Local stereo matching using combined matching cost and adaptive cost aggregation

Shiping Zhu[1]*, Zheng Li[1]
[1] Department of Measurement Control and Information Technology, School of Instrumentation Science and Optoelectronics Engineering, Beihang University, Beijing 100191, China
[e-mail: spzhu@163.com]
[e-mail: lizheng900911@163.com]
*Corresponding author: Shiping Zhu

Received August 31, 2014; revised November 10, 2014; accepted December 13, 2014; published January 31, 201

Abstract

Multiview plus depth (MVD) videos are widely used in free-viewpoint TV systems. The best-known technique to determine depth information is based on stereo vision. In this paper, we propose a novel local stereo matching algorithm which is radiometric invariant. The key idea is to use a combined matching cost of intensity and gradient based similarity measure. In addition, we realize an adaptive cost aggregation scheme by constructing an adaptive support window for each pixel, which can solve the boundary and low texture problems. In the disparity refinement process, we propose a four-step post-processing technique to handle outliers and occlusions. Moreover, we conduct stereo reconstruction tests to verify the performance of the algorithm more intuitively. Experimental results show that the proposed method is effective and robust against local radiometric distortion. It has an average error of 5.93% on the Middlebury benchmark and is compatible to the state-of-art local methods.

Keywords: Stereo matching; gradient matching cost; adaptive window; radiometric distortion

This project is funded by the National Natural Science Foundation of China (NSFC) under grants No. 61375025, No. 61075011 and No. 60675018, also the Scientific Research Foundation for the Returned Overseas Chinese Scholars from the State Education Ministry of China. We express our appreciations to the reviewers for their through review and very helpful comments, which help improving this paper.

http://dx.doi.org/10.3837/tiis.2015.01.013 **ISSN : 1976-7277**

1. Introduction

In recent years, three-dimensional TV (3DTV) and free-viewpoint TV (FTV) are promising technologies for the next generation of home and entertainment services. The key point in 3DTV and FTV is calculating depth information of the scenes or objects. Binocular stereovision is a popular technique for building a three dimensional description of a scene observed from two slightly different viewpoints. By finding correspondent pixels in the reference and target images, depth information can be gained through disparity. This process is called stereo matching. Stereo matching is a classical and challenging problem in computer vision, which has been a hot research focus for a long time. In the last decade, researchers had put forward a large number of algorithms to solve this problem, but because of the ill-posedness of such a problem, there is not a perfect solution yet. Most stereo matching algorithms focus on establishing an energy function and minimizing such an energy function to estimate disparities. So, stereo matching is essentially a problem of finding an optimized solution. The equation is conducted by establishing reasonable energy functions, adding some constraints and adopting an optimization algorithm, which is also the method for solving all ill-posed problems. A thorough survey and taxonomy of dense stereo techniques was provided by Scharstein and Szeliski [1]. They summarized the stereo matching process into four steps: matching cost computation, cost aggregation, disparity computation and disparity refinement. They also divided stereo matching algorithms into local methods and global methods respectively according to the way of cost aggregation. Global methods can generally acquire a higher accuracy, but with less efficiency. On the contrary, local methods are fast and easy to realize, while it is difficult to choose a proper matching cost function [2] and construct right support windows.

Matching cost is the similarity measure of corresponding points between the left and right images. Most stereo matching algorithms use intensity based similarity measures. For instance, the sum of absolute difference (SAD), sum of square difference (SSD) [1], Adapt Weight [3] and Segment Support [4] etc. are all in this category. For ideal images, they can produce results with high precision, but these methods are very sensitive to the image radiometric distortion. When the illumination condition and exposure time change, the accuracy will fall down quickly. Thus it is impossible to apply these methods to real images. Fortunately, there are some kinds of matching costs which are robust to radiometric distortion. The normalized cross-correlation (NCC), Gradient [5][6][7], Rank and Census transform [8][9] are the most commonly used ones.

Local stereo methods need to aggregate single pixels' matching costs in a support region which is defined by a window. Inevitably, they will run into problems when deciding the window size to be used. Small windows do not contain enough information and can lead to noisy results, while large windows contain enough texture information but encompass pixels at different depths near depth discontinuities, resulting the foreground fattening effect. Fusiello and Roberto [10] proposed to select a best window among multiple predefined windows as the support window; Veksler [11] presented a variable window choosing method by exploring a useful range of interesting window shapes and sizes; Zhang [12] constructed a cross-based adaptive window for every pixel according to the color correlation of adjacent pixels and achieved good results. Qu [13] developed a binary support window by calculating the mean intensity in a predefined fixed window, but this binary support window may have a disconnected structure and would degrade the accuracy.

Global stereo methods consider stereo matching as a labeling problem where the pixels of the reference image are nodes and the estimated disparities are labels. They typically skip the cost aggregation step and define a global energy function that includes a data term and a smoothness term. The former sums pixel-wise matching costs, while the latter supports piece-wise smooth disparity selection. The labeling problem is solved by energy function minimization, using dynamic programming, graph cuts, or belief propagation. Some newest global stereo matching algorithms can be found in [14][15][16][17].

To address the above matching cost computation and window size selection problems, this paper proposes a stereo matching algorithm based on an improved gradient cost and adaptive cost aggregation. Our main contributions are twofold: First, we improve the gradient matching cost by incorporating the phase information and proposed a hybrid cost function which combines gradient and color matching cost. Second, we develop a four-step disparity refinement method to eliminate mismatches.

The remaining portions of this paper are organized as follows: We first propose our method and describe the algorithm thoroughly in section 2. Section 3 presents the experimental results and we finally conclude our work in section 4.

2. Proposed Method

According to Scharstein and Szeliski's taxonomy, stereo matching process can be concluded into the following four steps: matching cost computation, cost aggregation, disparity computation and disparity refinement. We will follow this classification to describe our algorithm in detail. The outline of the proposed algorithm is shown in **Fig. 1**. Given two rectified images, we first calculate the corresponding gradient images, which is the prerequisite for computing matching cost. Then an adaptive window is constructed for every pixel to meet the need of cost aggregation. After this, by using the Winner-Takes-All strategy, the initial disparity maps are gained. At last, the final depth images are produced after disparity refinement.

