A Theoretical Analysis of Linear and Multi-linear Models of Image Appearance

Yilei Xu, Amit K. Roy-Chowdhury *
Department of Electrical Engineering
University of California, Riverside

Abstract

Linear and multi-linear models of object shape/appearance (PCA, 3DMM, AAM/ASM, multilinear tensors) have been very popular in computer vision. In this paper, we analyze the validity of these models from the fundamental physical laws of object motion and image formation. We rigorously prove that the image appearance space can be closely approximated to be locally multilinear, with the illumination subspace being bilinearly combined with the direct sum of the motion, deformation and texture subspaces. This result allows us to understand theoretically many of the successes and limitations of the linear and multi-linear approaches existing in the computer vision literature, and also identifies some of the conditions under which they are valid. It provides an analytical representation of the image space in terms of different physical factors that affect the image formation process. Experimental analysis of the accuracy of the theoretical models is performed as well as tracking on real data using the analytically derived basis functions of this space.

1. Introduction

Linear, multi-linear, and non-linear models of object shape/appearance have been very popular in computer vision. Examples include principal components analysis (PCA), active appearance/shape models (AAM/ASM) [9, 4], 3D morphable models (3DMM) [3], multi-linear models (MLM) [12, 13], non-linear manifolds [7], among others. To resolve questions about the effectiveness and accuracy of these methods, experimental evaluations have been carried out on larger and larger datasets. While these experiments are a very valuable contribution, it is also important to analyze the accuracy of these models from the fundamental physical laws of object motion and image formation. Such an analysis will allow us to understand the conditions under which each of them is valid. This paper is a rigorous theoretical study along that direction.

*The authors were partially supported by NSF grant IIS-0712253.

1.1. Overview of the Theoretical Results

• Starting from fundamental physics-based models governing rigid object motion, deformations, the interaction of light with the object and perspective projection, we derive a description of the mathematical space in which an image lies. Specifically, we prove that the image space can be closely approximated to be *locally* multilinear, with the illumination subspace being bilinearly combined with the direct sum of the motion, deformation and texture subspaces.
• This result allows us to justify theoretically the validity of many of the linear and multi-linear approaches existing in the computer vision literature, while also identifying some of the physical constraints under which they are valid. In fact, as explained in Section 3.2, we can now understand theoretically why some methods have worked well in some situations, and not so well in others.
• While assuming local linearity may be intuitive, we provide, possibly for the first time an analytical description of this image space in terms of different physical factors that affect the image formation process.
• We show that since we can analytically express the image space, we can estimate the motion, deformation and lighting parameters without needing a large number of training examples to first learn the characteristics of this space and the estimates are not a function of the learning data. This analytical expression can be used in future with learning-based methods for more efficient image modeling.

Relation to Existing work: The theoretical analysis in this paper builds on some recent work that have described image appearance in terms of mathematical models derived from fundamental physical laws. In describing the effect of lighting on an object, researchers have obtained descriptions of the illumination space, e.g., illumination cone [2] and basis illumination models [1, 10]. A more recent result showed that rigid motion and lighting were related bilinearly [14] in the image appearance space. In this paper, we consider a much more general condition than any of the above - an imaged object undergoing a rigid motion (i.e., pose change) while deforming and the illumination also changing randomly. The theoretical derivation is based on a few weak assumptions - a finite dimensional vector space representa-

tion of illumination, small time interval between two image instances, smooth 3D surface and texture of the object that are differentiable.

2. Theoretical Derivation of the Image Appearance Space

2.1. Problem formulation

Consider an object whose images are being captured by a perspective camera. We attach the world reference frame to the camera. Let the 3D surface of the object be described by $\mathcal{C}(u, v) \in \mathbb{R}^3$ in the object reference frame, where $\mathcal{C}$ is parameterized using u and v. Consider two time instances t_1 and $t_2 = t_1 + \Delta t$, between which the object can move rigidly and deform (see Fig. 1).

Let the pose of the object with respect to the camera reference frame before the motion to be defined as the translation $\mathbf{T}$ and rotation matrix $\mathbf{R}$. The rigid motion of the object is represented as the translation $\Delta\mathbf{T} = \mathbf{V}\Delta t$ of the centroid and the rotation $\Delta\boldsymbol{\Omega} = \omega\Delta t$ about the centroid of the object during the time interval Δt. $\boldsymbol{\Delta}\mathbf{R} = e^{\hat{\omega}\Delta t}$ is the rotation matrix due to $\Delta\boldsymbol{\Omega}$, and $\hat{\omega} \in \mathbb{SO}(3)$ is the skew-symmetric matrix corresponding to $\omega \in \mathbb{R}^3$. Deformation is defined in the object reference frame. While the object is deforming, its texture may also change and the illumination may be different at t_1 and t_2. Our goal is to express the image $\mathcal{I}_{t_2}$ mathematically as a function of motion $\Delta\mathbf{T}$ and $\Delta\boldsymbol{\Omega}$, deformation, illumination, and texture change.

We make the following assumptions, which are valid in most situations and will discuss them later in Section 3.1.
A1) Illumination is represented by a finite dimension linear orthogonal basis.
A2) Δt is small, which implies that the motion between t_1 and t_2 is small.
A3) $\mathcal{C}(u, v)$ is smooth and both the deformation and the change of texture are smooth, allowing $\frac{\partial^2\mathcal{C}}{\partial u\partial t} = \frac{\partial^2\mathcal{C}}{\partial t\partial u}$ and $\frac{\partial^2\mathcal{C}}{\partial v\partial t} = \frac{\partial^2\mathcal{C}}{\partial t\partial v}$.

For ease of explanation, we start from a fixed rigid object under varying illumination. Then we consider the problems of a moving rigid object under varying illumination and a fixed deforming object under varying illumination (**Theorem 1**). Next we consider a moving and deforming object under fixed illumination (**Theorem 2**), and a moving and deforming object under varying illumination (**Theorem 3**). We prove that the image space of a moving and deforming object under varying illumination is a locally multilinear. When we keep higher order terms, the image space become nonlinear.

