IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received August 11, 2019, accepted September 9, 2019, date of publication September 13, 2019,
date of current version September 30, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2941215

Cost Aggregation for Stereo Matching Using Total Generalized Variation With Fusion Tensor

EU-TTEUM BAEK AND HYUNG JEONG YANG
School of Electronics and Computer Engineering, Chonnam National University, Gwangju 61186, South Korea

Corresponding author: Hyung Jeong Yang (hjyang@jnu.ac.kr)

This work was supported in part by the Basic Science Research Program through the National Research Foundation of Korea (NRF) through the Ministry of Education under Grant NRF-2017R1A4A1015559, and in part by the National Research Foundation of Korea (NRF) Grant through the Korea Government (MSIP) under Grant NRF-2017R1A2B4011409.

ABSTRACT Stereo matching methods have achieved remarkable improvements by exploiting various attempts. However, most stereo matching algorithms still suffer from problems such as ambiguous region and inherent ambiguities. In particular, some problems affecting cost aggregation step have the greatest impact on depth results. To resolve the above-mentioned problems, we propose a new cost aggregation method using the modified total generalized variation with fusion tensor. First, two kinds of diffusion tensors are extracted from the guidance color image and the guidance depth map. They are incorporated into an energy functional to obtain the total generalized variation. After formulating the final energy functional, it is optimized via a primal-dual energy minimization method. The performance of the proposed method is experimentally verified by qualitatively and quantitatively comparing the results to those of other algorithms.

INDEX TERMS Stereo matching, cost aggregation, modified total generalized variation.

I. INTRODUCTION

Depth estimation has traditionally been one of the most crucial tasks in the field of computer vision. It is highly fundamental for various computer vision-based applications including 3D object recognition [1], extraction of information from aerial surveys [2], geometry extraction for 3D object mapping [3], self-driving cars, and obstacle estimation [4]. In general, depth information can be acquired by several methods such as active depth cameras and passive depth cameras. Active depth sensor resolves depth information using a physical sensor. It emits light onto the scene and derives depth information based on the known speed of light, whereas passive depth cameras measure the correlation of images captured from two or more cameras. Active depth camera ensures more accurate depth information than passive depth camera, and it provides depth data much faster than passive depth cameras. However, it is difficult to use it outdoors during the daytime because of the presence of infrared ray noise. In addition, active depth camera provides only a low-resolution depth map due to hardware limitations. In contrast, passive depth camera estimates depth information indirectly from 2D images. These cameras can be used outdoors during daytime and can generate a high-resolution depth map. Therefore, passive camera-based methods have been studied continuously. In this paper, we focus on the passive camera-based method, i.e., the stereo matching method.

Stereo matching is inherently an ill-posed inverse problem as it reconstructs 3D information from the pair of 2D plains, and stereo matching method has various difficulties in whole or in each matching step [5]. An ill-posed problem is the one that does not meet the three Hadamard criteria for being well-posed. These criteria are: having a solution, having a unique solution, and having a solution that continuously depends on the parameters or input data. Conversely, the ill-posed problem may have several incomplete solutions and solutions that depend discontinuously on the parameters or input data. Therefore, it is exceedingly difficult to tackle the ill-posed problems. These problems are separated into two groups, namely ambiguous region and inherent ambiguities [6] in corresponding method. Ambiguous pixels are similar to other pixels near the point of interest in the reference image. Similarly, matching ambiguity occurs when their pixel similar to the target pixel in the target image are present along the scan line. The matching ambiguity problem also arises when matching intrinsically symmetrical shapes. The inherent ambiguity contains two special cases: ambiguous pixel and matching ambiguity. The inherent ambiguities are caused by the following reasons. When pixels are saturated in the acquired image, there is a high probability that there are

The associate editor coordinating the review of this manuscript and approving it for publication was Feng Shao.

nonadjacent pixels. When the brightness is zero at a pixel, it is likely to create adjacent pixels with non-zero brightness.

The depth data acquisition with the binocular cue suffers from occlusion problem, which is an is a key challenge in stereo matching. Occlusion means that an occluded pixel is apparent in the source image, but there is no corresponding pixel in the target image. Because an object is obscured by the view of some objects or regions, occluded pixels are only visible in the reference image, but in the target image. Therefore, occlusions are a principal challenge for the accurate computation of visual correspondence.

Generally, stereo matching methods perform the following steps: 1) matching cost computation; 2) cost aggregation; 3) winner takes all (WTA)/ disparity optimization; 4) disparity refinement. There are several problems and difficulties in the process of each step. In particular, the ambiguous region problem affecting cost aggregation step has the greatest impact on depth results. To tackle these difficulties, several approaches have been addressed.

Local methods aggregate each slice of cost volume within finite windows to make implicit smoothness assumptions. In contrast, global approaches formulate an energy function with explicit smoothness constraints and optimize it via global optimization techniques such as Expectation-Maximum (EM) [7], dynamic programming [8], belief propagation [9], graph cut [10] and semi-global matching (SGM) [11]. Recently, a method of combining deep learning with the global optimization method has been studied. One such typical method is SGM-NET [12], which trains penalties of SGM. In practical applications, local approaches are preferred to the global ones owing to the formers' speed.

