Real-time disparity estimation using foreground segmentation for stereo sequences

Hansung Kim
Dong Bo Min
Shinwoo Choi
Kwanghoon Sohn, MEMBER SPIE
Yonsei University
Department of Electrical Engineering
Center for Advanced Broadcasting Technology
Seoul, 120-749, Korea
E-mail: khsohn@yonsei.ac.kr

Abstract. We propose a fast disparity estimation algorithm using background registration and object segmentation for stereo sequences from fixed cameras. Dense background disparity information is calculated in an initialization step, so that only disparities of moving object regions are updated in the main process. We propose a real-time segmentation technique using background subtraction and interframe differences, and a hierarchical disparity estimation using a region-dividing technique and shape-adaptive matching windows. Experimental results show that the proposed algorithm provides accurate disparity vector fields with an average processing speed of 15 frames/s for 320 × 240 stereo sequences on an ordinary PC. *© 2006 Society of Photo-Optical Instrumentation Engineers.* [DOI: 10.1117/1.2183667]

Subject terms: stereo matching; real-time disparity estimation; foreground segmentation; adaptive window.

Paper 040894R received Nov. 26, 2004; revised manuscript received Jun. 10, 2005; accepted for publication Jul. 25, 2005; published online Mar. 22, 2006.

1 Introduction

With the progress of multimedia systems, a rapidly increasing number of researchers are working on 3-D imaging systems. The applications for such a system are obviously plentiful. Immersive video conferencing can enhance the effectiveness of interpersonal communication,[1–3] 3-D TV and display systems increase the impact of news or movies and advertising,[4,5] and the 3-D mixed reality technique enables remote surgery or expert consultancy in the medical areas, and provides a means for remote maintenance in hazardous environments.[6,7]

The most important problem in realizing these kinds of systems is to reconstruct 3-D coordinates of captured scenes. Thus far, many active and passive methods have been proposed to recover depth information from a real scene. Active techniques utilize ultrasonic transducers or lasers to illuminate the work space, so that they yield fast and accurate depth information.[8–10] However, there are limitations to these techniques with respect to measurement range and hardware cost. Passive techniques based on computer vision are less sensitive to environments and typically require a simpler and less expensive setup for range sensing. Those approaches are capable of estimating depth information from acquired images and camera parameters.[5,11]

One of the most important problems in the depth estimation using passive techniques is to find the corresponding pair I_1 and I_2 of a single world point w in two separate image views as shown in Fig. 1. If we assume that the cameras are identical and the coordinate systems of both cameras are aligned in parallel, the determination of the disparity from I_1 to I_2 becomes finding a function $d(x,y)$ such that

$$I_2(x,y) = I_1(x + d(x,y), y). \qquad (1)$$

Considerable effort has been expended on the disparity estimation problem since the 1970s.[12] Recently, Scharstein and Szeliski discussed the taxonomy of existing stereo algorithms in their paper,[13] and Brown et al. reviewed advances in correspondence methods, methods for occlusion, and real-time implementation.[14] Disparity estimation algorithms can be classified into two categories: local methods, including area-based approaches[15,16] and feature-based approaches,[17,18] and global methods, such as energy-based[19,20] and DSI-based[21,22] approaches.

However, most of them have serious limitations in common applications, since many kinds of 3-D imaging system require real-time calculation of disparity fields for dynamic scenes. Most real-time implementations have made use of special-purpose hardware, such as digital signal processors (DSPs) or field-programmable gate arrays (FPGAs).[23–25] With increasing clock speeds, real-time stereo processing has been recently realized on ordinary desktop computers.[26–30] However, they generally show poor quality

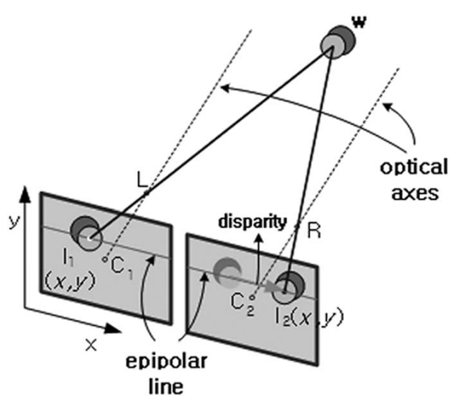


Fig. 1 Stereo geometry.

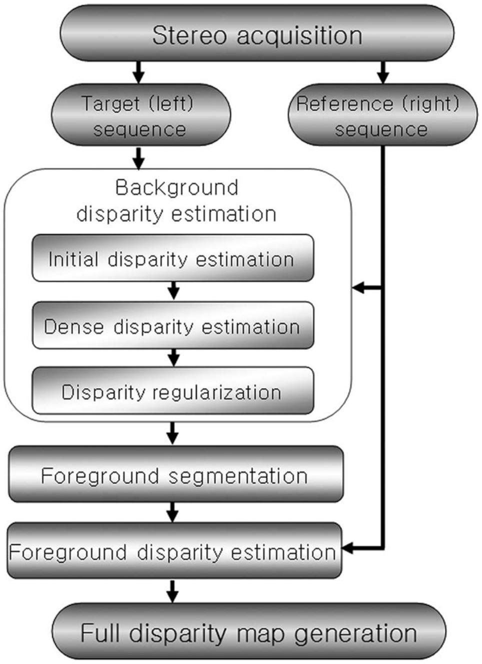

Stereo acquisition
Target (left) sequence
Reference (right) sequence
Background disparity estimation
Initial disparity estimation
Dense disparity estimation
Disparity regularization
Foreground segmentation
Foreground disparity estimation
Full disparity map generation

Fig. 2 Block diagram of the overall process.

in wide-ranging applications, since they use simple area-based algorithms such as the sum of absolute differences (SAD).

We have previously proposed a two-stage algorithm to find smooth and precise disparity vector fields in a stereo image pair.[31] The algorithm comprises dense disparity estimation and edge-preserving regularization. It resulted in a clean disparity map with good discontinuity localization, but the computational cost was so high that it did not work in real time.

In this paper, we propose a fast disparity estimation algorithm using background registration and foreground segmentation. We assume that the stereo camera does not move, and that no object moves for a few seconds in an initialization step for generating background information.

Figure 2 shows a block diagram of the proposed system. Accurate and detailed disparity information for the environment is estimated in advance, and then only disparities of moving object regions are calculated, using a fast algorithm and foreground segmentation. As a preprocessing, acquired image sequences are low-pass filtered to reduce noise effects and then rectified, since we assume that stereo images are captured in parallel stereo cameras for disparity estimation. We use a real-time stereo rectification function provided by Triclops SDK.[32]

The remainder of this paper is organized as follows. Proposed real-time foreground segmentation and disparity estimation techniques are described in Sec. 2 and Sec. 3, respectively. Then, Sec. 4 shows experimental results with evaluation. Finally, concluding remarks and plans for future work are presented in Sec. 5.