2.2. Fixed Rigid Object under Varying Illumination

In [1, 10], the authors showed that, when a rigid object is fixed with respect to the camera, the reflectance image $\mathcal{I}$ of size $P \times Q$ can be represented as

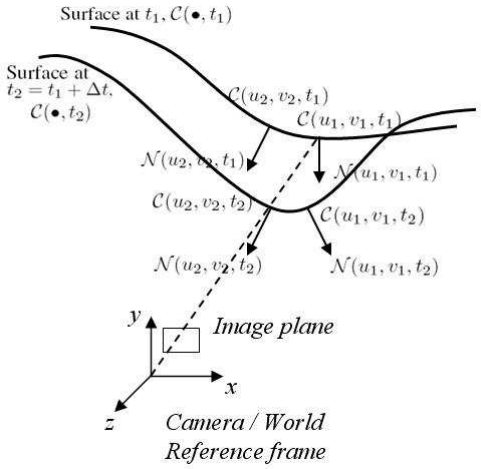


Figure 1. Pictorial representation depicting imaging framework.

$$\mathcal{I} = \mathbf{l}^{\mathbf{T}}\mathcal{B}_l(\mathbf{n}) = \mathcal{B}_l(\mathbf{n}) \times_1 \mathbf{l}, \tag{1}$$

where the 2D tensor $\mathcal{I} \in \mathbb{R}^{1\times P\times Q}$ is the reflectance image, $\mathbf{l} \in \mathbb{R}^{N_l\times 1}$ is the illumination coefficient vector determined by the illumination conditions, $\mathcal{B}_l \in \mathbb{R}^{N_l\times P\times Q}$ is the tensor version of a set of basis images, $\mathbf{n}$ is the unit norm vector at the reflection point, and $\times_n$ is called the *mode-n product* [6][1]. For a Lambertian object with attached shadows, $N_l \approx 9$. The bases for each pixel can be expressed as [1]

$$b_i(\mathbf{n}_j) = \rho_j r_i Y_i(\mathbf{n}_j), i = 0, 1, \ldots, \tag{2}$$

where ρ encrypts the surface reflectance property at the reflection point, Y_i is the spherical harmonics function, and r_i is a constant for each spherical harmonics order. For each pixel, b_i is a vector. Arranging the b_i for all the pixels together will give the tensor $\mathcal{B}_l$. When the Lambertian reflectance property is not satisfied, higher orders of the spherical harmonics functions will be needed [10].

2.3. Moving Rigid Object under Varying Illumination

Under assumptions (A1) and (A2), the authors in [14], proved that the image space can be approximated by a bilinear function of the illumination and rigid motion parameters, i.e.,

$$\mathcal{I}_{t_2} = (\mathcal{B}_{l|t_1} + \mathcal{B}_{m|t_1} \times_2 \mathbf{m}) \times_1 \mathbf{l}_{\mathbf{t_2}}, \tag{3}$$

where $\mathcal{B}_{m|t_1} \in \mathbb{R}^{9\times 6\times P\times Q}$ is the tensor version of the motion bases [2], and $\mathbf{m} = (\Delta\mathbf{T}^{\mathbf{T}}, \Delta\boldsymbol{\Omega})^{\mathbf{T}}$ is the motion parameter vector. However, this requires the object to be rigid.

[1]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

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

[2]The exact forms of b_i, $\mathcal{B}_l$, and $\mathcal{B}_m$ can be found in [1, 14]. Defining them precisely requires introducing a lot of notation for which we lack

2.4. Deforming Object at Fixed Pose and under Varying Illumination

Consider that the pose of the object is fixed with respect to the camera, but that is is deforming. The surface of the object is a function of time, i.e. $\mathcal{C}(u,v,t) : \mathbb{R}^2 \times [0,T) \rightarrow \mathbb{R}^3$. Assume that the evolution of the surface obeys the following PDE:

$$\frac{\partial \mathcal{C}(u,v,t)}{\partial t} = \beta(u,v,t)\mathcal{N}(u,v,t). \quad (4)$$

The derivation of this model can be found in Section 2.1 of [11]. Thus, given the parameterization (u,v), the deformation of the object is defined by the function $\beta(u,v,t)$, where $\mathcal{N}(u,v,t)$ is the surface normal at $\mathcal{C}(u,v,t)$. At the time instance t, $\beta(u,v,t)$ is a 2D function and can be decomposed using most of the 2D transformation techniques, including 2D unitary transforms, wavelet transforms, and B-spline basis, among others. Assuming the deformation of an object to be smooth over (u,v), most of the energy of $\beta(u,v,t)$ at time instance t would be concentrated in the low frequency components. Decomposing $\beta(u,v,t)$ using the top N_D bases, we have

$$\beta(u,v,t) = \mathbf{b}_d(t)^{\mathbf{T}}\Phi_d(u,v), \quad (5)$$

where $\Phi_d \in \mathbb{R}^{N_D \times 1}$ is the vector of the top N_D basis at (u,v), and $\mathbf{b}_d \in \mathbb{R}^{N_D \times 1}$ encrypts the deformation at (u,v) as a function of t. In the case that texture changes gradually while deforming, we can model the change of texture using a similar reasoning as

$$\frac{\partial \rho(u,v,t)}{\partial t} = \gamma(u,v,t), \quad (6)$$

and define $$\gamma(u,v,t) = \mathbf{b}_\rho(t)^{\mathbf{T}}\Phi_\rho(u,v), \quad (7)$$

where $\mathbf{b}_\rho \in \mathbb{R}^{N_\rho \times 1}$ and $\Phi_\rho \in \mathbb{R}^{N_\rho \times 1}$. Then we have the following theorem.

Theorem 1 *Under Assumptions (A1), (A2) and (A3), the image space of a fixed deforming object under varying illumination is locally bilinear, with the illumination subspace being bilinearly combined with the direct sum of the deformation and texture subspaces.*

Outline of the proof: Let A and B represent the same object before and after deformation respectively, as shown in Fig. 1. The ray from the optical center to a particular pixel (x,y) intersects with the surface of the object at some point. Before the object's deformation, the ray intersects with the surface at $\mathcal{C}(u_1,v_1,t_1)$ (on A), and after deformation, it intersects at $\mathcal{C}(u_2,v_2,t_2)$ (on B). During the deformation, $\mathcal{C}(u_2,v_2,t_1)$ (on A) evolves to $\mathcal{C}(u_2,v_2,t_2)$ (on B). Note that $\mathcal{C}(u_2,v_2,t_2)$ may not overlap with $\mathcal{C}(u_1,v_1,t_1)$ - they are just on the same projection ray.

sufficient space. Our interest is in the forms of the expression only. This paper can be understood without knowing the exact details of these terms.

