Temporal Consistent Stereo Disparity Search using Belief Propagation with Adaptive Weighted

Hsin-Yu Hou, Sih-Sian Wu, Da-Fang Chang, Liang-Gee Chen, *Fellow, IEEE*
DSPIC Lab, Department of Electrical Engineering
National Taiwan University, Taiwan
hsinyuhou13579@gmail.com, benwu@video.ee.ntu.edu.tw, b02901126@ntu.edu.tw, lgchen@ntu.edu.tw

***Abstract*—This paper introduces a temporal consistent stereo matching algorithm with the adaptive discontinuity penalty. Compared to previous temporal consistent methods for stereo matching, window-based similarity and adaptive weight for temporal cues are imposed on the energy function. We solve the energy function in a belief propagation framework. Consequently, the proposed algorithm outperforms previous methods in those stereo video datasets using for temporal consistency evaluation. As the result, the proposed method can solve problems inducing from previous methods like error propagation from previous occlusion region and ease the problem from repeated pattern area.**

Keywords*—*Temporal Consistency, Stereo Matching, Disparity Estimation.

I. Introduction

The requirement of accuracy information becomes more and more urgent according to the prospering development of the autonomous car, 3D interaction application and augmented reality (AR). There are many 3D capture techniques including active methods, structure light [7] and time-of-flight (ToF) [8], and passive methods, stereo matching and light-field [9]. Stereo matching method is most potential for mobile or wearable device since its efficient power consumption and relative low device requirement. In order to get the disparity map of the environment, many complex algorithms have been proposed to improve the performance such as [13][12][15] [3]. These algorithms focus at improving disparity image per image without taking temporal information into consideration. Although temporal propagation is mentioned in Patchmatch Stereo [3], it represents the constraint information in a same input image during iterative procedure rather than within different time frame images.

To leverage the character, we introduce the method combining stereo matching and time-consistency. For serial frames of films, the disparity map of each frames should be similar to their precedent frame. The original disparity map can be refined with the time-consistency attribute. Therefore, the disparity map with better performance and more stability can be provided by the proposed algorithm. The three main contribution of this paper are summarized as follows

1. A temporal cue is adopted in cost function which can be employed into conventional stereo matching methods seamlessly.

2. The proposed method is verified in both cases of the dynamic scene with static stereo camera and static scene with a moving one.

3. A stability function is proposed to represent the level of disparity flicker.

The rest of this paper is organized as follows: In Section III the proposed method is introduced. The experimental results and discussion are shown in Section IV. Finally, the paper is concluded in Section V.

II. Related works

Temporal consistent issue of disparity estimation has attracted more attention recently. It assumes that 3D information of following frames must have highly correlation. Pham et.al. [5] proposed a algorithm using spatio-temporal cues. Vretps et.al. [4] also proposed a techniques to filter outliers out with statistic distribution to improve temporal robust. A 3D bilateral volume for filtering is proposed to enhance temporal consistency [6] which achieved great improvement in time domain. It assume that the corresponding points in previous frames are located at the same position exactly.

Above assumption is not suitable for dynamic scenes since the proper corresponding points are not always in the same location. Optical flow algorithms [27] and [28] are induced to find proper correspondence points and adopted by [24] and [19], respectively.

However, forward and backward successive frames information are required causing this technique is not practical for our desire.

Window-based temporal consistent method with [6] aggregating supporting pixel within adjacent frames. To provide better performance global method are used to preserve temporal consistent. Energy function based methods adopting consistent cues into data term or smooth term. More complex components are considered such as mesh [18] and hyperplanes [21]. Lv et.al. [23] estimated the scene geometry after segment-based processing.

III. Proposed method

A. Algorithm Overview

Since the movement of the object in the video is smooth, it is expected that the disparity of the object will not change too much. In our proposed algorithm, the disparity of previous frame are used for guidance of temporal consistent. To improve the robust, we use patch similarity to adjust the proper penalty from temporal discontinuity. Our algorithm flow can be expressed as Fig.1.

