SCIENCE CHINA

• RESEARCH PAPERS •

February 2013 Vol. 0 No. 0: 0–0
doi:

Lighting virtual objects in a single image via coarse scene understanding

CHEN XiaoWu, JIN Xin* & WANG Ke

State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University Beijing 100191*, China*

Abstract

Achieving convincing visual consistency between virtual objects and a real scene mainly relies on the lighting effects of virtual-real composition scenes. The problem becomes more challenging in lighting virtual objects in a single real image. Recently, scene understanding from a single image has made great progress. The estimated geometry, semantic labels and intrinsic components provide mostly coarse information, and are not accurate enough to re-render the whole scene. However, carefully integrating the estimated coarse information can lead to an estimate of the illumination parameters of the real scene. We present a novel method that uses the coarse information estimated by current scene understanding technology to estimate the parameters of a ray-based illumination model to light virtual objects in a real scene. Our key idea is to estimate the illumination via a sparse set of small 3D surfaces using normal and semantic constraints. The coarse shading image obtained by intrinsic image decomposition is considered as the irradiance of the selected small surfaces. The virtual objects are illuminated by the estimated illumination parameters. Experimental results show that our method can convincingly light virtual objects in a single real image, without any pre-recorded 3D geometry, reflectance, illumination acquisition equipment or imaging information of the image.

Keywords illumination estimation, lighting virtual object, single image, scene understanding, photo editing

Citation Chen X W, Jin X, Wang K. Lighting virtual objects in a single image via coarse scene understanding.

1 Introduction

Rendering virtual objects in real scenes with real illumination can greatly increase the realism of and consistency between the virtual and the real. Lighting virtual objects using illumination from a real scene is a hot topic in the computer graphics community [1, 2, 3, 4, 5, 6, 7] and has wide application in film production, digital entertainment and photo editing, amongst others. For convenient usage, our objective is to light virtual objects in a single real image automatically through illumination estimation.

Previous works on illumination estimation require manually computed geometry and reflectance values (e.g., [8]) or light probes (e.g., [9]), which limit application. In recent studies of illumination estimation from only a single outdoor image, a sun and sky doom model is typically used [1, 2] and cast shadows via sparse representation are proposed [10]. Most recently, Chen *et al.* [5] applied geometry and intrinsic components to estimate scene illumination. However, since this method randomly selects the small surfaces from the coarse geometry model, the outliers introduce inappropriate surfaces, thereby affecting the illumination estimation accuracy. More constraints and optimization need to be taken into consideration.

*Corresponding author (email: jinxin@vrlab.buaa.edu.cn)

(a) Input (b) Result

Figure 1: Estimating the illumination of a scene to insert a virtual *helicopter* into a real scene. The lighting effects of the virtual *helicopter* match those in the existing image and convincing shadows are cast on the real scene rendered using the estimated illumination.

Great success has been achieved in video illumination estimation [3, 4, 6, 7]. Recently, Xing *et al.* [6] proposed an image-based framework to estimate on-line the illumination parameters that change dynamically in outdoor video sequences. They used an interactive initialization stage with a few brushes to select areas with the specified normal. Liu *et al.* [7] estimated the relative intensities of sunlight and skylight via a spare set of planar feature-points extracted from each frame. Although we cannot use information from multiple video frames in our automatic single image composition task, their work greatly inspired us to use normal and semantic constraints to refine the surface selection.

In recent years, scene understanding from a single image has attracted the attention of many researchers. Estimation of scene geometry (e.g. [11]), semantic labels (e.g., [12]), and intrinsic components (including shading and reflectance images, e.g., [13]) has achieved great success in some specific scenarios. Scene appearance is mostly determined by the scene geometry, reflectance and illumination. We were thus inspired by these scene understanding technologies to estimate scene illumination by means of current scene understanding technologies. The estimated geometry, semantic labels, and intrinsic components provide mostly coarse information, and are not accurate enough to re-render the whole scene. However, carefully integrating the estimated coarse information can lead to an estimation of the illumination parameters of the real scene.

Thus, in this paper we propose a novel method for illumination estimation to light virtual objects in a real scene by integrating the coarse information estimated by current scene understanding technologies. Panagopoulos *et al.* [14] jointly recovered the illumination environment and an estimate of the cast shadows in a scene from a single image. However, they required a coarse 3D geometry of the primary object in the image (such as a motorcycles or a car). On the contrary, the input image for our scenario is a larger scene without obvious primary objects. Whereas, Panagopoulos *et al.* employ a higher-order Markov Random Field (MRF) illumination model, which combines low-level shadow evidence with high-level prior knowledge, we use a simple local illumination model combining the estimated coarse geometry scene model and intrinsic components of the input scene image.

Lalonde *et al.* [2] used the three most evident appearance cues (the sky, the vertical surfaces, and the ground) to estimate the illumination in a scene, directly. Unlike this work, we first estimate the coarse scene geometry and semantic labels. Next, the shading and reflectance images are estimated using a simple non-linear regression method. Then, we use the normal and semantic constraints for triangle surface selection. The selected surfaces and the shading image are used for illumination estimation using RANSAC (RANdom SAmple Consensus) refinement [15]. Finally, we light the inserted virtual objects using the estimated illumination (see Figure 1). This can be regarded as a step-by-step method for estimating from the appearance cues, the geometry, semantic labels, and intrinsic components, and

finally the scene illumination. Experimental results show that our method can convincingly light virtual objects inserted into a single real image, without any pre-recorded 3D geometry, reflectance, illumination acquisition equipment, imaging information of the image or human interactions.

The main contribution of this paper is a novel method for single image-based illumination estimation to light virtual objects in a real scene. We integrate coarse information from scene understanding into the illumination estimation framework.

2 Related work

In this section, we briefly review related work on three aspects: illumination estimation, scene understanding, and object relighting.

