A Robust Temporal Depth Enhancement Method for Dynamic Virtual View Synthesis

Can Liu
National ASIC System
Engineering Research Center
Southeast University
Nanjing, 210096, P.R.China
liucan@seu.edu.cn

Weizheng Zhang
Chien-Shiung Wu College
Southeast University
Nanjing, 210096, P.R.China
213113460@seu.edu.cn

Zhi Qi
National ASIC System
Engineering Research Center
Southeast University
Nanjing, 210096, P.R.China
q_zhi@yahoo.com

Longxing Shi
National ASIC System
Engineering Research Center
Southeast University
Nanjing, 210096, P.R.China
Lxshi@seu.edu.cn

ABSTRACT

Depth-image-based rendering (DIBR) is a view synthesis technique that generates virtual views by warping from the reference images based on depth maps. The quality of synthesized views highly depends on the accuracy of depth maps. However, for dynamic scenarios, depth sequences obtained through stereo matching methods frame by frame can be temporally inconsistent, especially in static regions, which leads to uncomfortable flickering artifacts in synthesized videos. This problem can be eliminated by depth enhancement methods that perform temporal filtering to suppress depth inconsistency, yet those methods may also spread depth errors. Although these depth enhancement algorithms increase the temporal consistency of synthesized videos, they have the risk of reducing the quality of rendered videos. Since conventional methods may not achieve both properties, in this paper, we present for static regions a robust temporal depth enhancement (RTDE) method, which propagates exactly the reliable depth values into succeeding frames to upgrade not only the accuracy but also the temporal consistency of depth estimations. This technique benefits the quality of synthesized videos. In addition we propose a novel evaluation metric to quantitatively compare temporal consistency between our method and the state of arts. Experimental results demonstrate the robustness of our method for dynamic virtual view synthesis, not only the temporal consistency but also the quality of synthesized videos in static regions are improved.

Keywords

FTV, DIBR, Temporal consistency, Depth enhancement

1. INTRODUCTION

Virtual view synthesis is being considered as an important component of 3D Video (3DV) [1, 2] and Free-viewpoint Television (FTV) [3, 4, 5] systems. Depth-image-based rendering (DIBR) [6, 7] is one of the widely accepted key techniques of view synthesis.

The synthesis quality of DIBR system highly depends on the accuracy of depth map. For static scenarios, depth estimation has been well-studied using images provided by standard datasets, such as the Middlebury dataset [8]. Numerous methods [9, 10, 11] have presented impressively excellent results of depth estimation as published on the website of Middlebury stereo evaluation [12]. However, for dynamic scenarios, the depth maps estimated using even the most powerful method may still have the problem of temporal inconsistency [13], which leads to flickering artifacts in the synthesized video. There is a much higher possibility that these flickering artifacts exhibit in the challenging areas containing static texture-less regions or static scenes with non-Lambertian surfaces.

In order to remove the flickering artifacts mainly in these static regions, some temporal depth enhancement methods have been proposed. It was shown that spatiotemporal bilateral filter introduced by Richardt *et al.* [14] improved the temporal consistency of depth sequences, but it would spread depth errors to neighboring regions or subsequent frames for overlooking depth reliabilities. As an improvement, Cheng *et al.* [15] presented a quad-lateral filter that took the depth reliability into consideration through measuring the difference of

depth values between two pixels. This work has achieved alleviated depth contamination in neighboring regions to a certain extent, whereas its pixel based processing was quite time consuming. This kind of correction of depth may improve the depth consistency over the entire image. However, it is unable to maintain the correct depth variations due to moving objects in dynamic regions.

The work by Richardt *et al.* [16] and Min *et al.* [17] utilized optical flow to supply motion cues for temporal depth filter, thus they sustained the correct depth variations in motion regions. However, the optical flow analysis suffered from its unaffordable computational expenses and poor performance in texture-less regions. To improve the efficiency of motion extraction, Fu *et al.* [18] proposed a light block-based motion detection algorithm to introduce motion cues into temporal depth filter, which deployed high weight to stationary pixels and low weight to motion pixels. This fast and effective work has been used as the standard temporal depth enhancement framework by Moving Pictures Experts Group (MPEG) to View Synthesis Reference Software (VSRS) [19]. However, Fu's motion detection suffered from inaccuracy and inflexibility problems caused by fixed local window size and thresholds.

The previous methods [16, 17, 18], which only took advantage of the texture, might fail to detect the motion regions when there contained much less texture information. Lu *et al.* [20] used high quality of depth information from depth sensors instead to support correct detection of the scene static structure. Consequently, they successfully preserved the depth variations of dynamic regions in the output of temporal depth filter. However, Lu's work has only been tested with high quality depth videos from Kinect and ToF, while the performance with depth videos obtained by stereo matching methods is to be verified.

Most of the previous methods concentrated on how to precisely separate motion regions from static regions and remained correct depth variations in motion regions. However, for static regions, their depth enhancement methods could not satisfy both temporal consistency and depth accuracy.

In conclusion, the current methods exhibit the following limitations: 1) they sacrifice the quality of depth for temporal consistency, or they cannot maintain temporal consistency and depth quality at the same time; 2) most of the algorithms are computationally too expensive to be real-time.