The proposed energy function is represented as

$$E(p,d) = E_d(p) + E_s(p,d) + E_t(d,\overline{p}^{t-1}). \quad (1)$$

Where $E_d^t(p)$ means the unary term for pixel p in frame t preserving the appearance consistency. $E_s(p,d)$ is represented pairwise term for pixel p and neighboring pixels q maintaining the smoothness. The last element, $E_t(d,\overline{p}^{t-1})$, is the proposed temporal term keeping the temporal consistency.

In the first frame, the proposed method applies conventional stereo matching method to obtain a disparity map for the first left frame. In our implementation, the Adaptive Support-weight [1] is adapted. The filtered disparity by weighted median filter [2] is used to guide next frame disparity estimation. After doing the raw matching disparity, we adapt the optical flow algorithm from Lucas and Kanade [16] to find the corresponding in the previous frame of each pixel. With the corresponding points of each pixel, the temporal discontinuous penalty can be computed and the proper weight can also be determined by the proposed method. As a result, the desired disparity value with temporal consistency is generated.

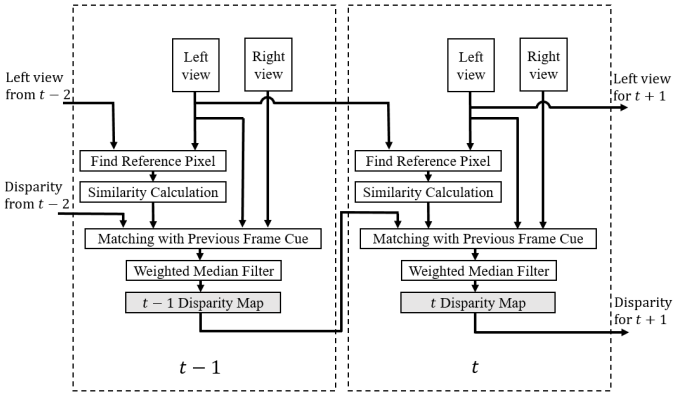


Fig. 1: The workflow of our proposed method.

B. Detail of the Proposed Algorithm

We focus in the temporal consistent term and how to fuse into the final energy function. Data term $E_d^t(p)$ is generated by using adaptive support-weight method [1]. For every possible disparity d_n, the proposed cost function $E(d_n)$ with temporal consistency cue is represented as

$$E(d_n) = \frac{\sum_{q\in N_p, \overline{q}_d \in N_{\overline{p}_d}} w(p,q)w(\overline{p}_d,\overline{q}_d)e(q,\overline{q}_d)}{\sum_{q\in N_p, \overline{q}_d \in N_{\overline{p}_d}} w(p,q)w(\overline{p}_d,\overline{q}_d)}, \quad (2)$$

where $w(p,q)$ is a weight including range weight and spatial weight which can be presented as

$$w(p,q) = exp(-\frac{\triangle c_{pq}}{\gamma_r}) \times exp(-\frac{\triangle g_{pq}}{\gamma_g}). \quad (3)$$

$\triangle c_{pq}$ is the intensity difference of the color of pixel p and q in RGB domain. γ_r and γ_g are parameters in adaptive support weight controlling the sensitivity of each difference. We set $\gamma_r = 5$ and $\gamma_g = 17.5$ with reference to [1]. $e(q,\overline{q}_d)$ can be any desired pixel wise matching cost. In our implementation, the pixel-wise matching cost is followed the setting in [17] which is defined as

$$e(q,\overline{q}_d) = \beta(I_c(q) - I_c(\overline{q}_d)) + (1-\beta)(\nabla I_{Census}(q) - \nabla I_c(\overline{q}_d)), \quad (4)$$

where β is 0.75, $I_c(q)$ is the AD cost and $\nabla I_{Census}(q)$ is the Census cost between pixel p and pixel q.

$$E_s(p,d) = \lambda \cdot min(|l_p - d|, T) \quad (5)$$

is the smoothing term also called pairwise term. It maintains the smoothness in the disparity map. We adopt the truncated linear model in [29]. A weighting according to the color difference between adjacent pixels are also imposed just like [30].

