# Temporal Consistent Stereo Disparity Search Using Belief Propagation with Adaptive Weight

***Abstract*—XXX 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. 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 accurate information becomes more and more urgent according to the prospered development of the autonomous car, 3D interaction application and augmented reality (AR). There are many 3D capture techniques are catoried into active methods [7][8] and passive methods. Among those, stereo matching method is the most potential for mobile or wearable devices because of 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 characteristic, 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 previous frame. The original disparity map can be refined with the time-consistency attribute. Therefore, the disparity map with better performance can be provided by the proposed algorithm. The two main contributions of this paper are summarized as follows:

1. A temporal term with adaptive weighted based on similarity is fused into a conventional energy function.

2. The proposed method is verified in stereo video sequences and outperforms others.

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 consistency issue of disparity estimation has attracted more attention recently. It assumes that 3D information of following frames must have highly correlated. 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 robustness. A 3D bilateral volume for filtering is proposed to enhance temporal consistency [6] which achieved great performance. It assumed that the corresponding points in previous frames are located exactly at the same position.

The assumption above is not suitable for dynamic scenes since the proper corresponding points are not always in the same location. Optical flow algorithms such as [26] and [27] are induced to find proper correspondent points and adopted by [24] and [19], respectively. However, both previous and upcoming frames are required in this method, which is not practical for our purpose for real-time application. In this paper, we focus on methods which use information only from precedent frames for depth estimation.

Window-based temporal consistent method [6] aggregates supporting pixel within adjacent frames. To provide better performance, global methods are used to preserve temporal consistent. Energy function based methods adopt consistent cues into data term or smooth term. Larsen et.al. [22] proposed the algorithm with pixel-wise similarity measurement. More complex components are considered such as mesh [18] and hyper-planes [21]. Lv et.al. [23] estimated the scene geometry after segment-based processing. [24] uses temporal discontinuity to scale the dataterm. [19] using binary temporal propagation in the proposed energy function.

## III. Proposed method

### A. Algorithm Overview

Since objects in the video move smoothly, 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. The framework of our algorithm ins shown in Fig.1. To estimate the depth map $D_t$ at time $t$ We first compute the optical flow of two consecutive frames $I_t$ and $I_{t-1}$ and XXX. We then pass these pixel pairs through similarity function to get the likelihood. After we get the reference pair and likelihood, we combined these information with previous depth map $D_t$ into BP energy minimization and calculate the energy cost. Finally, we used these altered energy cost into disparity determination, where we apply winner-take-all to estimate the final depth map. In addition, this depth map will be passed into the calculation of next frame.

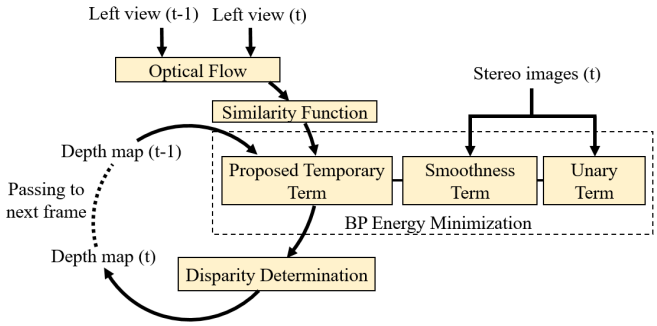


Fig. 1: The framework of our proposed method.

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 $E_d^t(p)$ is the unary term for pixel $p$ for label $d$ 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, which keeps the temporal consistency.

### *B. Detail of the Proposed Algorithm*

We focus on the temporal consistent term and the model to combine it 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 unary term, $E(d_n)$, 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. $\gamma_r$ and $\gamma_g$ are parameters in adaptive support weight controlling the sensitivity of each difference. $e(q,\overline{q}_d)$ can be any desired pixel wise matching cost. In our implementation, the pixel-wise matching cost is followed by 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 $\beta$ is the ratio parameter to fuse two matching cost, $I_c(q)$ is the AD cost and $\nabla I_{Census}(q)$ is the Census cost between pixel $p$ and pixel $q$.

