MODELING TIME-VARYING ILLUMINATION PATTERNS IN VIDEO

Yilei Xu, Amit K. Roy-Chowdhury

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

ABSTRACT

Recreating the temporal illumination variations of natural scenes has great potential for realistic synthesis of video sequences. In this paper, we present a 3D (model-based) approach that achieves this goal. The approach requires a training sequence to learn the time-varying illumination models, which can then be used for synthesis in another sequence. The motion and illumination parameters in the training sequence are estimated alternately by projecting onto appropriate basis functions of a bilinear space defined in terms of the 3D surface normals of the objects. The motion is represented in terms of 3D translation and rotation of the object centroid in the camera frame, and the illumination is represented using a spherical harmonics linear basis. We show video synthesis results using the proposed approach.

Index Terms**—** Illumination, motion, 3D model, video sequence

1. INTRODUCTION

Determination of the illumination conditions in a video sequence as a function of time has important applications in object recognition, video summarization and video synthesis. While many of the existing methods for estimating motion and shape of an object can handle significant changes in the illumination conditions by *compensating* for the variations, there do not exist many methods that can *recover* the *time-varying* illumination conditions from *video* sequences of moving objects. In this paper, we propose a 3D approach for learning the illumination conditions of video sequences. The recovered parameters are then used to synthesize new videos under the lighting conditions of the original ones.

Most of the advanced methods for modeling lighting have concentrated on the study of single images, e.g. shape from shading , photometric stereo , illumination cone . In one of the most important results on illumination modeling, Basri and Jacobs [1] and Ramamoorthi and Hanrahan [2], independently derived a 9D spherical harmonics based linear representation of the images produced by a Lambertian object with attached shadows. However, there has been little work on integrating the advances in illumination modeling with methods for shape and motion estimation. Some exceptions are [3, 4, 5]. Integrating illumination models with motion and shape would allow us to compute all the three together. This is different from applying image-based approaches like [1, 2] to every frame of a video sequence separately, becasue the image-based approaches would require estimating the pose of the object as it moves, which, in turn, would be made difficult by the fact that lighting is changing. Integrating motion and shape with the illumination models would overcome this problem by providing a representation of the image appearance in terms of all these three parameters.

In this paper, we propose a 3D approach for estimation of the time-varying illumination conditions from video sequences, while simultaneously tracking the objects in the video. It allows tracking, and hence illumination estimation, under large changes of pose. The method is built upon the approach in [4], where the authors showed that the set of all Lambertian reflectance functions of a moving object, at any position, illuminated by arbitrarily distant light sources, lies close to a bilinear subspace consisting of nine illumination variables and six motion variables. The lighting can consist of combinations of point and extended sources, and can change slowly or suddenly. This allows us to learn the illumination variations of natural scenes in indoor and outdoor environments. We demonstrate the ability of this method in learning the illumination variations in videos of natural scenes and synthesis of new sequences.

The rest of the paper is organized as follows. Section 2 presents an overview of the joint illumination and motion models for video sequences followed by the algorithm for learning the motion and illumination parameters from video using this model. In Section 3, video synthesis results with two approaches are presented. Section 4 concludes the paper and highlights future work.

2. INTEGRATING ILLUMINATION, 3D MOTION AND SHAPE MODELS IN VIDEO

2.1. Bilinear Model of the Motion and Illumination

In this section, we present the fundamental result on estimating the illumination and 3D motion parameters. Recent studies has shown that, for a fixed Lambertian object, the set of reflectance images *under distant lighting without cast shadows*

This work was partially supported by NSF grants 0551719 and 0622176.

1-4244-1437-7/07/$20.00 ©2007 IEEE

can be approximated by a linear combination of nine basis images, defined using spherical harmonics [1, 2]. Several papers have shown the suitability of this model for faces [1, 6].

