Multimed Tools Appl (2015) 74:5935–5954
DOI 10.1007/s11042-014-1898-1

View synthesis distortion elimination filter for depth video coding in 3D video broadcasting

Linwei Zhu · Yun Zhang · Xu Wang · Sam Kwong

Published online: 25 February 2014

Abstract Depth image based rendering (DIBR), which can generate synthesized images according to the users' demand, is a key technique for achieving 3D television. However, view synthesis by DIBR technique is very sensitive to depth coding distortion. Because depth distortion will lead to geometrical rendering position errors, and seriously affect the quality of synthesized images. In this paper, we propose an in-loop filter to minimize view synthesis distortion at the cost of transmitting extra filter parameters as supplementary information. And an adaptive parameter determination scheme is presented for the proposed filter. Then a good trade-off between bit rate and view synthesis distortion has been achieved by considering the spatial-temporal correlations of 3D video sequence. The simulation results reveal that the proposed view synthesis distortion elimination method can significantly improve the rate-distortion performance, which achieves Bjontegaard Delta Peak Signal-to-Noise Ratio (BDPSNR) gain from 0.41 to 1.09 dB compared with the benchmark.

Keywords 3D video · View synthesis · Depth image based rendering · Depth video coding

1 Introduction

For the requirement of visual enjoyment, three-dimensional video (3DV) [13, 18], as a novel type of multimedia, has attracted more and more attention. It is capable of providing the audiences with depth perception and interactivity. To enable the depth perception and interactive functionalities in 3DV system, multiview depth video is adopted to provide geometrical information for view synthesis based on the depth image based rendering (DIBR) [9] technique. According to this technique, it can generate arbitrary viewpoint of 3DV at the client without encoding and transmitting all the views at the server. In this technique, the accuracy of depth information is critical since the depth video is used for view synthesis instead of being

L. Zhu · Y. Zhang (✉)
Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China
e-mail: yun.zhang@siat.ac.cn

Y. Zhang · X. Wang · S. Kwong
Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong

X. Wang · S. Kwong
Shenzhen Research Institute, City University of HongKong, Shenzhen, China

displayed directly. In order to reduce the amount of coding bits, multiview depth videos should be encoded and transmitted to the client. However, traditional video coding standards are designed for compressing color video. Compressing depth video directly by these coding standards often introduces serious coding artifacts along object boundaries, which severely affects the subjective and objective quality of synthesized images.

To encode the depth video efficiently, many researchers devote their efforts to develop highly efficient depth video coding schemes and depth reconstruction algorithm. Oh et al. [10, 11] proposed a depth reconstruction filter which considered the occurrence frequency, similarity, and closeness of pixels. It was robust to noise and smoothness. It reduced the depth bit rate and improved the rendering quality. Liu et al. [6] proposed a depth coding algorithm by utilizing the structure similarity between depth and corresponding color video, in which the in-loop filter and coding mode were introduced. In [8], a weighted mode filter was proposed to suppress the coding artifacts, then the spatial resolution sampling and the dynamic range compression were used to reduce bit rate. Basically, in these schemes, kinds of in-loop filters are proposed. They all only focus on the post-processing of depth video coding for achieving better quality of synthesized images.

In the coding optimization, Yuan et al. [14] derived a concise distortion model for the synthesized virtual views, and optimized the bit allocation scheme between depth and color video based on the Lagrangian multiplier method. Hu et al. [5] presented a rate control (RC) scheme for multiview video coding which optimized the bit allocation problem including the joint depth/color bit allocation and frame level bit allocation. The objective function was the maximum sum of quality of all real and virtual views. Generally, these above mentioned algorithms mainly focus on the bit allocation optimization between color videos and depth videos, to achieve the optimal rate-distortion (R-D) performance.

In addition, there are researchers who analyze the priorities of depth video, and propose some algorithms to pre-process depth video for enhancing the object boundaries of depth video. For instance, Zhao et al. [19] proposed a depth no-synthesis-error model to exploit the depth redundancies, which was consequently applied to the depth pre-processing to improve the coding efficiency. In his finding, the depth value can fluctuate in a certain range without influence on the quality of synthesized images. To further improve the depth video coding efficiency and the quality of synthesized images, Yuan et al. [15] theoretically analyzed the relationship between depth coding distortion and view synthesis distortion in 3DV system, and proposed a new 3DV diagram containing Wiener filter on synthesized images to eliminate view synthesis distortion. The scheme efficiently improved the quality of synthesized images by utilizing additional filter coefficients. Though the quality of synthesized images could be improved in some case, these filter coefficients require a large transmission bandwidth. Furthermore, the spatial diversity and property of the 3DV were not considered when calculating the Wiener filter coefficients.

In this paper, in order to further improve the depth coding performance, we exploit spatial-temporal correlations of multiview video sequence, and propose a new scheme to reduce the filter coefficients based on Yuan's coding framework. The remainder of this paper is organized as follows. Section 2 analyzes depth distortion effect to view synthesis distortion. Section 3 presents the proposed scheme of eliminating view synthesis distortion for synthesized images. Experimental results and comparisons are presented in Section 4. Finally, conclusions are given in Section 5.

2 Analysis on depth distortion effect to view synthesis distortion

In a general 3DV system framework, multiview video plus depth has been the central data format of representing 3D world scene. Usually, multiview color videos will firstly be captured

by camera array and their corresponding depth videos could be captured by depth cameras or generated from depth estimation based on advanced stereo matching algorithms. Both color and depth videos will be encoded and transmitted to the client. Due to the limited number of cameras, the set of view is relatively sparse. To support the requirement of display, more virtual view color videos shall be generated by view synthesis algorithms at the client with reconstructed multiview color and depth videos. Finally, the reconstructed multiview color and synthesized videos are input to display for viewing. Figure 1 shows the general framework of the 3DV system.

To analyze depth distortion effect to view synthesis distortion, let (x,y) be the pixel location at image plane of real camera and (u,v) be the pixel location at image plane of virtual camera that rendered from pixel (x,y). It is assumed that the two cameras are parallel. Then the relationship between the two pixels can be represented as [14]

$$u = x + \frac{f \times l}{z}, v = y \tag{1}$$

where f represents the horizontal focal length of camera, l is the baseline distance between two cameras. In 3DV data format, depth video is usually represented by 8-bits value and nonlinear quantization is adopted for conversion from physical depth to 8-bit depth value [9], which is

$$\frac{1}{z} = \frac{d}{255} \times C_1 + C_2 \tag{2}$$

where d is depth value ranging from 0 to 255, $C_1 = \frac{1}{z_{\min}} - \frac{1}{z_{\max}}$, $C_2 = \frac{1}{z_{\max}}$, z_{min} and z_{max} are the nearest and farthest depth plane of 3D scene, respectively. Suppose Δd be the distortion of the pixel location (x,y) with original depth value d, which is introduced by depth video coding, then the geometrical position error Δu caused by Δd can be described as follow.

$$\Delta u = \frac{f \times l}{255} \times C_1 \times \Delta d \tag{3}$$

In Eq. 3, it implies depth distortion from coding that causes geometrical position error of video object, thus introduces view synthesis distortion and decreases the quality of synthesized image [17]. In addition, the geometrical position error and depth distortion are with direct proportional relation. To further analyze the influence of synthesized image quality from depth distortion, coding and view synthesis experiments from different distorted depth videos are