Most cost aggregation methods define a support window for each pixel and sum/average matching cost over the windows. Yoon *et al.* first proposed to filter the cost volume with a joint bilateral filter, which is extremely effective for preserving edges [14]. However, the bilateral filter is computationally expensive owing to its large kernel size. To speed up the cost aggregation, He *et al.* proposed a guided image filter [15], which has linear runtime along with the number of image pixels. This filter shows leading speed and accuracy performance [13]. Yang *et al.* presented a tree filter cost aggregation method, which enlarges the window size to the whole image [16]. The tree filter-based cost aggregation can be performed exceedingly fast by making minimum spanning tree derived from a graph. Recently, Zheng *et al.* proposed a cross-scale cost aggregation, which estimates accurate disparity values in homogeneous regions [17]. This method constructs a hierarchical structure to aggregate matching costs. However, conventional methods do not deal with ambiguous areas. Therefore, they generate low quality in the depth discontinuities and highly textured regions because of ambiguous regions. In addition, the texture is copied from the color image to the depth map [23]. To resolve the problem, we present a new cost aggregation method by integrating fusion tensor and the total generalized variation (TGV) method [18], which is used to measure image characteristics up to a certain order of differentiation.

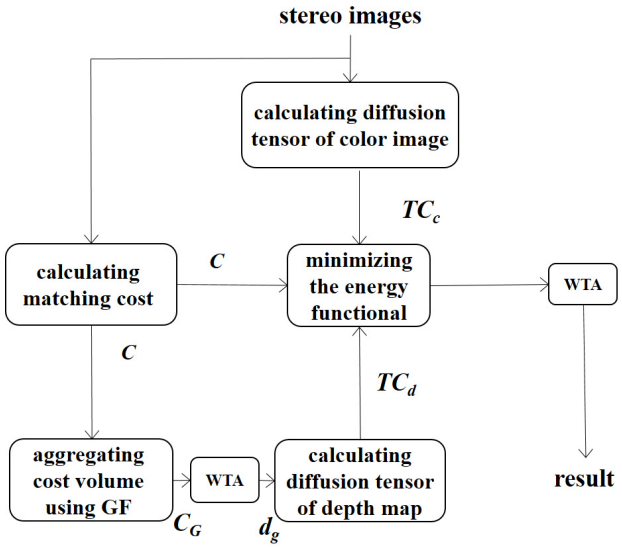


FIGURE 1. Procedure of the proposed method. C is the matching cost, C_G is an aggregated matching cost using GF, d_g is the guidance depth map, TC_c and TC_d are the diffusion tensor for the color image and the guidance depth map.

The rest of this paper is organized as follows: the cost matching computation is described in Section 2, the proposed cost aggregation is described in Section 3, and the experimental results regarding quantitative and qualitative criteria are presented in Section 4. Finally, this paper is concluded in Section 5.

II. MATCHING COST COMPUTATION

In this section, we explain the matching cost computation and guidance depth map generation. Figure 1 illustrates the overall procedure of the proposed method. First, we generate a guidance depth map using precomputed matching data and calculate the weighted sum of tensors, which contains the tensor of the guidance depth map and the color image. We incorporate the weighted sum of tensors into a total generalized variation method to formulate the proposed variation functional. After optimizing the variation functional, we apply a Winner-Takes-All strategy (WTA) to obtain a final depth map.

A. COST VOLUME GENERATION

In the matching cost computation step, the general stereo algorithm begins by calculating the matching cost at each position p for all disparities d under consideration. In other words, a 3D cost volume is generated by measuring matching costs for each pixel p at all possible disparity levels between the reference image and the target image. The commonly used method for computing the matching cost is the truncated absolute intensity differences and truncated absolute

difference of gradients in x-direction as

$$C(p,d)=\lambda\cdot\min(T_c, C_{AD}(d))+(1-\lambda)\cdot\min(T_g, C_{GD}(p,d)) \tag{1}$$

where $C(p, d)$ is a per-pixel matching cost of a pixel p for disparity value d. T_c and T_g are the truncation values, respectively. λ is a weight which is a constant value between 0 and 1. $C_{AD}(d)$ and $C_{GD}(p,d)$ are the cost value of absolute difference and the cost value of gradient difference in the x-direction, respectively. $C_{AD}(d)$ and $C_{GD}(p,d)$ are represented as

$$\begin{aligned} C_{AD}(d) &= |I_r(x,y) - I_t(x+d,y)| \\ C_{GD}(d) &= |G_r(x,y) - G_t(x+d,y)| \end{aligned} \tag{2}$$

where I_r and I_t are the reference and the target image, respectively. The absolute difference of gradients is computed as

$$G(x,y,d) = |\nabla_x(I_r(x,y)) - \nabla_x(I_t(x+d,y))| \tag{3}$$

where $\nabla_x(I(x,y))$ denotes the gradient in x-direction computed at pixel p.

