3D Reconstruction of Stereo Images for Interaction between Real and Virtual Worlds

Hansung Kim*, Seung-jun Yang** and Kwanghoon Sohn*

*Dept. of Electrical and Electronic Eng., Yonsei University
134 Shinchon-dong, Seodaemun-gu, Seoul, 120-749, Korea
**Omnitel, Inc. 719-24, Yeoksam-dong, Kangnam-gu, Seoul, 135-080, Korea
khsohn@yonsei.ac.kr

Abstract

Mixed reality is different from the virtual reality in that users can feel immersed in a space which is composed of not only virtual but also real objects. Thus, it is essential to realize seamless integration and interaction of the virtual and real worlds. We need depth information of the real scene to synthesize the real and virtual objects. We propose a two-stage algorithm to find smooth and precise disparity vector fields with sharp object boundaries in a stereo image pair for depth estimation. Hierarchical region-dividing disparity estimation increases the efficiency and the reliability of the estimation process, and a shape-adaptive window provides high reliability of the fields around the object boundary region. At the second stage, the vector fields are regularized with a energy model which produces smooth fields while preserving their discontinuities resulting from the object boundaries. The vector fields are used to reconstruct 3D surface of the real scene. Simulation results show that the proposed algorithm provides accurate and spatially correlated disparity vector fields in various kinds of images, and synthesized 3D models produce natural space where the virtual objects interact with the real world as if they are in the same world.

1. Introduction

Mixed reality (MR) is an environment in which the virtual and real environments are composed [1, 2]. In a mixed reality system, users can feel immersed and interact in a space which is composed of not only real objects but also computer-generated objects. Thus seamless integration and natural interaction of the virtual and real worlds are essential for the mixed reality. However, in most conventional MR systems, virtual objects are simply overlaid on the image of the real ones as if the virtual ones are placed in front of real ones. When the real object is placed in front of the virtual one, the image of the virtual object should be pruned before display [3, 4]. Moreover, when a virtual object collides with real objects, it should react on the event properly [5, 6]. We need depth information of the real scene to realize the interaction.

Many active and passive methods have been proposed to recover depth information from real scene. Active techniques utilize ultrasonic or laser to illuminate the work space, so that they yield fast and accurate depth information [7, 8]. However, they are disadvantageous in a limitation of measurement range and hardware cost. Conversely, passive techniques based on computer vision are less sensitive to environments and typically require a simpler and less expensive setup for range sensing. They estimate the depth information from the correspondence of acquired images and camera parameters [3, 4, 9, 10].

When two images are acquired by a stereo camera system, every physical point M yields a pair of 2D projections m_1 and m_2 on two images. If we know both the intrinsic and extrinsic parameters of the stereo system, we can reconstruct the 3D location of the point M from m_1 and m_2 [11, 12].

In the simple case of a parallel camera system as shown in Fig. 1, the depth of a point M can be simply calculated by:

$$Z = f - \frac{fB}{x_2 - x_1}, \tag{1}$$

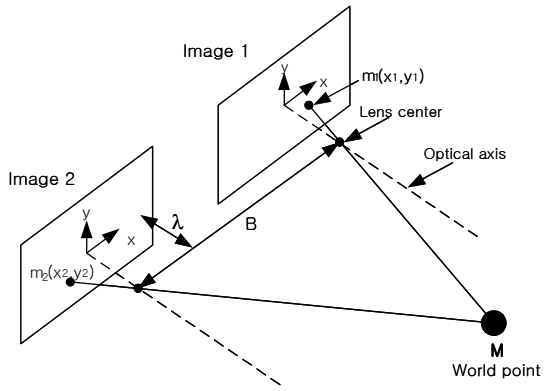


Figure 1. Parallel stereo camera geometry

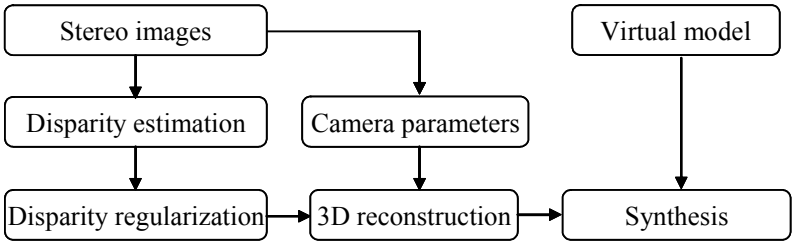


Figure 2. Block diagram of the synthesis system

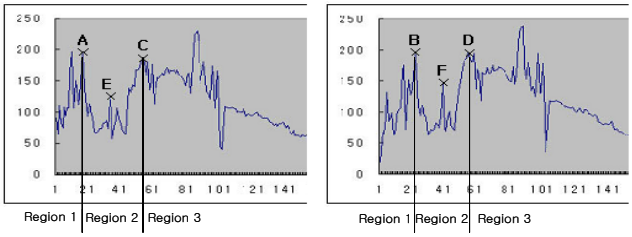
where B is the baseline distance between two cameras and f is the focal length of the camera. We assume the parallel camera geometry in the following process of this paper for simplicity.

