Active Lighting Learning for 3D Model Based Vehicle Tracking

Tingbo Hou
Computer Science Department
Stony Brook University
thou@cs.sunysb.edu

Sen Wang
Kodak Research Laboratories
Eastman Kodak Company
sen.wang@kodak.com

Hong Qin
Computer Science Department
Stony Brook University
qin@cs.sunysb.edu

Abstract

*Varying illumination is a challenging issue in many computer vision problems (*e.g.*, tagging, matching, and tracking), while in inverse rendering, people are interested in estimating illumination from rendered images or videos. Can these two techniques be combined together to form a unified framework for vehicle tracking and lighting learning? This paper gives probably the first thought in this joint problem, by presenting a framework to adaptively learn lighting from an image sequence while tracking the object (specifically, the vehicle) in it. We formulate the illumination model with both diffusion and specularity components using a frequency-space representation, and design a nonlinear model to estimate lighting coefficients in a low-dimensional subspace. The lighting learning and vehicle tracking are integrated in a unified Markov network, which can be solved by an iterative believe propagation (BP) method. The proposed framework can track a vehicle moving in a video, as well as transfer the learned lighting to other objects, which shows its potential in augmented reality.*

1. Introduction

Varying illumination is usually considered as an obstacle for many computer vision problems, such as tagging, matching and tracking, since the consistency assumption of intensity or appearance breaks down. Ambiguities will always occur when appearance changes (due to lighting variation) are incorrectly interpreted as motion effects. In inverse rendering, recent research [1, 9] shows that the appearance of an object can be described as a spherical convolution of the illumination and Bidirectional Reflectance Distribution Function (BRDF). It provides an enable tool to actively learn lighting from images or videos, which can assist matching and tracking problems.

In recent years, many techniques to handle varying illumination have been proposed and are roughly classified into two categories: the implicit methods and the explicit methods. The implicit methods do not probe into illumination model using analytic representations, including Intrinsic Images [12, 14], Illumination Ratio Map (IRM) [18], etc. The recovery of intrinsic images, which are decomposed into illumination images and reflectance images, require image sequence with the same viewpoint but different illumination [14] or a single image with prior knowledge of a camera response curve [12]. The IRM [18] considers homogeneous Lambertian surfaces, which have linear response to illumination changes and describes the changes by the ratio of intensities of corresponding pixels. The explicit methods focus on representing an analytic formula of the illumination model and factoring lighting from the reflectance light field. Recent work shows that by employing convolution representations, the reflected light field can be expressed mathematically as a product of spherical harmonics coefficients of the BRDF and the lighting. However, these methods have strong requirements on the prior knowledge of object geometry and texture: either known [1, 9] or approximated by a set of PCA bases obtained from a training dataset [13, 20].

In this paper, we propose a novel framework to adaptively learn lighting from an image sequence while tracking the object (specifically, the vehicle) in it. The contributions of this paper are as follows:

(1) We formulate the lighting learning with both diffusion and specularity components in a Heteroscedastic Errors-in-Variables (HEIV) model, where the lighting coefficients can be iteratively estimated.

(2) We present a unified Markov network to integrate lighting learning and vehicle tracking, which can be solved by an iterative belief propagation (BP) method.

(3) Although our framework has been used on the vehicle domain under the visible spectrum, it has a potential to be applied to other types of objects beyond the visible spectrum, such as near infrared (IR) lighting.

