A Physics-Based Analysis of Image Appearance Models

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

Abstract—Linear and multilinear models (PCA, 3DMM, AAM/ASM, and multilinear tensors) of object shape/appearance have been very popular in computer vision. In this paper, we analyze the applicability of these heuristic models from the fundamental physical laws of object motion and image formation. We prove that under suitable conditions, the image appearance space can be closely approximated to be multilinear, with the illumination and texture subspaces being trilinearly combined with the direct sum of the motion and deformation subspaces. This result provides a physics-based understanding of many of the successes and limitations of the linear and multilinear 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. Numerical analysis of the accuracy of the physics-based models is performed, and tracking results on real data are presented.

Index Terms—Image appearance models, theoretical analysis, multilinear, deformation, face tracking.

1 Introduction

Modeling the appearance of an image is a fundamental problem in computer vision. A large number of factors affect the image formation process, including object shape, albedo, pose, illumination, and camera models. A number of models, like active appearance/shape models (AAM/ASM) [17], [5], 3D morphable models (3DMM) [3], multilinear models (MLM) [26], [29], or nonlinear manifolds [15], have been used to construct and parameterize the image appearance manifold in terms of these factors. To resolve questions about the effectiveness and accuracy of these methods, experimental evaluations have been carried out on larger and larger data sets. 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 image formation. In this paper, we prove that under suitable conditions, the image space of a moving and deforming object under varying illumination can be closely approximated to be locally multilinear with the illumination subspace and the texture subspace being trilinearly combined with the direct sum of the motion and deformation subspaces. This result provides an *analytical* representation of the image space in terms of different physical factors that affect the image formation process. Under special circumstances, the image space can be simpler. We show applications on tracking faces in video using this physics-based model.

1.1 Related Work

Until a few years ago, the factors that affect the image formation process (e.g., motion, illumination, and object deformation) were usually studied separately. One of the classical methods for 2D motion estimation on the image plane is optical flow [11], which 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) problem [4], [27], is another classical research area in computer vision. While largely constrained to the analysis of rigid objects, it has been extended to nonrigid objects under orthographic projection [28]. However, most SfM algorithms do not take illumination variation into consideration. The authors in [35] proposed modeling the change of illumination in optical flow and combining it with structure from motion, photometric stereo, and multiview stereo in an optimization framework.

In the study of illumination, Shape from Shading (SfS) [9], [10], [19] 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 an image to estimate the 3D structure in a scene. Shashua [23] and Moses [18] showed that ignoring the effect of shadows, the set of images under varying illumination lies in a 3D linear subspace (photometric stereo). Belhumeur and Kriegman [2] showed that the set of images of an object under arbitrary illumination forms a convex cone in the space of all possible images. In [1] and [20], the authors independently derived an analytical 9D spherical harmonics-based linear representation of the images produced by a Lambertian object with attached shadows. For the specular objects, higher orders of the spherical harmonics functions with nonnegativity constraints were used [24].

Partial differential equations have been used for representing shape deformations [12] with a lot of success in tracking problem. Another common approach for modeling a deforming object is to use a linear combination of bases. 3D Morphable Models (3DMM) [3] decompose the 3D shape and texture of a face along the principle component directions, and is well-known in applications of face image synthesis and face recognition. Active Appearance Model (AAM) [5], [17] is applied to 2D shape and texture. Shape analysis has also been used to study deforming shapes, for example, in human activities [30]; however, it focuses on 2D shapes, and thus is not well designed for modeling pose and illumination variations. Linear dynamical systems have been proposed to model the texture variation in certain stationary stochastic processes, termed dynamic textures [7]. A multilinear extension of it using the higher order SVD has also been proposed for modeling texture variation with multiple factors [6] and further extended in [8].

To combine the effects of these various factors, linear, multilinear, and nonlinear models of object shape/appearance have been popularly used for modeling the image appearance. Principal Components Analysis (PCA) is one of the early attempts at modeling the image appearance variation due to the change of identity in face images, and later applied to model the variations due to the changes of illumination. AAM/ASM [5], [17] tried to model the appearance variation due to the changes of shape and texture. 3DMM [3] is similar to AAM in that it uses linear models for approximating the 3D shape and texture. However, the image appearance manifold is a highly nonlinear function of the parameters and becomes computationally expensive. MLM assumes the image space to be multilinear in the identity, pose, and illumination, and multilinear SVD can be applied to learn the bases [29]. Locally, linear models have been another approach for representing the image appearance space [16], [21], [25]. Nonlinear manifolds [15] have also been proposed for modeling the facial expression variations.

None of the above methods provides an analytical analysis of the validity of these models. A more recent result showed analytically that rigid motion and lighting were related bilinearly [31] in the image appearance space. However, this work assumes the object to be rigid and does not consider variations of texture.

- *Y. Xu is with Navteq Corp., 425 W. Randolph St., Chicago, IL 60606. E-mail: pkuelija@hotmail.com.*
- *A.K. Roy-Chowdhury is with the Department of Electrical Engineering, University of California, Riverside, Room 322 EBU-II, Riversde, CA 92521. E-mail: amitrc@ee.ucr.edu.*

Manuscript received 27 Feb. 2009; revised 27 Oct. 2009; accepted 22 Sept. 2010; published online 29 Nov. 2010.
Recommended for acceptance by K. Kutulakos.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-2009-02-0137.
Digital Object Identifier no. 10.1109/TPAMI.2010.216.

Thus, it cannot be used to model deformation, e.g., facial expression and identity variations. The result in [32] assumed that texture change is smooth in time and led to a result where the texture subspace is linearly combined with the rigid motion and deformation subspaces. It is a special case of this paper.

1.2 Overview and Contributions of the Paper

In this paper, we consider a general image formation process—an imaged object undergoing a rigid motion (i.e., pose change) while deforming and the illumination changing randomly. The theoretical derivation is based on a few weak assumptions that are usually applicable—a finite dimensional vector space representation of illumination, small motion between two consecutive frames, and a smooth differentiable 3D surface (shape and texture) of the object. The following are the main contributions of the paper:

- 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. While assuming local linearity may be intuitive, we provide an analytical description of this image space in terms of different physical factors that affect the image formation process.
- This result allows us to analyze theoretically the validity of many of the linear, locally linear, and multilinear 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 3.1, we can now understand theoretically why some methods have worked well in some situations, and not so well in others.
- 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. Some initial work in this direction was presented in [33].

