Computers and Electrical Engineering 40 (2014) 1344–1352

Contents lists available at ScienceDirect

Computers and Electrical Engineering

journal homepage: www.elsevier.com/locate/compeleceng

Efficient two-stage early SKIP mode termination for depth video coding ☆

Huanqiang Zeng [a,c,*], Yongtao Wang [b], Zhe Wei [c], Canhui Cai [a]

[a] School of Information Science and Technology, Huaqiao University, Xiamen, China
[b] Institute of Computer Science and Technology, Peking University, Beijing, China
[c] School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore

ARTICLE INFO

Article history:
Available online 31 October 2013

ABSTRACT

The *video plus depth* format has been widely-used in multi-view video systems. Therefore, low complexity depth video coding becomes very essential. For that, an efficient *two-stage early SKIP mode termination* (TESMT) algorithm is proposed in this paper. First, by using texture-depth correlation, our approach first checks whether current *macroblock* (MB) is motionless or slow-motion based on motion activity of corresponding region in texture video. If so, SKIP mode is selected as optimal mode and mode decision process is early terminated. Otherwise, our approach further checks whether the rate-distortion cost of SKIP mode is below an adaptive threshold, which is derived by exploiting spatial-temporal correlation between current MB and its adjacent MBs in depth video. Experimental results show that proposed algorithm significantly reduces computational complexity while keeping almost the same coding efficiency of depth video and quality of synthesized view, compared with exhaustive mode decision in multi-view video coding.

1. Introduction

The multi-view video systems, such as *three dimensional TV* (3DTV) and *free viewpoint TV* (FTV), have received more and more attentions, since these services offer the users a realistic scene with depth perception and free viewpoint selection [1,2]. Unlike the single-view video, the multi-view video is composed of more than one view of the observed scene, and these views are synchronously acquired by using multiple cameras from different viewpoints. Hence, one of the major challenges for the practical multi-view video systems is how to store and transmit the large amount of multi-view video. For this, an efficient data representation format for multi-view video, called *video plus depth*, has been standardized by the MPEG standard committee [3]. An example of this video plus depth format for the multi-view video sequence "Champagne_Tower" is shown in Fig. 1. This format allows to capture and transmit less viewpoint's data at the sender side and to synthesize the desired virtual views at the receiver side based on the neighboring texture video and depth video using the *depth-image-based rendering* (DIBR) technique [4]. Note that the depth video represents the depth information of video objects in the corresponding texture video and plays a key role in synthesizing the virtual view. Therefore, efficient video coding techniques for depth video are dispensable.

Multiple depth video coding methods can be found in the literatures [5,7,8]. Considering the sharp boundaries of video objects in depth video are crucial to the visual quality of the synthesized view and the traditional block-based codec (e.g., H.264/AVC) could not preserve them well, Merkle et al. [5] presented a platelet based depth map compression method. In this method, the quad-tree decomposition is used to divide the depth image into the variable-size blocks according to the

☆ Reviews processed and recommended for publication to Editor-in-Chief by Guest Editor Dr. Jing Tian.
* Corresponding author at: School of Information Science and Technology, Huaqiao University, Xiamen, China. Tel.: +86 595 22693016.
E-mail addresses: zeng0043@e.ntu.edu.sg (H. Zeng), wyt@pku.edu.cn (Y. Wang), weiz0002@e.ntu.edu.sg (Z. Wei), chcai@hqu.edu.cn (C. Cai).

0045-7906/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.compeleceng.2013.09.007

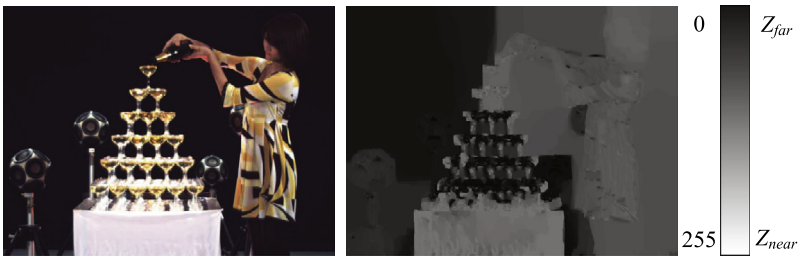

Fig. 1. An illustration of the *video plus depth* format (sequence "Champagne_Tower", the first frame of View 41), consisting of the texture video (left) and its associated depth video (right). Note that the depth range is normalized to the maximum Z_{far} and the minimum Z_{near} distance from the camera for the corresponding 3D points, corresponding to the pixel intensities 0 and 255, respectively.

boundaries and each block is approximated by one modeling function that contains one or two surfaces. However, this method yields a lower coding efficiency, since it does not make full use of the spatial and temporal correlations in depth video. Moreover, some depth video coding methods are developed based on *multi-view video coding* (MVC). The MVC is designed for high efficiency multi-view texture video coding and has been standardized as an amendment of the H.264/AVC–Annex H [6]. The depth video can be treated as a gray video and thus directly encoded using MVC. Among them, Liu et al. [7] presented a trilateral filter to remove the coding artifacts based on the pixel closeness and the similarity among pixels in depth video and texture video. Furthermore, a sparse dyadic mode is utilized to reconstruct the depth map by the sparse representations of depth blocks and derivation of edge information. Daribo et al. [8] shared one common *motion vector* (MV) field to reduce the amount of information that describes the motion of the texture video and the depth video so as to save the bit rate.