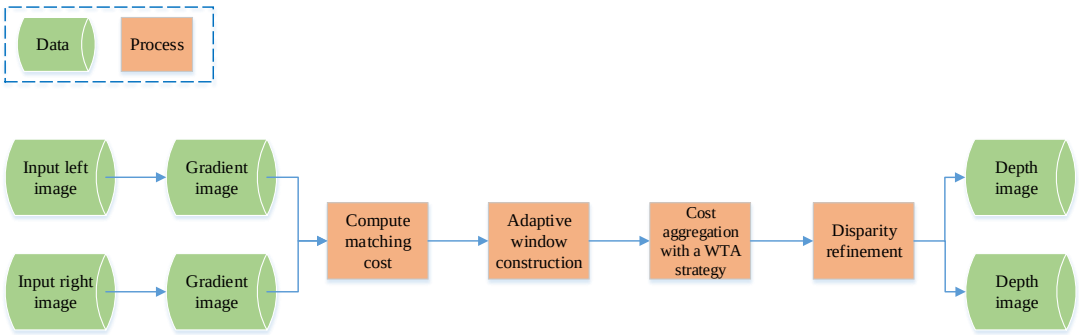


Fig. 1. Outline of the proposed algorithm.

2.1 Matching cost computation

Matching cost is the similarity measure of corresponding points between the left and right images. Using different cost functions will get different disparity discriminations. As we discussed before, gray or color intensity-based matching costs are very sensitive to radiometric

distortion and noise, while gradient-based matching costs are more robust to these factors and have been widely used.

The gradient of an image corresponds to the direction along which the gray value of the image changes most remarkably. In other words, the change of image intensity can be described by image gradient. Mathematically, image gradient is defined as the first-order partial derivatives of image intensity with respect to *x* and *y*, which are represented as a vector:

$$G = \begin{pmatrix} G_x \\ G_y \end{pmatrix} = \begin{pmatrix} \frac{\partial I}{\partial x} \\ \frac{\partial I}{\partial y} \end{pmatrix} \tag{1}$$

where $I(x, y)$ is the image intensity of an anchor pixel (x, y). In practical applications, G can be calculated by convolving the image with gradient masks. Here we just use the simplest gradient mask:

$$G_x = [-1\ 0\ 1],\quad G_y = \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix} \tag{2}$$

Thus, we can get the gradient images of both left and right images: $G_L = \left(G_{Lx}, G_{Ly}\right)^{\mathrm{T}}$, $G_R = \left(G_{Rx}, G_{Ry}\right)^{\mathrm{T}}$. For rectified images, supposing $p = (x, y)$ is a pixel in the left image, then $pd = (x - d, y)$ is the corresponding pixel in the right image with disparity d. Hence, the gradient matching cost function C_G can be defined as:

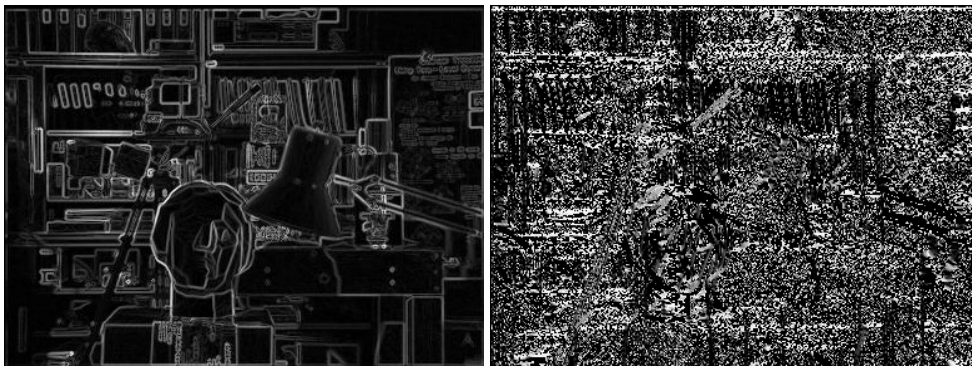
$$\begin{aligned} C_G(p, d) &= \sqrt{\Delta G_x^2 + \Delta G_y^2} \\ \Delta G_x &= \sum_{j=R,G,B} \left| G_{Lx}^j(p) - G_{Rx}^j(pd) \right| \\ \Delta G_y &= \sum_{j=R,G,B} \left| G_{Ly}^j(p) - G_{Ry}^j(pd) \right| \end{aligned} \tag{3}$$

The above cost function only considers the modulus information of the gradient vector. Here, we develop an improved cost function which incorporates the gradient phase, similar to [6]. Using the gradient vector's two components G_x and G_y, the modulus and the phase are computed as:

$$m = \sqrt{G_x^2 + G_y^2} \tag{4}$$

$$\varphi = \arctan\left(\frac{G_y}{G_x}\right) \tag{5}$$

Generally, the modulus m represents the rate of change and the phase φ represents its direction. To show them intuitively, **Fig. 2** gives an example of the computed m and φ for Tsukuba image. We can see that gradient values can reflect the image edges or skeleton to some extent as well as the differences between m and φ.



(a) (b)

Fig. 2. (a) Modulus of gradient; (b) Phase of gradient.

As m and φ provide different information about the neighborhood of a pixel, they have different invariance properties with respect to radiometric distortion. For instance, neither the modulus nor the phase is affected by additive (offset) changes in the input images, while multiple variations (gain) affect the modulus but not the phase. So, it is more proper to consider them separately. Our method is based on this idea. To make full use of the gradient information, we combine the modulus and phase linearly with a weight parameter α, forming our new cost function:

$$G(p,d)=\sum_{c\in\{\mathrm{R,G,B}\}}\alpha\left|m_l^c(p)-m_r^c(pd)\right|+f\left(\left|\varphi_l^c(p)-\varphi_r^c(pd)\right|\right) \tag{6}$$

where, m^c and φ^c are the modulus and phase of the gradient operator applied to each color band c∈{R, G, B} respectively; α is the weight of modulus with a range of [0, 1]. Considering the π-periodicity property of the phase, we employ f to normalize it into single period:

$$f(x)=\begin{cases}x & \text{if } 0\le x\le\pi\\ 2\pi-x & \text{if } \pi<x<2\pi\end{cases} \tag{7}$$

Because we have used a weight parameter α, it is easy to adjust the algorithm's performance by changing the value of α. This is important as different lighting and exposure time can lead to different degrees of radiometric distortion and noise. From (6), we can see that the larger α is, the bigger effect the modulus will have. On the contrary, the phase will dominate if α is small. According to the radiometric distortion degree, the proper value of α can be set empirically.

