Detection of Moving Objects and Cast Shadows Using a Spherical Vision Camera for Outdoor Mixed Reality

Tetsuya Kakuta*
The University of Tokyo

Lu Boun Vinh
The University of Tokyo

Rei Kawakami
The University of Tokyo

Takeshi Oishi
The University of Tokyo

Katsushi Ikeuchi
The University of Tokyo

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

Figure 1: *Process of segmentation and rendering. (a) Input image, (b) foreground extraction, (c) shadow removal, (d) synthesized image.*

Abstract

This paper presents a method to detect moving objects and remove their shadows for superimposing them on Mixed Reality (MR) systems. We cut out the foreground from a real image using a probability-based segmentation method. Using color, spatial, and temporal priors, we can improve the accuracy of the segmentation. Energy minimization is executed by graph cuts. Then we remove the shadow region from the foreground with F-value calculated from the pixel value and the spectral sensitivity characteristic of the camera. Finally we superimpose virtual objects using the stencil buffer, which is used to limit the area of rendering for each pixel. Synthesized images of an outdoor scene show the efficiency of the proposed method.

CR Categories: I.3.7 [Computer Graphics]: Three-Dimensional Graphics and Realism—Virtual Reality; I.4.6 [Image Processing and Computer Vision]: Segmentation—Pixel Classification;

Keywords: mixed reality, augmented reality, foreground extraction, shadow removal

1 Introduction

The objective of Mixed Reality (MR) is to add virtual objects to a real scene. In the video see-through MR systems, virtual objects are superimposed on the digital video stream. However, realistic image composition requires that the virtual objects be correctly occluded by foreground objects.

In this paper, we propose an effective foreground segmentation and shadow removal method to solve the occlusion problem in outdoor MR. As shown in Figure 2(c)(d), we have to estimate the depth of foreground objects and synthesize them in front or back of virtual objects. And also Figure 3 shows the need for removing shadow from the foreground area. We estimate depth of foreground objects from their position in the panoramic image obtained by a spherical vision camera. The advantage of using this kind of camera is that the illuminant information of the real scene is available from a single spherical image. Conventional methods [Sato et al. 1999; Kakuta et al. 2007] need another camera to obtain a spherical image for illuminant estimation. We apply the proposed method to a video sequence and show the accuracy of segmentation in a synthesized image.

The processing flow of the proposed method is described below. First, we extract foreground areas from an input image. Second, the shadow of foreground objects is removed. Third, we estimate the depth of each foreground object and compare these to the depth of virtual objects. Last, the virtual objects are superimposed on an input image in consideration of occlusion.

2 Related work

Several research groups have worked on this occlusion problem. Kanbara et al. and Kim et al. have suggested the stereo vision-based method for estimating depth information of real surroundings [Kanbara et al. 2001; Kim et al. 2003]. The method developed by Gordon et al. can correctly render interaction devices into the scene using foreground mask image extracted by the background subtraction method [Gordon et al. 2002]. Fischer et al. presented a real-time static occlusion handling method for medical intervention using preprocessed visual hull volume [Fischer et al. 2004]. A method for handling occlusion in non-real-time MR was proposed by Lepetit et al. [Lepetit and Berger 2000]. However, it is difficult to apply their method to an outdoor scene because of both the complicated geometry and changing illumination.

The foreground segmentation method from video sequences has been widely developed in the field of computer vision. Recently, accurate and robust segmentation and occlusion detection are achieved by graph cut technique [Boykov et al. 2001]. Kol-

