Proceeding of the IEEE
International Conference on Robotics and Biomimetics
Dali, China, December 2019

An Edge-constrained Iterative Cost Aggregation Method for Stereo Matching

Guanying Huo
College of IOT Engineering
Hohai University
Changzhou, China
huoguanying@hhu.edu.cn

Ying Luo
College of IOT Engineering
Hohai University
Changzhou, China
798782541@qq.com

***Abstract*—To improve the accuracy of stereo matching in low-textured or depth discontinuous regions of images, a new method using edge-constrained iterative cost aggregation is proposed in this paper. Firstly, a bilateral diffusion is used as the preprocessing to reduce inconsistency in the matching direction, which can make the follow-up matching more reliable. Secondly, a matching cost function that combines both color and edge is adopted, and a two-step iterative cost aggression based on the minimum spanning tree (MST) is then proposed. In the first cost aggregation, to avoid accumulation of small weights happened in low-textured regions, an enhanced weight function with coefficient adjustment is presented. And in the second cost aggregation, to appropriately provide weights to neighbors, an edge constraint is introduced. The edge information used in the second aggregation is produced by the random forest method from the disparity map obtained in the first cost aggregation. Left-right consistency check and epipolar constraint are both adopted to further eliminate false edge points. Finally, the disparity refinement is utilized to optimize the disparity map. The proposed method is conducted on Middlebury v.2 data set. Experimental results demonstrate that the proposed method can achieve higher matching accuracy, compared with other five state-of-art methods.**

***Keywords*—stereo matching, bilateral diffusion, edge-constrained, two-step iterative cost aggression, enhanced weight function**

I. INTRODUCTION

Stereo matching can produce a disparity map by finding the corresponding pixels in a pair of stereo images. The disparity map can then be used to calculate depth information, which is crucial for many applications such as binocular ranging [1,2], object detection and tracking [3,4], three-dimensional (3D) scene reconstruction [5-7], and autonomous driving systems [8]. As pointed out by Scharstein [9], stereo matching methods usually include four steps: matching cost computation, cost aggregation, disparity computation and disparity refinement, and can be generally divided into two categories: local methods and global methods.

Local methods [10-13] usually compute the matching cost between two regional blocks according to the similarity of intensity or other cost values at a local window. They usually have low computational complexity and thereby are suitable for real-time applications. However, the usual assumption of local methods is that the corresponding pixels which are determined by similarity of intensity and spatial information also have similar disparity map values, which will not hold in depth discontinuous regions. Therefore, the mismatching rates of local methods are significantly high in low-textured or depth discontinuous regions of images. Selecting an appropriate support window and a better cost function can somewhat improve the match. Yoon et al. [14] proposed an adaptive support-weight approach for stereo matching, which can achieve better results by adaptively adjusting the weights in the window according to the similarity of both color information and geometric distance.

Global methods obtain the optimal disparity map by minimizing a predefined energy function that consists of a data term and an explicit smoothness term, usually without the step of cost aggregation. In the energy function of global methods, the data term is employed to determine the similarity between the matching pixels, while the smoothness term is used to guarantee the consistency between adjacent pixels. Belief propagation (BP) [15-17], graph cut (GC) [18-20] and dynamic programming (DP) [21-24] are classical global methods, which have good performance. Compared with local methods, global methods can better deal with cases of low-textured or depth discontinuous, therefore can achieve higher matching accuracy. However, the computation is usually much expensive, which restrict their application. To improve the time performance, Mei et al. [25] designed a new global stereo matching system using parallel processing, the matching cost function of which combines the features of AD and Census. The cost aggregation is implemented over locally cross-based region followed by a scan-line optimization for achieving with high-precision disparity map. Although the method proposed by Mei is faster than most global methods, it is still slower for high-resolution images, and the parameters of this method are usually difficult to correctly choose.

Different from traditional local or global methods, Yang [26] proposed a non-local cost aggregation algorithm, in which the MST structure substitutes for local support window. In Yang's method, the reference image is regarded as an undirected tree structure, where the nodes correspond to all the image pixels and the edges correspond to all the edges between adjacent pixels. The similarity between any two nodes is determined by their shortest distance on the tree. By traversing the tree in two sequential passes, the cost value of every node receives weight supports from relevant parent or children nodes. Evaluations on Middlebury dataset demonstrate that this non-local cost aggregation method can significantly improve the accuracy of the stereo matching.

Mei et al. [27] proposed a segment-tree (ST) based algorithm integrating MST with graph-based segmentation. The initial disparity map as a supplement is blended into the weight function, which will then contribute to a more stable tree structure. Therefore, the final disparity map obtained is much smoother and more precise. Zhang et al. [28] proposed a generic cross-scale cost aggregation framework for multi-scale interaction by introducing regularization terms. A

This work was funded by the National Natural Science Foundation of China under Grant Nos. 41876097.

978-1-7281-6321-5/19/$31.00 ©2019 IEEE

variety of excellent cost aggregation methods have been incorporated into this framework. Using the color-depth weight, Yao et al. [29] iteratively rebuilt the tree structure and obtained an enhanced disparity map, with a remarkable improvement in matching accuracy in textureless regions. Gao et al. [30] proposed a modified initial cost function containing both vertical gradient information and Gemen-McClure function, which can produce a more stable initial disparity map. A more reasonable tree structure is developed for robust cost aggregation using an improved segmentation method that effectively reduces the matching error rates in discontinuous areas. Huang et al. [31] proposed a disparity refinement method with O(1) computational complexity based on the MST structure, where the additional operation is replaced by a maximum operation in belief propagation. According to experimental results on high-resolution images, both the computation efficiency and the matching accuracy have been greatly improved.

The MST-based methods mentioned above, except the ST method, usually perform cost aggregation on the entire image without any additional constraint. To deal with depth discontinuities, it will be better to take into consideration any constraint that can reduce irrelevant weights received from different depth areas. Moreover, in low-textured regions, the color differences are basically very small and quite close to zero. Consequently, large edge weights are inevitably accumulated along the traversing paths, which leads to the small-weight-accumulation problem [31].

