Acta Technica *62* (2017), No. 4A, 31–42 © 2017 Institute of Thermomechanics CAS, v.v.i.

Adaptive reference frame based fast mode decision for multi-view video coding

Fengsui Wang[2], Jinjin Li[2], Haiying Cheng[2], Shuming Zhu[2]

Abstract. The Multi-view Video Coding (MVC) adopts multiple reference frame technology and offers many prediction modes and exhaustive mode decision in order to significantly improve coding efficiency, resulting in increasing the computational complexity tremendously in MVC. In this paper, a fast encoding algorithm by jointly using early Direct mode decision and adaptive reference frame for MVC was proposed. Based on the observation that the Direct mode was dominant to be the optimal mode, the proposed method calculated the rate distortion cost of the Direct mode and compared with an adaptive threshold, to terminate early the mode decision process if possible. Then, the prediction reference frame structure is adaptively chosen based on its characteristic for the current group of picture of the given multi-view video. Consequently, the computational load reduction can be obtained via skipping those unnecessary inter-view predictions and unlikely prediction modes. Experimental results have demonstrated that the proposed algorithm not only can significantly reduce the encoding computational load, but also can keep a negligible loss of coding efficiency.

Key words. Multi-view video coding, adaptive reference frame, mode decision, temporal prediction.

1. Introduction

Recently, Three Dimensional (3D) video has received an increasing demand, since it provides users with a totally new stereoscopic vision and interactive multimedia experience. As typical new multimedia applications, 3D Television (3DTV), Free-viewpoint Television (FTV) and immersive teleconference etc. offer the viewers a realistic depth perception of the scene [1, 2]. However, how to capture Multi-View

[1]Acknowledgement - This work is supported by the Major Program of Scientific Research of Anhui Province, China (KJ2015A071), and the Natural Science Foundation of Anhui Province, China (1708085MF154).

[2]Workshop 1 - College of Electrical Engineering, Anhui Polytechnic University, Wuhu, 241000, China

Video (MVV) is a key factor for these new applications. MVV can be captured simultaneously from multiple different viewpoints cameras at the same time, in order to represent 3D real-world video content. Obviously, MVV needs a tremendously huge amount transmission bandwidth and storage space for compression, compared with the traditional single view videos. Therefore, developing a multi-view video coding algorithm with low complexity and high coding efficiency is indispensable for interactive 3D applications.

Currently, Joint Video Team (JVT) of ISO/IEC Moving Picture Experts Group (MPEG) and ITU-T Video Coding Experts Group (VCEG) has developed novel MVC encoding standard based on the state-of-the art H.264/AVC standard as the appendix H of the single view video coding, H.264/AVC [3], in order to improve compression efficiency and lower encoding complexity for efficient storage and transmission. In the MVC scheme, different from the single view coding standard (for example H.264/AVC), both motion estimation (ME) within a single view by using the traditional temporal prediction and disparity estimation (DE) between neighboring views by exploiting inter-view prediction are offered in order to reduce data redundancy. And that many new encoding techniques, such as hierarchical B pictures (HBP) prediction structure, various prediction modes, disparity estimation, and intricate prediction modes, and so on, are adopted to significantly eliminate the temporal, spatial and inter-view redundancies root in the MVV, at the cost of extremely high computational complexity, which is the bottleneck of enabling MVC into practical applications. Therefore, developing a fast mode decision algorithm, which can reduce the coding complexity while keeping almost the same coding performance, is the key issue for the MVC practical realization.

Many algorithms have been presented to lower the computational load and improve coding efficiency of MVC. In our opinion, they can be categorized into three types as follows. The first type aims to adopt a fast mode decision algorithm for reducing prediction direction and candidate modes required to be checked [4-9]. The second type is to adaptively change prediction structure for omitting the unnecessary temporal or inter-view prediction [10-12]. The third type jointly employs prediction direction, multi-reference frames selection, and various correlations . In the first type, time consuming and mode distribution in the inter-view mode selection can theoretically keep a good trade off between the computational load lowering and coding efficiency improving. In the second type, the computational complexity reduction gain is highly dependent on the content of the multi-view video sequences. Generally, methods having the higher time saving usually lead to poor encoding image quality, while methods with the better RD performance usually demand consuming more coding time poor. The above-mentioned methods can reduce the computational complexity to some extent. However, the encoding performance of the whole encoder still requires improvements with respect to computation complexity and RD performance.

