Journal of Applied Science and Engineering, Vol. 16, No. 3, pp. 225–232 (2013) DOI: 10.6180/jase.2013.16.3.01

A Novel Stereo Object Segmentation Algorithm Using Disparity and Temporal Information

Jing Chen[1,2]*, Can-Hui Cai[2] and Cui-Hua Li[1]

[1]School of Information Science and Engineering, Xiamen University
[2]Institute of Information Science and Engineering, Huaqiao University

Abstract

A novel stereoscopic object segmentation method based on disparity and temporal-spatial cues is proposed in this paper. First, a foreground layer map is generated from a patched disparity image and a motion object map is produced using an adaptive reference frame selection method. Then the intersection of these two maps is used to extract the foreground motion objects. In the foreground layer map generation stage, two intensity image sequences of the input stereo videos are separately mapped to the rank space by a rank transform to eliminate the parameter deviation of binocular cameras, and reduce the interference of the environment fluctuation and noises. A fast multi-window matching strategy is proposed to speed up the stereo matching process. Then, a disparity patching process is performed to bridge the disparity discontinuity in the smooth region to improve the accuracy of the disparity map. In the motion object map generation stage, an adaptive reference frame selection method is introduced to obtain salient motion and produce the motion mask of each view. Experimental results have indicated the good performance and the practicability of the proposed stereo segmentation algorithm.

Key Words: Disparity Estimation, Stereo Matching, Rank Transform, Disparity Map

1. Introduction

Stereo segmentation is an important topic in three-dimensional (3D) visual analysis. Generally, the stereo segmentation algorithms can be classified as region-based and object-based segmentation [1]. The former one aims to cluster the similar pixels into homogenous regions, while the latter one tries to extract the meaningful object and distinguish the foreground from the background [2]. The most important step in region-based algorithm is stereo matching. In [3], a stereo matching algorithm with an adaptive window was proposed to solve the problem of inaccurate local matching caused by the disparity discontinuity. Since the window searching was performed pixel by pixel, it was very time consuming. To improve the efficiency, a region growth based matching algorithm whose template window size is adaptive to the change of its textural information was proposed in [4]. To reduce the image ambiguity, [5] provided a window-based method for correspondence searching by using dynamic support-weights, which varied according to the color similarity and the geometric proximity. Considering that the intensity in different view may change due to the deviation of camera parameters or lighting inconsistency, in [6], the rank transform was applied to the stereo video sequences before the stereo matching to enhance its robustness. The trouble of this method is its high computational load. V. Kolmogorov et al. [7] described a bi-layer segmentation algorithm for stereoscopic video by fusing disparity with motion and edge cues; however, ambiguities encountered in stereo matching might yield unexpected artifacts. Wu [8] proposed another bi-layer segmentation method for stereo video sequences. However, the inappropriate reference frame selection in the

*Corresponding author. E-mail: chenjing8005@gmail.com

motion detection step might degrade its segmentation accuracy.

In order to further improve the segmentation performance, a novel foreground moving object segmentation method for stereoscopic video is proposed in this paper. The diagram of the proposed stereo object segmentation algorithm is illustrated in Figure 1. Two segmentation techniques, disparity segmentation in the rank domain (the upper left portion in Figure 1) and temporal-spatial segmentation via adaptive reference frame (the upper right portion in Figure 1) are exploited. In the former one, a multi-window stereo matching algorithm is designed to speed up the matching process in rank domain, and the disparity map is patched to mitigate the harm of disparity discontinuity in the smooth region. With the patched disparity map, the foreground layer objects can be well distinguished from the background layer. In the latter one, an adaptive reference frame selection method is applied to alleviate the influence of object motion speed on the extraction of salient motion, and refine the accuracy of moving object location. The foreground moving object can be extracted from the overlap of the disparity map and the temporal-spatial map.

The rest of the paper is organized as follows. Section 2 describes the proposed stereo object segmentation algorithm, which consists of four major steps: (1) Execute the stereo matching in rank domain; (2) perform disparity based segmentation and build the foreground map; (3) detect the salient motion objects in the temporal-spatial domain; (4) extract foreground motion objects from the intersection of the foreground map and the salient motion objects. Section 3 shows some simulation results to demonstrate the good performance of the proposed algorithm. Section 4 concludes the paper.