The paper is organized as follows: Section 2 presents the main result and an outline of the proof. Details of the derivation are given in the supplemental material, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2010.216. We discuss the implications of the result in face recognition in Section 3. In Section 4, the application of this result to tracking is presented. Experimental results are given in Section 5, where we analyze the numerical accuracy of the theoretical results. We also show some results of tracking a face with expression variations through pose and lighting changes. Finally, Section 6 concludes the paper and highlights future work.

2 Theoretical Derivation of the Image Appearance Space

2.1 Problem Formulation

Consider an object whose images are being captured by a static 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).

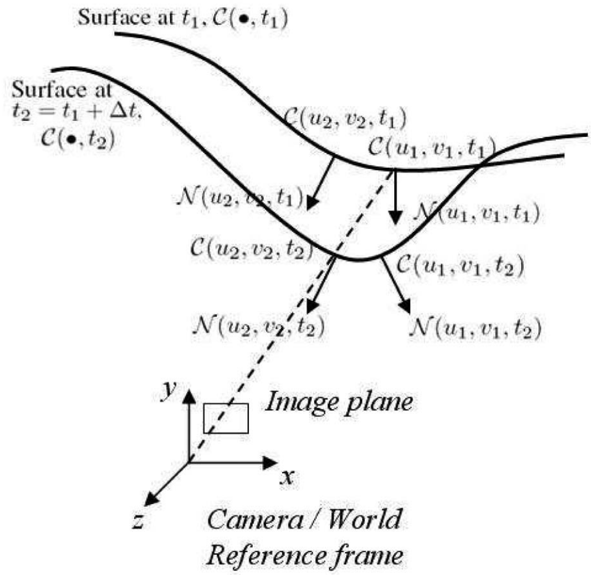

Surface at $t_1, \mathcal{C}(\bullet, t_1)$
Surface at $t_2 = t_1 + \Delta t$, $\mathcal{C}(\bullet, t_2)$
$\mathcal{C}(u_2, v_2, t_1)$
$\mathcal{C}(u_1, v_1, t_1)$
$\mathcal{N}(u_2, v_2, t_1)$
$\mathcal{N}(u_1, v_1, t_1)$
$\mathcal{C}(u_2, v_2, t_2)$
$\mathcal{C}(u_1, v_1, t_2)$
$\mathcal{N}(u_2, v_2, t_2)$
$\mathcal{N}(u_1, v_1, t_2)$
y
Image plane
x
z
Camera / World Reference frame

Fig. 1. Pictorial representation depicting imaging framework.

Let the pose of the object with respect to the camera reference frame before the motion be defined by the translation $\mathbf{T}$ and rotation matrix $\mathbf{R}$. The rigid motion of the object is defined as the translation $\Delta\mathbf{T} = \mathbf{V}\Delta t$ of the centroid and the rotation $\Delta\mathbf{\Omega} = \omega\Delta t$ about the centroid of the object during the time interval Δt. $\Delta\mathbf{R} = e^{\hat{\omega}\Delta t}$ is the rotation matrix due to $\Delta\mathbf{\Omega}$, and $\hat{\omega} \in so(3)$ 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 $\mathcal{I}_{t_1}$, motion $\Delta\mathbf{T}$ and $\Delta\mathbf{\Omega}$, deformation, illumination, and texture change. This will allow us to describe the image appearance variation in terms of the physical parameters.

We make the following assumptions: Assumption A1 is made since we are describing the local image appearance space. Assumptions A2 and A3 are valid in most practical situations.

A1. Δt is small, which implies that the rigid motion and deformation between t_1 and t_2 are small. Illumination change can be arbitrary.

A2. Illumination is represented by a finite dimensional linear orthogonal basis.

A3. $\mathcal{C}(u, v)$ is smooth and the deformation is smooth, allowing $\frac{\partial^2\mathcal{C}}{\partial u\partial t} = \frac{\partial^2\mathcal{C}}{\partial t\partial u}$ and $\frac{\partial^2\mathcal{C}}{\partial v\partial t} = \frac{\partial^2\mathcal{C}}{\partial t\partial v}$, and albedo ρ is spatially smooth.

We prove that under suitable situations, the image space of a moving and deforming object under varying illumination is locally multilinear. For ease of explanation, we start from a fixed rigid object under varying illumination. Then, we consider the problem of a moving rigid object under varying illumination. These are overviews of existing work. Next, we consider the main case of interest in this paper, a moving and deforming object under varying illumination (**Theorem 1**). Then, starting from this theorem, we derive two Corollaries of a fixed deforming object under varying illumination (**Corollary 1**) and a moving and deforming object under fixed illumination (**Corollary 2**).

2.2 Fixed Rigid Object under Varying Illumination (Review of [1])

In [1], [20], the authors showed that when a rigid object is fixed with respect to the camera, the reflectance image $\mathcal{I}$ of size $P \times Q$ pixels can be represented as

$$\mathcal{I} = \mathcal{B}_l(\mathcal{M}) \times_l \mathbf{l}, \tag{1}$$

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, $\mathcal{M}$ is the 3D model of the object, including both the 3D shape and texture, and $\times_l$ is the *mode-n product* [14] along the illumination dimension.[1] It has been shown in [1] that for a Lambertian surface object with attached shadows, more than 99.22 percent of the energy can be captured by the first nine bases when spherical harmonics functions are used, i.e., $N_l \approx 9$. When the Lambertian reflectance property is not satisfied, higher orders of the spherical harmonics functions will be needed [24].

2.3 Moving Rigid Object under Varying Illumination (Review of [31])

Under assumptions A1 and A2, the authors in [31] proved that the image space can be approximated by a bilinear function of the illumination and rigid motion parameters, i.e.,

$$\mathcal{I} = (\mathcal{B}_l + \mathcal{B}_{ml} \times_m \mathbf{m}) \times_l \mathbf{l}, \tag{2}$$

where $\mathcal{B}_{ml} \in \mathbb{R}^{9\times 6\times P\times Q}$ is the tensor version of the motion bases, $\mathbf{m} = (\Delta\mathbf{T}^{\mathbf{T}}, \Delta\mathbf{\Omega}^{\mathbf{T}})^{\mathbf{T}}$ is the motion parameter vector, where $\Delta\mathbf{T}$ is the translation of the centroid of the object and $\Delta\mathbf{\Omega}$ is the rotation about the centroid of the object. The exact forms of b_i, $\mathcal{B}_l$, and $\mathcal{B}_m$ can be found in [1], [31]. Although this theory incorporates motion into the framework, it requires the object to be rigid.

