Integrating the Effects of Motion, Illumination and Structure in Video Sequences

Yilei Xu, Amit K. Roy-Chowdhury *

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

Abstract

Most work in computer vision has concentrated on studying the individual effects of motion and illumination on a 3D object. 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, 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 is based on the intuitive notion that, given an illumination direction, the images of a moving surface cannot change suddenly over a short time period. We experimentally compare the images obtained using our theory with ground truth data and show that the difference is small and acceptable. We also provide experimental results on real data by synthesizing video sequences of a 3D face with various combinations of motion and illumination directions.

1. Introduction

Motion and illumination are two fundamental features used for analyzing video sequences. Traditionally, they have mostly been studied separately. Optical flow [6] methods deal with motion, but are based on the assumption that the intensity of the same point does not change over time. Structure from Motion (SfM) [3, 5] is used for 3D modelling by exploiting the motion information in a video sequence, but does not consider illumination changes. Other techniques like Shape from Shading (SfS) [4, 7, 11] rely on the illumination information in a *single* image to estimate the 3D structure of a scene. Photometric stereo [12] methods consider varying illumination but require the object to be fixed relative to the camera. Negahdaripour [10] proposed a method for optical flow computation by combining geometric and radiometric cues. Recently, Zhang et al [14] have combined motion and varying illumination by unifying Structure from Motion, Photometric Stereo, and multi-view stereo in an optimization framework. Their work is essentially based on optical flow, but does not provide an explicit expression relating the image, and the motion, structure and illumination variables. In [8], the authors proposed a method for estimating the shape and radiance of the scene, but under constant illumination.

A number of researchers have shown that the set of images that an object can produce under varying illumination lies in a low-dimensional linear subspace. Shashua [12] and Moses [9] derived a 3D linear representation of the set of images ignoring attached shadows. Belhumeur and Kriegman [2] have shown that the set of images of an object under arbitrary illumination forms a convex cone in the space of all possible images. Basri and Jacobs [1] analytically derived a 9D spherical harmonics based linear representation of the images produced by a Lambertian object with attached shadows. This 9D space is an approximation of the infinite dimensional space derived in [2]. However, these methods focus primarily on the problem of object recognition, and are restricted to the analysis of single images. Extending the work in [1] directly to video sequences would require repeating the processes described to 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 a joint illumination and motion space of video sequences.

In this paper, we derive, from first principles, 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 nine illumination variables and six motion variables. Our work extends the results in [1] to video sequences. We consider the case of continuous motion, and represent variations in surface norms and albedo upto 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

*The work was supported from the Initial Complement funds provided by the Bourns College of Engineering, University of California, Riverside.

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

with various combinations of motion and illumination directions.

The rest of the paper is organized as follows. Section 2 presents previous work on the Lambertian Reflectance Linear Subspace (LRLS) method for modelling 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 concludes the paper and highlights future work.

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,

$$I = A\rho \max(cos\theta, 0),$$

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 [1] 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% 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

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

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

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

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 [1] 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. However, it does not consider the motion of the object relative to the camera. For brevity, we will refer to the work in [1] as the Lambertian Reflectance Linear Subspace (LRLS) theory.

The LRLS theory, as described in [1], 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 it's 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 [1] 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 object so as to combine the effects of motion, illumination, and 3D structure into a single framework.

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 Basri and Jacobs' work [1] 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, and we will also assume that the focal length, f of the camera is the only intrinsic parameters[2]. Hence, 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 consider that it passes through the center of the image.

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 $l_{ij}^{t_1}$. 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:

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

Let us define the the motion of the object in the above reference frame as the translation $\mathbf{T} = (T_x, T_y, T_z)^T$ of the centroid of the object and the rotation $\mathbf{\Omega} = (\omega_x, \omega_y, \omega_z)^T$ 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 $l_{ij}^{t_2}$.

3.2 Computation of the new basis image

Let $\mathbf{A}$ and $\mathbf{B}$ 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

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

where $\mathbf{J_P}$ is the Jacobian matrix of the norm, $\mathbf{n_P}$, with respect to a point, $\mathbf{P} \in \mathbb{R}^3$ on the surface of the object, and $\mathbf{\Delta}$ is the difference in the coordinates of $\mathbf{P_2}$ and $\mathbf{P_1}$ (Henceforth we will refer to $\mathbf{\Delta}$ 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 affects 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 can be derived from the motion of the object.

