VIDEO STEREO MATCHING WITH TEMPORALLY CONSISTENT BELIEF PROPAGATION

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

The belief propagation (BP) technique is successful in image stereo matching problem. However, when we consider stereo matching for videos, directly applying the BP algorithm frame by frame results in unsatisfactory temporally inconsistent disparity maps. In this paper, we present the temporally consistent belief propagation for video stereo matching. We introduce a temporal term in traditional BP objective function and propose an adaptive weighting scheme to account for this temporal term. We show that the proposed algorithm performs favorably against previous methods in the stereo video datasets. Furthermore, the proposed method can solve problems induced by previous methods like error propagation from previously occluded regions.

Index Terms— Temporal Consistency, Stereo Matching, Disparity Estimation.

1. INTRODUCTION

The requirement of accurate depth information becomes more important in light of the prospered development of autonomous cars, 3D interaction and augmented reality. In order to get the disparity map of the environment, many complex algorithms have been proposed to improve the performance, such as [1–4]. These algorithms focus on improving disparity image per image without taking temporal information into consideration. Although temporal propagation is mentioned in PatchMatch Stereo[4], it represents the constraint information in the same input image during iterative procedure rather than between different time frame images.

The account of research on temporal consistency in disparity search is increasing. Most of them are using binary weights for the temporal term, which is not robust. Furthermore, previous and upcoming frames are both used to enhance the quality, making real-time application impractical. The original disparity map can be refined with the time-consistency attribute. The disparity map with better performance can be provided by the proposed algorithm with only one previous frame. The three main contributions of this paper are summarized as follows. First, we fuse a temporal term with adaptive weighting based on similarity into a conventional energy function. Second, the temporally consistent disparity search algorithm is required only one previous frame. Third, the proposed method is verified in stereo video sequences and outperforms others.

The rest of this paper is organized as follows: In Section 3 the proposed method is introduced. The experimental results and discussion are shown in Section 4. Finally, the paper is concluded in Section 5.

2. RELATED WORKS

There are two key techniques involved in temporal consistency stereo matching algorithms. First, the proper corresponding points and frames selection for the process. Second, the proper method to fuse the temporal term.

Correspnding points and frames selection: Pham *et.al.* [5] proposed a algorithm using spatio-temporal cues. Vretps *et.al.* [6] 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 by Chrisyian *et.al.*[7]. Above methods assume that the corresponding points in previous frames are located exactly at the same position. This is not suitable for dynamic scenes since the proper corresponding points are not always at the same location. Optical flow algorithms such as the Kanade-Lukas-Tomashi (KLT) method [8] and SIFT Flow [9] are introduced to find proper correspondents points and adopted by [10] and [11], respectively. However, both previous and upcoming frames are required in this method, which is not practical for real-time application. In this paper, we focus on methods which use information only from preceding frames for depth estimation.

Methods to fuse the temporal term: The window-based temporal consistency method in [7] aggregates supporting pixel within adjacent frames. Global methods are used to provide better performance. Larsen *et.al.* [12] proposed an algorithm with pixel-wise similarity measurement. More complex components are considered such as mesh [13] and hyperplanes [14]. Lv *et.al.*[15] estimated the scene geometry af-

[*] Joint first author

978-1-5386-1737-3/18/$31.00 ©2018 IEEE

ter segment-based processing. Baek *et.al.*[10] uses temporal discontinuity to scale the data term(TDT). This method concentrates at properly fusing the temporal term into an energy function. However, the method from [10] is easily affected by incorrect corresponding points. To solve this problem, we use an additive temporal cue with adaptive weights in the proposed algorithm.

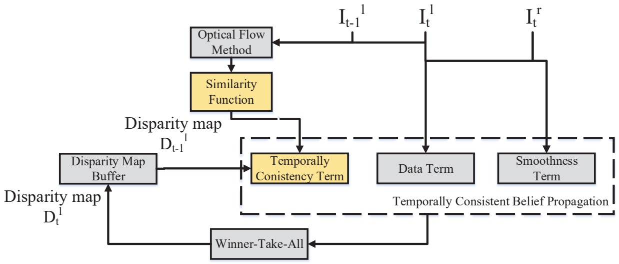


Fig. 1: The framework of our proposed method.