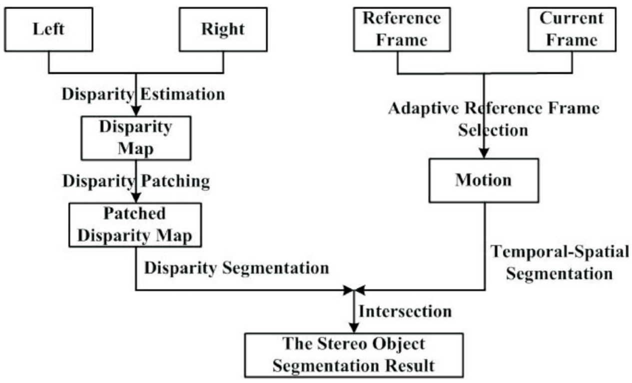


Figure 1. The diagram of the proposed algorithm.

2. The Stereo Object Segmentation Algorithm

2.1 The Stereo Matching

To obtain the disparity map of a stereo video, a fast multi-window based matching algorithm in rank domain is proposed to estimate the disparity. Figure 2 illustrates the stereo matching process. First, frames from left and right views are separately mapped to the rank space by the rank transform. Since the rank transform can be immune to slight deviation of diverse views and some kinds of noises (e.g., salt & pepper, Gaussian, etc.) [9], doing the multi-window stereo matching in rank space can mitigate the interferences due to the disturbance of 3D systems.

2.1.1 The Rank Transform

Let the window size be $M \times N$, X_{ij} denote the luminance of current pixel, and R_{ij} be the rank of X_{ij}. Then, the rank transform is defined as

$$R_{ij} = \sum_{m=i-(M-1)/2}^{i+(M-1)/2} \sum_{n=j-(N-1)/2}^{j+(N-1)/2} u(x_{ij} - x_{mn}) \tag{1}$$

where

$$u(x_{ij} - x_{mn}) = \begin{cases} 1, & x_{ij} > x_{mn} \\ 0, & x_{ij} \le x_{mn} \end{cases} \tag{2}$$

For example, the pixel P (highlighted in Figure 3) is in the middle of 5×5 rank transform window, with intensity of 109. The luminance of P is lager than 11 pixels in the window, so the rank of P is 11.

The rank transform only takes care of coefficient rank in the windows, therefore, when the global illumination changes, the rank transform result keeps unchanged. In other word, rank transform can effectively

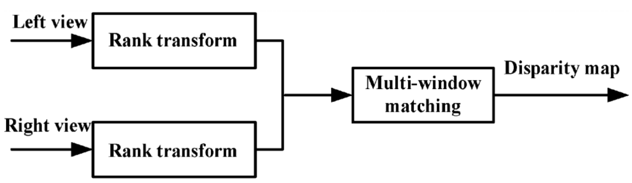


Figure 2. Diagram of the proposed stereo matching method.

prevent from the environmental illumination change. Figure 3(b) shows the rank transform result when the intensity of the 5 × 5 block in Figure 3(a) is increased by 20.

Figure 4 shows the experimental results on “cones” pairs [10] to further verify the immunity against illumination change. The rank transform window size used is 13 × 13. Figure 4(a) and (b) are the original left and right image, and Figure 4(c) and (d) are their rank transform images, respectively. Figure 4(e) is the brightened left image, in which the intensity of each pixel is increased by 20 (described as original left + 20 in the figure illustration), and Figure 4(g) is its rank transform image. Figure 4(f) is the darkened right image, in which the intensity of each pixel is decreased by 20 (described as original right – 20 in the figure illustration), and Figure 4(h) is its rank transform image. It can be seen from Figure 4 that the ranked views (c, d, g, h) can keep clearer edge information than the gray views (a, b, e, f) under global illuminating change or camera parameter inconsistent conditions.