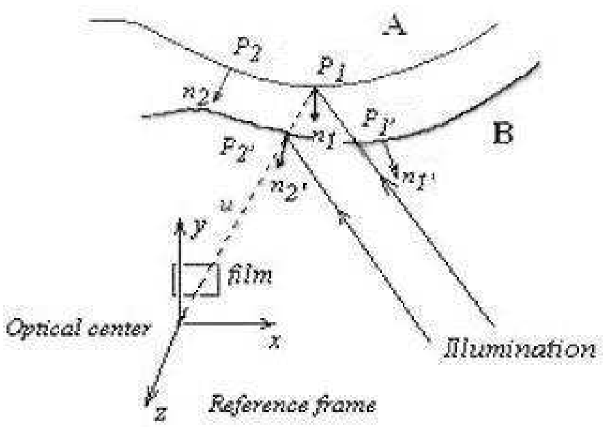


Figure 1: Pictorial representation showing the motion of the object and its projection.

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 $\mathbf{n_{P_1}}$ and $\mathbf{n_{P_2'}}$ is a function of the spatial (from $\mathbf{n_{P_1}}$ to $\mathbf{n_{P_2}}$) and temporal (from $\mathbf{n_{P_2}}$ to $\mathbf{n_{P_2'}}$) coordinates. Using the coordinate change $\mathbf{\Delta}$ and the Jacobian matrix of norm at $\mathbf{P_1}$, we are able to calculate the first order difference between $\mathbf{n_{P_1}}$ and $\mathbf{n_{P_2}}$. Using the motion information, we can obtain the difference between $\mathbf{n_{P_2}}$ and $\mathbf{n'_{P_2}}$. 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 upto a first order approximation as

$$\Delta\mathbf{n} = \mathbf{n_{P_2'}} - \mathbf{n_{P_1}} = \mathbf{J_{P_1}}\mathbf{\Delta} + \frac{\partial \mathbf{n_{P_2}}}{\partial t}\Delta t, \qquad (4)$$

and

$$\Delta\rho = \rho_{\mathbf{P_2'}} - \rho_{\mathbf{P_1}} = \nabla\rho_{\mathbf{P_1}}\mathbf{\Delta}, \qquad (5)$$

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

$$\begin{aligned} b_{ij}(\mathbf{n_{P_2'}}) &= (\rho_{\mathbf{P_1}} + \Delta\rho) r_i Y_{ij}(\mathbf{n_{P_1}} + \Delta\mathbf{n}) \\ &= b_{ij}(\mathbf{n_{P_1}}) + \nabla\rho_{\mathbf{P_1}} r_i Y_{ij}(\mathbf{n_{P_1}})\mathbf{\Delta} \\ &\quad + \rho_{\mathbf{P_1}} r_i \nabla Y_{ij}(\mathbf{n_{P_1}})\Delta\mathbf{n} + o(\mathbf{\Delta}). \qquad (6) \end{aligned}$$

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

[2] This assumption can be relaxed.

image is a linear function of $\mathbf{\Delta}$. ($\frac{\partial \mathbf{n_{P_2}}}{\partial t}$ is not a function of $\mathbf{\Delta}$, as we will show latter.)

3.3 Computation of coordinate change $\mathbf{\Delta}$

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.,

$$\mathbf{P_2}' - \mathbf{P_1} = k\mathbf{u}, \tag{7}$$

where

$$\mathbf{u} = \frac{1}{\sqrt{x^2+y^2+f^2}} \begin{pmatrix} x \\ y \\ f \end{pmatrix}, \tag{8}$$

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

$$\mathbf{P_2}' = \mathbf{R}(\mathbf{P_2} - \mathbf{T_0}) + \mathbf{T_0} + \mathbf{T}, \tag{9}$$

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

$$k\mathbf{u} = \mathbf{R}(\mathbf{P_2} - \mathbf{T_0}) + \mathbf{T_0} + \mathbf{T} - \mathbf{P_1}. \tag{10}$$