It should be pointed out that most of the multi-view video systems exploit the MVC to compress the multi-view video plus depth sequences due to better compatibility and higher coding efficiency. However, the consequent problem is the heavy computational complexity, because (1) Compressing the video plus depth format sequences nearly doubles the complexity, since the compression of depth video requires the similar time or resource to that of texture video; and (2) The MVC has tremendous computational complexity. Because it adopts many techniques to fully exploit the spatial and temporal correlations within a single view and the inter-view correlation between two neighboring views, such as various prediction modes, *hierarchical B picture* (HBP) prediction structure, and exhaustive mode decision. Therefore, how to reduce the computational complexity while maintaining almost the same video coding quality and the total bit rate has become the main concern of optimization techniques for any practical multi-view video plus depth codec realization.

In recent years, fast mode decision methods have been widely developed to reduce the computational complexity of MVC-based multi-view texture video coding [9–11]. But they can not be directly applied to MVC-based depth video coding, since the characteristics of depth video are different from that of texture video. For that, few fast mode decision methods for depth video coding are presented in [12–14]. Chiang et al. [12] utilized the Sobel edge detector to identify the edge and non-edge *macroblock*s (MBs) firstly. The obtained result is then used to skip those unlikely modes. Wang et al. [13] presented an early termination method based on the detection of difference. In this method, the differences between the current MB and the co-located MBs in the original frame and the reconstructed frame are computed respectively. If the difference is equal to zero, the current MB is considered as a static one and the SKIP mode is directly selected as the optimal mode to speed up the mode decision process. Zhang et al. [14] suggested a simple early SKIP mode, which derives the motion information of the current depth MB from its corresponding MB in texture video. This method significantly reduces the computational load at the expense of large loss of coding efficiency.

In this paper, an efficient mode decision method for depth video coding, called *two-stage early SKIP mode termination* (TESMT), is proposed based on the texture-depth and spatial-temporal correlations. In our approach, for the current MB in depth video, two early termination schemes are sequentially checked: (1) whether its motion activity is motionless or slow motion, which is inferred from that of the corresponding region in texture video; and (2) whether its *rate-distortion* (RD) cost of the SKIP mode is smaller than an adaptive threshold, which is derived by using the coding information of its spatial and temporal adjacent MBs. If any above-mentioned early termination condition is met, the SKIP mode will be selected as the optimal mode and the checking process of the remaining modes is skipped to save the computational complexity. Otherwise, the exhaustive mode decision is performed to find the optimal mode. Experimental results have shown that the proposed TESMT algorithm can greatly reduce the computational complexity while keeping almost the same coding efficiency of the depth video and the same quality of the synthesized view, compared with that of the exhaustive mode decision in MVC.

The rest of this paper is organized as follows. The overview of mode decision in MVC is described in Section 2. The proposed fast mode decision method, TESMT, is presented in Section 3 in detail. Extensive simulation results are documented and discussed in Section 4. Finally, conclusions are given in Section 5.

2. Overview of mode decision in MVC

First, the MVC adopts the HBP prediction structure [15] to efficiently explore the spatial, temporal and inter-view correlations inherited in multi-view video. Fig. 2 shows an example of the HBP prediction structure with 8 views and the length of

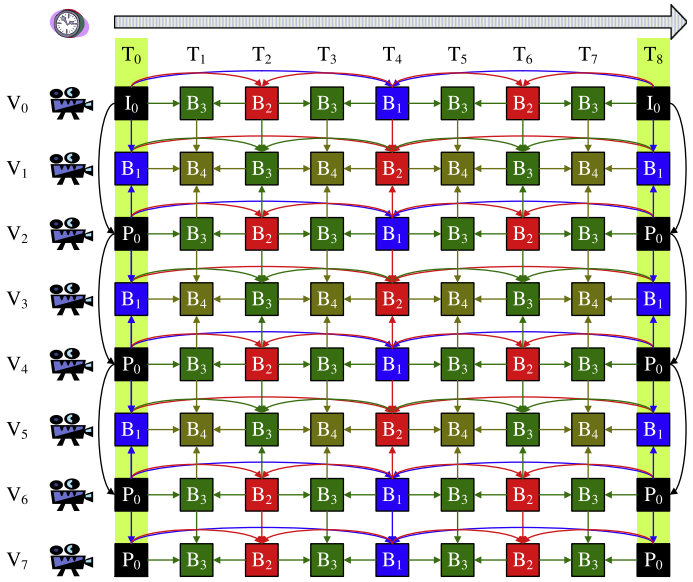

Fig. 2. An illustration of *Hierarchical B picture* (HBP) prediction structure in the MVC.