$E_t(d,\overline{p}^{t-1}$ is the temporal consistent term maintaining the smoothness in adjacent frames. The term can be defined as

$$E_t(d,\overline{p}^{t-1} = W_f(p_t, p_{t-1}) \quad (6)$$

Back to the Equ.(2), we use $t-1$ to represent the previous frame, and t is the current frame. d_n is a disparity hypothesis $d_n \in 0, ..., L-1$, where L is disparity range. d_{t-1} is the disparity of previous frame reference point p_{t-1}. We find the reference pixel by using optical flow algorithm from Lucas and Kanade [16]. We add a truncation T_f here to limit the cost from measuring the previous frame. The proposed method adjusts the weight of previous frame by $W_f(p_t, p_{t-1})$. To avoid the error guidance from occlusion or mismatch pixels, the weight is adaptive according to the patch similarity. This weight function can be written as

$$W_f(p_t, p_{t-1}) = \frac{\alpha}{2^{S(p_t,p_{t-1})}}, \quad (7)$$

where α is the parameter that should balance the cost from stereo matching and previous frame cue. $S(p_t, p_{t-1})$ is similarity representing the likelihood of the reference pixel p_{t-1} to be the same object as p_t. The higher similarity indicates that $W_f(p_t, p_{t-1})$ will become larger. In other words, the more similar the corresponding patches are, the disparity value becomes more temporal smoothing. Here we apply similarity measurement method we called Adaptive Support-weight (ASW) [1], which is

$$S(p_t, p_{t-1}) = \gamma \cdot \frac{\sum_{q_t \in N_t, q_{t-1}} w_q(p_{t-1}, q_{t-1})|I(q_t) - I(q_{t-1})|}{\sum_{q_t \in N_t} w_q(p_{t-1}, q_{t-1})}. \quad (8)$$

$w_q(p_{t-1}, q_{t-1})$ is the weight function from the color difference between the center pixel and ones within its supporting window, which can be shown as

$$w_q(p_{t-1}, q_{t-1}) = exp(\frac{-|I(q_t) - I(q_{t-1})|}{\gamma_c}), \quad (9)$$

where γ_c is 10. Either Equ.(**??**) or Equ.(8) can be substituted into weight of Equ.(7). Respective result and performance are presented and discussed in the next section. After calculating the weight function, we add the previous frame cost back in Equ.(2). As a result, d_t, is acquired from proposed cost function via Winner-Takes-All manner as

$$d_t = \arg\min_{d_n \in S_d} \{E(d_n)\} \quad (10)$$

IV. Experimental Result

We evaluate our methods by testing on stereo video sequence [25] and five different scenes from [26]

A. Results I (New Tsukuba)

We apply our method on New Tsukubsa dataset, which is synthetic image series with ground truth. New Tsukuba dataset [25] gives us many circumstance that the camera might move in the view. To evaluate our method on different condition, we split the dataset into two series, which are frame 1-30 and frame 70-100, for the comparison. In the 1st frame to the 30th frame, the camera rotates counter-clockwise seen from the top, and the movement is relatively slow. In the 70th frame to the 100th frame, the camera approaches the stone statue in the center, and the camera movement is faster than that in 1st to 30th. We evaluate the performance of our method by comparing the error rate of disparity map. We compare three methods here. First, original BP without previous frame cue. Second, the disparity map with previous frame cue by [**?**] in APSIPA. Third, our purposed method with linear combination of previous cue.

The method is run with the setting γ is 0.001, T_f is 24, α is 0.04, similarity window size(which is the size of N_t) is 5×5. The table shows the 1-30 image and 71-100 of New Tsukuba dataset. In the table, Err. is the abbreviation for error rate, Red. represents reduction or improvement.