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.,

$$\mathbf{n}_{\mathbf{P_1}}^T(\mathbf{P_1} - \mathbf{P_2}) = 0. \tag{11}$$

Using (10) and (11) and after some algebraic manipulation (please refer to Appendix), we can show that

$$\begin{aligned} \mathbf{\Delta} \;=\; & \left(\mathbf{R}^{-1} - \mathbf{I}\right)(\mathbf{P_1} - \mathbf{T_0}) - \mathbf{R}^{-1}\mathbf{T} \\ & -\mathbf{R}^{-1} \frac{\mathbf{n}_{\mathbf{P_1}}^T \left(\left(\mathbf{R}^{-1} - \mathbf{I}\right)(\mathbf{P_1} - \mathbf{T_0}) - \mathbf{R}^{-1}\mathbf{T}\right)}{\mathbf{n}_{\mathbf{P_1}}^T \mathbf{R}^{-1}\mathbf{u}} \mathbf{u}. \end{aligned} \tag{12}$$

The coordinate change, $\mathbf{\Delta}$, obtained in (12) 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. This will allow us to interpret the joint effect of the motion and illumination analytically, while sacrificing little in terms of accuracy. Using a series of mathematical approximations, we can obtain $\mathbf{\Delta}$ as a linear function of the motion variables (please refer to Appendix) as:

$$\begin{aligned} \mathbf{\Delta} \;\cong\; & \hat{\mathbf{P}}\mathbf{\Omega} + \mathbf{T} - \frac{1}{\mathbf{u}^T\mathbf{n}_{\mathbf{P_1}}}\mathbf{u}\mathbf{n}_{\mathbf{P_1}}^T\hat{\mathbf{P}}\mathbf{\Omega} - \frac{1}{\mathbf{u}^T\mathbf{n}_{\mathbf{P_1}}}\mathbf{u}\mathbf{n}_{\mathbf{P_1}}^T\mathbf{T} \\ =\; & \left(\mathbf{I} - \frac{1}{\mathbf{n}_{\mathbf{P_1}}^T\mathbf{u}}\mathbf{u}\mathbf{n}_{\mathbf{P_1}}^T\right)\left(\hat{\mathbf{P}}\mathbf{\Omega} - \mathbf{T}\right) \\ \triangleq\; & \mathbf{C}(\hat{\mathbf{P}}\mathbf{\Omega} - \mathbf{T}), \end{aligned} \tag{13}$$

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 $\mathbf{\Delta}_l$ and $\mathbf{\Delta}_{nl}$.

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:

$$\begin{aligned} \frac{\partial \mathbf{n_{P_2}}}{\partial t}\Delta t \;=\; & \frac{\partial(\mathbf{n}_{\mathbf{P_1}} + \mathbf{J}_{\mathbf{P_1}}\mathbf{\Delta})}{\partial t} = \mathbf{\Omega} \times (\mathbf{n}_{\mathbf{P_1}} + \mathbf{J}_{\mathbf{P_1}}\mathbf{\Delta}) \\ =\; & \mathbf{\Omega} \times \mathbf{n}_{\mathbf{P_1}} + o(\mathbf{\Omega}\mathbf{T}) \cong (-\mathbf{n}_{\mathbf{P_1}})^{\wedge}\mathbf{\Omega} \\ \triangleq\; & -\hat{\mathbf{N}}\mathbf{\Omega}. \end{aligned} \tag{14}$$

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 $\frac{\partial \mathbf{n_{P_2}}}{\partial t}$ is not a function of $\mathbf{\Delta}$, as claimed at the end of section 3.2.

3.5 Bilinear space of motion and illumination

Substituting (13) and (14) into (4), we get a linear expression for $\Delta\mathbf{n}$ as a function of motion variables, i.e.,

$$\Delta\mathbf{n} = \left(\mathbf{J}_{\mathbf{P_1}}\mathbf{C}\hat{\mathbf{P}} - \hat{\mathbf{N}}\right)\mathbf{\Omega} - \mathbf{J}_{\mathbf{P_1}}\mathbf{C}\mathbf{T}. \tag{15}$$

