

Implementation of Disparity Information-based 3D Object Tracking

Jung-Hwan Ko, Yong-Woo Jung, and Eun-Soo Kim*

Abstract

In this paper, a new 3D object tracking system using the disparity motion vector (DMV) is presented. In the proposed method, the time-sequential disparity maps are extracted from the sequence of the stereo input image pairs and these disparity maps are used to sequentially estimate the DMV defined as a disparity difference between two consecutive disparity maps Similarly to motion vectors in the conventional video signals, the DMV provides us with motion information of a moving target by showing a relatively large change in the disparity values in the target areas. Accordingly, this DMV helps detect the target area and its location coordinates. Based on these location data of a moving target, the pan/tilt embedded in the stereo camera system can be controlled and consequently achieve real-time stereo tracking of a moving target. From the results of experiments with 9 frames of the stereo image pairs having 256×256 pixels, it is shown that the proposed DMV-based stereo object tracking system can track the moving target with a relatively low error ratio of about 3.05 % on average.

Keywords : stereo object tracking, disparity motion vector, pan/tilt

1. Introduction

Many researches have been done on real time tracking technology [1-3] of moving target in many different related fields in for the last few decades, and recently, the interests in this technology has heightened with the advancement of the industrial and military thehnology develps and the increasing demand for automatated systems [1]. In 1998, T. Darrell et al. [2] presented a visual person tracking system using multi-modal integration, which consists of three primary visual processing modules: depth estimation, color segmentation and intensity pattern classification. T. Joshi et al. [3] also addressed the problem of estimating structure and motion of a smooth curved object from its silhouettes observed over time by a trinocular stereo rig under perspective projection in 1999. More recently, M. Han and T. Kanade [4] described a reconstruction method of multiple motion scenes from uncalibrated views and Lee et al. [5,6] suggested an opto-digital stereo object tracking system using block matching algorithm and optical correlator.

Principally, the stereo object tracking system can perceive three-dimensional (3D) feeling, just like a human eye. This is possible due to the binocular disparity [7-9] between the left and right eyes. But, if there is a disparity between the view point of a target object and the focusing points of the left and right eyes then, it will cause double vision, which in turn case the eyes to feel fatigue. Therefore, when there is stereo disparity in the target object, it should be removed and this process is called the convergence angle control. For this the stereo object tracking system requires two skills: the first one is the convergence angle control which removes the stereo disparity of tracking object while keeping the binocular disparity and the second is the camera's pan/tilt control which always ensure that the moving object is centralized on the camera's field of view (FOV) [10,11].

In this paper, a new stereo object tracking system is proposed, in which a moving object can be detected and tracked in real-time by using only the disparity motion vectors (DMV). Some experiments on the proposed DMV-based stereo object tracking using 9 frames of the stereo input image pair were carried out and the performance analysis will be discussed herein.

Manuscript received November 15, 2005; accepted for publication December 10, 2005.
* Member, KIDS.
This research was supported by the MIC(Ministry of Information and Communication), Korea, under the ITRC(Information Technology Research Center) support program supervised by the IITA(Institute of Information Technology Assessment)(IITA-2005-C1090-0502-0038).
Corresponding Author : Jung-Hwan Ko
Dept. of Electronic Eng., Kwangwoon University 447-1 Wolge-Dong, Nowon-Gu, Seoul 139-701, Korea.
E-mail : misagi@explore.kw.ac.kr **Tel** : +02 940-5118 **Fax** : +02 941-5979

2. Disparity Motion Vector-based Stereo Object Tracking

Fig. 1 shows a process flowchart of the proposed DMV-based stereo object tracking system, which consists of 3 steps: disparity estimation, detection of moving object & its location coordinates and stereo object tracking. In the first step, disparity maps are extracted from the sequence of the stereo input image pair by using a disparity estimation algorithm. In the second step, the disparity motion vectors are sequentially estimated from these disparity maps, and by analyzing the local changes of the disparity values in the DMV, the locational coordinates of the moving target can be extracted. In the third step, these locational data of the moving target are used to controll the pan/tilt em-bedded in the stereo object tracking system so as to achieve stereo tracking of a moving target.

In this paper, the disparity estimation scheme is used to extract the target and its locational coordinates for stereo tracking of a moving target. This disparity estimation may be considered as an extension of the conventional motion estimation algorithm for compressing the video sequences [12]. For that reason, instead of transmitting the stereo image pair, only one of the stereo image pair and its disparity vector need to be transmitted through the communication channel as a compressed image data as shown in Fig. 1, and in the receiver, the other image can be reconstructed using these disparity vector. Furthermore, multiview stereo images can be synthesized to obtain a real 3D display through the so-called intermediate view reconstruction (IVR) scheme [13]. Eq. (1) shows the MAD function to be used to extract the disparity vector from the input stereo image pair, where N_x, N_y and $I_L(m,n)$, $I_R(m+i,n+j)$ represent the matching block size in the x and y directions and the left and right images in the coordinates of (m, n) and $(m+ d_i, n)$, respectively. In reality, the human visual system(HVS) mainly responds to the horizontal disparity, and does not respond to vertical disparity as much. Thus, in this paper, we will consider only the horizontal disparity.

