Quality Preserving Depth Estimation in Sequential Stereo Images

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

***Abstract*— Computational complexity of the local stereo matching method is affected by disparity ranges. In case of depth estimation in sequential stereo images, high computational complexity is a problem in terms of real-time processing. In this paper, we propose a temporal correlation based stereo matching method in sequential images. Using temporal information in a sequential stereo matching method provides inaccurate disparity ranges, since the estimated depth map accuracy is gradually degraded. To preserve the depth map quality in temporal stereo matching procedure, we adopt the guided image filtering for matching cost aggregation. Since the guided image filtering has a similar structure with bilateral filter, it preserves an object boundary region even in restricted disparity search ranges. Inaccurately estimated disparity values from the temporal correlation are compensated by filtering based cost aggregation method. From the experiment results, we check that the proposed depth map acquisition method preserves the depth map quality in temporal domain stereo matching.**

I. Introduction

Depth information is widely used for many kind of computer vision applications. Especially, a depth information is generally acquired through hardware and software. Usually hardware (Kinect v.2 or ToF camera) based depth acquisition method has a limitation depending on the capturing circumstance. Additionally, ToF sensor only perceives the depth value within 4 to 5 meters. Contrary to previous methods, the software based method (stereo matching) estimates the depth information under the less restricted conditions.

The main issue of software based depth estimation is computational complexity problem. Software based method has two different type of depth estimation procedure. Considering all of pixel values while estimating a depth value is called a global stereo matching method. The other method only considers predefined window kernel for stereo matching. That method is called local stereo matching method. As we infer from the name of matching method, global stereo matching method takes a more time than local stereo matching method.

The goal of the global stereo matching method is finding a solution, which has a lowest energy function. Usually used global matching method is Markov Random Field (MRF) for energy function. Global matching is composed with two cost term, which data term and smoothness term. The data term represents a local matching cost (e.g. SAD, SSD, NCC) and the smoothness term considers a spatially smoothed edge region in stereo image. This cost function is minimized using global energy minimization techniques, such as graph cut and belief propagation(BP). As referred to the problem of global matching method, finding an optimal cost value takes a most of the time in the overall matching procedure. Also the global matching method is not well scaled to high-resolution image or wide label space. Fast cost optimization method [1] also does not generates an accurate depth image.

For fast stereo matching, using sequential images and the global based matching method are not proper in terms of computational complexity. To compensate the disadvantage of time consuming in a global matching, many GPU computing based energy optimization methods have been developed [2, 3]. Even though the computational complexity problem is treated, data and smoothness term has a drawback. The optical flow [3] based method has an effect only small valid displacements. The other problem is occurred in smoothness term. That might over smooth the overall cost function than our expectation. Because of the over smoothed region, the object boundary region will be collapsed while computing the matching costs.

As an alternative solution of the global based stereo matching method, the local based stereo matching approach is adopted for fast depth estimation. Since the accuracy of local matching based depth value is lower than global based matching result, we follow several steps to improve the depth image quality; 1) computation of matching cost; 2) cost aggregation; 3) disparity optimization; 4) disparity refinement.

For initial displacement result from stereo image, sum of absolute differences (SAD), sum of squared differences (SSD) and normalized cross correlation (NCC) pixel matching methods are used within the predefined window kernel in step 1. After computing the matching cost value, various cost aggregation methods [4, 5] are applied to previously obtained cost value in step 2. To find an optimal disparity value, we

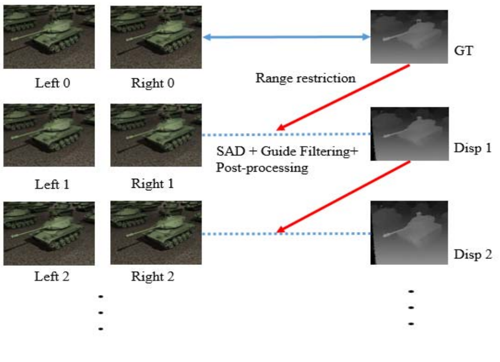


Fig. 1 Pipeline of the proposed method

compare each pixel's cost value along the disparity search range. Among the many cost values, we only select the smallest cost value for assign a disparity value for specific pixel in step 3. In step 4, to compensate disparity error and removing occlusion region, which oppositely visible in stereo image. For occlusion handling left and light image consistency check method [6] and hole filling method [7] are developed.