So far, we have expressed the coordinate and norm change as linear expressions of the motion variables. Substituting (13) and (15) into (1) and (6), which contain the illumination variables, we have

$$I(x,y,t_2) = \sum_{i=0,1,2}\ \sum_{j=-i,-i+1\ldots i-1,i} l_{ij}^{t_2} b_{ij}(\mathbf{n}_{\mathbf{P}_2'}), \tag{16}$$

where

$$b_{ij}(\mathbf{n}_{\mathbf{P}_2'}) = b_{ij}(\mathbf{n}_{\mathbf{P_1}}) + \mathbf{A}\mathbf{T} + \mathbf{B}\mathbf{\Omega}, \tag{17}$$

$$\mathbf{A} = -r_i\left(\nabla\rho_{\mathbf{P_1}}Y_{ij}(\mathbf{n}_{\mathbf{P_1}}) + \rho_{\mathbf{P_1}}\nabla Y_{ij}(\mathbf{n}_{\mathbf{P_1}})\mathbf{J}_{\mathbf{P_1}}\right)\mathbf{C},$$

and

$$\mathbf{B} = -\mathbf{A}\hat{\mathbf{P}} - r_i\rho_{\mathbf{P_1}}\nabla Y_{ij}(\mathbf{n}_{\mathbf{P_1}})\hat{\mathbf{N}}.$$

[3] We define the skew symmetric matrix of a vector $\mathbf{X} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$ as $\mathbf{X}^{\wedge} = \hat{\mathbf{X}} = \begin{pmatrix} 0 & -x_3 & x_2 \\ x_3 & 0 & -x_1 \\ -x_2 & x_1 & 0 \end{pmatrix}$.

In (17), $b_{ij}(\mathbf{n}_{\mathbf{P}'_2})$ are the basis images after motion. The first term, $b_{ij}(\mathbf{n}_{\mathbf{P}_1})$, are the original basis images before motion. It is only determined by the object model and does 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 $\mathbf{AT}+\mathbf{B\Omega}$, where the first term describes the effect of translation, and the second term describes the effect of rotation. Substituting (17) into (16), we see that the new image spans a bilinear space of the motion variables and illumination variables.

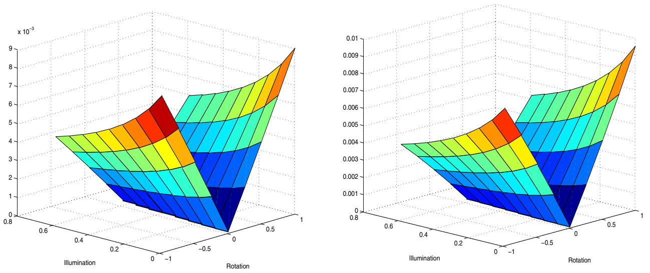
When the illumination changes gradually, the result can be simplified by using the Taylor series to approximate the illumination coefficients as $l_{ij}^{t_2} = l_{ij}^{t_1} + \Delta l_{ij}$. The bilinear space now becomes a combination of two linear subspaces, defined by the motion and illumination variables as

$$\begin{aligned} I(x,y,t_2) \quad = \quad & I(x,y,t_1) + \sum_{i=0,1,2} \sum_{j=-i,\ldots,i} l_{ij}^{t_1}(\mathbf{AT}+\mathbf{B\Omega}) \\ & + \sum_{i=0,1,2} \sum_{j=-i,\ldots,i} \Delta l_{ij} b_{ij}(\mathbf{n}_{P_1}). \qquad (18) \end{aligned}$$

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 the linear space of the motion variables, since the third term in (18) is zero.

3.6 Discussion

This bilinear space result integrates the effects of illumination and motion in generating an image from a 3D object. Moreover, 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, Equations (16) and (17) integrate motion, illumination, 3D structure, albedo and camera intrinsic parameters into one single framework. 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 Basri and Jacobs' work [1], 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 bilinear. Thus the joint illumination and motion space for a sequence of images is bilinear with (approximately) nine illumination variables and six motion variables. In [13], the authors assumed that face images lie in a multilinear space of illumination, view point, identity and expression variables, and then used Multilinear Independent Components Analysis (MICA) to learn and recognize faces. Our result provides a theoretical underpinning for this assumed model considering only pose and illumination.