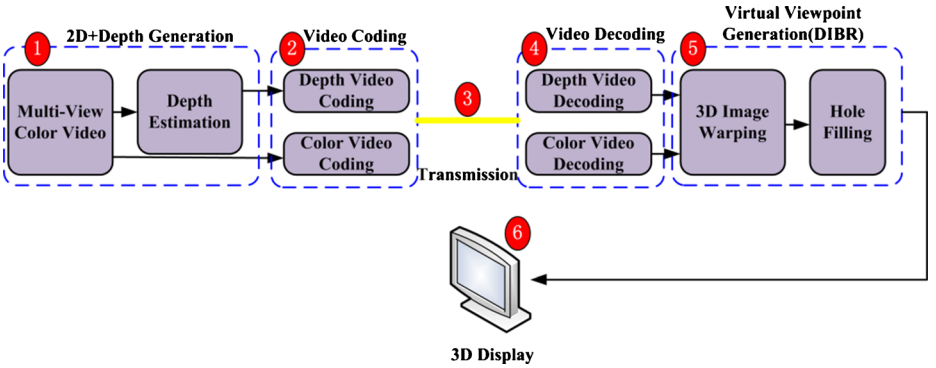


Fig. 1 The framework of 3D video system

analyzed. Figure 2 illustrates synthesized images from different depth videos, where Fig. 2(a) is rendered from original depth map, Fig. 2(b) to (e) are rendered from decoded depth maps which are coded with different Quantization Parameters (QPs). It is found that depth distortion will have important impacts on synthesized image quality compared with the benchmark (Fig. 2(a)). When *QP* increases, more distortion will be introduced in synthesized images, especially around the boundary of objects. From these synthesized images shown in Fig. 2, we can find that it is necessary to eliminate the view synthesis distortion for further improving the quality of synthesized images.

To reduce synthesized image quality degradation caused by depth distortion, Yuan [15] proposed the framework of view synthesis distortion elimination by extensively transmitting filter coefficient bits. In his framework, Wiener filter is adopted to eliminate view synthesis distortion under the assumption that compression distortion is a stationary White Noise. Figure 3 shows the coding diagram with view synthesis distortion elimination. In the scheme, Wiener filter coefficients are calculated frame by frame and transmitted to client for enhancing the quality of virtual view. Synthesized image quality is improved at the cost of transmitting filter coefficients. However, the number of filter coefficients is large. In this paper, according to Yuan's framework shown in Fig. 3 [15], we proposed an adaptive parameter determination scheme with the Wiener filter. Then, a good trade-off between bit rate and view synthesis distortion has been achieved by considering the spatial-temporal correlations of 3D video sequence.

3 Proposed view synthesis distortion elimination filter for depth coding

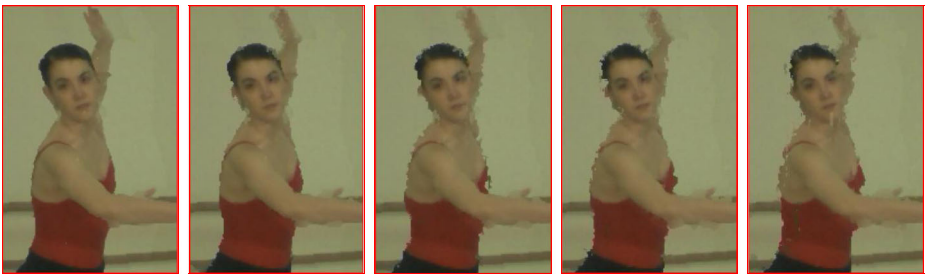
3.1 Filter coefficients determination

Let $\mathbf{I}_o$ and $\mathbf{I}_d$ represent synthesized images rendered by original depth video and compressed depth video, respectively. For the pixel located at (x,y), $I_d(x,y)$ and $I_o(x,y)$ has the relationship as

$$I_d(x,y) = I_o(x,y) + n(x,y), \tag{4}$$

where n represents rendering noise caused by quantization error of depth video coding. Based on the theory of image restoring, the distorted synthesized image can be restored by a Wiener filter denoted as $\mathbf{C}$ (with window size w). Suppose $\mathbf{I}_r$ represent restored image of $\mathbf{I}_d$, and then the restoration model can be described as

$$\mathbf{I}_r = \mathbf{C} \otimes \mathbf{I}_d, \tag{5}$$



(a) Original (b) QP=24 (c) QP=28 (d) QP=32 (e)QP=36

Fig. 2 Comparisons on synthesized images from different coding distorted depth maps

where '⊗' denotes convolution operator. As shown in Fig. 3, the statistical information of $\mathbf{I}_o$ and $\mathbf{I}_d$ are available at the encoder, then the Wiener filter coefficient matrix $\mathbf{C}$ can be determined by minimizing the least square error between $\mathbf{I}_o$ and $\mathbf{I}_r$, which is

$$\mathbf{C} = \underset{C}{\operatorname{argmin}} E\left(\left|\left|\mathbf{I}_r - \mathbf{I}_o\right|\right|^2\right), \tag{6}$$

where $E()$ is the expectation operator and $\|.\|$ is the L2-Norm. The determined coefficients of restore filter will be transmitted to client as auxiliary information.

3.2 Spatial-temporal correlations of 3D video sequence

In [15], each pixel in one frame is restored by a single filter coefficient matrix, which is a non-adaptive case. Based on image restoration model, each pixel can be restored adaptively by using a particular filter matrix with high restoration quality in the extreme condition. At the pixel level, the Wiener filter coefficient matrix corresponds to a scalar parameter, where the auxiliary information is very huge. To make a balance between restoration quality and auxiliary information (filter coefficient coding bit rate), spatial-temporal correlations of multiview video are analyzed in the following section. For easy illustration, an example that described the spatial-temporal correlations existing in video sequence is provided in Fig. 4. As observed in Fig. 4, the texture information in the red rectangle block is quite similar to its neighbor rectangle due to spatial correlation. On the other hand, the texture information in the blue rectangle block is quite similar in that of temporal successive frames due to temporal correlation [16]. Since color video is highly correlated, the filter coefficients that obtained by Eq. 6 are also highly correlated. To verify this assumption, the correlation coefficient ρ_i between Wiener filter coefficients $\mathbf{X}^i$ and $\mathbf{Y}^i$ is collected based on the statistical method and the average correlation coefficient ρ is defined as,

$$\begin{cases} \rho = \dfrac{1}{B}\displaystyle\sum_{i=1}^{B} \rho_i \\ \rho_i = \left| \dfrac{E[X^i Y^i] - E[X^i]E[Y^i]}{\sqrt{D[X^i]}\sqrt{D[Y^i]}} \right| \end{cases}, \tag{7}$$

where $E[\cdot]$ and $D[\cdot]$ are the expectation and standard deviation operators, B is the number of divided regions for each frame, i ($i \in [1,B]$) represents the index of the divided regions. For easy illustration, the statistical correlation coefficient analysis for temporal correlation for Wiener filter

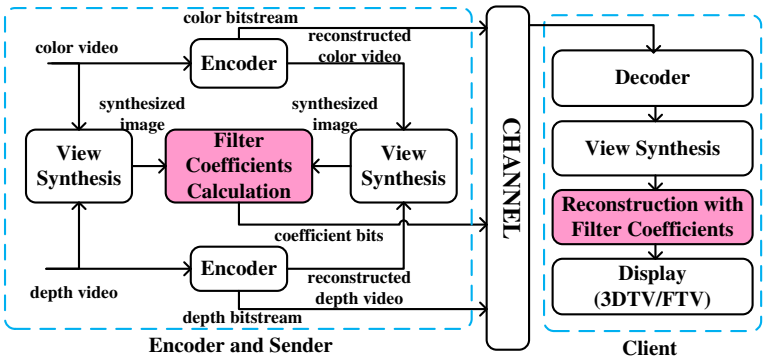


Fig. 3 Coding diagram of view synthesis distortion elimination [15]