As color intensities of an image directly reflect the brightness of pixels, using the gradient similarity alone may lose lots of details of the scene. Thus, we propose a combination of the color based SAD cost and the improved gradient cost, which is simple but very effective as it can yield more reliable similarity measure by compensating one another. The color based SAD matching cost can be represented as:

$$C(p,d)=\sum_{c\in\{\mathrm{R,G,B}\}}\left|I_l^c(p)-I_r^c(pd)\right| \tag{8}$$

Then we use a robust function to normalize the costs into [0, 1]:

$$\rho(x,\lambda)=1-\exp\left(-\frac{x}{\lambda}\right) \tag{9}$$

where λ is a controlling parameter. The final integrated matching cost of pixel p corresponding to disparity d is defined as:

$$e(p,d)=1-\exp\left(-\frac{G(p,d)}{\lambda_G}\right)+1-\exp\left(-\frac{C(p,d)}{\lambda_C}\right) \tag{10}$$

In this way, both $G(p,d)$ and $C(p,d)$ are in the range of [0, 1] and their contributions to the final cost can be adjusted by setting different values of λ_c and λ_G. The proper values of these parameters can be got empirically.

2.2 Adaptive window construction

As the identification ability of single pixel's matching cost is weak, we need to propagate the adjacent pixels' matching costs and aggregate them to improve accuracy. The neighborhood region is determined by a local support window and the pixels in the window will be included for aggregation. So, it is natural to ask how large the window should be. In fact, a fixed window can never get satisfactory results, because image regions with different characters need different windows. In textureless regions, larger windows are needed to provide enough pixels. On the contrary, regions with high texture and depth discontinuities need smaller windows to avoid being over-smoothed. To address this problem, Zhang proposed a cross-based adaptive window construction method which can alter the window's shape and size adaptively. Such a cross-based support region is achieved by expanding a cross-shaped skeleton around each pixel p to create four segments $\{h_p^-, h_p^+, v_p^-, v_p^+\}$, defining two sets of pixels H(p), V(p) in the horizontal and vertical directions. More details about the method can be found in [12]. In their original implementation, only one threshold for color similarity and one threshold for spatial closeness are used, which cannot satisfy all cases. Motivated by [18], we present a modification of the original cross-based support region approach in this paper.

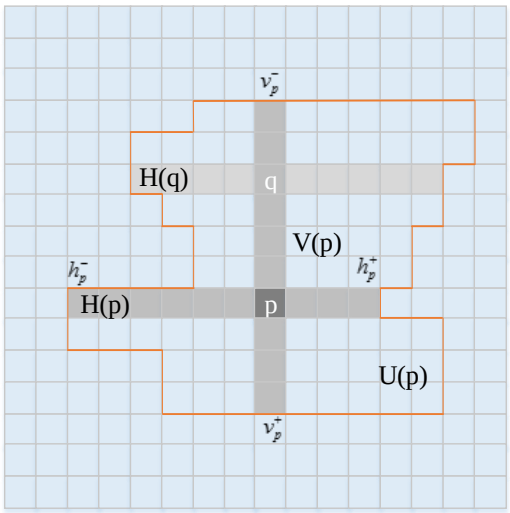


Fig. 3. Construction process of the adaptive window.

The key idea of the cross-based support region is to decide an upright cross for every pixel p in the input image, which is based on the color similarity and spatial closeness. As is shown in **Fig. 3**, the pixel-wise adaptive cross consists of two orthogonal line segments, intersecting at the anchor pixel p. We use H(p) and V(p) to represent the horizontal and vertical segments respectively. Thus, four arms: left, right, up and down are constructed for each pixel and represented as $\{h_p^-, h_p^+, v_p^-, v_p^+\}$. By changing the length of the arms adaptively, we can

effectively capture an adaptive support region for each pixel. Here, we use enhanced rules to decide each pixel's arm length. Just taking p's left arm as an example, it stops when it finds an endpoint pixel p_i that violates one of the three following rules:

1. $D_c(p_i, p) < \tau_1$ and $D_c(p_i, p_i + (1,0)) < \tau_1$;
2. $D_s(p_i, p) < L_1$;
3. $D_c(p_i, p) < \tau_2$, if $L_2 < D_s(p_i, p) < L_1$.

Where, $D_s(p_i, p)$ is the spatial distance between p_i and p; $D_c(p_i, p)$ represents the color difference, which is defined as $D_c(p_i, p) = \max_{c\in\{R,G,B\}}(|I_c(p_i) - I_c(p)|)$; $\tau_1 > \tau_2$, $L_1 > L_2$, are the predefined color thresholds and spatial thresholds. Rule 1 restricts the color difference between p_i and p as well as p_i and its predecessor p_i +(1, 0) on the same arm. This prevents the arm to span over the edges in the image. Rule 2 and 3 provide multiple choices for the arm length. In textureless regions, we use larger threshold L_1 and τ_1 to guarantee enough pixels. But when the arm length exceeds a smaller value L_2, Rule 3 will play its role by using a much stringent threshold τ_2 to make sure that the arm will extend only in regions with very similar colors.

After the above process, we can get the end pixels of the four arms: $\{h_p^-, h_p^+, v_p^-, v_p^+\}$, then H($p$) and V($p$) can be got by:

$$\begin{cases} H(p) = \{(x,y) \mid x \in [x_p - h_p^-, x_p + h_p^+], y = y_p\} \\ V(p) = \{(x,y) \mid x = x_p, y \in [y_p - v_p^-, y_p + v_p^+]\} \end{cases} \tag{11}$$

Finally, by iteratively applying this approach for every pixel q along V(p), we can get the local support window U(p):

$$U(p) = \bigcup_{q\in V(p)} H(q) \tag{12}$$

Fig. 4 shows an example of the adaptive local support windows, which approximates local image structures appropriately.