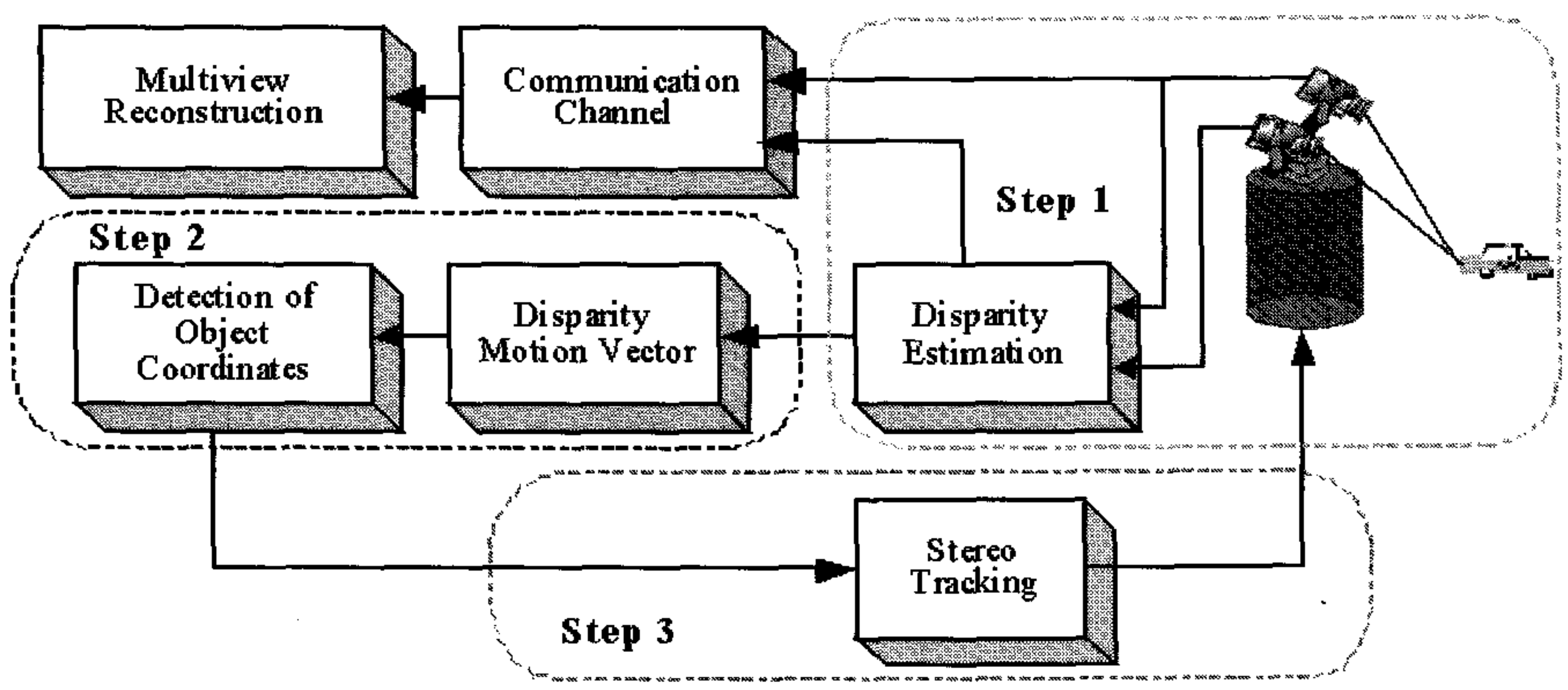


Fig. 1. The proposed DMV-based stereo object tracking system.

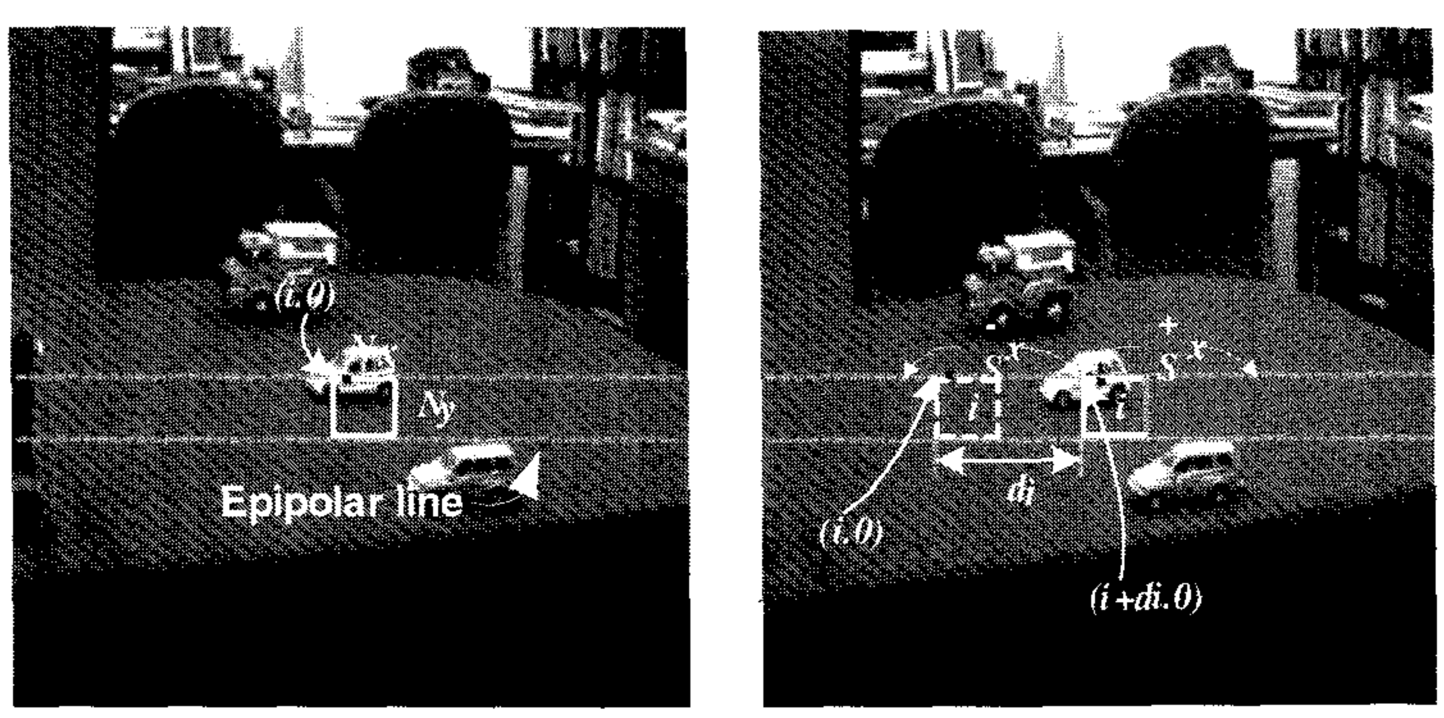


Fig. 2. Corresponding block on epipolar line.

$$\frac{1}{N_x N_y}\sum_{m=1}^{N_x}\sum_{n=1}^{N_y}\left|I_L(m,n)-I_R(m+d_i,n)\right| \quad (1)$$

Fig. 2 shows a stereo image pair to show the process of finding the correspondent matching block or pixel in the right image for an arbitrary block or pixel of the left image on the same homologous epipolar line. For the correspondent matching block or pixel, an optimal disparity vector d_i can be obtained by using the same matching criterion, the MAD function, in the search ranges of $\{(m,n): -S_x \le m \le S_x, n\}$ as shown in Eq. (2)

$$d_i = \arg\{\min_{(m,n)}[MAD(m,n)]\} \quad (2)$$

Once the disparity maps are extracted from the process in Step 1, the DMV, which is defined as a disparity difference between two consecutive disparity vectors of T-1 and T frames, is sequentially estimated from these disparity maps. Basically, this DMV experiences, relatively large changes in the disparity values in the regions where a target object is located before and after moving. On the other hand, it experiences almost no change in disparity values in the regions where there is no movement, such as in the background. As a result, the DMV can provide us with motion information of a moving target by showing a large disparity difference between the disparity vectors of T-1 and T frames in the target areas.

That is, by calculating the disparity differences between two consecutive disparity maps of T-1 and T frames, the DMV maps can be sequentially estimated. From each of these DMV maps, it can be confirmed that the areas having a large change in disparity values is the result of the target's movement. And, these areas are assumed to be the potential locations of the moving target being tracked. In this study, we assumed that there is only one moving target and there is almost no change in the background. Under this circumstance, two areas where a relatively large change of disparity values might be found in the DMV map, and each of these two areas represents the potential place where the target is located in the T-1 and T frames, respectively. One is the area where the target is in the T-1 frame and the other one is the area where the target is in the T frame.