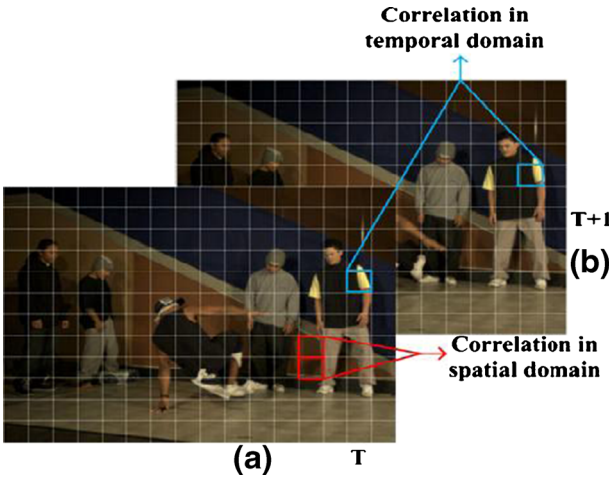


Fig. 4 Correlations in spatial-temporal domain

coefficients is provided in Fig. 5. As shown in Fig. 5(a), it is found that, the correlation coefficients are all larger than 0.8, and most of them are above 0.9 for different sequences and different frames. The correlation coefficient is usually high for slow motion sequence and it decreases as the motion being fast. For example, Breakdancers is an extreme fast motion sequence due to its low capturing frame rate and it maintains relative low correlation. Figure 5(b) shows the average value for different sequences and block partition, we can observe that the average correlations of different multiview video sequences are from 0.93 to 0.99 for different parameters, which are approaching to 1.0, especially for the slow motion sequences. Correlation coefficient approaching to 1 means that the filter coefficients are highly temporal correlated. In addition to the temporal correlation, there are still large correlations in the spatial and view domain [16] for the filter coefficients, which could be adopted to reduce filter coefficients.

Since video contents and their properties are different over regions, the Wiener filter coefficients may vary and not be consistently suitable for all these regions. Thus, frames can

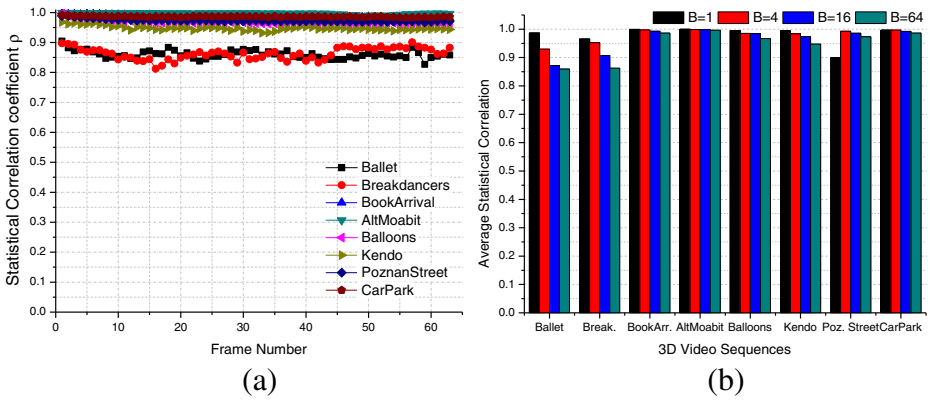


Fig. 5 Statistical correlation coefficient analyses for different sequences and parameters. **a** Statistical correlation coefficient variations along frames (B=64); **b** average statistical correlation coefficient

be divided into blocks and refined block division will lead to more precise filter coefficients. Meanwhile, since the frames along time direction are highly correlated, it has a probability of having similar filter coefficient matrix in temporal domain, which indicates that frames at different time can share the filter coefficient matrix. Therefore, to achieve a better trade-off between auxiliary filter coefficient bit rate and restoration quality, frames are divided into blocks to get the optimal coefficients. Also, the video sequence can be divided into groups of frames and coefficients can be shared for a group of frames instead of each frame.

3.3 Block-wise filter coefficients calculation based on temporal correlation

Since the basic unit for the filter might not be consistently optimal by using frame, the size of basic unit may influence the performance of algorithm. To determine the optimal basic unit, video sequence is divided into blocks. Two frames from reference and distortion video sequences at the same time form a Frame Couple (denoted as FC). A group of successive FCs is defined as a GROUP, and let G denote the number of frames in a group. In a FC, reference and distorted frames can be divided into several blocks, and two blocks with the same location are defined as a Block Couple which is shorten as BC, and let B denote the number of BCs in a FC. The BC is the smallest unit to calculate filter coefficient matrix in the proposed scheme. An example of video sequence decomposition is illustrated in Fig. 6.

After video sequence decomposition, taking BC as a basic unit, Eq. 6 becomes

$$\mathbf{C}^{*}_{k,j,i} = \arg\min_{C_{k,j,i}} E\left(\left|\left|\mathbf{I}_{r}^{k,j,i} - \mathbf{I}_{o}^{k,j,i}\right|\right|^{2}\right), \tag{8}$$

where i ($i \in [1, B]$) represents the index of a BC in the j-th ($j \in [1, G]$) FC of the k-th GROUP, $\mathbf{I}_{r}^{k,j,i}$ and $\mathbf{I}_{o}^{k,j,i}$ are restored block and original reference block in a BC. Then each GROUP will obtain B×G filter coefficient matrixes $\mathbf{C}^{*}_{k,j,i}$. Since these coefficients are highly correlated in the temporal color frames, there is also a high correlation between the filter coefficients for different frames as shown in Fig. 5. Thus, after getting filter coefficients matrix of each BC in a GROUP, the data of auxiliary information can be further reduced based on the temporal correlation by following equation

$$c^{*}_{k,i} = \frac{1}{\sum_{j=1}^{G} \omega_j} \times \sum_{j=1}^{G} \omega_j c^{*}_{k,j,i}, \tag{9}$$

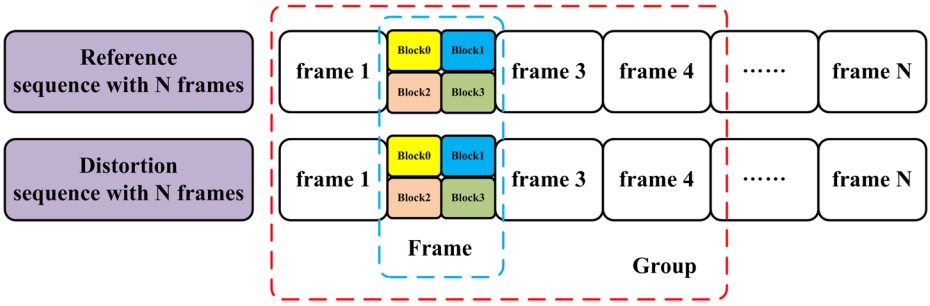


Fig. 6 Video sequence decomposition schematic

where $c^*_{k,j}$ represents the filter coefficient matrix of the i-th BC in the k-th GROUP. In other words, only B filter matrixes (auxiliary information) need to be transmitted to the client for each GROUP.