2.4 Mathematical Model of Deformation

Consider that the pose of the object is fixed with respect to the camera, but that it is deforming. The surface of the object is a function of time, i.e., $\mathcal{C}(u,v,t) : \mathbb{R}^2 \times [0,T) \to \mathbb{R}^3$. It can be shown that under proper parameterization, nonrigid deformation can be modeled such that each point on the surface is evolving only along its surface normal direction, with an amount $\beta(u,v,t)$ defined at this point, i.e,

$$\frac{\partial\mathcal{C}(u,v,t)}{\partial t} = \beta(u,v,t)\mathcal{N}(u,v,t), \tag{3}$$

where $\mathcal{N}(u,v,t)$ is the surface normal at $\mathcal{C}(u,v,t)$. The derivation of this model can be found in Section 2.1 of [22]. Thus, given the parameterization (u,v), a deformation of the object can be identified via $\beta(u,v,t)$. Although other models may also work, this model is chosen for its simplicity in describing the nonrigid deformation. 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. In (A3), we stated that the deformation is smooth, which means the function $\beta(u,v,t)$ (which describes the amount of shape change at each point (u,v)) is smooth. Thus, most of the energy of $\beta(u,v,t)$ at time instance t would be concentrated in the low frequency components, and can be decomposed using the top N_D bases as

$$\beta(u,v,t) = \Phi_d(u,v) \times_d \mathbf{b}_d(t), \tag{4}$$

where $\Phi_d \in \mathbb{R}^{N_D\times 1}$ is the vector of the top N_D basis at (u,v), $\mathbf{b}_d \in \mathbb{R}^{N_D\times 1}$ encrypts the deformation at (u,v) as a function of t, and $\times_d$ indicates the tensor product along the deformation dimension.

In (A3), we assumed the texture to be spatially smooth. Using the same parameterization, the texture function on the surface can be decomposed using top N_ρ bases as

$$\rho(u,v,t) = \Phi_\rho(u,v) \times_\rho \mathbf{b}_\rho(t), \tag{5}$$

where $\mathbf{b}_\rho \in \mathbb{R}^{N_\rho\times 1}$ and $\Phi_\rho \in \mathbb{R}^{N_\rho\times 1}$. Similarly to the above notation, $\times_\rho$ indicates the tensor product along the texture dimension. It is important to note a difference between (4) and (5). In (4), β is the rate of change of the surface curvature (shape), while (5) describes the change of albedo ρ itself. Thus, for ρ, we do not require the temporal change of it to be smooth (see assumption (A3)).

2.5 Main Results

Theorem 1. *The image space of a rigidly moving and deforming object under varying illumination is locally multilinear, with the illumination subspace and the texture subspace being trilinearly combined with the direct sum of the motion and deformation subspaces, i.e.,*

$$\mathcal{I}_t = \mathcal{B}_{l\rho md} \times_l \mathbf{l}(t) \times_\rho \mathbf{b}_\rho(t) \times_m \begin{pmatrix} \mathbf{V}\Delta t \\ \omega\Delta t \\ \mathbf{b}_d\Delta t \\ 1 \end{pmatrix}, \tag{6}$$

where $\mathcal{B}_{l\rho dm} \in \mathbb{R}^{N_l\times N_\rho\times(6+N_D+1)\times P\times Q}$ is the tensor version of the joint illumination, texture, rigid motion, and deformation bases.

The proof of the theorem and the detailed notations are given in the supplemental material, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2010.216. The theorem proves that the image space of a moving and deforming object under varying illumination is a locally trilinear function of the illumination, deformation, motion, and texture change parameters. The result uses the facts that the illumination space is known to be low-dimensional and a low-dimensional representation of the texture space is sufficient for nonvisualization applications. The result is valid in a local region around pose $(\mathbf{R}, \mathbf{T})$. The locality property comes because of the small time interval assumption in A1.

The theorem stated above describes a very general condition, i.e., all of the factors of the illumination, motion, deformation, and texture are changing. When only a few of the parameters are changing while the others are fixed, we get a few special conditions. When the pose, object shape, and texture are fixed and only illumination changes, by setting the $\mathbf{V}$, ω, $\mathbf{b}_d$ to be zero and $\mathbf{b}_\rho(t)$ to be a constant we get a linear subspace in the same form of (1). When the deformation and texture are fixed while both pose and illumination could change, then (6) degenerates into (2).

When the pose of the object is fixed and the shape, texture, and illumination can change (i.e., deforming object at fixed pose), by setting the $\mathbf{V}$, ω parameters in (6) to be zero we have the following:

Corollary 1. *Under Assumptions A1, A2, and A3, the image space of a fixed deforming object under varying illumination is locally trilinear in the illumination, deformation, and texture parameters, i.e.,*

$$\mathcal{I} = (\mathcal{B}_{\rho l} + \mathcal{B}_{d\rho l} \times_d \mathbf{b}_d\Delta t) \times_\rho \mathbf{b}_\rho \times_l \mathbf{l}, \tag{7}$$

where $\mathcal{B}_{d\rho l} \in \mathbb{R}^{N_D\times N_\rho\times N_l\times P\times Q}$ is the tensor version of the deformation, texture, and illumination basis, and $\mathcal{B}_{\rho l} \in \mathbb{R}^{1\times N_\rho\times N_l\times P\times Q}$ is the tensor version of the texture and illumination basis.

Similarly, when the illumination is fixed while the pose, shape, and texture of the object could change (i.e., rigid motion and deformation with fixed illumination), we can get the corollary below.

Corollary 2. *Under Assumptions A1, A2, and A3, the image space of a rigidly moving and deforming object under fixed illumination is a locally bilinear, with the texture subspace being bilinearly combined with the direct sum of the motion and deformation subspaces, i.e.,*

$$\mathcal{I} = \left(\mathcal{G}_\rho + \mathcal{G}_{md\rho} \times_m \begin{pmatrix} \mathbf{V} \\ \omega \\ \mathbf{b}_d \end{pmatrix} \Delta t\right) \times_\rho \mathbf{b}_\rho, \tag{8}$$

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

where $\mathcal{G}_\rho = \mathcal{B}_{\rho l} \times_l \mathbf{l}$ *and* $\mathcal{G}_{md\rho} = \mathcal{B}_{md\rho l} \times_l \mathbf{l}$. $\mathcal{B}_{\rho l} \in \mathbb{R}^{N_\rho \times N_l \times P \times Q}$ *is the tensor version of the texture and illumination basis, and* $\mathcal{B}_{md\rho l} \in \mathbb{R}^{6 \times N_D \times N_\rho \times N_l \times P \times Q}$ *is the tensor version of the joint rigid motion, deformation, texture, and illumination basis.*

2.6 Discussion of the Theoretical Results

The result in (6) implies that the illumination and texture subspaces are trilinearly combined with the union of the rigid motion and deformation subspaces. The result in (6) has two major contributions: First, it provides a physics-based, analytical representation of the multilinear bases for representing the image appearance space, while all previous methods have relied on learning such bases from data. Second, it shows that under a set of assumptions that often hold, it is possible to approximate the image appearance space to be multilinear (which can be simplified in special cases).

