Dense Shape Reconstruction of a Moving Object under Arbitrary, Unknown Lighting," *Proc. IEEE Int'l Conf. Computer Vision,* 2003.

[42] R. Szeliski and S. Kang, "Recovering 3D Shape and Motion from Image Streams Using Non-Linear Least Squares," *J. Visual Computation and Image Representation,* vol. 5, pp. 10-28, 1994.

[43] K. Thornber and D. Jacobs, "Broadened, Specular Reflection and Linear Subspaces," Technical Report 2001-033, NEC, 2001.

[44] C. Tomasi and T. Kanade, "Shape and Motion from Image Streams under Orthography: A Factorization Method," *IEEE Int'l J. Computer Vision,* vol. 9, no. 2, pp. 137-154, 1992.

[45] L. Torresani and C. Bregler, "Space-Time Tracking," *Proc. IEEE European Conf. Computer Vision,* 2002.

[46] M.A.O. Vasilescu and D. Terzopoulos, "Multilinear Independent Components Analysis," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* June 2005.

[47] Y. Xu and A.K. Roy-Chowdhury, "Integrating the Effects of Motion, Illumination and Structure in Video Sequences," *Proc. Int'l Conf. Computer Vision,* Oct. 2005.

[48] G. Young and R. Chellappa, "Statistical Analysis of Inherent Ambiguities in Recovering 3D Motion from a Noisy Flow Field," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 14, pp. 995-1013, 1992.

[49] 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," *Proc. Ninth IEEE Int'l Conf. Computer Vision,* Mar. 2003.

[50] L. Zhang and D. Samaras, "Face Recognition from a Single Training Image under Artibrary Unknowon Lighting Using Spherical Harmonics," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 28, no. 3, pp. 351-364, Mar. 2006.

[51] Z. Zhang, "Determining the Epipolar Geometry and Its Uncertainty: A Review," *Int'l J. Computer Vision,* vol. 27, pp. 161-195, 1998.

[52] S. Zhou and R. Chellappa, "Image-Based Face Recognition under Illumination and Pose Variations," *J. Optical Soc. Am., A,* vol. 22, pp. 217-229, Feb. 2005.

Yilei Xu received the BS degree in electrical engineering from Peking University, Beijing, China, in 2004 and the MS degree in electrical engineering from the University of California at Riverside in 2006, where he is now pursuing the PhD degree in the Electrical Engineering Department. His main research interests include computer vision, video processing and analysis, pattern recognition, and machine learning. He is currently working on illumination modeling and motion estimation from video sequences. He is a student member of the IEEE.

Amit K. Roy-Chowdhury received the ME degree in systems science and automation from the Indian Institute of Science, Bangalore, India. He received the PhD degree in 2002 from the University of Maryland, College Park, where he also worked as a research associate in 2003. He has been an assistant professor of electrical engineering at the University of California, Riverside since January 2004. His research interests are in the broad areas of image processing and analysis, computer vision, video communications, and machine learning. Currently, he is working on problems of pose and illumination invariant video-based object recognition, event analysis in large video networks, and multiterminal video compression. He has published more than 50 papers in peer-reviewed journals, conferences, and edited books. He is the author of the book *Recognition of Humans and Their Activities Using Video*. He is on the program committees of many major conferences in computer vision and image/signal processing and is a regular reviewer for the main journals in these areas. He is a member of the IEEE.