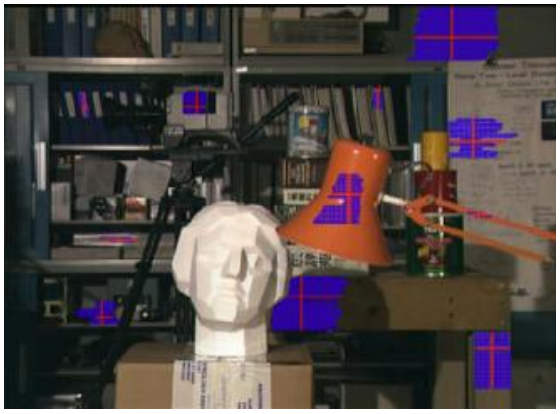


Fig. 4. Example of the adaptive local support windows

2.3 Cost aggregation

Traditional local algorithms only take the reference image's support region into account. In contrast, we will symmetrically consider support regions of both target and reference images. Considering two corresponding pixels p=(x, y) and pd=(x-d, y) in the reference and target

images, then we can acquire two local support regions U(p) and U′(pd). We will combine them to define the union support region:

$$U_d(p) = \{(x,y) \mid (x,y) \in U(p), (x-d,y) \in U'(pd)\} \tag{13}$$

After the support region being prepared, the aggregation matching cost of p is computed as follows:

$$E_d(p) = \frac{1}{N} \sum_{q \in U_d(p)} e(q,d) \tag{14}$$

where N is the number of total pixels in the support region $U_d(p)$, and $e(q, d)$ is the raw per pixel's matching cost corresponding to disparity d. At last, we employ the Winner-Takes-All (WTA) strategy to select the best disparity with the lowest matching cost in the disparity range:

$$d^0(p) = \arg \min_{0 \le d \le d_{\max}} E_d(p) \tag{15}$$

where $d \in [0,\ d_{\max}]$ represents the disparity range, $d^0(p)$ is chosen as the initial disparity of p.

2.4 Disparity refinement

The disparity maps obtained after the previous three processes still contain some mismatches and unreliable values. For further refinement, post-processing steps are required. Our post-processing consists of four steps:

First, we apply a 5×5 median filter to both d_L and d_R which represent the left and right disparity maps respectively for removing isolated outliers.

Second, we implement the common reliable tool: left-right consistency check. A pixel p is characterized as valid if the constraint: $d_L(p) = d_R(p - d_L(p),\ 0)$ holds true. Otherwise, p will be marked invalid and needs to be handled if the constraint is violated. Furthermore, the invalid disparities can be classified into two classes: occlusions and mismatches. We employ Hirschmüller's approach to decide an invalid point is either occlusion or mismatch [19].

Third, we present a disparity refinement method based on the local disparity histogram to recover the invalid disparities. For a pixel p in the disparity image, we build a local disparity histogram $\varphi_p(d)$ in the neighborhood region of p, and count the times that every disparity occurs. Thereby, there will be d_{max}+1 bins corresponding to each disparity. Here, we do not need to seek for a new neighborhood region, but to reuse the previous local support region U(p) for pixel p. Thus, this process will not add much computation cost. Let $H(i)$ be the length of the ith bin, i = 0 to d_{max}. We calculate d^* as a disparity with the maximum normalized histogram:

$$h(i) = \frac{H(i)}{\sum_i H(i)}, \quad i = 0 \ to \ d_{\max} \tag{16}$$

$$d^* = \arg \max_i h(i) \tag{17}$$

In statistic, this disparity value is the local optimal one, and $h(d^*)$ represents its confidence level. The initial disparity $d^0(p)$ of pixel p is replaced by the new value d^* if $h(d^*)$ is greater than τ_h; otherwise, it is left unchanged:

$$\begin{cases} d_p = d^* & h(d^*) > \tau_h \\ d_p = d_p & \text{otherwise} \end{cases} \tag{18}$$

where $\tau_h \in [0, 1]$ is a confidence threshold. This step is repeated iteratively until there are no more updates to disparities in the map.

At last, as the invalid disparities may remain unchanged in step 3, there are still some invalid points need to be filled. We then introduce an interpolation strategy which treats occlusion and mismatch points differently. Interpolation is performed by propagating valid disparities to neighboring invalid disparities areas. For invalid pixel *p*, we find the nearest valid pixels along 8 directions and their disparities d_{pi} are stored. The final disparity of *p* is created by:

$$d_p = \begin{cases} \text{seclow } d_{pi} & \text{if p is occluded,} \\ \text{med } d_{pi} & \text{if p is mismatched.} \end{cases} \tag{19}$$

If *p* is occluded, we select the second lowest value ($\text{seclow } d_{pi}$) to get rid of the preference to foreground or background. If *p* is mismatched, the median ($\text{med } d_{pi}$) is used which can maintain discontinuities in cases where the mismatched area is located at the boundary. Experiments show it can get better results.

3. Experimental Results and Discussions

3.1 Accuracy of the proposed algorithm

This section presents experimental results as we have programmed and implemented the algorithm in C++. To verify the performance of the proposed method, our experiments are based on the rectified stereo images from the Middlebury stereo benchmark [20]. It offers 4 pairs of stereo images: Tsukuba, Venus, Teddy and Cones, with the sizes of 384×288, 434×383, 450×375 and 450×375 respectively. The disparity ranges of them are also given, which are: 0-15, 0-19, 0-59 and 0-59 pixels correspondingly. By comparing the results with the ground truth disparity images, we can get the quantified errors and make objective evaluation. The parameters in the algorithm are set as in **Table 1**, which are kept constant if no special declaring.