(a) (b)

Figure 2: (a) shows the normalized error for one particular pixel between true intensity and the one with linear approximation of $\mathbf{\Delta}$ in (13). (b) shows the normalized error for the same pixel between the true intensity and the one with nonlinear approximation of $\mathbf{\Delta}$ in (12)

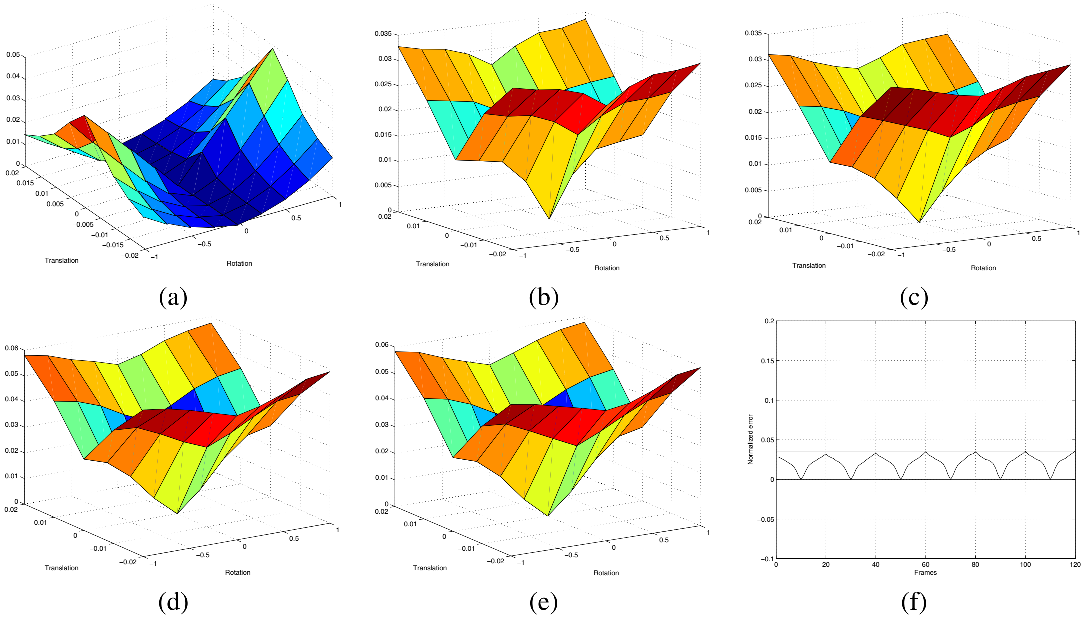
Fig. 2 shows the effect of approximating the nonlinear function $\mathbf{\Delta}$ in (12) with a linear approximation. We plot the difference in the image intensity at a particular point as a function of $l_{11}^{t_1}$ and ω_y. The rotation range is defined as in Section 4, and the illumination changes in a typical range. We take the intensity obtained from the LRLS method as the true value. The difference in Fig. 2(a) is computed between the true value and the intensity obtained with the linear expression of $\mathbf{\Delta}$ (using (13)) and normalized w.r.t. the true one. The difference in Fig. 2(b) is computed between the true value and the intensity obtained with the nonlinear expression of $\mathbf{\Delta}$ (using (12)) and also normalized w.r.t. the true one. As can be seen, there is no perceptible difference between the bilinear and nonlinear image spaces.

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 respect to the 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 (11) by making the tangent plane approximation. Next, the small motion assumption is used to obtain the linear approximation of (13). 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 y-axis on the change of albedo, norm and $\mathbf{\Delta}$. We also compare the synthesized image 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 initial 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



(a) (b) (c)

(d) (e) (f)

Figure 3: (a) Normalized difference of the linear and nonlinear coordinate change, $\frac{|\mathbf{\Delta}_l - \mathbf{\Delta}_{nl}|}{|\mathbf{\Delta}_{nl}|}$. (b) Normalized difference of the norm, $\frac{|\mathbf{n}_{\mathbf{P}_2'} - \bar{\mathbf{n}}_{\mathbf{P}_2'}|}{|\bar{\mathbf{n}}_{\mathbf{P}_2'}|}$, 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, $\frac{|\rho_{\mathbf{P_2}} - \bar{\rho}_{\mathbf{P_2}}|}{|\bar{\rho}_{\mathbf{P}_2'}|}$, 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, $\frac{|I(.,.,t_2)^l - I(.,.)^{LRLS}|}{|I(.,.)^{LRLS}|}$, where $I(.,.,t_2)^l$ is the image generated by linear coordinate change $\mathbf{\Delta}_l$, and $I(.,.)^{LRLS}$ is the image obtained with the LRLS theory. (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 $\mathbf{\Delta}_{nl}$. (f) Normalized error as in (d), plotted as a function of time for a video sequence.

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 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 of the extremities of the face (where our theory is not valid), there maybe a lot of points with large error. To avoid the bias caused by these points, we represent the total error using the median of the errors of all the points.

