International Journal of Signal Processing, Image Processing and Pattern Recognition
Vol.7, No.5 (2014), pp.229-238
http://dx.doi.org/10.14257/ijsip.2014.7.5.20

Spatio-Temporal Consistency Enhancement for Disparity Sequence

Haixu Liu[1], Chenyu Liu[1], Yufang Tang[1], Haohui Sun[1] and Xueming Li[2]

[1]School of Information and Communication Engineering,
Beijing University of Posts and Telecommunications, Beijing, China
[2]Beijing Key Laboratory of Network System and Network Culture, Beijing, China
liuhaixu@bupt.edu.cn

Abstract

Disparity estimation for still images has attracted much interest and acquired many promising results. However, simply applying these methods to produce a disparity sequence may suffer from the undesirable flickering artifacts. These errors not only distinctly decrease the visible quality of the synthesized video, but also significantly reduce the coding efficiency of the disparity sequence. In this paper, a novel temporal consistency enhancement algorithm based on Guided Filter and Temporal Gradient (GFTG) is proposed. The flickering artifacts and noises are effectively removed and the edges of objects are well preserved. Both quantitative and qualitative evaluations show that the spatio-temporal consistency has been highly improved by utilizing our approach.

Keywords: *disparity estimation, disparity sequence, spatio-temporal consistency, view synthesis*

1. Introduction

Disparity estimation has been widely studied over the past few decades and it is still one of the most active research topics in the field of computer vision [1]. It is a key and fundamental technology applied for synthesizing the virtual viewpoints for the systems of next-generation broadcasting services, such as Three-Dimensional Television (3DTV), Multi-View Video (MVV) and Free-viewpoint Television (FTV), *etc.*

In the last few years, a deal of disparity estimation algorithms has been developed for generating accurate disparity maps from a stereo image pair. A detailed overview for such methods and the state-of-the-art approaches can be found in [2-4]. Meanwhile, the unified platform [5] for comparing the performances of different algorithms also provides a review of latest approaches.

Generally, the task of disparity estimation for still images has been well studied. A number of impressive schemes have been reported, such as Graph-cut [6], Belief-Propagation [7], Adaptive support-weight [8] and CostFilter [9], *etc.*, however, these frame-by-frame approaches are usually conducted without considering the temporal consistency of the disparity sequence. Therefore, simply implementing them to the video sequences often results in annoying flickering-artifacts between consecutive frames. These flickers will significantly reduce the subjective quality of the sequence. It could cause a serious problem due to the fact that human visual system is highly sensitive to high-frequency visual artifacts, especially when it is utilized for synthesizing the virtual view. Furthermore, temporal inconsistency will also drastically decrease the efficiency of inter-frame coding of the disparity sequence, since more bits have to be allocated to deal with the significant changes of the disparity values.

Accordingly, how to alleviate the temporal inconsistency and maintain the high-accuracy of the disparity sequence has become a crucial task.

In this paper, an efficient spatial-temporal consistency enhancement algorithm based on Guided Filter and Temporal Gradient (GFTG) is proposed for smoothing disparity sequences. The goal of this work is to address the problem of flickering-artifacts which is caused by the incoherent disparity maps. Particularly, the proposed approach has the capability of integrating any state-of-the-art image-based disparity estimation technologies and extending them to the spatial-temporal domain. We first treat the video as a space-time volume, and then we smooth the original disparity maps with the guided filter frame by frame. After that, an adaptive filter based on temporal gradient is applied to the volume for improving the consistency of the sequence. Experiments show that, compared with the existed adaptive support-weight approach and spatial-temporal method, the proposed algorithm considerably increases the temporal consistency of the disparity sequence.

The rest of this paper is organized as follows. Firstly, we briefly introduce the related works in Section 2. Then details about the proposed approach are described in Section 3. Experimental results and analysis are shown in Section 4. Finally, the conclusions and future work are presented in Section 5.

2. Related Work

Although making the disparity sequence temporally smooth is vital and necessary for many applications, relatively fewer studies have been explored in this field. Ref. [10] proposes an energy minimal function to improve temporal consistency, and it uses an iterated dynamic programming scheme to solve the energy function. In addition, ref. [11] develops a coherence function to represent the temporal consistency between consecutive disparity maps. The coherence function is calculated according to the motion probability between frames. However, when the objects or cameras move fast, these approaches may not work well.