3.4 Parameters determination

Before estimating filter coefficient matrix, the parameters w (window size of filter), B (number of BCs in a FC) and G (number of FCs in a GROUP) should be determined. To determine the optimal parameters, statistical coding experiments are conducted. Recent video coding reference software JMVC 8.3 [2] is adopted. Depth maps are encoded and QP is set to 24. Total 32 frames are encoded for each sequence and each view. By using the decoded depth maps and original color images, virtual view image is generated by view synthesis reference software VSRS 3.5 [12]. Meanwhile, the restoration is performed to improve the virtual view image quality via given additional filter coefficients. For easy illustration, the relationship between parameters and visual quality of multiview video sequences "AltMoabit" and "PoznanStreet" are provided as examples and they are shown in Figs. 7 and 8, respectively. The image quality of virtual view from distorted image is measured with Peak Signal-to-Noise Ratio (PSNR) against the virtual view image generated by original depth map and original color image. The scattered symbols are real collected data and solid lines are the fitting results of the data.

Figures 7(a) and 8(a) show the relationship between PSNR and window size (w), which is obtained when G and B are fixed. It is observed that the PSNR value exponentially increases as the window size w increases and the PSNR values are also the same when w is larger than 5. Since the bit cost of transmitting filter coefficients will increase as w increases, w is set as 3 or 5 to have a good tradeoff between the bit size and synthesized image quality. Figures 7(b) and 8(b) show the relationship between B and PSNR value for the two sequences. We can observe that the PSNR value is in linear relationship with $\sqrt{B}$, thus, we can model the PSNR value and $\sqrt{B}$ as

$$PSNR = a \cdot B^{1/2} + \beta, \tag{10}$$

where α and β are model coefficients. Figures 7(c) and 8(c) show the relationship between G and the PSNR value of virtual view image. We find that PSNR value decreases a little bit (almost the same) when G increases. In terms of the filter coefficients, they reduce as G increases, namely

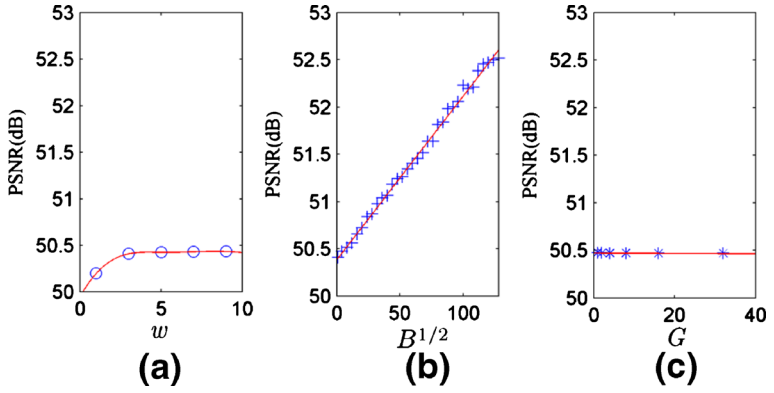


Fig. 7 Relationship between parameters and synthesized image quality for "AltMoabit" sequence. **a** PSNR versus w when G=4, B=1. **b** PSNR versus B when G=4, w=3. **c** PSNR versus G when w=3, B=16

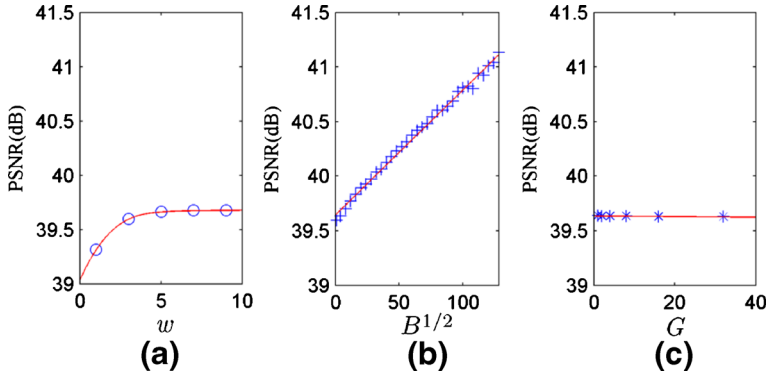


Fig. 8 Relationship between parameters and synthesized image quality for "PoznanStreet" sequence. **a** PSNR versus w when G=4, B=1. **b** PSNR versus B when G=4, w=3. **c** PSNR versus G when w=3, B=16

large G is preferred. However, in case of fast motion and large video scene changing, small G is preferred. Therefore, G is usually not large than 32.

For a video sequence, it contains N frames, the total number of filter coefficients can be calculated as

$$M = w^2 \times B \times \frac{N}{G}, \tag{11}$$

Based on extensive experiments, we find that the number of bits for encoding filter coefficients (auxiliary information) is proportional to the number of filter coefficients. Then, the total bits denoted as R_{tot} can be described as the sum of encoding bits of the video sequences and compressed filtering coefficients, which is

$$R_{tot} = R + cM, \tag{12}$$

where c is the compression ratio of filter coefficients and approximates to 0.008 by using lossless compression tools. To control the trade-off between bit and visual quality, the following cost function J is defined as

$$J = D + \lambda \times R_{tot}, \tag{13}$$

where λ is the Lagrangian multiplier and D denotes the distortion of the synthesized image quality measured by Mean Squared Error (MSE). To minimize the cost function J, its derivative is set to zero [16],

$$\frac{\partial J}{\partial R_{tot}} = \frac{\partial D}{\partial R_{tot}} + \lambda \equiv 0. \tag{14}$$

Based on the relationship between MSE and PSNR, Eq. 10 can be rewritten as

$$D = c_1 \times 10^{-\frac{\alpha\sqrt{B}+\beta}{10}}, \tag{16}$$

where c_1=255^2WH, W and H are the width and height of the image. According to Eqs. 11 and 12, we have the following

$$R_{tot} = R + c\frac{N}{G}w^2 B. \tag{17}$$

Substituting Eqs.16 and 17 into Eq.14 and solving it, we can obtain

$$\lambda = c_2 \alpha 10^{-\frac{\alpha\sqrt{B}+\beta}{10}} / \sqrt{B}, \tag{18}$$

where c_2 is a constant, $c_2 = \frac{255^2 WHG\ln(10)}{20cNw^2}$. The optimal B can be determined by solving the following Equation

$$B^* = \arg \min_B J. \tag{19}$$

The parameters of α and β can be estimated by pre-processing selected frames in the video sequence with changing the value of B. Table 1 and Table 2 provide the estimated parameters of part test video sequences and different QPs.

3.5 The proposed overall algorithm

Compared with Yuan's method [15], the local filter coefficients are utilized instead of global filter coefficients in our proposed method. At the same time, in order to achieve a great trade-off between bit rate and view synthesis distortion, we design an adaptive parameter determination scheme. In addition, the correlation in temporal is considered to reduce the extra bit rate.

The steps of the proposed overall algorithm are summarized as follows.

Step 1: First, the view synthesis is done with the original color video and depth video, and the result of view synthesis is denoted as reference sequence. Second, the original depth video is encoded with the multiview video coding (MVC) codec. The virtual view denoted as distortion sequence is synthesized by the reconstructed depth video and original color video. Then a FC is formulated by reference and distortion sequences as shown in Fig. 6.

Step 2: The parameters determination is implemented as described in Section 3.4. For the adaptive block scheme, if the current frame is the first frame of a GROUP, the model parameters, α and β, are determined by Eq. 10, then the optimal value of B is determined by Eq. 19. For the fixed block scheme, fixed B is used.

Step 3: The filter coefficients matrix $\mathbf{C}^*_{k,j,i}$ of each BC in the current FC is calculated by Eq. 8 and stored. If it is the last FC in a GROUP, go to **Step 4**; otherwise, go to **Step 1** for next FC.