One of the most important problems in depth estimation is to find corresponding points in the images, which is called disparity estimation. In this paper, we propose a two-stage algorithm to find smooth and detailed disparity fields with sharp object boundaries in a stereo image pair. The algorithm consists of dense disparity estimation using region-dividing technique and edge-preserving regularization. The estimated disparity fields are converted into depth information with camera parameters and the 3D model of the real scene is reconstructed. By synthesizing the reconstructed model of real scene with virtual models, seamless integration and interaction can be realized. Fig. 2 shows the overall process of the synthesis.

2. Dense disparity estimation

A lot of works have been done on the correspondence problem since the 1970's. In the IEEE Workshop on Stereo and Multi-Baseline Vision 2001, numerous state of the art algorithms for the disparity estimation are presented and evaluated. Recently, D. Scharstein and R. Szeliski provided taxonomy of existing stereo algorithms in their paper [13], and a test bed for the quantitative evaluation of the algorithms in their homepage [14].

In this paper, dense disparity fields of stereo images are obtained hierarchically by using a region-dividing technique and shape-adaptive matching windows which we proposed in Electronic Imaging 03 [15].



(a) Left scanline (b) Right scanline

Figure 3. Scanlines of stereo image pair

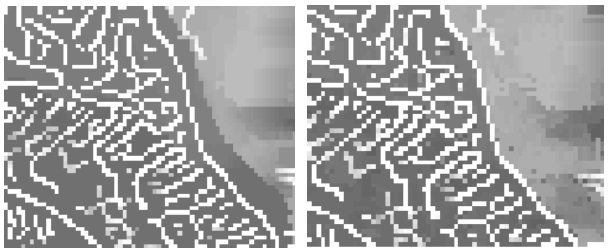
The region-dividing technique is based on the ordering constraint [16]. The technique performs point matching in order of the possibility of correct matching and divides the region into two sub-regions at the true matching point. For example, the Fig. 3 shows corresponding scanlines extracted from a pair of stereo images. If (A, B) and (C, D) are matching pairs, the point E must be matched in the region between B and D according to the ordering constraint. We establish the matching order according to edge intensities, and employ a simple mean absolute difference (MAD) function as the cost function to select the best match from a set of displacement candidates. After the region splits into sub-regions, the search ranges of points in each sub-region are restricted to the corresponding sub-region. It means that if we make an error in dividing regions, the subsequent matching process in the false region produces wrong results. Therefore, true correspondence must be carefully checked in matching process. In order to reject outliers, we perform a bi-directional consistency check for the matching points. According to the uniqueness constraints and the consistency constraints of stereo images [11], the disparity vector $d_l(x)$ estimated from the left image to right image and the corresponding disparity vector $d_r(x+d_l(x))$ estimated from the right image to left image have the same scalar values with opposite directions. With this property, we can evaluate the reliability of the estimation, and increase the whole reliability by eliminating unreliable matching in the region-dividing process. If the matching condition satisfies Eq. (2), the disparity is authorized to the true disparity and the region is divided into two sub-regions. Otherwise, we categorize the point into occlusion and skip to the next points. ε is a matching threshold for bi-directional matching.

$$\left| d_1(x + d_r(x + d_1(x))) \right| =: \delta < \varepsilon \qquad (2)$$

At the first step, we estimate initial disparity vectors in units of $N \times N$ block in low resolution images. The input stereo images are subsampled by a factor of two and split into $N \times N$ rectangular blocks. Then, disparity vector of each block is estimated from the images using the region-dividing block matching.