TABLE I: New Tsukuba Frame 1 to Frame 30

	Err.(%)	Err. Red.(%)
BP	10.4829	0
APSIPA	10.4484	0.33
Purposed	10.0893	3.75

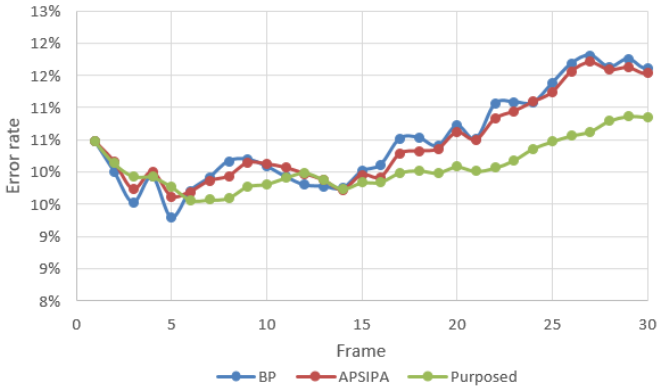


Fig. 2: Error comparison on 1st to 30th frame

TABLE II: New Tsukuba Frame 71 to Frame 100

	Err.(%)	Err. Red.(%)
BP	15.4072	0
APSIPA	14.9047	3.26
Purposed	12.7305	17.37

From Table I and Fig.2, the correctness of the disparity map can be improved for 3% as seen in reduction rate. In the first 10 frame, the movement is relatively slow compared to the following sequence. Consequently, these stereo frames will be similar to their previous frames respectively. We can say that the room for improvement is smaller using the temporal consistency method. We can also see that we get more improvement after 15th frame. On the other hand, from Table II and 3, our purposed method can achieve larger improvement. The result of the disparity map is shown in Fig.4 We also highlight the improvement on the sculpture in Fig. 5. Summarizing, we ensure our method can improve the quality of disparity map and reduce error pixels. The improvement is larger while the camera movement is faster.

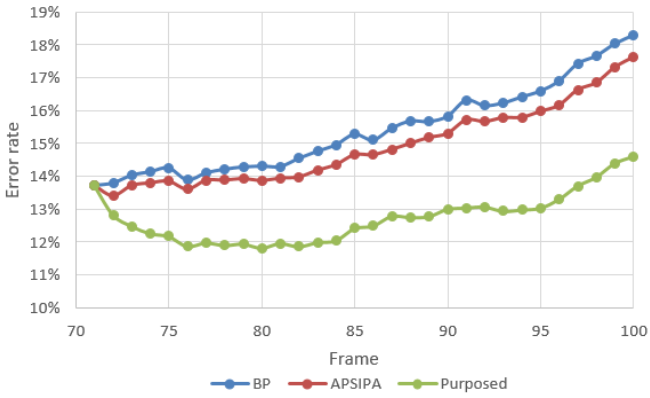
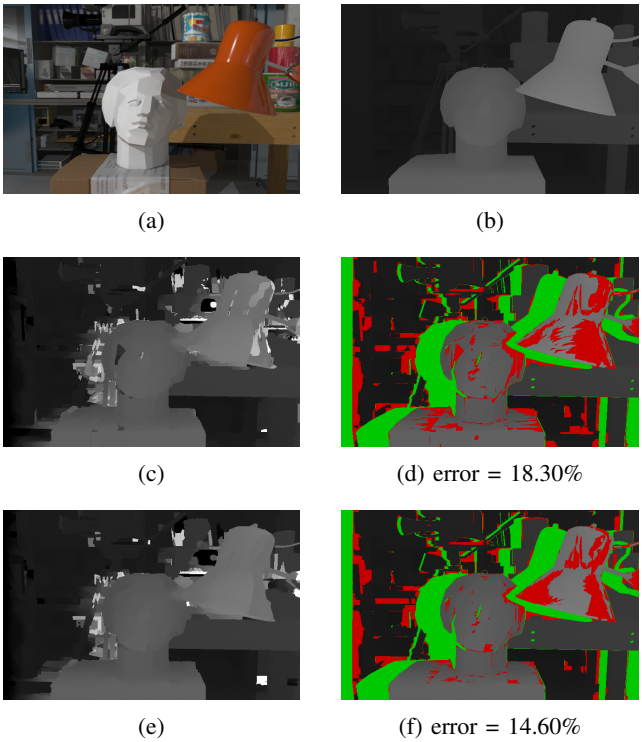


Fig. 3: Error comparison on 71st to 100th frame



(a) (b)

(c) (d) error = 18.30%

(e) (f) error = 14.60%

Fig. 4: Disparity comparison of frame 100. (a) Left image. (b) Ground truth. (c) BP (d) bad pixel(error >2) of BP (e) Our method (f) bad pixel(error >2) of our method

TABLE III: Comparison of New Tsukuba from the frame 71 to the frame 100.