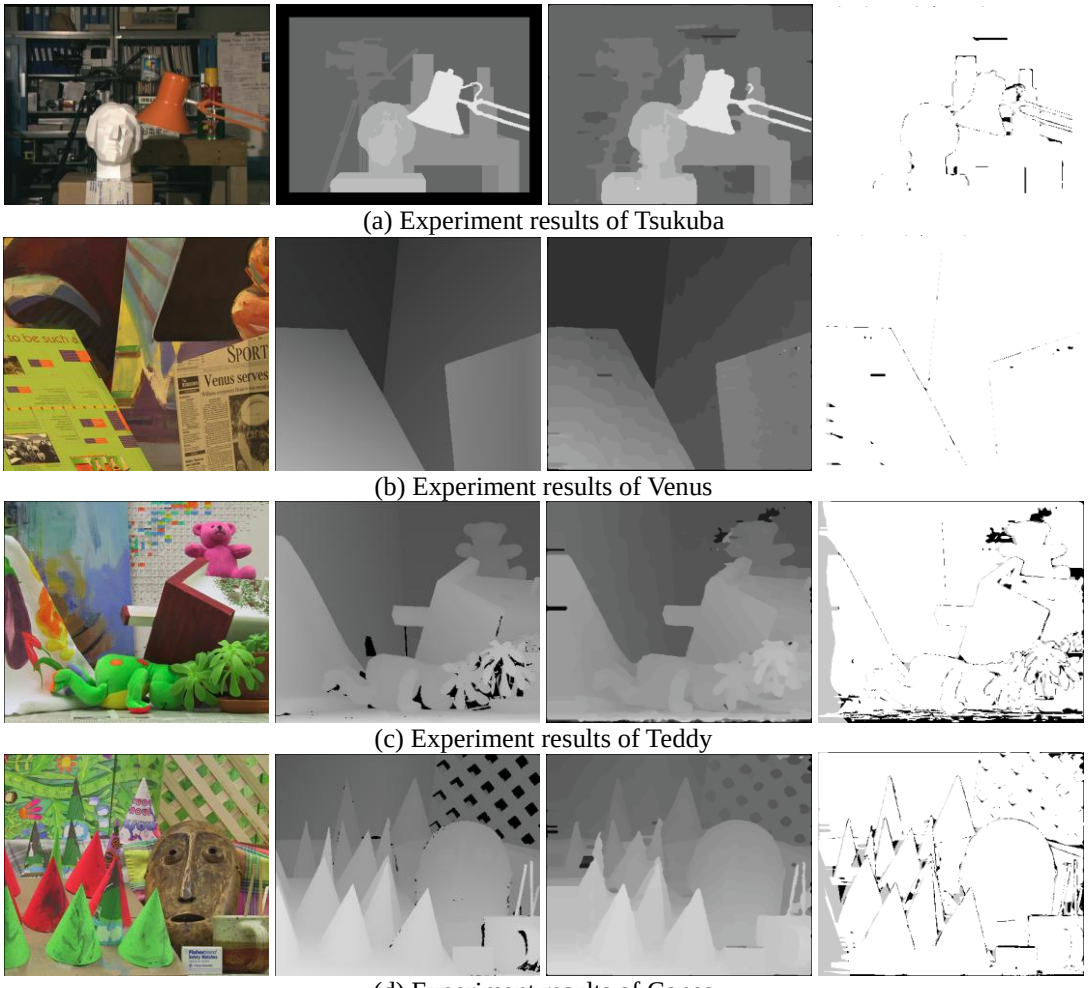
Table 1. Parameter settings for all experiments

α	λ_C	λ_G	L_1	L_2	τ_1	τ_2	τ_h
0.12	35	5	36	18	5	18	0.5

Fig. 5 shows the experimental results of our method on all four stereo pairs of the Middlebury stereo database. The left most column contains the left original images of the four stereo pairs. The ground truth disparity images are shown in the second column, our estimated disparity images are displayed in the third column, and the forth column gives the error maps computed with the ground truth. In the error maps, the white regions denote correctly calculated disparity values which do not differ for more than 1 pixel from the ground truth. Instead, if the estimated disparity differs for more than 1 pixel from the ground truth value, it is marked as an error and displayed in black and gray, where black represents the errors in the non-occluded regions, and gray represents errors in the occluded regions. **Table 2** lists the objective evaluation of ours and other methods with the error threshold: $\delta_d = 1$ pixel, which

means bad pixels are those whose absolute disparity errors are above 1 pixel. Columns Nonocc, All and Disc represent the percentage of bad pixels for pixels in non-occluded regions, for all pixels and for pixels in regions near depth discontinuities.

From overall performance, the proposed method achieves satisfactory results. Our algorithm correctly estimates the disparities of both textureless and textured surfaces. For instance, the large uniform surfaces in stereo pairs Venus and Teddy are successfully recovered while preserving the disparity edges well. For quantified comparison, the proposed method outperforms many classical global and local methods, like Enhanced BP [21], GC+occ [22], SemiGlob [18], AdaptWeight [3] and so on. Although the NonLocalFilter [2] and P-linearS [23] methods have lower average error than ours, but these methods have not consider image amplitude distortion and are sensitive to radiometric difference as they are intensity-based algorithms. In the next subsection, we will demonstrate our method's robustness to image radiometric distortion thoroughly.



(a) Experiment results of Tsukuba

(b) Experiment results of Venus

(c) Experiment results of Teddy

(d) Experiment results of Cones

Fig. 5. Experimental results on Middlebury datasets. From left to right in each row are the original left images, the ground truth disparity maps, the produced disparity maps by our algorithm and the error maps respectively.

Table 2. Objective evaluation of matching results.

Algorithms	Tsukuba			Venus			Teddy			Cones			Avg. Error
	nonocc	all	disc	nonocc	all	disc	nonocc	all	disc	nonocc	all	disc	
NonLocaltFilter[2]	1.47	1.85	7.88	0.25	0.42	2.60	6.01	11.6	14.3	2.87	8.45	8.10	5.48
P-LinearS[23]	1.10	1.67	5.92	0.53	0.89	5.71	6.69	12.0	15.9	2.60	8.44	6.71	5.69
Proposed	**1.46**	**1.92**	**6.80**	**0.36**	**0.53**	**2.41**	**6.61**	**12.1**	**15.3**	**4.08**	**9.99**	**9.55**	**5.93**
AdaptWeight[3]	1.38	1.85	6.90	0.71	1.19	6.13	7.88	18.3	18.6	3.97	9.79	8.26	6.67
Enhanced BP[21]	0.94	1.73	5.05	0.35	0.86	4.34	8.11	13.3	18.5	5.09	11.1	11.0	6.69
SemiGlob[18]	3.26	3.96	12.8	1.00	1.57	11.3	6.02	12.2	16.3	3.06	9.75	8.90	7.50
GC+occ[22]	1.19	2.01	6.24	1.64	2.19	6.75	11.2	17.4	19.8	5.36	12.4	13.0	8.26

To clarify the function of our improved gradient matching cost, we conduct a quantitative comparison test of the proposed method with the traditional method of only using modulus information. In addition, to eliminate interferences and show the effect of our four-step disparity refinement method, we use the results without disparity refinement. For simplicity, we only present the errors of the estimated disparities of non-occluded regions in **Table 3**. It is clear to see that our proposed matching cost imrpove the resulsts a lot. Also, compared with the results after disparity refinement in **Table 2**, the effectiveness of our refinement method is obvious too as the error percentages of disparity maps without refinement are much higher in non-occluded areas.