Step 4: Filter coefficients $\mathbf{C}^*_{k,i}$ is calculated based on Eq. 9 and these filter coefficients are encoded with lossless coding method. Go to **Step 1** for next GROUP."

At the client, the received filter coefficient matrix $\mathbf{C}^*_{k,i}$ could be used as a post restoration filter for the synthesized image to improve the image quality.

4 Experimental results and analyses

To verify the effectiveness of the proposed schemes for eliminating rendering distortion, Yuan's scheme is compared with the proposed schemes. Eight standard 3D video sequences,

Table 1 Parameters of test sequence (QP=24)

Sequence	Resolution	α	β	$\sqrt{B}$
Breakdancers	1,024×768	0.05151	45.680	15
AltMoabit	1,024×768	0.03213	48.879	9
Balloons	1,024×768	0.04399	43.343	11
PoznanStreet	1,920×1,088	0.02232	38.775	11

Table 2 Parameters of Balloons sequence

QP	Resolution	α	β	$\sqrt{B}$
24	1,024×768	0.04399	43.343	11
28	1,024×768	0.04824	42.715	8
32	1,024×768	0.04926	41.927	6
36	1,024×768	0.05019	41.073	4

including Ballet, Breakdancers [20], BookArrival, AltMoabit [4], Balloons, Kendo [7], PoznanStreet and PoznanCarpark [3], with different resolution, camera setting and video properties, are adopted for the coding and rendering process. The detailed information of these test sequences and rendered views is listed in Table 3. In Table 3, the values in the Camera Array column are the camera setting properties, where 1D-Arc denoting convergence and 1D-parallel denoting parallel camera setting. In addition, Reference and Rendered Views are the index of viewpoint in multiview sequences which are used to be referenced and rendered. Detailed information of the coding, rendering and filtering parameters is listed in Table 4. The recent multiview video coding and view synthesis reference software, JMVC 8.3 and VSRS 3.5 are adopted for depth video coding and view synthesis. In the process of depth coding and view synthesis, only two viewpoints and 64 frames per viewpoint are encoded and rendered. According to the analysis of Section 3.4, the parameters of G and w are set as 32 and 5 in the experiment, respectively. For each sequence, four different *QP*s (24, 28, 32 and 36) are tested and four schemes (original JMVC, Yuan [15] and two proposed schemes) are compared. The traditional encoding scheme using JMVC, denoted as "original", and Yuan's scheme are employed as benchmark schemes. Due to different block division settings, it derives two proposed schemes, which are fixed block scheme and adaptive block Scheme. B is set as 64 for the fixed block scheme. Note that only the depth video is encoded and original color video is used in the rendering process for both the benchmarks and proposed schemes. The rate-distortion (R-D) performance in terms of depth bit rate plus auxiliary information and the PSNR value of luminance component of synthesized image, i.e. Y-PSNR of synthesized image, is evaluated for different schemes. The BDPSNR (dB) and BDBR (%) [1] are employed in the experiments for comparison.

The R-D results for these schemes are summarized in Table 5. The positive BDPSNR results and the negative BDBR results in the table indicate that R-D performance of the

Table 3 Parameters for the multiview test sequences

Sequences	Provider	Resolution	Camera array	Reference views	Rendered views
Ballet	MSR [20]	1,024×768	1D-Arc	3,5	4
Breakdanceres		1,024×768	1D-Arc	4,6	5
Bookarrival	HHI [4]	1,024×768	1D-Parallel	8,10	9
AltMoabit		1,024×768	1D-Parallel	8,10	9
Balloons	Nagoya Univ. [7]	1,024×768	1D-Parallel	3,5	4
Kendo		1,024×768	1D-Parallel	3,5	4
PoznanStreet	Poznan Univ. [3]	1,920×1,088	1D-Parallel	3,5	4
PoznanCarpark		1,920×1,088	1D-Parallel	3,5	4

Table 4 Coding, rendering and filtering parameters

Coding parameter	
Coding platform	JMVC 8.3
Basis *QP*s	24,28,32,36
Coded views	2
Coded frames in each view	64
Filter parameters	
Number of frames in a GROUP, G	32
Number of blocks in a frame, B	64 for the fixed block scheme, adaptive for the adaptive scheme
Window size w	5
Coding for the filter parameter	Lossless coding
Rendering parameters	
Rendering software	VSRS 3.5
Rendering precision	Half pixel
Filter	(Bi)-Cubic
Viewblending	0

corresponding scheme is better than that of the original JMVC. As shown in Table 5, Yuan's scheme achieves 0.55 dB BDPSNR gain on average or reduces the BDBR 33.16 % on average. For the proposed fixed block scheme, it improves the BDBPSNR from 0.55 dB to 1.08 dB, 0.81 dB on average when compared to the original coding scheme. For the adaptive scheme, it achieves the BDPSNR gain from 0.43 dB to 1.09 dB, 0.80 dB on average. Compared with Yuan's scheme, the proposed two algorithms achieve BDPSNR gain from 0.11 dB to 0.54 dB more. While evaluated with BDBR, the two proposed schemes achieve 46.21 % and 48.33 % bit reduction on average, respectively, while compared with the original coding scheme. It means 13.05 % and 15.17 % more bit reduction are achieved by the proposed algorithms when compared with Yuan's scheme. The adaptive block scheme is the best one for most sequences, but for some sequences, the adaptive block scheme is a little bit inferior to the fixed block scheme, such as Kendo. The reason is that frames may not be evenly divided into integral blocks at the boundary for different resolutions, and the marginal area may cost additional bits for the adaptive block scheme. However, the fixed block scheme usually does not have this problem in this aspect. For better observation, R-D curves comparison for eight different sequences are shown in Fig. 9. The proposed adaptive and fixed block schemes achieve better R-D performance as compared with the benchmarks.

In addition to the R-D performance, the subjective quality of synthesized images is also compared. Figure 10 shows the subjective image comparison among synthesized images generated by different schemes. The upper row is for synthesized images and the bottom row shows enlarged regions of the synthesized images. The image shown in Fig. 10(a) is rendered from original depth and original color videos. Figure 10(b) is rendered from original color video and compressed depth video, where QP is 28. Figure 10(c) and (d) are restored images by using Yuan's scheme and the proposed scheme. From Fig. 10(b), we can observe that the distortion around the girl's boundary is quite large while it is generated by original JMVC with QP being 28. Yuan's scheme can generally improve the subjective quality of the synthesized image. However, some