The rest of the paper is organized as follows. In Section 2, the motivation is described. In Section 3, the proposed fast MVC algorithm using adaptive multi-reference frame for multi-view video coding is proposed in detail. Experimental simulation and discussed are shown in Section 4. Finally, conclusion is drawn in the

last section.

2. Motivation

In order to efficiently reduce the intra-view redundancy in each view and the inter-view redundancy between the two neighboring views, MVC exploits the HBP prediction structure in the MVC reference software– Joint multi-view video coding (JMVC) by the JVT. HBP prediction structure with 8 views is illustrated in Figure 1.

In each view, the HBP frame architecture is a group of pictures (GOP). They can be categorized into two classes of frames: the anchor pictures (i.e., the pictures at T0 and T8) and the non-anchor pictures (i.e., the pictures at T1, T2, ..., T7). Moreover, there exists three classes of the views: the base view (i.e., the picture V0); the main views (i.e., V2, V4 and V6), whose reference directions are only temporal prediction using ME in the non-anchor pictures; and the auxiliary views (i.e., V1, V3, V5 and V7), whose reference directions are both the temporal prediction by ME and the inter-view prediction by DE. As a result, MVC not only performs ME within the same view to exploit the temporal correlation, but also conducts DE among neighboring views at the same time using inter-view correlation besides ME in the auxiliary views, which results in extremely large computational complexity that obstructs MVC from practical application.

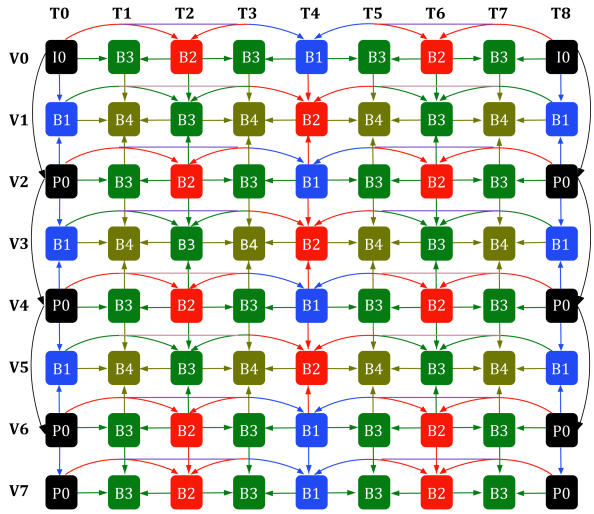

Fig. 1. Hierarchical B pictures prediction architecture for multi-view video coding

In order to achieve higher coding efficiency, JMVC employs ME and DE by a set of variable block sizes, including Direct, 16×16, 16×8, 8×16 and 8×8, where the 8×8 can be further divided into 8×8, 8×4, 4×8 and 4×4 sub-modes (jointly denoted as P8× in this work). The optimal mode is decided by full mode decision and

Lagrange rate distortion optimization (RDO) function in various prediction modes, that is, exhaustively checking all the prediction modes and calculating the RD cost of each mode, and then selecting the one with the minimum RD cost as the optimal mode. RD cost is shown in Equation (1).

$$J_{\mathrm{MODE}}(S_k, C_k|\lambda_{\mathrm{MODE}}) = \mathrm{SSD}(S_k, C_k) + \lambda_{\mathrm{MODE}} R(S_k, C_k) \tag{1}$$

where $J_{MODE}(S_k,\ C_k|\ \lambda_{MODE})$ is RD cost of MODE, S_k and C_k represent the kth original MB and the corresponding reconstructed MB, respectively. $\mathrm{SSD}(S_k, C_k)$ denotes the sum of squared difference between the original MB and the reconstructed MB. λ_{MODE} is the Lagrange multiplier, and $R(S_k, C_k)$ denotes the total bit rate after entropy coding.