Figure 5 shows the stereo matching results in rank domain (transform window is 11 × 11) and in gray-level space domain, respectively, with the same fixed window searching strategy. The window size is 9 × 9 and the matching criterion used is *Standard Absolute Deviation* (SAD). One can see from Figure 5 that under noisy environment, the outcomes of stereo matching in rank domain (Figure 5(a) and (c)) keep almost unchanged, but those in gray domain (Figure 5(b) and (d)) are degraded a lot. Table 1 documents the comparison of noise influences on stereo-matching in both rank domain and gray domain tested on “cones”, which indicates rank trans-

108	109	108	108	108
109	110	110	109	110
109	109	109	108	109
108	108	109	108	110
109	108	110	108	107

Rank transform ⇒

1	0	1	1	1
0	0	0	0	0
0	0	0	1	0
1	1	0	1	0
0	1	0	1	1

rank(*P*) =11

(a)

128	129	128	128	128
129	130	130	129	130
129	129	129	128	129
128	128	129	128	130
129	128	130	128	127

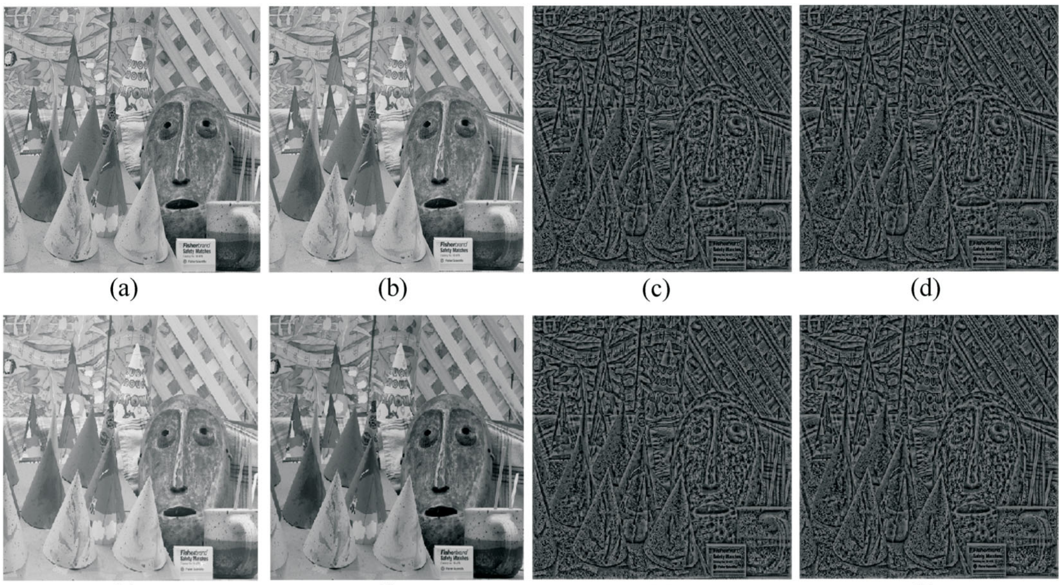
Rank transform ⇒

1	0	1	1	1
0	0	0	0	0
0	0	0	1	0
1	1	0	1	0
0	1	0	1	1

rank(*P*) =11

(b)

Figure 3. An example of the rank transform.



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

(e) (f) (g) (h)

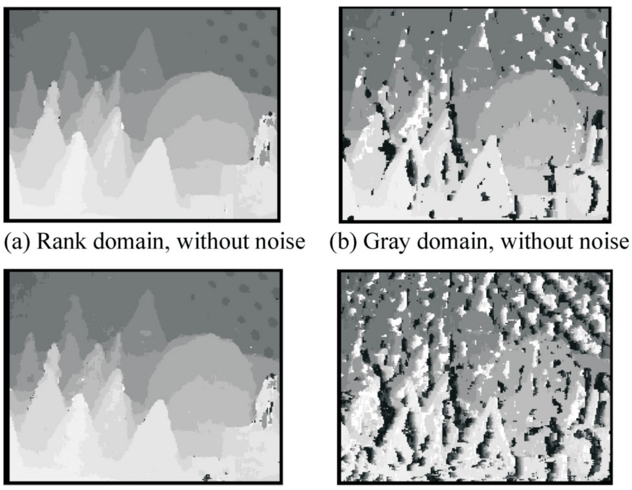
Figure 4. An illustration for anti-lighting change ability of the rank transform: (a) original left, (b) original right, (c) rank transform of (a), (d) rank transform of (b), (e) original left + 20, (f) original right – 20, (g) ranked transform of (e), (h) ranked transform of (f).

form based stereo-matching is more robust under noise interference.

