Cost Aggregation with Guided Image Filter and Superpixel for Stereo Matching

Eu-Tteum Baek and Yo-Sung Ho
Gwangju Institute of Science and Technology (GIST)
123 Cheomdangwagi-ro Buk-gu, Gwangju 61005, Republic of Korea
E-mail: {eutteum, hoyo}@gist.ac.kr

***Abstract*— Cost aggregation is one of the popular method for stereo matching due to efficiency and effectiveness. Their limitation is a high complexity and some error near the contour, which makes them not to implement in real time. Furthermore, the weakness makes them unattractive for many applications which require the accurate depth information. In this paper, we present a cost aggregation method using the superpixel-based edge-preserving filter and the guided image filter for stereo matching. First, we combine cost using a census transform and truncated absolute difference of gradients. The guided filter and the super pixel based smooth filter are exploited for the cost aggregation in order. In order to refine depth information, we apply occlusion handling and median filter. Consequently, the proposed method increases the accuracy of the depth map, and experimental results show that the proposed method generates more robust depth maps compared to the conventional methods.**

I. INTRODUCTION

Depth estimation has been one of the most important tasks in the computer vision field, and it is highly fundamental in applications such as 3D object recognition, extraction of information from aerial surveys, geometry extraction for 3D building mapping, and obstacle estimation. The depth map can be obtained by some methods such as passive sensor method [1], active depth sensor method [2], and hybrid depth sensor method. Passive depth cameras estimate a pair of corresponding points between two consecutive images taken from different viewpoints whereas Active depth sensor acquires depth map with a physical light sensor. Hybrid depth cameras combine a passive method and an active method to obtain more accurate depth data and to cover their weaknesses [3].

Depth estimation can be classified into global and local approaches according to the strategies exploited for estimation. Global approaches generally formulate an energy function with various constraints and optimize it via global optimization techniques such as dynamic programming [4], belief propagation [5], graph cut [6], and semi-global matching [7]. Local approaches obtain a disparity map by measuring correlation in local windows. The cost is aggregated over the window. Common local cost functions include the sum of absolute differences (SAD), the sum of squared differences (SSD), normalized cross correlation (NCC), and the census transform [8]. Local methods are much faster and more suitable for a practical implementation than global methods.

In recent years, many aggregation methods for stereo matching have been proposed. Yoon et al. first proposed to filter the cost volume with a joint bilateral filter [9]. The idea is that pixels having a color similar to the center pixel are likely to lie on the same object, therefore have similar depth, and effectively preserves depth boundaries. However, implementation of the bilateral filter is very slow. Therefore, many approximation methods have been proposed to accelerate the Yoon et al.'s method. He et al. presented a guided image filter [10], which has linear runtime in the number of image pixels. This filter shows leading speed and accuracy performance. Although the filter is very fast and generates relatively accuracy results, some error occurs.

In this paper, we propose a new cost-volume filtering method which is a two-step aggregation method. First, we employ guided image filter and apply superpixel-based filter to the cost volume. And post-processing method is also used to handle occlusions and textureless regions.

II. COST AGGREGATION IN STEREO

The goal of the proposed method is to improve the conventional stereo disparity estimation methods using the new cost aggregation method. Fig. 1 shows the overall framework of the proposed algorithm. our algorithm performs the following four steps: 1) constructing cost volume; 2) cost aggregation; 3) disparity selection; 4) disparity refinement.

A. Cost Volume

A 3D cost volume is generated by measuring matching costs for each pixel p at all possible disparity levels between the left image and the right image. In our implementation, we choose the census transform and truncated absolute difference of gradients. The census transform is represented as

$$D_s(d_s) = Hamming\ \ (I_c(p), \bar{I}_c(\bar{p}_d)) \quad (1)$$

where $I_c(p)$ and $\bar{I}_c(\bar{p}_d)$ are transformed vectors using census transform, which is a non-parametric local transform method. Hamming distance is the number of differences between two vectors as shown in Fig. 3(a). Let $I_c(p)$ denotes census transform of one point p. The center pixel's intensity value is