Some methods also exploit optical flow to solve the problem. Ref. [12] proposes a weighted mode filtering and extends it to the temporal domain. In the approach, the temporal neighbors are determined by the optical flow method, and a patch-based reliability measure is applied to deal with errors of the estimated optical flow. Additionally, M. Bleyer, *et al.,* [13] use a median filter to smooth the disparity maps. In order to overcome the problem caused by large motion, they compute the optical flow for all the pixels. The downside of these methods is that they have to sacrifice the time efficiency, since the process of dense flow field estimation is usually time consuming. Moreover, the estimation of the flow field may also introduce new errors into the whole system.

Fortunately, Richardt, *et al.,* [14] report the promising results, they extend their Dual-Cross-Bilateral grid method to the time dimension and show the real-time performance by using a reformulation of the adaptive support weights algorithm [8].Moreover, they produce five synthetic videos with ground truth disparities, which make it possible to quantitatively assess the estimation results of a disparity sequence. Despite excellent results have been shown by utilizing their approach, we demonstrate that more accurate disparity video can be generated by using our method.

3. Technical Details

As we described in the sections before, disparity sequences generated by existed image-based stereo matching algorithms are implicated with the problem of flickering-artifacts. To address this issue, we present a novel video disparity estimation algorithm. The flowchart of the proposed approach is shown in Figure 1. In general, there are two key stages in the presented method. Firstly, the original disparity maps are smoothed by using the guided filter frame by frame. The noises and inconsistent pixels in spatial domain are eliminated during this stage. Afterwards, the whole disparity sequence is treated as a space-time volume, and an adaptive filter based on temporal gradient is employed to smooth the volume for enhancing the temporal consistency of the sequence. In the following sections, each stage will be described in detail.

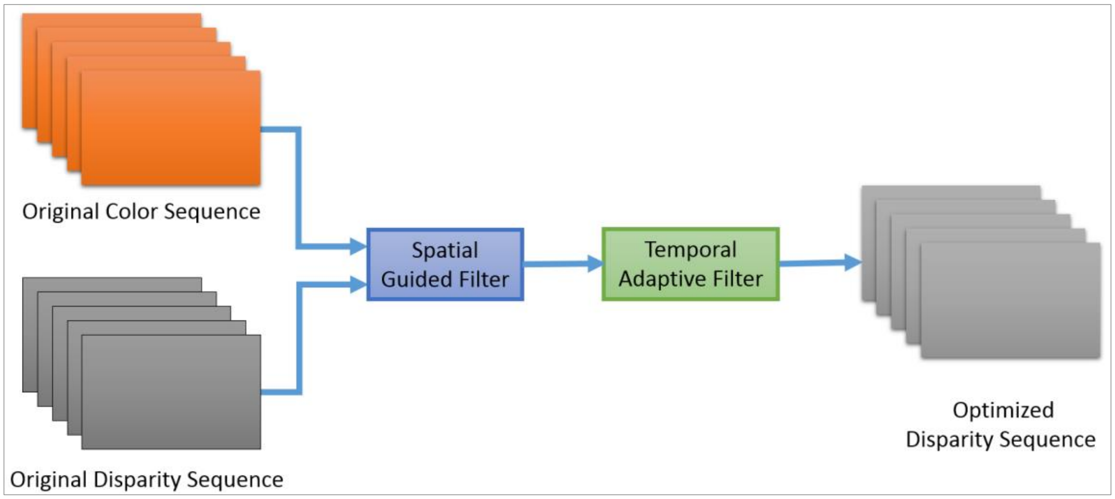


Figure 1. Flowchart of the Proposed Method

3.1. Image-based Disparity Map Refinement

We firstly seek to remove the spatial noises of the original disparity sequence. Considering that the disparity video and its corresponding color video should have similar image structure, the guided filter [15] is eventually adopted to smooth the disparity sequence frame by frame. The guided filter has shown well performance as an edge-preserving smoothing operator. Additionally, its computational complexity is independent of the filtering kernel size, which means it is much faster than other similar operators. In this case, we employ the original color image I as the guidance image, and utilize its corresponding original disparity map D as the guided image. Accordingly, the filtering result D' is defined as the linear weighted sum of all pixels in the original disparity map:

$$D'(m) = \sum_{n} W_{mn}(I) D(n) \tag{1}$$

where m and n are pixel indexes. The filter kernel W_{mn} is a function of guidance image$I$ and its weights can be expressed as:

$$W_{mn}(I) = \frac{1}{|\omega|^2} \sum_{k:(m,n)\in\omega_k} \left(\frac{(I(m)-\mu_k)\cdot(I(n)-\mu_k)}{\sigma_k^2 + \grave{o}} + 1\right) \tag{2}$$

here, $|\omega|$is the pixels number of$I$in a window $\omega_k$centered at pixel k, $\mu_k$and $\sigma_k^2$are the mean and variance of Iinω_k,respectively.ϵ is a threshold for smoothing.

One of the great advantages of the guided filter is that it can be rewritten as a fast and non-approximate linear-time algorithm, whose computation complexity is *O(N)*. Accordingly, the local linear model can be summarized as:

$$D'(m) = \frac{1}{|\omega|} \sum_{k:m\in\omega_k} (a_k I(m) + b_k) \tag{3}$$

$$a_k = (\Sigma_k + \grave{o}\mathrm{U})^{-1}\left(\frac{1}{|\omega|}\sum_{m\in\omega_k} I(m)D(m) - \mu_k \bar{D}(k)\right) \tag{4}$$

$$b_k = \bar{D}(k) - a_k^{\mathrm{T}}\mu_k \tag{5}$$

where Σ_k is the 3×3 covariance matrix of I in ω_k, U is a 3×3identity matrix, and $\bar{D}$ is the mean of D inω_k.

3.2. Spatio-temporally Consistent Disparity Map Estimation

In the previous work, we smoothed the original disparity sequence from the spatial domain. Next, we exploit an adaptive filter based on temporal gradient to reduce the transient errors in the temporal domain.

The filtering results are mainly decided by three factors: 1) the disparity map currently being processed, 2) the original color image corresponding to the current disparity map, and 3) the previous color maps.

To be specific, the filter can be expressed as:

$$D_{res}(p,i) = \frac{\omega_0 \cdot D'(p,i) + \sum_{l=1}^{N} \omega_g(p,i-l)\cdot\omega_d(p,i-l)\cdot D_{res}(p,i-l)}{\omega_0 + \sum_{l=1}^{N}\omega_g(p,i-l)\cdot\omega_d(p,i-l)} \tag{6}$$

where $D'(p,i)$and $D_{res}(p,i)$ represent the input disparity value and the result of filtering operation at pixel p in the ith frame, respectively. $\omega_0$is a constant to restrict the weight of the disparity map which is to be processed. Ndepicts the number of involved frames. The parametersω_g and $\omega_d$present the weight of previous frames from different aspects.

The computation of $\omega_g$is defined as:

$$\omega_g(p,i-l) = 1 - \frac{1}{3}\sum_{j=R,G,B}\frac{|\nabla_T(I^j(p,i-l))|}{I^j_{\max}(i) - I^j_{\min}(i-l)} \tag{7}$$

$$\nabla_T(I(p,i-l)) = I(p,i) - I(p,i-l) \tag{8}$$

here, $I$is the original color image, $\omega_g(p,i-l)$ describes the impact from temporal gradient difference between current ith frame and previous lth frame at pixel$p$. The smaller the difference, the higher the weight.

The definition of ω_d is shown as a negative exponential function:

$$\omega_d(p,i-l) = e^{-l} \tag{9}$$

Function ω_d indicates the correlation of distance between current frame and the previous one. Higher weight is calculated for the frame which is closer to the current one. As the increase of the distance between the two frames, the weight is reduced gradually.