From (1), we see that when the illumination coefficients, l_i, are known, only the norm and the reflectance of the surface point of interest affect the reflection intensity at a particular pixel. The difference between $\mathcal{N}(u_1,v_1,t_1)$ and $\mathcal{N}(u_2,v_2,t_2)$ consists of two parts. The first part is the change from $\mathcal{N}(u_1,v_1,t_1)$ to $\mathcal{N}(u_2,v_2,t_1)$, which can be approximated using a first order Taylor expansion at $\mathcal{C}(u_1,v_1,t_1)$, while the second part is due to the deformation from $\mathcal{N}(u_2,v_2,t_1)$ to $\mathcal{N}(u_2,v_2,t_2)$. Thus we can express the change in norm as

$$\begin{aligned}\Delta\mathcal{N} &= \mathcal{N}(u_2,v_2,t_2) - \mathcal{N}(u_1,v_1,t_1) \\ &= \mathbf{J}_{\mathcal{N}}|_{u_1,v_1,t_1}\boldsymbol{\Delta} + \frac{\partial \mathcal{N}(u_2,v_2,t)}{\partial t}|_{t_1}\Delta t, \quad (8)\end{aligned}$$

where $\mathbf{J}_{\mathcal{N}}|_{u_1,v_1,t_1}$ is the Jacobian matrix of the norm, $\mathcal{N}(u,v,t)$, with respect to the parameters (u,v) at point $\mathcal{C}(u_1,v_1,t_1)$, and $\boldsymbol{\Delta}$ is the difference in the surface parameters (u_2,v_2) and (u_1,v_1). The term $\frac{\partial \mathcal{N}(u_2,v_2,t)}{\partial t}\Delta t$ is due to the deformation. Note that since texture changes gradually, using a similar reasoning we have

$$\begin{aligned}\Delta\rho &= \rho(u_2,v_2,t_2) - \rho(u_1,v_1,t_1) \\ &= \nabla\rho|^{\mathbf{T}}_{u_1,v_1,t_1}\boldsymbol{\Delta} + \frac{\partial \rho(u_2,v_2,t)}{\partial t}|_{t_1}\Delta t. \quad (9)\end{aligned}$$

Thus, $\Delta\,\mathcal{N}$ and $\Delta\rho$ can be substituted into the expression for the basis images in (2), which can be rewritten as

$$\begin{aligned}b_i(u_2,v_2) &= (\rho(u_1,v_1,t_1) + \Delta\rho)r_iY_i(\mathcal{N}(u_1,v_1,t_1) + \Delta\mathcal{N}) \\ &= b_i(u_1,v_1) + \Delta\rho r_iY_i(\mathcal{N}(u_1,v_1,t_1)) \\ &\quad +\rho(u_1,v_1,t_1)r_i\nabla Y_i(\mathcal{N}(u_1,v_1,t_1))\Delta\mathcal{N} \\ &\quad +O(\Delta^2). \quad (10)\end{aligned}$$

The last term is a higher order term and we will discuss it in the later part of the derivation.

Let we now use a subscript $_w$ to denote the variables in the world reference frame. Since $\mathcal{C}_w(u_1,v_1,t_1)$ and $\mathcal{C}_w(u_2,v_2,t_2)$ are on the same ray (see Fig. 1), we can represent the difference between them using a unit vector $\mathbf{r}$ under the perspective camera model as

$$\mathcal{C}_w(u_2,v_2,t_2) - \mathcal{C}_w(u_1,v_1,t_1) = k\mathbf{r}. \quad (11)$$

The transformation between the world frame and the object frame can be written as

$$\begin{aligned}\mathcal{C}_w(u_1,v_1,t_1) &= \mathbf{R}\mathcal{C}(u_1,v_1,t_1) + \mathbf{T}, \\ \mathcal{C}_w(u_2,v_2,t_2) &= \mathbf{R}\mathcal{C}(u_2,v_2,t_2) + \mathbf{T}. \quad (12)\end{aligned}$$

Note that the pose of the object is fixed during the deformation). Using (4),(5),(6), and (7), the evolution of the object surface can be rewritten in a discrete format as

$$\begin{aligned}\mathcal{C}(u_2,v_2,t_2) &= \mathcal{C}(u_2,v_2,t_1) \\ &\quad +\mathbf{b}_d^{\mathbf{T}}(t_1)\Phi_d(u_2,v_2)\mathcal{N}(u_2,v_2,t_1)\Delta t, \\ \rho(u_2,v_2,t_2) &= \rho(u_2,v_2,t_1) + \mathbf{b}_\rho^{\mathbf{T}}(t_1)\Phi_\rho(u_2,v_2)\Delta t. \quad (13)\end{aligned}$$

Under Assumption (A2), which implies that the deformation between the two consecutive frames is small, the point $\mathcal{C}(u_2, v_2, t_1)$ should be close to the point $\mathcal{C}(u_1, v_1, t_1)$. Thus, we may alternatively consider that the new point $\mathcal{C}(u_2, v_2, t_1)$ is on the tangent plane that passes through the point $\mathcal{C}(u_1, v_1, t_1)$, i.e.,

$$\mathcal{C}(u_2, v_2, t_1) = \mathcal{C}(u_1, v_1, t_1) + \alpha_u \mathcal{T}_u|_{u_1,v_1,t_1} + \alpha_v \mathcal{T}_v|_{u_1,v_1,t_1}, \quad (14)$$

where $\mathcal{T}_u|_{u_1,v_1,t_1}$ represents the tangent $\mathcal{T}_u$ at (u_1, v_1, t_1). After a series of manipulations, we have (see Appendix)

$$\mathbf{A}_{t_1} \begin{pmatrix} \alpha_u \\ \alpha_v \end{pmatrix} = -\mathbf{b}_d^{\mathbf{T}}(t_1)\Phi_d \left(\mathbf{I} - \frac{\mathbf{R}^{-1}\mathbf{r}\mathcal{N}^{\mathbf{T}}}{\mathcal{N}^{\mathbf{T}}\mathbf{R}^{-1}\mathbf{r}}\right)\Delta t, \text{where}$$

$$\mathbf{A}_{t_1} = (\mathbf{I} - \frac{\mathbf{R}^{-1}\mathbf{r}\mathcal{N}^{\mathbf{T}}}{\mathcal{N}^{\mathbf{T}}\mathbf{R}^{-1}\mathbf{r}})(\mathbf{b}_d^{\mathbf{T}}(t_1)\Phi_d\mathbf{J}_{\mathcal{N}}\Delta t + \mathcal{N}\mathbf{b}_d^{\mathbf{T}}(t_1)\nabla\Phi_d\Delta t) + (\mathcal{T}_u|_{t_1}, \mathcal{T}_v|_{t_1}). \quad (15)$$