2 Foreground Segmentation

Real-time foreground segmentation is one of the most important components in the proposed system, because the performance of the segmentation decides the efficiency and quality of the final disparity fields. Conventional object segmentation algorithms are roughly classified into two categories according to their primary segmentation criteria. The first kind use spatial homogeneity as a criterion. Morphological filters are used to simplify the image, and then the watershed algorithm is applied for region boundary decision.[33,34] The segmentation results of these algorithms tend to track the object boundary more precisely than other methods because they use the watershed algorithm. The main drawback of these algorithms is their high computational complexity, since the watershed is a computationally intensive algorithm. The second approaches make use of change detection such as frame difference[35,36] or background mosaics.[37,38] These algorithms work very fast and distinguish semantic object regions from static background.

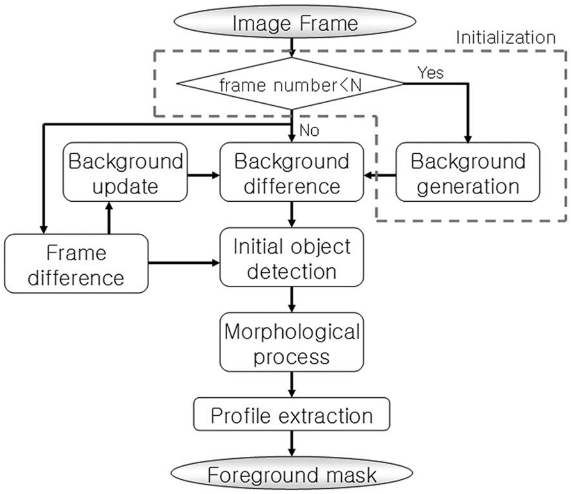


Fig. 3 Proposed foreground segmentation algorithm.

In this section, we propose a foreground segmentation technique based on the second approach, using both background subtraction and interframe differences. Figure 3 shows a diagram of the proposed foreground segmentation algorithm.

First, the background masks $I_{\min}(x,y)$ and $I_{\max}(x,y)$ are modeled with the minimum and maximum intensities of the first N frames, respectively, because the background information is very sensitive to noise and change of illumination. Then, the frame difference mask $I_{\mathrm{fd}}(x,y)$ is calculated as the difference between two consecutive frames. In the third step, an initial foreground mask is constructed from the frame difference and background difference masks by the OR process, that is, if a pixel of current frame satisfies one of the following equations, it is determined to be belonged to an initial foreground region:

$$I_{\mathrm{cur}}(x,y) < I_{\min}(x,y) - \mathrm{Th}_{\mathrm{tol}}, \tag{2a}$$

$$I_{\mathrm{cur}}(x,y) > I_{\max}(x,y) + \mathrm{Th}_{\mathrm{tol}}, \tag{2b}$$

$$I_{\mathrm{fd}}(x,y) > \mathrm{Th}_{\mathrm{fd}}. \tag{2c}$$

Here $\mathrm{Th}_{\mathrm{tol}}$ and $\mathrm{Th}_{\mathrm{fd}}$ mean the threshold values for the background and frame difference regions, respectively.

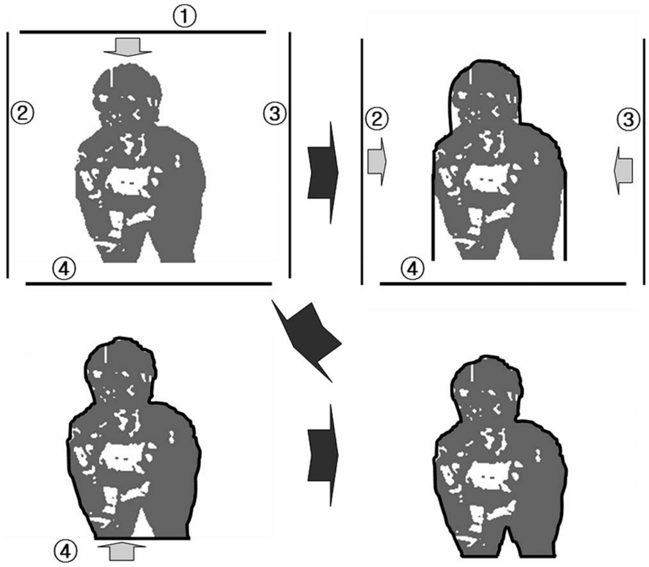


Fig. 4 Profile extraction.

However, due to the camera noise and irregular object motion, there exist some noise regions in the initial mask. One of the conventional ways to eliminate the noise regions is using the morphological operations to filter out small regions. Thus, we refine the initial mask by a closing process and eliminate small regions with a region-growing technique.[39]

Finally, in order to smooth the boundaries of the foreground and to eliminate holes inside the regions, we propose a profile extraction technique. This technique is adapted from Kumar's.[38] A weighted 1-pixel-thick drape moves from one side to the opposite side. The adjacent pixels of the drape are connected by elastic springs, so it covers the object but does not infiltrate into gaps whose widths are smaller than a threshold M. This process is performed on all four quarters, and the region wrapped by the four drapes is decided as a final foreground region. Figure 4 shows the profile extraction process applied to an initial object.

Changes of lighting or background objects, however, are serious problems in many background-registration-based segmentation algorithms. In order to overcome these problems, the long-term behavior of the object motion accumulated from past frames is observed. If a pixel is stationary for the past $\mathrm{Th}_{\mathrm{Bg}}$ frames, then the corresponding pixel and disparity fields in the background buffer are updated by those in the current frame.

Segmentation results by the proposed method are shown in Fig. 5. Figure 5(b) is the result of initial object detection from Fig. 5(a). Main objects are detected well, but they include noises on background and object boundaries. In Fig. 5(c), we can see that noises are eliminated and object surfaces are smoothed by the morphological process. However, many holes still exist inside the objects. Figure 5(d) is the final segmentation result. After applying the profile extraction technique, good semantic foreground regions are obtained.

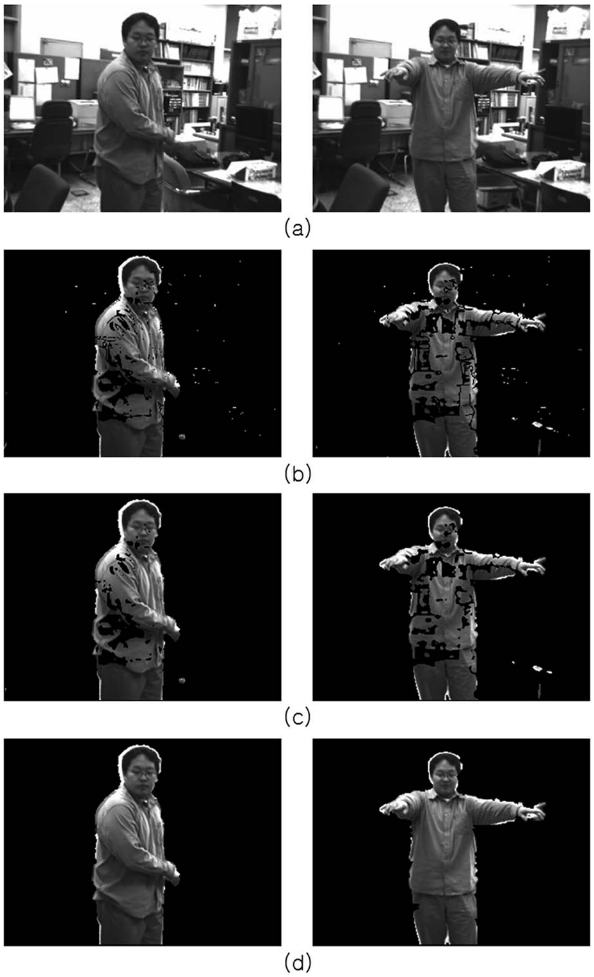


Fig. 5 Segmentation results: (a) original images, (b) initial foreground regions, (c) after morphological process, (d) final foreground regions.