Mainly for the above two reasons, in this paper, to improve the matching accuracy of stereo image pairs, a new stereo matching method is proposed. A bilateral diffusion in horizontal direction is firstly adopted, which can suppress clutter background and better guarantee matching consistency. A two-step iterative cost aggression based on the minimum spanning tree (MST) is then proposed. In the first cost aggregation, an enhanced weight function with coefficient adjustment is proposed for avoiding accumulation of small weights happened in low-textured regions. After the first cost aggregation, edge information will be carefully obtained and then used in the second cost aggregation. The random forest algorithm is utilized for edge detection on the initial disparity map obtained from the first cost aggregation. Left-right consistency (LRC) check and epipolar constraint are then performed for detecting leftmost occlusions. The corresponding edge points in the leftmost occlusions are discarded for eliminating irregular edges generated by the random forest algorithm. The generated edge will be incorporated into the cost function of the second cost aggregation, so that only pixels in the same region restricted by edges can provide enough weights to neighbors, which can improve the matching accuracy in depth discontinuous regions. Finally, after the second cost aggregation, a disparity map is produced and then optimized with the disparity refinement. Performance evaluation on Middlebury data sets demonstrates that the proposed method can achieve higher matching accuracy, compared with other five state-of-art methods. The main contributions of this paper are as follows.

(1) To reduce inconsistency in the matching direction, a lightweight bilateral diffusion in horizontal direction is firstly adopted, which can suppress clutter background.

(2) In the first cost aggregation, to avoid accumulation of small weights happened in low-textured regions, an enhanced weight function with coefficient adjustment is presented.

(3) In the second cost aggregation, to improve the matching accuracy in depth discontinuous regions, an edge constraint is introduced, so that only pixels in the same region restricted by edges can provide sufficiently weights to neighbors.

II. Materials and Methods

The proposed method consists of four steps: (1) bilateral diffusion, (2) matching cost computation, (3) two-step iterative cost aggregation, and (4) disparity refinement.

A. Bilateral Diffusion

To guarantee consistency in the matching direction, it is useful to suppress clutter background caused by noise or small texture. A bilateral diffusion in horizontal direction is hereby proposed for smoothing the stereo images. The diffused intensity of each pixel is determined by its two adjacent pixels on the left and right. The diffusion is applied to the original RGB image rather than to a converted grayscale image, which can fully consider three-channel information of images. Using the original RGB image as input, the diffusion adopts the following equation for processing:

$$I_s^i(p) = I^i(p) + \frac{1}{3}\lambda \sum_{j\in\{R,G,B\}} [h_1(p)\cdot\nabla_1 I^j(p) + h_2(p)\cdot\nabla_2 I^j(p)],\ i\in\{R,G,B\} \quad (1)$$

where $I^i(p)$ denotes the pixel intensity of pixel p in channel i of the input RGB image; the latter item is the average of the RGB three-channel image information, where λ controls the diffusion speed with a default value of 0.2 according to [32]; the symbol ∇ represents the derivative operation, which can be defined by Equations (2) and (3):

$$\nabla_1 I^j(p) = I^j(p-1) - I^j(p)\,, \quad (2)$$

$$\nabla_2 I^j(p) = I^j(p+1) - I^j(p)\,, \quad (3)$$

where pixel p-1 and p+1 are the two nearest pixels on the left and right side of pixel p. The conduction coefficient h denotes the weight of each difference and is calculated according to:

$$h_1 = \exp\left(\frac{-|\nabla_1 I^j(p)|}{\alpha}\right), \quad (4)$$

$$h_2 = \exp\left(\frac{-|\nabla_2 I^j(p)|}{\alpha}\right), \quad (5)$$

where the constant α can be fixed or be adaptively changed according to the noise level. In this paper, the constant α is set to be 0.1. When the pixels are on a high-contrast boundary, the differences between adjacent pixels are large and the exponential term h will be small. Therefore, the pixels on the other side of the boundary will hardly contribute to the diffused intensity values so that the boundary can be well kept after diffusion.

B. Matching cost computation

Matching cost function is the key of stereo matching algorithm, which will directly affect the accuracy of the final disparity map. In this paper, both color and gradient information are combined into the initial matching cost function $C(p, d)$. The three-channel RGB information is used to reduce mismatches caused by those pixels with the same grayscale but different colors when only using grayscale information. The absolute difference of the color $C_{AD}(p, d)$ and the gradient $C_{Grad}(p, d)$ after thresholding are formulated according to (6) and (7), respectively.

$$C_{AD}(p,d) = \min(\frac{1}{3}\sum_{i=R,G,B} |I_L^i(p) - I_R^i(p_d)|, th_1) , \tag{6}$$

$$C_{Grad}(p,d) = \min\left\{\begin{matrix} \frac{1}{3}(\sum_{i=R,G,B} |\sqrt{(\nabla_x I_L^i(p))^2 + (\nabla_y I_L^i(p))^2} \\ -\sqrt{(\nabla_x I_R^i(p_d))^2 + (\nabla_y I_R^i(p_d))^2} |), th_2 \end{matrix}\right\}, \tag{7}$$

where $I_L^i(p)$ and $I_R^i(p_d)$ respectively denote the pixel values in the left and right images of channel i; th_1 and th_2 are the two thresholds used for truncating the color and gradient differences, respectively; ∇_x and ∇_y represent derivative operation in x and y directions, respectively. The initial matching cost function can then be expressed as follows:

$$C(p,d) = w_1 C_{AD}(p,d) + w_2 C_{Grad}(p,d) , \tag{8}$$

where w_1 and w_2 are the two weights that balance the two cost terms of color and gradient, with the constraint $w_1 + w_2 = 1$. In this paper, th_1=7, th_2=2, and w_1=0.11 are adopted according to [26].

C. Two-step Iterative Cost Aggregation

As mentioned in Section 1, because the color differences are basically very small and quite close to zero, one shortcoming of the original MST-based cost aggregation is that large edge weights can be unreasonably accumulated in textureless or low-textured regions, which is called the small-weight-accumulation problem. Besides, because the cost aggregation is performed on the whole image without incorporating any prior constraints, irrelevant weights could be received from different depth areas. In this case, the depth discontinuities would be incorrectly destroyed. Therefore, based on the MST structure, an edge-constrained iterative cost aggression is proposed in our method. Firstly, the weight function with coefficient adjustment is presented in the first aggregation. After edge detection on the initial disparity map, an edge constraint is then applied in the second cost aggregation for better performance at depth discontinuities.