3. PROPOSED METHOD

3.1. Overview

We show the framework of our algorithm in Fig. 1. We use the left and right images at time t, I_t^l and I_t^r, the disparity map and the left image at time $t-1$, D_{t-1}^l and I_{t-1}^l as inputs to estimate D_t^l, the left disparity map at time t. Here, I_t^l and I_t^r are used for the data term as well as the smoothness term. I_{t-1}^l and I_t^l are used for estimating corresponding points in the previous frame by Kanade-Lukas-Tomashi (KLT) method [8]. These two images are also used to provide corresponding patches in respective frames. D_{t-1}^l indicates the temporal discontinuity within the disparity value hypothesis. The proposed energy function consists of the data term, the smoothness term and the proposed temporal term. The final proposed energy function is represented by

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

3.2. Detail of the proposed algorithm

We focus on the temporal consistency term and the model to combine it into the final energy function. The data term $E_d^t(p)$ is generated by using the adaptive support-weight (ASW) method [16]. For every possible disparity d_n, the data term, $E_d(d_n)$, is represented as

$$E_d(d_n) = \frac{\sum_{q\in\Omega_p,\overline{q}_d\in\Omega_{\overline{p}_d}} w(p,q)w(\overline{p}_d,\overline{q}_d)e(q,\overline{q}_d)}{\sum_{q\in\Omega_p,\overline{q}_d\in\Omega_{\overline{p}_d}} w(p,q)w(\overline{p}_d,\overline{q}_d)}, \quad (2)$$

where $w(p,q)$ accounts for range weight and spatial weight, as defined in ASW [16] and $e(q,\overline{q}_d)$ can be any pixel-wise matching cost. In our implementation, we follow the setting in [17] as

$$e(q,\overline{q}_d) = \beta\cdot(|I(q)-I(\overline{q}_d)|)+(1-\beta)\cdot(|I_C(q)-I_C(\overline{q}_d)|), \quad (3)$$

where β is the ratio parameter to fuse two matching costs, one is the AD cost and the other is the Census cost, I_C.

The smoothing term, which is also called the pairwise term, is defined as

$$E_s(p,d) = W_s\cdot\lambda\cdot\min(|l_p-d|,T). \quad (4)$$

We adopt the truncated linear model in [18]. We set the weight, W_s, according to the color difference between adjacent pixels in a similar approach and with similar parameters as in [19].

To preserve the strong data term, the temporal consistency term is added into the energy function instead of scaling the data term as [10]. The weight should be adaptive to increase the robustness. This term can be defined as

$$E_t(d,D_{t-1}^l(p)) = W_f(p_t,I_{t-1}^l)\cdot\min(|l_p-D_{t-1}^l|,T_f), \quad (5)$$

where $W_f(p_t,p_{t-1})$ are the proposed adaptive weights based on the similarity between patches in the present and the previous frame. D_{t-1} is the disparity map of the previous frame. We apply a truncated factor, T_f, to increase robustness as was shown to be helpful in [11]. $I_t^l(p)$ is the processed patch in the present frame, and $I_{t-1}^l(p')$ is the corresponding patch in frame $t-1$. The proposed method adjusts the weights of previous frame by evaluating the similarity to previous corresponding pixels estimated by Lucas and Kanade optical flow method [8]. The measured similarity weights of the temporal discontinuity for the temporal consistency term is defined as