We used three assumptions for deriving Theorem 1 and Corollaries 1 and 2. Assumption A1 is reasonable for most video sequences captured under frame rates between 15 and 30 fps, and can be used to validate the theoretical model. Assumption A2 essentially says that we use a basis illumination model. This is widely used. For Lambertian surfaces, the dimension is small, while a 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 reasonable for many objects and has been widely used for modeling deformation [22].

3 Modeling the Face Image Space

When confined to face images of a single person, the variations of the texture and shape are usually small while the change due to illumination may still be drastic. Thus, from theorem, by expanding the $\mathbf{b}_\rho$ around the mean face texture coefficient the image space of faces becomes bilinear with the illumination being bilinearly combined with direct sum of the motion, deformation, and texture parameters, i.e.,

$$\begin{aligned}\mathcal{I}_t &= \left(\mathcal{B}_{l\bar{\rho}md} + \mathcal{B}_{l\rho md} \times_{m\rho} \begin{pmatrix} \frac{\partial \mathbf{b}_\rho}{\partial t} \\ \mathbf{V} \\ \omega \\ \mathbf{b}_d \end{pmatrix} \Delta t \right) \times_l \mathbf{l}(t), \text{where } \mathcal{B}_{l\bar{\rho}md} \\ &= \mathcal{B}_{l\rho md} \times_\rho \mathbf{b}_{\bar{\rho}}. \end{aligned} \tag{9}$$

$\mathbf{b}_{\bar{\rho}}$ is the mean face texture coefficient. Thus, (9) models face appearance locally around the neutral mean shape and mean texture of faces at the cardinal poses $\mathbf{p}_j$, while modeling globally along the illumination dimension. This was the result we derived in [32], and is a special case of Theorem 1.

To construct the image space representing all possible pose and deformations, we divide the pose in pan and tilt directions uniformly into a set of regions, each region being identified with a cardinal point in pose and deformation space. The effects of 3D translation on image plane can be removed by centering and scale normalization, while in-plane rotation to a predefined pose can mitigate the effects of rotation about the z-axis. Thus the image of an object under arbitrary pose, $\mathbf{p}$, can always be described by the multilinear object representation at a predefined $(\mathbf{T}_x^{pd}, \mathbf{T}_y^{pd}, \mathbf{T}_z^{pd}, \mathbf{\Omega}_z^{pd},)$, with only $\mathbf{\Omega}_x$ and $\mathbf{\Omega}_y$ depending upon the particular pose. Thus, the image manifold under any pose can be approximated by the collection of a few tangent planes on distinct $\mathbf{\Omega}_x^j$ and $\mathbf{\Omega}_y^j$, denoted as $\mathbf{p}_j$.

Although the result in (9) is locally multilinear along the pose dimension, in the supplemental material, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2010.216, we show that this piecewise locally multilinear manifold can be embedded into a higher dimensional globally multilinear subspace of much higher dimension (as was used in [29]).

3.1 Relation to Existing Methods

This theoretical study provides an understanding of the applicability of many linear/multilinear models of object appearance/shape representation used commonly in computer vision. We can also understand the conditions under which these popular models can be applied. Below we provide such an analysis, taking face representation and recognition as an example (since all of the models have been applied to faces).

PCA. From (9) we can see that when the illumination and pose are fixed, the image space is linear in the shape 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 face recognition under fixed pose and illumination and poor performance when illumination is changing.

AAM/ASM. AAM/ASM [5] are 2D approaches that represent shape and appearance using two separate linear sets of basis vectors. Thus, a warping of the texture is needed to combine the shape and texture together. With this warping, the image becomes a nonlinear function of the shape, texture, and the pose parameters. In our approach, which is 3D-based, the texture is inherently coupled with the deformation model, and results in the multilinear formula in (6).

MLM. In MLM [26], [29], different factors (illumination, pose, identity) are assumed to be globally multilinearly combined. We show that lighting and texture are indeed trilinearly combined with the direct sum of the motion and deformation subspaces. Since this multilinearity property is local, MLM methods will be more efficient and accurate when modeling local regions of the image space. However, from the supplemental material, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2010.216, we see that a global MLM is also valid if we use higher dimensions.

Local linearization. Probabilistic Appearance Model (PAM) [16] uses a series of tangent planes along pose to approximate the manifold; thus, it is also locally linear. Our theoretical result provides an analytical description of this space. In [34], 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. The same reasoning is valid for locally linear models like [21], [25].

Nonlinear 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 detailed representation, but comes at the cost of higher computation due to optimization on a nonlinear manifold. Nonlinear manifold is also the approach taken in [15]. In a rough sense, we can say that the proposed approach is somewhere between the 3D morphable model and AAM. We are able to decouple the illumination, 3D motion, deformation, and texture variations, while not requiring the availability of large training data sets like both morphable models and AAM/ASMs. However, it will probably not provide high fidelity reconstructions like 3DMM.