2. Previous Work

Inverse rendering which measures rendering attributes (lighting, texture, and BRDF) from photographs, continues to be an active research area with interests from both

computer vision and computer graphics. In previous work, tremendous progress has been made in the recovery of these three rendering attributes with one or two unknowns [17]. In general cases where lighting, texture and BRDF are all unknown, this problem is ill-conditioned until strong assumptions and requirements on input data have been made. Ramamoorthi and Hanrahan [9] presented a signal processing framework for inverse rendering with known geometry and isotropic BRDFs. In their work, the reflected light field was expressed as a convolution of the lighting and BRDF using spherical harmonics. As a frequency-space convolution, spherical harmonics has been used as a tool to represent lighting. In [1], it is shown that the reflected light field from a Lambertian surface can be characterized using only its first 9 spherical harmonic coefficients, where geometry is assumed to be known. Xu and Roy-Chowdhury [15] integrated motion with illumination by deducing relationship between motion and basis images of spherical harmonics. Later, Zhang *et al*. [20] integrated the spherical harmonic illumination representation into the Morphable Model approach, by modulating the texture component with the spherical harmonic bases. They used PCA to initialize geometry and texture from a large set of training data, and estimate lighting and basis images independently in iteration. To release the requirements on geometry and texture, Wang *et al*. [13] proposed a subregion based framework that uses a MRF to model the statistical distribution and spatial coherence of texture. Though lighting in small region is more homogeneous, it is hard to divide an image to homogeneous regions. Their method still needs training data to compute PCA texture model and is limited to Lambertian surfaces. Recently, Hou *et al*. [6] proposed a vehicle matching method under varying illumination using spherical harmonics. By assuming the albedo of vehicle is unified, they can recover illumination parameters to match different vehicles.

Learning illumination while tracking is a new research area which integrates illumination invariant tracking [5, 7] and learning illumination from video sequences [15]. Recently, Zhang *et al*. [19] presented an algorithm for computing optical flow, shape, and lighting from an image sequence. They exploited a scaling variable to represent intensity variation under an assumption of Lambertian surfaces. Freedman and Turek [3] proposed an illumination insensitive tracking algorithm using graph cuts. They placed consistency constraints on the illumination model, assuming that the consistent scene intensities should propagate throughout all frames. Zhang *et al*. [18] integrated IRM with optical flow into a unified Markov network and solved it using BP. Finally, Xu and Roy-Chowdhury [15, 16] showed that the set of Lambertian reflectance functions of a moving object, illuminated by arbitrarily distant light sources, lies "close" to a bilinear subspace consisting of nine illumination variables and six motion variables. However, for non-Lambertian surfaces and dramatic changes of lighting, their method can not work well.

3. Vehicle tracking with lighting learning

3.1. Illumination as convolution

We start from the reflectance light field for a surface point x with its normal $\vec{n}$, given as

$$B(x, \vec{w}_o) = \int_\Omega T(x)\rho(\vec{w}_i, \vec{w}_o)L(x, \vec{w}_i)(\vec{w}_i \cdot \vec{n})dw_i, \quad (1)$$

Here, B is the reflected light field, expressed as a function of the surface position x and outgoing direction $\vec{w}_o$. The integrand is a product of terms: the texture $T(x)$, the BRDF $\rho(\vec{w}_i, \vec{w}_o)$ and the lighting $L(x, \vec{w}_i)$.

Assuming the camera has a linear sensitive function, its response to reflectance light can be interpreted as a part of texture. Therefore, we can approximate the image intensity I using frequency-space convolution of Eq. 1 in terms of spherical harmonic functions

$$I = \sum_{l=0}^{l^*} \sum_{m=-l}^{l} \Lambda_l L_{lm}(K_d\rho_{dl} + K_s\rho_{sl})Y_{lm}(n), \quad (2)$$

where Λ_l is the normalization constant, $Y_{lm}(n)$ is the spherical harmonic function of surface normal n in the local coordinate, L_{lm} is the lighting coefficients, ρ_d and ρ_s are diffuse and specular BRDF, K_d and K_s are diffuse and specular albedo (texture), respectively, and l^* is the cutoff for levels of spherical harmonics. For the Larmbertian BRDF, an analytic formula of $A_l = \Lambda_l\rho_{dl}$ can be computed as

$$A_0 = \pi, A_1 = \frac{2\pi}{3}, A_2 = \frac{\pi}{4}, A_3 = 0, A_4 = -\frac{\pi}{24}, ...,$$

numerically.