In this paper, we figure out a robust temporal depth enhancement (RTDE) method to improve depth quality especially in static regions, and keep the good results as temporally consistent as possible. First, we deploy an effective filtering strategy across the entire image based on motion block detection [21]. Given the motion detection, we then process depth values only in static regions and leave the correct depth variations in motion regions untouched. Second, in order to improve the robustness with regarding to the first limitation mentioned above, we introduce the depth reliability information as an important filtering weight into the temporal depth filter. Experimental results show that our method is fast and achieve satisfactory synthesized videos with regarding to both rendering quality and temporal consistency. In addition, inspired by the work of Fu *et al.* [18] and Solh *et al.* [22], we propose a simple and feasible evaluation metric to measure the temporal consistency quantitatively. We believe the similarity of Mean Square Difference (MSD) curves of adjacent frames is robust to illumination variations from different views, and reveals the degree of the temporal consistency when we compare the synthesized videos from the enhanced and original depth sequences.

The rest of this paper is organized as follows. Section 2 addresses the proposed robust temporal depth enhancement algorithm and temporal consistency evaluation metric. Experimental results are shown in Section 3, finally Section 4 concludes this paper.

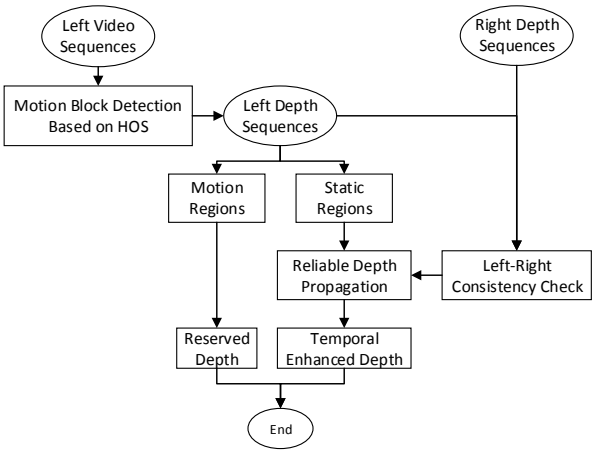


Figure 1: Enhancement flowchart of the left depth sequences

2. ALGORITHMS

In this paper, Depth Image-based Rendering (DIBR) deploys the framework that synthesizes the intermediate virtual view from both left and right reference views. Our proposed temporal enhancement method is a component of DIBR system as the depth preprocessing procedure. The flowchart illustrated in Figure 1 shows the preprocessing of left depth sequences, which is the same to the right depth sequences. Using motion block detection based on High-order Statistics (HOS) [21], we separate one input depth frame into static and motion regions. We only suppress depth inconsistency in static regions while maintain reasonable depth variations in motion regions. The depth values that pass through the

consistency check between the left and right images will be used to recover those spoiled depth values in the same static regions, not only in the current frame but also in succeeding frames. In such a manner, the reliable estimation of depth has been propagated.

The crucial steps of our proposed method will be elaborated subsequently. In the last part of this section, we will discuss the proposed objective evaluation metric for temporal consistency which is utilized in our experiments.

2.1 Motion detection based on HOS

As described in previous section, the depth inconsistency problem mainly occurs in the static regions, thus the motion detection should be conducted to supply motion cues to temporal depth filter for appropriate depth enhancement in static regions, without interfering reasonable depth variations in dynamic regions. Most current motion detection studies based on inter-frame difference, such as [18] and [25], may suffer from the inaccuracy and inflexibility in their results due to fixed local window size and thresholds. Because of the unavoidable misalignment between the depth map and color image [26], this triggers depth destruction after depth enhancement near moving object boundaries, as shown in Figure 2.

(a) (b)

Figure 2: Depth destruction near moving object boundaries caused by misalignment problem. In the 39th frame, the misalignment between the depth map and the color image, especially in the head and leg regions, as in shown in (a), leads to depth artifacts in the subsequent frame as shown in the yellow marked rectangle regions in (b).

Instead of using a nicely aligned motion segmentation result that is difficult and expensive to achieve, we prefer a bounding box (i.e. motion block) to separate the static regions from the dynamic areas in images. In this way, the misalignment problem can be effectively avoided and it also runs fast. Usually the accuracy of bounding box segmentation depends excessively on the threshold of inter-frame difference. Since a unified threshold is usually too coarse to capture the noise variation in video sequences, a step-forward processing on inter-frame difference map before binaryzation is expected to generate more accurate bounding box based motion detection.

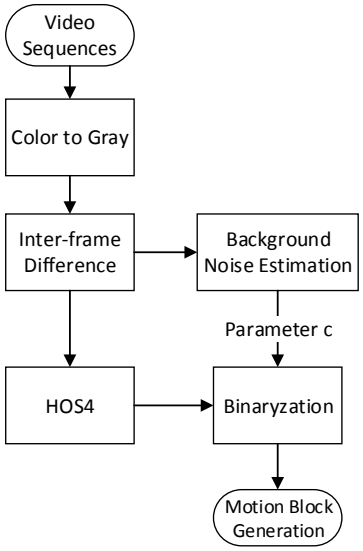
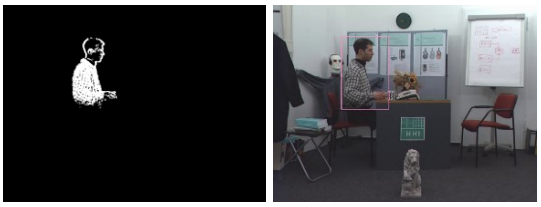