Most of computational complexity increase because of the disparity search range. For fast depth estimation, we propose temporal domain related disparity estimation method as represented in Fig. 1. We use ground truth(GT) information as an initial disparity image. If the GT is not provided, local stereo matching result also used as initial disparity image. Since the general SAD matching method does not generate a correct disparity image in terms of object boundary region, we adopt the guide filtering (GF) for cost efficient cost aggregation (step 2). The estimated disparity image is used for restriction of disparity search range in stereo matching of sequential images.

II. Temporal depth estimation

A. Matching cost computation

General local window based stereo matching method computes cost function by comparing with neighbor pixel values.

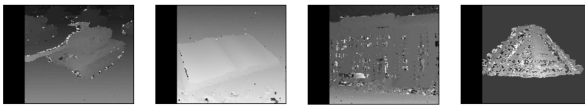
$$\sum_{(i,j)\in W} |I_1(i,j) - I_2(i,j+d)|$$

$$\sum_{(i,j)\in W} \left(I_1(i,j) - I_2(i,j+d)\right)^2 \qquad (1)$$

$$\frac{\sum_{(i,j)\in W} I_1(i,j)\cdot I_2(i,j+d)}{\sqrt[2]{\sum_{(i,j)\in W} I_1^2(i,j)\cdot \sum_{(i,j)\in W} I_2^2(i,j+d)}}$$

Widely used local stereo matching cost function is indicated in (1). From top to bottom each equation represents SAD, SSD and NCC. SSD has a higher computational complexity compared to SAD method as it involves multiplication operations. Furthermore, NCC is even more complex than both SAD and SSD method as it contains numerous multiplication, division and square root operations.

The purpose of proposed temporal domain depth estimation is reducing the computational complexity for each sequential image matching result, since we adopt the SAD cost computation function in our method. Fig. 2 shows estimated disparity map only using SAD matching cost.



Tanks Book Street Temple

Fig. 2 Initial disparity map generated by SAD function

B. Guide filtering for cost optimization

Local matching methods use winner-takes-all (WTA) to find an optimal cost value. The basic concept of WTA is selecting a minimum cost value which correlated to optimal disparity value for each pixel. But SAD method just computing a cost value within defined window kernel as indicated in (1). Inside of window kernel pixels does not consider object boundary region or homogeneous region with regard to overall image pixels. Since applying WTA to SAD result just select minimum cost value, the estimated disparity value is not accurate especially around object boundary and homogeneous region.

To overcome that kind of issue in cost aggregation, we adopt the guided image filtering [8] for initial disparity cost aggregation step. Through the SAD cost function, we get a different cost volume depending on a number of disparity search range. Cost volume *L* compose of different cost volume label *l,* a set of volume is defined as $L = \{1, \ldots, l\}$. The cost volume C includes three dimensional array, which related to cost for choosing optimum label *l* at each pixel.

All of the cost volume set *L* is filtered using guided image filtering. The output of guide filtering at pixel index *i* at set of labels is defined weighted average of all pixels within the same cost label. The weighted cost function is indicated in (2).

$$C' = \sum_j W_{i,j}(I)C_{i,j} \qquad (2)$$

where C' indicates filtered result of cost volume set and *W* is filter weight which related to guide filtering. Depending on input image *I*, guide filtering affect to result of initial cost aggregation. In our cost aggregation procedure, we use input stereo image as guidance image.