Table 3. Comparison of the proposed matching cost with traditional gradient cost

Methods (without refinement)	Tsukuba	Venus	Teddy	Cones
Proposed cost	3.05	2.25	9.51	5.09
Traditional cost	4.69	3.76	12.4	8.61

3.2 Sensitivity to radiometric distortion

To test stereo algorithms' sensitivity to radiometric differences, Hirschmüller and Scharstein [20] created 6 datasets: Art, Books, Dolls, Moebius, Laundry and Reindeer, which are shown in **Fig. 6** as well as their ground truth disparity maps. We also present the disparity maps produced by the proposed method. Each dataset is taken using three different exposures and under three different configurations of the light sources. Thus, there will be 9 different image combinations that exhibit significant radiometric differences. To demonstrate the performance under radiometric distortion of the proposed method, we keep the right image unchanged and alter the exposure and lighting conditions of the left image. Thus we can consider the two factors separately. We show the experimental results of "Reindeer" as an example in **Fig. 7**. Obviously, the qualities of the produced disparity maps are very stable throughout the experiments, which can show the strong robustness of the proposed method.

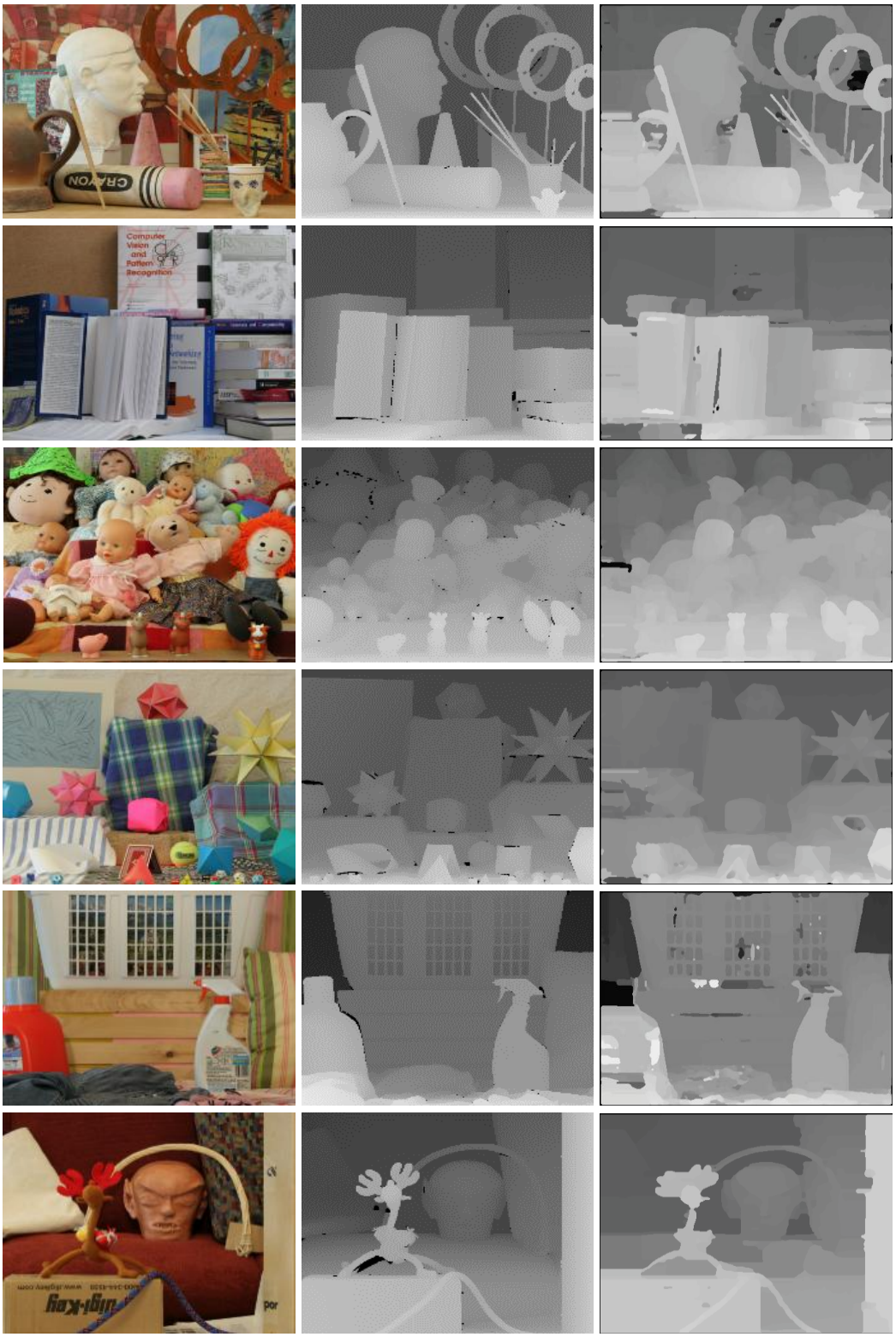


Fig. 6. More experimental results without radiometric difference. From top to down are accordingly the Art, Books, Dolls, Moebius, Laundry, and Reindeer stereo datasets. From left to right are the original color images, ground truth and disparity maps produced by the proposed method.

Fig. 7. Experimental results of the Reindeer pairs by the proposed method with radiometric difference. The first row are the left images under three different exposures and the second row are the cooresponding disparity maps. The third row are the left images under three different light conditions with the cooresponding disparity maps shown in the last row.