Table 5 R-D performance comparison among different schemes

Sequence	Scheme	Original JMVC		Yuan's scheme		Proposed fixed block scheme (B=64)		Proposed adaptive block scheme	
	QP	Bit rate (kbit/s)	PSNR (dB)	Bit rate (kbit/s)	PSNR (dB)	Bit rate (kbit/s)	PSNR (dB)	Bit rate (kbit/s)	PSNR (dB)
Ballet	24	717.07	40.34	717.18	41.11	727.68	41.34	739.89	41.46
	28	488.25	39.13	488.36	40.01	498.97	40.27	503.14	40.34
	32	322.93	38.03	323.03	38.93	333.80	39.22	333.80	39.22
	36	202.35	36.93	202.45	37.83	213.28	38.19	208.84	38.15
BDBR(%)/BDPSNR(dB)				−34.58	0.87	−41.57	1.08	−42.44	1.12
Breakdancers	24	854.02	43.83	854.13	44.39	864.30	44.58	883.68	44.61
	28	546.51	42.48	546.62	43.08	556.97	43.30	563.90	43.32
	32	326.21	41.23	326.31	41.83	336.78	42.10	335.65	42.08
	36	191.31	40.02	191.42	40.64	202.02	40.86	197.65	40.82
BDBR(%)/BDPSNR(dB)				−25.65	0.60	−32.08	0.77	−31.24	0.75
Bookarrival	24	650.58	48.58	650.69	48.72	659.68	49.16	667.20	49.16
	28	419.93	46.63	420.04	46.80	429.34	47.23	431.83	47.25
	32	260.47	45.20	260.58	45.37	270.18	45.87	267.15	45.82
	36	161.27	44.11	161.38	44.26	171.20	44.81	166.31	44.70
BDBR(%)/BDPSNR(dB)				−5.56	0.16	−17.40	0.55	−14.84	0.45
Alt Moabit	24	674.86	50.18	674.97	50.33	683.63	50.49	691.17	50.51
	28	457.45	48.14	457.55	48.22	466.47	48.66	469.05	48.68
	32	304.61	46.39	304.71	46.46	313.75	47.26	312.99	47.23
	36	202.66	44.99	202.76	45.04	211.95	46.29	208.68	46.20
BDBR(%)/BDPSNR(dB)				−2.09	0.08	−14.42	0.62	−11.71	0.53
Balloons	24	613.74	42.63	613.85	43.50	623.94	43.75	633.53	43.83
	28	367.21	41.99	367.32	42.83	377.56	43.07	378.82	43.08
	32	215.47	41.14	215.58	41.96	226.09	42.23	222.79	42.18
	36	128.14	40.62	128.24	41.35	138.88	41.62	131.97	41.50
BDBR(%)/BDPSNR(dB)				−82.07	0.82	−111.89	1.03	−128.38	1.07
Kendo	24	608.20	45.11	608.30	45.89	617.60	46.22	634.81	46.22
	28	402.27	44.40	402.38	45.15	411.80	45.49	416.57	45.50
	32	260.28	43.73	260.39	44.37	270.19	44.73	269.10	44.68
	36	171.82	43.18	171.93	43.74	181.70	44.11	177.67	44.05
BDBR(%)/BDPSNR(dB)				−48.28	0.69	−68.62	0.99	−66.97	0.97
Poznan Street	24	620.89	39.41	621.00	39.70	631.82	39.85	639.47	39.90
	28	354.36	38.80	354.47	39.08	365.46	39.26	364.17	39.25
	32	203.44	38.15	203.55	38.45	214.59	38.62	208.99	38.57
	36	120.89	37.41	121.00	37.69	132.14	37.92	124.81	37.83
BDBR(%)/BDPSNR(dB)				−27.30	0.29	−39.66	0.41	−43.59	0.45
Poznan Carpark	24	2368.67	37.64	2368.80	38.44	2379.48	38.53	2442.58	38.66
	28	1769.09	36.88	1769.22	37.76	1779.99	37.84	1824.32	37.98
	32	1262.03	36.06	1262.16	36.98	1273.10	37.09	1301.66	37.22
	36	840.19	35.10	840.33	35.99	851.41	36.30	866.57	36.47
BDBR(%)/BDPSNR(dB)				−39.77	0.89	−44.06	1.00	−47.49	1.09
Average BDBR(%)/BDPSNR(dB)				−33.16	0.55	−46.21	0.81	−48.33	0.80

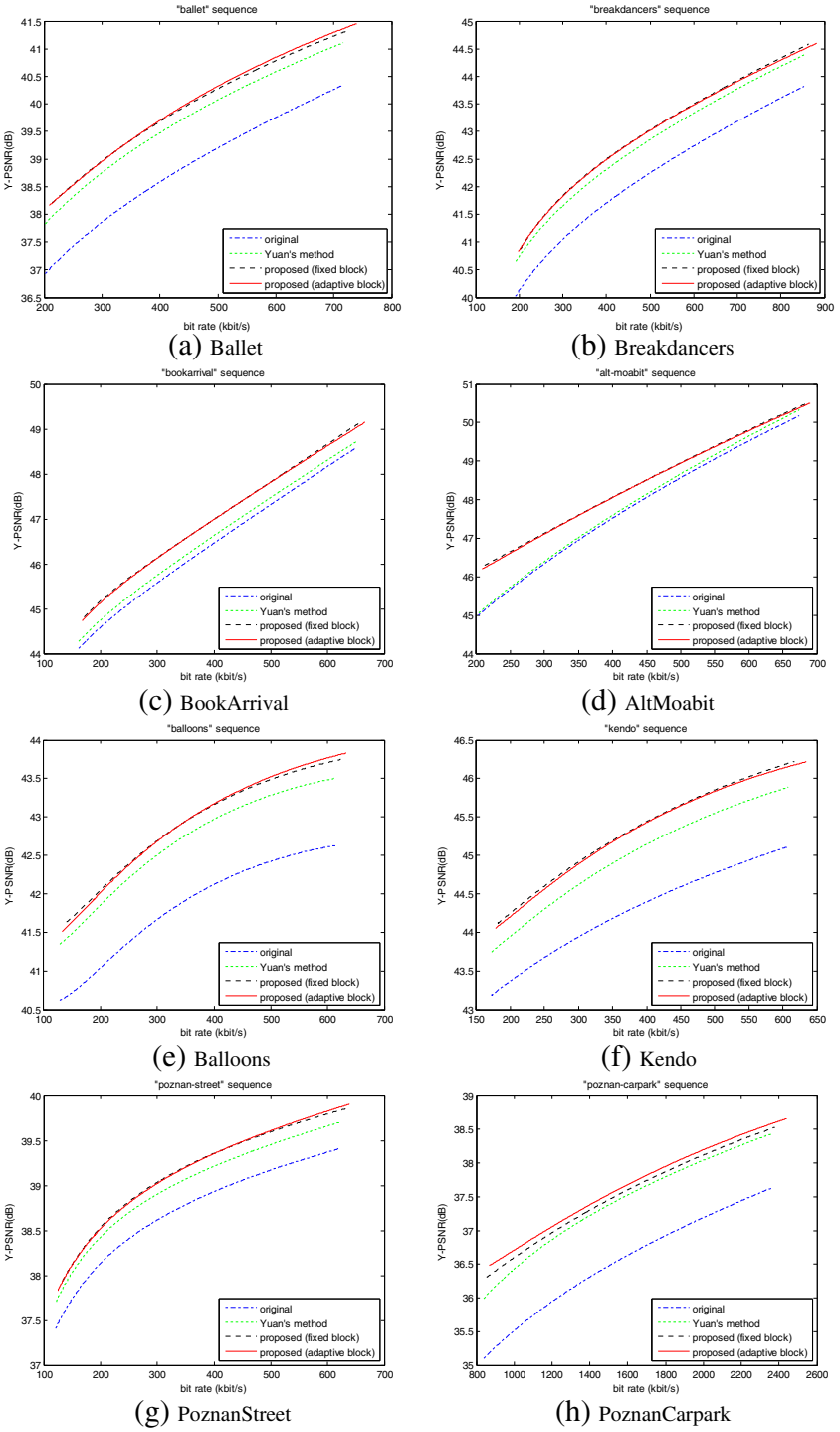


Fig. 9 R-D curves comparison among JMVC, Yuan's scheme and the proposed schemes

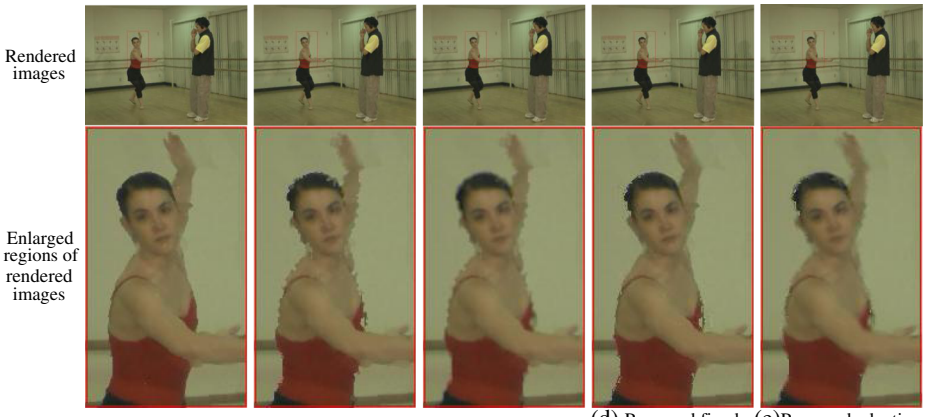


Fig. 10 Subjective image comparison among rendered images generated by different schemes. (Ballet)