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

is the smoothing term, which is also called pairwise term. It maintains the smoothness in the disparity map. We adopt the truncated linear model in [28]. We set our weighting, $W_f$, according to the color difference between adjacent pixels with a similar approach and parameter in [29]. $E_t(d,\overline{p}^{t-1})$ is the temporal consistent term maintaining the smoothness in adjacent frames. This term can be defined as

$$E_t(d,\overline{p}^{t-1}) = W_f(p_t,\overline{p}^{t-1}) \cdot min(|l_p - d_{t-1}|, T_f). \quad (6)$$

$W_f(p_t,p_{t-1})$ is the proposed adaptive weighting for the temporal inconsistent. We add a truncated factor, $T_f$, for robustness just like [19]. $\overline{p}^{t-1}$ is the corresponding patch in frame $t-1$. The proposed method adjusts the weight of previous frame by similarity with previous corresponding pixels estimated by Lucas and Kanade optical flow method [16]. The measured similarity weights the temporal incontinuity for temporal consistent term is defined as

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

where $\alpha$ 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 unlikelihood 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 more temporal smooth will be set into energy function. Less similarity with corresponding pixels may indicate mismatching or occlusion cases making weights smaller to avoid error propagated to next frame. Similarity measurement can be any methods. Here we apply range weighting in Adaptive Support-weight (ASW), which is represented as

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

$w_q(p_{t-1},q_{t-1})$ is the weighting 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 $\gamma_c$ is sensitivity factor for intensity difference. As a result, $d_t$, is acquired from proposed cost function via Winner-Takes-All manner as

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

## IV. Experimental Result

We evaluate our methods by testing on synthetic stereo video sequences which are New Tsukuba [25] and five different scenes from [**?**]. Similarity window size $N_t$ is $5\times5$. For the data term we adopted $7 \times 7$ Census window and $\beta$ is 0.25. $\{\gamma_r,\gamma_g\} = \{5, 17.5\}$. Size of ASW is $13 \times 13$

Fig. 2: Error rates of different algorithms. (a) Frame 1-30 and (b) Frame 71-100.

and $\gamma_d$ is 10. For the smoothness term we set parameters as $\{T, \lambda\} = \{\frac{L}{8}, 2\}$ where $L$ is the disparity range. For our purposed method of temporary term, we set $\gamma$ to 0.001, and truncation $T_f$ to 24.

### 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] is the static scene with a moving stereo camera. To evaluate our method on different condition, we chose two parts of the dataset, which are frame 1-30 and frame 70-100. From frame 1-30, the camera rotates counter-clockwise seen from the top, and the movement is relatively small. From frame 70-100, the camera approaches the statue in the center, and the camera movement is larger than that in 1-30. We evaluate the performance of our method by comparing the error rate of disparity map. We compare four methods here including ECCV 2010 [6], original BP without previous frame cue, APSIPA 2015 [24], and our purposed method. Note that all methods are implemented with same matching cost, weighted Census and AD costs. Last three methods are implemented based on the same BP-based optimization with different condition about temporal term: the original BP (BP Only) does not include the temporal term , APSIPA 2015 [24] alters the data term directly, and our purposed method uses adaptive weighting on temporary term. The result is shown in Table I. The respective error of each frames is shown in Fig.3 and Fig.4.