B. GUIDANCE DEPTH MAP GENERATION

To obtain a fusion tensor, a guidance depth map should be constructed in advance. Therefore, we generate a guidance depth map by exploiting the guided image filter (GF) [13] and WTA strategy. Given a guidance color image I_r, the GF aggregates the cost volume. The GF is represented as

$$C_G(p,d) = W_p^G C(p,d) \tag{4}$$

where $C_G(p, d)$ is an aggregated cost volume, and W_p^G indicates the kernel weight of a guided image filter. The guided image filter depends on local optimization while performing the WLS (weighted least square) filter. The filter weights are defined as

$$W_{i,j}=\frac{1}{|w|^2}\sum_{k:(i,j)\in wk}(1+(I_i-\mu_k)(\sum_k+\varepsilon U)^{-1}(I_j-\mu_k)) \tag{5}$$

where $|w|$ is the total number of pixels in a window w_k centered at pixel k, and ε is a smoothness parameter. $\sum_k$ and μ_k are the covariance and the mean of pixel intensities within w_k. μ_k is 3×1 vectors, while $\sum_k$ and the unary matrix U are the size of 3×3. The Winner-Takes-All strategy (WTA) is applied for $C_G(p, d)$ to generate the guidance depth map I_g.

III. TGV-BASED COST AGGREGATION

The goal of cost aggregation is to eliminate artifacts in a cost volume to obtain high-quality depth map. In the matching cost computation step, several artifacts and erroneous cost values occur due to matching ambiguities such as repetitive texture regions, homogenous, or occluded areas. To address this problem, we propose a new cost aggregation method using modified TGV with fusion tensor, which aggregates the erroneous cost values while preserving the primary structure in the cost volume.

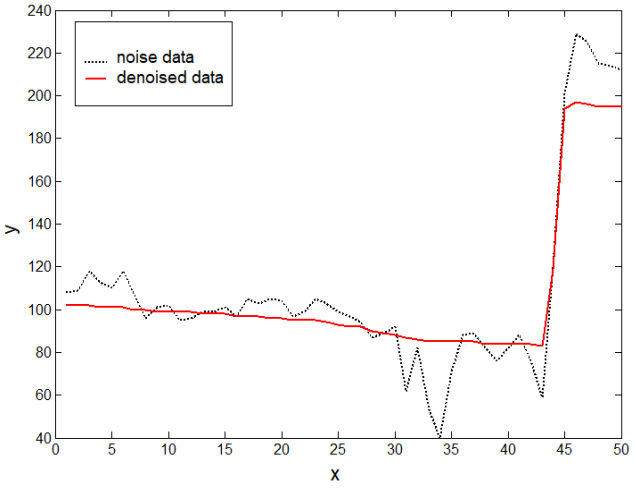


FIGURE 2. Application of 1D total variation denoising. Black dotted line is the original signal, red solid line is the denoised signal.

A. TOTAL VARIATION

The total variation-based energy functional directly deals with finding the optimal functions. Typically, the variational energy functional is composed of two terms, a data-driven energy term E_d, and a total variation regularizer TV in (6). The energy function incorporating the two terms is formulated by

$$\begin{aligned} E &= E_d(u,I) + \lambda\cdot E_s(u) \\ where\ E_d &= \sum_p (u-I)^2, \\ TV &= \sum_p \|\nabla u\|_2^2 \end{aligned} \tag{6}$$

where I is an original unobserved image, and u is a reconstructed image. The total variation (TV) regularizer is expressed as

$$TV = (\partial u/\partial x)^2 + (\partial u/\partial y)^2 \tag{7}$$

The total variation of a signal measures the amount by which a signal changes between signal values. In digital image denoising, the use of the total variation functional is common because the gradient strength can prevent the smoothness in the edge region. Given an input signal, the goal of total variation denoising is to find an approximation that has smaller total variation than the input signal but is "close" to the input signal. Figure 2 shows the graph of 1D total variation denoising where black dotted line is the original signal while red solid line is the denoised signal. However, it depicts the staircasing effects in case of smooth flows. To resolve this problem, we designed our regularization term as total generalized regularization (TGV) [18].

B. TOTAL GENERALIZED VARIATION

TV deals only with the first derivative, whereas TGV deals with a higher-order derivative. In other words, the total generalized variation (TGV) method is a functional that has the

ability to measure the image characteristics up to a certain order of differentiation [18]. Considering the time complexity of the algorithm, we utilized second-order derivatives of the guidance color image as a regulator term. The total generalized variation of the first and second-order can be represented as

$$TGV = \min_{v} \left\{ \alpha_1 \int_{\Omega} |w_d (\nabla u - v)| \, dx + \alpha_0 \int_{\Omega} |\nabla v| \, dx \right\} \tag{8}$$

where u denotes the result, v represents all the complex vector fields on Ω. This functional has weighting factors, α_0 and α_1, which balance the first- and second-order derivatives of the function. w_d indicates an anisotropic diffusion tensor, which is the weighted sum of the diffusion tensor.

Because TGV is the norm of Banach space, it is consistent with the mathematical theory of the convex optimization problem. Each function of the bounded variation results in a finite TGV value, thereby making the concept suitable for image processing. Additionally, TGV is translation invariant as well as rotationally invariant. Therefore, the images meet the requirement of being measured independently from the actual viewpoint.

C. FINAL ENERGY FUNCTIONAL

The proposed TGV-based energy model is composed of three terms. The first term is responsible for maintaining a similar solution at each cost level. The second and third terms are the first and second-order regularization terms, responsible for minimizing the first and second derivatives. Therefore, the energy functional can preserve the important structure while suppressing the texture or noise at the cost level.