As the sensitivity to radiometric distortion is mainly affected by the similarity measure or matching cost, we test three different matching costs including our proposed one. To highlight our proposed matching cost, all of the three compared methods use the adaptive window based cost aggregation to exclude the influence of aggregation ways. The resulting curves are shown in **Fig. 8**. The experiments cover all 3×3 combinations of exposure and light changes which are represented as 1/1 to 3/3. The error rates are the average of all 6 datasets. Seeing from the plots, in every exposure and lighting configuration, the proposed method has the best performance while the SAD method is the worst one. All of the 3 methods have better

performance when the two images are under the same exposure and lighting configurations than when they are under different exposure and lighting configurations. The SAD method is very sensitive to radiometric distortion as its error percentage rise dramatically when left/right images are under different configurations. The gradient method is much better but still not satisfactory. The proposed method is very robust to radiometric distortion as its error rates keep in a low level and vary little throughout when exposure and lighting condition differs. This is because SAD is an intensity based similarity measure and depends on pixels' color or gray intensities which are hypersensitive to radiometric difference. Instead, the proposed method utilizes the gradient information and designs a new matching cost function by integrating the gradient modulus and phase. Hence, our method is not sensitive to color variance and keeps strong robustness to radiometric distortion.

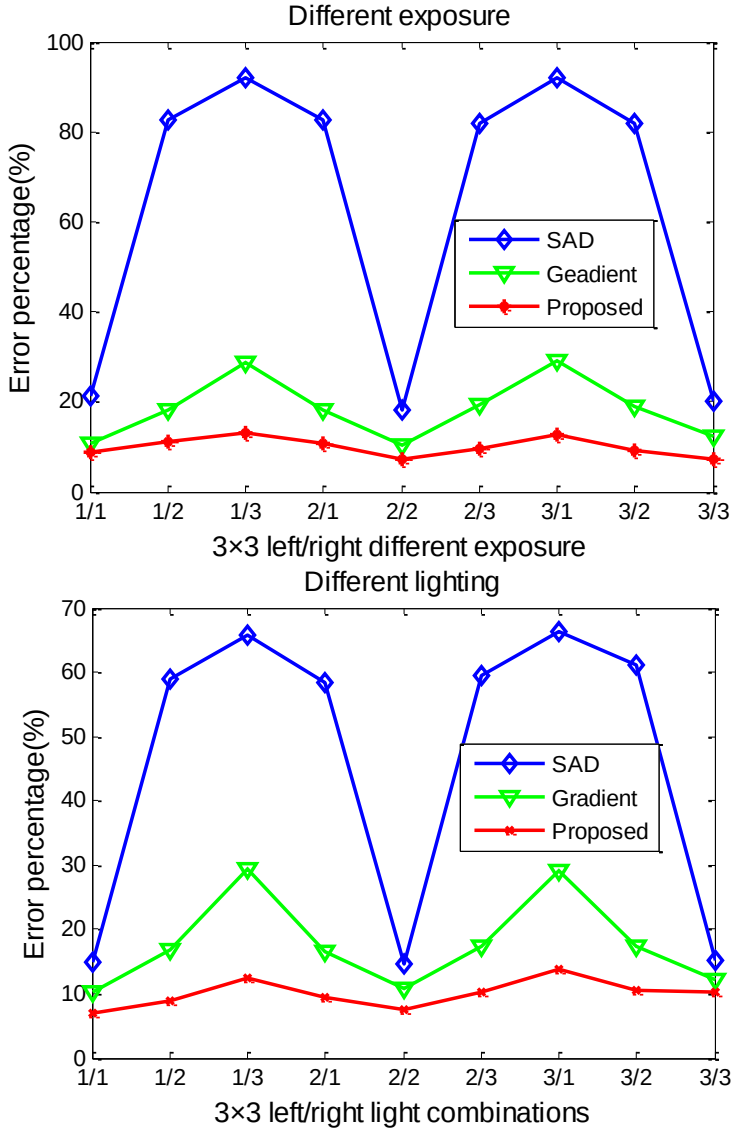


Fig. 8. Performance comparison under 3×3 left/right image combinations that differ in exposure and lighting conditions. (a). Different lightings; (b). Different exposures.

3.3 Stereo scene reconstruction

There are many applications for stereo matching, and three dimension (3D) scene reconstruction is an important one. By re-projecting an image pixel to the 3D space using its depth information, we can reconstruct a complete 3D object model from the 2D images. The quality of scene reconstruction is influenced by the accuracy of acquired depth map to a large extent. To illustrate the quality of the derived matching results, we present reconstructed views of the previous test images in **Fig. 9** in order to gain a further impression of the accuracy and details of the computed depth information. The reconstructing results show that our estimated depth maps are competent to 3D reconstruction tasks.

Fig. 9. 3D scene reconstruction results by using the produced disparity images.

5. Conclusion

This paper presents a novel stereo matching method based on a combined cost function and adaptive window cost aggregation. The improved cost function integrates both modulus and phase components of the gradient vector and then combines them with SAD cost, leading to a superior accuracy. In order to address the window size selecting problem, we introduce an adaptive window solution. The algorithm constructs an adaptive support region for every pixel according to the local color similarity and spatial closeness. Thus, every pixel can get a proper support region for aggregation. In addition, this support region can be reused in the later disparity refinement step. We explore a four-step refinement process, including median filter,

left-right consistence checking, invalid pixels recovering and holes filling. We evaluate our algorithm on the stereo pairs from the Middlebury database. The proposed algorithm matches textureless as well as textured surfaces equally well and can preserve depth discontinuities at the same time. The experimental result comparisons have demonstrated that the proposed method outperforms many local and global methods. Furthermore, the proposed algorithm handles well with radiometric differences, showing strong robustness to radiometric distortion of input images.

Though the proposed method achieves good performance, there are still some aspects to be improved, such as redundancy among the disparity search range, more sophisticated disparity refinement process and parallel implementation for the proposed method will be considered in the next step research.

References

[1] Daniel Scharstein and Richard Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *International Journal of Computer Vision*, vol. 47, no. 1, pp. 7-42, April, 2002. Article (CrossRef Link)

[2] Qinxiong Yang, "A non-local cost aggregation method for stereo matching," in *Proc. of IEEE International Conference on Computer Vision and Pattern Recognition*; pp. 1402-1409, June 16-21, 2012. Article (CrossRef Link)

[3] Kuk-Jin Yoon and In-So Kweon, 'Locally adaptive support weight approach for visual correspondence search," *IEEE Transactions on Pattern Analysis and Machine Intelligence*; vol. 28, no.4, pp. 924-931, April, 2006. Article (CrossRef Link)

[4] Federico Tombari, Stefano Mattoccia and Luigi Di Stefano, "Segmentation based adaptive support for accurate stereo correspondence," in *Proc. of the 2nd Pacific Rim conference on Advances in image and video technology*; no. 4872, pp. 427-438, December 17, 2007.