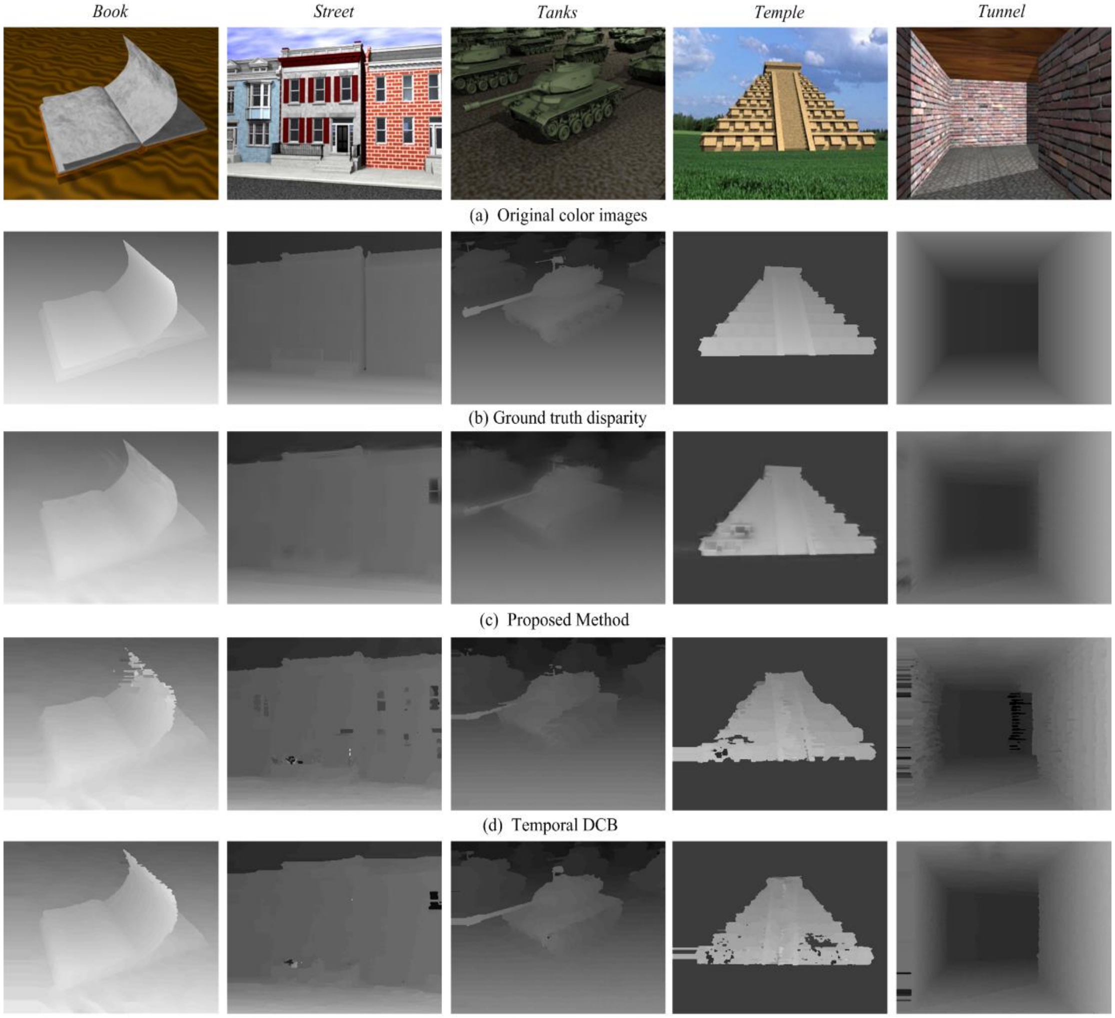
4. Experimental Results

To evaluate the proposed method, we implement the experiments on five public test sequences obtained from [14]. These sequences provide the ground-truth disparity maps which allow us to have a fair comparison. All of the images are 400×300 with 64 disparities. We compare the proposed approach with two state-of-the-art algorithms: the locally adaptive support-weight scheme [8] and the spatio-temporal method named as Temporal DCB [14]. In our experiments, we utilize the algorithm of [8] to obtain the original disparity sequence, however, note that our method can be applied as a post-processing of any depth estimation algorithm. It is also noted that we follow the implementation in [14] to generate the results of [8] and [14].The proposed method is evaluated using the same parameters for all images: $|\omega| = 28^2$,$\epsilon = 0.1^2$,$\omega_0 = 0.6$, and$N = 3$.

4.1. Subjective Evaluation

The subjective evaluation results are illustrated in Figure2. The original five color images (left view) along with their corresponding ground truth disparity maps are shown in Figure2 (a) and (b), respectively. The disparity images generated by our approach, Temporal DCB [14] and Yoon's method [8] are presented in Figure 2(c), (d) and (e), respectively. From the comparison, we could find out that the proposed algorithm can obtain more smooth disparities on the surface of objects while preserving the discontinuity property at the boundaries.