TABLE I: Comparison of New Tsukuba from frame 1-30 and frame 71-100.

| Methods | BP Only | ECCV 2010 [6] | APSIPA 2015 [24] | Proposed |
|---|---|---|---|---|
| 1st-to-30th | 10.48 | 18.2 | 10.45 | 10.09 |
| 71th-to-100th | 15.4 | 22.8 | 14.9 | 12.7 |

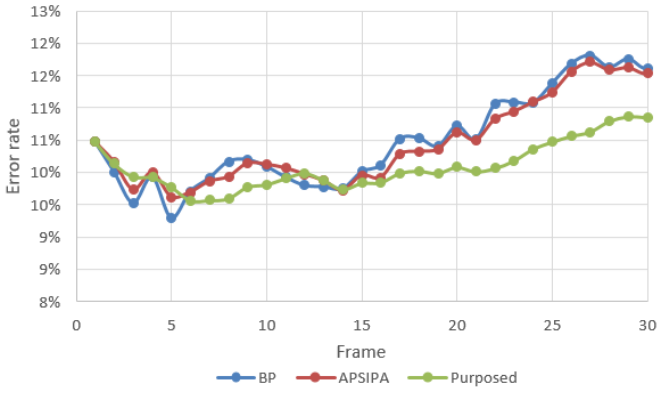


Fig. 3: Error comparison on 1st to 30th frame

The results of the disparity map are shown in Fig.7. We also highlight the improvement on the sculpture in Fig. 5.

From Table I and Fig.3, the overall correctness of the disparity map is better than ECCV 2010[6] , BP, APSIPA

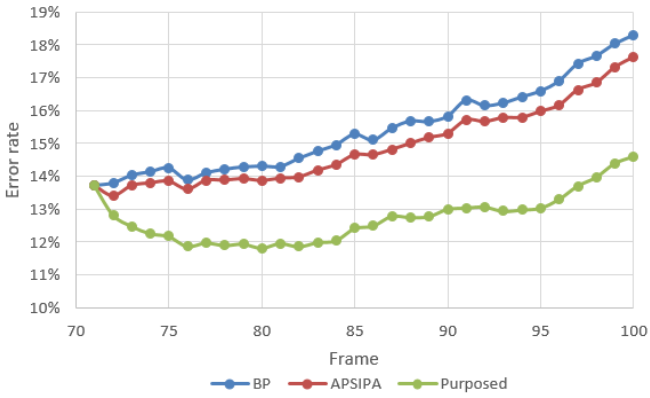


Fig. 4: Error comparison on 71st to 100th frame

2015[24]. In the first 10 frames, the movement is relatively small compared to the following sequence. We can also see that we get more improvement after 15th frame, where the camera begin to move faster. From Fig.4, which is Frame 71-100, our purposed method can achieve $14.8\%$ error reduction compared to [24].

From above, we can observe that the improvement room is related to the movement of the camera. If the movement is larger, such as frame 15-30 and frame 71-100, our method can improve the quality compared to smaller movement in frame 1-10. In summary, we ensure our method can improve the quality of disparity map and reduce error pixels. The improvement will be more significant with faster camera movement.

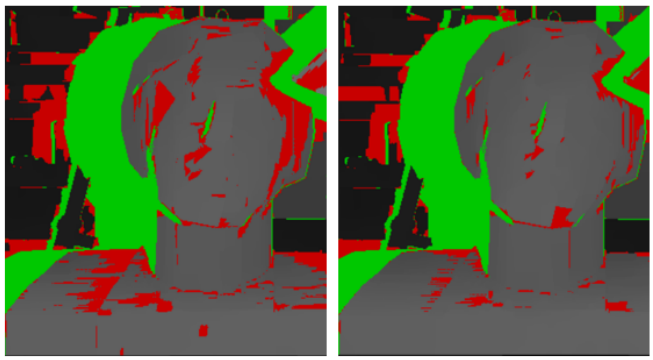

Fig. 5: Error comparison of sculpture in frame 100, green indicates occlusion, red indicates error pixel.(left: BP only, right: purposed)

### B. *Result II (DCB Grid Dataset)*

We use a synthetic dataset with five stereo sequences and ground truth provided by [6]. The result is presented in Fig.8 and Table II. These datasets are Book, Street, Tanks, Temple, and Tunnel. Note that we used three different $\alpha$ setting in DCB data to further discuss the characteristics of our method.