Now, the target area of the T-1 and T frames need to be determined and in which direction the target will be moving. This can be done through the following processes. First, two areas extracted from the DMV is mapped onto the disparity maps of T-1 and T frames, respectively. Then two potential target areas are marked in each disparity map. That is, four potential target areas in total are finally detected from these two disparity maps (herein after referred to as "candidate areas"). Among these, two are the real target areas in each frame and the others are the imaged target areas from the other frame (herein after referred to as "false areas"). By analyzing these 4 areas, the right target area in each frame can be determined.

Fig. 3 shows the finally detected four candidate areas in detail. Here, two pairs of *Cand1, Cand2* and *Cand1', Cand2'* denote the respective candidate target areas in each frame of T-1 and T, and $D_{T-1}(x_1, y_1)$, $D_{T-1}(x_{1+i}, y_{1+j})$ & $D_{T-1}(x_m, y_n)$, $D_{T-1}(x_{m+i}, y_{n+j})$ and $D_T(x_1, y_1)$, $D_T(x_{1+i}, y_{1+j})$ & $D_T(x_m, y_n)$, $D_T(x_{m+i}, y_{n+j})$ mean the location coordinates of the start and end points of each candidate area, *Cand1, Cand2, Cand1'* and *Cand2'*, respectively.

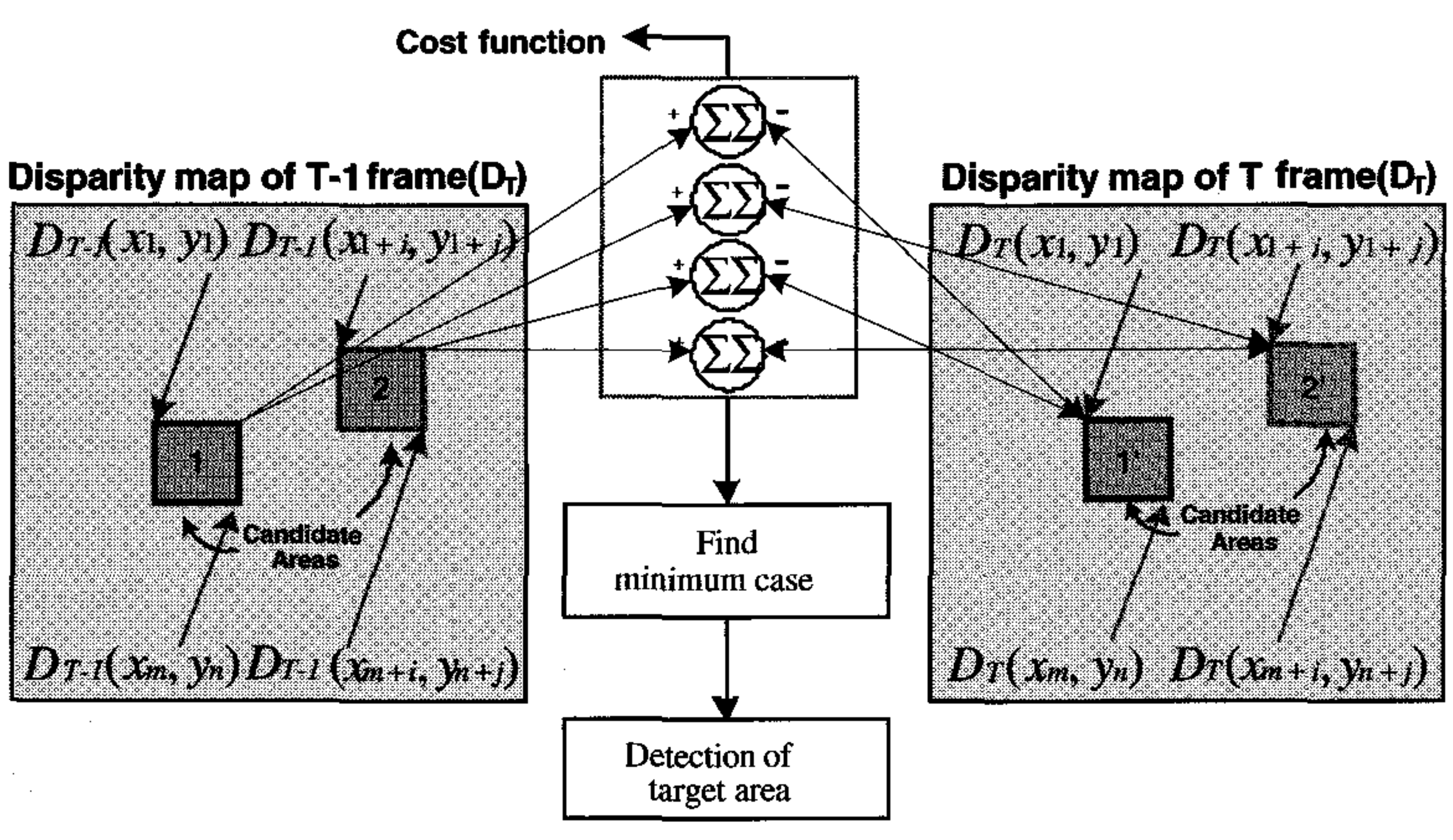


Fig. 3. Detection of a moving object from the candidate areas.

Fig. 4 shows an overall flowchart for finding the real target area and its locational coordinates in each frame and also the moving direction of the object. In general, the real target area and the false area in each frame can be distinguished from each other by calculating the similarities in disparity data between two candidate areas in the T-1 frame and the other two candidate areas in the T frame. Specifically, there is a higher correlation in disparity data between the real target areas in each frame than the other cases. Therefore, a pair of the candidate areas having a relatively higher correlation value between their disparity data is taken as the real target areas in each frame of T-1 and T and the other two areas are discarded as the false areas. Here, the MAD function is also used as a cost function to detect the real target area in each frame within the candidate areas that which has been used for disparity estimation in the previous section. That is, by calculating 4 cases of the cost functions as shown in Eqs. (3)~(6), which are formulated by the pair combination between the candidate areas (*Candl, Cand2*) in T-1 frame and the candidate areas (*Candl′, Cand2′*) in T frame, one case is selected to make this cost function value the smallest among them and then, a pair of the candidates areas to form this cost function is decided as the real target areas in each frame of T-1 and T. Once the target areas are determined, the locational coordinates of the target object in each frame and its relative moving distance between two consecutive frames can be obtained directly.