Compared with other methods, the proposed approach is able to generate much clearer boundaries (see the Book sequence). Moreover, the noises are eliminated well both in low texture areas (see the background of Temple sequence) and repetitive regions (see the wall of Tunnel sequence).



(a) Original color images

(b) Ground truth disparity

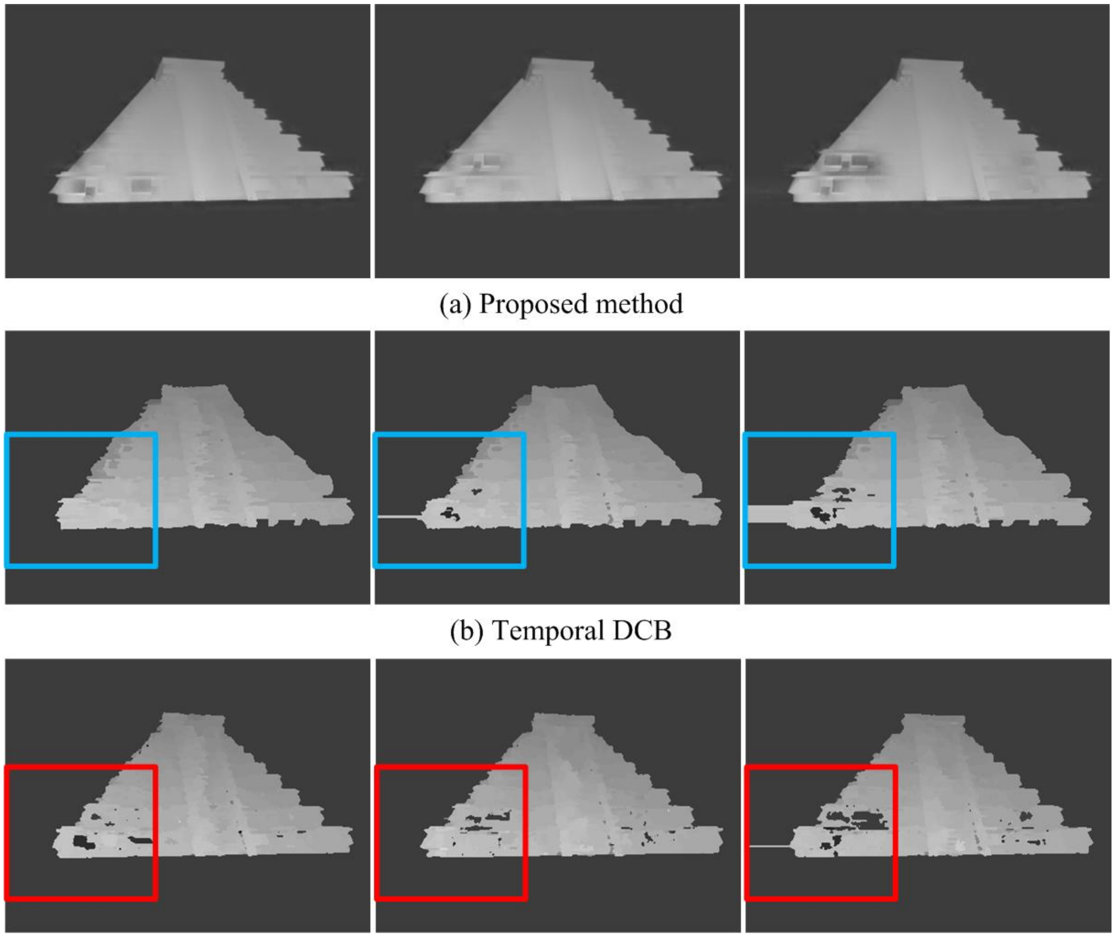
(c) Proposed Method

(d) Temporal DCB

(e) Yoon & Kweon

Figure 2. Subjective Comparison with Previous Methods

To further verify the effect of removing spatial-temporal noises, we show some consecutive frames of Temple sequence in Figure 3. From Figure 3(b) and (c), we can notice that the frames contain inconsistent disparity values in static regions, whereas the disparity values in Figure 3(a) are temporally consistent. The experimental results reveal that the proposed algorithm can guarantee more temporal smooth results.