In [4], the motion was taken into the consideration and it was shown that for moving objects it is possible to approximate the sequence of images by a bilinear subspace. It was proved that if the motion of the object from time t_1 to new time instance t_2 is small, then upto a first order approximation, the reflectance image $I(x, y)$ at t_2 can be expressed as [4]:

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

where

$$b_{ij}(\mathbf{n}_{\mathbf{P}'_2}) = b_{ij}(\mathbf{n}_{\mathbf{P_1}}) + \mathbf{AT} + \mathbf{B\Omega}. \tag{2}$$

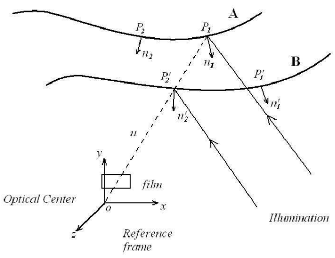


Fig. 1. Pictorial representation showing the motion of the object and its projection (reproduced from [4]).

In the above equations, $b_{ij}(\mathbf{n}_{\mathbf{P}'_2})$ and $l_{ij}^{t_2}$ are the basis images and illumination coefficients after motion, while $b_{ij}(\mathbf{n}_{\mathbf{P_1}})$ are the original basis images before motion. $\mathbf{A}$ and $\mathbf{B}$ contain the structure and camera intrinsic parameters, and are functions of pixel index (x, y). For one pixel (x, y), both $\mathbf{A}$ and $\mathbf{B}$ are N_l by 3 matrices, where $N_l \approx 9$ for Lambertian objects with attached shadow. Please refer to [4] for the derivation of (1,2) and explicit expression for $\mathbf{A}$ and $\mathbf{B}$. Substituting (2) into (1), we see that the new image spans a bilinear space of six motion and approximately nine illumination variables (for Lambertian objects with attached shadows). The basic result is valid for general illumination conditions, but require consideration of higher order spherical harmonics. In can also be used with other basis functions (e.g., wavelets [7, 8]) so long as the change of the bases is approximately linear in the 3D rigid motion terms.

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

$$\begin{aligned} I(x,y,t_2) &= I(x,y,t_1) + \sum_{i=0}^{2} \sum_{j=-i}^{i} l_{ij}^{t_1} (\mathbf{AT} + \mathbf{B\Omega}) \\ &\quad + \sum_{i=0}^{2} \sum_{j=-i}^{i} \Delta l_{ij} b_{ij}(\mathbf{n}_{\mathbf{P_1}}). \end{aligned} \tag{3}$$

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

We can express the result in (1) succinctly using tensor notation as

$$\mathcal{I} = \left(\mathcal{B} + \mathcal{C} \times_2 \begin{pmatrix} \mathbf{T} \\ \mathbf{\Omega} \end{pmatrix} \right) \times_1 \mathbf{l}, \tag{4}$$

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

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

For each pixel (p, q) in the image, $\mathcal{C}_{klpq} = [\begin{array}{cc} \mathbf{A} & \mathbf{B} \end{array}]$ of size $N_l \times 6$. Thus for an image of size $M \times N$, $\mathcal{C}$ is $N_l \times 6 \times M \times N$. $\mathcal{B}$ is a sub-tensor of dimension $N_l \times 1 \times M \times N$, comprising the basis images $b_{ij}(\mathbf{n}_{\mathbf{P_1}})$, and $\mathcal{I}$ is a sub-tensor of dimension $1 \times 1 \times M \times N$, representing the image. $\mathbf{l}$ is still the $N_l \times 1$ vector of the illumination coefficients.

2.2. Learning Joint Illumination and Motion Models from Video

Equation (1) provides us an expression relating the reflectance image I_{t_2} with new illumination coefficients $l_{ij}^{t_2}$ and motion variables $\mathbf{T}, \mathbf{\Omega}$, which lead to a method for estimating 3D motion and illumination as:

$$\begin{aligned} (\hat{\mathbf{l}}, \hat{\mathbf{T}}, \hat{\mathbf{\Omega}}) &= \arg\min_{\mathbf{l}, \mathbf{T}, \mathbf{\Omega}} \| I_{t_2} - \sum_{i=0,1,2} \sum_{j=-i}^{i} l_{ij} b_{ij}(\mathbf{n}_{\mathbf{P}'_2}) \|^2 \\ &\quad + \alpha ||\mathbf{m}||^2 \\ &= \arg\min_{\mathbf{l}, \mathbf{T}, \mathbf{\Omega}} \| \mathcal{I}_{t_2} - \left(\mathcal{B}_{t_1} + \mathcal{C}_{t_1} \times_2 \begin{pmatrix} \mathbf{T} \\ \mathbf{\Omega} \end{pmatrix} \right) \times_1 \mathbf{l} \|^2 \\ &\quad + \alpha ||\mathbf{m}||^2 \end{aligned} \tag{5}$$

where $\hat{x}$ denotes an estimate of x. Since the motion between consecutive frames is small, but illumination can change suddenly, we add a regularization term to the above cost function. It is of the form $\alpha ||\mathbf{m}||^2$, where $\mathbf{m} = \begin{pmatrix} \mathbf{T} \\ \mathbf{\Omega} \end{pmatrix}$.

Since the image I_{t_2} lies approximately in a bilinear space of illumination and motion variables (ignoring the regularization term for now), such a minimization problem can be achieved by alternately estimating the motion and illumination parameters by projecting the video sequence onto the appropriate basis functions derived from the bilinear space. This process guarantees convergence to a local minimum [10]. Assuming that we have tracked the sequence upto some frame for which we can estimate the motion (hence, pose) and illumination, we calculate the basis images, b_{ij}, at the current pose, and write it in tensor form $\mathcal{B}$. Unfolding[1] $\mathcal{B}$ and the image $\mathcal{I}$ along the first dimension, [9] which is the illumination dimension, the image can be represented as:

$$\mathcal{I}_{(1)}^T = \mathcal{B}_{(1)}^T \mathbf{l}. \tag{6}$$

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

$$\hat{\mathbf{l}} = (\mathcal{B}_{(1)}\mathcal{B}_{(1)}^T)^{-1}\mathcal{B}_{(1)}\mathcal{I}_{(1)}^T. \tag{7}$$

Keeping the illumination coefficients fixed, the bilinear space in equations (1) and (2) becomes a linear subspace, i.e.,

$$\mathcal{I} = \mathcal{B} \times_1 \mathbf{l} + \mathcal{G} \times_2 \begin{pmatrix} \mathbf{T} \\ \boldsymbol{\Omega} \end{pmatrix}, \tag{8}$$

where $\mathcal{G} = \mathcal{C} \times_1 \mathbf{l}$. Similarly, unfolding all the tensors along the second dimension, which is the motion dimension, and adding the effect of the regularization term, the cost function becomes:

$$\begin{aligned} (\hat{\mathbf{T}}, \hat{\boldsymbol{\Omega}}) &= \arg\min_{T,\Omega} \|\mathcal{I} - \left(\mathcal{B} \times_1 \mathbf{l} + \mathcal{G} \times_2 \begin{pmatrix} \mathbf{T} \\ \boldsymbol{\Omega} \end{pmatrix}\right)\|^2 \\ &\quad + \alpha\|\mathbf{m}\|^2 \end{aligned} \tag{9}$$

Substituting the definition of regularization term $\mathbf{m}$ into equation (9), it becomes a least square problem, and $\mathbf{T}$ and $\boldsymbol{\Omega}$ can be estimated as:

$$\begin{pmatrix} \hat{\mathbf{T}} \\ \hat{\boldsymbol{\Omega}} \end{pmatrix} = \left(\mathcal{G}_{(2)}\mathcal{G}_{(2)}^T + \alpha\mathbf{I}\right)^{-1} \mathcal{G}_{(2)}(\mathcal{I} - \mathcal{B} \times_1 \mathbf{l})_{(2)}^T, \tag{10}$$

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

Algorithm for estimating illumination from a video sequence: Consider a sequence of image frames $I_t, t = 0, ..., N-1$.

Initialization: Take one image of the object from the video sequence, register the 3D model onto this frame and map the texture onto the 3D model. Calculate the tensor of the basis images $\mathcal{B}_0$ at this pose. Use (7) to estimate the illumination coefficients. Now, assume that we know the motion and illumination estimates for frame t, i.e., $\mathbf{T}_t, \boldsymbol{\Omega}_t$ and $\mathbf{l}_t$.