In Fig. 3, (a) depicts the difference between $\mathbf{\Delta}_{nl}$ and $\mathbf{\Delta}_l$ normalized w.r.t. $\mathbf{\Delta}_{nl}$. Within the typical motion range defined above, the largest relative error of the linear solution (w.r.t. the nonlinear solution) is about five percent. Next, in Fig. 3(b) and 3(c), 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 [1] (see Section 2) and normalized w.r.t. the LRLS ones obtained for each image separately. Fig. 3(d) gives the normalized error of the image obtained with the bilinear approximation of (16), and (17). Typically, the motion between the consecutive frames is much smaller then the extremities of the above range; hence the difference in practice is about $2 \sim 3\%$. Moreover, if we consider only rotation, the error at the extremities of the above range is $1 \sim 2\%$. (See Fig. 2(a) and 2(b)). Fig. 3(e) computes the normalized error of the images obtained with the nonlinear expression of coordinate change in (12). The normalized error of the images obtained with linear coordinate change (13) in Fig. 3(d) and nonlinear coordinate change (12) in

Fig. 3(e) are very similar, which validate the approximations in the linearization part of the derivation.

Finally, we synthesized a video sequence of rotating face with our theory and LRLS theory respectively. Fig. 3(f) gives the normalized error of the video sequence synthesized with our theory. The maximum error is about 5%, 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 space of (16) and (17), the computation of which is very fast. In our experiment, generating each frame using LRLS theory will take $\sim 15 - 20$ seconds, while generating 20 frames with the bilinear space take only $\sim 20 - 25$ seconds.

Next, we applied our theory to a 3D face to synthesize image sequences for different combinations of motion and illumination directions using bilinear space theory. In Fig. 4, the pose of the 3D face is fixed and only illumination changes. From (16) and (17), $\mathbf{T}$ and $\mathbf{\Omega}$ are zero, basis images b_{ij} remain the same, and only l_{ij} change. Thus all 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 pose changes, thus l_{ij} are fixed and b_{ij} is a linear function of $\mathbf{T}$ and $\mathbf{\Omega}$; 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 image synthesized by the two methods.

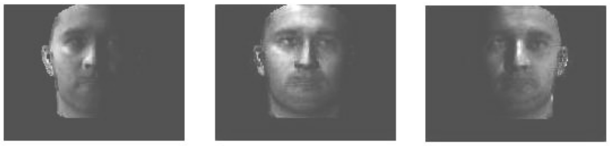


Figure 4: Reflectance images under fixed pose and rotating illumination. All these images lie in a linear subspace of illumination variables.

In Fig. 6, 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 $\mathbf{\Omega}$, and $I(x, y, t_2)$ is the combination of b_{ij} with varying coefficients l_{ij}. The generated images lie in a bilinear space of the illumination and motion variables.

5 Future Work and 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 effect of motion and illumination in the reflectance image. A detailed derivation of the bilinear space from first principles is presented. Experimental analysis of the theory and synthesized results of face images under varying motion and illumination are presented. Future work will involve joint estimation of 3D motion, illumination and structure from a video sequence, 3D model based tracking and object recognition across illumination and pose variations. We also to plan to extend the theory for the analysis of deformable objects in video sequences.

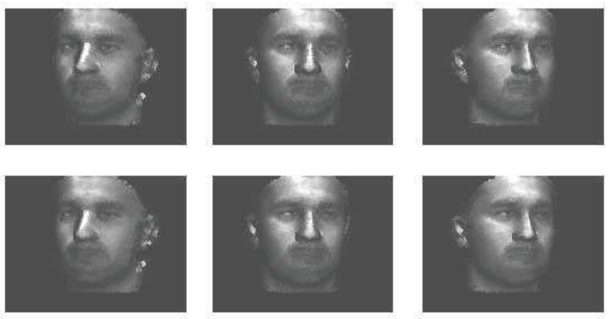


Figure 5: Reflectance images of the face rotating along the vertical axis under fixed illumination. The images in the upper row are generated by our theory, and the images in the lower row are generated by the LRLS theory repeated for each pose.