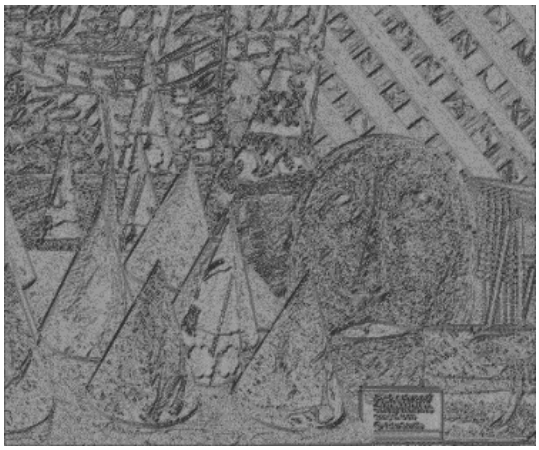
The conventional TGV model uses the diffusion tensor of the color to enhance the result. However, the direction of the diffusion tensor of any pixel is similar to that of the surrounding pixels in general. Moreover, there exists a phenomenon wherein the texture is copied a lot in the magnitude image for the color tensor, but it seems to acquire the information around the object precisely in the magnitude image for depth tensor, as illustrated in Fig. 3.

To overcome these limitations, this study employs a fusion tensor (or weighted sum of diffusion tensors), which combines the diffusion tensors for the color image and the guidance depth map. The weighted sum of the diffusion tensors is represented as

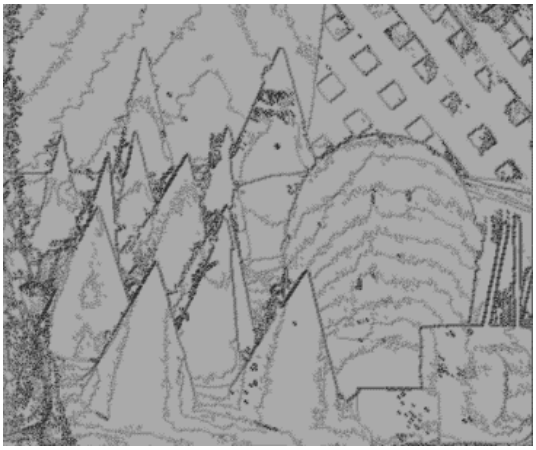
$$w_d = \alpha_T(z) TC_c + (1 - \alpha_T(z)) TC_d \tag{9}$$

Here, TC_c is the diffusion tensor for the color image and TC_d is the diffusion tensor for the guidance depth map, and $\alpha_T(z)$ is a weight function. z is calculated as follows:

$$\begin{aligned} z &= normalization\ of\ \ |\nabla I_r| \left|\nabla I_g\right| \\ where\ |\nabla I_r| &= \sqrt{\nabla_x I_r^2 + \nabla_y I_r^2} \\ \left|\nabla I_g\right| &= \sqrt{\nabla_x I_g^2 + \nabla_y I_g^2} \end{aligned} \tag{10}$$



(a)



(b)

FIGURE 3. Magnitudes of tensors. (a) is the diffusion tensor for the color image and (b) is the diffusion tensor for the guidance depth map.

where $|\nabla I_r|$ is the magnitude of the color gradient and $|\nabla I_g|$ is the magnitude of the guidance depth map. The normalization method is rescaling the range of features to scale the range of [0, 1]. z represents the amount of edge information contained in each pixel.

The weight function needs to preserve the edge region at higher weight values. A low weight of α significantly influences the diffusion tensor of the guidance depth map. The main aim is to determine when the cost aggregation method should depend on the color diffusion tensor and the depth diffusion tensor, respectively. To deal with the irrelevant textures and depth discontinuities, we formulated a measure to predict the color edges that are most likely to match the depth discontinuities. The weight function $\alpha_T(z)$ is represented as follows:

$$\alpha_T(z) = \frac{1}{1 + e^{-\varepsilon(z-\tau)}} \tag{11}$$

where ε controls the width of the transition area and τ determines a median value, as depicted in Fig. 4.

TABLE 1. Performance comparison.

Methods	Tsukuba		Venus		Cones		Teddy	
	Non occ	All	Non occ	All	Non occ	All	Non occ	All
Box	14.54	16.47	16.26	17.61	11.2	21.26	15.69	23.95
Bilateral [14]	6.08	7.11	2.03	2.82	7.14	15.47	9.37	16.75
Non-local [16]	5.46	6.54	2.58	3.38	7.19	16.4	8.21	15.6
Segmented [21]	6.31	7.53	3.18	4.07	8.22	17.56	9.43	16.9
Guided [15]	5.77	7.06	2.03	3.1	4.38	13.06	7.61	15.4
Proposed	3.23	4.08	1.02	1.9	5.32	12.88	7.01	13.5

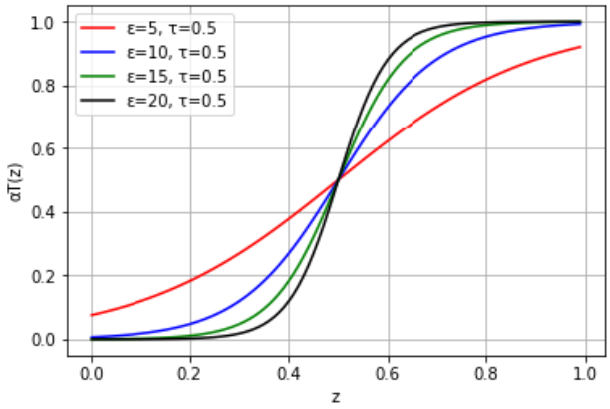


FIGURE 4. Graph of weight function.