For dataset Book, Street and Temple, our method with $\alpha$=0.08 can achieve the lowest error rate. For Tank dataset, our result will be superior with $\alpha$=0.02. We choose Tanks and

Temple theses two datasets to discuss how parameters in our method affect the performance.

TABLE II: Comparison results under dataset [6].

| | 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 | 9.08 |
| BP Only | 5.19 | 13.84 | 3.41 | 9.43 | 1.35 | 7.97 |
| APSIPA 2015 [24] | 5.20 | 12.33 | 3.36 | 7.89 | 1.14 | 5.98 |
| Proposed(0.08) | 4.82 | 10.21 | 4.09 | 6.38 | 1.84 | 5.47 |
| Proposed(0.04) | 4.89 | 11.20 | 3.48 | 7.11 | 1.19 | 5.57 |
| Proposed(0.02) | 5.03 | 12.11 | 3.37 | 8.21 | 1.19 | 5.98 |

$\alpha$ represents the balance of our temporary term. The energy cost of disparity candidate far from the previous reference disparity value will be punished accordingly to the setting of $\alpha$. However, the approaching camera creates more occlusion region than others. Thus, the temporary term from previous cue will be too large compared to BP matching term. We suggest that using our method on faster take-in scenery should set $\alpha$ to a lower value. The comparison of different $\alpha$ can be seen in Fig.6. In fact, our error rate in frame 25 can reduced from 5.06% to 3.36% with $\alpha = 0.02$ instead of $\alpha = 0.08$ for the best average accuracy.

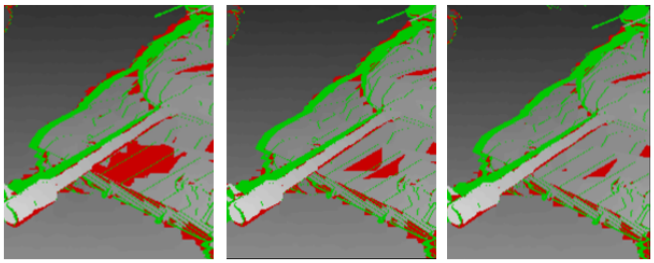

Fig. 6: Error comparison of different $\alpha$ setting in Tanks in frame 25th (from left to right: $\alpha$=0.08,$\alpha$=0.04,$\alpha$=0.02)

For the Temple dataset, the result shows that our method can deal with repeating texture. The costs in energy function of candidate depth is closed in repetitive texture, causing the difference of candidate to be smaller. Instead of weighting data term directly, a adaptive weight temporal consistent term is added. This make the label with strong representative data term is not reduced by similar corresponding pixels. In APSPIA 2015[24] method, the previous depth information is passed into current costs by multiplying a Gaussian weight. However, using multiplication will deteriorate the relativity of costs, causing the disparity determination will choose the wrong result. In comparison, a adaptive weighted temporal cue is adopted in our method, which can preserve the relativity from different candidate depth. As shonwn in Fig.8, the improvement of our method is better in four cases and achieve the best in average in DCB dataset. As for approaching camera scene, in Tanks dataset, setting $\alpha$ to a lower value can also achieve a compatible 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 $\alpha$ should not be a fix value for every scenery. We are currently testing how $\alpha$ 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 proposes an additive adaptive weight temporal term to provide temporal consistency. The proposed method can preserve the representative data term by the temporal term instead of weighting the data term directly. Our method can achieve top performance in two different synthetic stereo sequences. In New Tsukuba, the proposed method provides better improvement with a faster movement of camera. In DCB Grid dataset, our method can obtain best performance in most scenes. We also analyze the the characteristics of our method in different scenery and parameter setting.

## 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

[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).