Methods	ICASSP 2017 [21]	BP Only	ECCV 2010 [6]	APSIPA 2015 [?]	BMVC 2015 [19]	Proposed
71 to 100	N/A	15.4	22.8	14.9	N/A	12.7
1 to 30	N/A	18.2	18.2	N/A	N/A	N/A

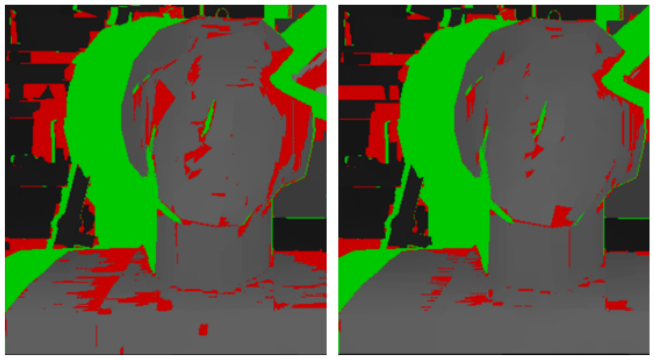


Fig. 5: Error comparison of sculpture in frame 100, green indicates occlusion, red indicates error pixel.(left: BP, right: purposed

B. Result II (DCB Grid)

We use a synthetic dataset with five stereo sequences and ground truth provided by [26].

TABLE IV: Comparison of with dataset [26].

	Book	Street	Tanks	Temple	Tunnel	Average
ICASSP 2017 [21]	NA	NA	NA	14	1.29	NA
ECCV 2010 [6]	8.61	16.39	7.15	11.98	2.05	9.084
BP Only	5.55	14.18	4.09	3.47	1.67	7.01
APSIPA 2015 [?]	5.56	12.87	3.44	8.24	2.87	6.60
BMVC 2015 [19]	3.08	3.66	4.65	5.91	0.21	
Proposed(0.08)	5.24	10.64	4.29	6.73	2.15	5.77
Proposed(0.04)	5.65	11.54	3.47	7.46	1.51	5.92

The result from DCB Grid datasets is presented in Fig.6 and Table IV. These datasets are Book, Street, Tanks, Temple, Tunnel.

Our purposed method can achieve best average quality. We will discuss the tank and tunnel dataset to conclude the characteristic of our method.

For tank dataset, the camera approaches the tank fast. The weight of previous cue will be too large and making a wrong estimation. We suggest that using our method on faster take-in scenary should set α lower. In fact, our error rate can reduced from 3.47% to 3.37% with $\alpha = 0.2$ instead of our standard setting $\alpha = 0.4$.

For tunnel dataset, the result shows that our method can deal with same texture with angles to the camera. The BP costs of candidate depth is closed in repetitive texture, causing the difference of candidate to be smaller. In ASPIA method, the previous depth information is passed into current costs by multiplying a Gaussian weight. Because we choose the depth by winner-take-all, the multiplication will deteriorate the relativity of costs and the result will be wrong. Our method uses linear combination of previous and current costs, which can preserve the relativity of cost from candidate depth.

Summarizing, the improvement of our method is better in four cases and achieve the best in average. As for approaching camera scene, in tank dataset, setting α to a lower value can also achieve a better result.

C. Future Work

Currently, we achieve our result with testing the balance between previous frame and current costs from stereo images. The tanks dataset shows that α should not be a fix value for every scenery. We are currently testing the how α effect the calculation in different dataset. We also plan to reconstruct the 3d model of the consecutive frame to find the camera motion, which will be helpful in balance configuration.

V. Conclusion

The paper proposed a stereo matching algorithm using disparity temporal continuity for temporal consistency. Our purposed method use optical flow to estimate the previous position, and design a linear combination method to improve the disparity map and preserve the relativity as well. We also purposed a ASW(Adaptive Support-weight) method, to calculate the similarity for measuring the confidence of the reference pixel in case the optical flow cannot find the correspondent point for occlusion. By observation, conventional methods without temporal consistent cue introduce flickering effect, the disparity sequence of the video will be unstable. We also discuss the balance between current matching cost and previous frame cue, and find a balance to achieve both the correctness and the stability. With proposed technique, the disparity is improved with the movement of camera.

References

[1] Kuk-Jin Yoon, "Adaptive Support-Weight Approach for Correspondence Search" *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.28, no. 4, pp. 650-656, April 2006.