$$W_f(p_t,p_{t-1}) = \frac{\alpha}{2^{S(p_t,p_{t-1})}}, \quad (6)$$

where α is the parameter that should balance the cost from stereo matching and the previous frame cue. $S(p_t,p_{t-1})$ is the similarity score representing the unlikelihood of the reference patch $I_{t-1}^l(p')$ to be the same object as $I_t^l(p)$. A 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 the disparity value is forced to be temporally smooth. Less similarity with corresponding pixels may indicate mismatching or occlusion cases making weights smaller to avoid the error being propagated to the next frame. We apply range weighting in ASW as our similarity measurement, which is represented as

$$S(p_t,p_{t-1}) = \gamma\cdot\frac{\sum_{q_t\in N_p^t} w_q(p_{t-1},q_{t-1})|I(q_t)-I(q_{t-1})|}{\sum_{q_t\in N_t} w_q(p_{t-1},q_{t-1})}. \quad (7)$$

N_p^t represents the local window for the similarity measurement between this patch and the corresponding patch at time

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

$$w_q(p_{t-1}, q_{t-1}) = exp(\frac{-|I_t^l(q) - I_{t-1}^l(q)|}{\gamma_c}). \qquad (8)$$

where γ_c is the sensitivity factor for intensity difference. As a result, d_t, is acquired from proposed cost function via Winner-Takes-All manner.

4. EXPERIMENTAL RESULT

We evaluate our methods by testing on synthetic stereo video sequences which are the new Tsukuba dataset[20] and five different scenes from the DCB Grid dataset [7]. The size of the similarity window is N_t is 5×5. For the data term we adopted a 7×7 Census window and β is set to 0.25. Furthermore, we set $\{\gamma_r, \gamma_g\} = \{5, 17.5\}$. The size of the ASW is 13×13 and γ_d is 10. For the smoothness term we set the parameters to $\{T, \lambda\} = \{\frac{L}{8}, 2\}$ where L is the disparity range. For our proposed method for the temporal term, we set γ to 0.001, the truncated factor T_f is 24.

4.1. New Tsukuba Dataset

We apply our method to the New Tsukubsa dataset, which is a series of synthetic images with ground truth. The new Tsukuba dataset [20] comprises a static scene with a moving stereo camera. To evaluate our method on different conditions, we choose two parts of the dataset, which are frames 1-30 and frames 70-100. In frames 1-30, the camera rotates counter-clockwise seen from the top, and the movement is relatively small. In frames 70-100, the camera approaches the statue in the center, and the camera movement is larger than that in frame 1-30.

We evaluate the performance of our method by comparing the error rates of the disparity maps. We compare four methods including DCB [7], the original BP algorithm without previous frame cue, TDT [10], and our proposed method. Note that all methods are implemented with the same matching costs, weighted Census and AD costs. The last three methods are implemented based on the same BP-based optimization with different condition about temporal term. The original BP algorithm[19] does not include the temporal term , TDT [10] alters the data term directly, and our proposed method uses adaptive weighting on the temporal consistency term. The result is shown in Table 1. The respective error of each frame is shown in Fig.2.

The results of the disparity map are shown in Fig.6. We also highlight the improvement on the sculpture in Fig. 3.

As Table 1 and Fig. 2(a) show, the overall correctness of the disparity map is better than DCB[7], BP[19] and TDT[10]. In the first 10 frames, the movement is relatively small compared to the following sequence. We can see that the proposed method yields a higher improvement after the 15th frame, where the camera begins to move faster. From Fig.2(b), representing frames 71-100, our proposed method can achieve 14.8% error reduction compared to [10].

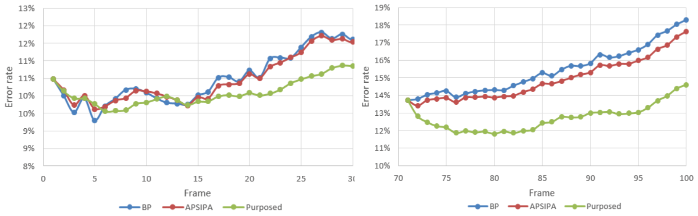
From the above, we can tell that the improvement is related to the movement of the camera. If the movement is larger, such as in frames 15-30 and frames 71-100, our method can provide a higher improvement as compared to the slower movement in frames 1-10. In summary, the improvement becomes more pronounced with a larger camera movement.