- • Step 1. Calculate the tensor form of the bilinear basis images $\mathcal{B}_t$ at the current pose using (2). Use (10) to estimate the new pose from the estimated motion.
- • Step 2. Assume illumination does not change, i.e. $\hat{\mathbf{l}}_{t+1} = \hat{\mathbf{l}}_t$. Compute the motion $\mathbf{m}$ by minimizing the difference between an input frame and the rendered frame $\|\mathcal{I}_{t+1} - (\mathcal{B}_t + \mathcal{C}_t \times_2 \begin{pmatrix} \hat{\mathbf{T}}_{t+1} \\ \hat{\boldsymbol{\Omega}}_{t+1} \end{pmatrix}) \times_1 \hat{\mathbf{l}}_{t+1}\|^2$, and estimate the new pose.
- • Step 3. Using the new pose estimate, re-estimate the illumination using (7). Repeat Steps 1 and 2 with the new estimated $\hat{\mathbf{l}}_{t+1}$ for that input frame.
- • Step 4. If the difference error between the input frame and the rendered frame can be reduced lower than an acceptable threshold, go to Step 5. Otherwise, perform a local optimization (using any gradient descent method) initializing with the estimates of motion and illumination.
- • Step 5. Set t = t + 1. Repeat Steps 1, 2, 3 and 4.
- • Step 6. Continue till t = N - 1.

In many practical situations, the illumination changes slowly within a sequence (e.g., cloud covering the sun). In this case, we use the expression in (3) instead of (1,2) in the cost function (5) and estimate Δl_{ij}.

3. EXPERIMENTAL RESULTS

In this section, we show the results for synthesizing new video sequences under the illumination conditions learned from the original ones.

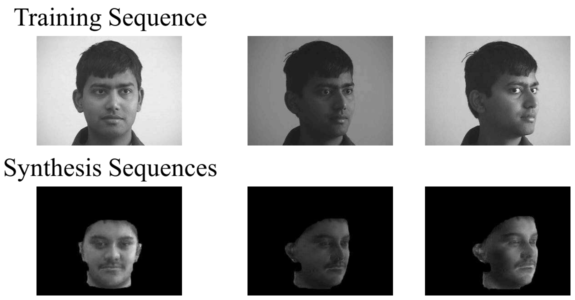


Fig. 2. Face synthesis with the motion and illumination models learned from training sequence. Motion and illumination are learned from the frames in the first row, and images in the second row are synthesized with the motion and illumination learned from the corresponding frames in the same column.

In Figures 3, we show examples of synthesizing a face using learned illumination and motion models. Motion and illumination are learned from the frames in the first and second rows respectively, and images in the third row are synthesized with the motion and illumination learned from the corresponding frames in the same column. This example shows how we can decouple lighting and motion. Also note the

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

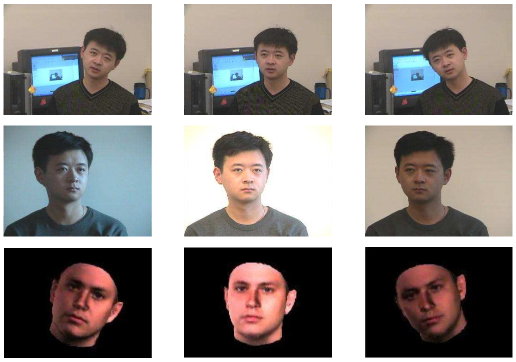
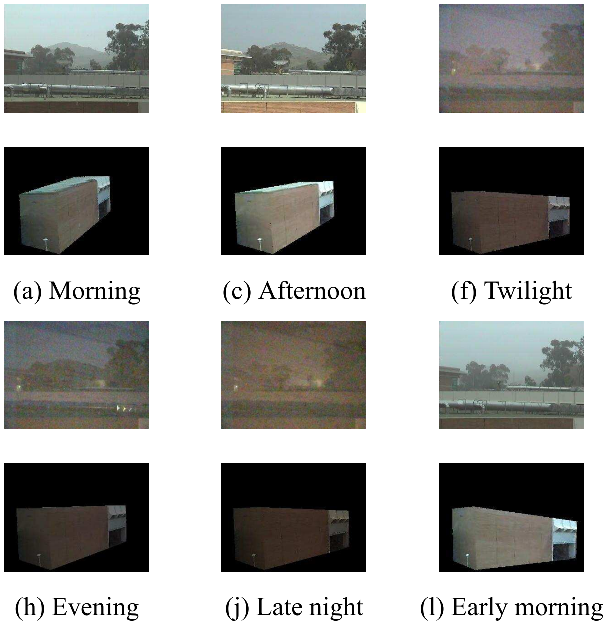