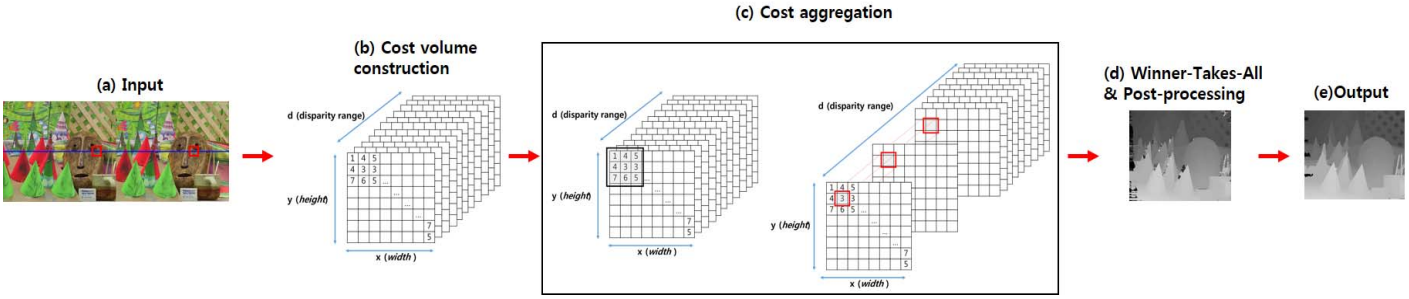


Fig. 1 Overall framework of our method

replaced by the bit string composed of the set of boolean comparisons such that in a square window and $I_c(p)$ is defined as

$$I_c(p) = \bigotimes_{q \in N_p} \xi(I(p), I(q)) \tag{3}$$

where $\otimes$ denotes concatenation, N_p is neighboring pixels in a window, and ξ denotes transform represented as

$$\xi(I(p), I(q)) = \begin{cases} 0, if I(p) < I(q) \\ 1, otherwise \end{cases} \tag{4}$$

Census transform converts the relative intensity difference to 0 or 1 in 1-dimensional vector form. Figure 3(b) represents an example of the census transform of a window with respect to the center pixel. The absolute difference of gradients is expressed as

$$G(p,d) = \left| \nabla_x(I_L(p)) - \nabla_x(I_R(p-d)) \right| \tag{5}$$

where $\nabla_x(I(p))$ denotes the gradient in x direction computed at pixel p. Final cost function is represented as

$$C(p,d) = \lambda \cdot \min(T_c, D(p,d)) + (1-\lambda) \cdot \min(T_g, G(p,d)) \tag{6}$$

where λ balances the census and gradient terms and T_c, T_g are the census and gradient truncation values.

B. *Cost Aggregation*

After constructing the cost volume, we exploiting the guided image filter [10] and superpixel [12] based smooth filter to filter each slice of the cost volume in order. Using a guidance image I, the guided image filter can be used to compute the aggregated cost as follows

$$C^g(p,d) = \sum_q W_{p,q} C(p,d) \tag{7}$$

where $C^g(p,d)$ denotes the aggregated cost using guided image filter, and $W_{p,q}$ is a filter weight. The filter weights are defined as

$$W_{i,j} = \frac{1}{|w|^2} \sum_{k:(i,j) \in w_k} (1 + (I_i - \mu_k)(\Sigma_k + \varepsilon U)^{-1}(I_j - \mu_k)) \tag{8}$$

where $|w|$ is the total number of pixels in a window w_k centered at pixel k, and ε is a smoothness parameter. Σ_k and μ_k are the covariance and mean of pixel intensities within w_k. I_s, I_t and μ_k are 3 × 1 vectors, while Σ_k and the unary matrix U are of size 3 × 3.

Fig. 2. Superpixels (k=400).

After using guided image filter to filter each cost slice, we exploit superpixel-based edge-preserving filter to aggregate cost again. Because superpixel ensures the boundary accuracy and the convergence to the real target boundary and low-complexity, we choose this algorithm. To obtain superpixel, we employ SLIC superpixel method which is a simple and efficient to decompose an image in visually homogeneous regions [12]. Given the k cluster points which the intervals of the image plane are fixed. The size of the interval is $S = \sqrt{N/k}$ so as to generate superpixel. The cluster center points move to the points where the texture maps have the lowest gradient. In limited regions, we calculate pixels similarity. In order to speed up the process of superpixel while almost maintaining accuracy, we choose the optimal search range. Typically, the size is used in a region 2S *2S around the cluster center points. Superpixel procedure exploits the CIELAB color space and the position of the center pixels to construct vectors $L_i=[\, l_i, a_i, b_i, x_i, y_i]^T$. For each cluster center L_i, computer the distance D between L_i for each pixel i. D is the distance measure, which pixel around

the cluster center L_i belong to the cluster. The distance measure D is represented as

$$d_c = \sqrt{(l_j - l_i)^2 + (a_j - a_i)^2 + (b_j - b_i)^2}$$
$$d_s = \sqrt{(x_j - x_i)^2 + (y_j - y_i)^2} \quad (9)$$
$$D = \sqrt{d_c^2 + (\frac{d_s}{S})^2 m^2}$$

Figure 2 shows the result of superpixel. The number of superpixels is 400.

(a) (b)

Fig. 3. Superpixel-based smooth filter. Green pixel is center pixel p, yellow pixels have the same label with center pixel q, and blue pixels have the different label with center pixel q.