2.1 Illumination estimation.

The literature is rich in methods addressing the problem of illumination estimation from images or videos [9, 16, 17, 1, 2, 3, 4, 5, 6, 7]. Several works estimate illumination with the help of a pre-recorded 3D geometry model and reflectance, such as [17]. Our work relies on only a single image without exact 3D information, such as geometry and specific reflectance models. Light probes such as a light sphere [9] or fisheye camera [16] are located in the real scene to record the scene illumination directly. However, since thousands of photos have already been taken, there may have not be an opportunity for placing light probes in the scene.

For illumination estimation from outdoor images, Lalonde *et al.* [2] estimated scene illumination from only a single outdoor image. They used a dataset of six million images to train the illumination inference model and estimated a sun and sky doom model particularly for outdoor images. The three most evident appearance cues (i.e., the sky, shadows on the ground and the varied intensities of the vertical surfaces to estimate the direction of light) were used directly to estimate the illumination in the scene. However, with the great achievements in scene understanding (such as geometry [18, 11, 12], semantic label [19, 12] and intrinsic component estimation [13, 20, 21, 22]), we believe that these scene understanding technologies can assist in estimating the scene illumination.

To the best of our knowledge, the most similar work to ours is that by Chen *et al.* [5]. They proposed a method to estimate the illumination from a single image. Firstly they estimate the coarse scene geometry and intrinsic components of the scene. Then, a sparse radiance map of the scene is inferred based on the scene geometry and intrinsic components. However, they randomly select small triangle surfaces from the coarse geometry model, which introduces inappropriate surfaces and makes the illumination estimation inaccurate. Additionally, the normal and semantic constraints are not taken into consideration.

Xing *et al.* [6] proposed an image-based framework to estimate dynamically changing illumination parameters of out door video sequences online for integrating a virtual object into the video of a real scene. This approach requires very simple interaction at the initialization stage with only a few brushes used to select areas with the specified surface normal, which are used to calculate the sunlight parameters. Liu *et al.* [7] proposed a full image-based approach for on-line tracking of outdoor illumination variations from videos captured with moving cameras. their key idea is to estimate the relative intensities of sunlight and skylight via a sparse set of planar feature-points extracted from each frame. To address the inevitable feature misalignments, a set of constraints are introduced to select the most reliable ones. Exploiting the spatial and temporal coherence of illumination, the relative intensities of sunlight and skylight are finally estimated through an optimization process. Although we can not use information from multiple video frames In our automatic single image composition task, the work by Liu *et al.* greatly inspired us to use normal and semantic constraints to refine the surface selection.

2.2 Scene understanding.

As surveyed in [5], there is great deal of literature that has addressed the problem of geometry estimation from a single image. Hoiem *et al.* [18] used features such as color, texture, edge, and location to recover

the surface layout (i.e. coarse surface orientation) from a single image. Saxena *et al.* [11] used similar features to estimate the 3D scene structure directly from a single image, with good performance shown for various test images. Liu *et al.* [12] estimated the depth of a single image with the help of predicted semantic labels, while Gupta *et al.* [19] recovered a 3D parse tree of a single image through physical reasoning. For our task, either of the methods in [11] and [12] can be used to output the coarse 3D scene structure from a single image.

Intrinsic image decomposition, which was first introduced in [23], decomposes an image into a pixel by pixel product of an illumination component and a reflectance component. This is an ill-posed problem and open challenge that has attracted the attention of many researchers, with recent works such as [13, 20, 21, 22]. Because this problem is ill-posed, automatic methods are challenged by the complexity of natural images. Thus, Bousseau *et al.* [21] proposed a user-assisted approach to specify regions of constant reflectance or illumination to guide intrinsic images decomposition. Although these methods can only output a coarse shading image and a coarse reflectance image of the scene, they are adequate for use in our illumination estimation method with the refinement procedure.

2.3 Object relighting.

Methods for lighting a synthetic object coherently with a real scene can be categorized as image data based or 3D model based (such as [24]). To render some specific objects (such as rain streaks [25], a human face [26], or the human body [27]) under various scene illumination conditions, researchers pre-capture the light field data of the object using a variety of lighting, and then render them for a specific scene illumination. For human face relighting, Peers *et al.* [28] employed a quotient image extracted from two faces in the reflectance field database to transfer illumination to the input face. Jin *et al.* [29] used local lighting contrast features to learn artistic lighting templates from portrait photos. However the template is designed for classification and artistic evaluation, it is not suitable for relighting. Chen *et al.* transferred the illumination of a single reference face image to the input face using edge-preserving filters [30] and transferred the artistic illumination of masterpiece portraits to the input face using artist-draw illumination templates [31].

However, for more general applications of object relighting, 3D models are often used. In the community of augmented reality, which renders 3D models into real scenes, researchers use a simplified version of the method in [9] to achieve real-time merging [32, 33, 34]. Pre-recording scene geometry or various light probe in the real scenes are often used. Haber *et al.* [35] relighted objects by recovering the reflectance of a static scene with known geometry from a collection of images taken under distant, unknown illumination. However, the geometry of the scene is estimated from many images containing nearly the same objects. Karsch *et al.* [36] proposed a method with only a small amount of user interaction to estimate scene geometry and illumination. On the other hand, in our work the geometry is estimated by a pre-trained classifier. The virtual object used in our work is a 3D model with textures, which is illuminated using the estimated illumination parameters.

3 Lighting virtual objects

The workflow of the proposed method is illustrated in Figure 2. The estimation of scene properties (geometry structure, semantic labels and intrinsic components), the ray-based illumination model, normal and semantic constraints, and illumination estimation to light virtual objects are described in this section.

3.1 Estimation of scene properties

We first estimate the coarse geometry structure, semantic labels and intrinsic components of the input real image.