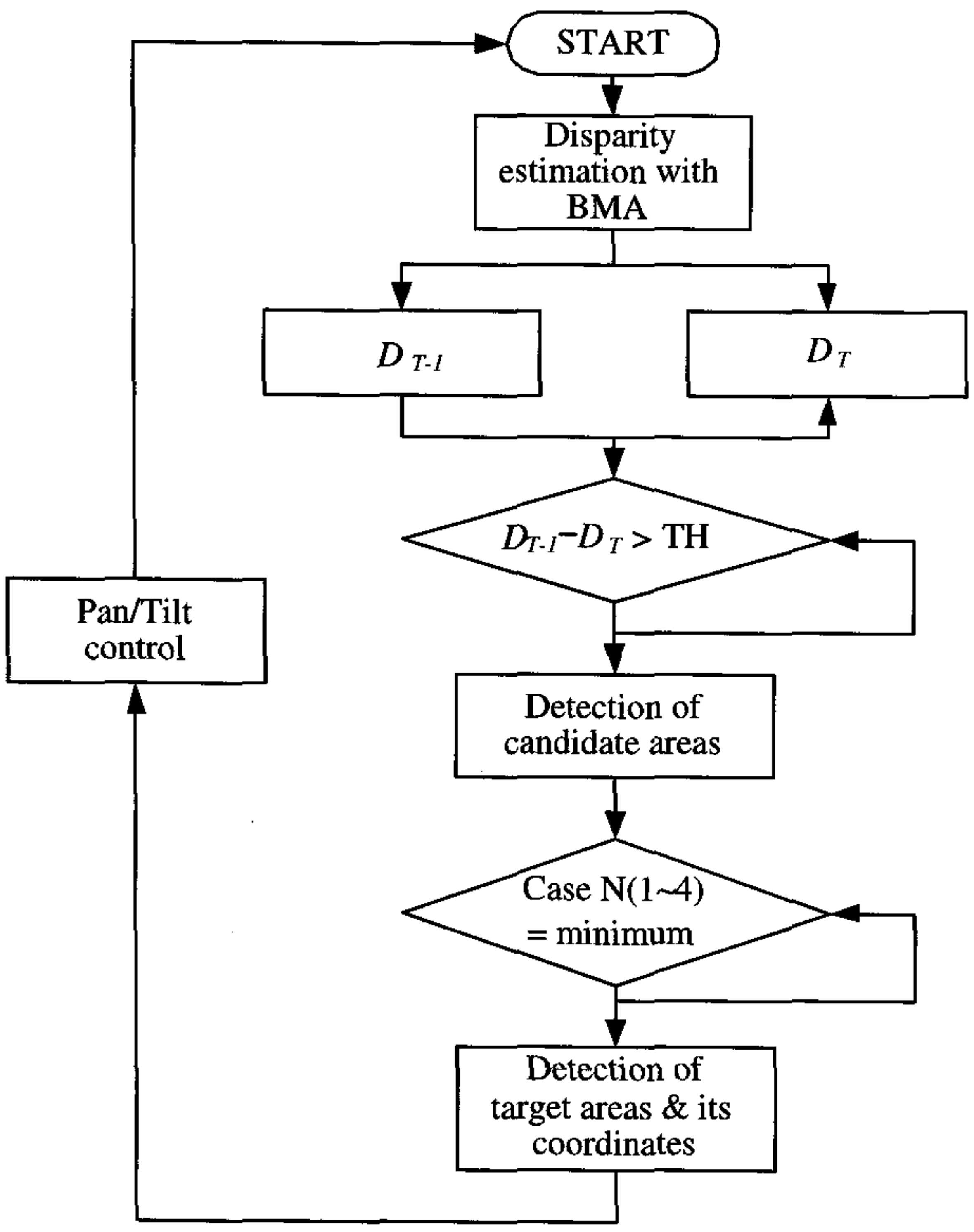


Fig. 4. Flowchart for detection of the moving target and its locational coordinates.

Case1:

$$Cand1 - Cand1' = \frac{1}{M \times N}\sum_{m=1}^{M}\sum_{n=1}^{N}\Big|\,|D_{T-1}(x_m, y_n)| - |D_T(x_m, y_n)|\,\Big| \quad (3)$$

Case2:

$$Cand1 - Cand2' = \frac{1}{M \times N}\sum_{m=1}^{M}\sum_{n=1}^{N}\Big|\,|D_{T-1}(x_m, y_n)| - |D_T(x_{m+i}, y_{n+j})|\,\Big| \quad (4)$$

Case3:

$$Cand2 - Cand1' = \frac{1}{M \times N}\sum_{m=1}^{M}\sum_{n=1}^{N}\Big|\,|D_{T-1}(x_{m+i}, y_{n+j})| - |D_T(x_m, y_n)|\,\Big| \quad (5)$$

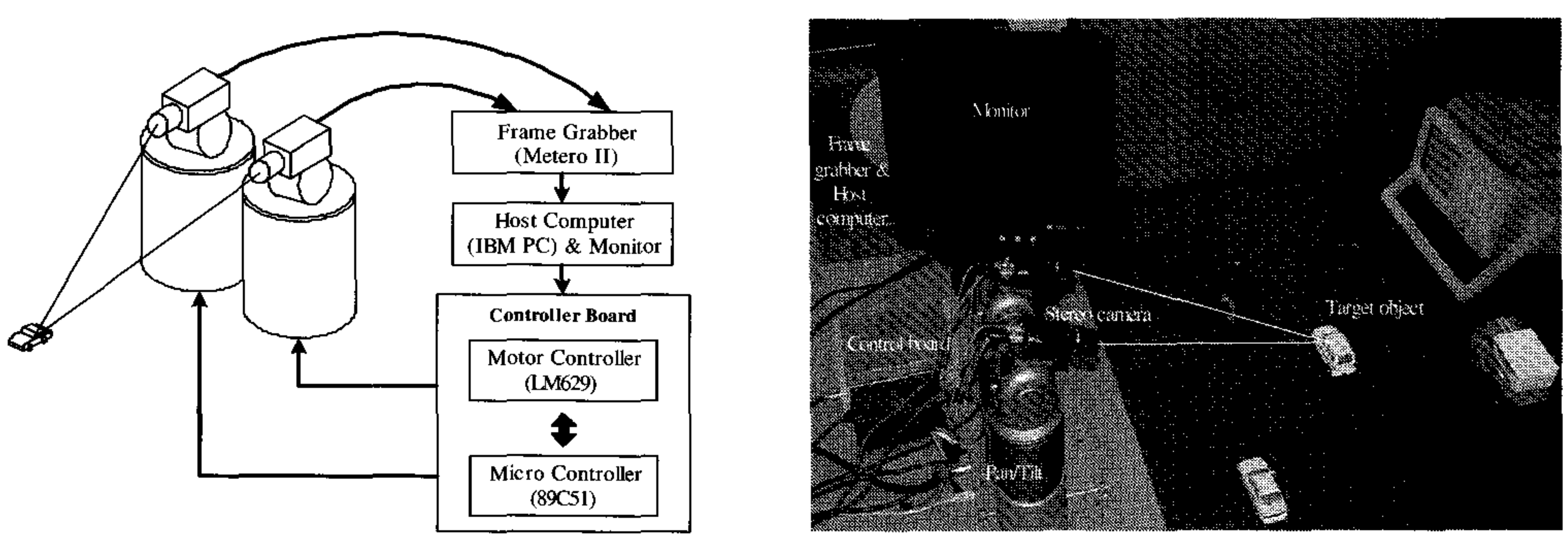


Fig. 5. Experimental setup for DMV-based 3D object tracking.