In the same label, we aggregate cost volume using Gaussian weights. As shown in Fig. 3, superpixel-based smooth filter can be used to compute the aggregated cost as follows

$$C^s(p,d) = \sum_q Z_{p,q} C^g(p,d) \quad (10)$$

where $C^s(p,d)$ is the aggregated cost using the superpixel-based smooth filter, and $Z_{p,q}$ is a filter weight. The filter weights are defined as

$$Z(x,y) = \frac{1}{W_Z} \sum_{k:(x,y)\in w_k} L(x,y) \cdot g(x,y)$$
$$where\ \ L(x,y) = \begin{cases} 1 & if \in S(p) = S(q) \\ 0 & otherwise \end{cases} \quad (11)$$

where W_Z is normalization factor, and $L(x,y)$ is an indicator function that is one if the center pixel and the neighbor pixel are in the same label and zero otherwise , and $g(x,y)$ is a Gaussian function.

$$g(x,y) = \frac{1}{2\pi\sigma^2} exp^{-\frac{x^2+y^2}{2\sigma^2}} \quad (12)$$

Before the Winner-Takes-All strategy (WTA) is applied, we apply cross-level cost aggregation inspired by [13]. We aggregate cost between levels (or slices) using a weighted mean filter. The weighted mean filter can be used to compute the aggregated cost as follows

$$C^d(p,d) = \frac{1}{4}(C^S(p,d-1) + 2 \cdot C^S(p,d) + C^S(p,d+1)) \quad (13)$$

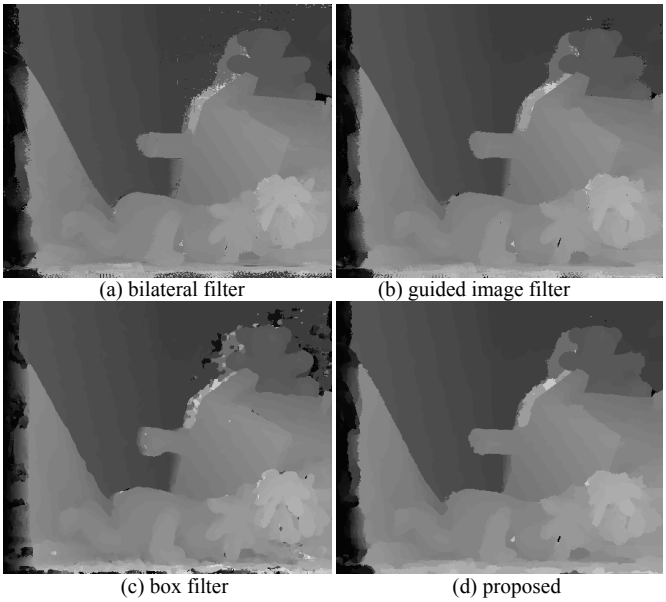
where $C^d(p,d)$ is the aggregated cost the weighted mean filter.

C. *Post processing*

Occlusion is an important and challenging problem in stereo depth estimation. The simplest method for occluded pixel detection and disparity estimation uses cross-checking. Given the estimated disparity, we apply the cross-checking test to detect occlusion, and fill the occlusion. The occlusion handling equation is defined as

$$O(s,d) = \arg\min_t \frac{1}{dist(s,t)} \exp(-\frac{dif_{s,t}}{\sigma^2})$$
$$dif_{s,t} = \sum_{c\in\{R,G,B\}} |I_c(s) - I_c(t)| \quad (14)$$

III. EXPERIMENTAL RESULTS



(a) bilateral filter (b) guided image filter

(c) box filter (d) proposed

Fig. 4. Visual comparison with conventional cost aggregation methods and proposed cost aggregation method.

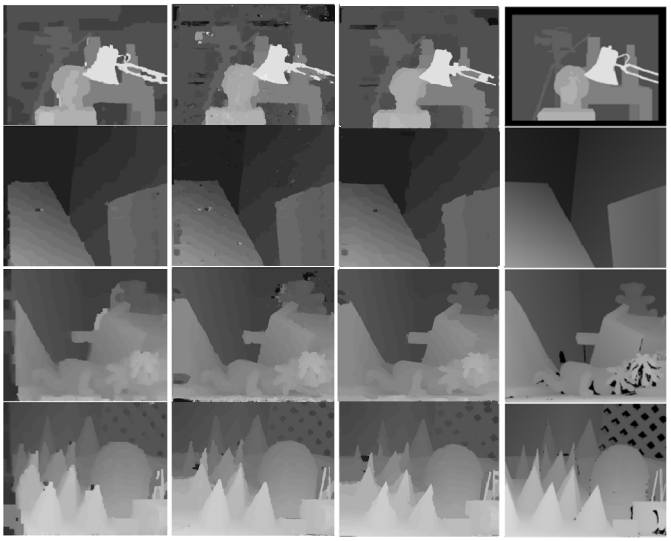
The data sets computed using conventional filter-based cost aggregation methods and the proposed cost aggregation method are presented in Fig. 4, which shows that the proposed cost aggregation method provided more accurate result and error is well removed.