blurring is also introduced. For the fixed block scheme, it has much better subjective and objective image quality than the benchmarks, especially for the object boundaries. Additionally, it is even better than the proposed adaptive block scheme for Ballet since it usually requires more bits for the filter coefficients. For the BookArrival sequence as shown in Fig. 11, the distortion was generated around the man's thumb in Fig. 11(b) when compared with the Fig. 11(a). The results of the proposed methods (Fig. 11(d) and (e)) are closer to the Fig. 11(a) compared with Fig. 11(b) and Yuan's method. Overall, the proposed schemes have much better subjective image quality for synthesized images, especially for the object boundary areas.

In addition to the R-D performance and image quality evaluation, we also analyze the computational complexity for the proposed algorithm. Tables 6 and 7 shows the time cost comparison at encoder and decoder sides, respectively. In this experiment, the depth videos of every sequence are encoded when QP is 28, the number of frames used in encoding and view synthesis is 64, and the proposed method uses the adaptive block scheme for parameter determination. It should be noticed that only the depth videos are encoded for every test sequences.

Let T_e and T_v be the time cost of depth encoding, view synthesis, respectively. $T_f(\varphi)$ stands for the time cost of filter coefficients calculation for scheme φ, $\varphi \in$ {Yuan's scheme, the proposed

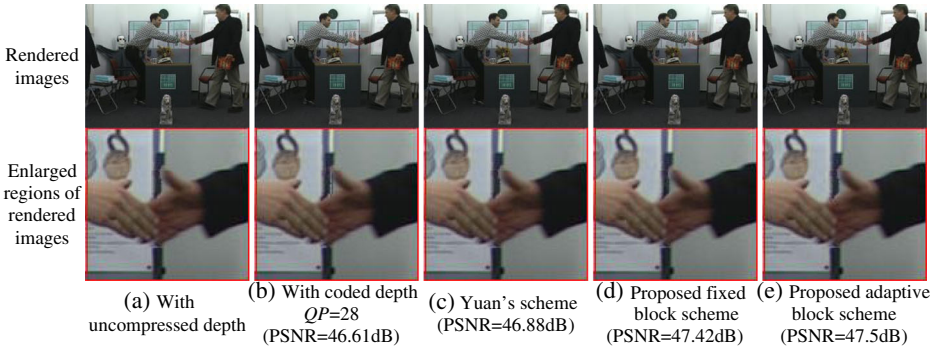


Fig. 11 Subjective image comparison among rendered images generated by different schemes. (BookArrival)

Table 6 Time cost comparison at encoder side (QP=28)

Sequence	Original time cost	Added time cost					
	T_e (s)	$2T_v$ (s)	Yuan's scheme		Proposed scheme		
			$T_f(\varphi)$ (s)	ΔT_{ENC} (%)	$T_f(\varphi)$ (s)		ΔT_{ENC} (%)
					Parameter estimation	Coefficient calculation	
Ballet	3605.3	628.6	188.7	22.66	156.3	36.8	22.79
Breakdancers	5650.7	588.1	192.1	13.80	156.4	42.6	13.92
BookArrival	3053.7	415.7	183.4	19.61	155.8	32.7	19.78
Altmoabit	2454.6	381	187.7	23.16	157.3	82.6	25.29
Balloons	3005.7	468.7	185.6	21.76	156.1	106.5	24.33
Kendo	4474.5	418	184.8	13.47	156.4	42.3	13.78
PoznanCarpark	5541.6	1,304	412.1	30.96	792.9	78.7	39.25
PoznanStreet	5964.7	1098.8	421.3	25.48	997.7	483.9	43.26
Average	4218.9	662.9	244.5	21.36	341.1	113.3	25.30

scheme}. In the proposed scheme, the time cost of filter coefficients calculation contains two parts, namely, the time cost of parameters estimation and the time cost of coefficients calculation. According to the framework shown in Fig. 3, compared with the original depth encoding, two view synthesis modules and filter coefficients calculation module are added at the encoder side. Therefore, the total time of the encoder side is $T_e+2T_v+T_f(\varphi)$. Thus, the ratio of added time cost to the original encoding time at the encoder side can be calculated as

$$\Delta T_{ENC} = \frac{2T_v + T_f(\varphi)}{T_e} \times 100\%. \tag{20}$$

Table 6 shows the time cost of every parts at the encoder side, including $T_e, T_v, T_f(\varphi)$. In Table 6, we can observe that Yuan's scheme increases the complexity about 21.36 % on

Table 7 Time cost comparison at the decoder side (QP=28)

Sequence	Original time cost		Added time cost			
			Yuan's scheme		Proposed scheme	
	T_d (s)	T_v (s)	$T_r(\varphi)$ (s)	ΔT_{DEC} (%)	$T_r(\varphi)$ (s)	ΔT_{DEC} (%)
Ballet	42.0	303.2	31.6	9.15	21.6	6.25
Breakdancers	45.3	284.8	31.8	9.63	26.1	7.90
Bookarrival	41.5	207.5	32.3	12.97	23.9	9.59
Altmoabit	41.7	200.4	31.9	13.17	25.7	10.61
Balloons	41.9	235.6	31.9	11.49	25.4	9.15
Kendo	43.5	208.6	32.4	12.85	26.3	10.43
PoznanCarpark	113.	685.2	88.5	11.08	68.3	8.55
PoznanStreet	114.8	526.4	88.5	13.80	67.9	10.58
Average	60.46	331.5	46.11	11.77	35.65	9.10

average compared with the original encoding time. For the proposed scheme, it increases about 25.30 % on average, which is 3.96 % more than that of Yuan's scheme.

In addition to the complexity of the encoder side, the complexity at the decoder side is also analyzed. Let T_d stand for the time cost of video decoding. For the original decoder side of a 3D video system, it includes the video decoder and view synthesis module, which are supposed to render the view images once. Thus, the complexity is T_d+T_v. For the filtering based new framework as shown in Fig. 3, the total time cost of the decoder side is $T_d+T_v+T_r(\varphi)$, where $T_r(\varphi)$ is the added time cost for reconstruction filtering for scheme φ, $\varphi\in\{$Yuan's scheme, the proposed scheme$\}$. Then the ratio of added time cost to the original time cost at the decoder side can be calculated as:

$$\Delta T_{DEC} = \frac{T_r(\varphi)}{T_d + T_v} \times 100\%. \tag{21}$$

Table 7 shows the complexity analysis of the decoder side. In Table 7, we find that Yuan's scheme increases the time cost about 11.77 % on average and the proposed scheme increases the time cost about 9.10 %. Basically, the time costs of the two schemes are quite similar. According to the above experimental results and analysis, we found the proposed method adds the computational complexity at encoder and decoder sides compared with the original 3D video coding process. Compared with Yuan's scheme, the proposed scheme has similar complexity at both the encoder and decoder sides. However, the good thing is the proposed scheme can improve more RD performance.