group of picture (GOP) = 8. In this HBP prediction structure, there are two kinds of pictures—the anchor pictures (i.e., the pictures at T_0 and T_8) and the non-anchor pictures (i.e., the pictures at T_i, for $i = 1, 2, \ldots, 7$). In addition, there are two kinds of views—the *main* views (namely, V_0, V_2, V_4 and V_6) and the *auxiliary* views (namely, V_1, V_3, V_5 and V_7). In the main views, the temporal prediction via *motion estimation* (ME) is performed for non-anchor pictures by referring the neighboring temporal pictures within the same view. In the auxiliary views, besides temporal prediction via ME, the inter-view prediction via *disparity estimation* (DE), which is a new feature of MVC, is further conducted by referring the neighboring pictures at the same time instant but from the neighboring views to improve the coding efficiency.

In addition to the HBP prediction structure, the MVC offers many prediction modes for effectively coding various video contents. To reduce the spatial redundancy, various intra prediction modes, including intra 4×4, intra 8×8 and intra 16×16 (jointly denoted as Intra in this work), are used [16]. To remove the temporal and inter-view redundancies, seven block sizes are adopted to conduct both ME and DE. As shown in Fig. 3, they are 16×16, 16×8, 8×16, 8×8, 8×4, 4×8 and 4×4, and the last four block sizes are jointly denoted as P8 $\times$ 8 in the MVC [17]. Among many prediction modes, the MVC exploits the exhaustive mode decision to select the optimal mode based on the criterion—Lagrangian *rate distortion optimization* (RDO) function [18]. More specifically, all the aforementioned modes require to be individually checked and the computed RD cost for each mode is compared to choose the mode with the minimum RD cost as the optimal mode. As a result, the computational complexity incurred by the exhaustive mode decision is extremely high, therefore, a fast mode decision algorithm is highly desirable for those real-time multi-view video plus depth applications.

3. Proposed two-stage early SKIP mode termination (TESMT)

3.1. Motivation

It is well-recognized that the SKIP mode in MVC is more suitable for coding the large homogeneous areas with no motion or a fairly slow motion [9,11]. And one can see from Fig. 1 that the depth video usually contains these scenes. Hence, it can be intuitively observed that the SKIP mode should be highly possible to be the optimal mode for depth video coding. To verify this intuition, we conduct extensive experiments to obtain the distribution of optimal mode by using the exhaustive mode

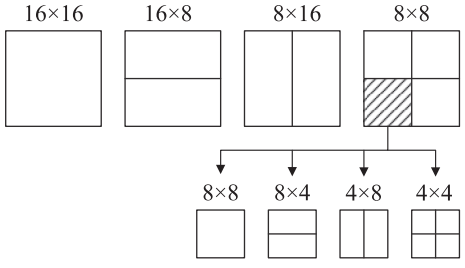

Fig. 3. An illustration of seven block sizes for ME and DE in the MVC.

decision in MVC on six depth video sequences, including "Alt_Moabit", "Book_Arrival", "Breakdancers", "Champagne_Tower", "Lovebird1" and "Newspaper" [19–21]. These depth video sequences are with various kinds of motion activities and generated by using *depth estimation reference software* (DERS) [22] developed by the MPEG's FTV standardization body. The experimental setup is described as follows: (1) 100 frames of each depth video sequence are encoded with GOP length = 8; (2) *Quantization parameter* (*QP*) = 24, 28,32 and 36; (3) HBP prediction structure is used; (4) RDO and *context-adaptive binary arithmetic coding* (CABAC) are enabled; and (5) the search range of the ME and DE is ±64. The corresponding statistical result averaged on these depth video sequences under different *QP* values is shown in Table 1.

From the results shown in Table 1, it can be clearly seen that the possibility that the SKIP mode is selected as the optimal mode is highest and increased with the *QP* values over various depth video sequences. On average, 78.07% MBs will select SKIP mode as the optimal mode while only 21.93% MBs will select the optimal mode from the remaining modes "16 × 16, 16 × 8, 8 × 16, P8 × 8, Intra". On the contrary, the checking process of the SKIP mode has very low complexity while that of the remaining modes are very time-consuming. Hence, if we can determine the SKIP mode as the optimal mode in advance, then the checking process of the remaining modes can be bypassed to save a lot of computational complexity. Therefore, it is very logical to develop an efficient early SKIP mode termination method for depth video coding to reduce the computational complexity.

3.2. Proposed method

The depth video reflects the depth information of video objects in the corresponding texture video and usually consists of large smooth regions with distinct boundary. Since similar image structure can be found between the texture and depth video, there exists a strong texture-depth correlation. Moreover, similar to the texture video, the depth video also contains a large amount of spatial and temporal correlations. By making full use of these correlations, we propose an efficient two-stage early SKIP mode termination method for depth video coding as follows.