4.2. DCB Grid Dataset

We use a synthetic dataset with five stereo sequences and ground truth provided by the DCB Grid Dataset[7]. The result

Table 1: Error rates (%) of algorithms for New Tsukuba dataset from frame 1-30 and frame 71-100.

Frame	BP[19]	DCB[7]	TDT[10]	Proposed
1-30	10.48	18.2	10.45	10.09
71-100	15.4	22.8	14.9	12.7



(a) Frame 1-30

(b) Frame 71-100

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

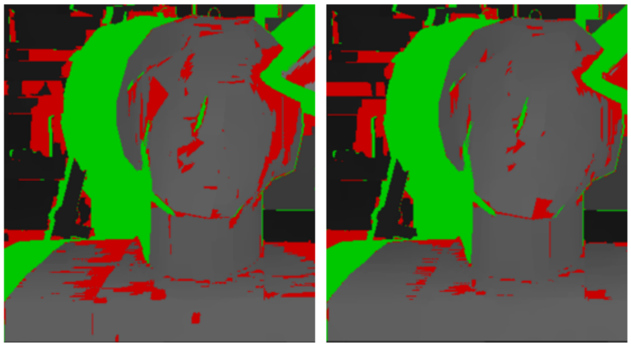


Fig. 3: Error comparison of BP (left) and the proposed method (right). Green pixels indicate occlusion regions and red pixels are error regions. With temporal information, the proposed method outperforms in the statue region.

Table 2: Comparison of results in dataset [7].

	Book	Street	Tanks	Temple	Tunnel	Avg.
DCB [7]	8.61	16.39	7.15	11.98	2.05	9.08
BP [19]	5.19	13.84	3.41	9.43	1.35	7.97
TDT [10]	5.20	12.33	3.36	7.89	1.14	5.98
Ours (0.08)	4.82	10.21	4.09	6.38	1.84	5.47
Ours (0.04)	4.89	11.20	3.48	7.11	1.19	5.57
Ours (0.02)	5.03	12.11	3.37	8.21	1.19	5.98

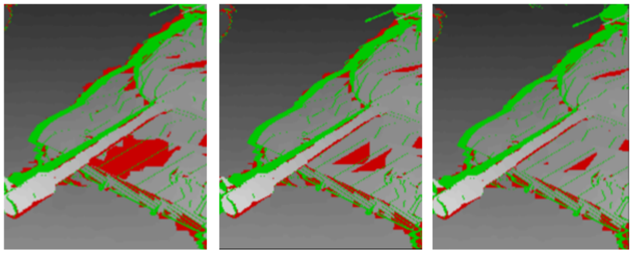


Fig. 4: Error comparison of different α setting in *Tanks* in frame 25. (from left to right: $\alpha = 0.08, 0.04, 0.02$)

is presented in Fig.5 and Table 2. Note that we used three different α settings in the DCB data to further discuss the characteristics of our method.

For the Book, Street and Temple sequences, our method can achieve the lowest error rate when $\alpha = 0.08$. For the Tank sequence, our results are superior for $\alpha = 0.02$. We choose Tanks and Temple to discuss how the parameters in our method affect the performance.

Here, α represents the balance of our temporal term. The energy cost of a disparity candidate far from the previous reference disparity value will be punished according to α. The approaching camera creates more occluded region than others. Thus, the temporal term from the previous cue will dominate the whole energy function and cause error propagation. We suggest that α should be set to a lower value when using our method on faster take-in scenery. The comparison for different α can be seen in Fig.4. In fact, the error rate in frame 25 is reduced from 5.06% to 3.36% when choosing $\alpha = 0.02$ instead of $\alpha = 0.08$, which is the best choice for average accuracy.