1) Coefficient Adjustment-based First Cost Aggregation

For MST-based cost aggregation, the reference image I can be regarded as a four-neighborhood, undirected graph $G(V, E)$, in which each node in V corresponds to one pixel in the image I and each edge in E connects the neighboring pixels. For an edge e that connects pixels s and r, its weight $w(s, r)$ is calculated according to:

$$w(s,r) = \max_{i\in\{R,G,B\}} |I^i(s) - I^i(r)| , \tag{9}$$

where $\max_{i\in\{R,G,B\}} |I^i(s) - I^i(r)|$ refers to the maximum absolute difference among the three RGB channels.

It's straightforward to define the similarity between pixels p and q according to the distance, which denotes the contribution of p to q on the aggregation path. The similarity function $S(p, q)$ is computed as follows:

$$S(p,q) = S(q,p) = \exp(-\frac{D(p,q)}{\sigma}) , \tag{10}$$

where $D(p, q)$ is the distance between p and q, which is the sum of all the edge weights $w(s, r)$ along the MST; σ is a user-specified scale parameter for adjusting the similarity between two pixels, the value of which is set to 0.1 in this paper. For MST method, because the intensities of pixels in low-textured or textureless regions are almost identical, the corresponding values of edge weights are consequently very small and close to zero, which leads to the small-weight-accumulation problem. In order to solve this problem, an enhanced weight function is proposed in this paper. In the proposed function, when the difference between neighboring pixels is smaller than 2, the value of edge weight is amplified by Π times of the original one. Because the similarity function is a decreasing function with regard to the weight, the similarity will decrease when the weight increases. Therefore, the similarity value is effectively suppressed in low-textured regions. The enhanced weight function is defined as follows:

$$w'(s,r) = w'(r,s) = \begin{cases} \max_{i\in R,G,B} |I^i(s) - I^i(r)| \times \Pi, & if \max_{i\in R,G,B} |I^i(s) \text{-} I^i(r)| \le 2 \\ \max_{i\in R,G,B} |I^i(s) - I^i(r)|, & \text{else} \end{cases} \tag{11}$$

where Π is the coefficient used for adjusting similarity values in low-textured regions.

The cost aggregation value $C_d^A(p)$ can be regarded as the sum of all the matching costs multiplied with their corresponding weights:

$$C_d^A(p) = \sum_{q\in I} S(p,q) C_d(q) , \tag{12}$$

where $C_d(q)$ denotes the cost value when a pixel q in the image is assigned to a disparity level d.

Cost aggregation need to be performed on all the pixels at all disparity levels. Moreover, all the cost values can be calculated effectively by traversing two sequential passes, and the aggregated cost value of each pixel is only related to its parent node and subtrees. Fig. 1 demonstrates the cost aggregation, which can be divided into two steps.

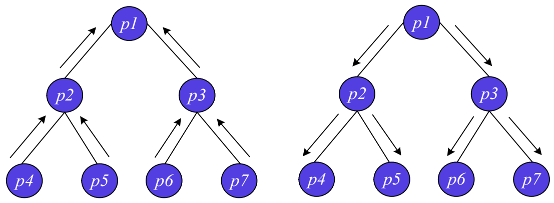
① The first step is performed from leaf nodes to root node, as can be seen in Fig. 1(a). For a pixel p, the aggregated value is updated once its subtrees have been traversed:

$$C_d^{A\uparrow}(p) = C_d(p) + \sum_{q\in Ch(p)} S(p,q) \cdot C_d^{A\uparrow}(q) , \tag{13}$$

② The second step is performed from the root node to leaf nodes, as can be seen in Fig. 1(b). For a pixel p, its final cost aggregation value is calculated with its parent Pr(p) as follows:

$$\begin{aligned} C_d^A(p) &= C_d^{A\uparrow}(p) + S(\mathrm{Pr}(p),p) \cdot [C_d^A(\mathrm{Pr}(p)) - S(\mathrm{Pr}(p),p) \cdot C_d^{A\uparrow}(p)] \\ &= S(\mathrm{Pr}(p),p) \cdot C_d^A(\mathrm{Pr}(p)(1 - S^2(\mathrm{Pr}(p),p)) \cdot C_d^{A\uparrow}(p) \end{aligned} , \tag{14}$$

where Pr(p) is the parent node of the pixel p.



(a) Leaf to root node (b) Root to leaf nodes

Fig. 1. The first cost aggregation based on MST.

2) Edge-constrained Second Cost Aggregation

After the first cost aggregation based on coefficient adjustment, better results can be achieved in low-textured regions. However, in the first cost aggregation, the tree is constructed only with color information, which may cause inevitable loss of 3D cues. Moreover, it is worth noting that each pixel receives support weights from all the neighboring pixels. In this case, those neighboring pixels in other adjacent regions will inevitably made adverse contributions. Because different regions are usually distinguished by edges, edge constraints are indispensable for dealing with ambiguities at depth discontinuities. Therefore, in the second cost aggregation, edges achieved by the random forest method [33] on the initial disparity map are used as a constraint.

Firstly, from the first aggregation the cost volumes obtained are obtained with the winner-takes-all (WTA) strategy, which is employed to look for the optimal disparity value. For a pixel p, its disparity value is defined as follows:

$$D_{IS}(p) = \arg\min_{d \in R}(C_d^A(p)) \,, \qquad (15)$$

where R is the searching scope of disparity, and $C_d^A(p)$ is the cost volume obtained from the first cost aggregation.

Secondly, the edge weight function is updated with color and depth cues. Then, a hybrid MST structure is re-established for the iterative cost aggregation:

$$w_H(s,r) = (1-k)\,|\,D_{IS}(s) - D_{IS}(r)\,| + k\,|\,I(s) - I(r)\,| \,, \qquad (16)$$

where D_{IS} denotes the disparity map obtained from the first cost aggregation, and k is a parameter balancing the two measures, with a default value of 0.5.

Thirdly, the second cost aggregation with the edge constraints is illustrated in Fig. 2. Different from the first aggregation, in the edge-constrained cost aggregation, the weights from the two pixels that belong to different regions are suppressed. The edge constraint is used for adjusting the aggregated cost value between any two nodes from different depth regions. If the two nodes belong to two different depth regions, the edge weight will be further suppressed by spatial information. That is, the larger the distance, the smaller the edge. Moreover, the edge-constrained cost aggregation develops in two sequential passes, i.e. from leaf nodes to the root node and from the root node to leaf nodes.