In the proposed TESMT algorithm, we first utilize the texture-depth correlation to design the first-stage early SKIP mode termination scheme. Intuitively, without a scene cut or any camera movement, for those motionless or slow-motion regions in the texture video, it is expected that the corresponding depth values across the adjacent frames of the depth video should be constant or consistent. It means that the corresponding depth region has similar motion activity to that of texture video. Considering this texture-depth correlation, the motion activity of the current MB in depth video can be inferred from that of its corresponding MBs in the texture video. Moreover, the motion activity of MBs in the texture video can be determined in advance based on their optimal modes, because (1) the optimal mode of an MB can accurately reflect its motion activity [17]; and (2) the texture video has been coded before depth video and its coding information can be re-used. Based on these observations, the first-stage early SKIP mode termination scheme is designed as below. Fig. 4 shows the current MB in depth video (i.e., MB_0) and its corresponding region in texture video, where MB_1 is the MB with the same position as MB_0 in the corresponding texture frame and MB_i for $i = 2, 3, \ldots, 9$ are the 8-neighbors of MB_1.

1. Determine the *stationary flag* (i.e., SF_i) for each MB_i in the texture frame as illustrated in Fig. 4. If one of the following two conditions is met: (1) the optimal mode of the MB_i is SKIP mode; (2) the optimal mode of the MB_i is 16 × 16 and its MV (i.e., $mv_i = (x_i, y_i)$) length, $L_i = |x_i| + |y_i| \leqslant T_1$, the motion activity of this texture MB can be determined as motionless or slow motion, and SF_i is set to 1; Otherwise, SF_i is set to 0.
2. For the corresponding texture region, the *sum of stationary flag* (*SSF*) is computed as $SSF = \sum_{i=1}^{9} SF_i$. If $SSF \geqslant T_2$, the texture region is more likely to be homogeneous region with motionless or slow motion and thus the current MB in depth video, MB_0, tends to fall in a stationary region. Therefore, the SKIP mode is selected as the optimal mode and the mode decision process is early terminated.
   To determine the thresholds T_1 and T_2, extensive experiments have been conducted on a set of commonly-used depth video sequences, including "Alt_Moabit", "Book_Arrival", "Champagne_Tower" and "Newspaper". The test conditions are listed as follows: 100 frames of each depth video sequence are encoded using HBP prediction structure with GOP length = 8, *QP* = 24, 28, 32 and 36, RDO and CABAC entropy coding are enabled, the search range of the ME and DE is ±64. By using the exhaustive mode decision in MVC under the above-mentioned test conditions, we study the probability that the SKIP mode is selected as the optimal mode under different MV length L_i. We found that when L_i is less than or equal to 1, the SKIP mode has high percentage (i.e., almost 90%) to be the optimal mode, thus T_1 is set to 1. Similarly, T_2 is determined as 6 by investigating the probability that the SKIP mode is chosen as the optimal mode under different sum of stationary flag *SSF* values.

Table 1
Distribution of optimal mode in depth video coding (%).

QP	SKIP	16 × 16	16 × 8	8 × 16	P8 × 8	Intra
24	69.92	12.00	4.92	5.49	6.31	1.36
28	75.27	9.98	4.29	4.40	5.03	1.03
32	81.21	7.97	3.38	3.62	3.24	0.58
36	85.89	6.04	2.44	2.75	2.56	0.32
Average	78.07	8.99	3.76	4.07	4.29	0.82

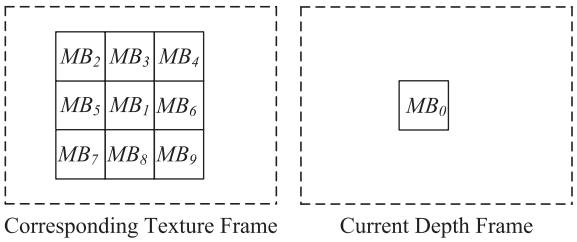

Fig. 4. The current MB, MB_0 and its texture-adjacent MBs.

If the above-mentioned first-stage early termination condition is not granted, we further exploit the spatial and temporal correlations in depth video to develop the second-stage early SKIP mode termination scheme. Fig. 5 shows the current MB in depth video, MB_0, and its spatial and temporal adjacent MBs, where MB_{10} is the MB with the same position as MB_0 in the previous coded depth frame. In this stage, for the current MB, the RD cost of the SKIP mode is computed to compare with an adaptive threshold. If this RD cost is smaller than the adaptive threshold, the SKIP mode is selected as the optimal mode and the checking process of the remaining modes is skipped. Considering the coding information (e.g., the optimal mode and the corresponding RD cost) of the current MB is highly correlated to that of its spatial and temporal adjacent MBs, this adaptive threshold is determined as the weighted average of the RD cost values of those adjacent MBs that choose the SKIP mode as the optimal mode:

$$T_{ST} = \frac{\sum_{i=10}^{13} B_i \cdot W_i \cdot RDcost(SKIP)_i}{\sum_{i=10}^{13} B_i \cdot W_i} \tag{1}$$

where $RDcost(SKIP)_i$ and W_i are the RD costs of the SKIP mode and the weights of the corresponding MB_i in Fig. 5, for $i = 10$, 11, 12, 13, respectively.