3 Disparity Estimation

In this section, we propose an efficient disparity estimation algorithm to find accurate and detailed disparity fields in a stereo image pair. The resulting disparity map should be smooth and detailed; continuous surfaces should produce smooth disparity fields while preserving the discontinuities that result from object boundaries. In order to satisfy both efficiency and accuracy, we propose a hierarchical approach. Initial disparity vectors of blocks are obtained from downsampled stereo images using a region-dividing disparity estimation technique. With the initial vectors, dense disparities are estimated with shape-adaptive windows in full-resolution images. In the case of background disparity estimation, vector field regularization is additionally performed to provide more detailed and reliable disparity fields.

3.1 *Hierarchical Disparity Estimation with Shape-Adaptive Windows*

At the first level, feature information is extracted from the input images. The Euclidean norm of the gradient,

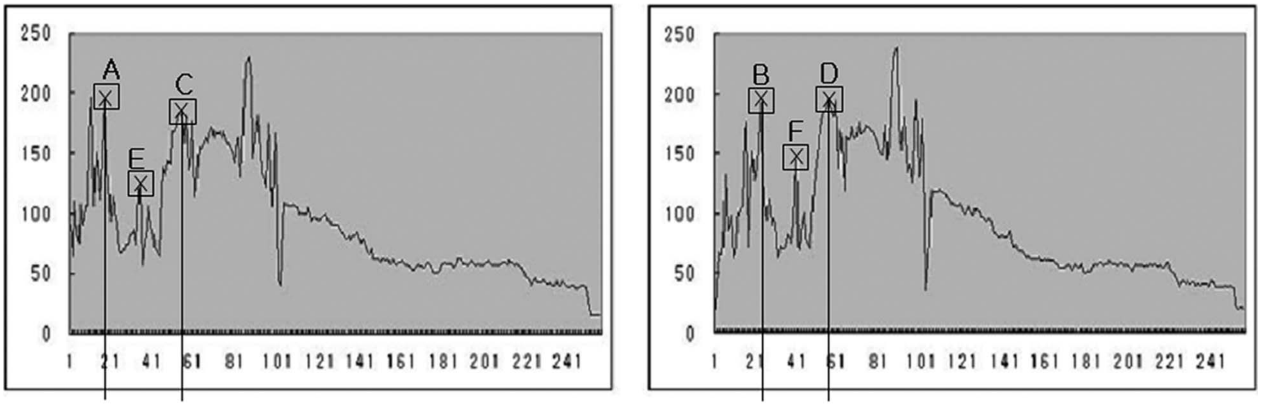


Fig. 6 Region-dividing technique.

$$|\nabla I| = \left[\left(\frac{\partial I}{\partial x} \right)^2 + \left(\frac{\partial I}{\partial y} \right)^2 \right]^{1/2}, \tag{3}$$

is used as a feature, because a gradient operator is better at estimating flows of the feature and detecting object boundaries. This information is also used in regularization process of background disparity generation.

A pair of stereo images is then subsampled by a factor of 2 and split into rectangular blocks $N \times N$ in size. The maximum value of $|\nabla I|$ in the partitioned block represents the feature intensity of the block.

Secondly, initial disparity vectors of blocks are estimated from the subsampled images using a region-dividing block-matching technique.[31] The region-dividing technique is based on the ordering constraint.[40] The technique performs point matching in the order of the possibility of correct matching and divides the region into subregions at the true matching point. It first establishes the matching order according to intensity of the feature of the blocks and performs bidirectional matching[41] in that order. If the block matching satisfies

$$|d_1(x) + d_r(x + d_1(x))| < \varepsilon, \tag{4}$$

the region is divided into two subregions and the search ranges of their blocks are restricted to each subregion. Otherwise, the process does not assign any disparity and skips to the next block.

For example, Fig. 6 shows corresponding scanlines extracted from a pair of stereo images. If (A,B) and (C,D) are matching pairs, 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 by the gradient operator of Eq. (3). We also employ a simple SAD function as a cost function to select the best match from a set of disparity candidates.

Thirdly, based on the initial vectors, dense disparity fields are estimated in full-resolution images. We perform a dense disparity estimation using the region-dividing technique and shape-adaptive windows. In order to cover all the probable disparity candidates, nine initial vectors (one from the current block and eight from neighboring blocks) are tested within a small search range α. When applying the region-dividing technique, unmatched points are considered to belong to an occluded region.

However, in the matching process, conventional rectangular windows yield a false result around strong features because the result is greatly influenced by the feature. For example, in the cases of points A and B in Fig. 7, although they belong to different regions, the same disparity vectors are assigned because of the strong edge between them. In order to avoid this type of problem, we propose a new matching window, which provides a high degree of reliability around the boundary region by deforming its shape according to the flow of the features. Let Ω denote the contour of the matching window. Starting from a sufficiently small contour Ω_0, the contour expands in the direction of nonincreasing $|\nabla I|$ until a maximum size $N \times N$ is reached. Figure 8 shows an example of window generation in the 1-D case. The window does not cross strong features, so that the correct sharp boundary of disparity vectors can be obtained, as shown in Fig. 9, where white lines represent the real edges of the object.

However, the adaptive window may decrease the matching power in highly textured regions. Thus, the shape-adaptive window is applied only for pixels in the block where the maximum difference of disparity from surrounding blocks is larger than ε (the same parameter as the bidirectional matching threshold).

The occlusion labels on each scanline are replaced by the nearest neighboring background disparity when a full disparity map is required.

3.2 *Background Disparity Refinement*

Dense disparity fields of background are initially estimated in the hierarchical way proposed in Sec. 3.1. The disparity

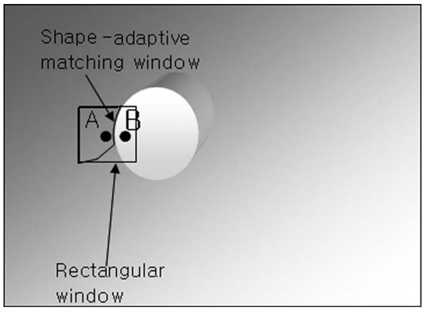


Fig. 7 Rectangular window and the proposed window.