TC_c and TC_d are represented as

$$TC_c = \exp(-\beta\,|\nabla I_r|^{\gamma})nn^T + n^{\perp}n^{\perp T}$$
$$TC_d = \exp(-\beta\,|\nabla I_g|^{\gamma})nn^T + n^{\perp}n^{\perp T} \quad (12)$$

where n is the normalized direction of the image gradient, and β and γ are scalar values, which adjust the magnitude and sharpness of the tensor, respectively. The weighted sum of the diffusion tensors is combined with the final energy functional, defined by

$$\min_{u,v}\left\{\int_{\Omega}|u - C(p,d)|\,dx + \alpha_1\int_{\Omega}|w_d\,(\nabla u - v)|\,dx + \alpha_0\int_{\Omega}|\nabla v|\,dx\right\} \quad (13)$$

where u denotes the aggregated result, v represents all the complex vector fields of the given image I_r on Ω, and $|\nabla v|$ represents the symmetrized derivative of u.

D. PRIMAL-DUAL OPTIMIZATION

This study utilizes the primal-dual energy minimization method to optimize the energy functional for each slice of a cost volume [19] because our optimization problem is convex but non-smooth. To apply the primal-dual energy minimization method, we first apply the Legendre-Fenchel transform to reformulate the convex and non-differentiable problem into a convex-concave saddle-point problem. The Legendre-Fenchel transform is a transform mathematical procedure that involves transforming convex and non-convex functions defined in a vectorial space, V, into convex functions defined in the dual vectorial space, V^*. The Legendre-Fenchel transform retransforms the original functional into a so-called primal-dual problem. The primal-dual energy functional involves the 2D vector field, p, and 4D vector field, q. q and p are the dual variables, which help in converting the two regularization terms into differentiable expressions. With the aid of p, the absolute value $|w|$ of a 2D vector w can be rewritten as

$$|\mathrm{w}| = \sup_{|\mathrm{p}|\le 1}\langle \mathrm{w},\mathrm{p}\rangle \quad (14)$$

where $<,>$ denotes the inner product. Utilizing (14), the total generalized variation can be rewritten as

$$\max_{|\mathrm{p}|\le 1,\,|\mathrm{q}|\le 1}\alpha_0 < w_d\,(\nabla u - v), \quad p > +\alpha_1 < \nabla v, q > \quad (15)$$

By substituting (15) into (9), the energy functional can be expressed as

$$\min_{u,v}\max_{|\mathrm{p}|\le 1,\,|\mathrm{q}|\le 1}\left\{\int_{\Omega}|u - C(p,d)|\,dx + \alpha_0 < w_d\,(\nabla u - v), p > +\alpha_1 < \nabla v, q >\right\} \quad (16)$$

We seek the minimum point of the energy in the $\boldsymbol{u}$, $\boldsymbol{v}$ directions as well as the maximum in the $\boldsymbol{p}$, $\boldsymbol{q}$ directions. The functional (16) is convex and differentiable in $\boldsymbol{u}$, $\boldsymbol{v}$, $\boldsymbol{p}$, and $\boldsymbol{q}$; it iteratively performs a gradient ascend in the $\boldsymbol{p}$, $\boldsymbol{q}$ directions, followed by a gradient descent in the $\boldsymbol{u}$, $\boldsymbol{v}$ directions, until convergence. The first optimization part of the iteration deals with the gradient ascent for $\boldsymbol{p}$, $\boldsymbol{q}$. The derivative of the final functional E_{final} on $\boldsymbol{p}$, $\boldsymbol{q}$ is $\partial E_{final}/\partial \boldsymbol{p}$ and $\partial E_{final}/\partial \boldsymbol{q}$. Therefore, we iteratively update $\boldsymbol{p}^{t+1} = \boldsymbol{p}^t + \lambda(\partial E_{final}/\partial \boldsymbol{p})$ and $\boldsymbol{q}^{t+1} = q^t + \lambda(\partial E_{final}/\partial \boldsymbol{q})$, with the learning size λ. However, $\boldsymbol{p}^{t+1}$ and $\boldsymbol{q}^{t+1}$ have to be back-projected onto the