For the specularity component of the BRDF, we use Phone illumination model with a good approximation given by $\Lambda_l\rho_{sl} \approx \exp\left[-\frac{l^2}{2s}\right]$, where s is Phone exponent that can be approximated according to certain materials or estimated independently by specifying a neighborhood region of specularity under directional source of light. This is a good approximation when $l^* \approx \sqrt{2s}$. Furthermore, we assume $K_d + K_s = 1$. Then we define basis functions $b(x)$ as a function of vector $x = [n_x, n_y, n_z, K_d]'$:

$$b_{lm}(x) = (K_d A_l + (1 - K_d)\exp\left[-\frac{l^2}{2s}\right])Y_{lm}(n), \quad (3)$$

and define l as the vector of lighting coefficients L_{lm}. Therefore, assuming s is known or under independent estimation, the Eq. 2 can be simplified as

$$I = b(x)l. \quad (4)$$

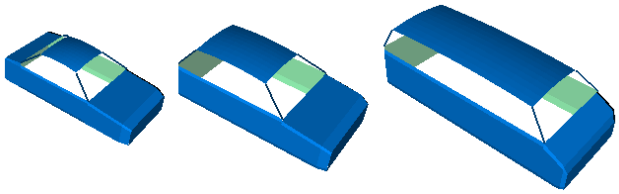

Figure 1. Three shapes of different types of vehicles (sedan, SUV, bus) generated by our generic shape model (color only for visualization purpose).

3.2. Initial settings of shape and texture

We notice that in Eq. 2, it involves the normal map of the given object but not the geometry. Following this observation, we adopt a generic model to initialize the normal map, and refine it during the later process of recovering illumination. In this paper, we choose the vehicle as our object domain for two main reasons: (1) vehicles are very commonly seen in our everyday lives; (2) vehicle matching and tracking under varying lightings is still a challenging problem with the wide range of applications in computer vision and surveillance. We segment vehicle models into 7 surfaces, and use 6 length parameters to generate control points for B-spline surfaces. Users can easily adjust those parameters to form an approximate model through a user interface. Fig. 1 shows three shapes of different types of vehicles (sedan, SUV, bus) generated by our generic model.

For the initial texture map, we adopt a color classifier similar with the work in [12] to group colors. We notice that the changes in color between pixels are due to either illumination or texture. When surfaces are Lambertian, any color changes due to illumination should affect all three color channels proportionally. When surfaces have low specularities, the chromatic changes due to illumination between two adjacent pixels should bring three color channels approximately to a proportional effect. Assuming two adjacent pixels with the same albedo have color c_1 and c_2, they should have close directions in the RGB color space. Otherwise, the chromaticity of colors is changed and color changes are caused by the changes of texture, not illumination.

To group colors associated with the same texture class, we treat each color as a vector in RGB space and normalize them. We then compute the angle between normalized color $\vec{c_1}$ and $\vec{c_2}$ as

$$(\vec{c_1}, \vec{c_2}) = \arccos(\vec{c_1} \cdot \vec{c_2}). \tag{5}$$

If the angle $(\vec{c_1}, \vec{c_2})$ is below a threshold, we assign them into the same group. Finally, we compute the average of each group of colors as the texture assigned to this group.

3.3. Estimation of lighting coefficients

To factorize lighting, BRDF and texture from Eq. 2 is ill-conditioned if all three variables are unknown, and thus we need some strategy to turn it into well-conditioned. Previous methods [13, 20] iteratively estimate the unknown coefficients and their strategy is to first estimate lighting coefficients under fixed normal and texture, and then estimate normal and texture under fixed lighting coefficients. Therefore, they highly rely on accurate input of normal and texture as initialization. In our framework, we design a HEIV model to estimate lightings and regress BRDF and texture simultaneously.