[2] Q. Zhang, L. Xu and J. Jia, "100+ Times Faster Weighted Median Filter (WMF)," 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, 2014, pp. 2830-2837.

[3] Bleyer, Michael, Christoph Rhemann, and Carsten Rother. "PatchMatch Stereo-Stereo Matching with Slanted Support Windows." Bmvc. Vol. 11. 2011.

[4] N. Vretos and P. Daras, "Temporal and color consistent disparity estimation in stereo videos," 2014 IEEE International Conference on Image Processing (ICIP), Paris, 2014, pp. 3798-3802.

[5] Pham, Cuong Cao, Vinh Dinh Nguyen, and Jae Wook Jeon. "Efficient spatio-temporal local stereo matching using information permeability filtering." Image Processing (ICIP), 2012 19th IEEE International Conference on. IEEE, 2012 pp. 2965-2968.

[6] Richardt, Christian, et al. "Real-time spatiotemporal stereo matching using the dual-cross-bilateral grid." European Conference on Computer Vision. Springer Berlin Heidelberg, 2010.

[7] https://msdn.microsoft.com/en-us/library/jj131033.aspx

[8] http://www.xbox.com/en-US/xbox-one/accessories/kinect

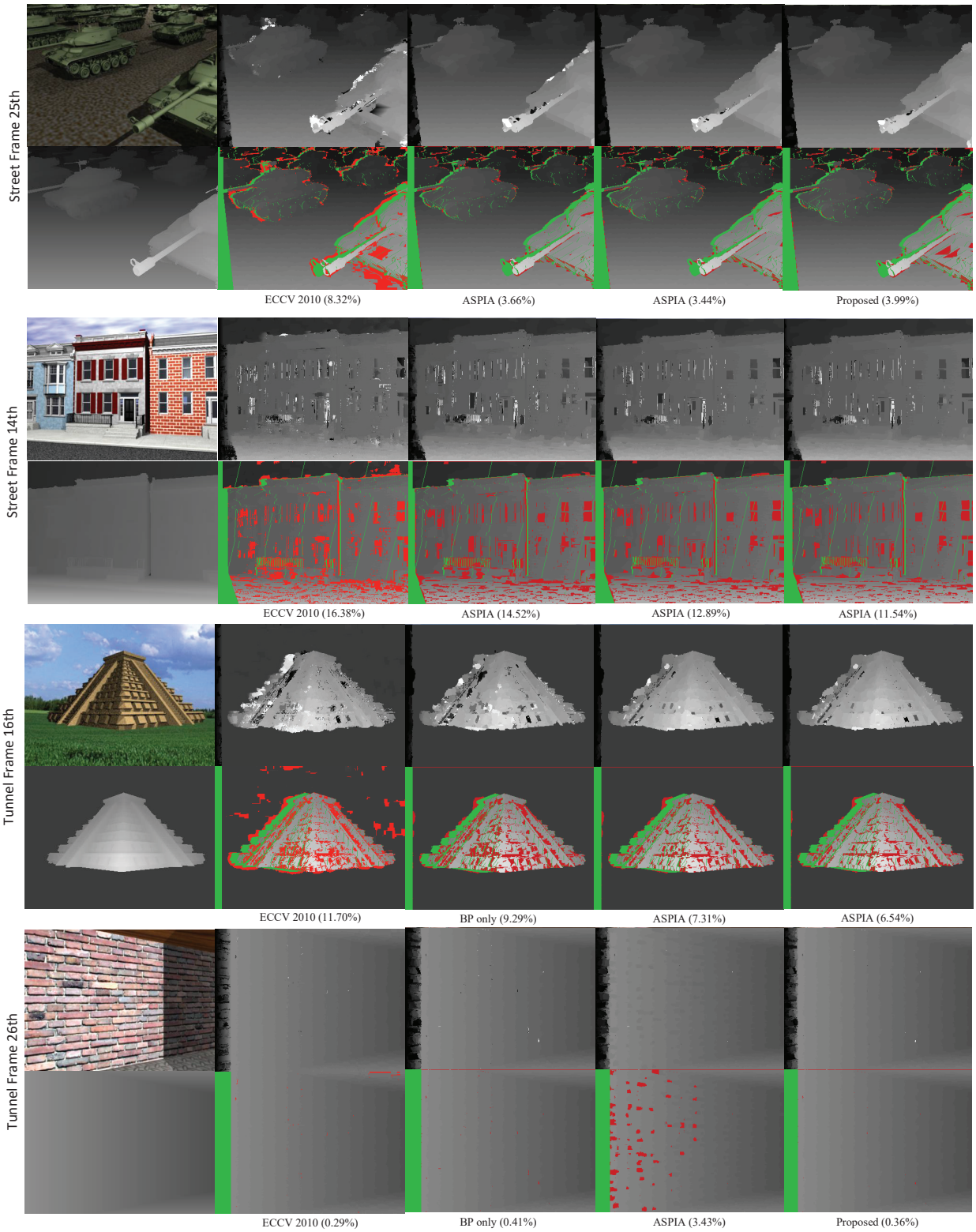


Fig. 6: Comparison figures.

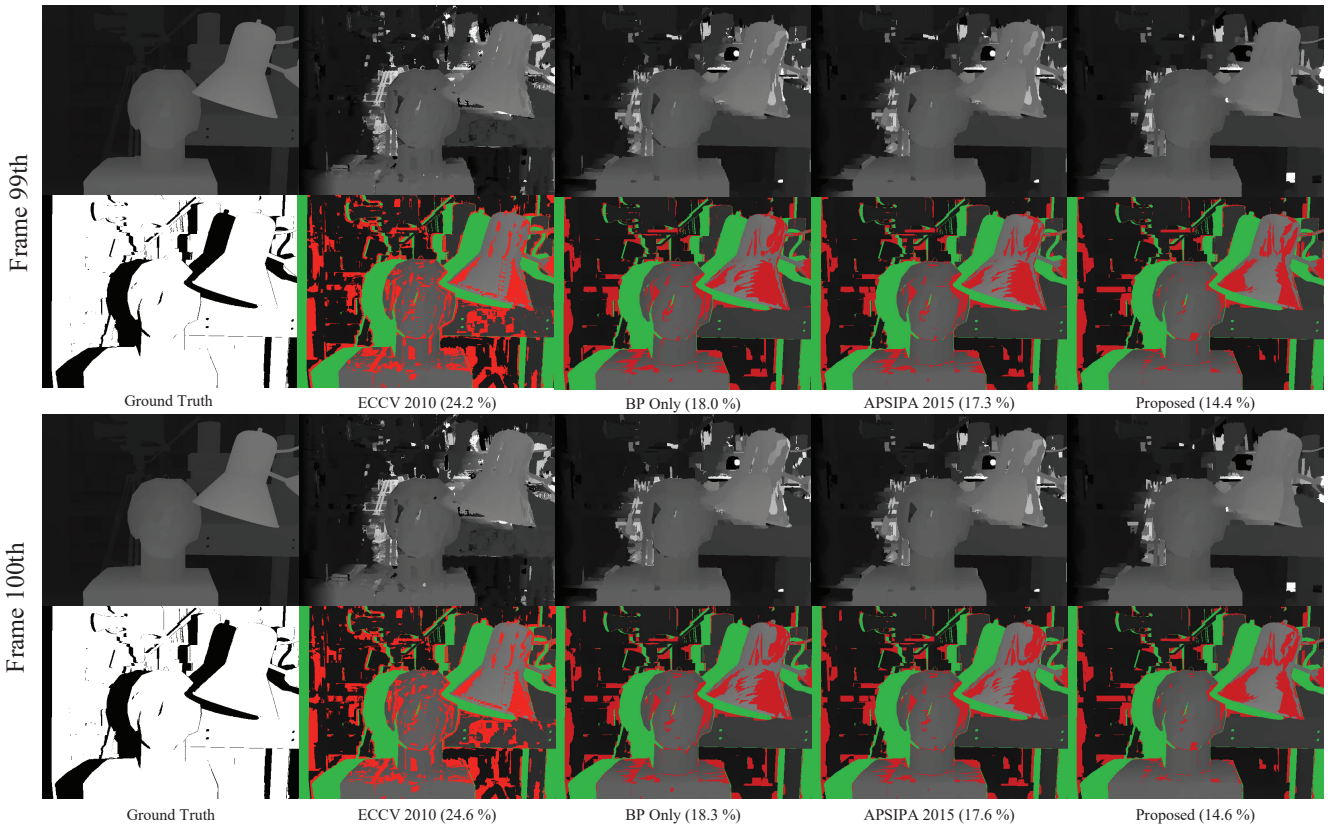


Fig. 7: Comparison figures in New Tsukuba dataset frame 71th to 100th.