Case4:

$$Cand2 - Cand2' = \frac{1}{M \times N}\sum_{m=1}^{M}\sum_{n=1}^{N}\left\| D_{T-1}(x_{m+i}, y_{n+j}) \right| - \left| D_{T}(x_{m+i}, y_{n+j}) \right\| \quad (6)$$

In addition, the moving direction of the target can also be determined from the polarity of the difference between the center coordinates of the target areas in each frame. That is, if its polarity is positive, then the moving direction of the object is determined to be the right and upper direction, and in the opposite case, it is the left and lower direction. Finally, these relative moving distance values in the x and y direction and its moving direction are transferred to the Step3 on the frame basis.

This moving distance values and the moving direction of the target calculated from step 2 are used for controlling the pan/tilt embedded in the stereo object tracking system and by controlling the pan/tilt embedded can real-time convergence angle control and stereo tracking of a moving target be achieved. For example, if Eq. (4) produces a minimum difference value, then the *Cand1* and the *Cand2'* are decided as the real target areas in the T-1 and T frames, respectively. From these two detected target areas, the starting coordinates, $D_{T-1}(x_1, y_1)$ of *Cand1* and the ending coordinates, $D_T(x_{1+i}, y_{1+j})$ of *Cand2'* can be extracted. Then, the position difference between these two is calculated and sent to the stereo object tracking system for convergence control and the pan/tilt control. The above process of Step 1 ~ Step 3 is repeated continuously until the moving object is under tracking.

According to the moving distance values obtained in Step 1 ~ Step 3 on the frame basis, the pan/tilt angles can be controlled. The moving distance in the x and y direction between two consecutive frames is just given by the values of Δx_i, Δy_i and these are used for controlling the pan/tilt systems to track the target. In order to control the pan/tilt systems to track the target, the relationship between image plane and 3D point must be established first, and this requires the computation of the moving angle per pixel. The relationship between pixel distance and pan/tilt angle in two images, taken from background compensation, has been derived by Murray [14]. For an initial inclination (α) of the camera system and pan and tilt rotations of θ and ϕ, the focal length of f, respectively, this relationship is given by Eq. (7).

$$x_{t-1} = f \cdot \frac{x_t + \theta \sin\alpha \cdot y_t + f \cdot \theta \cos\alpha}{-\theta \cos\alpha \cdot x_t + \phi \cdot y_t + f},$$
$$y_{t-1} = f \cdot \frac{-\theta \sin\alpha \cdot x_t + y_t - f \cdot \phi}{-\theta \cos\alpha \cdot x_t + \phi \cdot y_t + f} \quad (7)$$

Accordingly, with α, θ, ϕ, and f, the position (x_{t-1}, y_{t-1}) of the corresponding pixel in the previous image can be calculated for every pixel position (x_t, y_t) in the current image. After all, pan and tilt rotations of θ and ϕ can be derived from Eq. (7) as shown in Eq. (8).

$$\theta = \frac{f x_{t-1} y_t (y_{t-1} - y_t) + f(f^2 - y_{t-1} y_t)(x_{t-1} - x_t)}{f(y_t \sin\alpha - f\cos\alpha) + x_{t-1} x_t (y_{t-1} y_t \cos\alpha - y_t \sin\alpha + \cos\alpha)} \quad (8)$$
$$\phi = \frac{f x_t \cos\alpha (x_{t-1} - x_t)(f - y_{t-1}) + f(y_{t-1} y_t) x_{t-1} x_t \cos\alpha + f(y_t \sin\alpha + f\cos\alpha)}{(f^2 - y_{t-1} y_t)(x_{t-1} x_t \cos\alpha + f y_t \sin\alpha + f^2 \cos\alpha) - (1 - y_{t-1})(x_{t-1} x_t^2 \cos\alpha)}$$

Accordingly, once the location value of current and previous image is detected, pan and tilt rotations of θ and ϕ can be calculated by using Eq. (8), which can be derived as a control angle of a motor through the encoder of the pan/tilt controlling system.

Additionally, when this disparity vector is transmitted to the receiver with one of the stereo image pair in each frame through the communication channel, then the moving target can be shown in stereoscopic 3D display. Accordingly, the proposed method not only makes stereo tracking possible but also stereoscopic 3D display of a moving object by simply using the disparity information extracted from the sequence of the stereo input image pair.

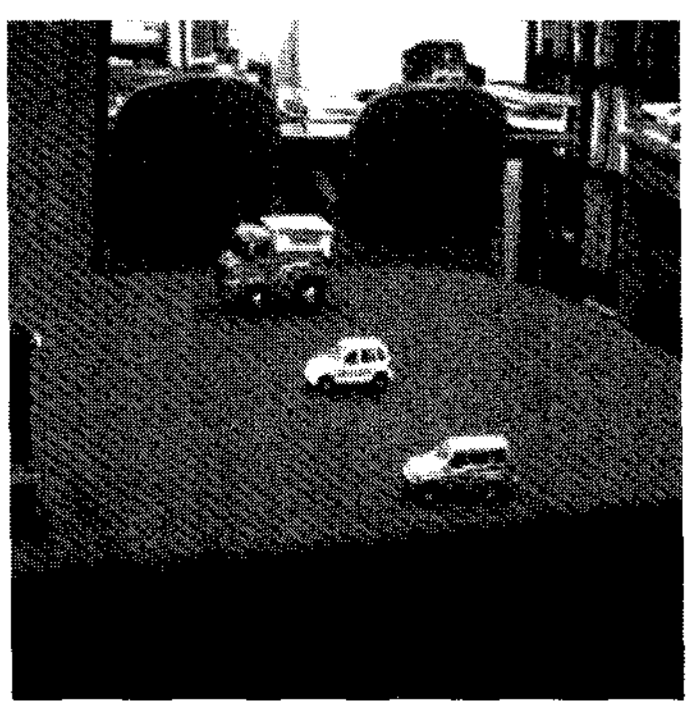
3. Experiments and its Results

Fig. 5 shows the experimental set-up for the proposed stereo object tracking system. In this study, experiments were done, using 9 frames of the input stereo image pair having 256×256 pixels as the test stereo images. The product of Dong Kyung Electronics(CS-82393BS) and the product of Hanwool Robotics(HWR-PT1) were also used as the stereo camera and the pan/tilt control system. As shown in Fig. 5, the sequential input stereo image pairs are caught up by using the stereo camera embedded on the pan/tilt system and transferred to the host computer through the frame grabber (Metero II/4 & Metero II MC/2). In the host computer, the locational coordinates of the target object in each frame are detected by executing the proposed DMV-based target extraction algorithm. Then, they are transmitted to the pan/tilt control board, in which the feedback control signals for the pan/tilt are generated using micro controller (89C51), and finally, the pan/tilt is controlled by using these signals through the motor controller (LM629).

Initially, the time-sequential disparity vectors are extracted from the sequence of the stereo input image pairs and the disparity motion vectors between the consecutive frames are calculated and this disparity motion vectors are used to detect the two candidate areas from each disparity map. Then, from the four detected candidate areas, two real areas where the object is located are identified through the disparity motion vectors and its location coordinates are finally detected. These location coordinate values are sent to the pan/tilt system embedded in the stereo object tracking system.