Figure 3: Workflow of motion detection based on HOS

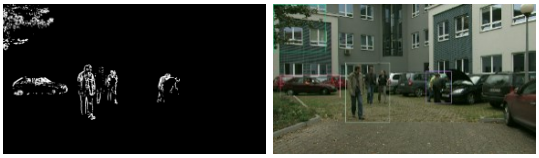
In our work, we use motion block detection based on HOS, which is adaptable to noise variations in video sequences thus it is able to receive precise motion segmentation results. The workflow is illustrated in Figure 3. The video sequences are firstly altered to gray-level images and then the inter-frame difference maps are calculated. Since the noise obeys Gaussian-distribution while moving objects have strong structure, which contributes to the high order statistics of frame difference. Consequently, we can accomplish motion block detection by separating the non-Gaussian signals form the Gaussian one. Considering the computation effort and accuracy, we calculate the 4-order moment of inter-frame difference as high order statistics (i.e. HOS4) and compare it with an adaptive threshold T, which is determined by the estimated background noise $\hat{\sigma}_d^2$ in static regions and a constant parameter c, i.e. $T = c(\hat{\sigma}_d^2)$. If the 4-order moment of a pixel is higher than T, we attribute this pixel to motion regions, and static regions on the contrary. The detected results of three datasets are shown in Figure 4. Additionally, parameter c and the calculation of the 4-order moment could be completed simultaneously, thus making motion detection more time-saving.



(a) Bookarrival

 ISBN 978-80-86943-65-7

(b) Pantomime



(c) Carpark

Figure 4: Detected motion blocks of three datasets

2.2 Robust temporal depth enhancement

The purpose of depth enhancement in static regions is to smooth depth map without causing any degradation in accuracy, however, this target is difficult to achieve. Most of the current methods [16][17][18], which achieve temporal depth consistency to a certain degree, overlook the reliability of depth values thus eventually lead to annoying perception, as illustrated in Figure 5.

2.2.1 Depth reliability check

Depth reliability is crucial to guarantee the consistent good quality of depth maps, especially in static regions, over the sequence of synthesized videos. Since it is unrealistic to judge the reliability of depth map using genuine depth information, in this paper we combine a left-right consistency checking (LRCC) method to detect then propagate the qualified depth values to successive frames.

Left-right consistency check is also called the two-view constraint [33]. It distinguishes outliers caused by occlusion, texture-less areas, and false match points. Assume that the disparity of a pixel p(x, y) in the left camera is $d_{LR}(x,y)$, and $d_{RL}(x,y)$ vice versa. The depth reliability mask is then defined as equation (1), where we set the threshold T to 1.

$$Mask(x,y)=\begin{cases}1, & |d_{LR}(x,y)\text{-}d_{RL}(x-d_{LR}(x,y),y)|<T\\ 0, & otherwise\end{cases} \quad (1)$$

2.2.2 Temporal depth filter

Depth values in the non-zero region of reliability mask that present good qualities are fed to our temporal depth filter to eliminate the depth contamination due to errors in neighboring area. This step is called Robust Temporal Depth Enhancement (RTDE). Specifically, our method handles three different conditions as described by the pseudo code of RTDE below.

In ***Case 1***, the depth pixel $\boldsymbol{p(n)}$ in the current frame $\boldsymbol{n}$ is reliable while the enhanced corresponding pixel in previous frame $\boldsymbol{p'(n-1)}$ is unreliable, in this occasion we reserve the depth value of current frame, therefore the high reliability depth value will not be contaminated by previous depth errors. In ***Case 2***, both the depth pixel $\boldsymbol{p(n)}$ and enhanced corresponding pixel $\boldsymbol{p'(n-1)}$ are detected to be reliable, the enhanced depth value of current frame $\boldsymbol{p'(n)}$ is calculated as the weighted sum of $\boldsymbol{p(n)}$ and $\boldsymbol{p'(n-1)}$, in this occasion we pledge the reliability of enhanced depth value and alleviate the negative effects caused by misdetection of LRCC. In ***Case 3***, the depth pixel $\boldsymbol{p(n)}$ in current frame $\boldsymbol{n}$ is detected as unreliable, we inherit the previous enhanced depth value $\boldsymbol{p'(n-1)}$, in this occasion we could guarantee both the consistency and quality of depth information.

for original depth pixel $\boldsymbol{p(x,y,n)}$ ***in current frame n, and*** $\boldsymbol{p'(x,y,n)}$ ***is the enhanced pixel of p, let*** $\boldsymbol{p(n)=p(x,y,n)}$, $\boldsymbol{p'(n)=p'(x,y,n)}$;

Case 1. ***if*** $\boldsymbol{p(n)}$ ***reliable &&*** $\boldsymbol{p'(n-1)}$ ***unreliable***

$\boldsymbol{p'(n)=p(n)}$;

end if

Case 2. ***else if*** $\boldsymbol{p(n)}$ ***reliable &&*** $\boldsymbol{p'(n-1)}$ ***reliable***