The weight W_i plays a key role in determining this adaptive threshold. It can be computed based on the intuition that the nearer the adjacent MB to the current MB, the larger the weight should be assigned. In other words, the weight W_i is inversely proportional to the distance (say, D_i) between the current MB (i.e., MB_0) and MB_i (where $i = 10$, 11, 12, 13):

$$W_i = \frac{1}{D_i} \tag{2}$$

To compute the distance D_i, the three-dimensional coordinate of MB_i shown in Fig. 5 is first defined based on its own spatial and temporal position as:

$$MB_i = (x_i, y_i, t_i) \tag{3}$$

where (x_i, y_i) and t_i denotes the spatial coordinates and the temporal index of MB_i respectively (where $i = 10$, 11, 12, 13). Then, considering the spatial and temporal correlations between the adjacent MBs and the current MB, the coordinate of each MB_i is empirically determined as:

$$\begin{aligned} MB_0 &= (0,0,0) \quad MB_{10} = (0,0,-1) \quad MB_{11} = (-1,0,0) \\ MB_{12} &= (0,-1,0) \quad MB_{13} = (1,-1,0) \end{aligned} \tag{4}$$

After that, we calculate the distance D_i between MB_i and MB_0 as:

$$D_i = \sqrt{(x_i - x_0)^2 + (y_i - y_0)^2 + (t_i - t_0)^2} \tag{5}$$

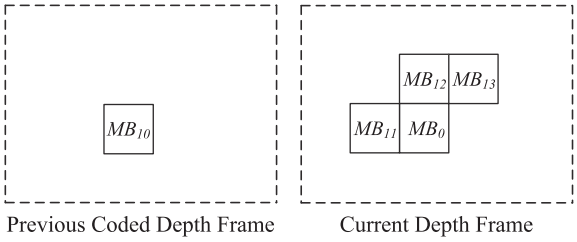

Fig. 5. The current MB, MB_0 and its spatial-temporal-adjacent MBs.

Table 2
The weights of the spatial and temporal adjacent MBs.

Index of MB i	10	11	12	13
W_i	0.27	0.27	0.27	0.19

Furthermore, it can be assumed that the sum of all the initial weights W_i is equal to 1, for $i = 10, 11, 12, 13$:

$$\sum_{i=10}^{13} W_i = 1 \tag{6}$$

By combining the Eqs. (2) and (6), we can calculate the values of the weights W_i and show them in Table 2.

Moreover, considering adjacent MB choosing non-SKIP mode as its optimal mode has little contribution to decide whether the SKIP mode is the optimal mode of the current MB, only the RD cost of those adjacent MBs in Fig. 5 that selected the SKIP mode as the optimal mode will be used. Hence, the B_i is defined as

$$B_i = \begin{cases} 1, & \text{if the optimal mode of } MB_i \text{ is SKIP mode;} \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

It should be pointed out that if all the B_i are equal to 0, the checking of the second-stage early termination scheme will be skipped for this special case.

In summary, the proposed TESMT algorithm consequently performs the above-mentioned first-stage and second-stage early termination schemes and its flowchart is presented in Fig. 6. If any early termination condition provided in these two stages is met, the SKIP mode is selected as the optimal mode and the checking process of the remaining modes

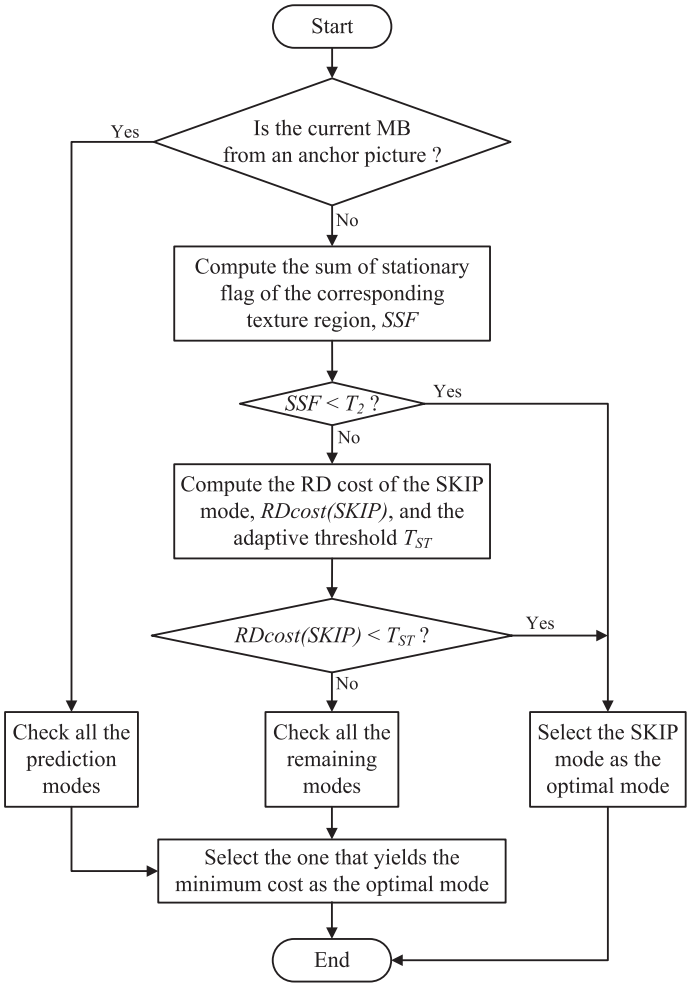

Fig. 6. Flowchart of the proposed TESMT algorithm.