Note that in (15), $\mathcal{T}_u, \mathcal{T}_v, \mathcal{N}, \mathbf{J}_{\mathcal{N}}, \mathbf{R}, \mathbf{r}$ are computed at t_1 and $\Phi_d, \nabla\Phi_d$ are constants in time. The first term $(\mathbf{I} - \frac{\mathbf{R}^{-1}\mathbf{r}\mathcal{N}^{\mathbf{T}}}{\mathcal{N}^{\mathbf{T}}\mathbf{R}^{-1}\mathbf{r}})(\mathbf{b}_d^{\mathbf{T}}\Phi_d\mathbf{J}_{\mathcal{N}}\Delta t + \mathcal{N}\mathbf{b}_d^{\mathbf{T}}\nabla\Phi_d\Delta t) \sim O(\Delta t)$, while the second term $(\mathcal{T}_u|_{t_1}, \mathcal{T}_v|_{t_1}) \sim O(1)$. Thus, using Assumption (A2) that Δt is small, the first term in the right hand side of the expression of $\mathbf{A}_{t_1}$ in (15) can be ignored with respect to the second term. Consequently, the solution of (α_u, α_v) can be written as

$$\begin{pmatrix} \alpha_u \\ \alpha_v \end{pmatrix} = \mathbf{B}_{t_1}\mathbf{b}_d(t_1)\Delta t, \text{where}$$

$$\mathbf{B}_{t_1} = -(\mathcal{T}_u, \mathcal{T}_v)^+(\mathbf{I} - \frac{\mathbf{R}^{-1}\mathbf{r}\mathcal{N}^{\mathbf{T}}}{\mathcal{N}^{\mathbf{T}}\mathbf{R}^{-1}\mathbf{r}})\mathcal{N}\Phi_d^{\mathbf{T}}, \quad (16)$$

and $(\mathcal{T}_u, \mathcal{T}_v)^+$ indicates the pseudo inverse of the non-square matrix $(\mathcal{T}_u, \mathcal{T}_v)$.

In (8), using Assumption (A2) to neglect the terms $O(\Delta t^2)$ with respect to $O(\Delta t)$ and Assumption (A3) for smooth deformation, we have (see Appendix)

$$\frac{\partial\mathcal{N}}{\partial t}|_{u_2,v_2,t_1}\mathbf{\Delta t} \approx -(\mathbf{J}_{\mathcal{N}}(\mathcal{C}|(u,v))\mathbf{J}_{\mathcal{N}}(\Phi_d|(u,v))^{\mathbf{T}}\mathbf{b}_d(t_1)\Delta t. \quad (17)$$

Thus, substituting (16) and (17) back into (8) and (9), the change of the norm and ρ can be expressed as

$$\Delta\mathcal{N} = (\mathbf{J}_{\mathcal{N}}|_{u_1,v_1,t_1}\mathbf{B}_{t_1} - \nabla\mathcal{C}|_{u_1,v_1,t_1}\nabla\Phi_d|^{\mathbf{T}}_{u_1,v_1,t_1})\mathbf{b}_d(t_1)\Delta t,$$
$$\Delta\rho = \nabla\rho|^{\mathbf{T}}_{u_1,v_1,t_1}\mathbf{B}_{t_1}\mathbf{b}_d(t_1)\Delta t + \Phi_\rho|^{\mathbf{T}}_{u_1,v_1,t_1}\mathbf{b}_\rho(t_1)\Delta t. \quad (18)$$

Thus, both $\Delta\mathcal{N}$ and $\Delta\rho$ are linear functions of $\mathbf{b}_d$ and $\mathbf{b}_\rho$. Substituting back into (10), and using tensor notation, we will have

$$\mathcal{I}_{t_2} = (\mathcal{B}_{l|t_1} + \mathcal{B}_{d\rho|t_1} \times_2 \begin{pmatrix} \mathbf{b}_d \\ \mathbf{b}_\rho \end{pmatrix}\Delta t) \times_1 \mathbf{l}_{t_2}, \quad (19)$$

where $\mathcal{B}_{d\rho|t_2} \in \mathbb{R}^{N_l\times(N_D+N_\rho)\times P\times Q}$ is the tensor version of the deformation and texture change basis. Thus, the image space is a locally bilinear function of the illumination parameters and the union of the deformation and texture change parameters. The locality property comes because this description is for a small deformation at a fixed pose. □

2.5. Moving and Deforming Object under Fixed Illumination

Theorem 2 *Under Assumptions (A1), (A2) and (A3), the image space of a rigidly moving and deforming object under fixed illumination is the direct sum of the motion, deformation and texture subspaces.*

Outline of the Proof: Reconsider Figure 1. We still have

$$\mathcal{C}_w(u_2, v_2, t_2) - \mathcal{C}_w(u_1, v_1, t_1) = k\mathbf{r}, \quad (20)$$
$$\mathcal{C}_w(u_1, v_1, t_1) = \mathbf{R}\mathcal{C}(u_1, v_1, t_1) + \mathbf{T},$$
$$\mathcal{C}_w(u_2, v_2, t_2) = \mathbf{\Delta R}\mathbf{R}\mathcal{C}(u_2, v_2, t_2) + \mathbf{\Delta T} + \mathbf{T}, \quad (21)$$

where $\mathbf{r}$ is the unit vector along the projection ray. Similarly, the deformation of the object can still be described using (13) as

$$\begin{aligned}\mathcal{C}(u_2, v_2, t_2) &= \mathcal{C}(u_2, v_2, t_1) + \beta(u_2, v_2)\mathcal{N}(u_2, v_2, t_1)\Delta t \\ &= \mathcal{C}(u_2, v_2, t_1) \\ &\quad + \mathbf{b}_d^{\mathbf{T}}(t_1)\Phi_d(u_2, v_2)\mathcal{N}(u_2, v_2, t_1)\Delta t. \quad (22)\end{aligned}$$

Because the time interval between the two consecutive frames is small, the motion and deformation are small. Using similar reasoning as used for deriving (14), we again have

$$\mathcal{C}(u_2, v_2, t_1) = \mathcal{C}(u_1, v_1, t_1) + \alpha_u \mathcal{T}_u|_{u_1,v_1,t_1} + \alpha_v \mathcal{T}_v|_{u_1,v_1,t_1}. \quad (23)$$

After a series of manipulations, we have (see Appendix)

$$\mathbf{A}_{t_1} \begin{pmatrix} \alpha_u \\ \alpha_v \end{pmatrix} = (\mathbf{I} - \frac{\mathbf{R}^{-1}\mathbf{r}\mathcal{N}^{\mathbf{T}}}{\mathcal{N}^{\mathbf{T}}\mathbf{R}^{-1}\mathbf{r}})(\hat{\mathbf{C}}_1\Delta\mathbf{\Omega} - \mathbf{R}^{-1}\Delta\mathbf{T} - \mathcal{N}\Phi_d^{\mathbf{T}}\mathbf{b}_d(t_1)\Delta t)),$$

where

$$\mathbf{A}_{t_1} = (\mathcal{T}_u, \mathcal{T}_v) + (\mathbf{I} - \frac{\mathbf{R}^{-1}\mathbf{r}\mathcal{N}^{\mathbf{T}}}{\mathcal{N}^{\mathbf{T}}\mathbf{R}^{-1}\mathbf{r}})(\mathbf{b}_d^{\mathbf{T}}(t_1)\Phi_d\mathbf{J}_{\mathcal{N}} + \mathcal{N}\mathbf{b}_d^{\mathbf{T}}(t_1)\nabla\Phi_d)\Delta t \quad (24)$$