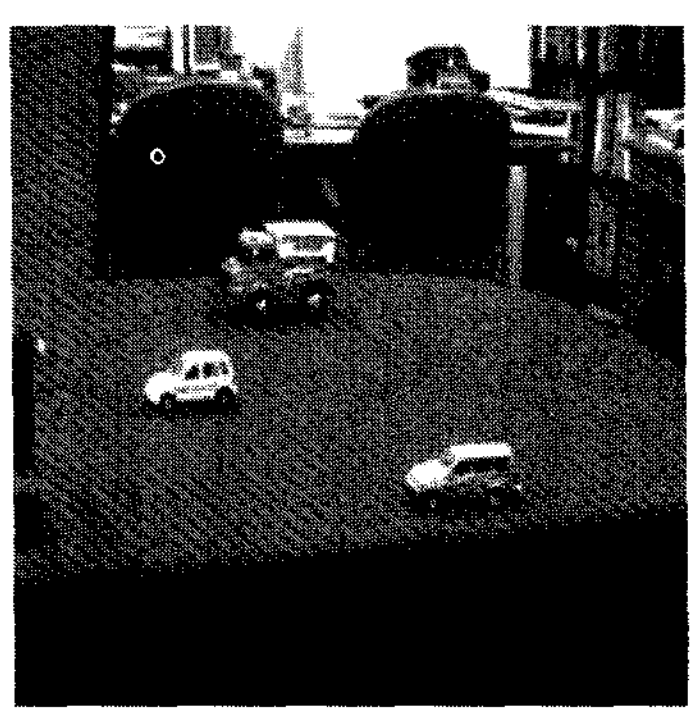
Fig. 6 shows the stereo image pairs of the 1^{st} (initial) and 2^{nd} frames. From this figure, a small car located in the center of the stereo input image is the target object and it is see to be moving from the right to the left direction with a dummy car and truck in the background. In the 2^{nd} frame, the target object is located away from the center of the stereo image by its moving, and so stereo tracking of a moving target is always required to keep it in the center of the stereo image in each frame.

Figs. 7(a) and (b) show the disparity maps extracted from the 1^{st} and 2^{nd} stereo image pairs, in which each gray level of the disparity maps represent the disparity value at that point. Here, the block size and the search range for disparity matching are given by $N_x \times N_y = 1 \times 1$ and $\pm S_x = \pm 32$, respectively. Then, the DMV is estimated by calculating the disparity difference between these 1^{st} and 2^{nd} disparity maps, and in this DMV map, two areas having a relatively large change of disparity values resulted from the target's movement are found as shown in Fig. 7(c).