The costs in the energy function of candidate depth is close in repetitive textures, causing the difference of candidate to be smaller. Instead of weighting the data term directly, a adaptive weight temporal consistency term is added. This makes the label with strong representative data term is not reduced by similar corresponding pixels. In the DCB method[10] the previous depth information is passed into the current costs by multiplying a Gaussian weight. Using multiplication may deteriorate the relative size of the data term, causing the disparity determination to choose the wrong result. In comparison, an adaptive weighted temporal cue is adopted in our method, which can preserve the relation of different candidate disparity values. As shown in Fig.5, the proposed method provide smoothest result in the flat regions compared to other algorithms. In the Tanks sequence, in which the camera is approaching a static scene, setting α to a lower value can also yield a competitive result.

5. CONCLUSION

We propose an additive adaptive weight temporal term to provide temporal consistency. The proposed method can preserve the representative data term through the temporal term instead of weighting the data term directly. Our method achieves state-of-the-art performance in two different synthetic stereo datasets. In New Tsukuba, the proposed method provides a higher improvement with for faster camera movement. In the DCB Grid dataset, our method can obtain the highest performance in most scenes. We also analyze the characteristics of our method in different scenery and parameter settings. In future work, we will extend the algorithms with an adaptive selection based on the video content and the scene change.

References

[1] Eric T Psota, Jedrzej Kowalczuk, Mateusz Mittek, and Lance C Perez, “Map disparity estimation using hidden markov trees,” in *IEEE International Conference on Computer Vision (ICCV)*, 2015.

[2] Chi Zhang, Zhiwei Li, Yanhua Cheng, Rui Cai, Hongyang Chao, and Yong Rui, “Meshstereo: A global stereo model with mesh alignment regularization for view interpolation,” in *IEEE International Conference on Computer Vision (ICCV)*, 2015.

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

[4] Michael Bleyer, Christoph Rhemann, and Carsten Rother, “Patchmatch stereo-stereo matching with slanted support windows.,” in *British Machine Vision Conference (BMVC)*, 2011.

[5] Cuong Cao Pham, Vinh Dinh Nguyen, and Jae Wook Jeon, “Efficient spatio-temporal local stereo matching using information permeability filtering,” in *IEEE International Conference on Image Processing (ICIP)*, 2012.

[6] Nicholas Vretos and Petros Daras, “Temporal and color consistent disparity estimation in stereo videos,” in *IEEE International Conference on Image Processing (ICIP)*, 2014.

[7] Christian Richardt, Douglas Orr, Ian Davies, Antonio Criminisi, and Neil A Dodgson, “Real-time spatiotemporal stereo matching using the dual-cross-bilateral grid,” in *European Conference on Computer Vision (ECCV)*, 2010.