[9] https://www.lytro.com/

[10] http://www.fujii.nuee.nagoya-u.ac.jp/multiview-data/

[11] http://cvlab-home.blogspot.tw/2012/05/h2fecha-2581457116665894170-displaynone.html

[12] Zhang, Chi, et al. "Meshstereo: A global stereo model with mesh alignment regularization for view interpolation." Proceedings of the IEEE International Conference on Computer Vision. 2015.

[13] Psota, Eric T., et al. "Map disparity estimation using hidden markov trees." Proceedings of the IEEE International Conference on Computer Vision. 2015.

[14] Besse, Frederic, et al. "Pmbp: Patchmatch belief propagation for correspondence field estimation." International Journal of Computer Vision 110.1 (2014): 2-13.

[15] Yang, Qingxiong. "A non-local cost aggregation method for stereo matching." Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. IEEE, 2012.

[16] Lucas, Bruce D., and Takeo Kanade. "An iterative image registration technique with an application to stereo vision." (1981): 674-679.

[17] Hosni, Asmaa, et al. "Fast cost-volume filtering for visual correspondence and beyond." IEEE Transactions on Pattern Analysis and Machine Intelligence 35.2 (2013): 504-511.

[18] Wang, Zongji, Xiaowu Chen, and Dongqing Zou. "Copy and Paste: Temporally Consistent Stereoscopic Video Blending." IEEE Transactions on Circuits and Systems for Video Technology (2017).

[19] Shin, Yong-Ho, and Kuk-Jin Yoon. "Spatiotemporal Stereo Matching with 3D Disparity Profiles." BMVC. 2015.

[20] Richardt, Christian, et al. "Real-time spatiotemporal stereo matching using the dual-cross-bilateral grid." European conference on Computer vision. Springer, Berlin, Heidelberg, 2010.

[21] H. Nakano, D. Sugimura and T. Hamamoto, "Disparity estimation in stereo videos using spatio-temporal disparity hyperplane models," 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, 2017, pp. 2017-2021.

[22] E. S. Larsen, P. Mordohai, M. Pollefeys and H. Fuchs, "Temporally Consistent Reconstruction from Multiple Video Streams Using Enhanced Belief Propagation," 2007 IEEE 11th International Conference on Computer Vision, Rio de Janeiro, 2007, pp. 1-8.

[23] Lv, Zhaoyang, et al. "A continuous optimization approach for efficient and accurate scene flow." European Conference on Computer Vision. Springer International Publishing, 2016.

[24] Baek, Eu-Tteum, and Yo-Sung Ho. "Temporal stereo disparity estimation with graph cuts." Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2015 Asia-Pacific. IEEE, 2015.

[25] Peris, M., Martull, S., Maki, A., Ohkawa, Y., Fukui, K. "Towards a simulation driven stereo vision system." Pattern Recognition (ICPR), 2012 21st International Conference on. IEEE, 2012.

[26] Christian Richardt, Douglas Orr, Ian Davies, Antonio Criminisi, Neil A. Dodgson. "Real-time Spatiotemporal Stereo Matching Using the Dual-Cross-Bilateral Grid" European Conference on Computer Vision, 2010.

[27] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision". Proceedings of Imaging Understanding Workshop, pages 121-130, 1981.

[28] Liu, Ce, Jenny Yuen, and Antonio Torralba. "Sift flow: Dense correspondence across scenes and its applications." IEEE transactions on pattern analysis and machine intelligence 33.5 pp. 978-994,2011.

[29] Felzenszwalb, Pedro F., and Daniel P. Huttenlocher. "Efficient belief propagation for early vision." International journal of computer vision 70.1 (2006): 41-54.

[30] Yang, Qingxiong, et al. "Stereo matching with color-weighted correlation, hierarchical belief propagation, and occlusion handling." IEEE Transactions on Pattern Analysis and Machine Intelligence 31.3 (2009): 492-504.