Guided image filtering simultaneously considers input image with initially estimated disparity image $C_{i,j}$. The concept of filtering is similar to bilateral filter (BF), but guided image filtering builds a linear model with regard to guidance image as indicated in (3).

$$\mathrm{E}(a_k, b_k) = \min_{(a,b)} \sum_i (aI_i + b - p_i)^2 + \varepsilon a^2 \qquad (3)$$

In (3), ε is a regularization parameter penalizing large a_k and p_i indicates the initially estimated disparity image at i^{th} label. By solving (3) through linear regression method the coefficient a_k and b_k are updated. After finding coefficient values, we solve the equation in terms of error minimization with initial disparity image and guidance image using filter kernel which explicitly expressed as (4). Specific process of filter kernel induction is derived in [8].

$$\mathrm{W}_{i,j} = \frac{1}{|\omega^2|} \sum_{k:(i,j)\in\omega_k} (1 + \frac{(I_i - \mu_k)(I_j - \mu_k)}{\sigma_k^2 + \epsilon}) \qquad (4)$$

where $\mu_k$and σ_k are indicate mean and variance of input image *I* in defined window ω_k with centered at pixel k. ϵ is a smoothness parameter, which correctly distinguish the object boundary region between different pixel values. Since general matching cost function SAD does not preserving edge information of stereo image, we adopt the guided image filter.

To perceive how to guided image filtering preserve boundary region, we firstly check Fig. 3. The main objective of guided image filtering in our temporal matching method is preserving the boundary region between two different object.

Fig. 3 shows test image *Tanks*, which include tanks and ground region. The numerator $(I_i - \mu_k)(I_j - \mu_k)$ in (4) has a positive sign if I_j is located on same side with I_i, however it has negative sign when they located in different side. As indicated in Fig. 3, if I_i and I_j has same side on the tank region, they has positive sign, however I_j indicates the ground region then that has a negative sign.

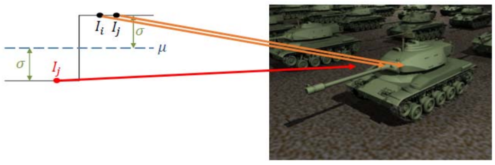


Fig. 3 1-D step in test image

Guide filtering also taken an effect from smoothness factor ϵ in (4). It controls the strength of averaging rate of initial disparity value. When σ_k^2 is very smaller than ϵ , then numerator in (4) is much smaller than denominator, so that average factor μ_k also automatically diminished similar to I_i and I_j. Hence, the guide filter kernel in (4) converge to low-pass filter. The low-pass filter only passes low frequency domain (inside of object region), hence the object boundary region is smoothed than other region. Because of that reason, using a proper ϵ is critical issues. For our experiment, we apply ϵ value 0.001.

III. Disparity Selection and Refinement

From the guided image filtering results, initially estimated disparity volumes(L) are filtered based on input stereo image. Generally used SAD matching method, a disparity value is determined by choosing a smallest cost value align with the disparity search range along horizontal line. However, guide filtered image result does not determine an optimal cost value for each pixel among disparity volumes. The guide filtered cost volumes are shown in Fig. 4.

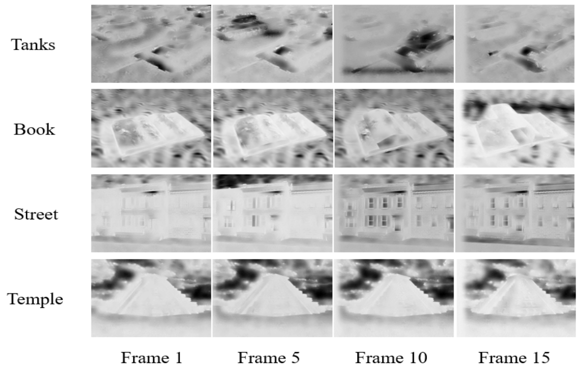


Fig. 4 Guide image filtered volume

To select an optimal disparity value among disparity values, WTA algorithm is applied to disparity volumes. For this purpose, the matching cost values are calculated for each disparity candidates independently and sequentially. As a result of that, the candidates with minimum cost value are assigned to the corresponding pixel coordinate. The cost computation procedure includes aggregation and optimization step. As a result of that procedure, middle of Fig. 5 shows optimized disparity images obtained by WTA method.