(a) Proposed method

(b) Temporal DCB

(c) Yoon & Kweon

Figure 3. Disparity Results of Consecutive Frames for Temple Sequence

4.2. Objective Evaluation

In objective evaluation, all of the disparity results are compared with the ground truth pixel by pixel. The performances of different methods are mainly evaluated by the Picture Signal-to-Noise Ratio (PSNR) and the Structural SIMilarity (SSIM) index. The results are illustrated in Figure 4 and Figure 5. The best performance is produced by the proposed scheme, which can be significantly observed on all datasets. Note that the index values of Tunnel sequence changed more drastically than other sequences, this is due to its rich texture and fast switching scenarios， nevertheless, our method obtains satisfactory results both in qualitative and quantitative evaluations for this video.

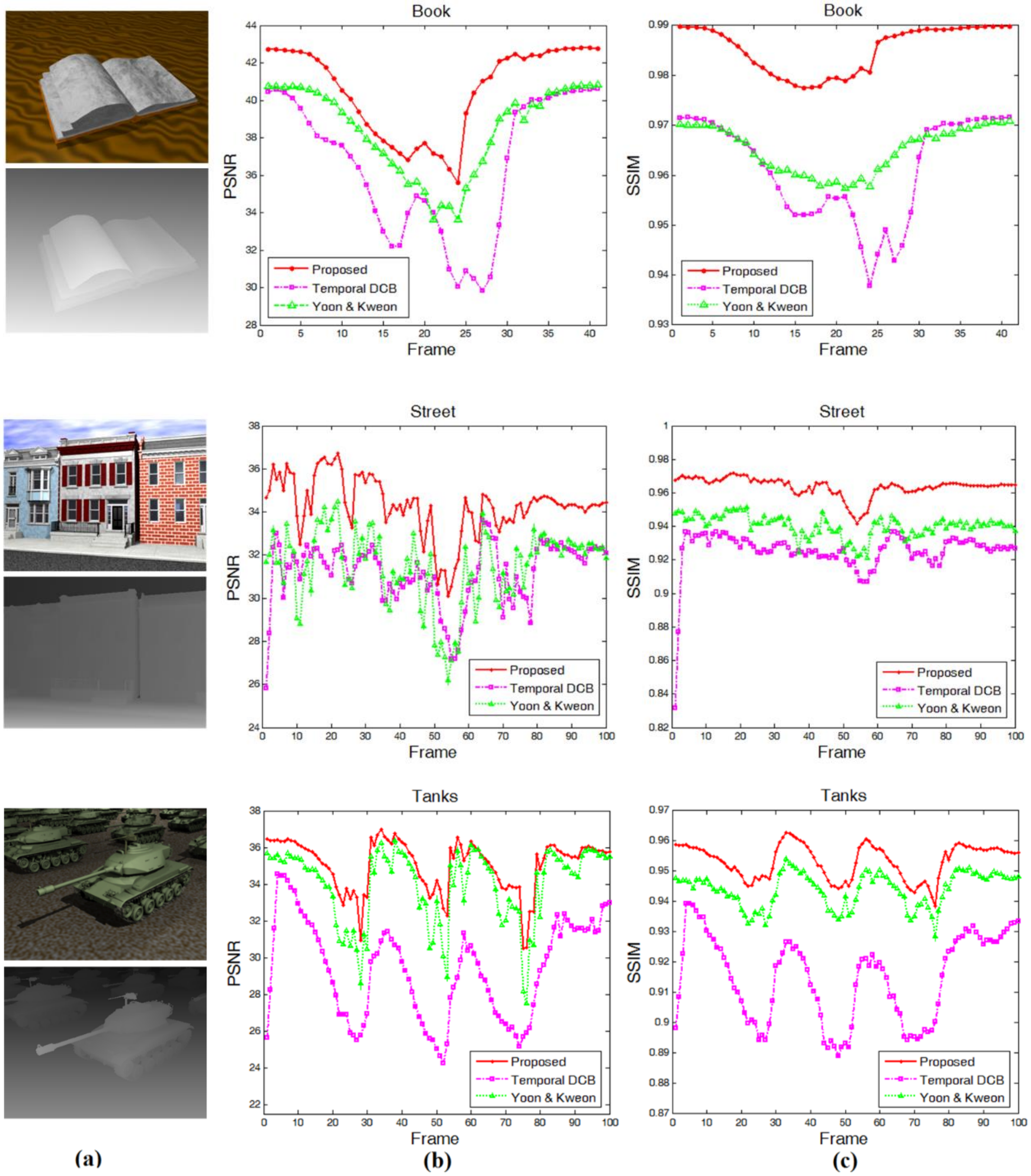


Figure 4. Quantitative Comparison of the Proposed Approach and with the Previous Methods on Book, Street and Tanks Datasets (a) One Color Frame and its Corresponding Disparity Map (b) PSNR and (c) SSIM Index for Different Algorithms