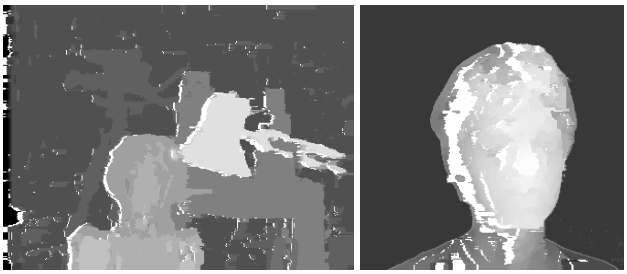
At the second step, based on the initial vectors, dense disparity vectors are estimated using the region-dividing technique with shape-adaptive matching windows in full-resolution images. The shape-adaptive window provides high reliability of the disparity fields around the boundary region by deforming its shape according to the flow of features so that the matching window does not cross strong features of the image. Starting from sufficiently small contour, the contour of the window expands to the direction of non-increasing magnitude of image gradient $|\nabla I|$ until a maximum size $N \times N$ is reached. The shape-adaptive window provides correct sharp object boundary of disparity fields as shown in Fig. 4, where white lines represent real edges of the object.

Fig. 5 shows the estimated disparity maps of the left images from "Head and lamp" and "Man" stereo image pairs used in simulation. White regions represent occlusion, and in case of the "Man" images, we extracted a textureless background region from the image by a simple region-growing technique and unified vectors in the region with the smallest disparity.



(a) Rectangular window (b) Proposed window

Figure 4. Estimation results around object boundary



(a) "Head and lamp" (b) "Man"

Figure 5. Estimated disparity maps

3. Edge-preserving disparity regularization

The disparity vectors estimated by the foregoing estimation method have highly reliable information. However, the spatial correlation of the estimated vector fields is not considered, so that false vectors can be yielded by wrong estimation or error propagation as we can see at the right side of the plaster figure in Fig. 5 (a). Therefore we propose to regularize vector fields by minimizing energy functional involving a fidelity term and a smoothing term such as:

$$E(d) = \int_{\Omega} (I_l(x,y) - I_r(x + d(x,y), y))^2 \, dxdy + \lambda \int_{\Omega} \psi(\nabla d(x,y), \nabla I_l(x,y)) dxdy \quad , \tag{3}$$

where Ω is an image plane, λ is a Lagrange multiplier, and $\psi(\nabla d, \nabla I_l)$ is a potential function whose gradient is given by:

$$\nabla(\psi(\nabla d, \nabla I_l)) = g(|\nabla I_l|^2)\nabla d \quad , \text{where} \quad g(s^2) = \frac{1}{(1+s^2)^2}. \tag{4}$$

We can solve the minimization problem by solving the associated Euler-Lagrange equation and the following corresponding asymptotic state of the parabolic system [17].

$$\frac{\partial d}{\partial t} = \lambda \, div(g(|\nabla I_l(x,y)|^2)\nabla d(x,y)) + (I_l(x,y) - I_r(x+d,y)) \frac{\partial I_r(x+d,y)}{\partial x} \tag{5}$$

This PDE corresponds to the nonlinear diffusion equation with additional reaction term [18], and $g(|\nabla I_l|^2)$ is a diffusivity function which takes a role of discontinuity marker as shown in Fig. 6. Therefore, the diffusion process makes the disparity vector map smooth on continuous surfaces and preserves its discontinuities at the object boundaries.

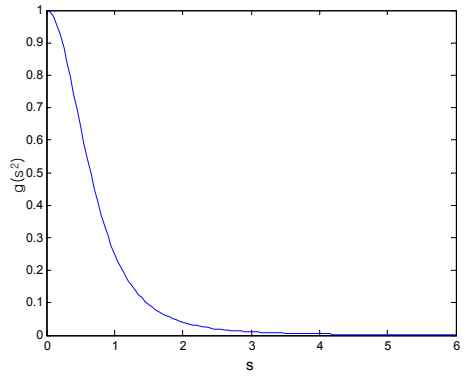


Figure 6. Diffusivity function

In order to solve the Eq. (5), we discretize the parabolic system by finite differences. All spatial derivatives are approximated by forward differences as Eq. (6), and the computationally expensive solution of the nonlinear system is avoided by using the first-order Taylor expansion in an implicit discretization as shown in Eq. (7).

$$\frac{\partial I(x,y)}{\partial x} := (I(x+1,y) - I(x,y))/\delta \tag{6}$$

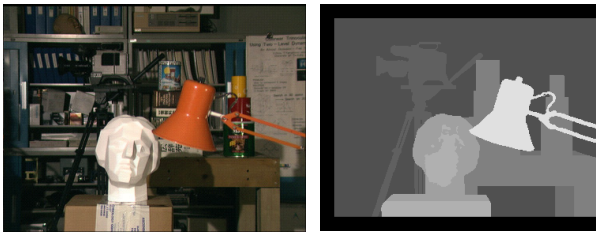
$$\begin{aligned} I(x+d^{k+1},y) &\approx I(x+d^{k},y) \\ &+ (d^{k+1}-d^{k})\frac{\partial I(x+d^{k},y)}{\partial x} + e^{k}(x+d^{k},y) \end{aligned} \tag{7}$$

Then, we can find the regularized disparity field in a recursive way by updating the field with Eq. (8).