The motivation of the proposed fast algorithm based on adaptive multi-reference frame is based on two observations, as follows. 1) As a novel feature of the MVC, HBP prediction structurre inter-view reference frame prediction via DE besides spatial and temporal predictions, which spontaneously results in tremendous computational complexity. Since most of the multi-view video sequences in real-world video content only contain slow motion or motionless objects and homogeneous background, their temporal reference frame prediction is much stronger than their inter-view reference frame prediction. In other words, the inter-view reference frame prediction is actually rarely used in many cases. Therefore, the proposed algorithm demands that the prediction reference frame can be adaptive to multi-view video content. 2) Within a GOP for an multi-view video sequence, the temporal correlation between current frame and its temporal reference frame is often decreased as their reference frame distance increases. That is, the temporal correlation varies with the frame distance between the current frame and its reference frame Similarly, the contribution of the inter-view prediction via DE to the improvement of coding efficiency changes from the frame distance between the current frame and its reference frame in inter-view direction. To verify these two observations, extensive experiments have been conducted by using various kinds of multi-view video sequences listed in Table 1 to obtain the distribution of temporal and inter-view reference frame predictions as the varying multi-reference frame distance. For this purpose, the frame distance D between the current frame and its multiple reference frames is defined as follows: the anchor frame is defined as the basic reference frame, i.e., $D = 0$, and other non-anchor frames are respectively defined as $D = 1, 2, 3, 4$, which is determined based on the frame distance between the current frame and its multiple forward reference frames. The bigger D of the current frame, the nearer distance between the current frame and its forward reference one will be.

Table 1. Multi-view video sequences

Sequences	From	Frames	Resolution
Ballroom	MERL	250	640×480
Vassar	MERL	250	640×480
Exit	MERL	250	640×480
Race1	KDDI	300	640×480
Ballet	Microsoft	100	1024×768
Breakdancers	Microsoft	100	1024×768
Doorflowers	HHI	150	1024×768
Jungle	HHI	250	1024×768

The experimental conditions are as follows: each test sequence is encoded using the HBP prediction structure under GOP=12, QP = 28, 30, 32, 36), rate distortion optimization (RDO) and context-adaptive binary arithmetic coding (CABAC) are enabled, and the search range of the ME and DE is ±64. The distribution of optimal prediction is shown in Figure 2 and Figure 3.