4 Application in Face Tracking

The locally multilinear image appearance model derived in Section 2 provides us with a method for tracking deforming objects under deformation and varying illumination conditions. We start by

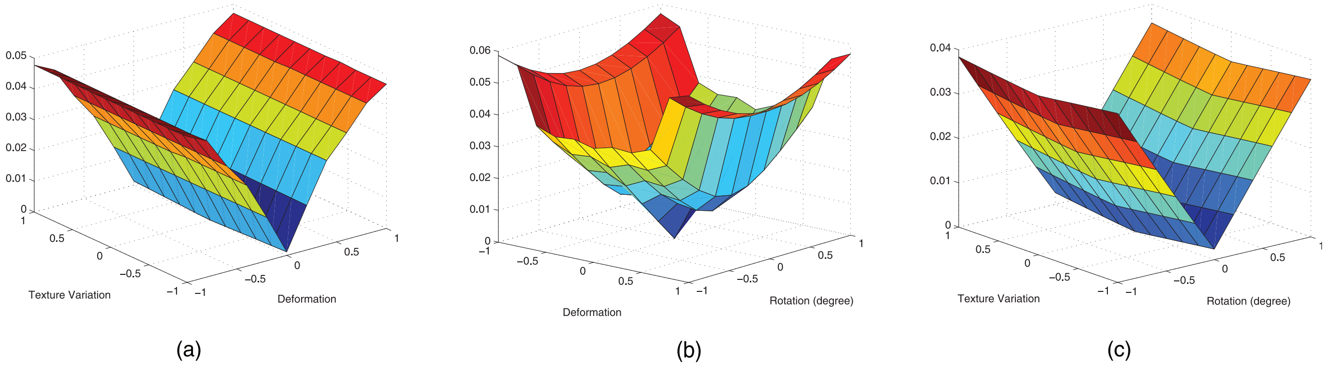


Fig. 2. Accuracy analysis of the theoretical model. The error is computed as the squared difference between the theoretically predicted pixel intensities and the true pixel intensities, normalized by the true values, and taking its mean over the face region.

registering a generic 3D face model to the first image of the video sequence. For most of the video sequences, the time interval between two consecutive frames is small; thus, assumption A1 is valid. Letting $\mathbf{m} = (\Delta\mathbf{T}, \Delta\mathbf{\Omega})^{\mathbf{T}}$ represent the rigid 3D motion parameter, we can estimate $\mathbf{m}$ and other parameters as

$$(\hat{\mathbf{l}}, \hat{\mathbf{m}}, \hat{\mathbf{b}}_d, \hat{\mathbf{b}}_\rho) = \arg \min_{\mathbf{l},\mathbf{m},\mathbf{b}_d,\mathbf{b}_\rho} \left\| \mathcal{I} - \mathcal{B}_{l\rho md} \times_l \mathbf{l} \times_\rho \mathbf{b}_\rho \times_m \begin{pmatrix} \mathbf{m} \\ \mathbf{b}_d \Delta t \\ 1 \end{pmatrix} \right\|^2 + \alpha_m \|\mathbf{m}\|^2 + \alpha_d \|\mathbf{b}_d\|^2, \tag{10}$$

where $\hat{x}$ denotes an estimate of x. Since the motion and the deformation between consecutive frames are small, we add regularization terms to the above cost function in the form of $\alpha_m \|\mathbf{m}\|^2$ and $\alpha_d \|\mathbf{b}_d\|^2$. The L2 norm over the image tensor $\mathcal{I}$ is taken over the pixels within the object region, which have nonzero values in the image appearance bases. Thus, the proposed method is robust to clutter in the background. Since the image $\mathcal{I}$ lies approximately in a locally multilinear space of illumination, rigid motion, deformation, and texture variables, such a minimization problem can be achieved by alternately minimizing over each parameter.

Although the sequence of parameter estimation can be altered, the speed of convergence will be affected. According to the assumptions (A1, A3), the time interval between the consecutive frames should be small; thus, the rigid motion, shape deformation, or albedo (texture)[2] change relatively slowly along time t when compared with lighting $\mathbf{l}$. Due to this reason, we first estimate the illumination parameter $\mathbf{l}$, then estimate other parameters using the estimated $\hat{\mathbf{l}}_t$ to make the convergence faster. When illumination changes gradually, the sequence of estimation can be altered without much loss of convergence speed.

5 Experimental Results

5.1 Numerical Accuracy Analysis

To evaluate the theory quantitatively, we performed a numerical error 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 1 second, the deformation will not exceed 5 percent 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 percent of the square of the mean value of the original texture. For the rotation, we assume the maximum degree the object can rotate in 1 second is 30 degrees, which means 1 degree between two consecutive frames.

In Fig. 2, we plot the normalized error versus (Fig. 2a) deformation and texture variation, (Fig. 2b) deformation and rigid motion, and (Fig. 2c) 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. 2 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 percent. This is the worst-case performance and happens when the object is deforming and rotating.

5.2 Application to Tracking

As an application of the theory, we use it for tracking faces under illumination and expression variations (Figs. 3 and 4). We use the face image space of Section 4.[3] We use the method described in Section 4 for estimating the pose, illumination, and deformation parameters.[4] In the first row of Fig. 3, we show the tracking of a rigid face under varying pose and illuminations, while in the first row of Fig. 4, we show the tracking of a face with changing expressions. The 2D locations of the face are shown and the pose parameters are represented as the Euler angle of the face with respect to the frontal one, following the "z-x-z" convention [13]. In Figs. 3b, 3c, 4b, and 4c, we show the norm of the estimated illumination parameters and the estimated 3D pose in Euler angles, respectively. The key frames shown in the first rows are marked using the dashed lines.

5.2.1 Accuracy of the Estimates

Since we do not have access to ground truth, the estimation accuracy can be judged by comparing against the original images and the back projection of feature points from the 3D model to the image plane. The norm of the illumination coefficient shows the intensity of the illumination, and the shift in Fig. 3b around the 160th frame is due to the sudden switching on and off of the local light source. This correlates with the images. Such a change is not

[2] We use the terms "texture" and albedo interchangeably.

[3] To construct the multilinear bases, we first define a local coordinate on the object surface (u, v) and choose 2D transformation bases defined on this local coordinate system (in our experiment, we use 2D DCT bases), which gives the $\phi_d(u, v)$. Then, this $\phi_d(u, v)$ is then substituted into (6) and (11) in the supplemental material. Substituting (6) and (11) into (9) in the supplemental material, the multilinear bases can be computed.

[4] For modeling deformation, we use 144 2D DCT bases (12 by 12); it is possible to to use more sophisticated bases to compact the facial deformation and texture variation into a much lower order. As a 3D point cloud model is used, there are in total about 30,000 points on the model. The tracking is performed by using only the pixels from the regions around the 2D projections of the 3D mesh. The 3D model is registered to the first frame for computing the tensor bases at the initial position.