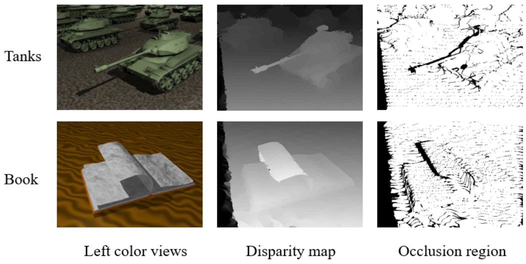


Fig. 5 Filtered disparity map and cross-check results.

Even the disparity volumes are optimized using guide filtering, occlusion area which does not detected for left and right viewpoints simultaneously. Occlusion area affect to accuracy of disparity value while performing stereo matching.

Those occluded region has to be handled such a way that reliable information correctly allocated geometrically accurate disparity value. To find an occluded region in filter disparity image, cross-check is conducted between left and right viewpoint based disparity images. The result of cross-check is demonstrated in last column of Fig. 5.

Occlusion area usually removed by assigning lowest disparity value of the spatially closest non-occluded pixel which lie on the same horizontal line. A simple occlusion handling strategy generates streak artifact on occlusion handling result. To handle the occlusion area with efficient strategy, we apply a weighted median filter for accurate disparity value. The weighted median filter uses weight term to improve the normal median filter. We can use guided image filtering as a filter weight, but we already apply the guided image filtering on disparity volume aggregation step. The guided image filtering requires many times of multiplication and matrix inversion operation. Those kinds of operation increase the computational complexity for overall processing.

Our purpose of proposed method in sequential image stereo matching is providing an accurate depth image throughout low computational complexity. Hence, we apply the bilateral filter as indicated in (5).

$$W_{i,j}^{bf} = \frac{1}{K_i}\exp(-\frac{|i-j|^2}{\sigma_s^2})\exp(-\frac{|I_i - I_j|^2}{\sigma_c^2}) \qquad (5)$$

where I indicates image intensity value and σ_s and σ_c are spatial and color similarity, K is a normalization factor.

IV. Experiment Results

We use Cambridge provided test sequences; *Tanks*, *Book*, *Temple* and *Street*. For bilateral filter parameter $\{\sigma_s, \sigma_c\}$={9, 0.1} and SAD matching window kernel size is predefined 5.

The proposed method is implemented using C++ with Intel i-7 5960X processor. Our method takes about 4 second for 15 sequential test image stereo matching. The computational complexity is reduced in SAD matching procedure, since we restrict the disparity search range based on previously obtained disparity value. However, guide image filtering and filter based occlusion handling strategy offset the computational complexity reduction. Since we aggregate the disparity volume using guided image filtering, the error propagation effect is restricted in sequentially estimated disparity results. The experiment results of sequential matching are indicated in Fig. 6.

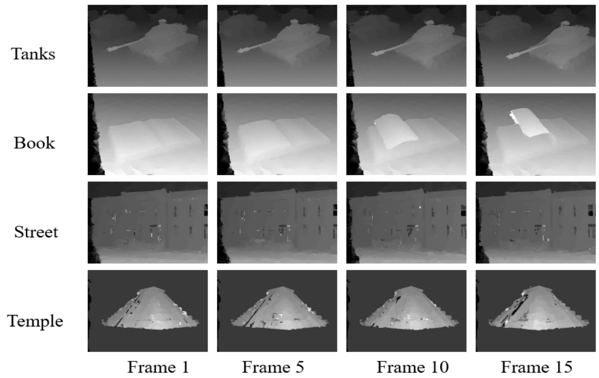


Fig. 6 Final disparity image in sequential test sets

Since we apply the guide filtering for disparity optimization, the overall process of time become more complex than conventional method. Table 1 shows computational complexity between proposed method and conventionally used temporal domain stereo matching strategies [9]. As we expected, the computational complexity of proposed method is higher than conventional methods.