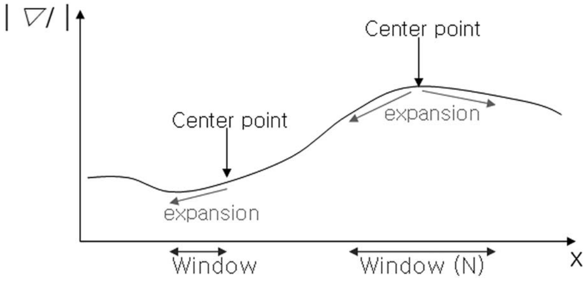


Fig. 8 Window generation in 1-D case.

vectors estimated by that method provide generally reliable information. However, spatial correlation of the estimated vector fields is not considered. The disparity fields of the background are estimated only once (at an initialization step) in the whole process, so we refine the fields in the continuous domain by regularization in order to provide more detailed and reliable background disparity fields. We use the energy-based disparity regularization we have previously proposed.[31] The energy functional consists of a fidelity term and a smoothing term such as

$$E(d) = \int_{\Omega} [I_l(x,y) - I_r(x + d(x,y),y)]^2 \, \mathrm{d}x \, \mathrm{d}y + \lambda \int_{\Omega} \psi(\nabla d(x,y), \ \nabla I_l(x,y)) \, \mathrm{d}x \, \mathrm{d}y, \tag{5}$$

where Ω is an image plane, λ a weighting factor of the smoothing term, and $\psi(\nabla d, \nabla I_l)$ a potential function whose gradient is given by

$$\nabla(\psi(\nabla d, \nabla I_l)) = \frac{1}{(1 + \nabla I_l^2)^2} \nabla d. \tag{6}$$

The minimization problem can be solved by solving the associated Euler-Lagrange equation and the following corresponding asymptotic state of the parabolic system:

$$\frac{\partial d}{\partial t} = \lambda \operatorname{div}\left[\frac{1}{(1 + \nabla I_l^2)^2} \nabla d(x,y)\right] + [I_l(x,y) - I_r(x + d,y)] \frac{\partial I_r(x + d,y)}{\partial x}. \tag{7}$$

This partial differential equation corresponds to the nonlinear diffusion equation with an additional reaction term,[42] and $1/(1+\nabla I^2)^2$ is a diffusivity function, which plays the role of a discontinuity marker, as shown in Fig. 10. Therefore, the diffusion process leads to a disparity vector map with smooth continuous surfaces and preserves its discontinuities at the object boundaries.

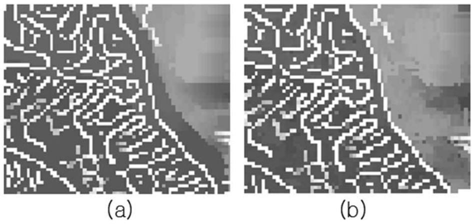


Fig. 9 Matching results using (a) a rectangular window and (b) the proposed window.

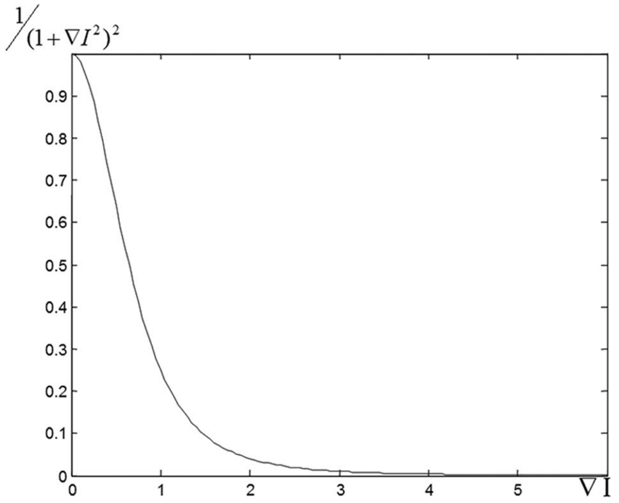


Fig. 10 Diffusivity function.

In order to solve Eq. (7), we discretize the parabolic system by finite differences, and find the regularized disparity field in recursive manner by updating the field using

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

3.3 *Foreground Disparity Estimation*

The most important requirement of foreground disparity estimation is processing speed, because the fields must be updated in every frame.

Hierarchical disparity estimation as in Sec. 3.1 is applied to the blocks that include foreground regions. Initial search ranges are also restricted by the neighbor background disparities, since the foreground objects always exist in front of the background region. The following equations show the search range decision, where $\mathrm{SR}_{\max}$ and $\mathrm{SR}_{\min}$ mean the maximum and minimum search ranges, respectively, and d_{ln} and d_{rn} are the left and right neighboring background disparities of the foreground region on the same scanline:

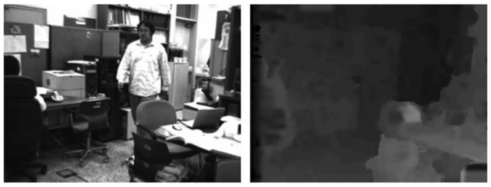


Fig. 11 Test sequence and estimated background disparity.

for $L \rightarrow R$ disparity,

$$\mathrm{SR}_{\max} = \min(d_{ln}, d_{rn}), \tag{9a}$$

for $R \rightarrow L$ disparity,

$$\mathrm{SR}_{\min} = \max(d_{ln}, d_{rn}). \tag{9b}$$

As a result, search ranges are restricted by three factors: background disparity, region-dividing technique, and hierarchical estimation. Thus, the processing time of foreground estimation is greatly reduced.

In background disparity estimation, wrong disparities around boundary regions are corrected by energy-based regularization. However, the regularization process involves such high computational complexity that it cannot be applied to foreground estimation. Moreover, segmentation errors may cause errors around the border of the foreground in the estimation. Therefore, we check the reliability of the disparity for the pixels in boundary blocks that include the boundary between background and foreground. The final disparities of the pixels in boundary blocks are determined by the following conditions, where d_{fore} is an estimated disparity and d_{back} is the disparity of the background at the same position:

$$\begin{aligned} &\text{if } (|I_r(x,y) - I_l(x + d_{\text{fore}}(x,y),y)| < |I_r(x,y) \\ &\qquad - I_l(x + d_{\text{back}}(x,y),y)|) \\ &d_{\text{final}}(x,y) = d_{\text{fore}}(x,y) \\ &\text{else} \\ &d_{\text{final}}(x,y) = d_{\text{back}}(x,y). \end{aligned} \tag{10}$$

4 Simulation Results

Figure 11 shows the left image captured by a stereo camera and the estimated background disparity map. We can see that the proposed algorithm results in a clean map with good discontinuity localization. However, the results from the video are difficult to evaluate objectively, since there is no ground truth. Therefore, we first applied our algorithm to stereo image sets whose ground truth disparity fields are known, then tested it with sequences from a stereo camera.

The parameters used in the simulation are listed in Table 1. Most parameters were selected experimentally, and the same set of parameters was used for all the experiments described in this section.