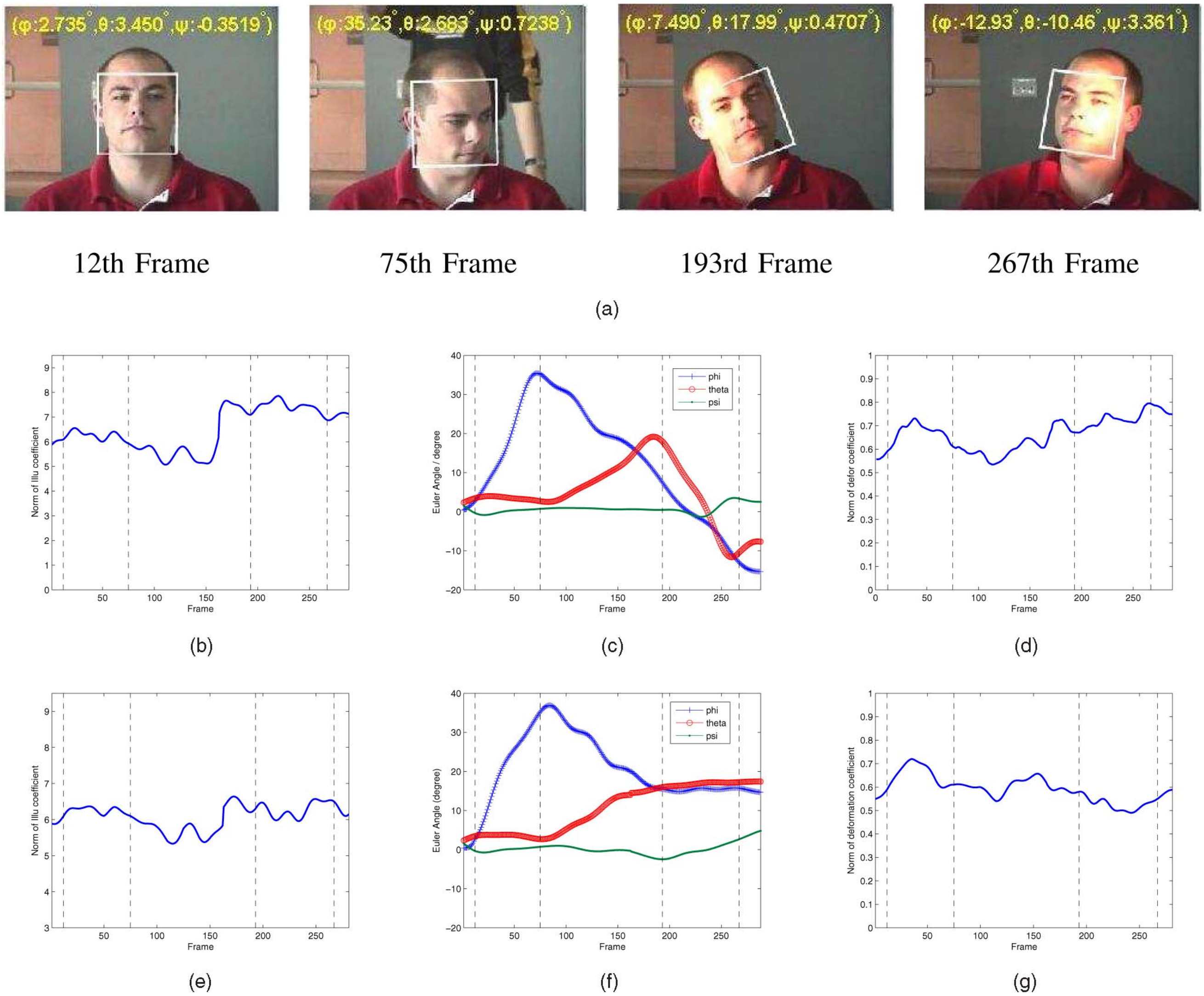


Fig. 3. One example of tracking using the theoretical model on real data under changes of pose and lighting. In (a), some key frames with the 2D locations and 3D Euler angles are shown. In (b), we show the norm of the estimated illumination coefficients as a function of time, (c) shows the estimated 3D pose represented with Euler angle, while (d) shows the norm of the estimated deformation parameter $\mathbf{b}_d(t)$ during the same period of time. The key frames shown in the first row are marked on the plots in (b), (c), and (d) with dotted lines. In (e), (f), and (g), we show the norm of the estimated illumination coefficients, 3D pose in Euler angle, and the norm of the estimated deformation parameter using the Linear model discussed in Section 5.3.

seen in the estimates using the linear model described below. In Fig. 4b, we see that the illumination change is more significant and this is in accordance with the images shown in Fig. 4a. In Figs. 3d and 4d, we plot the norm of the estimated deformation parameter $\mathbf{b}_d(t)$ in the video sequence shown in the first row of Figs. 3 and 4, respectively. The maximum of the norm of $\mathbf{b}_d(t)$ within the sequence is normalized to be 1. The larger the norm is, the larger the expression is from the mean face. In Fig. 4d, the large magnitude of $\mathbf{b}_d$ from about the 50th frame to the 125th frame corresponds to the yawn, while the plateau after 200th frame corresponds to the frown, as shown in the key frames. In Fig. 3d, the magnitude of $\mathbf{b}_d(t)$ has only small variations, in accordance with the steady expression on the face. These results show that we can decouple and estimate the 3D pose, illumination, and deformation parameters given a video sequence.