In order to evaluate objectively the performance of our method, we exploit the percentages of mismatching pixels (BPR) with known ground truth disparity. Table I shows the percentage of the bad matching pixels between the results of the proposed method and ground truths. This measure is computed for three subsets of the image: non-occluded, whole, and discontinuity regions, denoted as “nonocc”, “all”, and “disc”, respectively. The results exhibit robust performance compared to conventional global methods.

Figure 5 represents a visual comparison with conventional methods and proposed method. The results generated by our proposed method exhibit fewer artifacts than the disparity maps generated by the conventional method.

TABLE I
PERFORMANCE COMPARISON

Algorithm		CSBP	GC + occ	CCH + SegAggr	AdaptAggrDP	Proposed method
Tsukuba	nonocc	2.00	1.19	1.74	1.57	**4.57**
	all	4.17	2.01	2.11	3.50	**5.4**
	disc	10.50	6.24	9.23	8.27	**16.6**
Venus	nonocc	1.48	1.64	0.41	1.53	**1.19**
	all	3.11	2.19	0.94	2.69	**1.93**
	disc	17.70	6.75	3.97	12.4	**11.62**
Teddy	nonocc	11.10	11.20	8.08	6.79	**6.27**
	all	20.20	17.40	14.3	14.3	**11.44**
	disc	27.50	19.80	19.80	16.2	**18.81**
Cones	nonocc	5.98	5.36	7.07	5.53	**6.15**
	all	16.50	12.40	12.90	13.2	**13.77**
	disc	16.00	13.00	16.30	14.8	**17.3**



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

Fig. 5. Visual comparison with conventional methods and proposed method. (a) CSBP, (b) box aggregation + post processing, (c) proposed method, (d) ground truth.

IV. CONCLUSIONS

In this paper, we proposed the cost aggregation for the stereo matching method. The proposed method exploited the guided image filter and the segmentation based edge-preserving filter to aggregate cost volume. After filtering cost volume between slices, we applied occlusion handling to enhance the accuracy of the disparity. According to the experimental results, we have confirmed that the proposed method produces more accurate disparity maps compared to other methods in terms of bad pixel rates.

ACKNOWLEDGMENT

This work was supported by the 'Civil-Military Technology Cooperation Program' grant funded by the Korea government.

REFERENCES

[1] W.S. Jang, Y.S. Ho, "Efficient disparity map estimation using occlusion handling for various 3D multimedia applications," Trans. Consumer Electronics, vol. 57, no. 4, pp. 1937–1943, 2011.

[2] A. Frick , F. Kellner , B. Bartczak and R. Koch, "Generation of 3DTV LDV-content with time of flight camera," Int'l Conf. on 3DTV, pp. 45–48, 2009

[3] E.K. Lee, Y.S. Ho, "Generation of high-quality depth maps using hybrid camera system for 3-D video," J. Visual Comm. Image Represent, vol. 22 no. 1, pp. 73–84, 2011.

[4] S. Birchfield and C. Tomasi. Depth discontinuities by pixel-to-pixel stereo. J. Computer Vision, vol. 35 no. 3 pp. 269–293, 1999.

[5] V. Kolmogorov and R. Zabih, "Computing visual correspondence with occlusions using graph cuts," Conf. Computer Vision, pp. 508–515, 2001.

[6] J. S. Yedidia, W. T. Freeman, and Y. Weiss, "Understanding belief propagation and its generalizations," Exploring Artificial Intelligence in the New Millenium, pp. 239–269, 2003.

[7] H. Hirschmueller, "Stereo vision in structured environments by consistent semi-global matching," Proc. Computer Vision and Pattern Recognition, vol. 2, pp. 2386–2393, 2006.

[8] R. Zabih and J. Woodfill. "Non-parametric local transforms for computing visual correspondence," Conf. Computer Vision, pp. 151–158, 1994.

[9] K. J. Yoon and I. S. Kweon. "Adaptive support-weight approach for correspondence search," PAMI, vol. 28, no. 4, pp. 650–656, 2006.

[10] K. He, J. Sun, and X. Tang. Guided image filtering. ECCV, pp. 1–14, 2010.

[11] E. Baek and Y. Ho, "Temporal stereo disparity estimation with graph cuts," APSIPA, pp. 184-187, 2015.

[12] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," Trans. Pattern Analysis and Machine Intelligence, pp. 2274-2282, 2012.

[13] K. Zhang, Y. Fang, D. Min, L. Sun, S. Yang, S. Yan, and Q. Tian, "Cross-scale cost aggregation for stereo matching," Conf. Computer Vision and Pattern Recognition, pp. 1590-1597, 2014.