Stereo Matching Using Relative Total Variation and Entropy

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*— Depth estimation methods suffer from problems such as homogeneous region, depth discontinuous region, repetitive texture, and occlusion. In this paper, we propose a depth estimation method considering depth discontinuous, textureless region, and occlusion. First, we formulate MRF-MAP based energy function. The data term in the proposed energy function combines an intensity similarity term, a gradient differences term, and a relative total variation (RTV) term. A weight function on information of color entropy and gradients is defined. After optimizing the energy function via alpha expansion, we refine occlusion region and errors exploiting the initial depth map. To evaluate the performance of our proposed method, we measure the percentages of mismatching pixels (BPR). Consequently, the proposed method increases the accuracy of depth estimation, and experimental results show that the proposed method generates more robust depth maps compared to the conventional methods.

I. Introduction

Stereo matching is the one of the most extensively researched topics in the computer vision field, and has a long history, because stereo vision is highly important to many 3D applications such as 3D reconstruction, 3D printing, object detection, and 3D movie. In general, depth information can be acquired by several methods such as active depth cameras, passive depth cameras, and hybrid depth cameras. Active depth sensor resolves depth information with a physical sensor. It emits their light onto the scene, and derive its depth information based on the known speed of light, whereas passive depth cameras measure the correlation of images captured from two or more cameras. Hybrid depth cameras associate two methods to generate more accurate depth data and to cover their weaknesses. Active depth cameras and hybrid depth cameras ensure more accurate depth information than passive depth cameras, provide depth data much faster than passive depth cameras. However, Active and hybrid depth cameras can be used in indoor environments, and these provide only low-resolution images due to hardware limitations.

Stereo estimation methods can be separated into two broad classes: local and global methods. Given two images, local methods calculate the disparity with intensity values in a window. The disparity with the minimum aggregated cost is selected after aggregating the cost over the window. Common local costs function includes the sum of absolute differences (SAD), the sum of squared differences (SSD), normalized cross correlation (NCC), and the census transform. Global methods consider depth estimation as a labeling problem. The problem gets resolved by global optimization techniques such as such as dynamic programming, graph cuts, belief propagation, and semi-global matching for avoiding unexpected local minima.

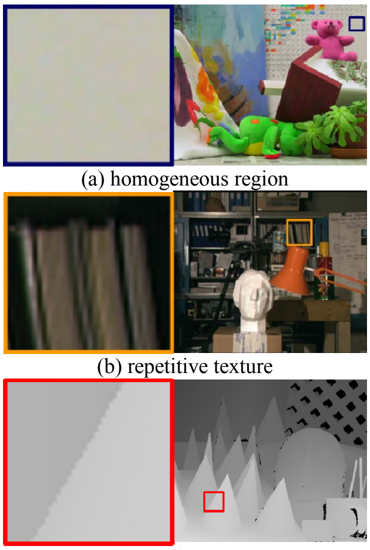

(a) homogeneous region
(b) repetitive texture
(c) depth discontinuous

Figure 1. Problems of stereo matching

Wei et al. introduced a region-based stereo matching for refining inherently ambiguous in homogeneous and occluded areas [2]. The Initial regions significantly affect results in the method. If very limited initial reliable regions are found, the final results are not visually satisfactory. Singh et al. proposed a depth estimation from stereo images based on adaptive weight [3]. To handle depth discontinuous and homogeneous region, they use segmentation method to give higher weight to pixels which lie on the same segment. Liu et al. presented an occlusion handling method using a two-step local method [11]. To compute an initial matching cost, they use contrast contest histogram descriptors. Then, disparity estimation is performed via two-pass weighted cost aggregation considering segmentation based adaptive support weights.