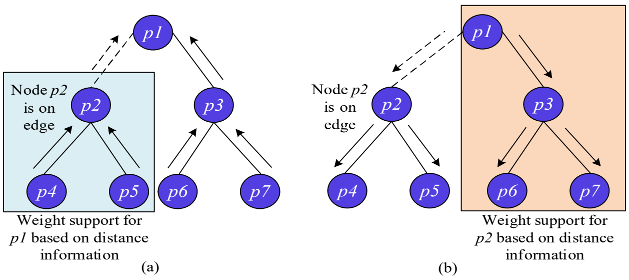

p1
Node p2 is on edge
p2
p3
p4
p5
p6
p7
Weight support for p1 based on distance information
(a)
p1
Node p2 is on edge
p2
p3
p4
p5
p6
p7
Weight support for p2 based on distance information
(b)

Fig. 2. Edge-constraint based second cost aggregation, where *p1* is the root node. (a) Leaf to root node; (b) root to leaf nodes.

① From leaf nodes to root node. As shown in Fig. 2 (a), if the child nodes (*p2- p7*) locate in same depth region as the root node (*p1*), the aggregation strategy is the same as the first cost aggregation. When child nodes (*p2, p4, p5*) indicated in the blue rectangle belong to same depth edge but other nodes (*p1, p3, p6, p7*) do not, additional edge constraints with regard to distance information will be imposed on the contribution from *p2* to *p1*. For each pixel p, its cost aggregation value is defined as follows:

$$C_d^{A\uparrow\uparrow}(p) = \begin{cases} C_d^A(p) + \sum\limits_{q \in Ch(p)} S_H(p,q) \cdot C_d^{A\uparrow\uparrow}(q), & q \in \text{edge} \\ C_d^A(p) + \exp(-\dfrac{L(p,q)}{\beta}) \sum\limits_{q \in Ch(p)} S_H(p,q) \cdot C_d^{A\uparrow\uparrow}(q), & q \notin \text{edge} \end{cases}, \qquad (17)$$

where $C_d^A(p)$ denotes the cost value of pixel p from the first aggregation; $Ch(p)$ is the set of children nodes of p; $S_H(p,q)$ is the similarity function in which the edge weight is replaced by Equation (16); $C_d^{A\uparrow\uparrow}(q)$ is the second aggregated cost value; and $L(p,q)$ denotes the distance between the two nodes p and q along the hybrid MST structure.

② From root node to leaf nodes. As shown in Fig. 2(b), when performing the aggregation from root node *p1* to leaf node *p2*, the node *p2* will receive support from its parent node *p1* indicated in the orange rectangle. Because *p2* is on depth edges, the support from *p1* to *p2* will be similarly suppressed by the distance information between the two nodes. For each pixel p, its final aggregated cost value is defined as follows:

$$C_d^{A'}(p) = \begin{cases} C_d^{A\uparrow\uparrow}(p) + S_H(\mathrm{Pr}(p),p) \cdot [C_d^{A'}(\mathrm{Pr}(p)) - S_H(\mathrm{Pr}(p),p) \cdot C_d^{A\uparrow\uparrow}(p)], \quad \mathrm{Pr}(p) \in \text{edge} \\ C_d^{A\uparrow\uparrow}(p) + \exp(-\dfrac{L(p,\mathrm{Pr}(p))}{\beta}) \cdot S_H(\mathrm{Pr}(p),p) \cdot [C_d^{A'}(\mathrm{Pr}(p)) - \\ S_H(\mathrm{Pr}(p),p) \cdot C_d^{A\uparrow\uparrow}(p)], \quad \mathrm{Pr}(p) \notin \text{edge} \end{cases} ,(18)$$

where $\mathrm{Pr}(p)$ is the parent nodes of the pixel p; and β is the parameter used to adjust the degree of distance constraints.

In the second cost aggregation, the edge constraint used is obtained by the random forest method on the initial disparity map. The random forest method can take full advantage of the structure in local image patches to learn an accurate and computationally efficient edge detector. And the edges of the disparity map can be well depicted by this method. However, because the existence of the leftmost occluded pixels in the left image, there are erroneous edges when edge detection is performed on the initial disparity map obtained from the first cost aggregation. The erroneous edges would cause adverse effects when performing the edge-constrained cost aggregation. Therefore, to eliminate this effect, the LRC check and the epipolar constraint are together used to detect leftmost occluded pixels.

The left image is selected as the reference image and LRC is used to detect those unstable pixels according to the following constraint:

$$|\,D_L(x,y) - D_R(x - D_L(x,y), y)\,| > 1 \,, \qquad (19)$$

where D_L and D_R represent the left and right disparity maps, respectively. The disparity values of those unstable pixels are set to 0. The unstable pixels are further classified into mismatches and occlusions [34] in term of epipolar constraint.

For an unstable pixel $p(x_l, y_l)$ in the left image, the epipolar line in the right image $e(x,y,d)$ is defined as follows:

$$e(x,y,d) = \{(x,y)_{Right} \mid x = x_l - d, y = y_l, d \in [0,R]\} \,. \qquad (20)$$

If the epipolar line $e(x,y,d)$ does not intersect with $D_R(x,y)$, p is considered as an occluded pixel, otherwise it is a mismatched pixel. Then the occlusions are further divided into the leftmost and non-border occlusions. According to the maximum disparity range of image pairs, the leftmost occlusion can be obtained. After that, all the pixels in the leftmost occlusion of the edge image are regard as invalid pixels, of which the intensities are set to 0.

D. Disparity Refinement

After cost aggregation, a raw disparity map is calculated by the WTA strategy. However, the disparities in occluded and depth discontinuous regions are not available in the raw disparity map. Therefore, disparity refinement is necessary in these regions. Firstly, the LRC check is implemented and all the pixels are divided into stable and unstable pixels. Then the cost values of unstable pixels are set to 0 while others are adjusted using the function in [26]. For a pixel p with disparity d, the new cost value is defined as follows:

$$C_d^{new}(p) = \begin{cases} |d - D(p)|, & p \text{ is stable} \\ 0, & p \text{ is unstable} \end{cases}, \tag{21}$$

where $C_d^{new}(p)$ denotes the new cost value when pixel p is assigned to a disparity d. To prevent over-suppression of cost volume, the original cost aggregation is implemented on the new cost volume. Because unstable pixels can still obtain propagated disparity values from stable pixels, the disparity values in occluded and mismatched regions can be largely corrected. Finally, a dense and more accurate disparity map can be obtained.