Under similar reason used in deriving (16), we can again neglect the second term in the expression of $\mathbf{A}_{t_1}$ in (24), and the solution to $(\alpha_u, \alpha_v)^{\mathbf{T}}$ can be obtained as

$$\begin{pmatrix} \alpha_u \\ \alpha_v \end{pmatrix} = -(\mathcal{T}_u, \mathcal{T}_v)^+(\mathbf{I} - \frac{\mathbf{R}^{-1}\mathbf{r}\mathcal{N}^{\mathbf{T}}}{\mathcal{N}^{\mathbf{T}}\mathbf{R}^{-1}\mathbf{r}})(\hat{\mathbf{C}}_1\Delta\mathbf{\Omega} - \mathbf{R}^{-1}\Delta\mathbf{T} - \mathcal{N}\Phi_d^{\mathbf{T}}\mathbf{b_d}(\mathbf{t_1})\Delta t)$$
$$\triangleq \mathbf{D}\Delta\mathbf{\Omega} + \mathbf{E}\Delta\mathbf{T} + \mathbf{F}\mathbf{b}_d(t_1)\Delta t. \quad (25)$$

However, when there exists both rigid motion and deformation, the temporal change of $\mathcal{N}(u_2, v_2)$ from t_1 to t_2 consists of two parts: one due to the deformation, and one due to the rotation. In Appendix, we derived the temporal change of norm from (4), which is purely due to deformation, i.e.,

$$\frac{\partial\mathcal{N}}{\partial t}|_{u_2,v_2,t_1}\mathbf{\Delta t}|_{\Delta\mathbf{\Omega}=0} \approx -\nabla\mathcal{C}|_{u_1,v_1,t_1}\nabla\Phi_d|^{\mathbf{T}}_{u_1,v_1,t_1}\mathbf{b}_d(t_1)\Delta t. \quad (26)$$

The temporal change of normal due to the rigid rotation by $\Delta\boldsymbol{\Omega}$ is

$$\frac{\partial \mathcal{N}}{\partial t}|_{u_2,v_2,t_1}\Delta t|_{\mathbf{b}_d=0} \approx -\hat{\mathcal{N}}|_{u_1,v_1,t_1}\Delta\boldsymbol{\Omega}. \quad (27)$$

Thus, substituting (25), (26) and (27) back into (8) and (9), the change of the norm and ρ can be expressed as

$$\begin{aligned}\Delta\mathcal{N} &= (\mathbf{J}_{\mathcal{N}}|_{u_1,v_1,t_1}\mathbf{D} - \hat{\mathcal{N}}|_{u_1,v_1,t_1})\Delta\boldsymbol{\Omega} + \mathbf{J}_{\mathcal{N}}|_{u_1,v_1,t_1}\mathbf{E}\Delta\mathbf{T} \\ &\quad +(\mathbf{J}_{\mathcal{N}}|_{u_1,v_1,t_1}\mathbf{F} - \nabla\mathcal{C}|_{u_1,v_1,t_1}\nabla\Phi_d|^{\mathbf{T}}_{u_1,v_1,t_1})\mathbf{b}_d\Delta t, \\ \Delta\rho &= \nabla\rho|^{\mathbf{T}}_{u_1,v_1,t_1}(\mathbf{D}\Delta\boldsymbol{\Omega} + \mathbf{E}\Delta\mathbf{T} + \mathbf{F}\mathbf{b}_d(t_1)\Delta t) \\ &\quad +\Phi_\rho|^{\mathbf{T}}_{u_1,v_1,t_1}\mathbf{b}_\rho(t_1)\Delta t. \end{aligned} \quad (28)$$

Thus, both $\Delta\mathcal{N}$ and $\Delta\rho$ are linear functions of $\mathbf{b}_d$ and $\mathbf{b}_\rho$. Substituting back into (10), and using tensor notation, we will have

$$\mathcal{I}_{t_2} = (\mathcal{B}_{l|t_1} + \mathcal{B}_{md\rho|t_1} \times_2 \begin{pmatrix} \mathbf{V} \\ \omega \\ \mathbf{b}_d \\ \mathbf{b}_\rho \end{pmatrix} \Delta t) \times_1 \mathbf{l}_{t_1}, \quad (29)$$

where $\mathcal{B}_{md\rho|t_1} \in \mathbb{R}^{N_l\times(6+N_D+N_\rho)\times P\times Q}$ is the joint deformation, rigid motion and texture basis obtained by substituting (28) into (10). □

2.6. Moving and Deforming Object under Varying Illumination

Theorem 3 *The image space of a rigidly moving and deforming object under varying illumination is locally multi-linear, with the illumination subspace being bilinearly combined with the direct sum of the motion, deformation and the texture subspaces.*

Outline of the Proof: When illumination is represented as a function of t as $\mathbf{l}_t$, using augmented variables we can have the following directly from (29):

$$\mathcal{I}_{t_2} = \mathcal{B}_{lmd\rho|t_1} \times_1 \mathbf{l}_{t_2} \times_2 \begin{pmatrix} \mathbf{V} \\ \omega \\ \mathbf{b}_d \\ \mathbf{b}_\rho \\ 1 \end{pmatrix} \Delta t, \quad (30)$$

where $\mathcal{B}_{ldm\rho|t_1} \in \mathbb{R}^{(N_L+1)\times(6+N_D+N_\rho+1)\times P\times Q}$ is the tensor version of the joint illumination, deformation, rigid motion and texture bases. Thus, when illumination, $\mathbf{l}$, is fixed, the local image space is a linear function of the union of the motion, deformation and texture change parameters. The result is valid in a local region around pose $(\mathbf{R}, \mathbf{T})$. □

3. Discussion of the Theoretical Results

3.1. Implications of the Assumptions

We used three assumptions for deriving Theorems 1, 2, and 3. Assumption (A1) essentially says that we use a basis illumination model. This is widely used. For Lambertian surfaces, the basis dimension is small, while non-Lambertian surface requires higher dimensions. Also, the basis function can be represented using spherical harmonics, wavelets, and other orthogonal representations. Our derivation does not need a specific choice, only that it is a function of the surface normal. Assumption (A3) is again not difficult to satisfy for most object surfaces. Assumption (A2) requires that the time interval between two consecutive frames to be small, which means that the motion, deformation, and texture change between the two frames is small. This assumption means that the theoretical result describes an image space in a local region, e.g. images in video sequences captured under frame rates between 15 and 30 fps. It is used to approximate the 3D surface in a small neighborhood by a tangent plane and to neglect higher powers of Δt with respect to lower powers. If higher order terms of Δt are retained, we can show the following:

Theorem 4 *If the second order terms of Δt are retained, the image space of a rigidly moving and deforming object under varying illumination will not be multi-linear but nonlinear.*

Outline of the Proof: In equation (10), we kept the first order term of Δ and ignored the higher order terms. As $\Delta \sim O(\Delta t)$, when we keep the higher order terms of Δt, Δ^2 terms needs to be kept. From (25), we know that keeping the term Δ^2 will introduce not only the cross terms between $\mathbf{T}$, $\Delta\boldsymbol{\Omega}$, $\mathbf{b}_d$, and $\mathbf{b}_\rho$, but also their squares, leading the image space to be not multi-linear, but completely nonlinear. □

Due to the assumption (A2), the result in (30) is modeling the local variation in the image appearance space. A collection of such locally multi-linear manifolds can be used to represent the global space.

3.2. Relation to Existing Methods

This theoretical study provides a rigorous proof for the validity of many linear/multi-linear models of object appearance/shape representation used recently in computer vision. We can also understand the conditions under which these popular models can be applied. We provide below such an analysis, taking face representation and recognition as an example (since these models have been applied to faces).

PCA: From (29) we can see that, when the illumination and pose are fixed, the image space is linear in the shape (deformation) and texture parameters, which encrypt the identity. This proves the validity of the use of PCA under such scenarios. It explains the relatively good performance of PCA when applied to the face recognition problem under fixed pose and illumination and poor performance when illumination is changing. Although our derivation leads to a locally linear model while PCA is globally linear, locally

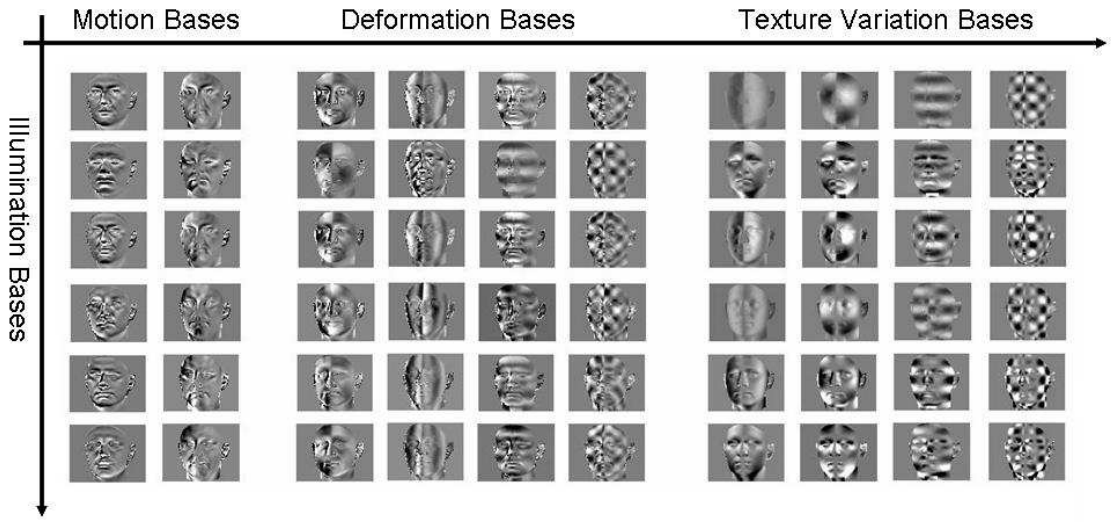


Figure 2. Some representative illumination, motion, deformation and texture variation basis images of a 3D face model.

linear subspaces can be described by using a higher dimensional global linear subspace. Thus, while the theory might predict a small number of bases, PCA will need more bases.

AAM/ASM: AAM/ASM [4] represent shape and appearance using a linear set of basis vectors, which are then mapped non-linearly to the image space. Using our analytically derived bases, the image space can be obtained as in (30) even with pose and illumination variations. This is a simpler form than the AAM/ASM models.

MLM: In MLM [12, 13], different factors are assumed to be multi-linearly combined. We proved that lighting is indeed bilinearly combined with the direct sum of the motion, deformation and texture subspaces. Direct sum is a special case of the multi-linear model with the coefficients of the cross terms being zero. Thus, the multi-linear models are valid but they will end up approximating the direct sum of the linear subspaces. Since this multi-linearity property is local, MLM methods will be more accurate when modeling local variations of the image space. It can be easily shown that the collection of local multi-linear manifolds can be embedded into a higher dimension globally multi-linear manifold, which provides the theoretical validation for MLM.

Local Linearization: Probabilistic Appearance Model (PAM) [8] uses a series of tangent planes along pose to approximate the manifold - thus it is also locally linear. Our theoretical result not only validates this assumption, but also provides an analytical description of this space. In [5], the authors locally linearize the appearance manifold for tracking, but they obtain the linearized basis from a learning algorithm. Again, we provide an analytical description of this linear subspace, which can be used to obtain the bases in a manner that is not dependent on the training data.

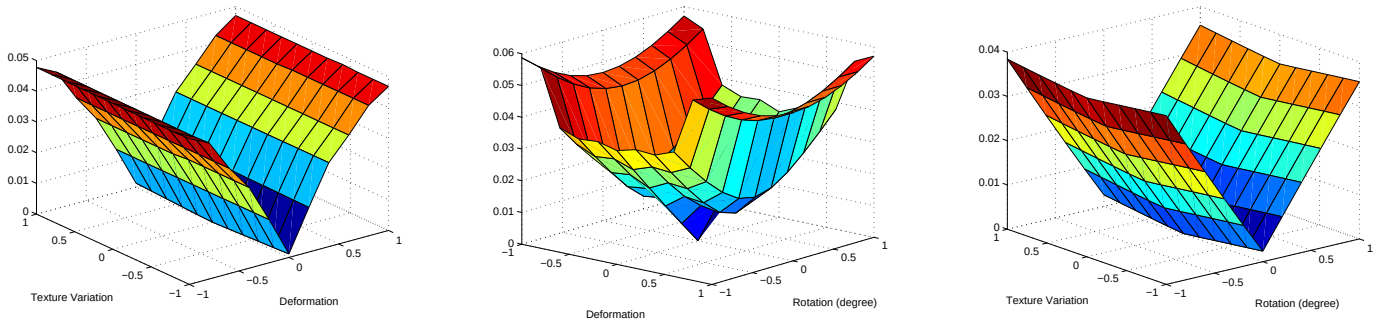
Non-linear approaches: In 3DMM, once the textured 3D shape is obtained, it is combined with the illumination and camera projection model, and thus the image pixel intensities are nonlinear in the shape and texture coefficients. This is a more accurate representation (**Theorem 4**), but comes at the cost of higher computation due to optimization on a non-linear manifold. Non-linear manifolds is also the approach taken in [7].