The HEIV model, introduced from statics by Matei and Meer [8], has already been applied to many problems in computer vision. It is a non-linear regression model, where values of variables are considered to be corrupted by error. We take l as the parameter and x as the variable, and then the model can be presented as

$$f(x, l) = b(x)l - I \tag{6}$$

where l is the illumination parameter to be estimated, and x is the variable with its value corrupted by error. The function $b(x)$ is named the carrier of variables. We assume variables x_i are affected by independent error

$$x_i = x_{io} + \delta x_i, \quad \delta x_i \sim G(\mu_{x_i}, \sigma^2 C_{x_i}),$$

where x_{io} denotes the true value of x_i, μ_{x_i} and C_{x_i} are the mean and the covariance matrix of the variables error δx_i. We assume the error δx_i subject to the Gaussian distribution. For the color image with three channels, there should be three vectors for each pixel with respect to RGB components of albedo. This will generate more constraints, which will be considered in later derivations.

Out estimator is based on an objective function that contains three components, given by

$$E(x, l) = \alpha E_x(x) + \beta E_f(x, l) + \gamma E_s(x), \tag{7}$$

where α, β, γ are weighted coefficients. The first term of Eq. 7 is the Mahalanobis Distance (MD) between x_i and x_{io}, shown as

$$E_x(x) = \sum_{i=1}^{n} (x_i - x_{io})^T C_x^+ (x_i - x_{io}). \tag{8}$$

The second term is the weighed quadratic sum of residual error generated by constraints of Eq. 6

$$E_f(x, l) = \sum_{i=1}^{n} \eta_i f_i(x_i, l), \tag{9}$$

where $\eta_i = w_i f_i$ is the Lagrange multiplier, and w_i is a weighted coefficient corresponding with each x_i, defined as $w_i = \left[\left(\frac{\partial f_i}{\partial x_i} \right)^T C_x \left(\frac{\partial f_i}{\partial x_i} \right) \right]^+$, where the superscript '+' denotes the pseudo inverse to account for the case of rank

deficient covariance matrices. The third term is a smooth function

$$E_s(x) = \sum_{i=1}^{n} \sum_{j \in N} (x_i - x_j)^2, \tag{10}$$

where N is the neighbor of x_i. The smooth function constraints that the estimated parameters are smooth in its local patch.

Taking the derivative of objective function E with respect to l, yields

$$\frac{\partial E}{\partial l} = 2[S_l - C_l]l = 2\sum_{i=1}^{n} w_i b_i I_i \tag{11}$$

with the weighted scatter matrix $S_l = \sum_{i=1}^{n} w_i b_i^T b_i$, and the covariance matrix $C_l = \sum_{i=1}^{n} \eta_i^T \left(\frac{\partial b}{\partial x}\right)^T C_x \left(\frac{\partial b}{\partial x}\right) \eta_i$. Finally, the optimal estimation of l is given by,

$$\widehat{l} = \left([S_l - C_l]^T [S_l - C_l]\right)^{-1} \sum_{i=1}^{n} w_i b_i I_i. \tag{12}$$

On the other hand, we also take the derivative with respect to x_i to compute its estimation, which yields

$$\frac{\partial E_i}{\partial x_i} = 2\alpha\delta x_i C_x^+ + \beta(\eta_i \frac{\partial f_i}{\partial x_i} + \frac{\partial \eta_i}{\partial x_i} f_i) + 2\gamma \sum_{j \in N} (x_i - x_j). \tag{13}$$

By ignoring the quadratic derivative, we can obtain the optimal estimation of x_i, given by

$$\widehat{\delta x_i} = -C_x(\frac{\beta}{\alpha}\eta_i \frac{\partial f_i}{\partial x_i} + \frac{\gamma}{\alpha} \sum_{j \in N} (x_i - x_j)). \tag{14}$$

From Eq. 12 and Eq. 14, we can iteratively recover lighting coefficients, and leave the de-lighted images (shape, texture) ready for tracking purpose. And the estimated lighting can be directly transfered to other objects.