5 Conclusions

In view synthesis of 3DV system, the quality of synthesized images is very sensitive to depth distortion introduced by depth coding. The depth distortion will lead to geometrical rendering position error, and seriously affects the quality of synthesized images. In this paper, we propose a new in-loop filter to minimize the rendering distortion at cost of transmitting extra filter coefficients as supplement information. A good trade-off between extra bit rate for filter parameters and view synthesis distortion has been achieved by considering the spatial property and temporal correlation of the 3DV sequence. Then, adaptive optimal parameter determination scheme is also presented. The simulation results show that the proposed scheme can significantly improve the depth coding efficiency as well as the quality of synthesized images.

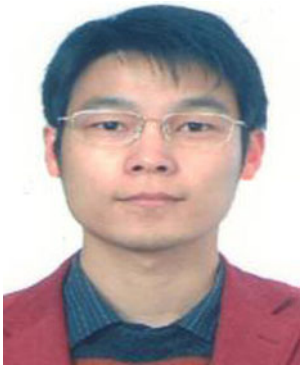
Acknowledgments This work was supported in part by the Natural Science Foundation of China under Grants 61102088 and 61272289, Shenzhen Emerging Industries of the Strategic Basic Research Project under Grant JCYJ20120617151719115, and the Guangdong Nature Science Foundation under Grant S2012010008457.

References

1. Bjontegaard G (2001) Calculation of Average PSNR Differences between RD-curves, ITU-T Video Coding Experts Group (VCEG), document M33, Austin, TX
2. Chen Y, Pandit P, Yea S, Lim CS (2009) Draft reference software for MVC, ITU-T JVT –AE207
3. Domański M, Grajek T, Klimaszewski K, Kurc M, Stankiewicz O, Stankowski J, Wegner K (2009) Poznań multiview video test sequences and camera parameters, ISO/IEC JTC1/SC29/WG11 MPEG 2009/M17050, Xian, China
4. Feldmann I, Mueller M, Zilly F, Tanger R, Mueller K, Smolic A, Kauff P, Wiegand T (2008) HHI test material for 3D Video, SO/IEC JTC1/SC29/WG11 MPEG2008/M15413, Archamps, France
5. Hu S, Kwong S, Zhang Y, Kuo C–CJ (2013) Rate-distortion optimized rate control for depth map based 3D video coding. IEEE Trans Image Process 22(2):585–594

6. Liu S, Lai P, Tian D, Chen CW (2011) New depth coding techniques with utilization of corresponding video. IEEE Trans Broadcast 57(2):551–561
7. Nagoya University FTV test sequences. [Online]. Available: http://www.tanimoto.nuee.nagoya-u.ac.jp/
8. Nguyen V, Min D, Do MN (2013) Efficient techniques for depth video compression using weighted mode filtering. IEEE Trans Circ Syst Video Technol 23(2):189–202
9. Nya PN, Köppel M, Doshkov D, Lakshman H, Merkle P, Müller K, Wiegand T (2011) Depth Image-based rendering with advanced texture synthesis for 3D video. IEEE Trans Multimed 13:453–465
10. Oh K, Yea S, Vetro A, Ho Y (2009) Depth reconstruction filter and down/up sampling for depth coding in 3-D Video. IEEE Signal Process Lett 16(9):747–750
11. Oh K, Yea S, Vetro A, Ho Y (2011) Depth coding using a boundary reconstruction filter for 3-D video systems. IEEE Trans Circ Syst Video Technol 21(3):350–358
12. Tanimoto M, Fujii T, Suzuki K (2009) View Synthesis Algorithm in View Synthesis Reference Software 3.0 (VSRS 3.0), Tech. Rep. Document M16090, ISO/IEC JTC1/SC29/WG11
13. Tanimoto M, Tehrani MP, Fujii T, Yendo T (2012) FTV for 3-D spatial communication. Proc IEEE 100(4): 905–917
14. Yuan H, Chang Y, Huo J, Yang F, Lu Z (2011) Model-based joint bit allocation between texture videos and depth maps for 3-D video coding. IEEE Trans Circ Syst Video Technol 21(4):485–497
15. Yuan H, Liu J, Xu H, Li Z, Liu W (2012) Coding distortion elimination of virtual view synthesis for 3D video system: theoretical analyses and implementation. IEEE Trans Broadcast 58(4):558–568
16. Zhang Y, Jiang G, Yu M, Ho Y (2009) Adaptive multiview video coding scheme based on spatiotemporal correlation analyses. ETRI J 31(2):151–161
17. Zhang Y, Kwong S, Xu L, Hu S, Jiang G, Kuo C–CJ (2013) Regional bit allocation and rate distortion optimization for multiview depth video coding with view synthesis distortion model. IEEE Trans Image Process 22(9):3497–3512
18. Zhang L, Tam WJ (2005) Stereoscopic image generation based on depth images for 3DTV. IEEE Trans Broadcast 51(2):191–195
19. Zhao Y, Zhu C, Chen Z, Yu L (2011) Depth no-synthesis-error model for view synthesis in 3-D video. IEEE Trans Image Process 20(8):2221–2227
20. Zitnick CL, Kang SB, Uyttendaele M, Winder S, Szeliski R (2004) High-quality video view interpolation using a layered representation. ACM SIGGRAPH and ACM Trans. Graphics 23(3):600–608

Linwei Zhu received his B.S. degree in applied physics from Tianjin University of Technology, China, in 2010, and M.S. degree in Signal and Information Processing from Ningbo University, China, in 2013. From Sep. 2011 to Jun. 2013, he is a visiting student in Shenzhen Institutes of Advanced Technology (SIAT), Chinese Academy of Science (CAS). After graduation from Ningbo University, he joins the SIAT as a research assistant. His research interests mainly include depth image based rendering, depth estimation and video transcoding.



Yun Zhang received the B.S. and M.S. degrees in electrical engineering from Ningbo University, Ningbo, China, in 2004 and 2007, respectively, and the Ph.D. degree in computer science from Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS), Beijing, China, in 2010. From 2009 to 2010, he was a Visiting Scholar with the Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong. In 2010, he joined in the Shenzhen Institutes of Advanced Technology (SIAT), CAS, as an Assistant Professor. Since 2012, he serves as Associate Professor in SIAT, CAS. His research interests include 3D/Multi-view video coding and standardization, 3D video pre/post-processing.

Xu Wang received the B.S. degree from South China Normal University, Guangzhou, China, in 2007, and M.S. degree from Ningbo University, Ningbo, China in 2010. He is now working toward the Ph. D degree of the Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong. His current research interests include stereoscopic image/video quality assessment, bit allocation and rate control for 3D video.

Sam Kwong (M'93–SM'04) received the B.S. and M.S. degrees in electrical engineering from the State University of New York at Buffalo in 1983, the University of Waterloo, Waterloo, ON, Canada, in 1985, and the Ph.D. degree from the University of Hagen, Germany, in 1996. From 1985 to 1987, he was a Diagnostic Engineer with Control Data Canada. He joined Bell Northern Research Canada as a Member of Scientific Staff. In 1990, he became a Lecturer in the Department of Electronic Engineering, City University of Hong Kong, where he is currently a Professor in the Department of Computer Science. His research interests are video and image coding and evolutionary algorithms.