2.1.2 The Multi-Window Stereo Matching Algorithm in Rank Domain

As mentioned before, the performance of region-based stereo matching algorithm greatly depends on the selected support region. Therefore, the design of searching window is essential. An unapt window size can cause disparity discontinuity and edge blurring. For that, a multi-window technique [11] is used in this work. To reduce the computational complexity, the technique of sliding window summations and reuse of intermediate results [12] is adopted to avoid lots of redundant computations introduced by multi-window technique. The process of the proposed multi-window stereo matching algorithm in rank domain is as follows.

- Map left frame and right frame separately to the rank space by the rank transform.
- Choose one frame as current frame and the other one as the reference frame.



(a) Rank domain, without noise (b) Gray domain, without noise (c) Rank domain, salt & pepper (d) Gray domain, salt & pepper

Figure 5. The anti-noise feature of stereo matching in rank transform domain.

Table 1. The error match rate comparison tested on "cones" in rank and Gray domains

"Cones"	Error match rate (%)			
Noise type	Salt & pepper	Gaussian	Poisson	Speckle
Rank domain	18.41	20.24	19.02	20.82
Space domain	21.74	23.05	21.69	23.16

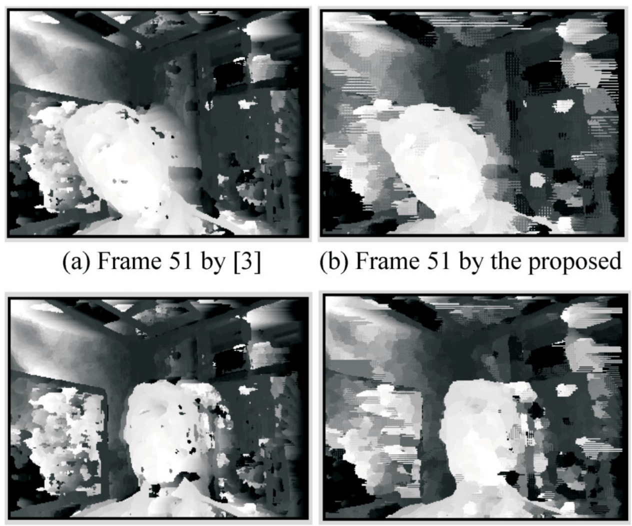
- Search for the best match for each pixel by sliding window summation method to build the disparity map.

Figure 6 shows the comparison of experimental results on "AC" stereo video [13] by the adaptive window matching method [3] and the proposed one. Figure 6(a) and (b) are experimental results on frame 51, and Figure 6(c) and (d) are results on frame 84. From Figure 6, it can be seen that the proposed method produces smoother disparity map in the inner parts of objects and clearer contour of edges. By using the fast multi-window searching method, the disparity map can be built in within 2.6 seconds by our method, while it takes 11.4 seconds by the method proposed in [3] and 4.3 seconds by the method proposed in [9].

2.1.3 Disparity Patching

It can be seen from Figure 6 that disparity discontinuity exits in the smooth area, presented as some "holes" in the disparity map, which will influence the stereo matching accuracy of smooth areas. A patching procedure is applied to the disparity map to overcome the disparity discontinuity in the smooth area. Firstly, a smooth area map $S(x, y)$ is set up by using (3), and then the disparity map is patched.

$$S(x,y)=\begin{cases}0 & f(x,y):\left|SAD(f(x,y))\geq\delta\right| \\ 1 & f(x,y):\left|SAD(f(x,y))<\delta\right|\end{cases} \tag{3}$$



(a) Frame 51 by [3] (b) Frame 51 by the proposed (c) Frame 84 by [3] (d) Frame 84 by the proposed

Figure 6. The comparisons of experimental results on "AC" by [3] and the proposed method.

where, δ is the threshold, SAD is redefined as

$$SAD(f(x,y)) = \frac{1}{mn} \sum_{(w_x,w_y)\in W} \left| f(x+w_x, y+w_y) - \overline{A}_w(f(x,y)) \right|$$

and $\overline{A}_w$ is the mean of gray values in the window, which is defined as

$$\overline{A}_w(f(x,y)) = \frac{1}{mn} \sum_{(w_x,w_y)\in W} f(x+w_x, y+w_y)$$

where, m, n are the length and width of the window, and W denotes the area of the window.