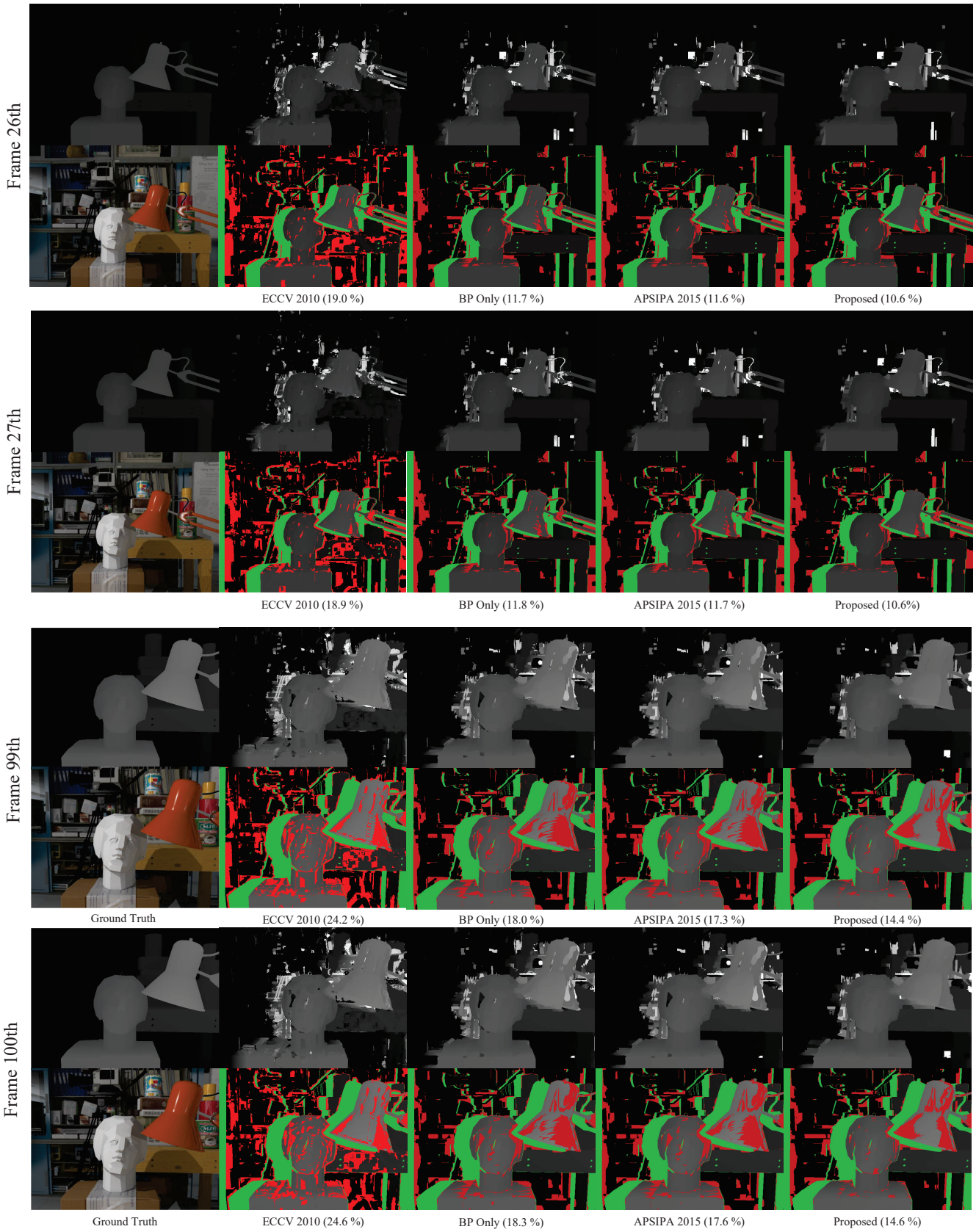


Fig. 7: Comparison figures in New Tsukuba dataset frame 71th to 100th.