978-1-5386-1542-3@2017 APSIPA

II. Stereo Matching

The goal of our work is to enhance stereo matching method considering homogeneous region, depth discontinuous, and repetitive texture as shown in Fig. 1. First, we employ relative total variation to remove repetitive texture and obtain meaningful edges. To enhance depth discontinuous regions, we calculate entropy and gradients in a window. After refining occlusion region, we obtain a final result. The overall framework of the proposed algorithm. is illustrated in Fig. 2.

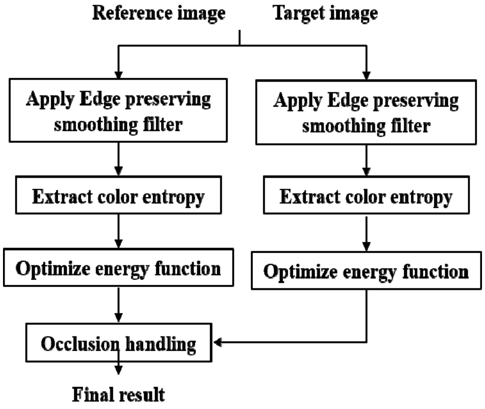
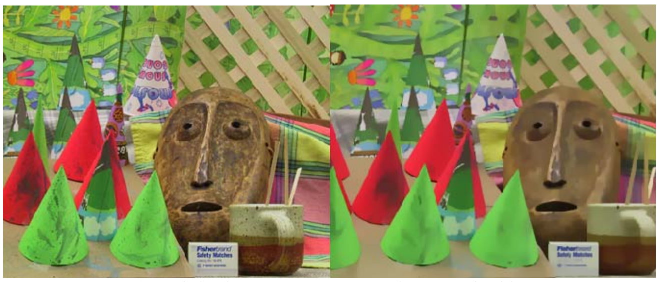


Figure 2. Block diagram of proposed stereo matching algorithms.

A. Edge-preserving image smoothing

The first step in the work-flow is to suppress textures from a reference image and a target image while preserving significant image structures. Since contours of objects are a significant cue for stereo matching and depth discontinuous occur around objects, extracting meaningful contours of objects are very important. In this paper, we employ structure extraction from texture via a relative total variation smoothing method to obtain a smoothed image [5].



(a) original (b) smoothed image

Figure 3. Edge-preserving image smoothing

Relative total variation (RTV) is a simple and efficient method to remove texture while conserving meaningful edges. The inherent characteristic difference between textural and structural edges provides a cue to distinguish those edges. RTV uses the ratio between the sum of absolute variation and the absolute sum of variation. The energy function is represented as

$$E_{RTV} = \sum_{p} (\hat{R}_p - I_p)^2 + \lambda \cdot RTV(p) \tag{1}$$

where $\hat{R}$ is a reconstructed image, I denotes an input image, ant $RTV(p)$ is RTV regularizer shown as

$$RTV(p) = \frac{D_x(p)}{L_x(p) + \varepsilon} + \frac{D_y(p)}{L_y(p) + \varepsilon}$$

$$\text{where}$$

$$D_x(p) = \sum_{q \in R(p)} g_{p,q} \left| (\partial_x \hat{R})_q \right|, D_y(p) = \sum_{q \in R(p)} g_{p,q} \left| (\partial_y \hat{R})_q \right| \tag{2}$$

$$L_x(p) = \left| \sum_{q \in R(p)} g_{p,q} (\partial_x \hat{R})_q \right|, L_y(p) = \left| \sum_{q \in R(p)} g_{p,q} (\partial_y \hat{R})_q \right|$$

where $g_{p,q}$ is Gaussian convolution, and ε is a small positive number. Figure 3 shows the suppressed image. Minor texture edges are removed as Fig. 3(b) is shown.