III. Results and Discussion

To demonstrate the effectiveness of the proposed method, in this section, the proposed method will be tested on Middlebury stereo benchmark [35] extended data. The experimental results of the proposed method will be compared with those of other five state-of-the-arts stereo matching algorithms, and both quantitative and qualitative evaluations are given here. The proposed method is implemented using C++ language and all the experiments are conducted on a PC with Intel (R) Core (TM) i5-7500 CPU @4.00GHz.

To evaluate experiments quantitatively, the mismatching rate B is calculated as follows:

$$B = \frac{1}{N} \sum_{p \in \mathrm{R}_g} |d_E(p) - d_T(p)| > \sigma_d , \tag{22}$$

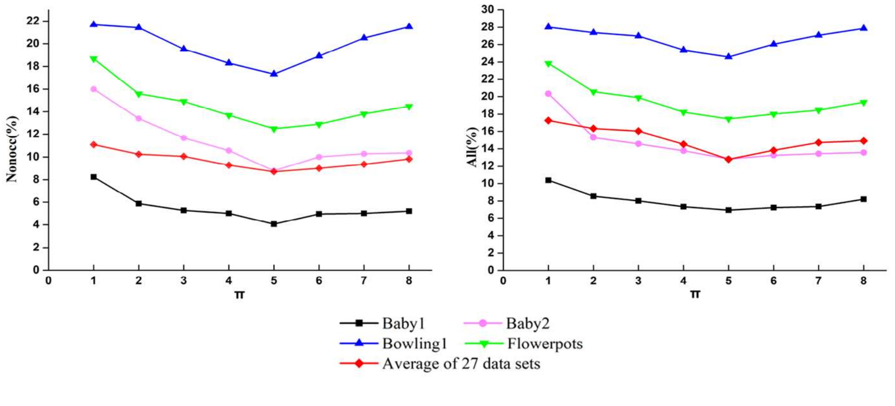
where $d_E(p)$ is the estimated disparity value of pixel p; $d_T(p)$ is the real disparity value; N denotes the total number of pixels in the tested region R_g of the image; σ_d is the disparity threshold with a value of 1 used in this paper for better results.

A. Parameter settings

Appropriate parameters are of great importance for better results. For the proposed method, the empirical values of most of the parameters have been given in this paper, and other parameters are carefully determined according to the algorithm evaluation. Because the coefficient Π in the improved weight function is a key parameter, specific experimental tests are carried out to choose a suitable value for it. We mainly show the test results on 27 data sets. According to the distribution of the maximum disparity ranges, four representative images pairs *Baby1*, *Baby2*, *Flowerpots* and *Bowling1* are selected for detailed analysis. The maximum disparity values are 46, 52, 61 and 77, respectively. In this section, we conducted experiments with the weight coefficient Π changing from 1 to 8 to choose an appropriate value. The percentages of mismatched pixels varying with different weight coefficients are shown in Fig. 3. It can be seen that the error matching rate is the lowest when $\Pi = 5$, whether in non-occluded regions or not. If the coefficient Π becomes larger or smaller, the matching error rates will increase gradually. When the weight coefficient Π decreases from 1 to 5, the mismatching rate drops to almost half of the original value for image pairs *Baby1* and *Baby2*. Therefore, $\Pi = 5$ is used in this paper.

B. Comparison and evaluation of six algorithms

In order to evaluate the performance of the proposed method, a comparison of the proposed method and five state-of-the-art methods is also given here. The five excellent methods with high matching accuracy are MST method [26], Segment-Tree (ST-2) [27], Cross-Scale (CS-MST) [28], Weighted Cost Propagation with Smoothness Prior (WCPSP) [36], Iterative Color-Depth (MST-CD2) [29]. The MST method is a non-local cost aggregation method based on the minimum spanning tree. By regarding the whole image as a tree structure, the MST method breaks the limitation of traditional window-based methods and greatly improves the efficiency and matching accuracy. By performing tree filtering in segmentation regions, the ST-2 method further improves the matching accuracy. The CS-MST method, which integrates the MST method and designs a generalized cost aggregation model by adopting multi-scale interactive information, can perform better than the MST method in many cases. The WCPSP method combines weighted cost aggregation of local methods and smoothness constraints of global methods to obtain more accurate depth information. The MST-CD2 method, which adopts iterative aggregation based on color-depth information, can improve the matching accuracy in textureless regions.



(a)

(b)

Fig. 3. The percentage of error matching with respect to different values of coefficient $\Pi \in [1,8]$. (a) Results in non-occluded regions; (b) Results in all regions.