4. Experimental Results

Computation of Bases: We used a 3D face model obtained from the 3DMM dataset to compute the analytically derived bases, $\mathcal{B}_{lmd\rho}$, in (30). We show some representative basis images in Fig. 2. The first column in the motion bases shows the bases for translation along the vertical axis, while the second column shows in-plane rotation bases. For the deformation and texture bases, we use 2D DCT basis functions for Φ_d and Φ_ρ, and show a few representative ones.

Image Synthesis using Theorem 3: In Fig. 3, we show the comparison between the images synthesized with our theory, and the ones synthesized by simulating the PDEs in (4) and (6). We use a face model with uniform texture, fix the illumination and pose, and then apply deformations on the cheeks and around mouth using 2D DCT basis functions. The texture change is effected over the entire face. In the second row, we show the images synthesized using (30), with the deformation bases, the texture variation bases, and a combination of the two. For comparison, we also show the corresponding images synthesized by simulating PDEs (4) and (6) in the third row. There is very little visual difference between the two.

Numerical Accuracy Analysis: To evaluate the theory in a more precise manner, we performed a numerical er-



(a) (b) (c)

Figure 4. Accuracy analysis of the theoretical model. The error is computed as the squared difference between the theoretically predicted pixel intensities and the true pixel intensities, normalized by the true values, and taking its mean over the face region.

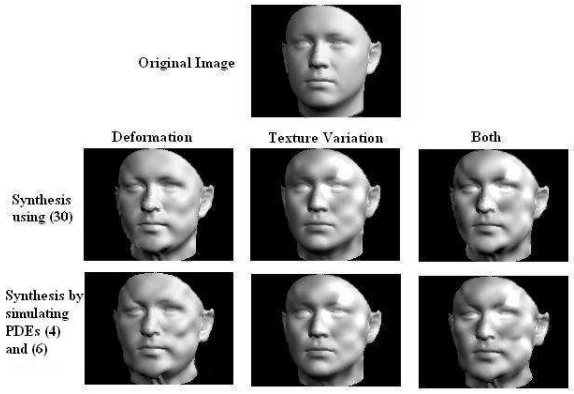


Figure 3. Comparison between the images synthesized with our theory, and the ones synthesized by simulating the PDEs in (4) and (6).

ror analysis. We chose some typical range of rigid motion, deformation, and texture variation between two consecutive frames in a video sequence. We computed the difference between the theoretically predicted pixel intensities and the true pixel intensities, normalized by the true values, and took the mean of this normalized error over the face region in the image. Assuming the face to be a hemisphere, we assumed that in one second, the deformation will not exceed 5% of the radius of this hemisphere, and set $\frac{5\%}{30 \text{ frames}}$ as one unit on the axis of deformation. Similarly, for the texture change, we assume the variance of the change will not exceed 5% of the square of the mean value of the original texture. For the rotation, we let that the maximum degree the object can rotate in one second to be $30°$, which means $1°$ between two consecutive frames.

In Fig. 4, we plot the normalized error versus (a) deformation and texture variation, (b) deformation and rigid motion, and (c) texture variation and motion. We choose rotation along the vertical axis for the motion (as that is a common motion of the face in video). Fig. 4 indicates that, within a typical range of motion, deformation, and texture variation, the normalized error between the predicted value and the true value will not exceed 6%. This is the worst case performance and happens when the object is deforming and rotating. This is justified by the theory since we neglect higher order changes due to deformation and rigid motion in (15) and (24).

Application to Tracking: As an application of the theory, we use it for tracking faces under illumination and expression variations (Figure 5). We use Levenberg-Marquardt method for minimizing the difference between each input frame and the predicted one from (30), and alternatively minimizing over the illumination subspace and the direct sum of the pose, deformation and texture subspace. In Fig. 5, we show the 2D location of the face and the pose parameters. Pose parameters are represented as the Euler angle of the face with respect to the frontal one, following the "z-x-z" convention. Illumination and deformation parameters are not shown due to lack of space.

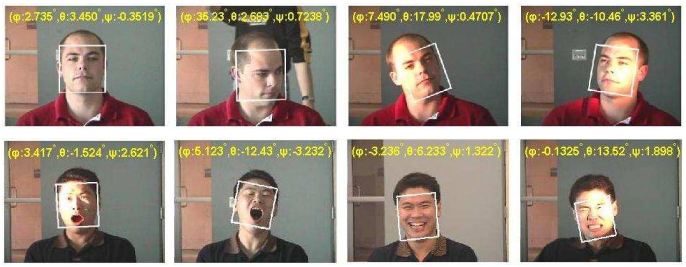


Figure 5. Examples of tracking using the theoretical model on real-data under changes of pose, lighting and expressions. In addition to the 2D location, we also obtain the 3D pose (as shown in the figure), illumination and deformation parameters.

5. Conclusions

In this paper, we analyzed the accuracy of linear and multi-linear object representation models from the fundamental physical laws of object motion and image formation. We proved that the image appearance space is *lo-*

cally multilinear, with the illumination subspace being bilinearly combined with the direct sum of the motion, deformation and texture subspaces. When higher order terms are not neglected, the image space becomes nonlinear. Using this result, we discussed the validity of many of the linear and multi-linear approaches existing in the computer vision literature, including PCA, AAM/ASM, PAM, MLM, and 3DMM. Experimental accuracy analysis of the theoretical results were also presented.