Table 1 Parameters used in simulations.

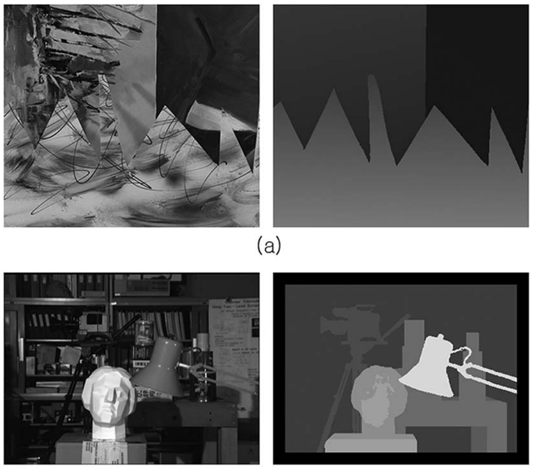
Stage	Parameter	Value
Foreground segmentation	Background generation	$N=50$
	Background difference	$\mathrm{Th}_{\mathrm{tol}}=10$
	Frame difference	$\mathrm{Th}_{\mathrm{fd}}=5$
	Profile extraction	$M=5$
	Background update	$\mathrm{Th}_{\mathrm{Bg}}=300$
Disparity estimation	Block size	$B=8$
	Bidirectional matching	$\varepsilon=1$
	Dense disparity range	$\alpha=2$
Disparity regularization	Lagrange multiplier	$\lambda=2000$
	Time step size	$\tau=0.0001$
	Number of iteration	$T=150$

4.1 *Results from Still Image Sets*

We measured the performance by applying the algorithm to the two still stereo image sets in Fig. 12, provided on Scharstein's home page with ground truth disparity maps.[43] We compared the proposed algorithm with the following four fast algorithms.

1. multiwindow[44]—multiple-window-based method
2. max-surface[45]—3-D maximum-surface techniques
3. real-time DP[46]—real-time dynamic programming
4. MMHM[28]—correlation-based method.

For the objective evaluation of the proposed algorithm, we used two measures of quality. The first is the bad matching percentage (BMP) of the estimated disparity map employed by Zitnick and Kanade,[47] which is defined as



(a)

(b)

Fig. 12 Test image sets: Left image and ground truth disparity of (a) "Sawtooth," (b) "Head and Lamp."

Table 2 Comparison of disparity estimation results.

Algorithm	Bad matching percentage (%) Unoccluded	Bad matching percentage (%) Depth discontinuity	RMSE (pixels), whole region
	(a) “Sawtooth”		
Multiwindow	1.16	19.84	1.2973
Max-surface	4.82	37.68	1.6933
Real-time DP	4.30	34.92	1.7542
MMHM	2.13	25.34	1.2069
Hierarchical	1.20	19.17	1.1340
Regularization	1.24	19.50	0.9032
	(b) “Head and Lamp”		
Multiwindow	3.58	16.84	1.3980
Max-surface	7.49	40.29	1.5294
Real-time DP	2.43	20.97	1.1255
MMHM	6.63	26.03	1.6242
Hierarchical	2.76	20.58	1.0235
Regularization	2.81	19.89	0.9165

$$B=\frac{1}{N}\sum_{(x,y)}\delta(d_e(x,y),d_T(x,y)),\quad \text{where}$$

$$\delta(a,b)=\begin{cases}1 & \text{if } |a-b|>1,\\ 0 & \text{else.}\end{cases}\tag{11}$$

The second is root-mean-squared error (RMSE) of the estimated map. The RMSE between the estimated map $d_e(x,y)$ and the ground truth map $d_T(x,y)$ can be calculated by

$$\text{RMSE}=\left\{\frac{1}{N}\sum_{(x,y)}[d_e(x,y)-d_T(x,y)]^2\right\}^{1/2}.\tag{12}$$

The proposed algorithm and other comparative algorithms do not deal with an image boundary problem; thus most of them show serious errors in the boundary regions. Even the ground truth disparity map of “Head and Lamp” in Fig. 12(b) does not provide true disparity around image boundaries, for that reason. Therefore, a border of 20 pixels was excluded from the evaluation.

Table 2 shows comparative performances of the algorithms. In the table, the “Hierarchical” and “Regularization” rows mean the results before and after regularization, respectively. That is, we can regard the former as the performance of background estimation, and the latter as the performance of foreground estimation, though the effect of the segmentation is not considered. We first measured the BMP in the unoccluded region of the “Sawtooth” and the “Head and Lamp” image pairs in order to appraise the general performance of the proposed algorithm. The unoccluded regions of both images are displayed in Fig. 13(a). As shown in Table 2, the proposed algorithms were good seconds to the multiwindow algorithm with “Sawtooth” and to the real-time DP algorithm with “Head and Lamp.” We then examined the performance in the depth discontinuity region, as shown in Fig. 13(b), in order to test the reliability of the fields around the object boundary region. Table 2 shows that the performance of the proposed algorithms is best with “Sawtooth,” and second best with the “Head and Lamp.” Finally, we measured the RMSE of the images over the entire regions. The results show that the proposed methods efficiently restrict errors, even in mismatched pixels. We can also observe that the regularization process provides very smooth and refined disparity fields, for it improved the performance in RMSE evaluation but not in BMP evaluation.

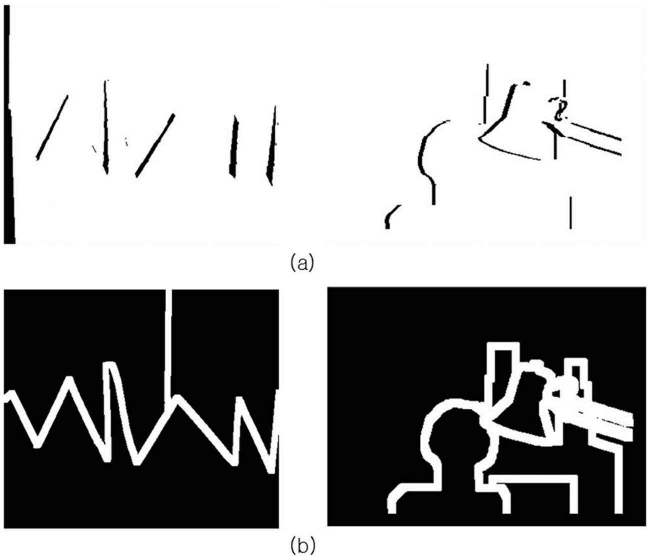


Fig. 13 Test region of the “Sawtooth” and “Head and Lamp”: (a) unoccluded region, (b) depth discontinuity region.