Saxena *et al.* [11] and Liu *et al.* [12] inferred a pixel wise depth map and 3D geometry structure of scenes from a single image. They used different image features with a similar MRF model. Saxena *et al.* [11] used features of color, texture, edge, location, and so on, while Liu *et al.* [12] applied semantic and

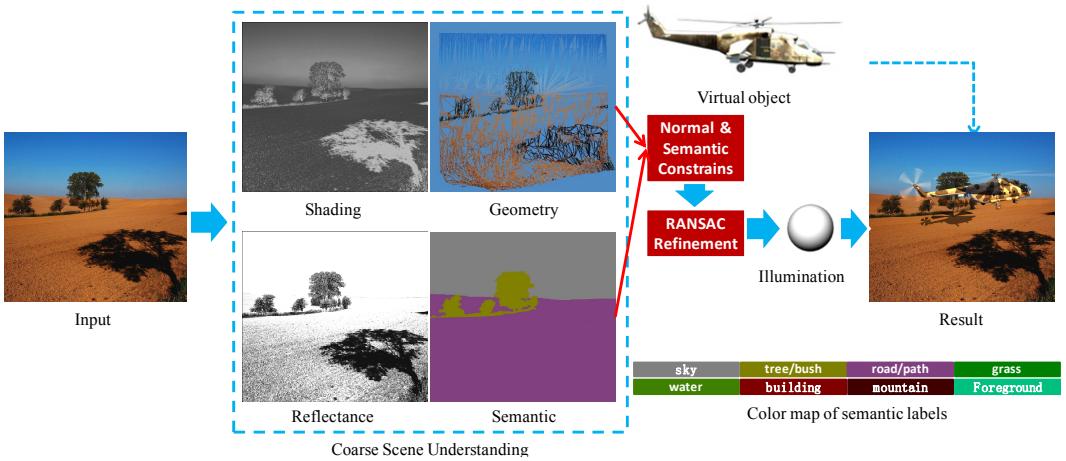


Figure 2: Workflow of our method. First we estimate the coarse geometry model and semantic labels of the input image. The input image is decomposed into intrinsic components including a shading image and a reflectance image. Then, we combine the coarse geometry model, the semantic labels and the shading image of the scene to estimate the illumination parameters of a ray based illumination model. The normal constraints and the semantic constraints are used to select the appropriate small surfaces from the coarse geometry model. Finally, the virtual object is illuminated using the estimated illumination. The virtual *helicopter* matches the input image in terms of lighting effects and casts convincing shadows on the real ground. Although the estimated geometry, semantic labels, and intrinsic components are not accurate in every pixel, the estimated illumination using via our illumination estimation algorithm is basically correct.

geometry constraints in a simple linear regression method and use the same training set of [11]. Similar to [5], in our implementation, either of the methods in [11] and [12] can be used to output the 3D scene structure from a single image. The coarse semantic labels predicted in [12] can be used for our semantic constraints in the surface selection stage. An example of the estimated geometry model and the semantic labels is shown in Figure 3.

Our aim is to estimate illumination automatically from a single image. Thus, we adopt certain automatic methods for intrinsic component estimation [13, 20, 22], similar to that in [5]. Each of these methods can be leveraged for our task of intrinsic component estimation . Thereafter, we refine the estimation results based on the normal and semantic constraints by means of RANSAC refinement. An example of the estimated shading image and reflectance image is shown in Figure 3.

We adopt the ray-based illumination model and *sparse radiance map* (SRM) as described in [5] (see Figure 4)). We make the assumption that these sparse light sources can approximate the real scene by combining the estimated light intensity. In an outdoor environment, the main light source is the sun. Thus, a small number m of virtual light sources is sufficient to estimate the sun direction. In an indoor environment, we can use a larger m to simulate multiple main light sources. We use the estimated ambient light value as the value of the remaining points in the sparse radiance map to simulate the sky outdoors and the other weak light sources indoors.

As shown in Figure 4, the light ray R in 3D space can be defined as:

$$R = I_L L \tag{1}$$

where I_L is the intensity of light ray R and L represents the unit normal vector in the ray direction. Suppose that the irradiance on surface *sur* caused by ray R is a combination of the irradiance caused by ray $R_1 = I_{L_1} L_1$ and ray $R_2 = I_{L_2} L_2$:

$$I_{L_1} L_1 \cdot N + I_{L_2} L_2 \cdot N = (I_{L_1} + I_{L_2}) \cdot N = I_L L \cdot N \tag{2}$$

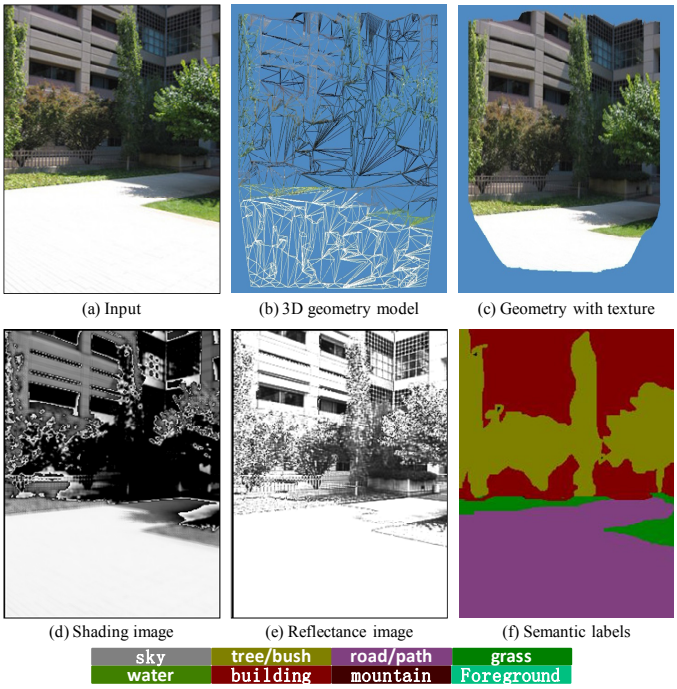


Figure 3: Estimation example: (a) input image, (b) 3D geometry model estimated by the method in [12]. (c) model with the input image as its texture. (d) and (e) shading image and reflectance image, respectively, estimated by the method in [13], and (f) predicted semantic labels according to [12].