Table 1. presents the average values of PSNR and SSIM for all the sequences. We can see the quality of disparity maps is considerably improved by the proposed approach, the average PSNR is effectively increased at least 2.38dB, while the average SSIM is improved at least 0.02.

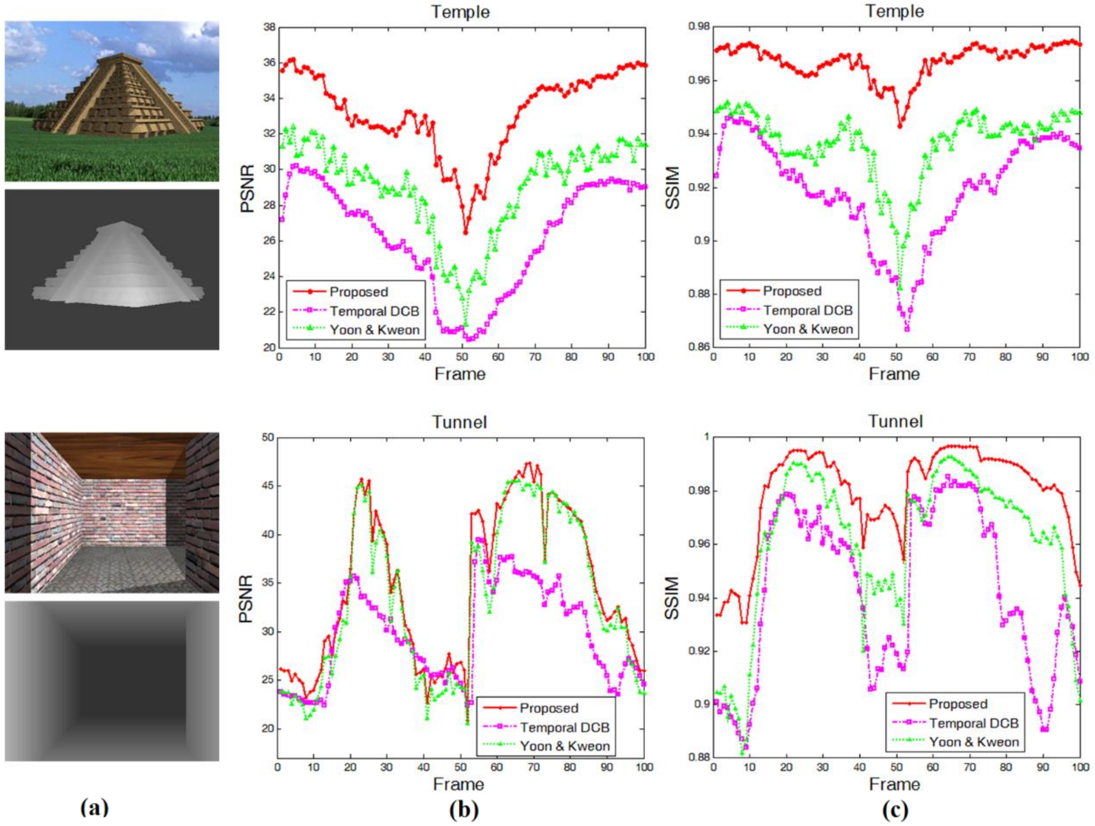


Figure 5. Quantitative Comparison of the Proposed Approach and with the Previous Methods on Temple and Tunnel Datasets (a) One Color Frame and its Corresponding Disparity Map (b) PSNR and (c) SSIM Index for Different Algorithms