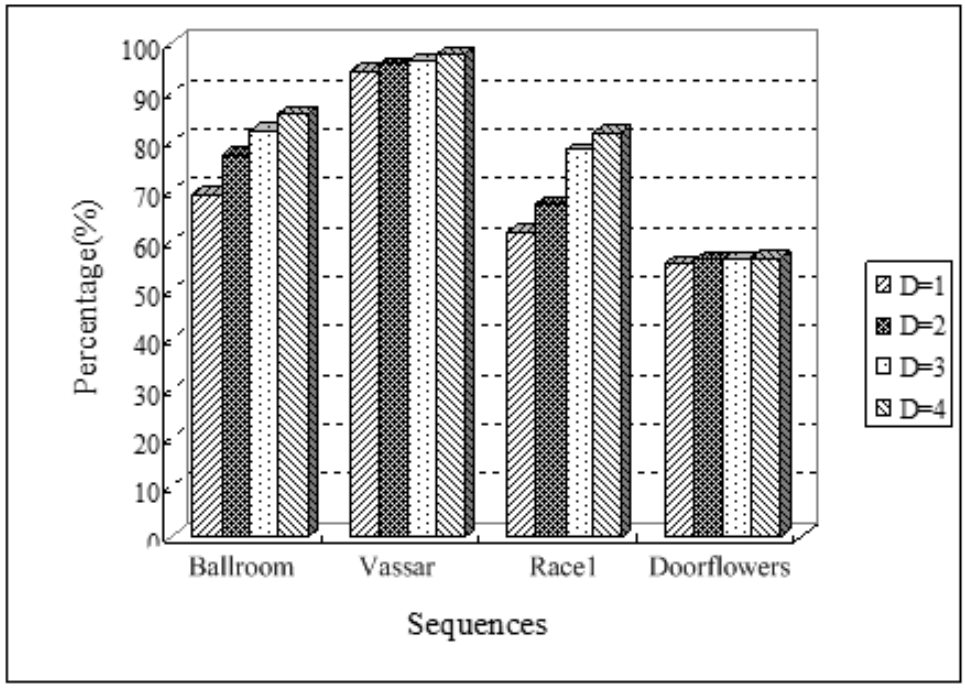

Fig. 2. Rate of temporal reference frame selected as the optimal prediction

From Figure 2 and Figure 3, we can see that both temporal and inter-view reference frame predictions change not only from sequence to sequence, but also from frame distance D to D. Moreover, the rate of temporal reference frame to be

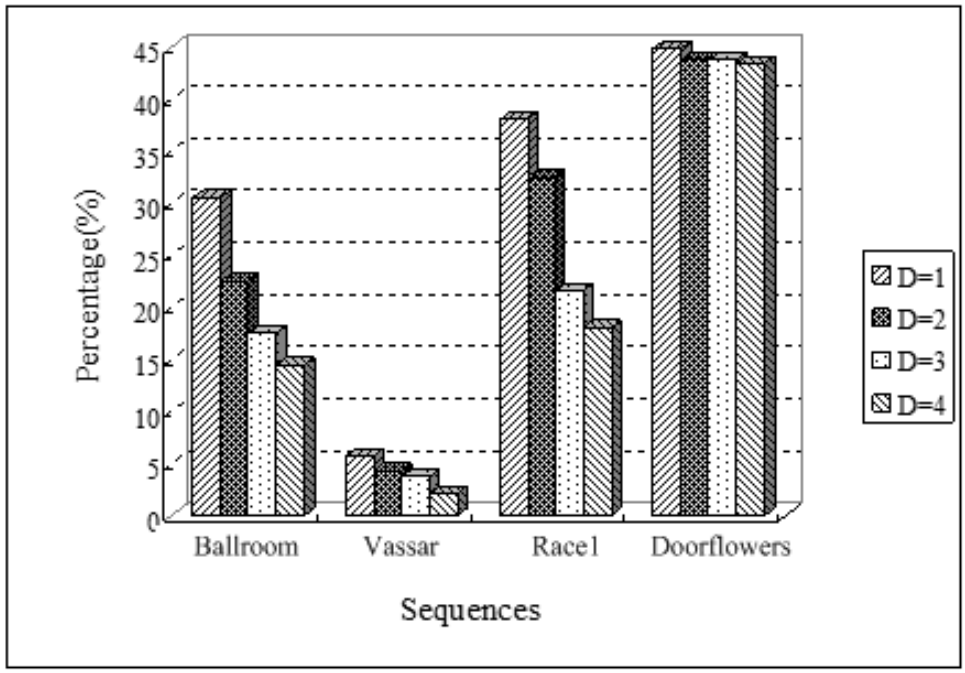

Fig. 3. Rate of inter-view reference frame selected as the optimal prediction

selected as the optimal prediction increases when the frame distance D increases. On the contrary, the rate of inter-view reference frame to be selected as the optimal prediction increases when the frame distance D decreases. For example, the frame distance D is very small for all D_i (i = 1, 2, 3, 4). This means the contribution of the inter-view prediction via DE to the improvement of coding efficiency is very limited. In other words, the inter-view prediction direction is rarely selected as the optimal prediction direction. Under the circumstances, skipping time-consuming inter-view prediction by using DE will only result in negligible image quality loss but significantly reduce encoding time.

3. Proposed Algorithm

3.1. Early Direct mode decision

The coding complexity of the Direct mode is quite small, but it provides good RD performance. Since the multi-view video sequences with homogeneous background and slow motion or motionless are often encountered in the real-world video content and Direct mode is suitable for coding those video sequences, the Direct mode intuitively should occupy the majority of the optimal modes. To verify this intuition, the statistics and analysis of optimal modes are performed by extensive experiments. The statistical results are shown in Figure 4.