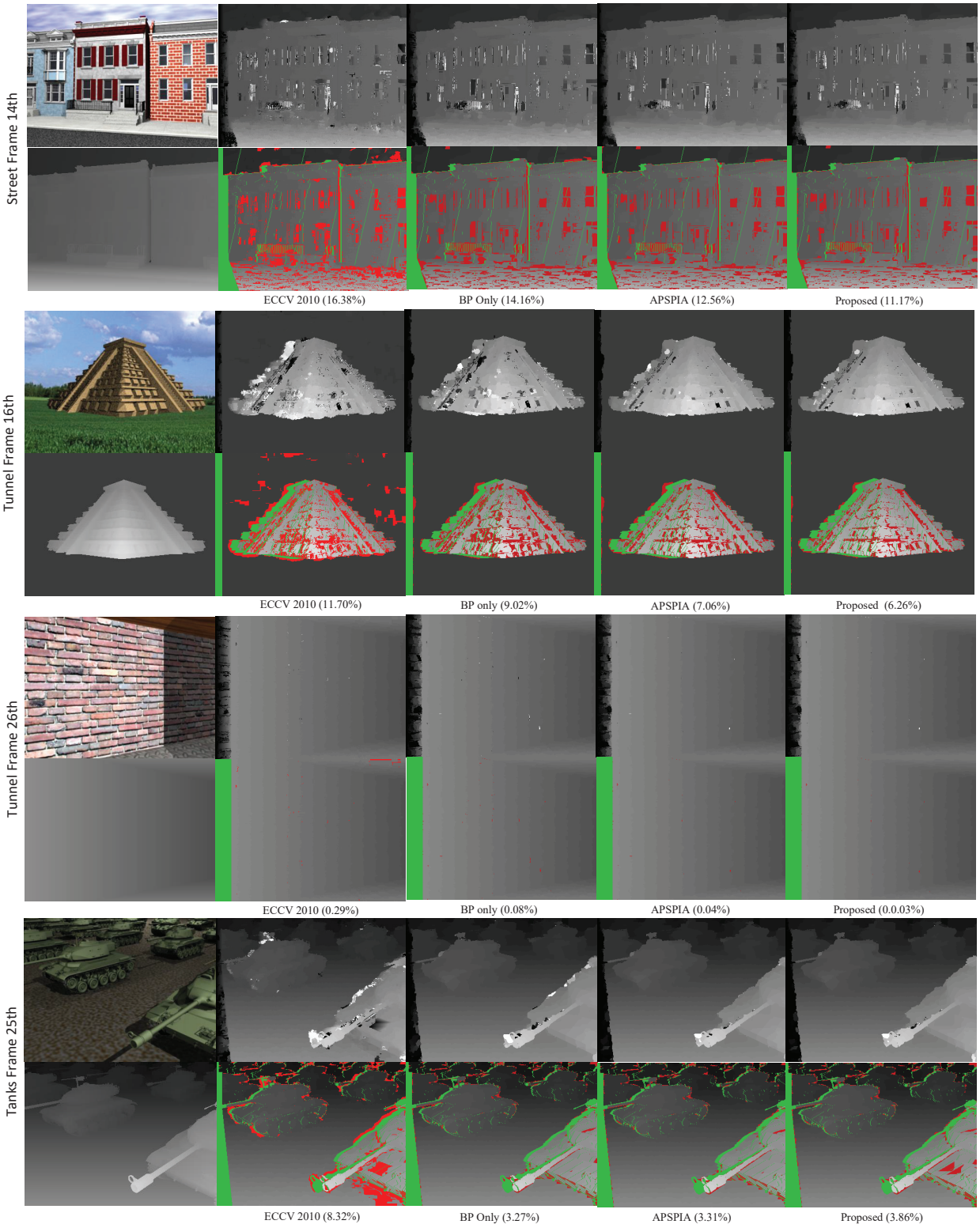


Fig. 8: Comparison figures.

[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] 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.

[27] 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.

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

[29] 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.