3.4. Illumination invariant tracking

Based on the lighting learning above, we proposed a unified Markov network to learn illumination and track an object under varying lighting. Previous research [2, 5] shows that optical flow can be modeled as a probabilistic graphics model. Here, we develop a Markov network integrating optical flow and lighting. Given two consecutive frames I_R, I_T (often notated as "reference" and "target"), and factorized lighting coefficient l_R and basis function b_R for the reference frame, our goal is to find the maximum a posteriori probability (MAP) of the disparity map D and lighting coefficient l_T, expressed as

$$P(D, l_T | b_R, l_R, I_T) \propto \prod_{i,j \in N} \Psi_{i,j} \prod_i \Phi_i, \tag{15}$$

where N is the size of the neighborhood of node x_i, Φ_s is the evidence function describing the probability of how the estimated D and l_T fit I_T, $\Psi_{i,j}$ is the compatibility function that smooths the estimated D and l_T along the network. We further define the evidence function of node x_i in the network as

$$\Phi_i(D(x_i), l_T) = \exp(-|b_R(D(x_i))l_T - I_T(x_i)|), \tag{16}$$

and define the compatibility function $\Psi_{i,j}$ between node $[x_i, x_j]$ as

$$\Psi_{i,j}(D) = \exp(-|D(x_i) - D(x_j)|). \tag{17}$$

To find the MAP of the joint Markov network, we use belief propagation [4, 10, 11] as an efficient tool to approach the solution. The message m_{ij} propagated from node x_i to node x_j is computed as

$$m_{ij}(x_j) \leftarrow \alpha \max_{x_i} \Psi_{ij} \Phi_i \prod_{k \in N \setminus j} m_{ki}(x_i), \tag{18}$$

where N is the size of the neighborhood of node x_i, and α is a normalization factor. The final belief for each node is determined by the local evidence function and the incoming messages from its neighbors:

$$b_i \leftarrow \alpha \Phi_i \prod_{k \in N} m_{ki}(x_i). \tag{19}$$

In our framework, we consider that the changes of lighting between two consecutive frames are not dramatic. Therefore, we use l_R as the initial value for l_T, and propagate belief on the network. When the propagation stops, we obtain an estimated basis function $\widehat{b_T}$, and we can update the l_T by Eq. 12. Then we can propagate belief again on the network. With this iterative procedure, we finally obtain the lighting and motion of each frame while robustly track the object through the video.

4. Near-IR illumination

Surfaces of objects have similar reflectance properties under active Near-IR illumination and visible illumination. Therefore, the framework proposed above can also be applied to images under Near-IR illumination. One benefit of using Near-IR illumination is that it can provide constant illumination, and work in low illumination environment without conspicuous light. The disadvantage of Near-IR illumination is that it does not have any color information. It is not significant to objects like the human face, but is important to vehicles, since vehicles have plentiful colors that can be discriminated easily in visible illumination. Near-IR images are very close to gray images under the visible spectrum. Objects may appear unnatural under IR illumination,

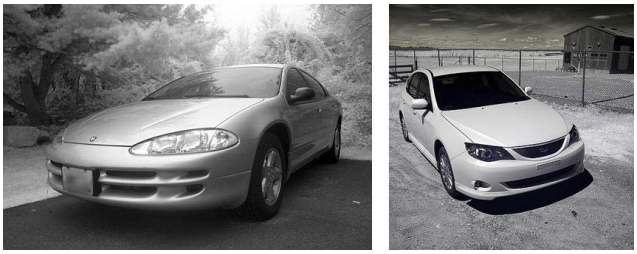

Figure 2. Vehicle images under Near-IR illumination.

since many materials do not have the same reflectance fraction under the visible or Near-IR spectrum, as shown in Fig. 2. For example, surfaces with a green color become brighter in Near-IR illumination than in the visible spectrum. Therefore, it is not realistic to compare vehicles across spectral bands. In visible illumination, the texture of a surface has RGB three channels, while in Near-IR illumination, the reflectance property has only one channel. As a potential, our framework can be applied to Near-IR illumination.