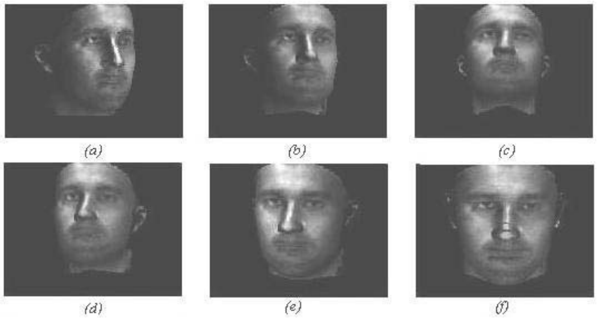


Figure 6: Reflectance images of a moving face with changing illumination directions. Illumination changes in the same way as pose, and always comes from the front of the face. The images are generated using the bilinear space of motion and illumination variables.

Appendix A Derivation of (12)

Equation (12) is the nonlinear solution of $\mathbf{\Delta}$ from the Equations (10) and (11) in Section 3. Substituting the expression

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

of $\mathbf{P_2}$ from (10) into (11), we can solve for k as

$$k = -\frac{\mathbf{n}_{\mathbf{P_1}}^T((\mathbf{R}^{-1}-\mathbf{I})(\mathbf{P_1}-\mathbf{T_0})-\mathbf{R}^{-1}\mathbf{T})}{\mathbf{n}_{\mathbf{P_1}}^T\mathbf{R}^{-1}\mathbf{u}}. \quad (19)$$

Substituting back into (10), $\mathbf{P_2}$ can be expressed as

$$\begin{aligned}\mathbf{P_2} &= -\mathbf{R}^{-1}\frac{\mathbf{n}_{\mathbf{P_1}}^T((\mathbf{R}^{-1}-\mathbf{I})(\mathbf{P_1}-\mathbf{T_0})-\mathbf{R}^{-1}\mathbf{T})}{\mathbf{n}_{\mathbf{P_1}}^T\mathbf{R}^{-1}\mathbf{u}}\mathbf{u} \\ &\quad +(\mathbf{R}^{-1}-\mathbf{I})(\mathbf{P_1}-\mathbf{T_0})-\mathbf{R}^{-1}\mathbf{T}+\mathbf{P_1}. \quad (20)\end{aligned}$$

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

$$\begin{aligned}\mathbf{\Delta} &= \left(\mathbf{R}^{-1}-\mathbf{I}\right)(\mathbf{P_1}-\mathbf{T_0})-\mathbf{R}^{-1}\mathbf{T} \\ &-\mathbf{R}^{-1}\frac{\mathbf{n}_{\mathbf{P_1}}^T\left(\mathbf{R}^{-1}-\mathbf{I}\right)(\mathbf{P_1}-\mathbf{T_0})}{\mathbf{n}_{\mathbf{P_1}}^T\mathbf{R}^{-1}\mathbf{u}}\mathbf{u}-\mathbf{R}^{-1}\frac{\mathbf{n}_{\mathbf{P_1}}^T\mathbf{R}^{-1}\mathbf{T}}{\mathbf{n}_{\mathbf{P_1}}^T\mathbf{R}^{-1}\mathbf{u}}\mathbf{u}. \quad (21)\end{aligned}$$

from which (12) follows.

Appendix B Derivation of (13)

When the motion is small, the inverse of the Rodrigues Rotation matrix, $\mathbf{R}^{-1}$, can be obtained from $-\mathbf{\Omega}$. So, the first term in the RHS of (21) can be rewritten as

$$\left(\mathbf{R}^{-1}-\mathbf{I}\right)(\mathbf{P_1}-\mathbf{T_0}) \cong (\mathbf{P_1}-\mathbf{T_0})^{\wedge}\begin{pmatrix}\omega_x\\ \omega_y\\ \omega_z\end{pmatrix} \triangleq \hat{\mathbf{P}}\mathbf{\Omega}. \quad (22)$$

For the third term in (21), we have

$$\begin{aligned}&\mathbf{R}^{-1}\frac{\mathbf{n}_{\mathbf{P_1}}^T\left(\mathbf{R}^{-1}-\mathbf{I}\right)(\mathbf{P_1}-\mathbf{T_0})}{\mathbf{n}_{\mathbf{P_1}}^T\mathbf{R}^{-1}\mathbf{u}}\mathbf{u} \cong \mathbf{R}^{-1}\frac{\mathbf{n}_{\mathbf{P_1}}^T\hat{\mathbf{P}}\mathbf{\Omega}}{\mathbf{n}_{\mathbf{P_1}}^T\mathbf{R}^{-1}\mathbf{u}}\mathbf{u} \\ &= \frac{1}{\mathbf{n}_{\mathbf{P_1}}^T\mathbf{R}^{-1}\mathbf{u}}\mathbf{R}^{-1}\mathbf{u}\mathbf{n}_{\mathbf{P_1}}^T\hat{\mathbf{P}}\mathbf{\Omega}. \quad (23)\end{aligned}$$