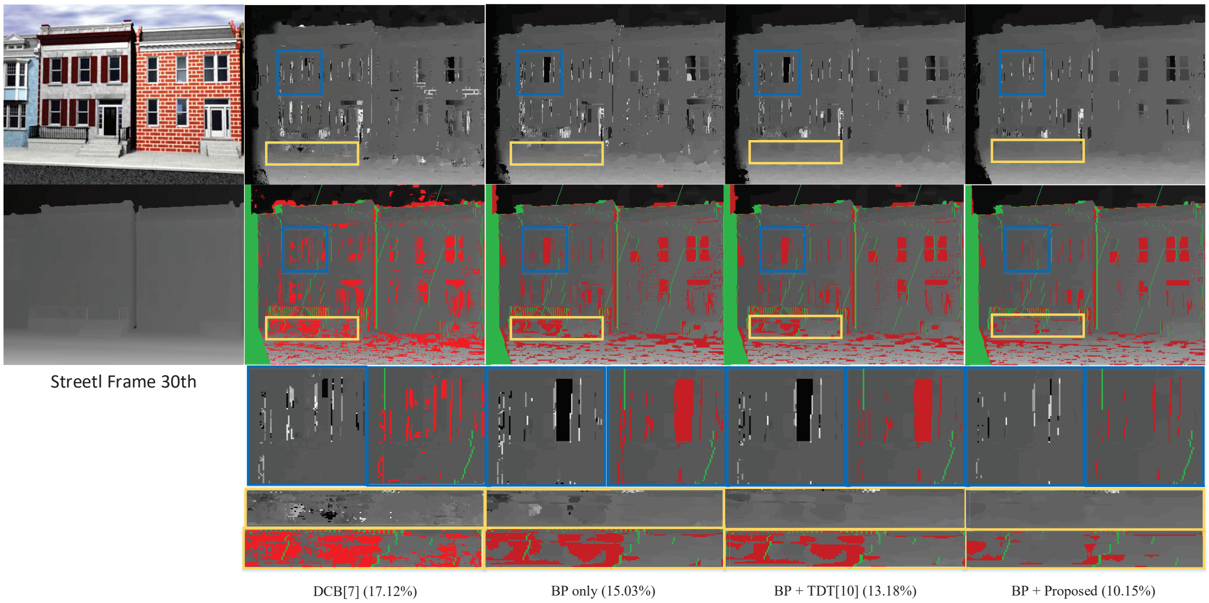


Fig. 5: Comparison figures of DCB Grid dataset where green pixels are occluded region and red pixels are error pixels with threshold as 1.0.

[8] Bruce D. Lucas and Takeo Kanade, "An iterative image registration technique with an application to stereo vision," in *International Joint Conference on Artificial Intelligence (IJCAI)*, 1981.

[9] Ce Liu, Jenny Yuen, and Antonio Torralba, "Sift flow: Dense correspondence across scenes and its applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 33, no. 5, pp. 978–994, 2011.

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

[11] Yong-Ho Shin and Kuk-Jin Yoon, "Spatiotemporal stereo matching with 3d disparity profiles.," in *British Machine Vision Conference (BMVC)*, 2015.

[12] E Scott Larsen, Philippos Mordohai, Marc Pollefeys, and Henry Fuchs, "Temporally consistent reconstruction from multiple video streams using enhanced belief propagation," in *IEEE International Conference on Computer Vision (ICCV)*, 2007.

[13] Zongji Wang, Xiaowu Chen, and Dongqing Zou, "Copy and paste: Temporally consistent stereoscopic video blending," *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 2017.

[14] Hiroki Nakano, Daisuke Sugimura, and Takayuki Hamamoto, "Disparity estimation in stereo videos using spatio-temporal disparity hyperplane models," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[15] Zhaoyang Lv, Chris Beall, Pablo F Alcantarilla, Fuxin Li, Zsolt Kira, and Frank Dellaert, "A continuous optimization approach for efficient and accurate scene flow," in *European Conference on Computer Vision (ECCV)*, 2016.

[16] Kuk-Jin Yoon and In So Kweon, "Adaptive support-weight approach for correspondence search," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 28, no. 4, pp. 650–656, 2006.

[17] Asmaa Hosni, Christoph Rhemann, Michael Bleyer, Carsten Rother, and Margrit Gelautz, "Fast cost-volume filtering for visual correspondence and beyond," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 35, no. 2, pp. 504–511, 2013.

[18] Pedro F Felzenszwalb and Daniel P Huttenlocher, "Efficient belief propagation for early vision," *International Journal of Computer Vision (IJCV)*, vol. 70, no. 1, pp. 41–54, 2006.

[19] Qingxiong Yang, Liang Wang, Ruigang Yang, Henrik Stewénius, and David Nistér, "Stereo matching with color-weighted correlation, hierarchical belief propagation, and occlusion handling," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 31, no. 3, pp. 492–504, 2009.

[20] Martin Peris, Sara Martull, Atsuto Maki, Yasuhiro Ohkawa, and Kazuhiro Fukui, "Towards a simulation driven stereo vision system," in *IEEE International Conference on Pattern Recognition (ICPR)*, 2012.