*e-mail: kakuta, lbvinh, rei, oishi, ki@cvl.iis.u-tokyo.ac.jp

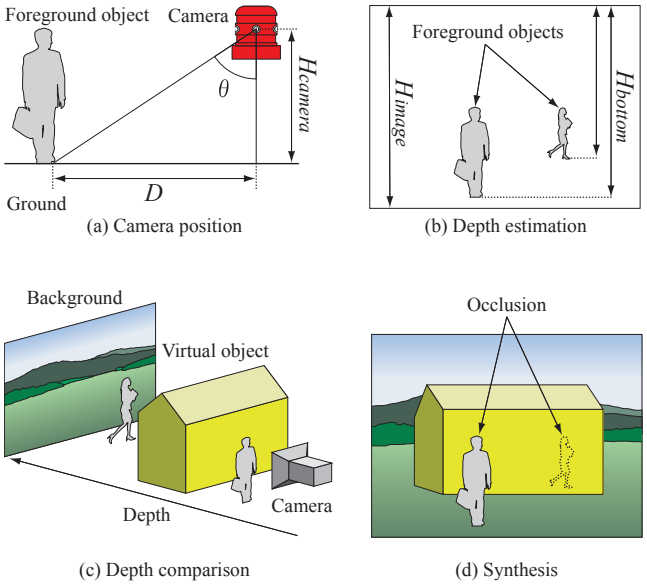

Figure 2: *Depth estimation of foreground objects.*

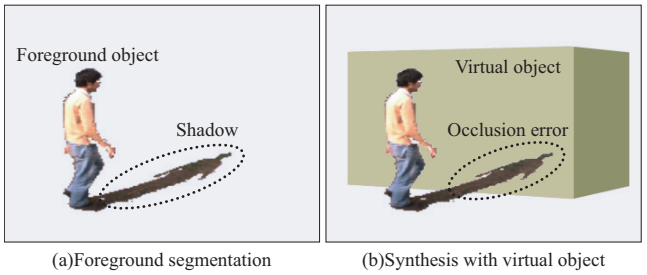

Figure 3: *Problem of shadow region.*

mogorov et al. suggest an algorithm capable of real-time segmentation of foreground from background in monocular video sequences using a probabilistic model based on temporal continuity, spatial continuity, color likelihood, and motion likelihood [Kolmogorov et al. 2006]. Sun et al. develop the real-time foreground layer extraction algorithm using color and contrast cues [Sun et al. 2006]. But the shadow region included in the foreground is not considered in these methods.

The detection of shadow regions is also an important topic in computer vision. In many applications, shadows interfere with segmentation and tracking tasks. The method for detecting and modeling moving cast shadows from a Gaussian mixture model (GMM) is proposed by Martel-Brisson et al [Martel-Brisson and Zaccarin 2005]. Kawakami et al. introduced a method removing the illumination color of outdoor scenes with a single image. However, these methods are not applicable for MR application because of the limitation of computation time.

3 Foreground extraction

In this chapter, we describe the foreground extraction method based on a probabilistic model. Our method is extended from the method proposed by Shiota et al. [Shiota 2007], which is combined Kolmogorov's method [Kolmogorov et al. 2005] and Sun's method [Sun et al. 2006]. It is assumed that the background image is known. In addition to the color and contrast information, the method uses

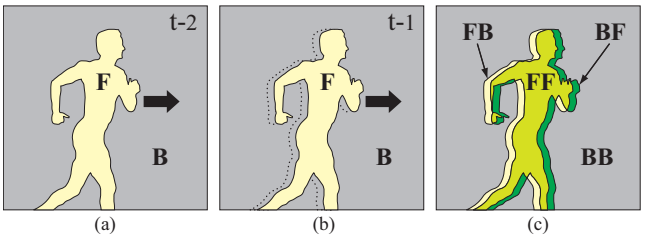

Figure 4: *Temporal transition.*

Figure 5: *Contrast of neighboring pixel-pairs.*

temporal probability as a prior knowledge for improving accuracy.

3.1 Probabilistic model

For the foreground/background segmentation, we should minimize the energy term E^t at time t according to [Shiota 2007].

$$E^t = \alpha V^T(X^t, X^{t-1}, X^{t-2}) + \beta V^S(X^t, I^t) + U^C(X^t, I^t), \quad (1)$$

where $I^t = (I_1^t, I_2^t, \cdots, I_N^t)$ is the image extracted from video sequences at time t and $X = \{x_1, \cdots, x_N\}$ is the label in the input image. V^T is the temporal prior term and V^S is the spatial prior term, and they are defined by the prior probability. U^C is the color likelihood term and can also be defined by the color likelihood observed from an input image. α, β are the weight variables between energies and are given experimentally. We describe the overview of these terms in the remainder of this section.

Temporal prior term: $V^T(X^t, X^{t-1}, X^{t-2})$ is computed under the assumption that the prior knowledge of the labeling is represented by the second-order Markov chain model for temporal domain. The labeling result at a pixel is moving to temporal direction when the foreground object moves in an image as shown in Figure 4. This term imposes a tendency to temporal continuity of segmentation labels.