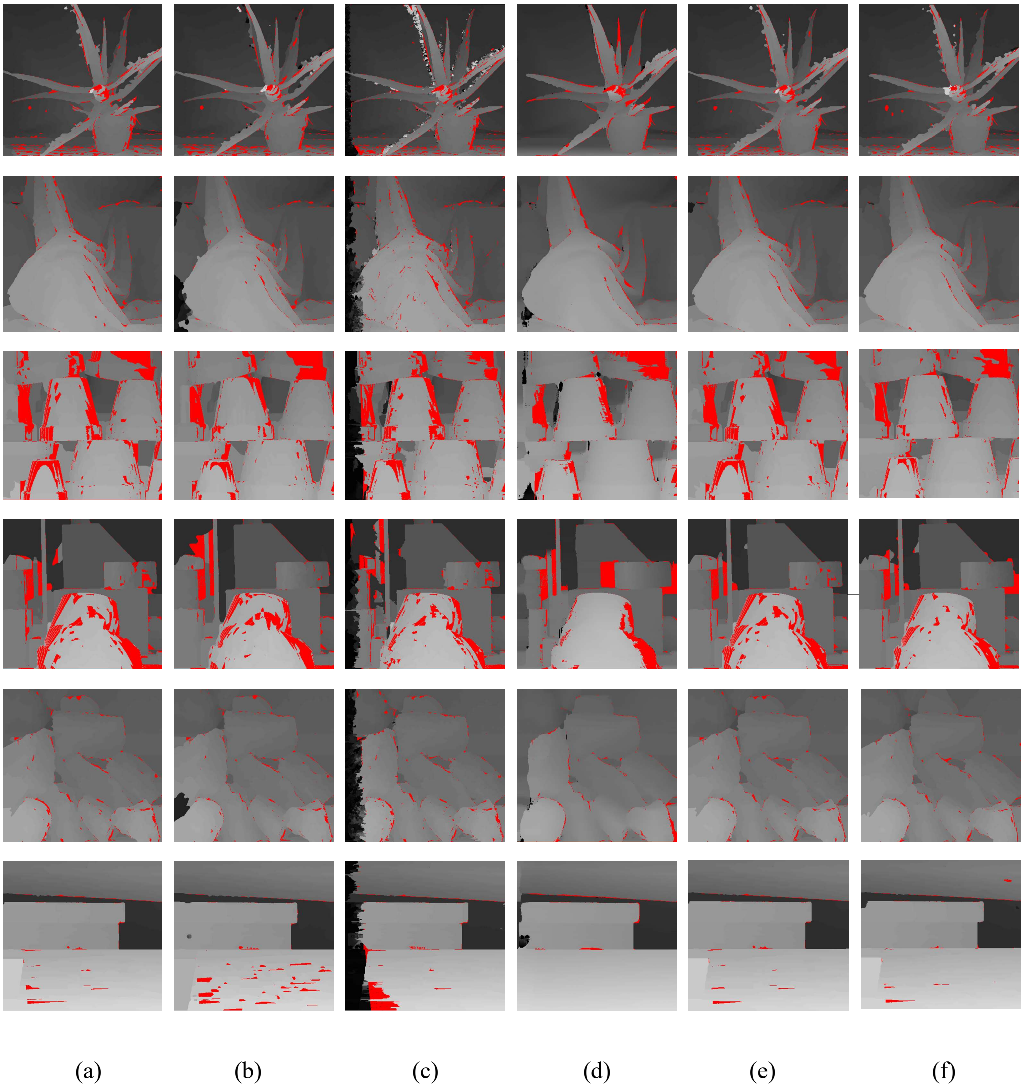


Fig. 4. Disparity maps obtained by six algorithms. (a) MST; (b) ST-2; (c) CS-MST; (d) WCPSP; (e) MST-CD2; (f) The proposed method.

Six representative pairs of images, namely *Aloe*, *Cloth3*, *Flowerpots*, *Lampshade1*, *Rock2*, and *Wood2*, are selected for visual comparison among the methods mentioned above. The corresponding maximum disparity values are 71, 56, 61, 65, 57, and 73 pixels for these six images. Fig. 4 demonstrates the disparity maps of the six pairs of images obtained by the six methods, in which the red pixels represent the mismatched ones in the non-occluded regions. Compared with the other five methods, the proposed method usually has fewer mismatches and can achieve higher matching accuracy. The proposed method also has good performance in the typically tough regions such as the low-textured and discontinuous regions. The disparity maps obtained by the proposed method are generally smoother and more reliable.

For *Aloe*, the proposed method has fewer mismatched pixels on the edges and centers of the leaves compared with other methods. With the edge-constrained aggregation, the proposed method can also well deal with the edges of the flower pot. As for *Flowerpots*, while the other methods fail to give clear boundaries of flowerpots, the proposed method can recover the accurate disparities at the edges of the middle bottom flowerpot to a great extent. For *Lampshade1*, owing to the uneven illumination, it is also very difficult to achieve accurate disparity maps. Nevertheless, the proposed method can still achieve more reliable result. For example, in the regions of yellow trapezoidal wedge and the V-shaped block, there are almost no mismatched pixels in the result obtained by using our method. Although the WCPSP can also work well in these regions, it produces many mismatched pixels on both sides of the V-shaped block. This is because the foreground block and the background wedge have similar color, which causes incorrect cost values propagated on a horizontal tree.

For *Cloths4* and *Rock2*, owing to the iterative cost aggregation, the mismatched pixels of the disparity map in depth discontinuous regions are significantly reduced compared with the other methods. Therefore, more reliable disparity maps without black holes are obtained by the proposed method. For *wood2*, because the whole image has a similar color, it's quite a challenge to achieve reliable result in the depth discontinuous regions. Compared with the MST and ST-2 methods, our method performs much better in the low-textured region of the bottom wood. Meanwhile, the CS-MST and WCPSP methods, especially the former, produce some unreasonable black regions in the bottom wood region. It can be clearly seen that for *wood2*, among all the methods, the proposed method and the MST-CD2 method can achieve much better results.

To further compare the six methods, the quantitative evaluation results on Middlebury *v.2* data set of the six