Appendix

Derivation of (15) Substituting (12) into (11), we have

$$\mathcal{C}(u_2, v_2, t_2) - \mathcal{C}(u_1, v_1, t_1) = k\mathbf{R}^{-1}\mathbf{r}. \quad (31)$$

Substituting (13) into (31), we have

$$\alpha_u \mathcal{T}_u + \alpha_v \mathcal{T}_v + \mathbf{b}_d^{\mathbf{T}} \Phi_d(u_2, v_2)\mathcal{N}(u_2, v_2, t_1)\Delta t = k\mathbf{R}^{-1}\mathbf{r}. \quad (32)$$

Applying Taylor expansion, we have

$$\mathbf{b}_d^{\mathbf{T}} \Phi_d(u_2, v_2) = \mathbf{b}_d^{\mathbf{T}} \Phi_d(u_1, v_1) + \mathbf{b}_d^{\mathbf{T}} \nabla\Phi_d|_{u_1, v_1, t_1} \begin{pmatrix} \alpha_u \\ \alpha_v \end{pmatrix},$$

$$\mathcal{N}(u_2, v_2, t_1) = \mathcal{N}(u_1, v_1, t_1) + \mathbf{J}_{\mathcal{N}|u_1, v_1} \begin{pmatrix} \alpha_u \\ \alpha_v \end{pmatrix}. \quad (33)$$

Thus, $\mathbf{b}_d^{\mathbf{T}} \Phi_d(u_2, v_2)\mathcal{N}(u_2, v_2, t_1)$ can be expressed as

$$\left(\mathbf{b}_d^{\mathbf{T}} \Phi_d + \mathbf{b}_d^{\mathbf{T}} \nabla\Phi_d \begin{pmatrix} \alpha_u \\ \alpha_v \end{pmatrix}\right)\left(\mathcal{N} + \mathbf{J}_{\mathcal{N}} \begin{pmatrix} \alpha_u \\ \alpha_v \end{pmatrix}\right) = \mathbf{b}_d^{\mathbf{T}} \Phi_d \mathcal{N}$$
$$+\mathbf{b}_d^{\mathbf{T}} \Phi_d \mathbf{J}_{\mathcal{N}} \begin{pmatrix} \alpha_u \\ \alpha_v \end{pmatrix} + \mathcal{N}\mathbf{b}_d^{\mathbf{T}} \nabla\Phi_d \begin{pmatrix} \alpha_u \\ \alpha_v \end{pmatrix} + o(\alpha_u, \alpha_v), \quad (34)$$

where all the terms are computed at (u_1, v_1, t_1), and the last term is a high order term thus can be ignored. Substituting (34) into (33) and after some algebraic manipulations, we have (15).

Derivation of (17) Starting with (4) and using different differential geometric properties of a surface, we get

$$\frac{\partial \mathcal{N}}{\partial t} = \frac{\beta_u \mathcal{N} \times \mathcal{C}_v + \beta_v \mathcal{C}_u \times \mathcal{N}}{\|\mathcal{C}_u \times \mathcal{C}_v\|}. \quad (35)$$

Therefore, using (5) to compute β_u and β_v, we prove (17). The detailed proof is uploaded as supplementary material.

Derivation of (24) Substituting (21) into (20), we have

$$\mathbf{\Delta R}(\mathcal{C}(u_1, v_1, t_1) + \mathbf{b}_d^{\mathbf{T}} \Phi_d(u_2, v_2)\mathcal{N}(u_2, v_2, t_1)\Delta t$$
$$+\alpha_u \mathcal{T}_u + \alpha_v \mathcal{T}_v) - \mathcal{C}(u_1, v_1, t_1) = k\mathbf{R}^{-1}\mathbf{r} - \mathbf{R}^{-1}\mathbf{\Delta T}. \quad (36)$$

Using (34) to approximate $\mathbf{b}_d^{\mathbf{T}} \Phi_d(u_2, v_2)\mathcal{N}(u_2, v_2, t_1)$, we have

$$\left(\mathbf{\Delta R}(\mathcal{T}_u, \mathcal{T}_v) + \mathbf{b}_d^{\mathbf{T}} \Phi_d \mathbf{\Delta R} \mathbf{J}_{\mathcal{N}} \Delta t + \mathbf{\Delta R}\mathcal{N}\mathbf{b}_d^{\mathbf{T}} \nabla\Phi_d \Delta t\right) \begin{pmatrix} \alpha_u \\ \alpha_v \end{pmatrix}$$
$$= (\mathbf{I} - \mathbf{\Delta R})\mathcal{C}(u_1, v_1, t_1) - \mathbf{b}_d^{\mathbf{T}} \Phi_d \mathbf{\Delta R}\mathcal{N}\Delta t - \mathbf{R}^{-1}\mathbf{\Delta T} + k\mathbf{R}^{-1}\mathbf{r}, \quad (37)$$

where all the $\mathcal{N}, \mathbf{J}_{\mathcal{N}}, \Phi$and$\nabla\Phi$ are at (u_1, v_1, t_1) and subscripts are discarded. Solving and substituting back into (37), we have (24).

References

[1] R. Basri and D. Jacobs. Lambertian Reflectance and Linear Subspaces. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 25(2):218–233, February 2003. 1, 2

[2] P. Belhumeur and D. Kriegman. What is the set of images of an object under all possible lighting conditions? In *IEEE Conf. Computer Vision and Pattern Recognition*, 1996. 1

[3] V. Blanz and T. Vetter. Face recognition based on fitting a 3D morphable model. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 25(9):1063–1074, September 2003. 1

[4] T. Cootes, G. Edwards, and C. Taylor. Active Appearance Models. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 23(6):681–685, June 2001. 1, 6

[5] G. W. J.-M. F. Hua Yang, Marc Pollefeys and A. Ilie. Differential Camera Tracking through Linearizing the Local Appearance Manifold. *IEEE International Conference on Computer Vision and Pattern Recognition*, 2007. 6

[6] L. D. Lathauwer, B. D. Moor, and J. Vandewalle. A Multilinear Singular Value Decomposition. *SIAM J. Matrix Anal. Appl.*, 21(4):1253–1278, 2000. 2

[7] C. Lee and A. Elgammal. Nonlinear shape and appearance models for facial expression analysis and synthesis. *IEEE Conference on Computer Vision and Pattern Recognition*, I:313–320, 2003. 1, 6

[8] K. Lee, J. Ho, M. Yang, and D. Kriegman. Video-based face recognition using probabilistic appearance manifolds. In *Computer Vision and Pattern Recognition*, pages I: 313–320, 2003. 6

[9] I. Matthews and S. Baker. Active Appearance Models Revisited. *International Journal of Computer Vision*, 60(2):135–164, Nov. 2004. 1

[10] R. Ramamoorthi and P. Hanrahan. A signal-processing framework for inverse rendering. In E. Fiume, editor, *SIGGRAPH 2001, Computer Graphics Proceedings*, pages 117–128. ACM Press / ACM SIGGRAPH, 2001. 1, 2

[11] G. Sapiro. *Geometric Partial Differential Equations and Image Analysis*. Cambridge University Press, 2001. 3

[12] J. B. Tenenbaum and W. T. Freeman. Separating style and content with bilinear models. *Neural Computation*, 12(6):1247–1283, 2000. 1, 6

[13] M. Vasilescu and D. Terzopoulos. Multilinear Independent Components Analysis. In *Computer Vision and Pattern Recognition*, 2005. 1, 6

[14] Y. Xu and A. Roy-Chowdhury. Integrating Motion, Illumination and Structure in Video Sequences, With Applications in Illumination-Invariant Tracking. *PAMI*, May 2007. 1, 2