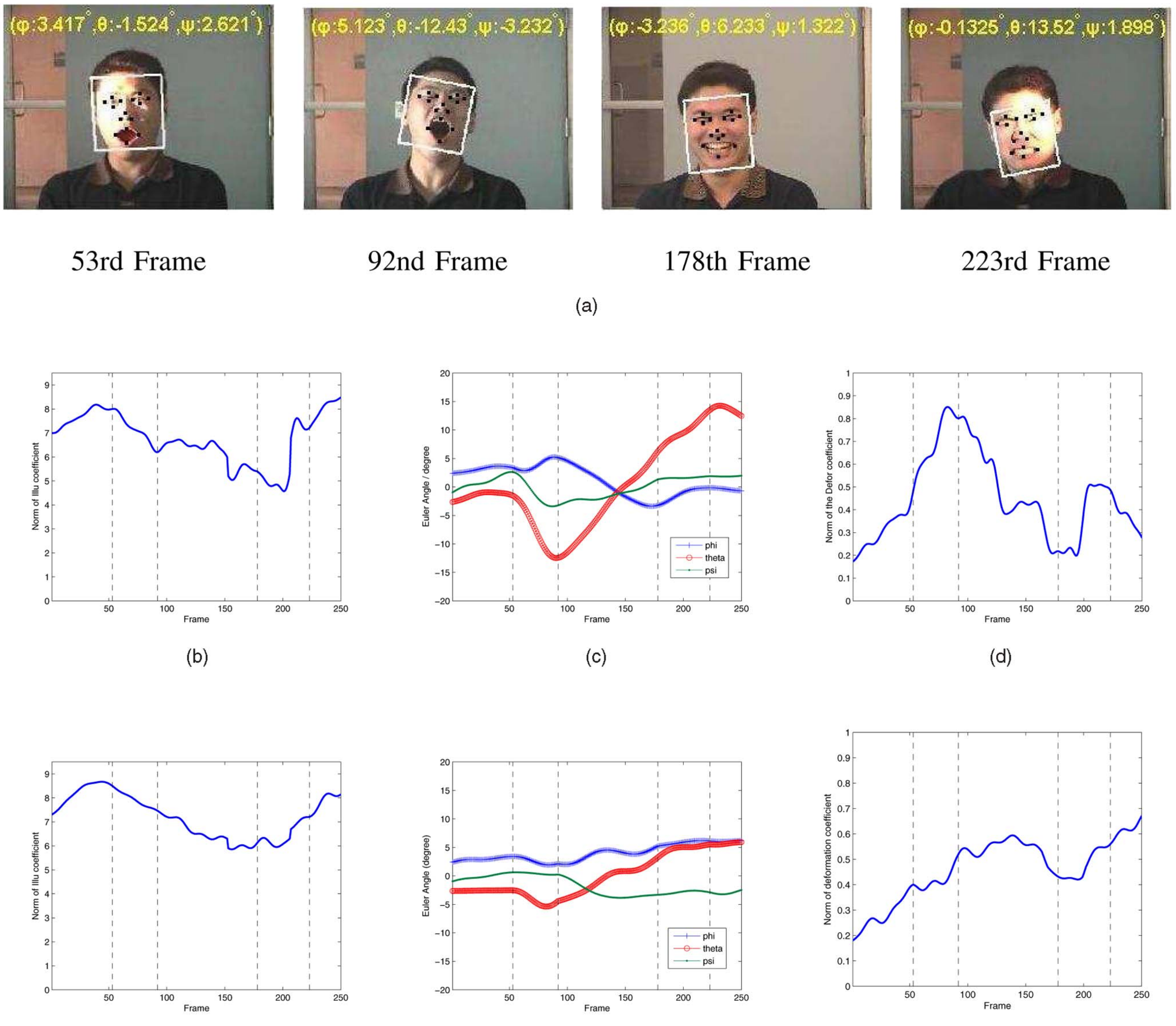
5.3 Comparison Against Linear Model

A linear model is mostly used for describing the image space due to a single variation factor, like identity, illumination, pose, etc. Mostly such models are obtained by applying machine learning techniques onto a collection of training data. Such a model is a poor fit for describing the image space due to multiple factors (e.g., pose, lighting, and deformation) as it mixes different factors together and lacks a clear physical interpretation of the basis functions. We compare the proposed model in (6) against a linear model to show that the multilinear model is indeed more accurate and robust than a simple linear model.

From (6), by rewriting $\mathbf{l}(t)$ and $\mathbf{b}_\rho(t)$ in incremental forms of $\mathbf{l}_{t_1} = \mathbf{l}_{t_0} + \Delta\mathbf{l}$ and $\mathbf{b}_{\rho t_1} = \mathbf{b}_{\rho t_0} + \Delta\mathbf{b}_\rho$, we can easily obtain an analytical linear expansion of the image space in terms of the change of the illumination, pose, deformation, and texture variation as

$$\mathcal{I} = \mathcal{B}_{l\rho} \times_l \Delta\mathbf{l} \times_\rho \mathbf{b}_{\rho t_0} + \mathcal{B}_{l\rho} \times_l \mathbf{l}_{t_0} \times_\rho \Delta\mathbf{b}_\rho + \mathcal{B}_{l\rho md} \times_l \mathbf{l}_{t_0} \times_\rho \mathbf{b}_{\rho t_0} \times_m \begin{pmatrix} \mathbf{V}_{t_0}\Delta t \\ \omega_{t_0}\Delta t \\ \mathbf{b}_{dt_0}\Delta t \\ 1 \end{pmatrix}. \tag{11}$$

Thus, we obtain a tangent plane of the image appearance manifold in terms of the change of the parameters at the specific illumination $\mathbf{l}_{t_0}$, texture ρ_{t_0}, shape $\mathbf{b}_{dt_0}$, and pose $(\mathbf{V}_{t_0}, \omega_{t_0})$, which is

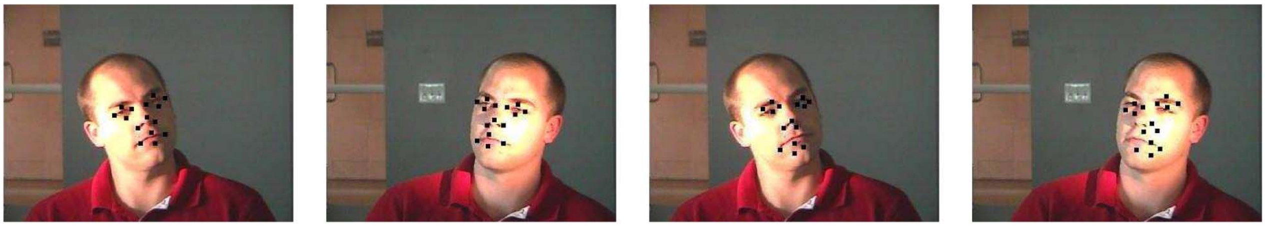


(a)

(b) (c) (d)

(e) (f) (g)

Fig. 4. Another example of tracking using the theoretical model on real data under changes of pose, lighting, and expressions. In (a), some key frames with the 2D locations and 3D Euler angles are shown, (b) shows the norm of the estimated illumination coefficients as a function of time, (c) shows the estimated 3D pose represented with Euler angle, and (d) shows the norm of the estimated deformation parameter $\mathbf{b}_d(t)$ during the same period of time. The larger the norm of $\mathbf{b}_d(t)$ is, the larger the expression deformation is. The key frames shown in the first row are marked on the plots in (b), (c), and (d) with dotted lines. In (e), (f), and (g), we show the norm of the estimated illumination coefficients, 3D pose in Euler angle, and the norm of the estimated deformation parameter using the Linear model discussed in Section 5.3.