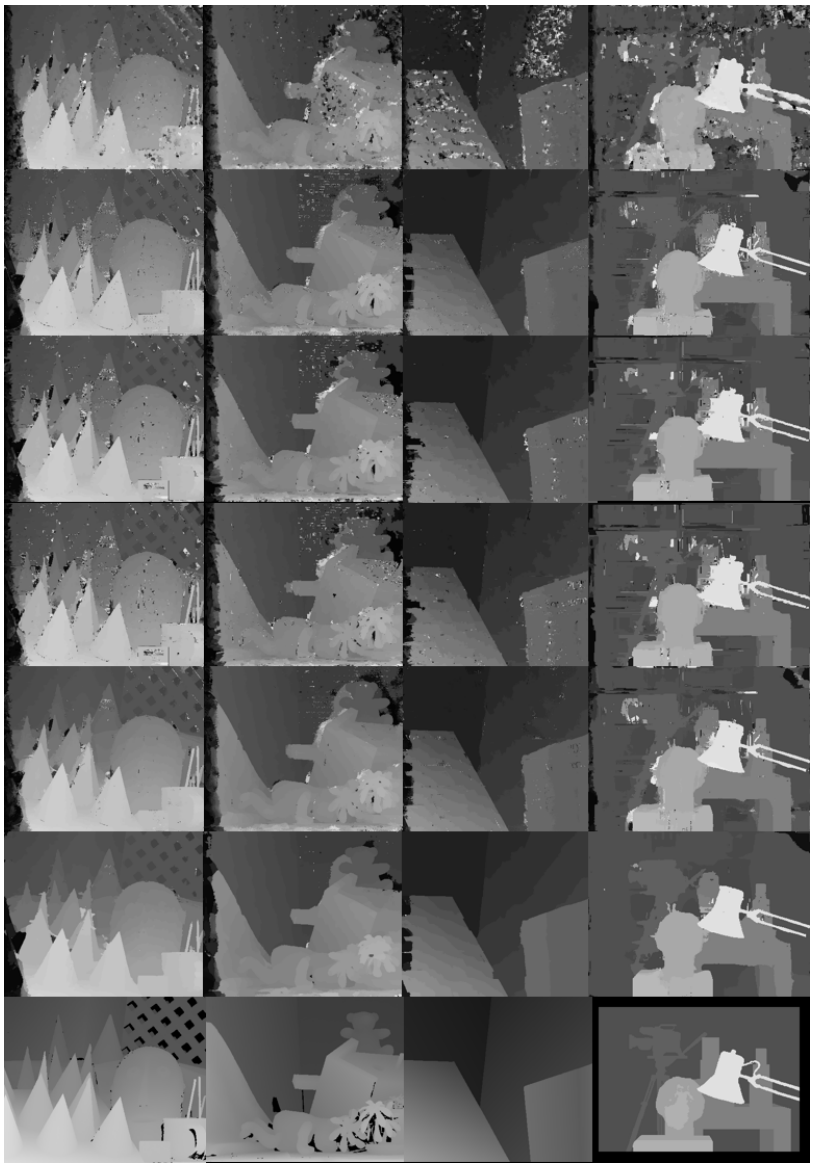


FIGURE 5. Experimental results on the Middlebury dataset. The first row images are the results of the box filter, second row images are the results of the bilateral filter, third row images are the results of the non-local aggregation, fourth row images are the results of the segmented tree aggregation, fifth row images are the results of the guided filter aggregation, sixth row images are the results of the proposed aggregation, and last row images are the ground truth.

unit circle to ensure that $|\boldsymbol{p}| \leq 1$ and $|\boldsymbol{q}| \leq 1$. Therefore, the final gradient ascent step can be defined as

$$\begin{aligned} \mathrm{p}^{\mathrm{t}+1} &= \frac{\mathrm{p}^{\mathrm{t}} + \lambda(\partial \mathrm{E}_{\mathrm{final}}/\partial \mathrm{p})}{\max(1, \mathrm{p}^{\mathrm{t}} + \lambda(\partial \mathrm{E}_{\mathrm{final}}/\partial \mathrm{p}))} \\ \mathrm{q}^{\mathrm{t}+1} &= \frac{\mathrm{q}^{\mathrm{t}} + \lambda(\partial \mathrm{E}_{\mathrm{final}}/\partial \mathrm{q})}{\max(1, \mathrm{q}^{\mathrm{t}} + \lambda(\partial \mathrm{E}_{\mathrm{final}}/\partial \mathrm{q}))} \end{aligned} \tag{17}$$

$\boldsymbol{u}$ and $\boldsymbol{v}$ can also be iteratively updated using the gradient descent method. After convergence for each slide, the winner-takes-all strategy (WTA) is exploited to generate the final result.

IV. EXPERIMENTAL RESULTS

In this study, we performed an exhaustive evaluation regarding the quantitative and qualitative comparison using the Middlebury dataset [20]. In the per-pixel cost computation step, the parameters were fixed as follows: $\lambda = 0.11$, $T_c = 0.02745$, and $T_g = 0.00784$. In the aggregation using GF step, a 19×19 local window was used, the smoothness parameter, ε, was set to 0.0001. In the optimization step, the parameters of the diffusion tensors β, γ were set to 9, 0.85 for all the scaling factors and images, weighting factors α_0 and α_1 were set to 5, 1, respectively, and τ was set to 0.0001.