Fig. 3. Another example of video synthesis with learned motion and illumination models. Motion and illumination are learned from the first row and second row respectively, and the corresponding frames in the third row are synthesized with the learned motion and illumination parameters.



(a) Morning (c) Afternoon (f) Twilight

(h) Evening (j) Late night (l) Early morning

Fig. 4. Synthesis of a building with illumination models learned from a natural outdoor scene. All the building frames are synthesized with the illumination learned from the corresponding frames on the top row. The illumination was learned throughout the day by looking at a mountain which is shown in the top row. Motion on the building is applied artificially.

effect of local illumination changes in the synthesized sequences, which would be difficult using purely image processing-based methods.

In another experiment, we observed a hill from our lab over a 24 hour period. A portion of this scene was used to learn the illumination model. There was no tracking involved here as the scene was static. We then synthesized a building on the campus using the learned illumination models for different times of the day. The motion was added artificially. Examples of the synthesized images are shown in Figure 4.

4. CONCLUSIONS

In this paper, we presented a approach for estimating illumination and motion simultaneously from a video sequence. The proposed method uses spherical harmonics based illumination representation, and allows us to learn time-varying models of illumination. The models can be used for video indexing and summarization as well as to recreate new objects of the same class under the illumination conditions in the training sequences. We showed video relighting synthesis results using this approach and analyzed its effectiveness on a number of examples.

5. REFERENCES

[1] R. Basri and D.W. Jacobs, “Lambertian Reflectance and Linear Subspaces,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 25, no. 2, pp. 218–233, February 2003.

[2] R. Ramamoorthi and P. Hanrahan, “On the relationship between radiance and irradiance: determining the illumination from images of a convex Lambertian object,” *Journal of the Optical Society of America A*, vol. 18, no. 10, Oct 2001.

[3] L. Zhang, B. Curless, A. Hertzmann, and S.M. Seitz, “Shape and Motion under Varying Illumination: Unifying Structure from Motion, Photometric Stereo, and Multi-view Stereo,” in *Proc. of IEEE International Conference on Computer Vision*, 2003.

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

[5] J. Lim, J. Ho, M. Yang, and D. Kriegman, “Passive photometric stereo from motion,” in *Proc. of IEEE International Conference on Computer Vision*, 2005.

[6] R. Ramamoorthi, “Modeling Illumination Variation With Spherical Harmonics,” in *Face Processing: Advanced Modeling and Methods*. Academic Press, 2005.

[7] R. Ng, R. Ramamoorthi, and P. Hanrahan, “Wavelet triple product integrals for all-frequency relighting,” in *SIGGRAPH*, 2004, pp. 475–485.

[8] T. Okabe, I. Sato, and Y. Sato, “Spherical harmonics vs. haar wavelets: Basis for recovering illumination from cast shadows,” in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, 2004, pp. I: 50–57.

[9] L. D. Lathauwer, B. D. Moor, and J. Vandewalle, “A Multilinear Singular Value Decomposition,” *SIAM J. Matrix Anal. Appl.*, vol. 21, no. 4, pp. 1253–1278, 2000.

[10] L. Ljung and T. Soderstrom, *Theory and Practice of Recursive Identification*, MIT Press, 1987.