(i.e., 16 × 16, 16 × 8, 8 × 16, P8 × 8, Intra) is skipped to significantly reduce the computational complexity. Otherwise, the exhaustive mode decision is performed to select the mode that corresponds the minimum RD cost as the optimal mode.

4. Experimental results and discussions

In order to evaluate the performance, we implement the proposed TESMT algorithm in the MVC reference software—*joint multi-view video coding* (JMVC 8.5) [23]. In our experiments, the proposed TESMT algorithm is tested on eight depth video sequences as shown in Fig. 7, which covers a wide range of motion activities and resolutions. Note that three views of each depth video are chosen for experiments. The first and third views are used as the reference views (i.e., the forward and backward views), respectively. The second view is used as the auxiliary view. The experimental setup is described as below: (1) 100 frames of each depth video sequence is encoded using the HBP prediction structure with GOP length = 8; (2) *QP*s are set as 24, 28, 32 and 36, respectively; (3) RDO and CABAC entropy coding are enabled; and (4) the search range of ME and DE is ±64. Moreover, the performance of the proposed method is compared with that of the exhaustive mode decision in MVC and measured by using the following indexes:

1. The BDPSNR (in dB) and BDBR (in percentage) as suggested in [24] are used to measure the averaged PSNR and bit rate difference between the RD performance produced by the proposed method and the MVC reference software, respectively.
2. The time saving ΔT is computed according to:

$$\Delta T = \frac{T_{Proposed} - T_{MVC}}{T_{MVC}} \times 100\% \tag{8}$$

where $T_{Proposed}$ and T_{MVC} denote the total encoding time of the proposed method and the MVC reference software, respectively.

Table 3 shows the experimental results of the proposed fast mode decision algorithm, TESMT, on a set of depth video sequences. It can be seen that the proposed TESMT algorithm is able to significantly reduce the computational complexity (say, 68.64% on average) while maintaining almost the same coding efficiency (only 0.07 dB loss in PSNR and 1.66% increment in the total bit rate), compared with the outcomes resulted by applying the exhaustive mode decision in MVC.

In addition, we compare the proposed TESMT algorithm with the method reported in reference [13] and document the results in Table 4. Note that the results here are averaged on the first and third views with the *QP* values 16, 22, 28 and 32 according to reference [13]. It can be observed that the proposed method consistently outperforms reference [13] in terms

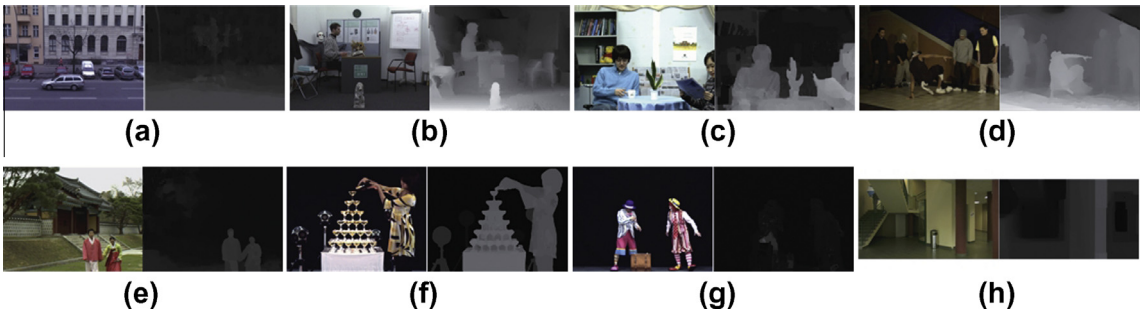

Fig. 7. An illustration of multi-view video plus depth sequences (the first frame): (a) Alt_Moabit (View 9), (b) Book_Arrival (View 8), (c) Newspaper (View 2), (d) Breakdancers (View 0), (e) Lovebird1 (View 6), (f) Champagne_Tower (View 39), (g) Pantomime (View 39), and (h) Pozan_Hall2 (View 5).

Table 3
Experimental results of the proposed TESMT algorithm on multiple depth video sequences.