$\boldsymbol{p'(n)=\alpha\times p(n)+(1-\alpha)\times p'(n-1)}$;

end if

Case 3. ***else***

$\boldsymbol{p'(n)=p'(n-1)}$;

end if

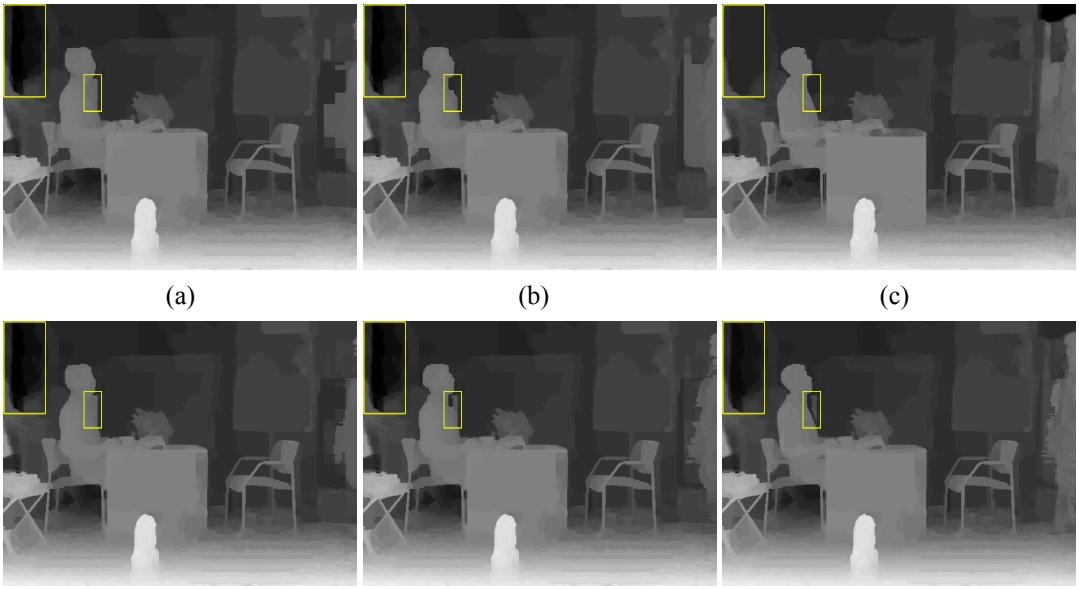
end for

In our enhancement scheme, the unreliable depth values in current frame are continuously replaced by reliable depth values in previous frames, in the meanwhile, the reliable depth values in current frame are preserved and propagated to the following frames to upgrade the depth quality of static regions. Moreover, the continuous delivery of dependable depth values used in ***Case 2*** and ***Case 3*** also suppress the fluctuation of depth values in the temporal domain. After being enhanced, the rectified depth maps will be utilized in further DIBR schemes.

Compared with other temporal enhancement methods, the proposed RTDE concentrates on both the temporal consistency and the depth quality improvement in static regions. The experimental results in Section 3 will demonstrate the robustness of our method.

2.3 Evaluation of Temporal Consistency

Current temporal consistency evaluation of the synthesized video is mainly prioritized to subjective evaluation in the absence of simple but efficient objective assessment metrics. Inspired by [18, 22], in this paper we introduce a novel objective metric to assess temporal consistency of the synthesized video. The idea is to compare the similarity of Mean Square Difference (MSD) curves of adjacent frames between



(a) (b) (c)

(d) (e) (f)

Figure 5: Depth quality degradation illustration. (a)-(c) corresponds to original depth for frame 9 to 11 of "Bookarrival", (d)-(f) corresponds to enhanced depth by TDE. As shown in (f), the yellow rectangle marked regions inherit errors of previous frames, leading to depth quality degradation.

Datasets	Frames	Resolution	View Synthesis (left, right → center)	c	T	α
Bookarrival	1→100	1024*768	(10,6) → 8	100	1	0.25
Pantomime	1→150	1280*960	(37,41) → 39	85	1	0.25
Carpark	1→100	1920*1088	(5,3) → 4	125	1	0.25

Table 1: Experimental parameters

the synthesized video and the original one. The temporal consistency is satisfying when two MSD curves are coherent. The similarity is measured as the Standard Deviation (STD) of differences of MSD values.

Assume $I_{ori}(x,y,k)$ and $I_{syn}(x,y,k)$ represent the intensity of pixel (x,y) in the original and synthesized frame respectively, k is the frame number. The MSD of $I_{ori}(x,y,k)$ and $I_{syn}(x,y,k)$ is determined by equation (2) and (3), where m, n represent the width and height of selected evaluation region.

$$MSD_{ori}(k) = \frac{1}{mn}\sum_{x=0}^{m-1}\sum_{y=0}^{n-1}|I_{ori}(x,y,k) - I_{ori}(x,y,k-1)|^2 \quad (2)$$

$$MSD_{syn}(k) = \frac{1}{mn}\sum_{x=0}^{m-1}\sum_{y=0}^{n-1}|I_{syn}(x,y,k) - I_{syn}(x,y,k-1)|^2 \quad (3)$$

The Temporal Inconsistency Index (TII), or the similarity of MSD curve is determined by equation (4), where STD is determined by the difference of MSD, written in expression (5).

$$TII = \frac{1}{Similarity} = STD(Differenece\ of\ MSD) \quad (4)$$

$$Difference\ of\ MSD(k) = MSD_{ori}(k) - MSD_{syn}(k) \quad (5)$$

As described in equation (4), lower STD value indicates higher similarity and better temporal consistency of synthesized video. Experimental results in next section will demonstrate that our objective evaluation metric accesses the temporal consistency of synthesized video as efficiently as subjective human perceptions do.