Spatial prior term: V^S indicates the tendency that the same label is assigned to the neighboring pixels in an image. Generally, the labels are spatially continuous on the foreground area but they are different on the segmentation boundaries. Therefore, we define the energy term V^S due to the contrast between neighboring pixels. Figure 5 shows the diagram of pixel-pairs. When the labels of pixel-pairs are different, the contrast of this pixel-pair becomes larger.

Color likelihood term: $U^C(X^t, I^t)$ is the likelihood that the color is I_n when the label of $n \in \mathcal{V}$ is x_n. The likelihood is computed based on the color distribution of foreground and background. These color distributions are represented by Gaussian Mixture Models [Blake et al. 2004; Kolmogorov et al. 2005] and are estimated by using the Expectation Maximization (EM) algorithm. The color distribution of the background region is learned from the background image I^B. A combined model of the color distribution over an input image and the distribution of each pixel are used to improve the local accuracy and robustness.

3.2 Energy minimization

We compute the labels $(\hat{X}^1, \cdots, \hat{X}^t)$ that minimize the energy $E(X^1, \cdots, X^t, I^1, \cdots, I^t)$ shown in equation 1. However, it is

impossible to compute all $(\hat{X}^1, \cdots, \hat{X}^t)$ at one time for the reason of the limit of computation time. We compute current label $\hat{X}^t$ using old labels $\hat{X}^1, \cdots, \hat{X}^{t-1}$ that are already estimated.

$$\hat{X}^t = \arg\min E(X^t, \hat{X}^{t-1}, \hat{X}^{t-2}, I^1) \tag{2}$$

Optimum label $\hat{X}^t$ can be estimated by graph cut [Boykov et al. 2001]. First, non-directed graph $\mathcal{G} = (\mathcal{V}^{'}, \mathcal{E})$ consisting of two nodes is generated from the input image. Here each edge $\mathcal{E}$ has non-negative weight w_e that corresponds to the energy function. Then we divide the graph $\mathcal{G}$ in order to separate two nodes using edge class $\mathcal{C} \subset \mathcal{E}$. Therefore the graph $\mathcal{G}(\mathcal{C}) = (\mathcal{V}^{'}, \mathcal{E} - \mathcal{C})$ consists of two graph where nodes are separated. The cost of graph cut $|\mathcal{C}|$ is represented as the sum of edge weights belonging to class $\mathcal{C}$. Then we compute the graph cut that minimizes the cost $|\mathcal{C}|$ and obtains the segmentation result $\hat{X}^t$ with this minimum cut.

3.3 Shadow removal

As shown in Figure 3, the shadow in the foreground will cause an occlusion problem when it is overlaid on virtual objects. Therefore, we remove those shadows from a foreground by introducing an illumination-invariant value, which can be obtained from pixel values [Kawakami et al. 2005].

Let us define *chromaticity* as:

$$i_r = \frac{I_R}{I_B} \quad , \quad i_g = \frac{I_G}{I_B}. \tag{3}$$

where I_R, I_G, I_B are the image intensity. Then we assume

$$i_r = s_r e_r \tag{4}$$

$$i_g = s_g e_g \tag{5}$$

where s_c and e_c correspond to the chromaticities of the surface reflectance and the illumination.

Here, we introduce an assumption that blackbody radiation can predict the daylight illumination colors [Finlayson and Schaefer 2001; Judd et al. 1964]. If we assume that the camera sensitivity function is sufficiently narrow and apply Wien's approximation to Planck's formula, the illumination chromaticity (e_r, e_g) can be formulated in a simple manner [Finlayson and Hordley 2001; Marchant and Onyango 2000]:

$$e_r = w e_g^A \tag{6}$$

where $A = \left(\frac{1}{\lambda_R} - \frac{1}{\lambda_B}\right) / \left(\frac{1}{\lambda_G} - \frac{1}{\lambda_B}\right)$, $w = \frac{\lambda_G^{5A} / \lambda_B^{5A}}{\lambda_R^5 / \lambda_B^5}$, and both are constant numbers characterizing the camera. λ_c (where index $c = \{R, G, B\}$) is the center wavelength of the camera sensitivity, which can be obtained using a monochromator and a spectrometer [Vora et al. 1997].

If we substitute Eq. (6) into Eqs. (4) and (5), we obtain:

$$i_r / (i_g)^A = s_r / (s_g)^A \equiv F. \tag{7}$$

The equation means that if the surface color is identical, the value of $i_r/(i_g)^A$ will always be the same, regardless of the illumination color; the pixels of the ground have the same value whether in shadow or sunlight. The value $s_r/(s_g)^A$ is referred to as F [Marchant and Onyango 2000].