5. Experimental results

Several constraints are imposed in our experiment. First, the normal $[n_x, n_y, n_z]'$ are constrained to be a unit vector with the length $|n| = 1$. Second, we take the average of estimated normal on three color channels as the final estimation of normal n in each iteration. Third, when the estimation is under a white light source, we use the average of estimated illumination coefficients on RGB channels as the final estimation of l in each iteration, assuming lightings on three color channels are equal.

For image sequences, we first show an experimental result of learning illumination and motion in two adjacent frames in Fig. 3, where (a) is the reference image, (b) is the target image, (c) is the disparity map by BP without illumination estimation, (d) is the tracking result by BP with IRM in [18], (e) and (f) are the disparity map and de-lit image of (b) by our framework. Under varying illumination, the ideal tracking result is a consistent disparity map of the moving vehicle, somehow like (e) where intensities indicate the disparities. For an object with large textureless areas like a vehicle, the recovery of motion without model estimation is very challenging as shown in Fig. 3(d). The BP in our framework, integrating illumination and model-based motion, has greatly improved performance to the previous method in [18].

In our experiment, image sequences are captured from a toy car under dramatically changing lighting in an indoor environment. Fig. 4 shows the experimental result of illumination learning and synthesis with tracking. Despite dramatically changing lighting, our framework can track the car through the image sequences and generate a new car motion sequence with normalized illumination, as shown in

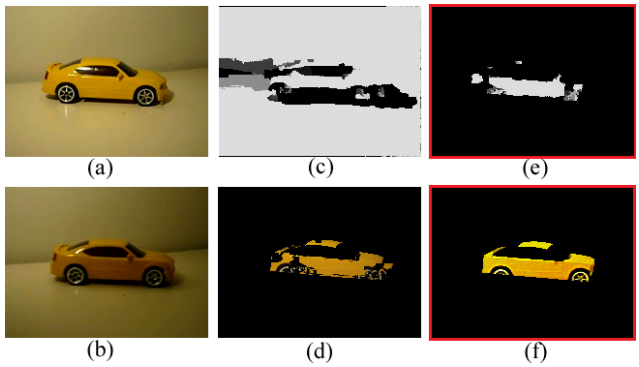

Figure 3. Learning illumination and motion between two frames: (a) is the reference image, (b) is the target image, (c) is the disparity map by BP without illumination estimation, (d) is the tracking result by BP with IRM in [18], (e) and (f) are the disparity map and de-lit image of (b) by our framework.

the second row of Fig. 4. Furthermore, we can also synthesize new video under controllable illumination learned from a given image sequence, as shown in the third row of Fig. 4, where the Max Planck model is re-lighted by the illumination learned from the first row. It is hard to quantitatively evaluate the estimation of lighting. However, in terms of visual appearances, the rendering results have similar lighting effects with the input video.

To evaluate the accuracy of lighting factorization, we compute the error of de-lit images through 105 frames by averaging the texture color and comparing them to the de-lit image in the first frame. As shown in Fig. 5, x-coordinate denotes the number of frame, y-coordinate denotes the error of de-lighting, the red dashed curve is computed by the method of IRM in [18], and the blue curve is the result of our framework. From the results, we can see that under dramatic illumination changes, our framework continues to generate more robust results on texture recovery and illumination learning.

6. Conclusion

We have proposed a method that integrates lighting learning and vehicle tracking, combining the two techniques into a unified framework. The proposed method for simultaneous illumination learning and object tracking contains a unified Markov network, which can be solved by a iterative BP method. This method can be applied to Near-IR images, while due to limited resources, we are only able to show experiments in visible illumination. In the future, we plan to further improve the results by incorporating more accurate reflectance models with cast shadows and to extend the current work to learn lighting with more real vehicle videos under different (visible or non-visible) illumination conditions.

Figure 4. Learning illumination with tracking. First row: frame 6, 23, 38, 46, 71 of the original video sequence. Second row: de-lit images by our framework. Third row: re-lit images of Max Planck model under corresponding illumination learned from images in the first row.