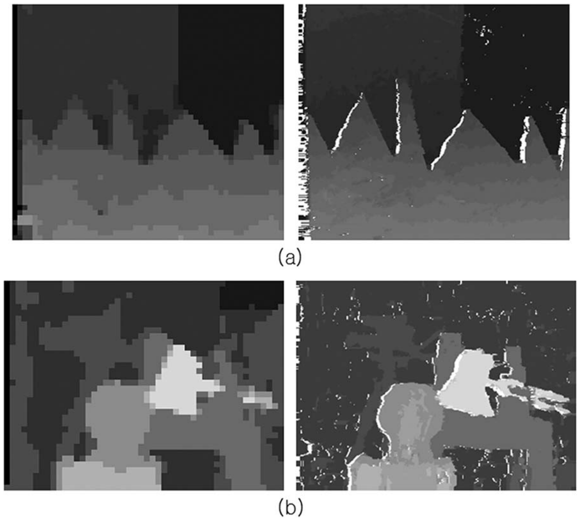


Fig. 14 Initial disparity maps and dense maps with occlusion: (a) “Sawtooth,” (b) “Head and Lamp.”

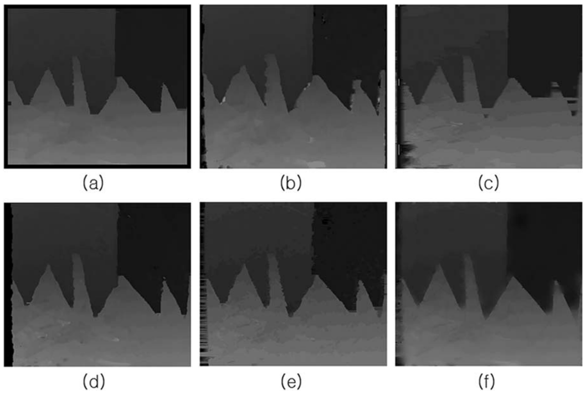


Fig. 15 Disparity maps of "Sawtooth": (a) multiwindow, (b) max-surface, (c) real-time DP, (d) MMHM, (e) hierarchical, (f) regularization.

Figure 14 shows the results of initial disparity estimation and dense disparity estimation with occlusion detection. In both results, we can see that the proposed hierarchical technique produces very reliable disparity vectors.

Figures 15 and 16, respectively, show estimated disparity maps and differences from the ground truth disparity of the "Sawtooth" image. In the difference images, correct matches appear in medium gray (128), and brighter and darker pixels show the extent of deviation from the ground truth. In examining the results, the multiwindow and the real-time DP algorithms are superior in finding discontinuities, but they have problems in error propagation in the horizontal direction. The max-surface algorithm shows a clean map, but serious errors appear around the object boundary region. This indicates that the algorithms have problems in detecting disparity discontinuity.

Figures 17 and 18 show the same types of results for the "Head and Lamp" image pair. The same problems occur in the results of comparative algorithms. The MMHM algorithm shows a good result with "Sawtooth," but produces prominent errors in some regions with "Head and Lamp." The proposed algorithm results in reasonably clean maps with good discontinuity localization. However, the algorithm fails to find disparity in a narrow background such as the area between the arms of the lamp.

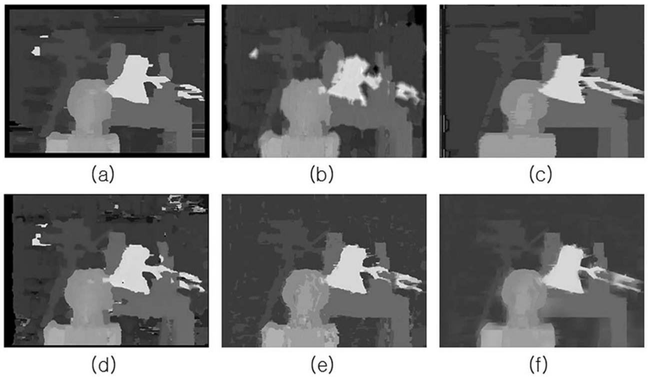


Fig. 17 Disparity maps of "Head and Lamp": (a) multiwindow, (b) max-surface, (c) real-time DP, (d) MMHM, (e) hierarchical, (f) regularization.

4.2 *Results from Stereo Sequence*

The proposed algorithm was applied to stereoscopic sequences captured by Digiclops®, which provides a rectified stereo sequence with a speed of 24 frames/s.[32] The size of images is 320×240, and we simulated the algorithm on a PC with a Pentium IV 3.0-GHz CPU, 512-Mbyte RDRAM, and Visual C++ with Intel C++ compiler 8.1 on a Windows XP operating system.

Figure 19 shows several frames of the resulting sequences; the left column is left images, the middle columns segmented foregrounds, and the right column final disparity fields. The image sequences are captured in absolutely natural condition without any special lighting equipment or any arrangement of objects.

In the segmentation results, we can see that most moving objects are segmented without holes. However, some parts of the background are included in the foreground when an object moves fast or two objects are overlapped in a scene because of frame difference or a profile extraction, respectively. Once in a while, infiltration of background into an object is also observed. In the results of final disparity fields, we can easily imagine the 3-D structure of the scene from the fields.

Table 3 shows the average-running-time analysis of our algorithm when one person moves in a scene. The system requires about 6 to 7 s for initialization before it works.

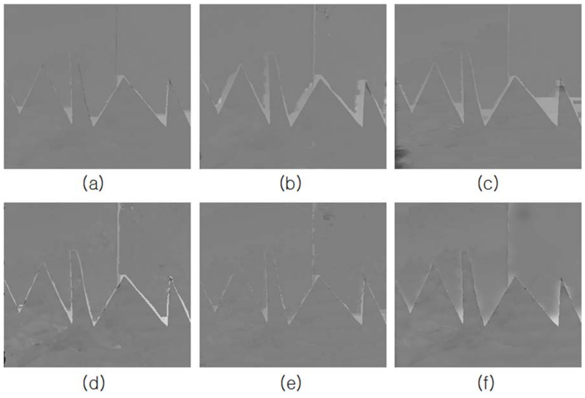


Fig. 16 Difference images of "Sawtooth": (a) multiwindow, (b) max-surface, (c) real-time DP, (d) MMHM, (e) hierarchical, (f) regularization.

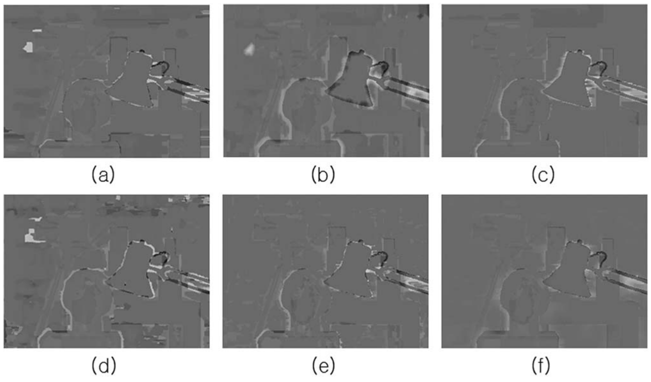


Fig. 18 Difference images of "Head and Lamp": (a) multiwindow, (b) max-surface, (c) real-time DP, (d) MMHM, (e) hierarchical, (f) regularization.