Table 1. Resultsof PSNR and SSIM for All the Sequences

Sequence	Average PSNR/dB				Average SSIM			
	Yoon &Kweon	Temporal DCB	Proposed Method	ΔPSNR (MIN)	Yoon &Kweon	Temporal DCB	Proposed Method	ΔPSNR (MIN)
Book	38.44	36.51	40.57	2.13	0.97	0.96	0.99	0.02
Street	31.46	31.17	34.29	2.84	0.94	0.93	0.96	0.02
Tanks	33.94	29.32	35.08	1.14	0.94	0.92	0.95	0.01
Temples	28.98	26.22	33.20	4.22	0.94	0.92	0.97	0.03
Tunnel	33.10	29.70	34.67	1.58	0.96	0.94	0.98	0.02
Average				**2.38**				**0.02**

5. Conclusions and Future Work

In this paper, we proposed a novel temporally consistent disparity estimation algorithm which is implemented based on the Guided Filter and Temporal Gradient (GFTG). It is proposed for processing the disparity sequence in order to enhance its spatial-temporal consistency. In our approach, the original disparity video is firstly smoothed by the guided filter, and then it is processed by an adaptive temporally filter based on the temporal gradient information. Both the subjective and objective evaluations demonstrate that the spatial-temporal consistency of the disparity sequence is highly improved by utilizing the proposed approach. The flickering artifacts and the noises are effectively eliminated while the boundaries of the objects are well preserved.

We employed diverse indexes to evaluate the performance of the proposed algorithm, however, how to quantitatively assess the temporal coherence of a sequence is still a challenging problem so far. For the future work, we would like to investigate this topic and try to explore a more suitable evaluation mechanism.

References

[1] F. Mroz and T. P. Breckon, “An empirical comparison of real-time dense stereo approaches for use in the utomotive environment”, EURASIP Journal on Image and Video Processing, vol. 2012, no. 1, **(2012),** pp. 1–19.

[2] R. Szeliski, “Computer vision”, algorithms and applications, Springer, **(2010)**.

[3] D. Scharstein and R. Szeliski, “A taxonomy and evaluation of dense two-frame stereo correspondence algorithms”, International Journal of Computer Vision, vol. 47, **(2002),** pp. 7–42.

[4] M. Z. Brown, D. Burschka and G. D. Hager, “Advances in computational stereo”, IEEE Transactions on Pattern Analysis and Machine Intelligence , vol. 25, no. 8, **(2003),** pp. 993–1008.

[5] http://vision.middlebury.edu/stereo/.

[6] Y. Boykov, O. Veksler and R. Zabih, “Fast approximate energy minimization via graph cuts”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, no. 11, **(2001),** pp. 1222–1239.

[7] J. Sun, N. Zheng and H. Shum, “Stereo matching using belief propagation”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 25, no. 7, **(2003),** pp. 787–800.

[8] K. J. Yoon and I. S. Kweon, “Adaptive support-weight approach for correspondence search”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 28, no. 4, **(2006),** pp. 650–656.

[9] A. Hosni, C. Rhemann, M. Bleyer, C. Rother and M. Gelautz, “Fast cost-volume filtering for visual correspondence and beyond”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 2, **(2013),** pp. 504 – 511.

[10] C. Leung, B. Appleton, B. C. Lovell and C. Sun, “An energy minimization approaches to stereo-temporal dense reconstruction.Proceedings of the 17th International Conference on IEEE Pattern Recognition, vol. 4, **(2004)** August 23-26, pp. 72–75, Cambridge, UK.

[11] D. Min, S. Yea and A. Vetro, “Temporally consistent stereo matching using coherence function”, in 3DTV-Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON), **(2010)** June 7-9, pp. 1–4, Tampere, Finland.

[12] D. Min, J. Lu and M. N. Do, “Depth video enhancement based on weighted mode filtering”, IEEE Transactions on Image Processing, vol. 21, no. 3, **(2012)**, pp. 1176–1190.

[13] M. Bleyer and M. Gelautz, “Temporally consistent disparity maps from uncalibrated stereo videos”, Proceedings of 6th International Symposium on. IEEE Image and Signal Processing and Analysis, **(2009)** September 16-18, pp. 383–387, Salzburg, Austria.

[14] C. Richardt, D. Orr, I. Davies, A. Criminisi and N. A Dodgson, “Real-time spatiotemporal stereo matching using the dual-cross-bilateral grid”, Computer Vision–ECCV 2010, Springer Berlin Heidelberg, **(2010)**, pp. 510-523.

[15] K. He, J. Sun and X. Tang, “Guided image filtering”, Proceedings of European Conference of Computer Vision, **(2010)** September 5-11, pp. 1–14, Crete, Greece.