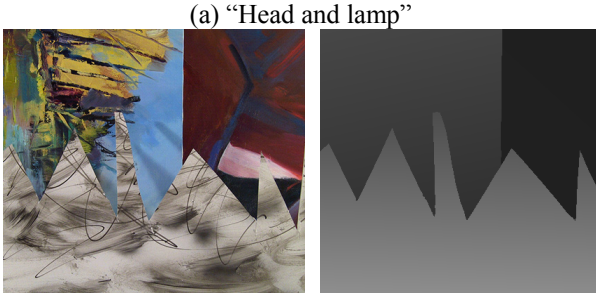
4. Evaluation of the proposed algorithm

Three stereo image pairs with different properties of contents and photographing environments are used to evaluate the performance of the proposed disparity estimation algorithm. A "Head and lamp" image pair shown in Fig. 7 (a) has maximum disparity of 16 pixels in the size of 384×288, and a ground truth disparity map is scaled by the factor of 16 in order to be shown in gray scale. A "Sawtooth" image pair shown in Fig. 7 (b) includes planar objects with much texture information. Their maximum disparity is 20 pixels in the size of 434×380, and the ground truth disparity map is scaled by 8. A "Man," the third image pair shown in Fig. 8, is a video conferencing scene with low textured contents and captured with extremely large baseline distance of 80cm. The maximum disparity of the image pair is 65 pixels in the size of 256×256. However, the only subjective evaluation is performed for the "Man" image pair because the ground truth disparity map is not provided. Actually, the "Man" images were captured by toed-in camera system so that epipolar lines of them are not exactly parallel, but we applied the same horizontal scanline search for simplicity because the images are small and the main object is placed at the center of the image where the epipolar line distortion is not serious.

The parameters used in simulation are listed Table 1. Most parameters are selected intuitively, but the same set



(a) "Head and lamp"



(b) "Sawtooth"

Figure 7. Left images and ground truth disparity maps of test sets

(a) Left image (b) Right image

Figure 8. "Man" stereo image pair

of parameters is used for all the experiments conducted in this section. In the case of gradient step size, we applied different sizes for the gradient of an image and that of disparity maps, because the values of the disparity maps are more sensitive to results than those of images. Finally, iterations for solving the PDE in the regularization stage are 150 times for the "Head and lamp" and the "Man" image pairs, and 600 times for the "Sawtooth" image pair.

Fig. 9 shows estimated disparity maps and differences to the ground truth disparity maps of the "Head and lamp" and the "Sawtooth" images. In the difference images, correct matches appear in medium grey (128), and brighter or darker pixels show the deviation from the

$$\begin{aligned} \frac{d^{k+1}(x,y)-d^{k}(x,y)}{\tau} &= \lambda\left\{\frac{\partial}{\partial x}\left(g\left(\left|\frac{\partial I_l(x,y)}{\partial x}\right|^2\right)\times\frac{\partial d^{k}(x,y)}{\partial x}\right)+\frac{\partial}{\partial y}\left(g\left(\left|\frac{\partial I_l(x,y)}{\partial y}\right|^2\right)\times\frac{\partial d^{k}(x,y)}{\partial y}\right)\right\} \\ &+\left(I_l(x,y)-I_r(x+d^{k}(x,y),y)\right)\times\frac{\partial I_r(x+d^{k}(x,y),y)}{\partial x}+\left(d^{k}(x,y)-d^{k+1}(x,y)\right)\times\left(\frac{\partial I_r(x+d^{k}(x,y),y)}{\partial x}\right)^2 \end{aligned} \tag{8}$$

Table 1. Parameters used in disparity estimation

Stage	Parameter	Values
Disparity estimation	Block and window size	N=8
	Bidirectional matching threshold	ε=1
	Search range of dense disparity	α=2
Disparity regularization	Lagrange multiplier	λ=2000
	Gradient step size	δ_I=3 / δ_d=1
	Time step size	τ=0.0001

ground truth. The proposed algorithm results in so clean map with good discontinuity localization. However, the algorithm fails to find disparity in narrow background such as the area between arms of the lamp and give ride to errors around object boundaries because of the leakage of diffusion.