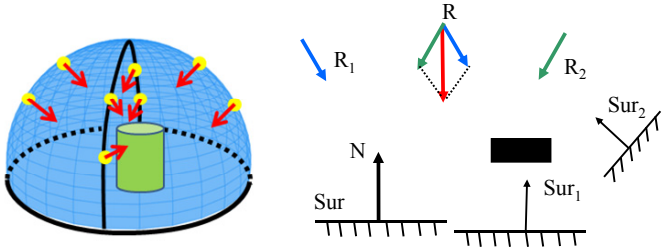


Figure 4: Sparse radiance map and ray combination (Eqs. 1, 2, 6, and 7). The sparse radiance map contains m sparse and discrete directional light sources evenly distributed on half a sphere around the scene and directed to the center point of the ground circle.

where N is the unit normal vector of the surface sur.

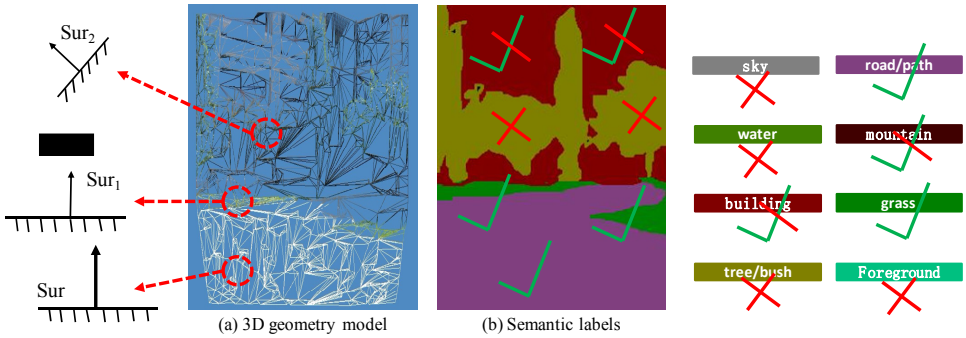
As described in [5], according to the intrinsic component decomposition, the intensity of an image scene pixel on a surface can be approximately decomposed into the irradiance collected by the surface at that point and the reflectance of the surface:

$$I = S * K \tag{3}$$

where I, S and K are the pixel values of the input scene image, the shading image and the reflectance image, respectively. For a Lambertian surface, the irradiance can be represented as [37]:

$$S = I_a + \sum_{i=1}^{m} I_{L_i} L_i \cdot N \tag{4}$$

where I_a is the ambient light of the scene. I_{L_i} and L_i are the intensity and direction, respectively, of the $i-th$ ray reaching surface sur. m is the number of rays reaching surface sur, and N and K are the



(a) 3D geometry model (b) Semantic labels

Figure 5: Normal and semantic constrains. We first use the normal constraint to eliminate those surfaces that do not obey Eq. 6 and Eq. 7. Then, for the selected ones, we use the semantic constraint rules to select more appropriate surfaces.

normal and the reflectance estimated in Section 3.1. Then we employ Levenberg-Marquardt algorithm [38] to obtain the solution of minimization between the shading image and the estimated irradiance:

$$\arg \max_{(I_a, I_1, I_2, \ldots, I_m)} \sum_{j=1}^{n_s} (S_j - (I_a + \sum_{i=1}^{m} I_{L_i} L_i \cdot N)) \tag{5}$$

where S_j is the value of the shading image of the $j-th$ 3D triangle surface estimated in Section 3.1, and n_s is the number of triangle surfaces used for illumination estimation.

3.2 Normal and semantic constraints

In real scenes, not all the surfaces satisfy Eq. 2. The surface should be visible to R, R_1 and R_2 (Eq. 6). The cosine of the angles between the three rays and the surface normal N should be above zero (Eq. 7). Thus, to use sparse light sources to approximate the illumination in real scenes, we would mostly choose the surfaces satisfying Eq. 6 and Eq. 7 to estimate the sparse radiance map.

$$\text{Vis}(sur, L) = \text{Vis}(sur, L_1) = \text{Vis}(sur, L_2) = 1 \tag{6}$$

$$L_1 \cdot N > 0, L_2 \cdot N > 0, L \cdot N > 0 \tag{7}$$

Shadow and highlight areas often do not satisfy Eq. 6 and Eq. 7. Most light sources are located immediately above of the scene. Thus, we select the surfaces whose normal orientations are almost directly above by setting a threshold T_N for the angle between the surface normal and the normal perpendicular to the horizontal ground (see the left part of Figure 5). To prune some highlight surfaces, we simply set a threshold T_H for the pixel value in the input image.

In addition, in outdoor scenes, semantic labels can be used to add more constraints. For example, surfaces on the road often face the light sources directly, whereas the lower parts of a building are always occluded from the sun. Water is not an ideal Lambert area. Thus, taking the semantic labels of [12] as an example, we add more semantic constraints according to the following rules (see the right part of Figure 5):

(1)Do not select any surface on sky, water, trees/bushes or the foreground;

(2)Select surfaces on a road/path and the grass;

(3)Select the surface on the upper parts of buildings and mountains if there is sky above them.

3.3 Illumination estimation for lighting virtual objects

However, the normal constraint and the semantic rules are not sufficient for pruning all the surfaces that do not satisfy Eq. 6 and Eq. 7. This is mainly because the estimation errors can occur in the geometry, semantic labels, and intrinsic component described in Section 3.1. Thus, we leverage the Random sample consensus (RANSAC) algorithm [15] together with our surface selection scheme to refine the illumination estimation by kicking out certain outliers.

We briefly describe the entire illumination estimation method combining the sparse radiance map, the normal and semantic constraints, the Levenberg-Marquardt algorithm, and the RANSAC algorithm in Algorithm 1.

The estimated $I_1, I_2, \ldots, I_m$ from the sparse radiance map are considered as the main light sources. We use the estimated ambient light value I_a as the value of the remaining points in the sparse radiance map. Using the estimated sparse radiance map, we light objects in a real scene using off-the-shelf rendering software. An example of the rendering result is shown in Figure 6.