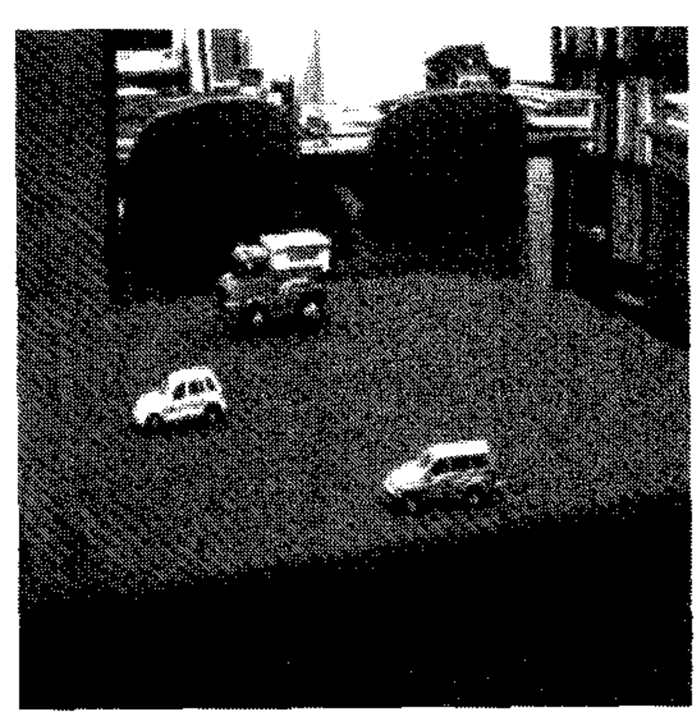


(a) Left image of 1^{st} frame

(b) Right image of 1^{st} frame



(c) Left image of 2nd frame



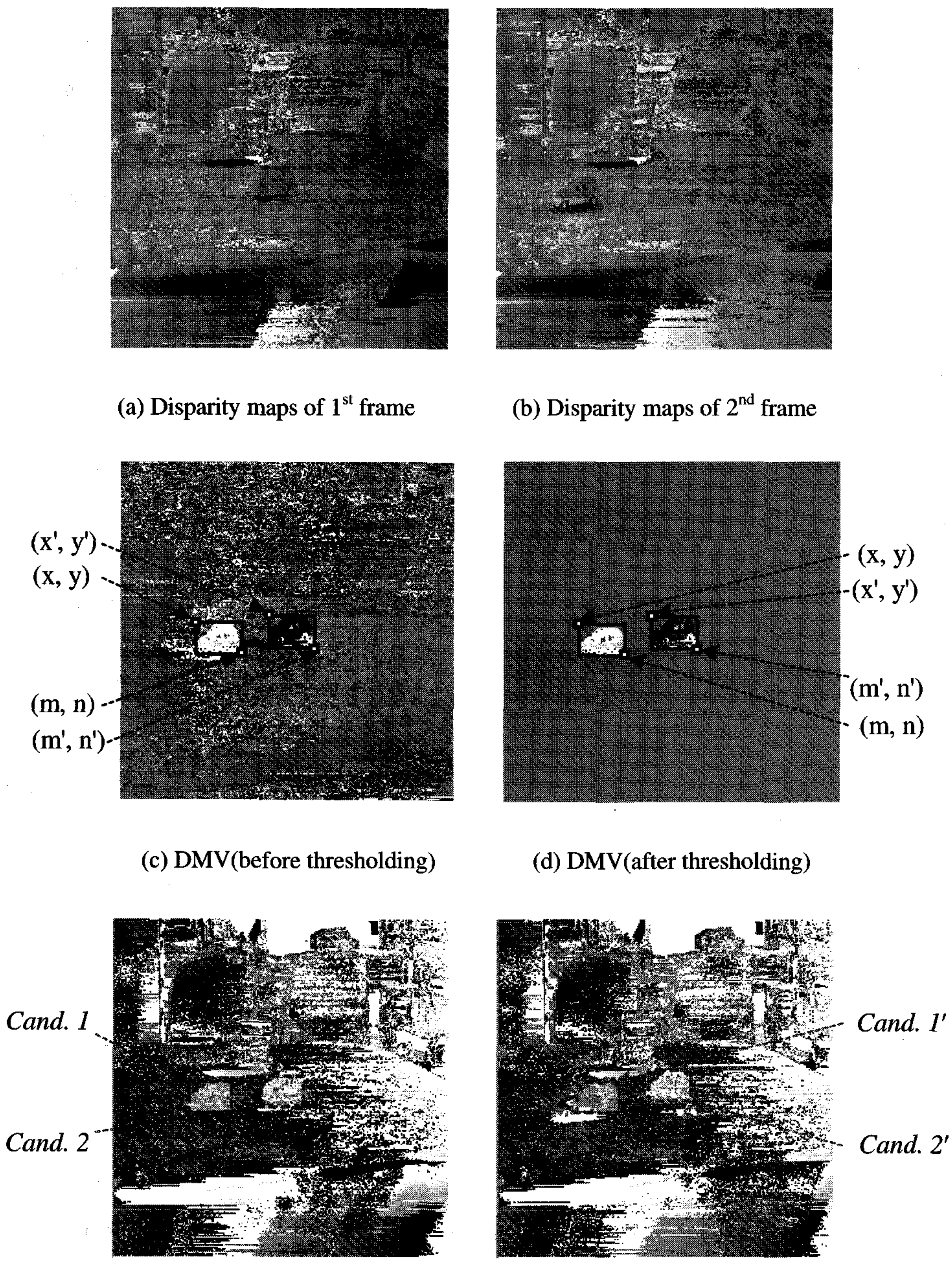
(d) Right image of 2nd frame

Fig. 6. Stereo image pairs of 1st, 2nd frames.

Moreover, these areas can be effectively extracted through a process called threshold. A thresholded version of the DMV map is given in Fig. 7(d). By mapping this DMV to each of the 1st and 2nd disparity maps, four candidate areas, *Cand1, Cand2, Cand1′* and *Cand2′*, are finally obtained as shown in Fig. 7(e) and (f).

Table 1 shows the four candidate areas found by using the proposed DMV-based target extraction method, in which *Start (x, y), End (x, y)* and *Center (x, y)* mean the starting, ending and center coordinates of the obtained candidate areas, respectively.

As shown in Table 1, two pairs of the candidate areas, *Cand1, Cand2*and *Cand1' and Cand2'* have the same location coordinates. In each pair of the candidate areas, one candidate represents the real target area and the other one represents the false target area, the image of which was



(a) Disparity maps of 1st frame (b) Disparity maps of 2nd frame

(c) DMV(before thresholding) (d) DMV(after thresholding)

(e) Candidate areas of 1st frame (f) Candidate areas of 2nd frame

Fig. 7. Four candidate areas extracted from the disparity maps of 1st, 2nd frames.

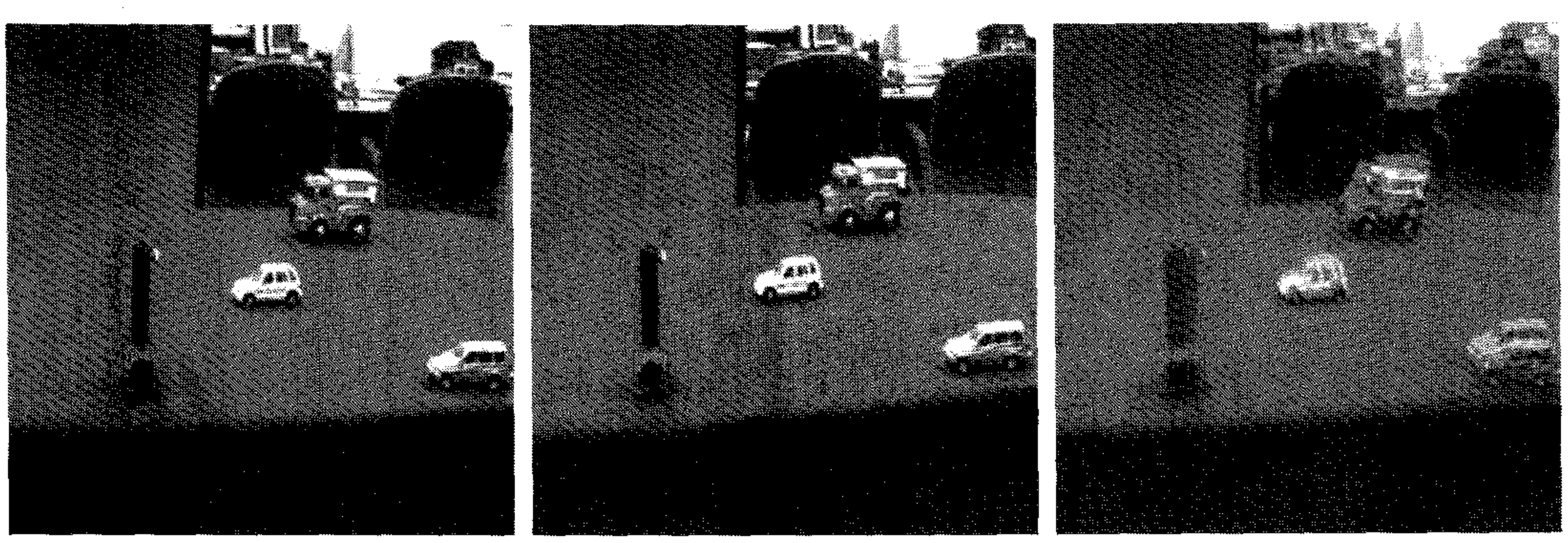
obtained from the other frame. By calculating 4 cases of the cost functions as shown in Eqs. (3)~(6) by using 4 kinds of combinational pairs of the candidate areas, the right target areas in each frame are determined and their location coordinates are also obtained. Basically, the real target areas in each frame have almost the same disparity vector values such that, the cost function for a pair of the real target areas in each frame become smaller than the other cases. Accordingly, among these 4 cases of the cost functions,