Fig. 10 is disparity maps of the "Man" image pair estimated by the proposed algorithm. We can see that dense disparity vectors are estimated with high reliability in spite of large displacements. Although epipolar lines are not horizontally parallel in the "Man" images, horizontal scan provides satisfactory results.

We compared the performance of the proposed algorithm with the following 4 algorithms by using root-mean-squared error (RMSE) of the estimated disparity fields.

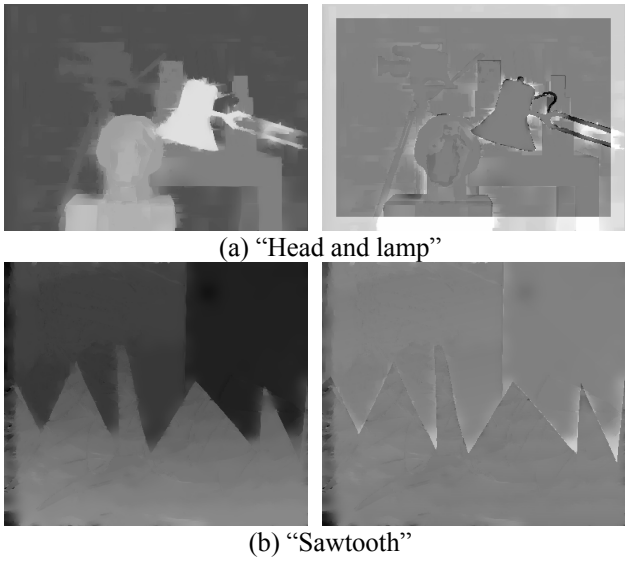
(1) Graph cut [19] – global optimization method based on 2D MRF
(2) Pixel-to-pixel stereo algorithm [20] – advanced scanline method
(3) Cooperative algorithm [21] – iterative method by diffusion
(4) MMHM [22] - correlation based method, fast approach

RMSE between the estimated disparity field $d_e(x,y)$ and the ground truth field $d_T(x,y)$ is calculated by:

$$RMSE = \left(\frac{1}{N} \sum_{(x,y)} (d_e(x,y) - d_T(x,y))^2 \right)^{1/2}. \qquad (9)$$

The proposed algorithm does not deal with a boundary problem so that we exclude a border of 20 pixels in the image in the evaluation.

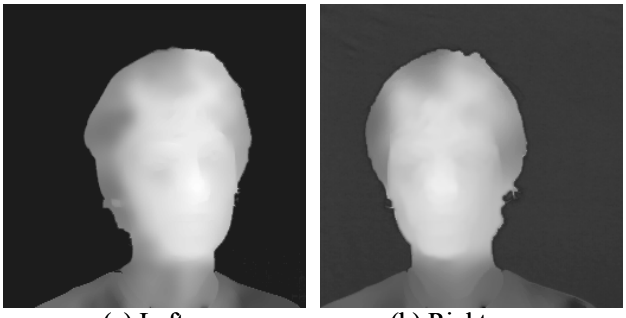
Fig. 11 shows comparative performance of the algorithms. We can see that the proposed algorithm shows the best results.



(a) "Head and lamp"

(b) "Sawtooth"

Figure 9. Disparity maps and difference images



(a) Left map (b) Right map

Figure 10. Disparity maps of the "Man" images

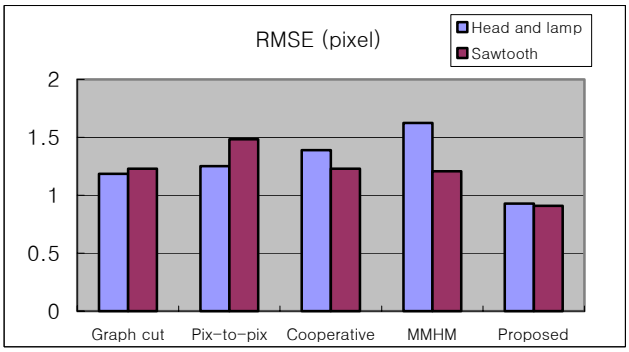


Figure 11. Performance of the algorithms

Finally, we check the computational efficiency of the algorithms by measuring operation time. The computation time of the proposed method on a Pentium IV machine running Windows XP operating system is listed in Table 2.