Algorithm 1 Illumination estimation

Input:

The input image.

The estimated geometry model, semantic labels, shading image and reflectance image according to Section 3.1.

Maximum iterations M, and error threshold T_E.

Number of light sources to be estimated m.

Output:

The sparse radiance map of the input image.

i.e. $(I_a, I_1, I_2, \ldots, I_m)$.

1: Set min_error = MAXERROR. $iterations = 0$.

Select n triangle surfaces whose angle with the normal perpendicular to the horizontal ground is greater than T_N and whose pixel value in the input image is less than T_H from the estimated geometry model.

Use the three semantic rules to selects surfaces. Set appropriate T_N, T_H, and semantic constraint parameters to ensure $n/2 \geqslant m + 1$.

2: Randomly select $n/2$ surfaces in 1. And use Levenberg-Marquardt algorithm to get the best-fit $(I_a, I_1, I_2, \ldots, I_m)$ described in Eq. 5. by setting $n_s = n/2$.

Use the other $n/2$ surfaces to compute the fit error by setting $n_s = n/2$:

$Error = \sum_{j=1}^{n_s}(S_j - (I_a + \sum_{i=1}^{m} I_{L_i} L_i \cdot N))$

$iterations + 1$. If $Error < min\_error$, set $min\_error = Error$.

3: If $iterations > M$ or $min\_error < T_E$

return $(I_a, I_1, I_2, \ldots, I_m)$ with the min_error.

Else go to 2.

4 Experiments

In this section, we present the estimated and rendered results using a varying number of light sources in the sparse radiance map, the rendering results of various input images, and comparisons with random surface selection [5] and the work of Lalonde *et al.* [2].

4.1 Number of virtual light sources

As depicted in Figure 7 and Figure 8, denote m as the number of the virtual light sources in the sparse radiance map. Obviously, the larger m is, the greater the time is needed for the estimation process and the closer the estimated sparse radiance map is to the real illumination distribution. An appropriate

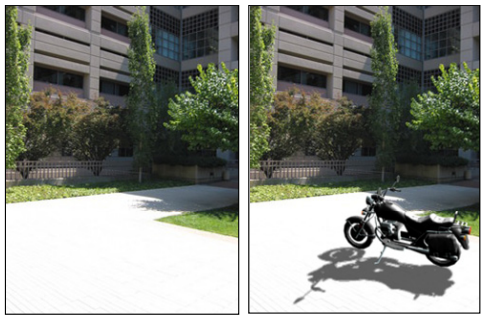


Figure 6: Example of the estimation and rendering result. (a) input image, and (b) rendering a virtual object in the image in (a).

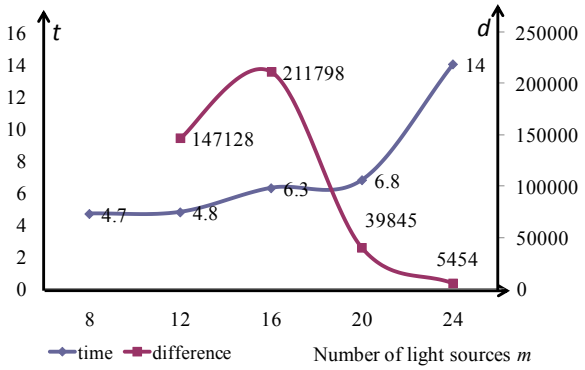


Figure 7: The time cost compared with the difference in the estimated sparse radiance maps. Note that, the differences between $m = 20$ and $m = 16$, and $m = 24$ and $m = 20$ are relatively small. But when $m = 24$, the time cost is much larger than when $m = 20$. The scene used for this analysis is depicted in Figure 8.

Figure 8: Estimated and rendering results with a varying number of light sources in the sparse radiance map and the time costs. Note that the differences of (c), (d), and (e) are quite small.

m should be chosen according to the task. In the experiments, we observed that typically a small m is sufficient to simulate the scene illumination in consumer photos. Increasing m increases the time cost of the illumination estimation (Algorithm 1), although the difference between the rendered results decreases. We tested $m = 8, 12, 16, 20, 24$. The rendering differences were calculated with respect to $m = 12$, i.e. the Sum of Squared Difference (SSD) between the white balls rendered with the estimated sparse radiance maps as shown in Figure 8 with $m = 12$ and $m = 8$.

Figure 9: Comparison between using the methods in [11] and [12]. In the outdoor images, with semantic constrains, the ground estimated in [12] is flatter than that in [11], which is more suitable for the illumination estimation and shadow rendering in our task. In the user study, the subjects were invited to rank the results as first or second according to the realism of the illumination effects of the virtual objects. The average rank scores are shown, where lower ranks are better.

4.2 Surface selection and RANSAC refinement

For geometry structure recovery, Saxena *et al.* [11] used features of color, texture, edge, location, and so on, while Liu *et al.* [12] adopted semantic and geometry constraints for simple linear regression method. In our experiments, we observe that, with semantic constraints, more reasonable results can be inferred (see Figure 9). With the surface selection scheme and RANSAC refinement described in Section 3.3 and Algorithm 1, surfaces whose normal orientations are not almost above, the highlight surfaces, surfaces in opposition to the semantic rules and some outlier surfaces are pruned. Although certain geometry, semantic and intrinsic component estimation errors exist, the estimated results are more convincing with such a refinement (see Figure 10). More results are shown in Figure 11.

User study. We carried out a small human factor experiment to ascertain the credibility of the rendered results. We invited 22 subjects who come from different background (6 females and 16 males aged between 18 and 30, some of whom were professional in applying visual effects for movie production) to evaluate our results. In Figure 9 and Figure 10, three test images are used in each figure for the user study.

For Figure 9, the subjects were invited to rank the results as first or second according to the realism of the illumination effects of the virtual objects. The user study result shows that with semantic constraints [12], more reasonable results can be inferred.