B. MRF-based energy function

Depth estimation is a kind of multi-labelling problem. Given stereo images, the stereo matching task is to estimate best labels for each pixel. Therefore, we use MRF-MAP based energy function represented as

$$E(d) = \sum_{s} D_s(d_s) + g(x,y) \sum_{s,t \in N(s)} S_{s,t}(d_s, d_t) \tag{3}$$

where $D_s(\ )$ is the data term and $S_{s,t}(\ )$ is the smoothness term. d_s represents disparity or label for each pixel. The data term is the term for how well the pixels match up for different disparities. The matching cost as data term is represented as

$$D_s(d_s) = \alpha C_c(d_s) + \beta C_g(d_s) + \lambda C_{RTV}(d_s) \tag{4}$$

where *α, β,* and *λ are* weights. $C_c(d_s)$ is a sum of absolute intensity differences using intensity, $C_g(d_s)$ denotes a sum of absolute gradient differences. The information of the gradient tells how quickly an image is changing, and gives a directional change in the intensity or color in the image. $C_{RTV}(d_s)$ is a sum of deburred intensity differences using RTV. For a point in one image, there are many possible corresponding points in the second image, especially in repetitive regions. To handle the repetitive texture problem, we add RTV term in the energy function. Each term are represented as

$$\begin{aligned} C_c(d_s) &= \sum_{i,j \in W} I_r(x+i, y+j) - I_t(x+i+d_s, y+j)|, \\ C_g(d_s) &= \frac{1}{2} \left(\left| \nabla_x I_r(x+i, y+j) - \nabla_x I_t(x+i+d_s, y+j) \right| \right. \\ &\quad \left. + \left| \nabla_y I_r(x+i, y+j) - \nabla_y I_t(x+i+d_s, y+j) \right| \right), \\ C_{RTV}(d_s) &= I_{RTV\ r}(x,y) - I_{RTV\ t}(x+d_s, y)| \end{aligned} \tag{5}$$

where I_r and I_t are reference and target images respectively, and I_{RTV} is a smoothed image. Smoothness term is restricted to measuring the differences between neighboring pixels' disparities. We use smoothness term as truncated L2-norm defined by

$$S_{s,t}(d_s, d_t) = \min(\lambda \mid d_s - d_t \mid, T_s) \qquad (6)$$

where Ts is the truncation value to constrain the high-cost increase. The depth discontinues regions, such as edge region, which leads to ineffective corresponding matching, and corresponding matching often fails due to its ambiguity in the homogeneous region. Therefore, handling the corresponding ambiguities and depth discontinuities are extremely important. Therefore, we exploit normal entropy of color in a window and partial derivatives respectively in x- direction and y-direction to formulate weight function. Entropy is the average amount of information contained in each message received. Here, message stands for an event, sample or character drawn from a distribution. Entropy thus characterizes an uncertainty about our source of information. When taken from a finite sample, the entropy can explicitly be written as

$$H(X) = -\sum_i p(x_i)\log_b p(x_i) \qquad (7)$$

Entropy can be normalized as the entropy is divided by the maximum entropy. The normalized entropy is shown as

$$\eta(X) = -\sum_{i=1}^{n} \frac{p(x_i)\log_b p(x_i)}{\log_b p(n)} \qquad (8)$$
$$\textit{where } Entropy_{\max} = \log_b p(n)$$

where $\eta(X)$ is entropy, and $Entropy_{max}$ is maximum entropy in a window. We exploit normalized entropy and partial derivatives respectively in x- direction and y- direction to formulate the weight functions respectively written as

$$g_x(x,y) = \exp\left(\frac{-\mid \nabla_x I(x,y) \mid}{256}\right)\exp\left(\frac{-\eta(x,y)}{Entropy_{\max}}\right)$$
$$g_y(x,y) = \exp\left(\frac{-\mid \nabla_y I(x,y) \mid}{256}\right)\exp\left(\frac{-\eta(x,y)}{Entropy_{\max}}\right) \qquad (9)$$

where $g_x(x,y)$ and $g_y(x,y)$ are the weight functions in x-direction and y- direction respectively. $\nabla f_x(x,y)$ and $\nabla f_y(x,y)$ are partial derivatives. The weight functions adjust smoothness weights. If color entropy in a window or gradient are high, smoothness term has more influence on our results. After formulating the energy functions, we employ the multi label algorithm which calculates optimal solution, is called the "alpha-expansion" algorithm [6].