Figure 7 depicts the smooth area maps of the 51st and 84th frames of "AC" sequence with the non-smooth area in black and smooth area in white. The smooth window size is 3 × 3, and the threshold of smooth is δ = 5.

The disparity patch can effectively fill holes in the disparity map and eliminate the influence of mismatch in local area. The disparity patching algorithm is summarized as follows.

- Calculate the mean disparity of each smooth area.
- Replace the disparity of the smooth area with the mean disparity.

Figure 8 shows the patching results in the former disparity map (Figure 6(b) and (d)). From this figure, one can see that the error match part in the disparity map is reduced and a clearer disparity map is produced.

2.2 The Disparity Based Segmentation

Based on the disparity map, a threshold is generated iteratively as follows.

(1) Initialization. Let $t_0 = (d_{\min} + d_{\max})/2$ be the initial threshold, where $d_{\min}$ and $d_{\max}$ is the minimum and the maximum value of the disparity. And set $i = 0$.

(2) Clustering. Separate disparity into two classes, "object" and "background" according to the current threshold value t_i. Compute the mean of "object", $\overline{d}_O$, and the mean of "background", $\overline{d}_B$, using

$$\overline{d}_O = \frac{\sum_{d(x,y)\ge t_0} d(x,y)\times w(x,y)}{N_O} \tag{4}$$

$$\overline{d}_B = \frac{\sum_{d(x,y)< t_0} d(x,y)\times w(x,y)}{N_B} \tag{5}$$

where, $d(x, y)$ represents the disparity value of pixel (x, y); $w(x, y)$ is the weight of (x, y), which is generally set to 1; N_O and N_B denote the number of object pixels and background pixels, respectively.

(3) Iteration. Let $t_{i+1} = (\overline{d}_O + \overline{d}_B)/2$, if $|t_{i+1} - t_i| < \delta$, then t_i is the best threshold, stop the thresholding process; otherwise, let $t_i = t_{i+1}$, go to clustering.

The segmentation is performed based on this threshold, and a foreground layer map is produced then.

2.3 The Salient Motion Detection in the Temporal Domain

To effectively locate the motion object, the interval between current and reference frames should be changed according to the motion state of the video. If objects move fast, the interval should be small; otherwise, the interval should be large enough to figure out the moving object. In this light, an adaptive interval inter-frame differential method is proposed in this paper. The diagram of the proposed method is shown in Figure 9. The initial interval is set to 1. If the mean of frame difference between current and reference frames is smaller than a

(a) Frame 51 (b) Frame 84

Figure 7. The smooth area map of "AC".

(a) Frame 51 (b) Frame 84

Figure 8. The patched disparity map of "AC".

threshold, the distance of current and reference frames is increased by 1; otherwise, the reference frame is selected. The differential between current frame and the selected frame generates the salient motion map of the object, which further benefits the segmentation.

Figure 10 illustrates the advantage of the proposed adaptive reference selection based salient motion detection algorithm. For 7^{th} frame of "AC" stereo video, if the close neighbor frames are selected as its references, the boundary of the object is not obvious (Figure 10(a), (b)), because the object motion is slow in the first 10 frames. By using the adaptive forward and backward reference frame selection method, the distance of compared frames is larger than one (e.g., forward distance is set to 6, and backward distance equals to 7). The differentials of selected frames are shown in Figure 10(c) and (d) with clearer boundary of the object and less background noise.

2.4 The Proposed Stereoscopic Object Segmentation Algorithm

The foreground motion object segmentation is achieved by intersecting the foreground layer map and the salient motion map. For the current input frames, the proposed stereoscopic object segmentation algorithm is summarized as follows.

(1) Generate the patched disparity map by using the fast multi-window stereo matching method in rank domain, and produce the foreground layer map.
(2) Detect the motion objects by the salient motion detection, and generate the motion mask.
(3) Extract the foreground layer motion objects by intersecting the motion mask and the foreground layer map.