For Figure 10, the subjects are invited to rank the results between on and four according to the realism of the illumination effects of the virtual objects. The average rank score of each result is shown below. This user study shows that results with full constraints and RANSAC refinement are more convincing than those obtained from previous steps.

Figure 10: RANSAC refinement results (the stone lion and the dark yellow chair are virtual objects) (a) result of random surface selection [5]. (b) result after some highlight surfaces are pruned in those surfaces with normal orientations not directly above. (c) result using semantic constraints, and (d) result with RANSAC refinement. The shadows on the ground in (d) are more convincing than those in (a), (b), and (c). The selected surfaces of the images in the third line in (a), (b), (c), and (d) are shown below them. In the user study, the subjects were invited to rank the results between on and four according to the realism of the illumination effects of the virtual objects. The average rank score of each result is shown below (the lower, the better).

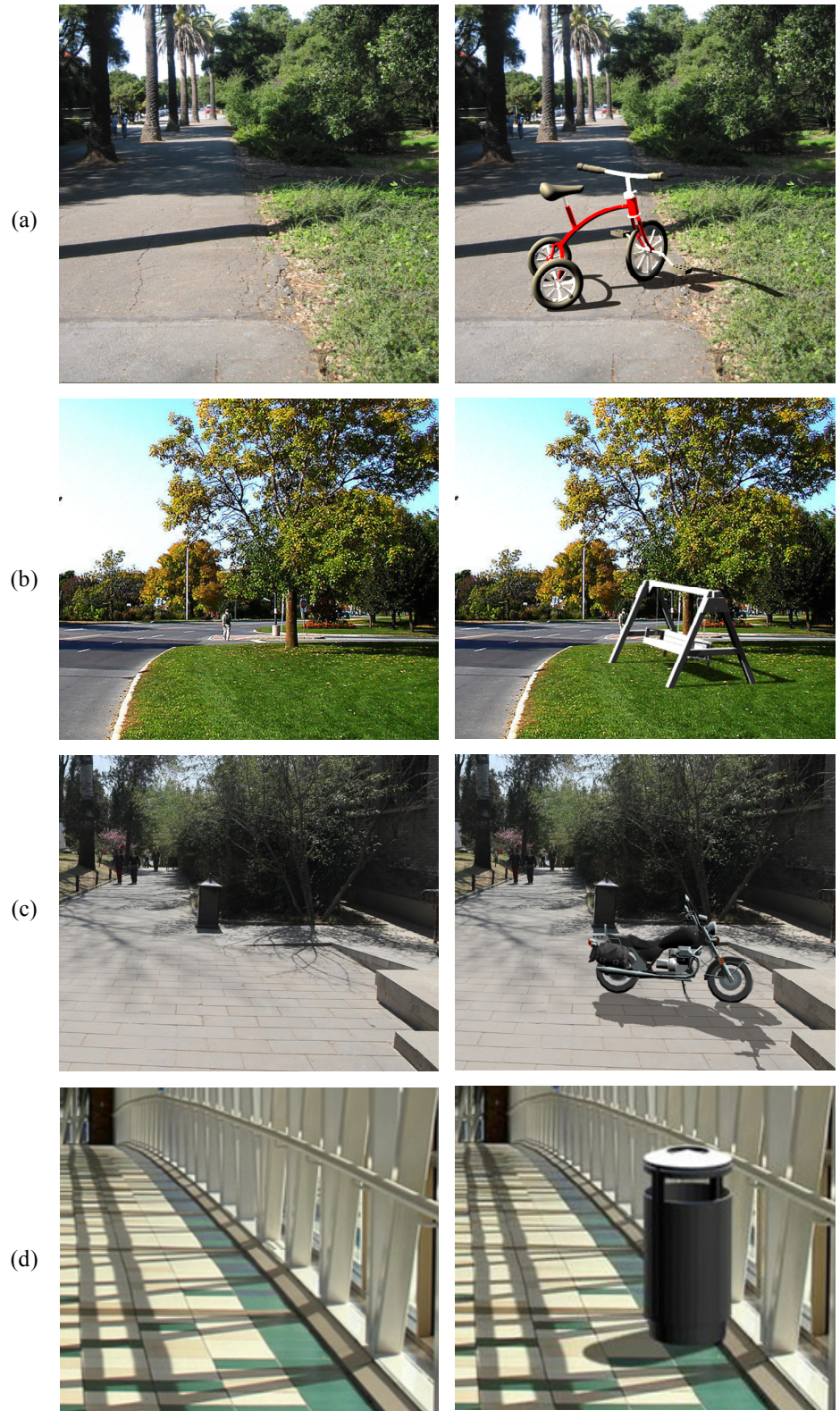


Figure 11: More results of various input images. The virtual objects inserted in to the real scenes are: (a) *bicycle*, (b) *swing* , (c) *motorcycle*, and (d) *trash can*.

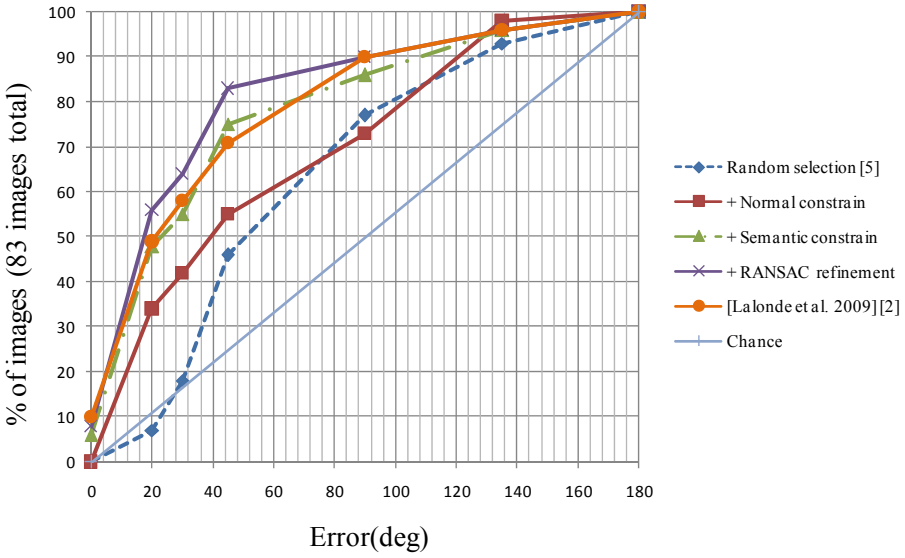


Figure 12: Quantitative evaluation using 83 images taken from the Webcam Clip Art Dataset [39]. Cumulative sun position error (angle between estimate and ground truth directions) for various refinement steps. The final RANSAC refinement outperforms the previous steps and those by Lalonde *et al.* [2].