3. EXPERIMENTAL RESULTS

In this section, we compare the proposed RTDE method with the temporal depth enhancement (TDE) method by Fu *et al.* [18] that is implemented in MPEG-VSRS. We test out method on the databases of 'Bookarrival' [27], 'Carpark' [28], and 'Pantomime' [29]. To our knowledge, the depth maps of 'Bookarrival' and 'Carpark' are generated using DERS [30], while 'Pantomime' is generated by the depth estimation software [31] from Nagoya University. All three datasets share the problem of temporal inconsistency in their depth sequences, especially in static regions, which leads to flickering artifacts in synthesized video sequences. To synthesize video sequences, we deploy the software VSRS provided by MPEG. Parameters of our experiment are concluded in Table 1, where **c** is the constant parameter in motion block detection based on HOS(see in Section 2) to adjust noise threshold. **T** is the parameter in LRCC (see in Equation 1). And **α** is the parameter in RTDE (see in Figure 5).

(a) (b) (c)

Figure 6: Temporal depth enhancement in static regions. (a) The original depth map without being processed. (b) The enhanced result using TDE. (c) Our result.

To evaluate our test results, we assess both rendering quality and temporal consistency of synthesized videos. Videos from multiple cameras may suffer from uneven illumination in the environment, thus making PSNR an unstable quality measurement for its sensitivity to variances of illumination. In our case, we combine SSIM [32] with PSNR as the assessment criteria. The comprehensive judgment with both criterions provides a fair enough measurement of similarity between the reference view and the virtual view regardless of annoying illumination condition. For temporal consistency, we use the proposed Temporal Inconsistency Index (TII), which is defined in equation (4).

3.1 The evaluation in static regions

In this section, we mainly analyze the robustness of our method in static regions. In the next subsection, we will evaluate the performance of our method across the entire image.

Figure 6 illustrates the specific results of depth processing in static regions of dataset Bookarrival (frame No. 79). In this frame depth errors in the marked static regions are inherited by TDE method. We achieve better results than TDE in static regions since we replace the unreliable depth values in the current frame by previous reliable depth values.

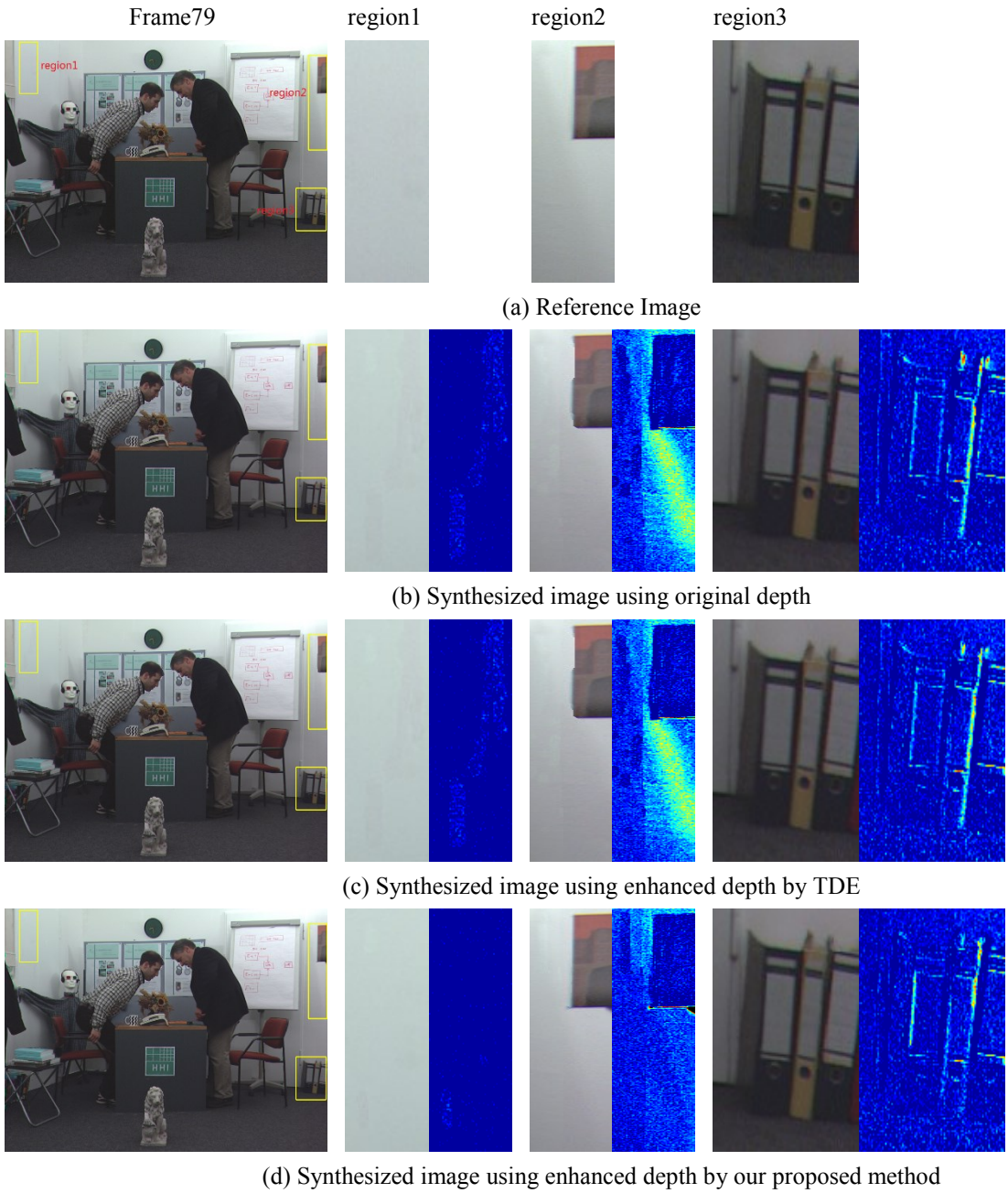
Figure 7 illustrates the qualitative analysis of synthesis quality in static regions. We separate the entire image into several regions, among which we mainly concentrate our analysis on region 1, 2, and 3. The reason we isolate these static regions is that the depth values in these regions share an apparent fluctuation, i.e. depth inconsistency in timeline, thus they are suitable to demonstrate the capability of our analysis. In Figure 7(c), the synthesized result is unsatisfying and similar to the result using original depth information. This is because TDE inherits errors from previous frames shown in Figure 6. In the proposed method, we replace the poor depth estimation with the reliable depth information from previous frames, therefore our results are better in quality, both perceptually and numerically (see Table 2).