We can remove shadowed pixels by using this F-value. If each pixel satisfies the conditions below, the pixel should be shadowed.

1. The difference between F-values of the foreground and background is less than the threshold t_F.
2. The brightness of the pixel of the foreground is less than that of the background.

4 Depth estimation and synthesis

To express correct occlusion, we then estimate the depth of foreground objects. We proposed a handy depth estimation method using a camera whose external parameters are known. Figure 2(a) shows the geometrical setup of the camera. We assume the ground is totally flat and the optical axis of the camera is parallel to the ground. When there are moving objects on the ground, referring to the input image, their depth can be computed from the bottom line of each foreground region as shown in Figure 2(b). If it is supposed that the scene is projected equidistantly by the lens and there is no distortion, we can compute the depth of moving objects D as,

$$D = H_{camera} tan\{\frac{\pi(H_{image} - H_{bottom})}{H_{image}}\} \tag{8}$$

where H_{camera} is the height of the camera, H_{image} is the height of the input image and H_{bottom} is the height of the bottom line of the foreground region.

Figure 2 (c) illustrates the comparison of depth of foreground objects and virtual objects. Then, as shown in Figure 2 (d), we synthesize them by considering their occlusion. At the stage of synthesis, the background image is rendered to the color buffer first. Then a mask image is generated from foreground regions which are in front of virtual objects. The mask image is rendered to the stencil buffer in order to avoid rendering virtual objects to this area. Finally, virtual objects are rendered to the color buffer and we can express collect occlusion. The shading and shadows of virtual objects are computed from the spherical image obtained from a camera using the method proposed by [Kakuta et al. 2007].

5 Experimental result

Our MR-system is based on a spherical vision camera Ladybug2 developed by Point Grey Research Inc. Ladybug2's head unit consists of six CCD cameras, and it enables the system to collect video from more than 75% of the full sphere. The main advantages of using this spherical vision camera are that we can obtain illuminant information at one time.

We set this camera on a tripod in an outdoor scene horizontally using a level gauge. The height of the camera from the ground is 1.5 meters, which is almost as high as a human's eye. In the experiment, a video sequence of an outdoor scene has been obtained with the camera Ladybug2 with a resolution of 1024×512 pixels. We superimpose virtual objects and apply the proposed method to this video sequence offline. The spec of the PC is, OS: Windows XP (SP2), CPU: Core2Duo E6850 3.0GHz, RAM: 4GB, GPU: nVIDIA GeForce8800GTS 640MB.

Figure 1 shows each process of segmentation and rendering. We set the parameters of GMM as $K^b = 15, K^B = 2, K^f = 5$. As shown in Figure 1 (b), we extract foreground area from the input image first (Figure 1 (a)). Then we remove shadow from the foreground area using F-value(Figure 1 (c)). Second, the depth of the foreground is estimated from the segmented image and rendered to the stencil buffer. Finally, we render virtual objects and generate the synthesized image. Figure 1 (d) shows the accurate occlusion of real and virtual objects. The shadow of foreground objects is

correctly removed and the shading of virtual objects is match to the illumination of the scene.

The proposed method can process 897.4 msec/frame on the MR-system noted above. Especially the spatial prior term for segmentation seems to takes much computation time. We are trying to make the process faster and handle the occlusion of MR in real time.

6 Conclusion

We have presented a novel occlusion method for MR. The proposed approach uses a probabilistic model-based segmentation and shadow removal with F-value. The depths of foreground objects are estimated from their position in the panoramic image obtained from a spherical vision camera. Making use of the stencil buffer, we can express the occlusion of foreground objects and virtual objects correctly. The spherical vision camera makes it possible to obtain the illuminant information from a single spherical image. Therefore we can estimate the distribution of illumination at the same time and express correct shading and shadow of virtual objects. We have shown the effectiveness of the approach in an outdoor scene with complex illumination. The result clearly showed the advantage of the occlusion method. Our approach is of value in increasing the reality of synthesized images in MR-systems.

Acknowledgement

This research was, in part, supported by the Ministry of Education, Culture, Sports, Science and Technology, under the program, "Development of High Fidelity Digitization Software for Large-Scale and Intangible Cultural Assets."