The MMHM algorithm is the fastest among the comparative algorithms that is near to realtime processing,

and the processing times of scaneline methods are less than 10 seconds. In the case of graph cut algorithm, the processing time is varied from 20 seconds to 700 seconds according to the parameters and options. The running time of the proposed algorithm is acceptable, but it is not proper for real-time processing yet. Improving computational efficiency of the estimation algorithm is the perspective of our work.

Table 2. Computation time

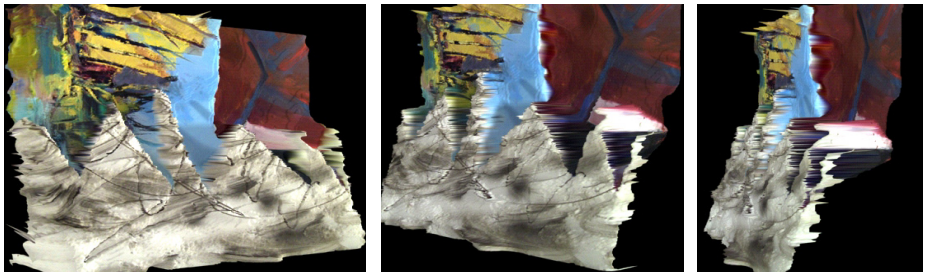
Images	Time (sec)
Head and lamp	6.765
Sawtooth	29.188
Man	5.406

5. 3D reconstruction and synthesis

Dense disparity vectors can be converted into depth information with camera parameters. We assumed the parallel camera system so that the depth information can be easily acquired by Eq. (1).

We reconstructed 3D model of the test image pairs with 3D MAX by using the estimated depth maps as displace maps and original images as diffuse maps. Fig. 12 shows the scenes of the reconstructed models from several viewpoints. Stereo image pairs do not provide entire texture information of the model, so we diffused the texture of the foreground to the background in order to heighten the 3D effect. As we can see in the results, 3D scenes of the test sets are naturally generated. If this

(a) "Head and lamp"



(b) "Sawtooth"

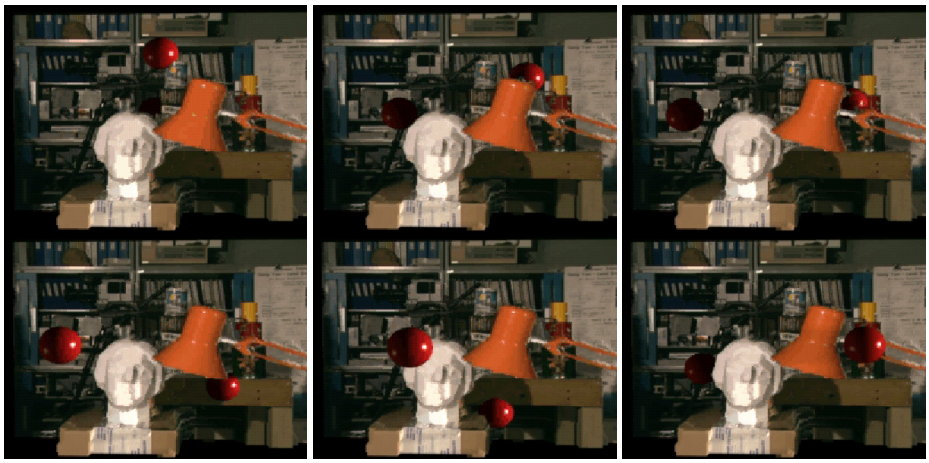
(c) "Man"

Figure 12. Results of 3D reconstruction

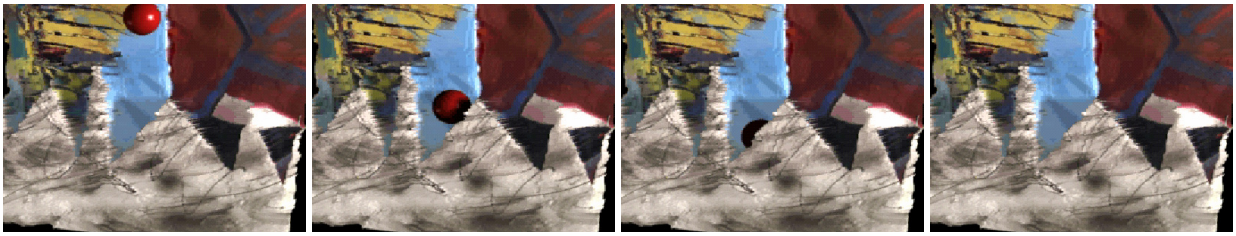
technique is applied to multi-view images, we can reconstruct the nearly complete 3D model by compensating for missing structure and texture information.