[5] Daniel Scharstein, *View synthesis using stereo vision*, Phd thesis, January, 1997.

[6] Leonardo De-Maeztu, Arantxa Villanueva and Rafael Cabeza, "Stereo matching using gradient similarity and locally adaptive support-weight," *Pattern Recognition Letters*, vol. 32, no. 13, pp. 1643-1651, October, 2011. Article (CrossRef Link)

[7] Xiaozhou Zhou and Pierre Boulanger, "Radiometric invariant stereo matching based on relative gradients," in *Proc. of IEEE International Conference on Image Processing*, pp. 2989-2992, September 30-October 3, 2012. Article (CrossRef Link)

[8] Ramin Zabih and John Woodfill, "Non-parametric local transforms for computing visual correspondence," in *Proc. of European Conference on Computer Vision*, pp. 151-158, May2-6，1994.

[9] Martin Humenberger, Christian Zinner, Michael Weber, Wilfried Kubinger and Markus Vincze, "A fast stereo matching algorithm suitable for embedded real-time systems," *Computer Vision and Image Understanding*, vol. 114, no. 11, pp. 1180-1202, November, 2010. Article (CrossRef Link)

[10] Andrea Fusiello, Vito Roberto and Emanuele Truco, "Efficient stereo with multiple windowing," in *Proc. of IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 858-863, June 17-19, 1997. Article (CrossRef Link)

[11] Olga Veksler, "Fast variable window for stereo correspondence using integral image," in *Proc. of IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 556-561, June 18-20, 2003. Article (CrossRef Link)

[12] Kang Zhang, Jiangbo Lu and Gauthier Lafruit, "Cross-based local stereo matching using orthogonal integral images," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 19, no. 7, pp. 1073-1079, July, 2009. Article (CrossRef Link)

[13] Qu Yufu, Jiang Ji Xiang and Deng Xiangjin et al, "Robust local stereo matching under varying radiometric conditions," *IET Computer Vision*, vol. 8, no. 4, pp. 263-276, July, 2014. Article (CrossRef Link)

[14] Frederic Besse, Carsten Rother, Andrew Fitzgibbon and Jan Kautz, "PMBP: PatchMatch belief propagation for correspondence field estimation," *International Journal of Computer Vision*, vol. 110, no. 1, pp. 2-13, October, 2014. Article (CrossRef Link)

[15] Liang Wang and Ruigang Yang, "Global stereo matching leveraged by sparse ground control points," in *Proc. of IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 3033-3040, June 20-25, 2011. Article (CrossRef Link)

[16] Nafise Barzigar, Aminmohammad Roozgard, Samuel Cheng and Pramood Verma, "SCoBeP: Dense image registration using sparse coding and belief propagation," *Journal of Visual Communications and Image Representation*, vol. 24, no. 2, pp. 137-147, February 2013. Article (CrossRef Link)

[17] Nicolas Papadakis and Vicent Caselles, "Multi-label depth estimation for graph cuts stereo problems," *Journal of Mathematical Imaging and Vision*, vol. 38, no. 1, pp. 70-82, September, 2010. Article (CrossRef Link)

[18] Xing Mei, Xun Sun, Mingcai Zhou, Shaohui Jiao, Haitao Wang and Xiaopeng Zhang, "On building an accurate stereo matching system on graphics hardware," in *Proc of IEEE International Conference on Computer Vision Workshops*, pp. 467-474, November 6-13, 2011. Article (CrossRef Link)

[19] Heiko Hirschmüller, "Stereo processing by semiglobal matching and mutual information," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 2, pp. 328-341, February, 2008. Article (CrossRef Link)

[20] Daniel Scharstein and Richard Szeliski, "The middlebury stereo vision page," *http://vision. Middlebury.edu/stereo/*, June, 2014.

[21] E. Scott Larsen, Philippos Mordohai, Marc Pollefeys and Henry Fuchs. "Temporally consistent reconstruction from multiple video streams using enhanced belief propagation," in *Proc. of IEEE International Conference on Computer Vision*, pp. 1-8, October 14-21, 2007. Article (CrossRef Link)

[22] Vladimir Kolmogorov and Ramin Rabih, "Computing visual correspondence with occlusions using graph cuts," in *Proc. of IEEE International Conference on Computer Vision*, pp. 508-515, July 7-14, 2001. Article (CrossRef Link)

[23] Leonardo De-Maeztu, Stfano Mattoccia, Arantxa Villanueva and Rafeal Cabeza. "Linear stereo matching," in *Proc. of IEEE International Conference on Computer Vision*, pp. 1708-1715, November 6-13, 2011. Article (CrossRef Link)

Shiping Zhu received the B.Sc. and M.Sc. degrees from Xi'an University of Technology, Xi'an, China, in 1991 and 1994, and the Ph.D. degree from Harbin Institute of Technology, Harbin, China, in 1997. From 1997 to 1999, he was a Postdoctoral Fellow with Beihang University, Beijing, China. From 2000 to 2002, he was a Postdoctoral Fellow with the Brain and Cognition Research Center, Université Paul Sabatier, Toulouse, France. From 2002 to 2004, he was a Postdoctoral Fellow with the Department of Computer Science and Department of Electrical and Computer Engineering, Université de Sherbrooke, Sherbrooke, QC, Canada. Since 2005, he has been an associate professor with the Department of Measurement Control and Information Technology, School of Instrumentation Science and Optoelectronics Engineering, Beihang University, Beijing, China. (E-mail: spzhu@163.com)

Zheng Li received the B.Sc. degree in Measurement and Control Technology and Instrumentation from China University of Geosciences, Wuhan, China in 2012, and he is currently pursuing the M.Sc. degree in Instrumentation Science and Technology at Beihang University, Beijing, China. His research interests include stereo vision, view synthesis and image processing. (E-mail: lizheng900911@163.com)