C. Occlusion handling

Occlusion means that occluded pixel is apparent to the source image, but there is no corresponding pixel in the target image. Therefore, occlusion is the key and challenging problem with stereo matching. Figure 4 represents the occlusion and the non-occlusion. Because an object is obscured by the view of some objects or regions, occluded pixels are only visible from the reference image, but in the target image. We exploit occlusion detection and occlusion refinement algorithms to handle the occlusion region [7].

Given the initial disparity maps, we first apply a median filter to remove noise. And, we exploit the ordering constraint, uniqueness constraint, and color similarity constraint to estimate occlusion region. Ordering constraint predicts candidates of occluded pixels. Pixels in the reference image are projected to the target image. In the case of many-to-one mapping, a pixel which possesses the largest disparity value is selected as the visible pixel, but the rest of the matching pixels become occluded pixels. Uniqueness constraint evaluates the mutual consistency from both disparity maps and both color images. If a particular pixel in the image is not an occluded pixel, the disparity values from the left and the right disparity maps should be consistent. Color similarity mean two pixels between reference and target images should be similar. The color similarity is measured by the Euclidean distance. After detecting occlusion region, the reasonable disparity value should be filled to the occluded pixel. The occlusion hole-filling equation is defined as

$$O(s,d) = \operatorname*{argmin}_t \frac{1}{dist(s,t)}\exp\left(-\frac{dif_{s,t}}{\sigma^2}\right)$$
$$dif_{s,t} = \sum_{c\in\{R,G,B\}} \mid I_c(s) - I_c(t) \mid \qquad (10)$$

where dist(s, t) is Euclidean distance from s to t, and $\text{dif}_{s,\,t}$ denotes color dissimilarity.

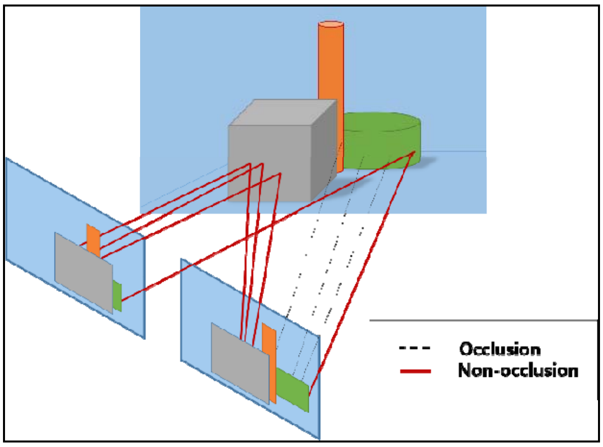


Figure 4. Occlusion and non-occlusion: The red lines are only visible in the left and right images

TABLE I
PERFORMANCE COMPARISON

Algorithm		CSBP	GC + occ	CCH + SegAggr	Jang's method	Proposed method
Tsukuba	nonocc	2.00	1.19	1.74	1.42	**3.38**
	all	4.17	2.01	2.11	2.30	**4.75**
	disc	10.50	6.24	9.23	7.94	**15.96**
Venus	nonocc	1.48	1.64	0.41	0.91	**1.10**
	all	3.11	2.19	0.94	1.54	**1.78**
	disc	17.70	6.75	3.97	12.71	**17.57**
Teddy	nonocc	11.10	11.20	8.08	6.34	**6.49**
	all	20.20	17.40	14.3	13.62	**11.43**
	disc	27.50	19.80	19.80	17.59	**20.73**
Cones	nonocc	5.98	5.36	7.07	4.96	**4.59**
	all	16.50	12.40	12.90	12.70	**12.21**
	disc	16.00	13.00	16.30	14.44	**13.62**

III. EXPERIMENTAL RESULTS

To evaluate the performance of our method objectively, we exploit the percentages of mismatching pixels (BPR) with known ground truth disparity, provided by [8]. 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 method. Figure 5 represents the results of the proposed method, conventional method, and ground truth. The results generated by our proposed method exhibit fewer artifacts than the disparity maps generated by the conventional method. In the experiment, α, *β,* and *λ* weights for data term are set to 0.4, 0.3, and 0.3 respectively.