We inserted virtual balls to the reconstructed model in order to show the interaction between real and virtual models. Fig. 13 shows snapshots of the synthesized animation. In Fig. 13 (a), two balls revolve round the foreground objects. When the balls move around to the back of the objects, they are gradually concealed by the real objects and disappeared into the space between the fore- and background. Fig. 13 (b) shows a similar situation. When the ball falls to the ground, it disappears naturally behind the front board. Fig. 13 (c) is the example of the active interaction between real and virtual world. The virtual ball is pitched to front board. When the ball strikes on the board, it bounces back to the mirror direction. The outgoing angle to the surface normal vector equals the incoming angle.

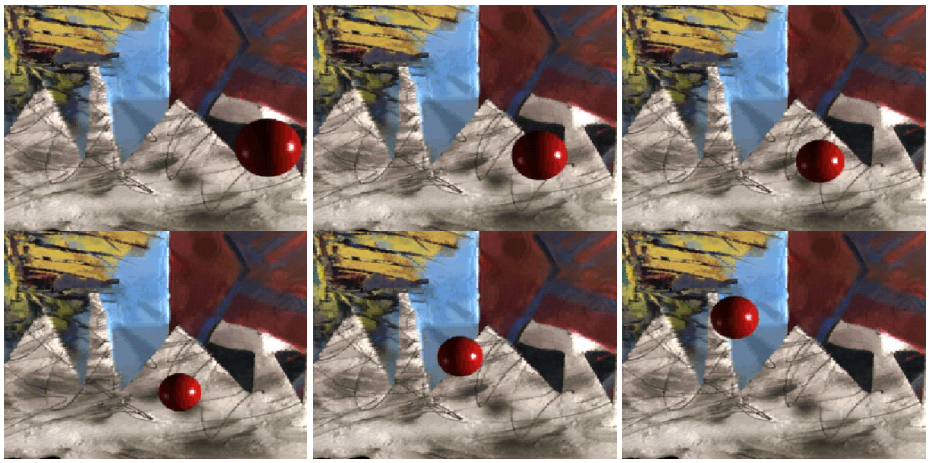
By rendering the synthesized model with virtual stereo



(a) Revolution around the "Head and lamp"



(b) Hiding behind the "Sawtooth"



(c) Bouncing back from the "Sawtooth"

Figure 13. Interaction between real and virtual world

camera and displaying on a stereoscopic display monitor, we verified that the rendered stereoscopic images reproduced a good depth sensation so that we could observe the mixed scene with considerably natural sensation.

6. Conclusion

In this paper, we proposed the depth estimation technique from a stereo image pair for 3D reconstruction in order to solve the mutual occlusion and interaction problem between real and virtual objects in MR system. The proposed two-stage disparity estimation algorithm finds smooth and precise disparity vector fields in a stereo image pair for depth reconstruction. The hierarchical disparity estimation using the region-dividing technique and the shape-adaptive window provides remarkably reliable disparity vectors, and the vector fields are regularized with energy-based edge-preserving regularization technique. As shown in the simulation results, our algorithm provides accurate and spatially correlated disparity vector fields in various environmental images.

We synthesized virtual object models with the reconstructed model of the real world. Rendered scenes show that the real world and the virtual objects interacts each other as if they are in the same world.

The perspective of work will be improving computational efficiency. The proposed algorithm was not integrated into an MR system yet, because the algorithm does not work in real-time. It can be used to reconstruct environment of real scene, but real-time estimation must be accomplished for applying to an MR system. For stereo image sequences, joint estimation with motion tracking can enhance the efficiency. We are also investigating the way to solve the PDE using finite element method (FEM) which can handle complicated geometry, general boundary conditions and nonlinear property more easily and efficiently [23]. Moreover, the FEM-based technique may be coupled to the 3D modeling technique.

It is also planned to adapt the algorithm to the images captured in the non-parallel camera setup with camera calibration. The proposed algorithm only works on image pairs captured in a parallel camera system. We are researching a competent calibration technique and 2 dimensional extensions of all algorithms and equations.

Acknowledgements

We would like to thank Dr. D. Scharstein and Dr. R. Szeliski for supplying the ground truth data on their homepage, and Dr. Y. Ohta and Dr. Y. Nakamura for the imagery from the University of Tsukuba.