(a) (b) (c) (d)

Fig. 5. The back projection of the the feature points from the 3D face model onto the image plane using the estimated pose parameters for the 193rd and 267th frames shown in Fig. 3. (a) and (b) are obtained with the pose estimation results using the multilinear mode in (6), while (c) and (d) are obtained with the ones using the linear model in (11).

the point at which the tensor bases $\mathcal{B}_{l\rho md}$ is computed. To evaluate the accuracy of such a model against the proposed one in (6), we apply it to track the same sequences the multilinear model has been applied to, and show the estimated illumination, pose, and deformation parameters in Figs. 3e, 3f, 3g, 4e, 4f, and 4g. In Figs. 3e, 3f, and 3g, we can see clearly that up to the 160th frame, in which period only the pose changes, the linear model is able to track the sequence. However, after the 160th frame, where both illumination and pose change, the linear model begins to lose track. Even the change in the illumination is not captured clearly by the linear model. The loss of track can be observed more obviously in Fig. 5, in which we compare the back projection of the feature points on the 3D face model onto the image plane using the estimated pose parameters obtained with the multilinear model in Fig. 5a and Fig. 5b against the ones using the linear model in Fig. 5c and Fig. 5d. In Figs. 4e, 4f, and 4g, the linear model begins to lose track very soon after the 50th frame due to the fact that all of the parameters are changing simultaneously and cannot be decoupled from each other.

6 Conclusions

In this paper, we analyzed the accuracy of linear and multilinear object representation models from the fundamental physical laws of object motion and image formation. We proved that the image appearance space is multilinear, with the illumination and texture subspaces being trilinearly combined with the direct sum of the motion and deformation subspaces. Using this result, we discussed the applicability of many of the linear and multilinear approaches existing in the computer vision literature, including PCA, AAM/ASM, and MLM, locally linear models, and 3DMM. We provided the application of this proposed model in illumination invariant deformable object tracking. Experimental accuracy analysis of the theoretical results was also presented. Future work will focus on the application of this result in object recognition.

Acknowledgments

The work was partly supported under US National Science Foundation (NSF) grant IIS 0712253.

References

[1] R. Basri and D. Jacobs, "Lambertian Reflectance and Linear Subspaces," *IEEE Trans. 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?" *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 1996.

[3] 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.

[4] T. Broida and R. Chellappa, "Estimating the Kinematics and Structure of a Rigid Object from a Sequence of Monocular Images," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 13, no. 6, pp. 497-513, June 1991.

[5] T. Cootes, G. Edwards, and C. Taylor, "Active Appearance Models," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 23, no. 6, pp. 681-685, June 2001.

[6] R. Costantini, L. Sbaiz, and S. Ssstrunk, "Higher Order SVD Analysis for Dynamic Texture Synthesis," *IEEE Trans. Image Processing,* vol. 17, no. 1, pp. 42-52, Jan. 2008.

[7] G. Doretto, A. Chiuso, Y. Wu, and S. Soatto, "Dynamic Textures," *Int'l J. Computer Vision,* vol. 51, no. 2, pp. 91-109, 2003.

[8] G. Doretto and S. Soatto, "Dynamic Shape and Appearance Models," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 28, no. 12, pp. 2006-2019, Dec. 2006.

[9] R. 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.

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

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

[12] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active Contour Models," *Int'l J. Computer Vision,* pp. 321-331, 1988.

[13] L.D. Landau and E.M. Lifschitz, *Mechanics,* third ed., Pergamon Press, 1976.

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

[15] C. Lee and A. Elgammal, "Nonlinear Shape and Appearance Models for Facial Expression Analysis and Synthesis," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* vol. I, pp. 313-320, 2003.

[16] 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,* vol. I, pp. 313-320, 2003.

[17] I. Matthews and S. Baker, "Active Appearance Models Revisited," *Int'l J. Computer Vision,* vol. 60, no. 2, pp. 135-164, Nov. 2004.

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

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

[20] 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. of Am.,* vol. 18, no. 10, Oct. 2001.

[21] S. Roweis and L. Saul, "Nonlinear Dimensionality Reduction by Locally Linear Embedding," *Science,* vol. 290, no. 5500, pp. 2323-2326, Dec. 2000.

[22] G. Sapiro, *Geometric Partial Differential Equations and Image Analysis.* Cambridge Univ. Press, 2001.

[23] 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.

[24] S. Shirdhonkar and D. Jacobs, "Non-Negative Lighting and Specular Object Recognition," *Proc. 10th IEEE Int'l Conf. Computer Vision,* vol. I, pp. 1323-1330, Oct. 2005.

[25] J.B. Tenenbaum, V. de Silva, and J.C. Langford, "A Global Geometric Framework for Nonlinear Dimensionality Reduction," *Science,* vol. 290, no. 5500, pp. 2319-2323, Dec. 2000.

[26] J.B. Tenenbaum and W.T. Freeman, "Separating Style and Content with Bilinear Models," *Neural Computation,* vol. 12, no. 6, pp. 1247-1283, 2000.

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

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

[29] M. Vasilescu and D. Terzopoulos, "Multilinear Independent Components Analysis," *IEEE CS Conf. Computer Vision and Pattern Recognition,* June 2005.

[30] A. Veeraraghavan, A. Roy-Chowdhury, and R. Chellappa, "Matching Shape Sequences in Video with Applications in Human Motion Analysis," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 27, no. 12, pp. 1896-1909, Dec. 2005.

[31] Y. Xu and A. Roy-Chowdhury, "Integrating Motion, Illumination and Structure in Video Sequences, with Applications in Illumination-Invariant Tracking," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 29, no. 5, pp. 793-806, May 2007.

[32] Y. Xu and A. Roy-Chowdhury, "A Theoretical Analysis of Linear and Multilinear Moels of Image Appearance," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition,* 2008.

[33] Y. Xu and A. Roy-Chowdhury, "Learning a Geometry Integrated Image Appearance Manifold from a Small Training Set," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition,* 2008.

[34] H. Yang, M. Pollefeys, G. Welch, J.-M. Frahm, and A. Ilie, "Differential Camera Tracking through Linearizing the Local Appearance Manifold," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition,* 2007.

[35] L. Zhang, B. Curless, A. Hertzmann, and S. Seitz, "Shape and Motion under Varying Illumination: Unifying Structure from Motion, Photometric Stereo, and Multiview Stereo," *Proc. IEEE Int'l Conf. Computer Vision,* 2003.