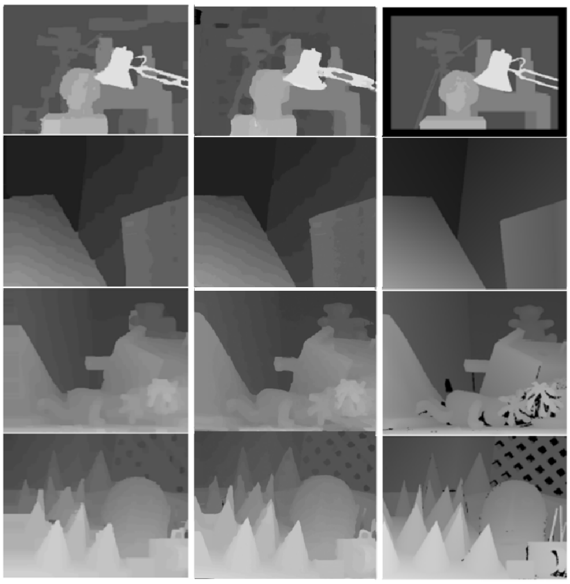


Figure 5. Final disparity maps. (a) Jang's method [2] (b) Proposed method (c) Ground truth

IV. CONCLUSIONS

This paper proposed a method for estimating disparity handling the corresponding ambiguities. The proposed method exploited the MAP-MRF model to formulate an energy function considering correspondence problems such as homogeneous region, depth discontinuous, repetitive texture, and occlusion. After optimizing the energy function via alpha expansion, we applied occlusion handling to enhance the accuracy of disparity. From the experimental results, we have confirmed that the proposed method produces more accurate disparity map compared to other methods regarding bad pixel rates.

ACKNOWLEDGMENT

This research was supported by the ‘Cross-Ministry Giga KOREA Project’ of the Ministry of Science, ICT and Future Planning, Republic of Korea (ROK). [GK17C0100, Development of Interactive and Realistic Massive Giga- Content Technology]

REFERENCES

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

[2] Y. Wei, and L. Quan, “Region-based progressive stereo matching,” Computer Vision and Pattern Recognition, vol. 1, pp. 106-113, 2004.

[3] R. Verma, H. S. Singh, and A. K. Verma, “Depth estimation from stereo images based on adaptive weight and segmentation,” The Institution of Engineers, vol. 93 no. 4, pp. 223-229. 2012.

[4] T. Liu, P. Zhang, and L. Luo, “Dense stereo correspondence with contrast context histogram, segmentation-based two-pass aggregation and occlusion handling,” Advances in Image and Video Technology, pp. 449–461, 2009.

[5] L. Xu, Q. Yan, Y. Xia, and J. Jia, “Structure extraction from texture via relative total variation.” ACM Trans. Graph., vol. 31, no. 6, pp.139:1 -139:10, 2012.

[6] Y. Boykov, O. Veksler and R. Zabih, "Faster approximate energy minimization via graph cuts", Pattern Analysis and Machine Intelligence, vol 23, no.11, pp 1222-1239, 2001.

[7] E. Baek and Y. Ho, “Occlusion detection for stereo matching and hole-filling using dynamic programming,” IS&T/SPIE Electronic Imaging, pp. 1-6, 2016.

[8] D. Scharstein and R. Szeliski. Middlebury Data Sets [Online]. Available: http://vision.middlebury.edu/stereo