References

BLAKE, A., ROTHER, C., M.BROWN, P.PEREZ, AND P.TORR. 2004. Interactive image segmentation using an adaptive gmmrf model. *In Proceedings of European Conference on Computer Vision*.

BOYKOV, Y., VEKSLER, O., AND ZABIH, R. 2001. Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence Vol. 23(11)* (November), 1222 – 1239.

FINLAYSON, G. D., AND HORDLEY, S. D. 2001. Color constancy at a pixel. *Journal of Optics Society of America A. Vol. 18*, No. 2, pp. 253–264.

FINLAYSON, G. D., AND SCHAEFER, G. 2001. Solving for color constancy using a constrained dichromatic reflection model. *International Journal of Computer Vision 42*, 3, 127–144.

FISCHER, J., BARTZ, D., AND STRAβER, W. 2004. Occlusion handling for medical augmented reality using a volumetric phantom model. In *Proc. Symp. on Virtual Reality Software and Technology (VRST'04)*, 174–177.

GORDON, G., BILLINGHURST, M., BELL, M., WOODFILL, J., KOWALIK, B., AND ERENDI, A. 2002. The use of dense stereo data in augmented reality. In *Proc. Int. Symp. on Mixed and Augmented Reality (ISMAR02)*, 14–23.

JUDD, D. B., MACADAM, D. L., AND WYSZECKY, G. 1964. Spectral distribution of typical daylight as a function of correlated color temperature. *Journal of the Optical Society of America 54*, 8, 1031–1040.

KAKUTA, T., OISHI, T., AND IKEUCHI, K. 2007. Real-time soft shadows in mixed reality using shadowing planes. In *Proc. IAPR Conference on Machine Vision Application (MVA2007)*, 195–198.

KANBARA, M., FUJII, H., TAKEMURA, H., AND YOKOYA, N. 2001. A stereo vision-based mixed reality system with natural feature point tracking. In *Proc. Int. Symp. on Mixed Reality (ISMR01)*, 56–63.

KAWAKAMI, R., TAN, R. T., AND IKEUCHI, K. 2005. Consistent surface color for texturing large objects in outdoor scene. In *Proc. Int. Conf. on Computer Vision (ICCV'05)*, 1200–1207.

KIM, H., YANG, S., AND SOHN, K. 2003. 3d reconstruction of stereo images for interaction between real and virtual worlds. In *Proc. IEEE and ACM Int. Symp. on Mixed and Augmented Reality (ISMAR03)*, 169–176.

KOLMOGOROV, V., CRIMINISI, A., BLAKE, A., CROSS, G., AND ROTHER, C. 2005. Bi-layer segmentation of binocular stereo video. *In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) Vol.2* (Jun), 407 – 414.

KOLMOGOROV, V., CRIMINISI, A., BLAKE, A., CROSS, G., AND ROTHER, C. 2006. Bilayer segmentation of live video. *In Proceedings of the 2006 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Vol.1* (June), 53 – 60.

LEPETIT, V., AND BERGER, M. O. 2000. A semi-automatic method for resolving occlusion in augmented reality. In *Proc. Int. Symp. on Mixed and Augmented Reality (ISMAR00)*, 174–177.

MARCHANT, J. A., AND ONYANGO, C. M. 2000. Shadow-invariant classification for scenes illuminated by daylight. *Journal of Optics Society of America A. Vol. 17*, No. 11, pp. 1952–1961.

MARTEL-BRISSON, N., AND ZACCARIN, A. 2005. Moving cast shadow detection from a gaussian mixture shadow model. In *Proc. Conf. on Computer Vision and Pattern Recognition (CVPR'05)*, 643–648.

SATO, I., SATO, Y., AND IKEUCHI, K. 1999. Acquiring a radiance distribution to superimpose virtual objects onto a real scene. *IEEE Trans. on Visualization and Computer Graphics*.

SHIOTA, K. 2007. *Probabilistic model-based foreground extraction and application to 3D shape reconstruction*. Master thesis, The University of Tokyo (in Japanese), March.

SUN, J., ZHANG, W., TANG, X., AND SHUM, H.-Y. 2006. Background cut. *ECCV Vol.2*, 628 – 641.

VORA, P. L., FARRELL, J. E., TIETZ, J. D., AND BRAINARD, D. H. 1997. Digital color cameras - 2 - spectral response. *HP Technical Report*.