TABLE I
BPR COMPARISON RESULTS

Test Sequences	Computational complexity(sec)		
	[9] Iterative	[9] Given	Proposed
Tank	3.43	3.14	4.12
Street	3.91	3.71	4.33
Temple	4.14	3.92	4.56
Book	3.07	2.86	3.38

However, in case of BPR value, conventional temporal stereo matching methods show about 32.1% BPR value [9], but, proposed method extremely reduces the average BPR value to 7.83%. Fig. 7 shows BPR variation of each test sequences.

As indicated in Fig. 7, the BPR is not constant but increasing rate is not linear when it compares to conventional temporal domain matching methods. Since we only take a restricted disparity search range, the cost optimization is conducted regardless of window kernel. Because of that reason, we can improve the accuracy of estimated disparity values in temporal domain stereo matching.

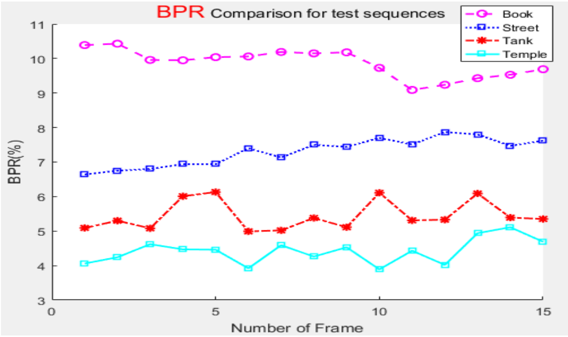


Fig. 7 BPR comparison of test sequences

V. CONCLUSION

To prevent the error propagation artifact in sequential images, we adopt the guided image filtering for cost volume aggregation. Since the guided image filtering preserves the object boundary region using input stereo image, the optimized disparity value is more accurate than conventionally obtained result. Even though the disparity values are optimized, occlusion area does not have a correct disparity value. To efficiently handle that region, the occlusion region is detected using a cross-check strategy. The extracted occlusion region is removed using a weighted median filtering. Although the computational complexity is increased than conventional methods, we can get an accurate disparity value in sequential image sets.

ACKNOWLEDGEMENT

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

REFERENCES

[1] O. Veksler, “Stereo correspondence by dynamic programming on a tree,” *CVPR*, vol. 2, pp. 384-390, June 2005.

[2] T. Pock, T. Schoenemann, G. Graber, H. Bischof, and D. Cremers, “A perceptually motivated online benchmark for image matting,” *CVPR*, pp. 1826-1833, June 2009.

[3] P. Gwosdek, H. Zimmer, S. Grewenig, A. Bruhn, and J. Weickert, “A highly efficient GPU implementation for variational optical flow based on the euler-lagrange framework,” *CVGPU* Workshop, vol. 6554, pp. 372-383, Sept. 2010.

[4] K. J. Yoon and I. S. Kweon, “Adaptive support-weight approach for correspondence search,” *TPAMI*, pp. 650-656, April 2006.

[5] K. Zhang, J. Lu and G. Lafruit, “Cross-based local stereo matching using orthogonal integral images,” *TCSVT*, vol. 19, no. 7, pp. 1073-1079, April, 2009.

[6] S. D. Cochran and G. Medioni, “3-d surface description from binocular stereo,” *TPAMI*, vol. 14, no. 10, pp. 981-994, Oct. 1992.

[7] S. Birchfield and C. Tomasi, “A pixel dissimilarity measure that is insensitive to image sampling,” *TPAMI*, vol. 20, no. 4, pp. 401-406, Apr. 1998.

[8] K. He, J. Sun and X. Tang, “Guide image filtering,” *PAMI*, vol. 35, no. 6, pp. 1397-1409, Oct. 2012.

[9] J.H. Mun and Y. S. Ho, “Temporally consistence depth estimation from stereo video sequences,” *PCM*, pp. 579-588, Sept. 2015.