Figure 8 illustrates the temporal consistency in static regions of the synthesized video, where x axis represents frame number and y axis indicates the Difference of MSD defined in Equation (5). The distortion in y axis represents inconsistency of video sequences. For region 1 in Figure 8(a), the inconsistency problem of original depth sequences mainly occurs in frame 11, 51 and 81 to 83, thus causing flickering artifacts in the synthesized video. After being processed by TDE, the fluctuation of depth is weakened, while inconsistency still remains. Since TDE just smooths the depth sequences which only reduces the range of depth fluctuation. In our proposed method, the fluctuation could be reduced to a great extent by continuously delivering reliable depth in the temporal domain. In Figure 8(a), the inconsistency around frame 11 cannot be suppressed, because depth values in region one are not reliable in preceding frames. In the case of the first condition when applying the temporal depth filter, the reliable depth values in frame 11 are preserved, furthermore, they are propagated to the subsequent frames. For region 2 and region 3, we only analyze frame 40 to 100 since obvious motion appears in the previous 39 frames, the results are shown in Figure 8(b) and (c). Overall, our proposed method keeps the best temporal consistency compared with other methods in static regions. The statistical data collection shown in Table 2 tells us that our method provides a better quality, which ensures the temporal consistency in static regions.

In Table 2, we analyze quantitatively both the rendering quality and the temporal consistency in three different static regions as presented in Figure 7. It is shows that our method obtains much better results than those of the original method or TDE.

3.2 The evaluation across the entire image

Although RTDE is not deployed to dynamic regions in this paper, experimental results indicate that our method achieves the best trade-off between the rendering quality and the temporal consistency across the entire image among synthesized results generated



(a) Reference Image

(b) Synthesized image using original depth

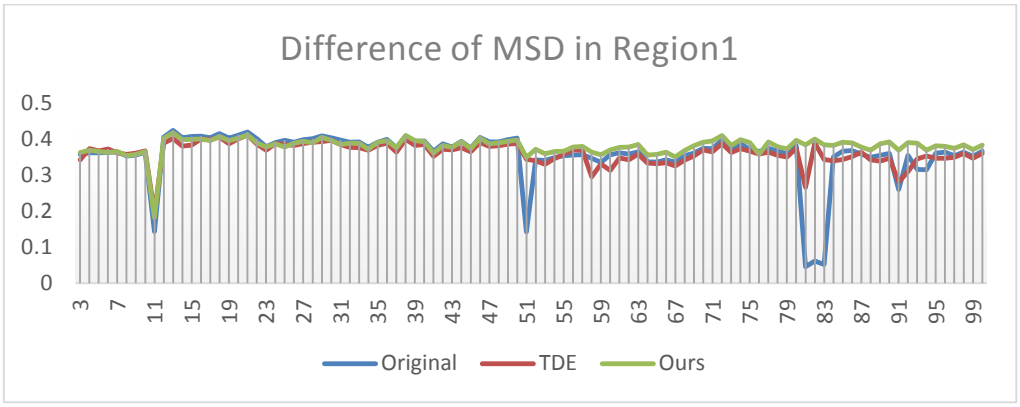
(c) Synthesized image using enhanced depth by TDE

(d) Synthesized image using enhanced depth by our proposed method

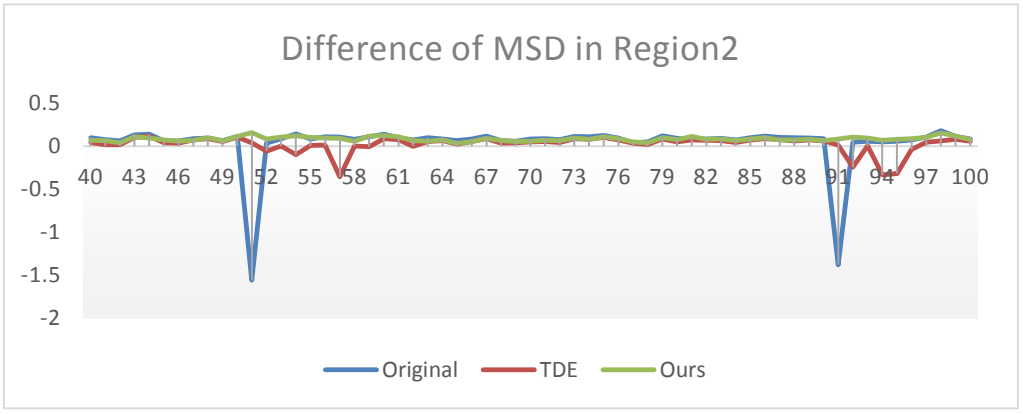
Figure 7: Qualitative analysis of synthesis quality in the static regions. (a) represents the reference image of Frame 79. (b) is the synthesized result using original depth. (c) indicates the synthesized image using TDE. (d) is the result of our method. The 2th, 4th and 6th columns show the synthesized results of local static regions. The error heatmaps between synthesized results and reference images are displayed in the 3th, 5th and 7th columns . Warm colors mean large errors, and cold colors mean small errors.

by original depth and enhanced depth from TDE (see Table 3).