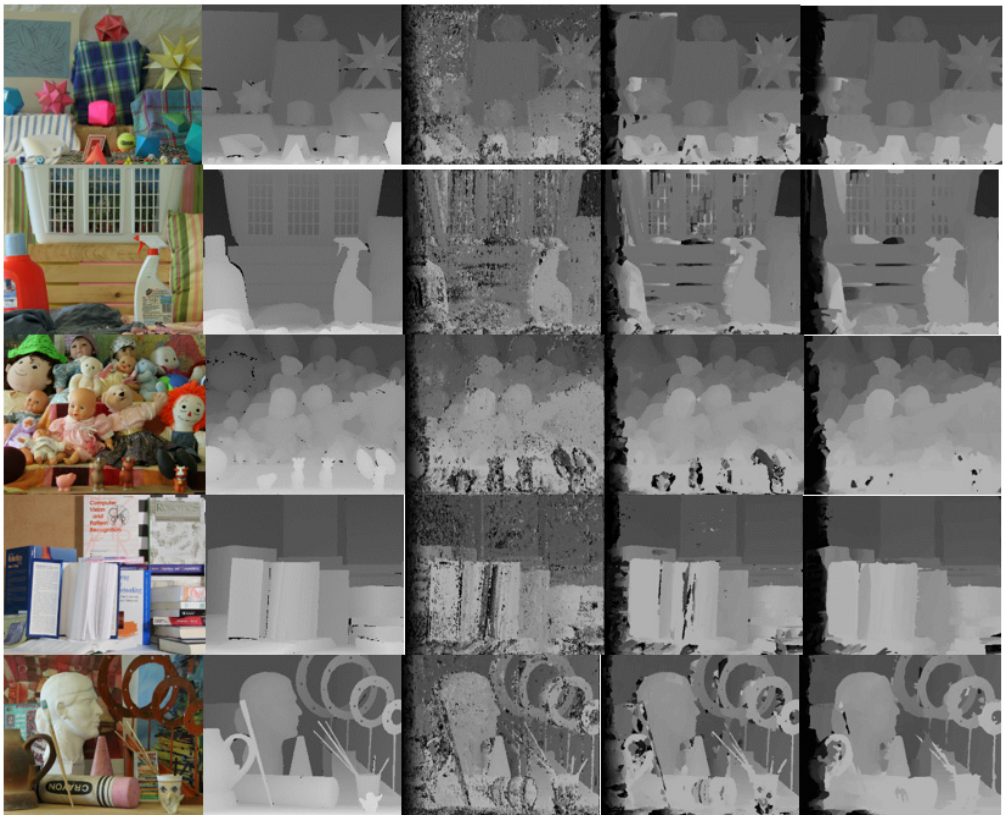


FIGURE 6. Results of the proposed stereo matching method. First column images are the color images, second column images are the ground truth images, third column images are the results of the bilateral cost aggregation, fourth column images are the results of the guided cost aggregation, and final column images are the results of the proposed method.

To evaluate the performance of the proposed method objectively, the percentage of mismatching pixels (BPR) was exploited, which can be defined as follows:

$$\begin{aligned} \mathrm{BPR}[\%] &= \left(\sum_{i=1}^{n} \delta(i)/n\right) \times 100 \\ \textit{where} \quad \delta(i) &= \begin{cases} 1, & \textit{if } |x_{gnd}(i) - x_{result}(i) > 1 \\ 0, & \textit{otherwise} \end{cases} \end{aligned} \tag{18}$$

where $x_{gnd}(i)$ and $x_{result}(i)$ are the ith pixels in the ground truth and the result, respectively. BPR is the total number of pixels in a depth map. Furthermore, no disparity refinement technique was employed to compare the different cost aggregation methods fairly.

Fig. 5 depicts the experimental results. The box filter results have many artifacts in the texture and homogeneous regions. Moreover, errors may occur in the depth discontinuous regions. The bilateral filter generates accurate depth information in the homogeneous and discontinuity regions; however, there are many artifacts in the repetitive texture regions. The segmented tree aggregation and non-local aggregation methods reduce the number of artifacts; however, the results contain noise in the homogeneous regions. Nevertheless, the proposed method performs better in the texture and homogeneous regions than the conventional algorithms.

Table 1 presents the percentage of the bad matching pixels for the proposed method and conventional aggregation methods, such as the bilateral filtering [14], guided image filtering approach [15], non-local approach [16], and segmented tree aggregation [21]. The proposed method outperforms the conventional cost aggregation methods with respect to bad pixel rate. For comparison purposes, additional experiments of stereo matching were conducted using the Middlebury dataset. To qualitatively verify the performance of the proposed method, we conducted the experiments on other stereo images. Fig. 6 depicts the results of the proposed stereo matching method. The proposed method generates more accurate depth maps in the texture regions than the conventional methods.

Next, we examined the performance comparison of the fusion tensor and basic tensor at the optimization stage. To demonstrate that the fusion tensor is more accurate than the conventional methods, we calculated the t-value. The t-value is a test statistic, which is a result of a statistical test to measure how far apart the two means are. The t-test formula can be defined as

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{19}$$

TABLE 2. T-values.

	Tsukuba	Venus	Cones	Teddy
TC_C	602.5432	605.7542	540.4139	518.5139
w_d	591.3595	392.2234	258.8972	458.3721

FIGURE 7. Enlarged stereo matching results.

where $\bar{x}_1$ and $\bar{x}_2$ are the sample means, s_1^2 and s_2^2 are the standard deviations, and n_1 and n_2 are the sample sizes. Table 2 lists the t-values between the tensor of ground truth and the tensor of color, TC_C, and between the tensor of ground truth and the fusion tensor, w_d. The proposed fusion tensor has a lower t-value for Tsukuba, Venus, Cones, and Teddy. A low t-value indicates that the two distributions are closer. As expected, using the proposed method was more advantageous for obtaining good results.

Fig. 7 illustrates some parts of the stereo matching results of Tsukuba, Venus, Books, and Cones. The conventional methods are not concerned with the ambiguous areas [13], [15], [16]. Therefore, the conventional cost aggregation methods cause the texture copying problem. However, enlarged depth maps demonstrate that the proposed method can solve the texture copying problem of the conventional cost aggregation methods.