TABLE I. QUANTITATIVE EVALUATION OF THE SIX METHODS ON 27 IMAGE PAIRS

Image pairs	**MST**	**ST-2**	**CS-MST**	**WCPSP**	**MST-CD2**	**Proposed method**
Aloe	5.02_5	4.42_4	4.88_6	4.08_2	4.19_3	$\mathbf{3.58}_1$
Art	10.24_5	9.98_4	10.69_6	8.38_2	9.04_3	$\mathbf{7.77}_1$
Baby1	10.50_6	4.20_3	8.21_4	$\mathbf{3.34}_1$	8.49_5	4.05_2
Baby2	17.96_6	15.96_5	13.54_3	$\mathbf{3.25}_1$	15.28_4	8.79_2
Baby3	7.34_6	5.16_3	5.59_4	$\mathbf{2.60}_1$	5.98_5	4.58_2
Books	10.09_5	10.03_4	10.66_6	$\mathbf{6.43}_1$	9.67_3	8.63_2
Bowling1	22.56_6	21.72_5	19.56_4	$\mathbf{11.37}_1$	18.49_3	17.33_2
Bowling2	11.60_6	11.04_5	10.11_4	5.86_2	9.62_3	$\mathbf{5.73}_1$
Cloth1	0.54_4	0.51_3	0.63_5	0.69_6	0.45_2	$\mathbf{0.29}_1$
Cloth2	4.19_5	3.59_4	4.35_6	2.44_2	3.01_3	$\mathbf{2.35}_1$
Cloth3	2.16_5	1.63_2	2.90_6	1.64_3	1.66_4	$\mathbf{1.29}_1$
Cloth4	1.50_4	1.29_3	1.88_6	1.67_5	1.10_2	$\mathbf{0.93}_1$
Dolls	6.16_6	5.10_4	5.89_5	$\mathbf{4.07}_1$	4.92_3	4.38_2
Flowerpots	22.26_6	15.42_3	16.79_4	13.55_2	18.52_5	$\mathbf{12.49}_1$
Lampshade1	12.81_6	11.71_5	10.10_2	10.33_3	10.44_4	$\mathbf{8.25}_1$
Lampshade2	12.29_5	13.30_6	12.08_4	$\mathbf{7.26}_1$	10.75_3	9.63_2
Laundry	11.42_3	11.93_5	11.92_4	12.07_6	10.94_2	$\mathbf{10.81}_1$
Midd1	23.15_4	21.23_2	24.43_5	28.83_6	22.49_3	$\mathbf{20.36}_1$
Midd2	32.76_5	$\mathbf{20.41}_1$	20.57_2	35.14_6	30.87_4	29.03_3
Moebius	7.87_5	7.64_4	7.57_3	10.53_6	$\mathbf{7.38}_1$	7.41_2
Monopoly	20.25_4	$\mathbf{19.03}_1$	21.03_5	28.16_6	19.45_3	19.09_2
Plastic	46.69_6	38.77_2	45.02_4	43.25_3	45.88_5	$\mathbf{30.40}_1$
Reindeer	9.67_5	7.12_3	9.79_6	$\mathbf{6.12}_1$	8.80_4	6.98_2
Rocks1	2.76_5	2.35_4	3.35_6	1.79_2	2.25_3	$\mathbf{1.62}_1$
Rocks2	2.03_5	1.66_4	2.28_6	1.35_2	1.60_3	$\mathbf{1.25}_1$
Wood1	11.92_6	4.87_2	10.18_4	$\mathbf{2.85}_1$	10.30_5	7.24_3
Wood2	1.10_4	2.82_5	3.17_6	0.78_2	0.89_3	$\mathbf{0.67}_1$
Avg. Error (%)	12.11_6	10.11_3	11.01_5	9.55_2	10.83_4	$\mathbf{8.70}_1$
Avg. Rank	5.11_6	3.56_4	4.67_5	2.78_2	3.44_3	$\mathbf{1.51}_1$
Avg. time(s)	$\mathbf{0.80}_1$	1.17_2	4.21_6	3.62_5	2.29_4	1.89_3

methods are presented in Table 1. The 27 extend data sets of *v.2* Middlebury are tested in non-occluded regions. In Table 1, The normal numbers are percentages of the mismatched pixels of disparity maps with the error threshold set to1.0, while the subscript numbers in each row are the relative ranks of different methods. The last three columns are the average error rate, average ranking and average running time. It can be seen from Table 1 that our method performs the best in terms of the overall accuracy and the average ranking.

The table reveals intuitively some momentous characteristics. Among all the methods, MST is the most fundamental one, which usually has the lowest rank in the average error rate. The CS-MST method has made some improvements on the MST method. Compared with MST and CS-MST methods, the other methods have more improvements, and therefore perform even better for the 27 stereo pairs. Among all the other methods, the WCPSP method is the most excellent one. For the WCPSP method, the cost volumes are aggregated on the horizontal tree structure, which is quite different from that of the MST method. The WCPSP method ranks 2*nd* in term of the average error rate among all the methods. It can also be seen that our method ranks 1*th* most and the average error rate of our method is the lowest with 8.70%. More precisely, our method reaches 15/27 of top performance on 27 image pairs with the lowest error rate among all the six methods. The percentages of the mismatched pixels produced by our method decline distinctly in *Bowling2*, *Flowerpots* and *Lamdshade1.* According to the quantitative evaluation results in Table 1 and the demonstrated disparity maps in Fig. 4, it can be known that the proposed method can achieve more accurate and smoother results than the other methods in most cases.

For practical applications such as autonomous driving, time performance is also very important. The average running times of different methods are also given in Table 1. Among all the methods, the MST method is the fastest, and the ST-2 method is almost as fast. The proposed method is a little slower than the MST and ST-2 methods, and faster than the MST-CD2, WCPSP and CS-MST methods. More precisely, the proposed method takes a little more than twice as long as the MST method, due to the additional bilateral diffusion and the second edge-constrained cost aggregation, which contribute to more accurate results. Compared with the WCPSP method, the proposed method saves half the time while achieving more accurate results.

C. Conclusions

In order to reduce mismatches in regions like low-textured or depth discontinuous, an edge-constrained iterative cost aggregation method for stereo matching is proposed in the paper. Different from the original MST method, the proposed method adopts a two-step iteration strategy: in the first iteration, an enhanced weight function with coefficient adjustment is presented to deal with the small-weight-accumulation problem happened in the low-textured region; meanwhile in the second iteration, an edge constraint is introduced so that only pixels in the same region restricted by edges can provide sufficiently weights to neighbors, which can thereby reduce mismatches in depth discontinuous regions. Moreover, the proposed method also adopts a lightweight bilateral diffusion in horizontal direction to better reduce inconsistency in the matching direction. Experimental results demonstrate that the proposed method can achieve more accurate results with less mismatches in low-textured or depth discontinuous regions, when compared with other five state-of-the-art methods. Among all the six methods, although the proposed method is slight slower than the MST and ST-2 methods, it is still fast and better than the other three method. Compared with the WCPSP method, which is also very excellent in the matching accuracy, the proposed method saves half the time while achieving more accurate results.

REFERENCES

[1] Q. L. Wang, J. Y. Li, H. K. Shen, T. T. Song, and Y. X. Ma, "Research of multi-sensor data fusion based on binocular vision sensor and laser range sensor," Key Engineering Material, vol. 693, pp.1397–1404, 2016.

[2] S. Guo, S. Chen, F. Liu, X. Ye, and H. Yang, "Binocular vision-based underwater ranging methods," 2017 IEEE International Conference on Mechatronics and Automation (ICMA), Takamatsu, 2017, pp.1058–1063.

[3] X. Yufeng, Z. Qiuyu, Y. Sai, and Z. Baozhu, "Human target detection and tracking under parallel binocular cameras," 2015 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA), Shenzhen, 2015, pp. 1–5.

[4] J. Hui, Y. Yang, Y. Hui, and L. Luo, "Research on identify matching of object and location algorithm based on binocular vision," Journal of Computational & Theoretical Nanoscience, vol. 13, pp. 2006–2013, 2016.

[5] Z. Chao, L. Wei, S. Hongwei, and L. Hong, "Three-dimensional surface reconstruction based on binocular vision," 2017 2nd International Conference on Robotics and Automation Engineering (ICRAE), Shanghai, 2017, pp. 389–393.