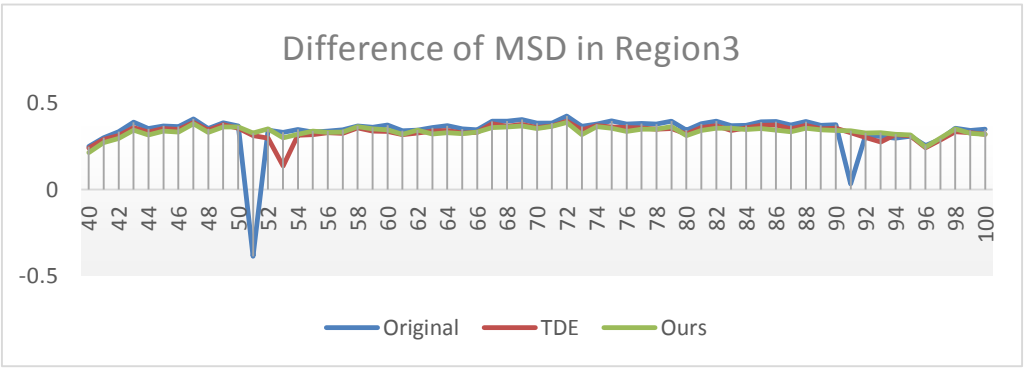
In Table 3, the temporal consistency of our proposed method in Bookarrival and Carpark datasets are a little lower than TDE but higher than the original one. This is because RTDE focuses merely on static regions, while TDE concentrates on the entire image so the temporal consistency in motion regions will also be improved. However, TDE provides the worst synthesis quality in Bookarrival and Carpark datasets, because the inaccuracy of motion detection and misalignment problem will degrades depth quality in motion regions or near the boundaries of moving objects. On the contrary, our RTDE could provide the best synthesis quality compared with the other two methods across the entire image of three datasets.



(a)



(b)



(c)

Figure 8: Temporal consistency evaluation in the static regions

Bookarrival	Region1(1-100)			Region2(40-100)			Region3(40-100)		
	Original	TDE	Ours	Original	TDE	Ours	Original	TDE	Ours
PSNR	49.08307	49.82602	**50.20146**	30.57667	30.74217	**31.55368**	34.54613	34.34882	**35.09781**
SSIM	0.946841	0.947205	**0.947711**	0.886019	0.900981	**0.936692**	0.920664	0.923001	**0.925322**
TII	0.068823	0.031144	**0.025312**	0.281997	0.098828	**0.025767**	0.109639	0.040748	**0.029024**

Table 2: Quantitative evaluation of synthesis quality and temporal consistency in the static regions. The PSNR and SSIM data are the average values of all frames. It should be noted that lower TII values represent higher temporal consistency and the bold figures indicate the best results among three methods.

 ISBN 978-80-86943-65-7

	Bookarrival			Pantomime			Carpark		
	Original	TDE	Ours	Original	TDE	Ours	Original	TDE	Ours
PSNR	35.76804	35.76522	**35.92959**	39.48054	39.53301	**39.76341**	30.94377	30.89608	**30.96458**
SSIM	0.923868	0.923806	**0.923939**	0.960488	0.960489	**0.963027**	0.890607	0.889084	**0.890761**
TII	0.053621	**0.029822**	0.035902	0.122023	0.101601	**0.095132**	0.297778	**0.058146**	0.128696

Table 3: Quantitative evaluation of synthesis quality and temporal consistency across the entire image

Moreover, for the dataset of Pantomime, our method keeps the best rendering quality as well as the temporal consistency. Because, there exists tremendous depth fluctuation in static regions of Pantomime. TDE only reduces the range of depth fluctuation, while our method could suppress them to a great extent.

4. CONCLUSIONS

In this paper, we proposed a robust temporal depth enhancement method for dynamic virtual view synthesis, which includes an effective and efficient motion block detection and a robust temporal depth filter aided by depth reliability check. Experimental results prove the robustness of our method that temporal consistency and synthesis quality can be both improved in static regions. Moreover, we proposed a comprehensive objective evaluation metric which is efficient and reasonable in assessing the temporal consistency of synthesized video sequences. However, as described in previous sections, in this paper we concentrate our method merely in static regions and only temporal filtering is conducted. In the future, we will extend our idea and method to motion regions and a spatiotemporal filter will be utilized to enhance depth sequences. Additionally, the accuracy of depth reliability check will also be considered.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China (Grant No.61204023, Grant No.61203251 and Grant No. 61404028).

REFERENCES

[1] A. Redert, M.O. de Beeck, C. Fehn, W. Ijsselsteijn, M. Pollefeys, L. Van Gool, E. Ofek, I. Sexton, and P. Surman, Advanced three-dimensional television system technologies, 3D Data Processing Visualization and Transmission, 2002. Proceedings. First International Symposium on, 2002, pp. 313-319.

[2] C. Zhu, Y. Zhao, L. Yu, and M. Tanimoto, 3D-TV System with Depth-Image-Based Rendering, Springer New York, 2013.