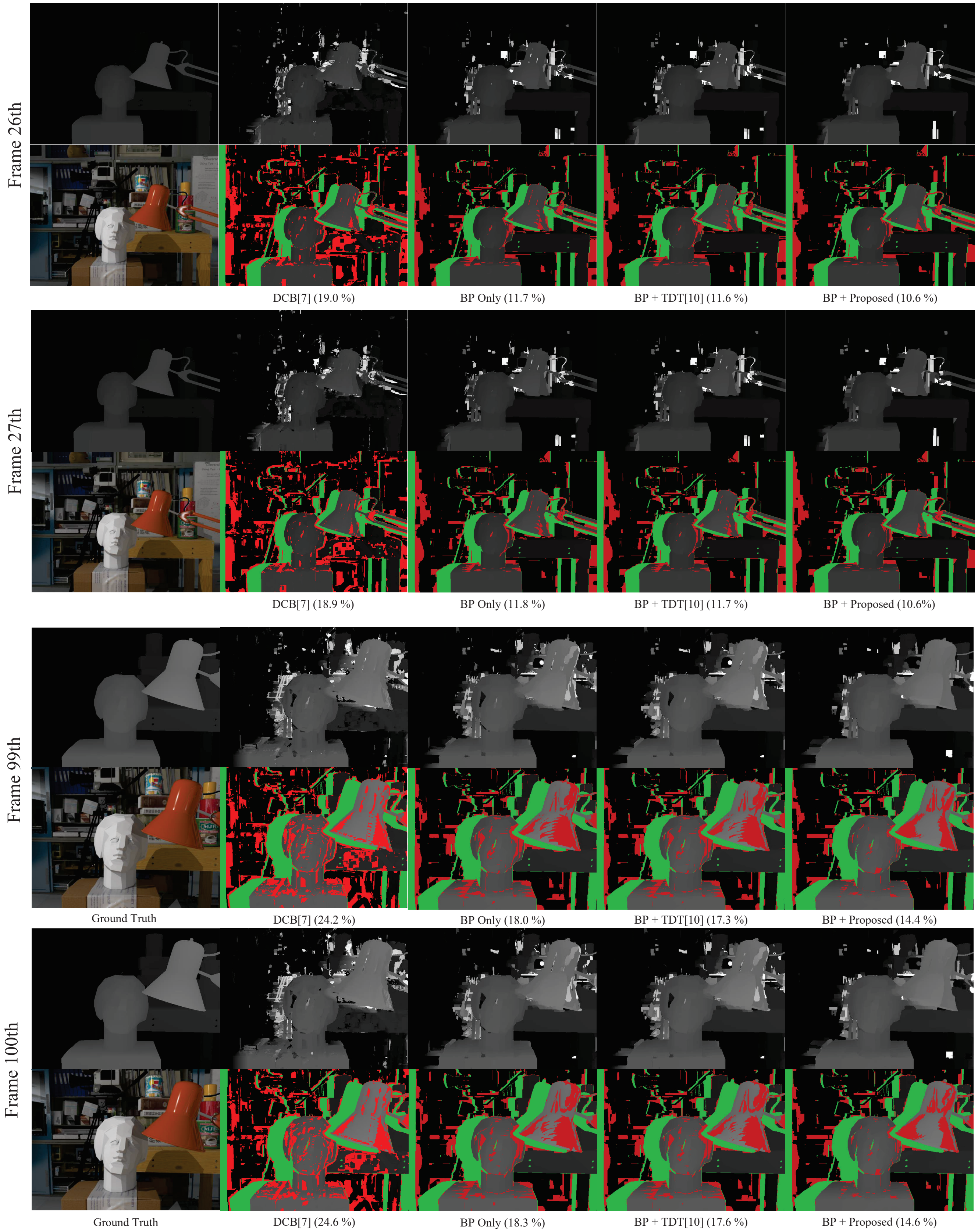


Fig. 6: Comparison figures in New Tsukuba dataset where green pixels are occluded region and red pixels are error pixels. Rows from top to down are frame 26, 27,99 and 100.