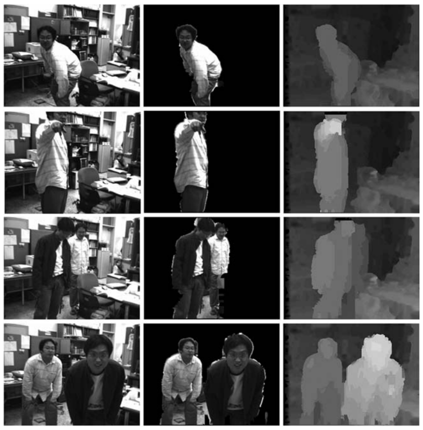


Fig. 19 Results of foreground segmentation and final disparity.

Table 3 Processing speed.

Stage	Step	Time (ms)
Initialization	Background generation	1852
	Background disparity estimation	5156
Main processing	Capturing and rectification	28.26
	Initial segmentation	9.69
	Morphological process	4.64
	Silhouette extraction	6.45
	Disparity estimation	17.81
	Total	66.85

After that, our algorithm shows an average speed of 15 frames/s. According to the referenced papers, multiwindow shows about 5 frames/s, max-surface 2 frames/s, real-time DP 8 frames/s without MMX optimization, and MMHM 5 frames/s.

Considering both processing speed and quality of disparity fields, the proposed algorithm shows the best results.

5 Conclusion

In this paper, we have proposed a near-real-time disparity estimation algorithm using background registration and object segmentation. Dense background disparity information is calculated in advance, and only disparities of moving object regions are updated in the main process. For efficient and accurate estimation, a real-time segmentation algorithm, hierarchical disparity estimation using a region-dividing technique, and shape-adaptive matching windows for disparity estimation are proposed.

The performance of the proposed algorithm was evaluated in objective and subjective ways. The computation time mainly depends on the image size, and it was about 67 ms per image pair having a resolution of 320×240 on an ordinary PC. If we embody the algorithm on a field-programmable gate array (FPGA) or a digital signal processor (DSP), the processing time can be further reduced.

In future work, we have to develop a more powerful segmentation algorithm. The performance of the segmentation decides the efficiency and quality of the final disparity fields. When the object regions are oversegmented, the processing slows down, but the accuracy of the disparity fields is not affected, because the disparity fields of the object regions are recalculated. However, undersegmented foreground regions classified as background due to wrong segmentation lead to serious errors in final fields, since the fields are not updated. The second goal of our work will be to improve the accuracy of disparity fields in object boundary regions.

It is also planned to develop a complete 3-D modeling algorithm from multiple stereo cameras. We are currently investigating a depth-field merging algorithm with camera calibration.

Acknowledgments

We would like to thank Dr. D. Scharstein and Dr. R. Szeliski for supplying the ground truth data on their home page, and Dr. Y. Ohta and Dr. Y. Nakamura for the imagery from the University of Tsukuba. This work was partly supported by the Ministry of Information and Communication, Korea, under the Information Technology Research Center support program supervised by the Institute of Information Technology Assessment and partly supported by the Ministry of Education and Human Resources Development, the Ministry of Commerce, Industry and Energy, and the Ministry of Labor through the fostering project of the Lab of Excellency.

References

1. J. Ohm, K. Gueneberg, E. Hendriks, E. Izquierdo, D. Kalivas, M. Karl, D. Papadimatos, and A. Redert, “A realtime hardware system for stereoscopic videoconferencing with viewpoint adaptation,” in *Proc. IWSNCH3DI ’97*, pp. 147–150 (1997).
2. E. Izquierdo, “Stereo matching for enhanced telepresence in 3D videocommunications,” *IEEE Trans. Circuits Syst. Video Technol.* **7**, 629–643 (1997).
3. O. Schreer and P. Sheppard, “VIRTUE—the step towards immersive tele-presence in virtual video conference systems,” in *Proc. eWorks 2000* (2000).
4. S. Gibbs, C. Arapis, C. Breiteneder, V. Lalioti, S. Mostafawy, and J. Speier, “Virtual studios: an overview,” *IEEE Multimedia* **5**, 18–35 (1998).
5. C. Fehn, E. Cooke, O. Schreer, and P. Kauff, “3D analysis and image-based rendering for immersive TV applications,” *Signal Process. Image Commun.* **17**(9), 705–715 (2002).
6. F. Betting, J. Feldmar, N. Ayache, and F. Devernay, “A new framework for fusing stereo images with volumetric medical images,” in *Proc. CVRMed ’95*, N. Ayache (Ed.), pp. 30–39, Springer (1995).
7. N. Navab, B. Bascle, M. Appel, and E. Cubillo, “Scene augmentation via the fusion of industrial drawings and uncalibrated images with a view to markerless calibration,” in *Proc. 2nd IEEE and ACM Int. Workshop on Augmented Reality*, pp. 125–133 (1999).
8. S. Feiner, B. MacIntyre, and D. Seligmann, “Knowledge-based augmented reality,” *Commun. ACM* **36**(7), 53–62 (1993).
9. G. J. Iddan and G. Yahav, “3D imaging in the studio and elsewhere,” in *Three-Dimensional Image Capture and Applications IV*, *Proc. SPIE* **4298**, 48–55 (1994).
10. M. W. Scott, “Range imaging laser radar,” U.S. Patent No. 4,935,616 (1990).
11. T. Kanade, A. Yoshida, K. Oda, H. Kano, and M. Tanaka, “A stereo machine for video-rate dense depth mapping and its new applications,” in *Proc. CVPR’96*, pp. 196–202, IEEE (1996).
12. W. Förstner, “Image matching,” in R. M. Haralick and L. G. Shapiro Eds., *Computer and Robot Vision*, Vol. **II**, Addison-Wesley (1993).
13. D. Scharstein and R. Szeliski, “A taxonomy and evaluation of dense two-frame stereo correspondence algorithms,” *Int. J. Comput. Vis.* **47**,