[3] T. Fujii, and M. Tanimoto, Free viewpoint TV system based on ray-space representation, ITCom 2002: The Convergence of Information Technologies and Communications, International Society for Optics and Photonics, 2002, pp. 175-189.

[4] M. Tanimoto, Overview of free viewpoint television. Signal Processing: Image Communication 21 (2006) 454-461.

[5] Y. Mori, N. Fukushima, T. Yendo, T. Fujii, and M. Tanimoto, View generation with 3D warping using depth information for FTV. Signal Processing: Image Communication 24 (2009) 65-72.

[6] C. Fehn, A 3D-TV approach using depth-image-based rendering (DIBR), Proc. of VIIP, 2003.

[7] C. Fehn, Depth-image-based rendering (DIBR), compression, and transmission for a new approach on 3D-TV, Electronic Imaging 2004, International Society for Optics and Photonics, 2004, pp. 93-104.

[8] D. Scharstein, and R. Szeliski, A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. International journal of computer vision 47 (2002) 7-42.

[9] X. Mei, X. Sun, M. Zhou, H. Wang, and X. Zhang, On building an accurate stereo matching system on graphics hardware, Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on, IEEE, 2011, pp. 467-474.

[10] C. Rhemann, A. Hosni, M. Bleyer, C. Rother, and M. Gelautz, Fast cost-volume filtering for visual correspondence and beyond, Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, IEEE, 2011, pp. 3017-3024.

[11] Q. Yang, A non-local cost aggregation method for stereo matching, Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE, 2012, pp. 1402-1409.

[12] http://vision.middlebury.edu/stereo/eval/

[13] R. Khoshabeh, S.H. Chan, and T.Q. Nguyen, Spatio-temporal consistency in video disparity estimation, Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on, IEEE, 2011, pp. 885-888.

[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, 2010, pp. 510-523.

[15] C. Cheng, S. Lin, and S. Lai, Spatio-temporally consistent novel view synthesis algorithm from video-plus-depth sequences for autostereoscopic displays. Broadcasting, IEEE Transactions on 57 (2011) 523-532.

[16] C. Richardt, C. Stoll, N.A. Dodgson, H.P. Seidel,

and C. Theobalt, Coherent spatiotemporal filtering, upsampling and rendering of RGBZ videos, Computer Graphics Forum, Wiley Online Library, 2012, pp. 247-256.

[17] D. Min, J. Lu, and M.N. Do, Depth video enhancement based on weighted mode filtering. Image Processing, IEEE Transactions on 21 (2012) 1176-1190.

[18] D. Fu, Y. Zhao, and L. Yu, Temporal consistency enhancement on depth sequences, Picture Coding Symposium (PCS), 2010, IEEE, 2010, pp. 342-345.

[19] M. Tanimoto, T. Fujii, and K. Suzuki. View synthesis algorithm in view synthesis reference software 3.5 (VSRS3. 5) Document M16090, ISO/IEC JTC1/SC29/WG11 (MPEG). 2009.

[20] S. Lu, N.N. King, L. Chern-Loon, and L. Songnan, Online Temporally Consistent Indoor Depth Video Enhancement via Static Structure. Image Processing, IEEE Transactions on 24 (2015) 2197-2211.

[21] L.M. Garth, and H.V. Poor, Detection of non-Gaussian signals: A paradigm for modern statistical signal processing [and prolog]. Proceedings of the IEEE 82 (1994) 1061-1095.

[22] M. Solh, G. AlRegib, and J.M. Bauza, 3VQM: A vision-based quality measure for DIBR-based 3D videos, Multimedia and Expo (ICME), 2011 IEEE International Conference on, Barcelona, Spain, 2011, pp. 1-6.

[23] S.D. Cochran, and G. Medioni, 3-D surface description from binocular stereo. IEEE Transactions on Pattern Analysis and Machine Intelligence 14 (1992) 981-994.

[24] P. Fua, A parallel stereo algorithm that produces dense depth maps and preserves image features. Machine vision and applications 6 (1993) 35-49.

[25] S. Lee, and Y. Ho, Temporally consistent depth map estimation using motion estimation for 3DTV, International Workshop on Advanced Image Technology, 2010.

[26] X. Xu, L. Po, K. Cheung, L. Feng, and C. Cheung, Watershed based depth map misalignment correction and foreground biased dilation for DIBR view synthesis, Image Processing (ICIP), 2013 20th IEEE International Conference on, IEEE, 2013, pp. 3152-3156.

[27] I. Feldmann, *et al.* HHI Test Material for 3D Video. ISO/IEC JTC1/SC29/WG11, M15413, April 2008.

[28] ftp://multimedia.edu.pl

[29] http://www.fujii.nuee.nagoya-u.ac.jp/multiview-data/mpeg/mpeg_ftv.html

[30] M. Tanimoto, T. Fujii, M. P. Tehrani, M. Wildeboer Depth Estimation Reference Software (DERS) 4.0, ISO/IEC JTC1/SC29/WG11, MPEG 2008/M16605, 2009.

[31] M. Tanimoto, T. Fujii, and K. Suzuki. Reference software of depth estimation and view synthesis for FTV/3DV. ISO/IEC JTC1/SC29/WG11 M 15836, 2008.

[32] Z. Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity. Image Processing, IEEE Transactions on 13 (2004) 600-612.

[33] C. Georgoulas, G.C. Sirakoulis, and I. Andreadis, Real-time Stereo Vision Applications, INTECH Open Access Publisher, 2010.