This research was supported by University IT Research Center Project.

References

[1] R. T. Azuma, "A survey of Augmented Reality," *Presence*, vol.6, no.4, 1997, pp. 335-385.

[2] S. Benford, C. Greenhalgh, G. Reynard, C. Brown and B. Koleva, "Understanding and Constructing Shared Spaces with Mixed-Reality Boundaries," *ACM Trans. on Computer-Human Interaction*, vol. 5, no. 3, 1998, pp.185-223.

[3] M. Wloka and B. Anderson, "Resolving Occlusion in Augmented Reality," *Proc. Interactive 3D Graphics*, New York, August 1995, pp.5-12.

[4] M. O. Berger, "Resolving Occlusion in Augmented Reality: a Contour Based Approach without 3D Reconstruction," *Proc. CVPR '97*, Puerto Rico, June 1997, pp.91-96.

[5] D. Breen, E. Rose and R. Whitaker, "Interactive Occlusion and Collision of Real and Virtual Objects in Augmented Reality," *Tech. report ECRC-95-02*, 1995.

[6] G. Klinker, K. Ahlers, et al. "Confluence of Computer Vision and Interactive Graphics for Augmented Reality," *Presence: Teleoperations and Virtual Environments*, vol. 6, no. 4, August 1997, pp. 433-451.

[7] S. Feiner, B. Maclntyre and D. Seligmann: "Knowledge-based Augmented Reality," *Comm. of the ACM*, vol. 36, no. 7, 1993, pp.53-62.

[8] G. J. Iddan and G. Yahav, "3D Imaging in the Studio (and elsewhere…)," *SPIE vol. 42983D SMPTE Journal*, 1994.

[9] T. Kanade, A. Yoshida, K. Oda, H. Kano, H. and M. Tanaka, "A stereo machine for video-rate dense depth mapping and its new applications," *Proc. CVPR '96*, 1996, pp.196-202.

[10] M. Kanbara, H. Takemura, N. Yokoya and T. Okuma, "A Stereoscopic Video See-Through Augmented Reality System Based on Real-Time Vision-Based Registration," *IEEE Virtual Reality 2000 Conference,* 2000, pp. 255-262.

[11] O. Faugeras, *Three-Dimensional Computer Vision: A Geometric Viewpoint*, The MIT Press, London, 2001.

[12] E. Trucoo and A. Verri*, Introductory Techniques for 3-D Computer Vision*, ch. 7, Prentice Hall, New Jersey, 1998.

[13] D. Scharstein and R. Szeliski, "A Taxonomy and Evaluation of Dense Two-frame Stereo Correspondence Algorithms," *IJCV,* vol. 47, April-June 2002, pp. 7-42.

[14] http://www.middlebury.edu/stereo

[15] H. Kim and S. Sohn, "Hierarchical Depth Estimation for Image Synthesis in Mixed Reality," *SPIE vol. 5006B Electronic Imaging 03*, Santa Clara, USA, 2003, pp.544-553

[16] A. L.Yuille and T. Poggio. “A Generalized Ordering Constraint for Stereo Correspondence,” *A.I. Memo 777*, AI Lab, MIT, 1984.
[17] G. Aubert and P. Kornprobst, Mathematical Problems in Image Processing: Partial Differential Equations and the Calculus of Variations, ch. 3, Springer, 2002.
[18] J. Weickert, “A Review of Nonlinear Diffusion Filtering,” *Lecture Notes in Comp. Science*, vol. 1252, 1997, pp. 3-28.
[19] Y. Boykov, O. Veksler and R. Zabih, “Fast Approximate Energy Minimization via Graph Cuts,” *IEEE Trans. PAMI* vol. 23, no. 11, 2001, pp. 1222-1239.
[20] S. Birchfield and C. Tomasi, “Depth Discontinuities by Pixel-to-pixel Stereo,” *IJCV*, vol. 35, no. 3, 1999, pp. 269-293.
[21] L. Zitnick and T. Kanade, “A Cooperative Algorithm for Stereo Matching and Occlusion Detection,” *IEEE Trans. PAMI*, vol. 22, no. 7, 2000, pp. 675-684.
[22] K. Mühlmann, D. Maier, J. Hesser and R. Männer, “Calculating Dense Disparity Maps from Color Stereo Images, an Efficient Implementation,” *IJCV*, vol. 47, 2002, pp. 79-88.
[23] C. Johnson, *Numerical solution of partial differential equations by the finite element method*, Cambridge university press, 1990.