Sequences	Views	Resolutions	BDPSNR (dB)	BDBR (%)	ΔT (%)
Alt_Moabit	9, 10, 11	1024 × 768	−0.02	+0.67	−69.46
Book_Arrival	8, 9, 10	1024 × 768	−0.06	+1.51	−73.46
Newspaper	2, 3, 4	1024 × 768	−0.10	+2.46	−67.33
Breakdancers	0, 1, 2	1024 × 768	−0.05	+1.32	−56.17
Lovebird1	6, 7, 8	1024 × 768	−0.12	+2.74	−79.81
Champagne_Tower	39, 40, 41	1280 × 960	−0.08	+1.97	−74.04
Pantomime	39, 40, 41	1280 × 960	−0.04	+0.86	−67.05
Pozan_Hall2	5, 6, 7	1920 × 1088	−0.07	+1.72	−61.82
Average			−0.07	+1.66	−68.64

Table 4
Performance comparison between reference [13] and the proposed TESMT algorithm.

Sequences	Reference [13]			Proposed		
	BDPSNR (dB)	BDBR (%)	ΔT (%)	BDPSNR (dB)	BDBR (%)	ΔT (%)
Book_Arrival	−0.05	+1.03	−25.35	−0.05	+1.44	−70.68
Newspaper	−0.49	+10.77	−53.55	−0.09	+2.21	−63.67
Lovebird1	−0.01	+0.10	−75.20	−0.10	+2.48	−77.53
Champagne_Tower	−0.54	+7.95	−63.30	−0.06	+1.63	−72.25
Average	−0.27	+4.96	−54.35	−0.08	+1.94	−71.03

Table 5
Performance comparison on the quality of the synthesized review resulted from the proposed TESMT algorithm and the exhaustive mode decision in MVC, respectively.

QP	24	28	32	36
Sequences	$\Delta PSNR$ (dB)			
Alt_Moabit	+0.01	+0.00	+0.00	+0.00
Book_Arrival	+0.01	+0.01	+0.01	+0.00
Newspaper	+0.00	−0.01	−0.01	−0.01
Champagne_Tower	+0.00	−0.02	−0.01	+0.00

of computational complexity reduction and coding efficiency maintenance. More specifically, compared with reference [13], 16.68% higher complexity reduction, 0.19dB BDPSNR improvement and 3.02% BDBR bit rate reduction are achieved by the proposed TESMT method.

Furthermore, we compare the quality of the synthesized view resulted from the proposed TESMT method and the exhaustive mode decision in MVC, respectively. In our experiments, the reconstructed neighboring texture and depth video (i.e., the first and third views) are utilized as the input to synthesize the virtual view (i.e., the second view) by using the *view synthesis reference software* (VSRS) [25]. The PSNR of the synthesized view is calculated by comparing the synthesized view with the ground truth (i.e., the original texture video captured at the same viewpoint). The corresponding results of four sequences are shown in Table 5 as examples. In this table, $\Delta PSNR = PSNR_{Proposed} - PSNR_{MVC}$ is used to measure the PSNR change of the synthesized view produced by the proposed method and the exhaustive mode decision in MVC, respectively. One can see that compared with the exhaustive mode decision in MVC, the proposed TESMT algorithm is able to maintain almost the same quality of the synthesized view.

5. Conclusions

The proposed TESMT algorithm makes full use of the texture-depth and spatial-temporal correlations to design an efficient two-stage early termination scheme for depth video coding. Using the proposed algorithm, we can effectively determine the SKIP mode as the optimal mode earlier so that those unnecessary time-consuming ME, DE and intra prediction processes can be skipped to significantly save the computational complexity. Experimental results have shown that compared with that of the exhaustive mode decision in MVC, the proposed algorithm can, on average, achieve 68.64% computational complexity reduction with only 0.07 dB loss in PSNR and 1.66% increment in the total bit rate while keeping almost unchanged quality of the synthesized view. In addition, the proposed algorithm is consistently superior to the method in [13] with 16.68% higher complexity reduction, 0.19 dB PSNR improvement and 3.02% bit rate reduction.

Acknowledgment

This paper is partially supported by the National Natural Science Foundation of China under the Grant 61372107.

References

[1] Muller K, Merkle P, Wiegand T. 3-D video representation using depth maps. Proc IEEE 2011;99(4):643–56.
[2] Tanimoto M, Tehrani MP, Fujii T, Yendo T. Free-viewpoint TV. IEEE Signal Process Mag 2011;28(1):67–76.
[3] Bourge A, Fehn C. White paper on ISO/IEC 23002-3 auxiliary video data representations. ISO/IEC JTC1/SC29/WG11, MPEG N8039; April 2006.
[4] Fehn C. Depth-image-based rendering (DIBR), compression and transmission for a new approach on 3D-TV. In: Proceedings of SPIE stereoscopic displays and virtual reality systems XI; January 2004. p. 93–104.
[5] Merkle P, Morvan Y, Smolic A, Farin D, muller K, de with PHN, et al. The effects of multiview depth video compression on multiview rendering. Signal Process: Image Commun 2009;24(1–2):73–88.
[6] ITU-T Recommendation H.264 and ISO/IEC 14496-10 AVC. Advanced video coding for generic audiovisual services; 2010.
[7] Liu S, Lai P, Tian D, Chen CW. New depth coding techniques with utilization of corresponding video. IEEE Trans Broadcast 2011;57(2):551–61.
[8] Daribo I, Tillier C, Pesquet-Popescu B. Motion vector sharing and Bitrate allocation for 3D video-plus-depth coding. EURASIP J Adv Signal Process 2009:1–14.