4.3 Comparison with related work

Chen *et al.* [5] randomly selected small surfaces without taking the normal and semantic labels into consideration, thus limiting the performance of their illumination estimation. Lalonde *et al.* [2] used a dataset of six million images for training an illumination inference model and estimated a sun and sky doom model for outdoor illumination. They used the three most evident appearance cues directly to estimate the illumination in a scene. A quantitative evaluation using 83 images taken from the Webcam Clip Art Dataset [39] is shown in Figure 12. The experiments show that, our method outperforms the methods in [5] and [2], with our cues and our entire estimation process being quite different.

5 Conclusion and discussion

In this paper, we proposed a novel method for single image-based illumination estimation to light virtual objects in real scenes. The main contribution of our work is the integration of the coarse information estimated by scene understanding to estimate scene illumination. Using current scene understanding technologies and normal and semantic constraints, we have shown convincing results comparable with the state of the art.

Discussion and future work. In our current work, we use a simple local illumination model. Although, light source directions are estimated well, the shadow chromaticity is sometimes not very convincing compared with the directions (see Figure 6 and Figure 10). The global illumination model can be employed to estimate more complex illumination. However, more parameters such as materials and greater estimation accuracy are required using future scene understanding technologies. This is a trend for illumination estimation work.

The illumination parameters of a scene including lighting intensity and direction are objective quantities. Synthesized images with known illumination parameters provide good ground truths. However, the training dataset of the estimation method for coarse geometry, semantic labels, and intrinsic components are all real images. Thus, in our future work, we will use synthesized data for our new training dataset to validate our method. Video illumination estimation based on scene understanding is also seen as a future work.

Acknowledgements

We would like to thank the anonymous reviewers for their help in improving the paper. This work was partially supported by NSFC (60933006), 863 Program (2012AA011504), and ITER (2012GB102008).

References

1 C. B. Madsen and M. Nielsen, "Towards probe-less augmented reality - a position paper," in *GRAPP*, 2008, pp. 255–261.
2 J.-F. Lalonde, A. Efros, and S. Narasimhan, "Estimating natural illumination from a single outdoor image," in *IEEE International Conference on Computer Vision*, October 2009.
3 Y. Liu, X. Qin, S. Xu, E. Nakamae, and Q. Peng, "Light source estimation of outdoor scenes for mixed reality," *The Visual Computer*, vol. 25, no. 5-7, pp. 637–646, 2009.
4 Y. Liu, X. Qin, G. Xing, and Q. Peng, "A new approach to outdoor illumination estimation based on statistical analysis for augmented reality," *Journal of Visualization and Computer Animation*, vol. 21, no. 3-4, pp. 321–330, 2010.
5 X. Chen, K. Wang, and X. Jin, "Single image based illumination estimation for lighting virtual object in real scene." in *CAD/Graphics'2011*, 2011, pp. 450–455.
6 G. Xing, Y. Liu, X. Qin, and Q. Peng, "On-line illumination estimation of outdoor scenes based on area selection for augmented reality." in *CAD/Graphics'2011*, 2011, pp. 439–442.
7 Y. Liu and X. Granier, "Online tracking of outdoor lighting variations for augmented reality with moving cameras," *IEEE Trans. Vis. Comput. Graph.*, vol. 18, no. 4, pp. 573–580, 2012.
8 S. R. Marschner and D. P. Greenberg, "Inverse lighting for photography," in *In Fifth Color Imaging Conference*, 1997, pp. 262–265.
9 P. Debevec, "Rendering synthetic objects into real scenes: bridging traditional and image-based graphics with global illumination and high dynamic range photography," in *Proceedings of the 25th annual conference on Computer graphics and interactive techniques*, ser. SIGGRAPH '98. New York, NY, USA: ACM, 1998, pp. 189–198.
10 X. Mei, H. Ling, and D. W. Jacobs, "Illumination recovery from image with cast shadows via sparse representation," *IEEE Transactions on Image Processing (TIP)*, no. 8, pp. 2366–2377, 2011.
11 A. Saxena, M. Sun, and A. Y. Ng, "Make3d: Learning 3d scene structure from a single still image," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, pp. 824–840, May 2009.
12 B. Liu, S. Gould, and D. Koller, "Single image depth estimation from predicted semantic labels," in *CVPR*, 2010, pp. 1253–1260.
13 M. F. Tappen, E. H. Adelson, and W. T. Freeman, "Estimating intrinsic component images using non-linear regression," *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, vol. 2, pp. 1992–1999, 2006.
14 A. Panagopoulos, C. Wang, D. Samaras, and N. Paragios, "Illumination estimation and cast shadow detection through a higher-order graphical model," in *CVPR*, 2011, pp. 673–680.
15 M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, pp. 381–395, June 1981.
16 J. michael Frahm, K. Koeser, D. Grest, and R. Koch, "Markerless augmented reality with light source estimation for direct illumination," in *In Conference on Visual Media Production CVMP*, 2005.
17 R. Basri, D. Jacobs, and I. Kemelmacher, "Photometric stereo with general, unknown lighting," *Int. J. Comput. Vision*, vol. 72, pp. 239–257, May 2007.
18 D. Hoiem, A. A. Efros, and M. Hebert, "Recovering surface layout from an image," *Int. J. Comput. Vision*, vol. 75, pp. 151–172, October 2007.
19 A. Gupta, A. A. Efros, and M. Hebert, "Blocks world revisited: Image understanding using qualitative geometry and mechanics," in *European Conference on Computer Vision(ECCV)*, 2010.
20 L. Shen, P. Tan, and S. Lin, "Intrinsic image decomposition with non-local texture cues," *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, vol. 0, pp. 1–7, 2008.
21 A. Bousseau, S. Paris, and F. Durand, "User assisted intrinsic images," *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia 2009)*, vol. 28, no. 5, 2009.
22 X. Jiang, A. J. Schofield, and J. L. Wyatt, "Correlation-based intrinsic image extraction from a single image," in *Proceedings of the 11th European conference on Computer vision: Part IV*, ser. ECCV'10. Berlin, Heidelberg: Springer-Verlag, 2010, pp. 58–71.
23 H. Barrow and J. Tenenbaum, "Recovering intrinsic scene characteristics from images," in *Computer Vision systems'78*, 1978, pp. 3–26.
24 P. Huang, Y. Gu, X. Wu, Y. Chen, and E. Wu, "Time-varying clustering for local lighting and material design," *Science China Information Sciences*, vol. 52, no. 3, pp. 445–456, 2009.
25 K. Garg and S. K. Nayar, "Photorealistic rendering of rain streaks," in *ACM SIGGRAPH 2006 Papers*, ser. SIGGRAPH '06. New York, NY, USA: ACM, 2006, pp. 996–1002.
26 A. Wenger, A. Gardner, C. Tchou, J. Unger, T. Hawkins, and P. Debevec, "Performance relighting and reflectance transformation with time-multiplexed illumination," in *ACM SIGGRAPH 2005 Papers*, ser. SIGGRAPH '05. New York, NY, USA: ACM, 2005, pp. 756–764.
27 C.-F. Chabert, P. Einarsson, A. Jones, B. Lamond, W.-C. Ma, S. Sylwan, T. Hawkins, and P. Debevec, "Relighting human locomotion with flowed reflectance fields," in *ACM SIGGRAPH 2006 Sketches*, ser. SIGGRAPH '06. New