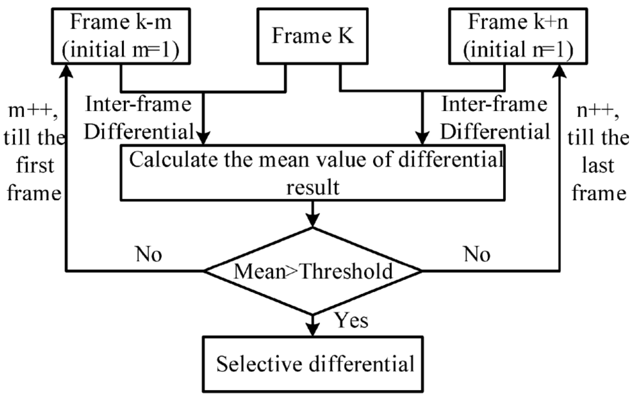


Figure 9. The diagram of the proposed adaptive inter-frame differential method.

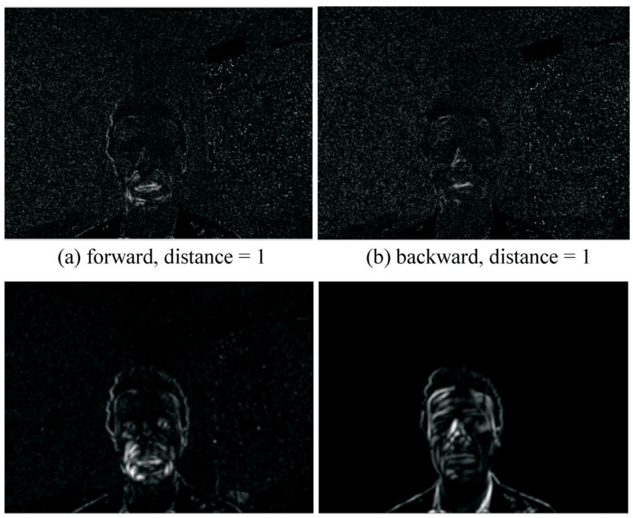
3. Experimental Results and Discussions

Experiments are performed on the "AC" stereo video sequence to verify the effectiveness of the proposed stereo object segmentation algorithm. In disparity segmentation part, the window size of the rank transform is 13×13; the smooth window size in the disparity patching step is 3×3, the threshold of smooth $\delta = 5$. In temporal-spatial segmentation part, the inter-frame differential method based on the adaptive forward and backward frame selection is applied, which helps to generate the motion mask of each view. The fusion of two parts is done by the intersection.

The final segmentation results by the proposed algorithm are as shown in Figure 11. From this figure, one can see that by exploiting both the disparity and temporal-spatial information, the proposed algorithm can efficiently and accurately perform the stereo object segmentation.

4. Conclusions

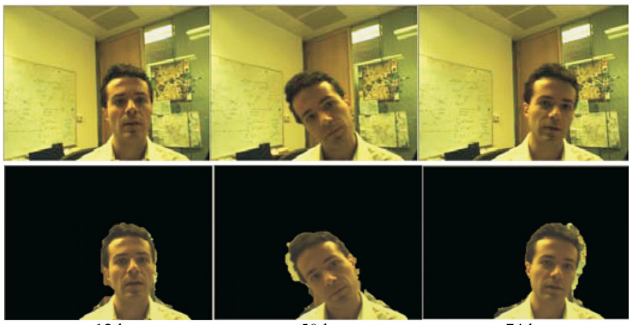
This paper has proposed an effective stereo video segmentation algorithm by combining motion object cues with disparity cues embedded in stereo videos. Consequently, the algorithm consists of three parts: the foreground layer map generation (i.e., disparity-based segmentation), the motion object map generation (i.e., tem-



(a) forward, distance = 1 (b) backward, distance = 1

(c) forward, distance = 6 (d) backward, distance = 7

Figure 10. The experimental results of salient motion detection.