Since $\mathbf{u}$ is a unit vector, each of it's 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, $\mathbf{R}^{-1}\mathbf{u} \cong \mathbf{u}$. Substituting back into (23), we have

$$\mathbf{R}^{-1}\frac{\mathbf{n}_{\mathbf{P_1}}^T\left(\mathbf{R}^{-1}-\mathbf{I}\right)(\mathbf{P_1}-\mathbf{T_0})}{\mathbf{n}_{\mathbf{P_1}}^T\mathbf{R}^{-1}\mathbf{u}}\mathbf{u} \cong \frac{1}{\mathbf{n}_{\mathbf{P_1}}^T\mathbf{u}}\mathbf{u}\mathbf{n}_{\mathbf{P_1}}^T\hat{\mathbf{P}}\mathbf{\Omega}. \quad (24)$$

Using similar reasoning for the fourth term in (21), we have

$$\mathbf{R}^{-1}\frac{\mathbf{n}_{\mathbf{P_1}}^T\mathbf{R}^{-1}\mathbf{T}}{\mathbf{n}_{\mathbf{P_1}}^T\mathbf{R}^{-1}\mathbf{u}}\mathbf{u} = \frac{1}{\mathbf{n}_{\mathbf{P_1}}^T\mathbf{u}}\mathbf{u}\mathbf{n}_{\mathbf{P_1}}^T\mathbf{R}^{-1}\mathbf{T}. \quad (25)$$

Finally, $\mathbf{R}^{-1}\mathbf{T} \cong \mathbf{T}$ by neglecting the terms which are products of the component of $\mathbf{\Omega}$ and $\mathbf{T}$. Substituting back into (21), we have

$$\begin{aligned}\mathbf{\Delta} &\cong \hat{\mathbf{P}}\mathbf{\Omega}-\frac{1}{\mathbf{n}_{\mathbf{P_1}}^T\mathbf{u}}\mathbf{u}\mathbf{n}_{\mathbf{P_1}}^T\hat{\mathbf{P}}\mathbf{\Omega}-\mathbf{T}-\frac{1}{\mathbf{n}_{\mathbf{P_1}}^T\mathbf{u}}\mathbf{u}\mathbf{n}_{\mathbf{P_1}}^T\mathbf{T} \\ &= \left(\mathbf{I}-\frac{1}{\mathbf{n}_{\mathbf{P_1}}^T\mathbf{u}}\mathbf{u}\mathbf{n}_{\mathbf{P_1}}^T\right)\left(\hat{\mathbf{P}}\mathbf{\Omega}-\mathbf{T}\right). \quad (26)\end{aligned}$$

References

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

[2] P. Belhumeur and D. Kriegman, "What Is the Set of Images of an Object Under All Possible Lighting Conditions?", *IEEE Conf. Computer Vision and Pattern Recognition*, pp.270-277, 1996.

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

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

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

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

[7] 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.

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

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

[10] S. Negaudaripour, "Revised Definition of Optical Flow: Integration of Radiometric and Geometric Cues for Dynamic Scene Analysis", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.20, no.9, pp.961-979, 1998.

[11] J. Oliensis and P. Dupuis, "Direct method for reconstructing shape from shading", *Proc. SPIE Conf. 1570 on Geometric Methods in Computer Vision*, pp. 116-128, 1991.

[12] A. Shashua, "On Photometric Issues in 3D Visual Recognition from a Single 2D Image", *Int'l J. Computer Vision*, vol.21, no.1-2, pp.99-122, 1997.

[13] M.A.O. Vasilescu and D. Terzopoulos, "Multilinear independent components analysis", *In Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, vol.I, pp.547-553, San Diego, CA, Jun. 2005.

[14] 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", *Proceedings of the 9th IEEE International Conference on Computer Vision*, pp.618-625, 13-16 Oct., 2003.