[6] C. Wang, M. Zhao, J. Yan, S. Zhou, and Y. Zhang, "Three-dimensional reconstruction of maize leaves based on binocular stereovision system," Transactions of the Chinese Society of Agricultural Engineering, vol. 26, pp. 198–202, 2010.

[7] J. Zhang, W. Pan, H. Shi, D. Zhang, and W. Li, "Three-dimensional reconstruction method based on target recognition for binocular humanoid robot," 2016 International Conference on Progress in Informatics and Computing (PIC), Shanghai, 2016, pp. 355–359.

[8] Z. Xiao, B. Dai, T. Wu, L. Xiao, and T. Chen, "Dense scene flow based coarse-to-fine rigid moving object detection for autonomous vehicle," IEEE Access, vol. 5, pp. 23492–23501, 2017.

[9] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," International Journal of Computer Vision, vol. 47, pp.7–42, 2002.

[10] S. Zhu, Z. Wang, X. Zhang, and Y. Li, "Edge-preserving guided filtering based cost aggregation for stereo matching," Journal of Visual Communication and Image Representation, vol. 39, pp. 107–119, 2016.

[11] M. Gerrits and P. Bekaert, "Local Stereo Matching with Segmentation-based Outlier Rejection," The 3rd Canadian Conference on Computer and Robot Vision (CRV'06), Quebec, Canada, 2006, pp. 66–69.

[12] A. Hosni, C. Rhemann, M. Bleyer, C. Rother, and M. Gelautz, "Fast cost-volume filtering for visual correspondence and beyond," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 2, pp. 504–511, 2013.

[13] K. Zhang, J. Lu, and G. 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, 2009.

[14] Kuk-Jin Yoon and In So Kweon, "Adaptive support-weight approach for correspondence search," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 28, no. 4, pp. 650–656, 2006.

[15] Jian Sun, Nan-Ning Zheng, and Heung-Yeung Shum, "Stereo matching using belief propagation," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 25, no. 7, pp. 787–800, 2003.

[16] A. Klaus, M. Sormann, and K. Karner, "Segment-Based Stereo Matching Using Belief Propagation and a Self-Adapting Dissimilarity Measure," 18th International Conference on Pattern Recognition (ICPR'06), Hong Kong, 2006, pp. 15–18.

[17] X. Wang, Y. Su, L. Tang, and J. Tan, "A combined back and foreground-based stereo matching algorithm using belief propagation and self-adapting dissimilarity measure," International Journal of Pattern Recognition and Artificial Intelligence, vol. 32, pp.1850019, 2018.

[18] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, no. 11, pp. 1222–1239, 2001.

[19] H. Wang, M. Wu, Y. Zhang, and L. Zhang, "Effective stereo matching using reliable points based graph cut," 2013 Visual Communications and Image Processing (VCIP), Kuching, 2013, pp. 1–6.

[20] T. Taniai, Y. Matsushita, and T. Naemura, "Graph cut based continuous stereo matching using locally shared labels," 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, 2014, pp. 1613–1620.

[21] M. Gong and Y. Yang, "Real-time stereo matching using orthogonal reliability-based dynamic programming," IEEE Transactions on Image Processing, vol. 16, no. 3, pp. 879–884, 2007.

[22] M. Bleyer, M. Gelautz, "Simple but effective tree structures for dynamic programming-based stereo matching," 2008 International Conference on Computer Vision Theory and Applications (VISAPP), Funchal, Portugal, 2008, pp. 415–422.

[23] B. Salehian, A. M. Fotouhi, and A. A. Raie, "Dynamic programming-based dense stereo matching improvement using an efficient search space reduction technique," Optik, vol. 160, pp. 1–12, 2018.

[24] Jae Chul Kim, Kyoung Mu Lee, Byoung Tae Choi, and Sang Uk Lee, "A dense stereo matching using two-pass dynamic programming with generalized ground control points," 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 2005, pp. 1075–1082

[25] X. Mei, X. Sun, M. Zhou, S. Jiao, H. Wang, and X. Zhang, "On building an accurate stereo matching system on graphics hardware," 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), Barcelona, 2011, pp. 467–474.

[26] Q. Yang, "Stereo matching using tree filtering," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 37, no. 4, pp. 834–846, 2015.

[27] X. Mei, X. Sun, W. Dong, H. Wang, and X. Zhang, "Segment-tree based cost aggregation for stereo matching," 2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, 2013, pp. 313–320.

[28] K. Zhang, Y. Fang, D. Min, L. Sun, S. Yang, and S. Yan, "Cross-scale cost aggregation for stereo matching," IEEE Transactions on Circuits and Systems for Video Technology, vol. 27, no. 5, pp. 965–976, 2017.

[29] P. Yao, H. Zhang, Y. Xue, M. Zhou, G. Xu, and Z. Gao, "Iterative color-depth MST cost aggregation for stereo matching," 2016 IEEE International Conference on Multimedia and Expo (ICME), Seattle, WA, 2016, pp. 1–6.

[30] S. Gao, H. Ge, and H. Zhang, "A nonlocal method with modified initial cost and multiple weight for stereo matching," Journal of Sensors, vol. 2017, pp. 1–16, 2017.

[31] X. Huang and Y. J. Zhang, "An O(1) disparity refinement method for stereo matching," Pattern Recognition, vol. 55, pp. 198–206, 2016.

[32] M. O"Byrne, V. Pakrashi, F. Schoefs, and B. Ghosh, "A stereo-matching technique for recovering 3d information from underwater inspection imagery," Computer-Aided Civil and Infrastructure Engineering, vol. 33, pp. 193–208, 2017.

[33] P. Dollár and C. L. Zitnick, "Fast Edge Detection Using Structured Forests," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 37, no. 8, pp. 1558–1570, 2015.

[34] H. Hirschmuller, "Stereo processing by semiglobal matching and mutual information," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 30, pp. 328–341, 2007.

[35] D. Scharstein, R. Szeliski, and H. Hirschmüller Middlebury stereo vision page [EB/OL]. http://vision.Middlebury.edu/stereo/.

[36] Q. Yang, "Local Smoothness Enforced Cost Volume Regularization for Fast Stereo Correspondence," IEEE Signal Processing Letters, vol. 22, no. 9, pp. 1429–1433, 2015.