[9] Ding LF, Tsung PK, Chien SY, Chen WY, Chen LG. Content-aware prediction algorithm with inter-view mode decision for multiview video coding. IEEE Trans Multimedia 2008;10(8):1553–63.
[10] Zeng HQ, Ma K-K, Cai CH. Motion activity-based block size decision for multi-view video coding. Picture Coding Symposium (PCS); 2010. p. 166–9.
[11] Zeng HQ, Ma K-K, Cai CH. Fast mode decision for multi-view video coding using mode correlation. IEEE Trans Circ Syst Video Technol 2011;21(11):1659–66.
[12] Chiang PT, Chen YC. Software and hardware design for coding depth map sequence with texture motion information. In: IEEE international symposium on circuits and systems (ISCAS); May 2009. p. 1052–5.
[13] Wang M, Jin X, Goto S. Difference detection based early mode termination for depth map coding in MVC. Picture Coding Symposium (PCS); December 2010. p. 502–5.
[14] Zhang Q, An P, Zhang Y, Shen L, Zhang Z. Low complexity multi-view video plus depth coding. IEEE Trans Consum Electron 2011;57(4):1857–65.
[15] Merkle P, Smolic A, Muller K, Wiegand T. Efficient prediction structure for multiview video coding. IEEE Trans Circ Syst Video Technol 2007;17(11):1461–73.
[16] Zeng HQ, Ma K-K, Cai CH. Hierarchical intra mode decision for H.264/AVC. IEEE Trans Circ Syst Video Technol 2010;20(6):907–12.
[17] Zeng HQ, Cai CH, Ma K-K. Fast mode decision for h.264/avc based on macroblock motion activity. IEEE Trans Circ Syst Video Technol 2009;19(4):491–9.
[18] Wiegand T, Schwarz H, Joch A, Kossentini F, Sullivan GJ. Rate-constrained coder control and comparison of video coding standards. IEEE Trans Circ Syst Video Technol 2003;13(7):688–703.
[19] Ho YS, Lee EK, Lee C. Multi-view video test sequence and camera parameters. ISO/IEC JTC1/SC29/WG11, MPEG M15419; April 2008.
[20] Tanimoto M, Fujii T, Fukushima N. 1D parallel test sequences for MPEG-FTV, ISO/IEC JTC1/SC29/WG11, MPEG M15378; April 2008.
[21] Feldmann I, Mueller M, Zilly F, Tanger R, Muller K, Smolic A, et al. HHI test material for 3D video, ISO/IEC JTC1/SC29/WG11, MPEG M15413; April 2008.
[22] Tanimoto M, Fujii T, Tehrani MP, Wildeboer M. Depth estimation reference software (DERS) 5.0, ISO/IEC JTC1/SC29/WG11, MPEG M16605; June 2009.
[23] Joint video team, multi-view video coding reference software—joint multi-view video coding (JMVC 8.5); March 2011. <http://www.garcon.ient.rwth-aachen.de>.
[24] Bjontegaard G. Calculation of average PSNR differences between RD-curves. In: Document VCEG-M33, VCEG 13th meeting; April 2001.
[25] Tanimoto M, Fujii T, Suzuki K. View synthesis algorithm in view synthesis reference software 2.0 (VSRS2.0). ISO/IEC JTC1/SC29/WG11, MPEG M16090; February 2009.

Huanqiang Zeng received the B.S. and M.S. degrees from Huaqiao University, Xiamen, China and the Ph.D. degree from Nanyang Technological University, Singapore, all in electrical engineering. His current research interests include the areas of image processing, video communication, multiview video processing and computer vision.

Yongtao Wang received Ph.D. degree of pattern recognition and intelligent system in 2009 from Huazhong University of Science and Technology, China. During 2010, he was a research scientist of Temasek Laboratories, Nanyang Technological University, Singapore. He is currently an assistant professor of Institute of Computer Science & Technology, Peking University, China. His research interests involve image processing and computer vision.

Zhe Wei received the B.S. degree (computer science) and M.S. degree (computer science) from Huaqiao University, Xiamen, China, and the Ph.D. degree (electrical and electronic engineering) from Nanyang Technological University, Singapore. His research interests include multimedia processing, image/video compression, multiple description coding and multimedia communication.

Canhui Cai received B.S. degree from Xidian University, Xian, China, M.S. degree from Shanghai University, Shanghai, China, and Ph.D. degree from Tianjin University, Tianjin, China, all in electronic engineering. He is currently a Professor of School of Information Science and Technology, Huaqiao University, Xiamen, China. His current research interests include video communications, image/video processing, and computer vision.