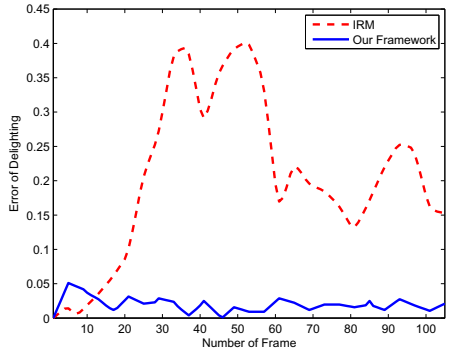

Figure 5. Error of de-lighting: the red dashed curve is computed by the method of IRM in [18], and the blue curve is the result of our framework.

References

[1] R. Basri and D. Jacobs. Lambertian reflectance and linear subspaces. *TPAMI*, 25(2):218–233, 2003.

[2] P. Felzenszwalb and D. Huttenlocher. Efficient belief propagation for early vision. In *CVPR*, pages 261–268, 2004.

[3] D. Freeman and M. Turek. Illumination-invariant tracking via graph cuts. In *CVPR*, pages 10–17, 2005.

[4] W. Freeman, E. Pasztor, and O. Carmichael. Learning low-level vision. *IJCV*, 40(1):25–47, 2000.

[5] H. Haussecker and D. Fleet. Computing optical flow with physical models of brightness variation. *PAMI*, 23(6):661–673, 2001.

[6] T. Hou, S. Wang, and H. Qin. Vehicle matching and recognition under large variations of pose and illumination. In *6th IEEE Workshop on Object Tracking and Classification Beyond and in the Visible Spectrum (in conjunction with CVPR)*, 2009.

[7] H. Jin, P. Favaro, and S. Soatto. Real-time feature tracking and outlier rejection with changes in illumination. In *ICCV*, pages 684–689, 2001.

[8] B. Matei and P. Meer. A general method for errors-in-variables problems in computer vision. In *CVPR*, pages 18–25, 2000.

[9] R. Ramamoorthi and P. Hanrahan. A signal-processing framework for inverse rendering. In *SIGGRAPH*, pages 117–128, 2001.

[10] E. Sudderth, M. Mandel, W. Freeman, and A. Willsky. Distributed occlusion reasoning for tracking with nonparametric belief propagation. In *NIPS*, pages 1369–1376, 2004.

[11] M. Tappen and W. Freeman. Comparison of graph cuts with belief propagation for stereo, using identical mrf parameters. In *ICCV*, pages 900–906, 2003.

[12] M. Tappen, W. Freeman, and E. Adelson. Recovering intrinsic images from a single image. In *NIPS*, pages 1343–1350, 2002.

[13] Y. Wang, Z. Liu, G. Hua, Z. Wen, Z. Zhang, and D. Samaras. Face re-lighting from a single image under harsh lighting conditions. In *CVPR*, pages 1343–1350, 2007.

[14] Y. Weiss. Deriving intrinsic images from image sequences. In *ICCV*, pages 68–75, 2001.

[15] Y. Xu and A. Roy-Chowdhury. Integrating the effects of motion, illumination and structure in video sequences. In *ICCV*, pages 1675–1682, 2005.

[16] Y. Xu and A. Roy-Chowdhury. Inverse compositional estimation of 3d pose and lighting in dynamic scenes. *PAMI*, 30(7):1300–1307, 2008.

[17] Y. Yu, P. Debevec, J. Malik, and T. Hawkins. Inverse global illumination: Recovering reflectance models of real scenes from photographs. In *SIGGRAPH*, pages 215–224, 1999.

[18] J. Zhang, L. McMillan, and J. Yu. Robust tracking and stereo matching under variable illumination. In *CVPR*, pages 871–878, 2006.

[19] 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 *ICCV*, pages 618–625, 2003.

[20] L. Zhang, S. Wang, and D. Samaras. Face synthesis and recognition from a single image under arbitrary unknown lighting using a spherical harmonic basis morphable model. In *CVPR*, pages 209–216, 2005.