York, NY, USA: ACM, 2006.

28 P. Peers, N. Tamura, W. Matusik, and P. Debevec, "Post-production facial performance relighting using reflectance transfer," in *ACM SIGGRAPH 2007 papers*, ser. SIGGRAPH '07. New York, NY, USA: ACM, 2007.

29 X. Jin, M. Zhao, X. Chen, Q. Zhao, and S. C. Zhu, "Learning artistic lighting template from portrait photographs," in *ECCV (4)*, 2010, pp. 101–114.

30 X. Chen, M. Chen, X. Jin, and Q. Zhao, "Face illumination transfer through edge-preserving filters," in *In Proceedings of the 24th IEEE International Conference on Computer Vision and Pattern Recognition*, 2011, pp. 281–287.

31 X. Chen, X. Jin, Q. Zhao, and H. Wu, "Artistic illumination transfer for portraits," *Computer Graphics Forum*, vol. 31, no. 4, pp. 1425–1434, 2012.

32 P. Supan, I. Stuppacher, and M. Haller, "Image based shadowing in real-time augmented reality," *IJVR*, vol. 5, no. 3, pp. 1–7, 2006.

33 J. Pilet, A. Geiger, P. Lagger, V. Lepetit, and P. Fua, "An all-in-one solution to geometric and photometric calibration," in *Proceedings of the 5th IEEE and ACM International Symposium on Mixed and Augmented Reality*, ser. ISMAR '06. Washington, DC, USA: IEEE Computer Society, 2006, pp. 69–78.

34 T. Jensen, M. S. Andersen, and C. B. Madsen, "Real-time image based lighting for outdoor augmented reality under dynamically changing illumination conditions," in *GRAPP*, 2006, pp. 364–371.

35 T. Haber, C. Fuchs, P. Bekaert, H.-P. Seidel, M. Goesele, and H. P. A. Lensch, "Relighting objects from image collections," in *CVPR*, 2009, pp. 627–634.

36 K. Karsch, V. Hedau, D. Forsyth, and D. Hoiem, "Rendering synthetic objects into legacy photographs," in *Proceedings of the 2011 SIGGRAPH Asia Conference*, ser. SA '11, 2011, pp. 157:1–157:12.

37 B. T. Phong, "Illumination for computer generated pictures," *Commun. ACM*, vol. 18, pp. 311–317, June 1975.

38 J. J. Moré, "The Levenberg-Marquardt algorithm: Implementation and theory," in *Numerical Analysis*, G. A. Watson, Ed. Berlin: Springer, 1977, pp. 105–116.

39 J.-F. Lalonde, A. A. Efros, and S. G. Narasimhan, "Webcam clip art: Appearance and illuminant transfer from time-lapse sequences," *ACM Transactions on Graphics (SIGGRAPH Asia 2009)*, vol. 28, no. 5, December 2009.

40 A. Saxena, S. H. Chung, and A. Y. Ng, "Learning depth from single monocular images," in *NIPS*, 2005.

41 J. T. Barron and J. Malik, "Shape, albedo, and illumination from a single image of an unknown object," in *CVPR*, 2012.

42 K. Hara, K. Nishino, and K. Ikeuchi, "Determining reflectance and light position from a single image without distant illumination assumption," in *Proceedings of the Ninth IEEE International Conference on Computer Vision - Volume 2*, ser. ICCV '03. Washington, DC, USA: IEEE Computer Society, 2003, pp. 560–.

43 I. Sato, Y. Sato, and K. Ikeuchi, "Illumination from shadows," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, pp. 290–300, 2003.

44 R. Grosse, M. K. Johnson, E. H. Adelson, and W. T. Freeman, "Ground-truth dataset and baseline evaluations for intrinsic image algorithms," in *International Conference on Computer Vision*, 2009, pp. 2335–2342.