Table 1. Location coordinates for the detected four candidate areas

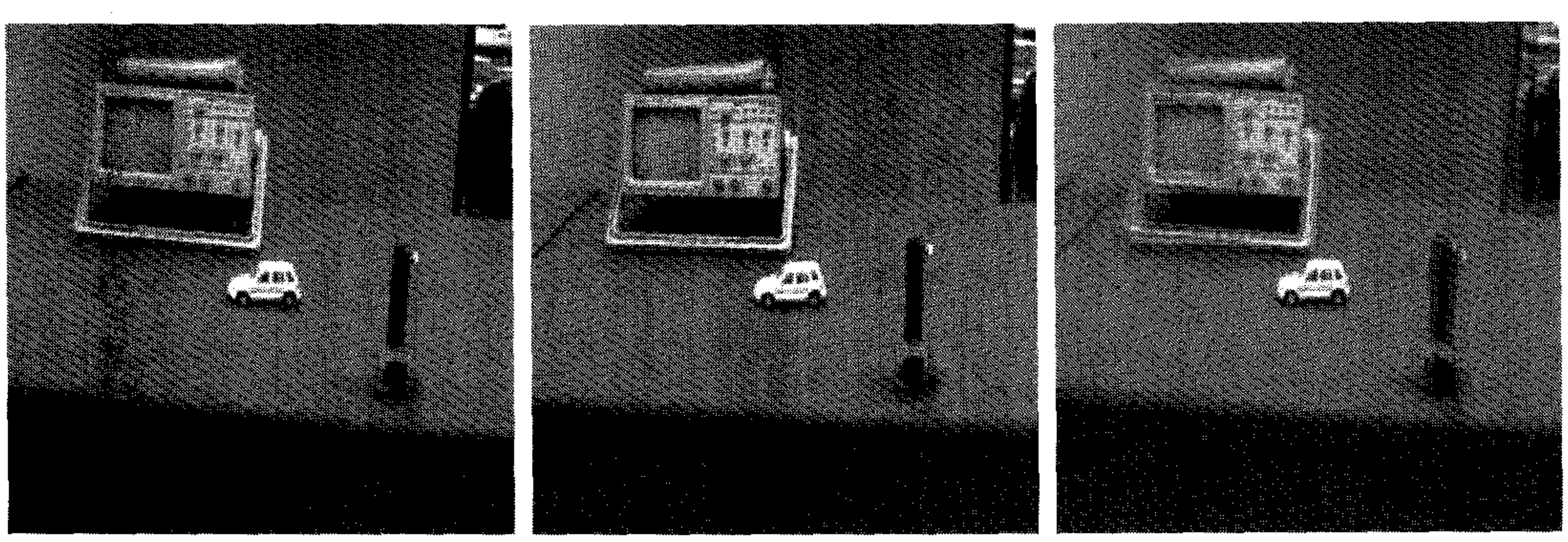
Areas / Coordinates	*Cand1*	*Cand2*	*Cand1'*	*Cand2'*
Start (x, y)	(61, 125)	(125, 118)	(61, 125)	(125, 118)
End (x, y)	(80, 141)	(144, 136)	(80, 141)	(144, 136)
Center (x, y)	(71, 133)	(135, 127)	(71, 133)	(135, 127)

the case that makes the value of the cost function smallest is selected as a combination pair of the real target areas. From this selected case, the real target areas in each frame can be finally obtained.

Figs. 8 (a) and (b) show the stereo image and the composite image of 2nd frame after tracking by adjusting the position of the pan/tilt using this control signal of (-64, 6). Here, the moving direction of the target object can be decided as follows. That is, if the difference between the center coordinates of the target areas in the 1st and 2nd frames is the positive value, then the target object is assumed to move to the right direction in the x-coordinate and to the upper direction in the y-coordinate, whereas, for the opposite case, it is assumed to move to the left direction in the x-coordinate and to the lower direction in the y-coordinate. In the above case, the difference between the center coordinates of the *Cand2* and *Cand1'* is found to be (-64, 6), which means the pan/tilt had moved 62 pixels to the left direction and 6 pixels to the upper direction. By



(a) Stereo image of 2nd frame after tracking (b) Composite image (2nd frame)



(c) Stereo image of 3rd frame after tracking (d) Composite image (3rd frame)

Fig. 8. Stereo image pairs and its composite images after tracking.

comparing these results with those of Fig. 8 (c), it is found that each target object in the left and right images in Fig. 8 (a) is adjusted to be located at the center of the input image plane after tracking, and the composite image of Fig. 8 (b) also shows that the stereo disparity between the left and right input images is adjusted almost to be zero after tracking, which means the convergence angle and the pan/tilt of the stereo object tracking, system were adequately controlled by the proposed method. In addition, the tracking results of the 3rd frame are also shown in Figs. 8 (c) and (d).

Table 2. Moving distances of the target object

Results / Tracking Frames	Actual results		Experimental results	
	Relative moving distance	Absolute moving distance	Relative moving distance	Absolute moving distance
1	(0, 0)	(0, 0)	(0, 0)	(0, 0)
2	(62, -6)	(62, 6)	(64, -6)	(64, 6)
3	(63, -2)	(125, 10)	(61, -1)	(125, 11)

Table 3. Location error ratio of the target object

Frames / Errors	2	3	4	5
Error ratio (%)	3.21	3.55	2.02	3.41
Average error ratio (%)	3.05			

Table 2 shows the actual and detected moving distances of the target object for 5 consecutive frames of the stereo input image pair. Table 2 shows that there is only an error of 1~2 pixels between the actual and detected moving distance values of the moving object. That is, in the 2nd frame, the object moves from the origin to (64, 6) and moves to (125, 11) in the 3rd frame. Table 3 shows the tracking error ratio between the actual and detected location coordinates of the target object which is about 3.05% on average. Here, tracking error ratio is defined as the value of the difference between the actual and detected moving distances divided by the actual moving distance. From these experimental results, we can see that the proposed DMV-based stereo tracking method can be applied to the practical implementation of the stereo object tracking system. But, further investigation need to be conducted to find ways compensate the mechanical error of the pan/tilt system and to develop some sophisticated algorithms for extracting the disparity information with greater accuracy.

4. Conclusions

In this paper, a new stereo object tracking system based-on the DMV is proposed. In the proposed method, the moving target and its location coordinates are effectively obtained using the DMV. This location information of the moving object are then used for controlling the pan/tilt embedded in the stereo object tracking system and to consequently stereo tracking of a moving object has been realized in real-time. From the results of the experiments conduced with the 9 frames of the stereo image pairs having 256×256 pixels, we confirmed that the proposed DMV-based stereo object tracking system can track the moving target with a relatively low error ratio of about 3.05 % on average between the detected and actual location coordinates of the moving target. This indicates a possibility in realizing stable real-time stereo object tracking and a practical atomic industrial stereo robot vision system based-on the disparity motion vector.

References

[1] S. Mandelbaum, R. Hansen, M. Burt, and P. Baten, *CIRA* (1998), p. 671.

[2] T. Darrell, G. Gordon, M. Harville, and J. Woodfill, *Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (1998), p. 601.

[3] T. Joshi, N. Ahuja, and J. Ponce, *Int. Journal of Computer Vision*, **31**, 31 (1999).

[4] M. Han and T. Kanade, *IEEE Conference on Computer Vision* (2001), p. 163.

[5] J. S. Lee, J. H. Ko, and E. S. Kim, *Optics Communications*, **191**, 191 (2001).

[6] J. S. Lee, C. W. Seo, and E. S. Kim, *Optics Communications*, **200**, 73 (2001).

[7] D. J. Coombs and C. M. Brown, *IEEE Control Systems Magazine*, **11**, 24 (1991).

[8] D. Coombs and C. Brown, *Int. Journal of Computer Vision*, **11**, 147 (1993).

[9] V. S. Grinberg, G. Podnar, and M. W. Siegel, *Proc. of SPIE* (1994), p. 56.

[10] M. G. Perkins, *IEEE Trans. on Communications*, **40**, p. 684 (1992).

[11] C. H. Yang, *Rapport technique de l'INRS-T'el'ecommunications*, **95**, p.12 (1995).

[12] Z. L. He and M. L. Liou, *IEEE Transactions on Circuits and Systems on Video Technology*, **7**, 819 (1997).

[13] K. H. Bae and E. S. Kim, *Optical Engineering*, **42**, 6 (2003).

[14] D. Murray and A. Basu, *IEEE Transaction on Pattern Analysis and Machine Intelligence*, **16**, 449 (1994).