V. CONCLUSION

In this study, a new cost aggregation method for the depth estimation method was proposed. The proposed method aggregated the slice of the cost volume by optimizing the energy functional. Because the direction of the diffusion tensor of any pixel is similar to that of the surrounding pixels, this study employed the fusion tensor to increase the correlations between the neighboring pixels and to reduce the texture copying from the color image. The experimental results verified that the combination of the two different techniques, TGV and image-guided cost volume filtering, can be an effective solution for acquiring accurate disparity maps. Moreover, the proposed method produces more accurate disparity maps compared to the conventional aggregation methods with respect to the bad pixel rate.

REFERENCES

[1] P. Wohlhart and V. Lepetit, "Learning descriptors for object recognition and 3D pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3109–3118.
[2] N. Haala, M. Rothermel, and S. Cavegn, "Extracting 3D urban models from oblique aerial images," in *Proc. Joint Urban Remote Sens. Event (JURSE)*, Mar./Apr. 2015, pp. 1–4.
[3] J. Rock, T. Gupta, J. Thorsen, J. Gwak, D. Shin, and D. Hoiem, "Completing 3D object shape from one depth image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 2484–2493.
[4] J. Ziegler *et al.*, "Making bertha drive—An autonomous journey on a historic route," *IEEE Intell. Transp. Syst. Mag.*, vol. 6, no. 2, pp. 8–20, Apr. 2014.
[5] S. I. Kabanikhin, "Definitions and examples of inverse and ill-posed problems," *J. Inverse Ill-Posed Problems*, vol. 16, no. 4, pp. 317–357, 2008.
[6] R. Manduchi and C. Tomasi, "Ambiguity in stereo matching," Dept. Comput. Sci., Stanford Univ., Stanford, CA, USA, Tech. Rep. 1, 1997.
[7] C. Strecha, R. Fransens, L. Van Gool, "Combined depth and outlier estimation in multi-view stereo," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2006, pp. 2394–2401.
[8] S. Birchfield and C. Tomasi, "Depth discontinuities by pixel-to-pixel stereo," *Int. J. Comput. Vis.*, vol. 35, no. 3, pp. 269–293, 1999.
[9] V. Kolmogorov and Z. Ramin, "Computing visual correspondence with occlusions via graph cuts," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2001, pp. 508–515.
[10] J. Yedidia, W. T. Freeman, and Y. Weiss, "Understanding belief propagation and its generalizations," *Exploring Artif. Intell. New Millennium*, vol. 8, pp. 239–269, Jan. 2003.
[11] H. Hirschmuller, "Stereo vision in structured environments by consistent semi-global matching," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2, Jun. 2006, pp. 2386–2393.
[12] A. Seki and M. Pollefeys, "SGM-Nets: Semi-global matching with neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 231–240.
[13] K. He, J. Sun, and X. Tang, "Guided image filtering," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 1–14.
[14] K.-J. Yoon and I. S. Kweon, "Adaptive support-weight approach for correspondence search," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 4, pp. 650–656, Apr. 2006.
[15] A. Hosni, M. Bleyer, C. Rhemann, M. Gelautz, and C. Rother, "Real-time local stereo matching using guided image filtering," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2011, pp. 1–6.
[16] Q. Yang, "A non-local cost aggregation method for stereo matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 1402–1409.
[17] K. Zheng, Y. Fang, D. Min, L. Sun, S. Yang, S. Yan, and Q. Tian, "Cross-scale cost aggregation for stereo matching," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1590–1597.
[18] K. Bredies, K. Kunisch, and T. Pock, "Total generalized variation," *SIAM J. Imag. Sci.*, vol. 3, no. 3, pp. 492–526, 2010.
[19] A. Chambolle and T. Pock, "A first-order primal-dual algorithm for convex problems with applications to imaging," *J. Math. Imag. Vis.*, vol. 40, no. 1, pp. 120–145, 2011.
[20] D. Scharstein, R. Szeliski. *Middlebury Data Sets*. Accessed: 2014. [Online]. Available: http://vision.middlebury.edu/stereo
[21] X. Mei, X. Sun, W. Dong, H. Wang, and X. Zhang, "Segment-tree based cost aggregation for stereo matching," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 313–320.

[22] D. Ferstl, C. Reinbacher, R. Ranftl, M. Ruether, and H. Bischof, "Image guided depth upsampling using anisotropic total generalized variation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 993–1000.
[23] K.-H. Lo, K.-L. Hua, and Y.-C. F. Wang, "Depth map super-resolution via Markov random fields without texture-copying artifacts," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 1414–1418.

EU-TTEUM BAEK received the B.S. degree in computer science and engineering from Chonbuk National University, South Korea, in 2012, and the M.S. degree in information and communication engineering and the Ph.D. degree in electrical engineering and computer science from the Gwangju Institute of Science and Technology (GIST), South Korea, in 2015 and 2019, respectively. He is currently a Researcher with Chonnam National University, Gwangju, South Korea. His research interests include 3D digital image processing, depth estimation, deep learning, and medical imaging.

HYUNG JEONG YANG received the B.S., M.S., and Ph.D. degrees from Chonbuk National University, South Korea. She is currently an Associate Professor with the Department of Electronics and Computer Engineering, Chonnam National University, Gwangju, South Korea. Her main research interests include multimedia data mining, pattern recognition, artificial intelligence, e-learning, and e-design.