7–42 (2002).
14. M. Brown, D. Burschka, and G. D. Hager, "Advances in computational stereo," *IEEE Trans. Pattern Anal. Mach. Intell.* **25**(8), 993–1008 (2003).
15. R. E. H. Franich, "Disparity estimation in stereoscopic digital images," PhD Thesis, Technische Universiteit Delft (1996).
16. S. S. Kumar and B. N. Chatterji, "Stereo matching algorithms based on fuzzy approach," *Int. J. Pattern Recognit. Artif. Intell.* **16**(7), 883–899 (2002).
17. W. E. L. Grimson, "Computational experiments with a feature based stereo algorithm," *IEEE Trans. Pattern Anal. Mach. Intell.* **7**(1), 17–34 (1985).
18. J. Y. Goulermas and P. Liatsis, "Hybrid symbiotic genetic optimisation for robust edge-based stereo correspondence," *Pattern Recogn.* **34**, 2477–2496 (2001).
19. L. Robert and R. Deriche, "Dense depth map reconstruction: a minimization and regularization approach which preserves discontinuities," *Lect. Notes Comput. Sci.* **1064**, 439–451 (1996).
20. L. Alvarez, R. Deriche, J. Sanchez, and J. Weickert, "Dense disparity map estimation respecting image discontinuities: a PDE and scale-space based approach," *J. Visual Commun. Image Represent* **13**, 3–21 (2002).
21. Y. Ohta and T. Kanade, "Stereo by intra- and intra-scanline search using dynamic programming," *IEEE Trans. Pattern Anal. Mach. Intell.* **7**, 139–154 (1985).
22. M. Mozerov, V. Kober, and T. Choi, "Improved motion stereo matching based on a modified dynamic programming," *Opt. Eng.* **40**(10), 2234–2239 (2001).
23. T. Kanade, A. Yoshida, K. Oda, H. Kano, and M. Tanaka, "A stereo machine for video-rate dense depth mapping and its new applications," in *Proc. CVPR'96*, pp. 196–202 (1996).
24. S. Kimura, T. Shinbo, H. Yamaguchi, E. Kawamura, and K. Naka, "A convolver-based real-time stereo machine (SAZAN)," in *Proc. CVPR'99*, Vol. **1**, pp. 457–463 (1999).
25. A. Darabiha, J. Rose, and W. J. MacLean, "Video-rate stereo depth measurement on programmable hardware," in *Proc. CVPR'03*, pp. 203–210 (2003).
26. S. Taraglio and A. Zanela, "Improving a real-time neural-based stereo vision system," *Real-Time Imag.* **7**, 59–76 (2001).
27. O. Schreer, N. Brandenburg, and P. Kauff, "Real-time disparity analysis for applications in immersive teleconference scenarios—a comparative study," in *Proc. ICIAP*, pp. 346–351, IEEE Computer Soc. (2001).
28. K. Mühlmann, D. Maier, J. Hesser, and R. Männer, "Calculating dense disparity maps from color stereo images, an efficient implementation," *Int. J. Comput. Vis.* **47**, 79–88 (2002).
29. L. D. Stefano and S. Mattoccia, "Real-time stereo within the VIDET Project," *Real-Time Imag.* **8**, 439–453 (2002).
30. J. Schmidt, H. Niemann, and S. Vogt, "Dense disparity maps in real-time with an application to augmented reality," in *Proc. WACV*, pp. 225–230, IEEE (2002).
31. H. Kim, Y. Choe, and K. Sohn, "Disparity estimation using region-dividing technique with energy-based regularization," *Opt. Eng.* **43**(8), 1882–1890 (2004).
32. http://www.ptgrey.com/.
33. J. Fan, G. Fujita, M. Furuie, T. Onoye, I. Shirakawa, and L. Wu, "Automatic moving object extraction toward compact video representation," *Opt. Eng.* **39**, 438–452 (2000).
34. S. Chien, Y. Huang, and L. Chen, "Predictive watershed: a fast watershed algorithm for video segmentation," *IEEE Trans. Circuits Syst. Video Technol.* **13**(5), 453–461 (2003).
35. A. Neri, S. Colonnese, G. Russo, and P. Talone, "Automatic moving object and background separation," *Signal Process.* **66**, 219–232 (1998).
36. S. Chien, S. Ma, and L. Chen, "Efficient moving object segmentation algorithm using background registration technique," *IEEE Trans. Circuits Syst. Video Technol.* **12**(7), 577–586 (2002).
37. K. Bhat, M. Saptharishi, and P. Khosla, "Motion detection and segmentation using image mosaics," in *Proc. ICME 2000*, Vol. **3**, pp. 1577–1580, IEEE (2000).
38. P. Kumar, K. Sengupta, and S. Ranganath, "Real time detection and recognition of human profiles using inexpensive desktop cameras," in *Proc. ICPR'00*, pp. 1096–1099, IEEE Computer Soc., (2000).
39. R. C. Gonzalez and R. E. Wood, *Digital Image Processing*, Addison-Wesley (1993).
40. A. L. Yuille and T. Poggio, "A generalized ordering constraint for stereo correspondence," A.I. Memo 777, AI Lab, MIT (1984).
41. E. Izquierdo, "Stereo image analysis for multi-viewpoint telepresence applications," *Signal Process. Image Commun.* **11**(3), 231–254 (1998).
42. J. Weickert, "A review of nonlinear diffusion filtering," *Lect. Notes Comput. Sci.* **1252**, 3–28 (1997).
43. http://www.middlebury.edu/stereo.
44. H. Hirschmüller, "Improvements in real-time correlation-based stereo vision," in *Proc. CVPR stereo Workshop*, pp. 141–148 (2001).
45. C. Sun, "Fast stereo matching using rectangular subregioning and 3D maximum-surface techniques," *Int. J. Comput. Vis.* **42**(1), 7–42 (2002).
46. S. Forstmann, J. Ohya, Y. Kanou, A. Schmitt, and S. Thuering, "Real-time stereo by using dynamic programming," in *Proc. CVPR Workshop*, p. 29, IEEE (2004).
47. L. Zitnick and T. Kanade, "A cooperative algorithm for stereo matching and occlusion detection," *IEEE Trans. Pattern Anal. Mach. Intell.* **22**(7), 675–684 (2000).

Hansung Kim received the BS degree in radio communication engineering in 1998, and the MS and PhD degrees in electronic and electrical engineering from Yonsei University, Korea, in 2001 and 2005, respectively. He is currently a researcher at the Media Information Science Lab at Advanced Telecommunications Research Institute International, Japan. His research interests include three-dimensional image processing, computer vision, and mixed reality. He is a member of IEEE, SPIE, ACM, Korean Society of Broadcast Engineers, and Institute of Electronics Engineers of Korea.

Dong Bo Min received the BS degree from the Department of Electrical and Electronics Engineering in 2003 and the MS degree in electronic and electrical engineering from Yonsei University, Korea, in 2005. He is currently pursuing his PhD in electronic and electrical engineering. His research interests include stereo vision, 3-D modeling, and view synthesis.

Shinwoo Choi received the BS degree from the Department of Electrical and Electronics Engineering in 2004 and the MS degree in electronic and electrical engineering from Yonsei University, Korea, in 2006. His research interests include stereo vision, 3-D modeling, and camera calibration.

Kwanghoon Sohn received the BE degree in electronics engineering from Yonsei University, Seoul, Korea, in 1983, the MSEE degree in electrical engineering from University of Minnesota in 1985, and the PhD degree in electrical and computer engineering from North Carolina State University in 1992. He was employed as a senior member of the research staff in the Satellite Communication Division at Electronics and Telecommunications Research Institute, Daeduk Science Town, Korea, from 1992 to 1993. Also, he was employed as a postdoctoral fellow at the MRI Center in the Medical School of Georgetown University. He is currently a professor in the School of Electrical and Electronic Engineering at Yonsei University. His research interests include three-dimensional image processing, computer vision, image communication, and neural networks. Dr. Sohn is a member of IEEE, the Korean Institute of Communications Science, and the Korean Institute of Telematics and Electronics.