12th 50th 74th

(a) Left views and segmentation results

12th 50th 74th

(b) Right views and segmentation results

Figure 11. The segmentation results of proposed algorithm.

poral-spatial based segmentation), and the fusion of both segmentation results. In disparity segmentation part, the rank transform is first exerted to alleviate the interference introduced by parameter deviation and noises of binocular stereo video. Second, the multiple windowing stereo matching strategy is applied to improve the matching accuracy, and the sliding window method is used to speeds up the stereo matching process. Third, the disparity patch in smooth area is employed to improve the accuracy of the disparity map. In temporal-spatial based segmentation part, an inter-frame differential method based on the adaptive forward and backward reference frame selection method is proposed to get a salient motion of each view. Finally, both resultant segmentation maps are intersected by a logical AND operation to produce the map of foreground motion objects. The experimental results have proved the proposed algorithm can achieve good performance for foreground motion object segmentation.

Acknowledgements

This work is partially supported by the National Natural Science Foundation of China (No. 61250009), the National Defense Basic Scientific Research Program of China, the Specialized Research Fund for the Doctoral Program of Higher Education of China (No. 20110121110020), the National Defense Science and Technology Key Laboratory Foundation, the Fujian Province Natural Science Foundation (No. 2011J01354) and the Fundamental Research Funds for the Central Universities of Huaqiao University (No. 11QZR03).

References

[1] Zhang, Q. and Ngan, K. N., "Segmentation and Tracking Multiple Objects under Occlusion from Multiview Video," *IEEE Transactions on Image Processing*, Vol. 20, No. 11, pp. 3308–3313 (2011). doi: 10.1109/TIP.2011.2159228

[2] Tola, E., Lepetit, V. and Daisy, F. P., "An Efficient Dense Descriptor Applied to Wide-Baseline Stereo," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 32, pp. 815–830 (2010). doi: 10.1109/TPAMI.2009.77

[3] Kanade, T. and Okutomi, M., "A Stereo Matching Algorithm with an Adaptive Window: Theory and Experiment," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 16, pp. 920–932 (1994). doi: 10.1109/34.310690

[4] Tang, L., Wu, C. and Chen, Z., "Image Dense Matching Based on Region Growth with Adaptive Window," *Pattern Recognition Letters*, Vol. 23, pp. 1169–1178 (2002). doi: 10.1016/S0167-8655(02)00063-6

[5] Yoon, K. and Kweon, I., "Adaptive Support-Weight Approach for Correspondence Search," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 24, pp. 650–656 (2006). doi: 10.1109/TPAMI.2006.70

[6] Banks, J. and Bennamoun, M., "Reliability Analysis of the Rank Transform for Stereo Matching," *IEEE Transactions on Systems Man and Cybernetics*, Part B, Vol. 31, Issue 6, pp. 870–880 (2001). doi: 10.1109/3477.969491

[7] Kolmogorov, V., Criminisi, A., Blake, A., Cross, G. and Rother, C., "Probabilistic Fusion of Stereo with Color and Contrast for Bilayer Segmentaion," *IEEE Transactions on Pattern Analysis and Machine Intel-*

ligence, Vol. 28, pp. 1480–1492 (2006). doi: 10.1109/TPAMI.2006.193

[8] Wu, Y., Wang, P. P. and Li, J., "Bi-Layer Segmentation from Stereo Video Sequences by Fusing Multiple Cues," *IEEE Int. Conf. on Multimedia and Expo (ICME)*, Vol. 1, pp. 1353–1356 (2008). doi: 10.1109/ICME.2008.4607694

[9] Chen, Y. and Cai, C., "Two-Level Stereo Matching Based on Rank Transform Domain," *Computer Engineering and Applications*, Vol. 44, No. 34, pp. 188–190 (2008).

[10] Middlebury. http://vision.middlebury.edu/stereo.

[11] Fusiello, A., Roberto, V. and Trucco, E., "Efficient Stereo with Multiple Windowing," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 858–863 (1997). doi: 10.1109/CVPR.1997.609428

[12] Karsten, Dennis, M., Hesser, D. and Reinhard, J., "Calculating Dense Disparity Maps from Color Stereo Images, an Efficient Implementation," *IEEE Workshop on Stereo and Multi-Baseline Vision*, pp. 30–36 (2001). doi: 10.1109/SMBV.2001.988760

[13] I2idatabase. http://research.microsoft.com/en-us/projects/i2i/data.aspx.

Manuscript Received: Feb. 27, 2013
Accepted: Jun. 28, 2013