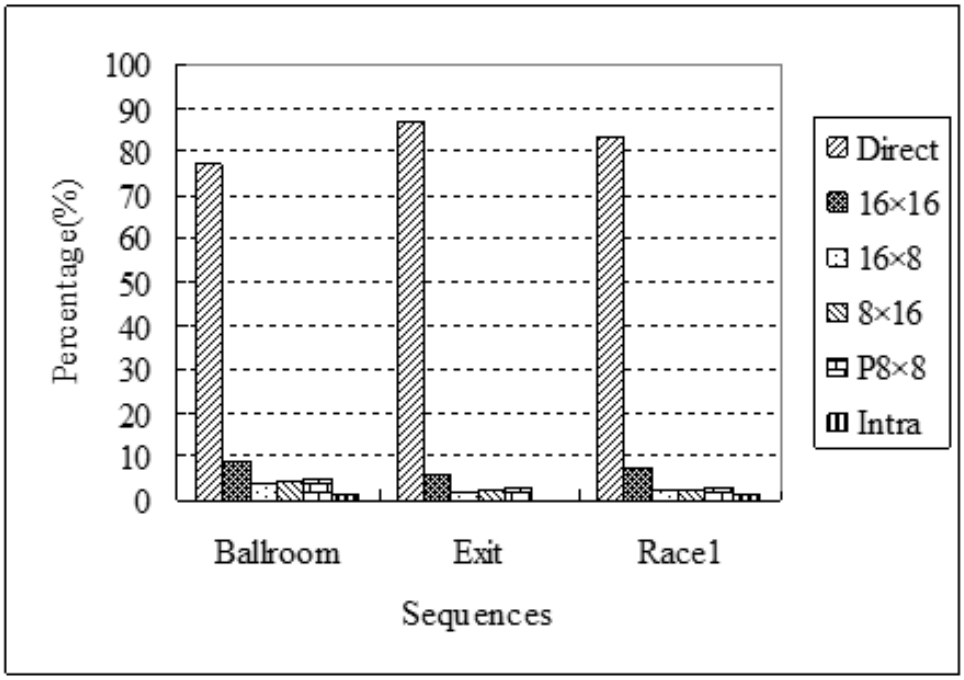

Fig. 4. Distribution of optimal mode in MVC

We can be seen from Figure 4 that Direct mode occupies the majority proportion of the optimal modes for various multi-view video sequences (more than 80% on average), while the sum of the other prediction modes selected to be the best mode are much less. On the other hand, the Direct mode consumes negligible coding time owing to it being without ME and DE. Therefore, if we can pre-decide whether the Direct mode is the optimal mode or not, great computational complexity will be saved.

It should be pointed out that mode decision in MVC checks all the prediction modes in accordance with following the order of "Direct, 16×16, 16×8, 8×16, P8×8, Intra". As the above analysis, Direct mode consume very small encoding time while mode decision of other modes is very time-consuming due to conducting sophisticated ME and DE. Hence, if we can design an algorithm for MVC to early determine whether Direct mode is the optimal mode or not, the time- consuming ME, DE and Intra prediction will be skipped.

As shown in Figure 5, MB0 denotes the current MB. MB1 is the corresponding co-located MB of the current MB in previously coded forward picture in temporal direction or forward view in inter-view direction, while MBi for $i = 2, 3, \ldots, 9$ are the eight neighboring MBs of MB1.

In proposed algorithm, the adaptive threshold for early termination is determined by using the summation of the average value of RDO, J_{avg}, and the minimal value of the RD cost of Inter16×16 in the corresponding coded frame, J_{min}. J_{avg} can be

calculated from the RD cost of the whole corresponding co-located MBs by making full use of the temporal correlation and inter-view correlation between the current MB and its corresponding MBs, among temporal-adjacent MBs in forward picture or inter-view-adjacent MBs in forward view. J_{avg} can be computed using equation (2).

$$J_{avg} = \frac{1}{N}\sum_{i=1}^{N} J_i \tag{2}$$

where N is the number of the MBs. J_i represents the RD cost of MBi. The adaptive threshold T can be derived from the following equation (3).

$$T = J_{avg} + J_{\min} \tag{3}$$

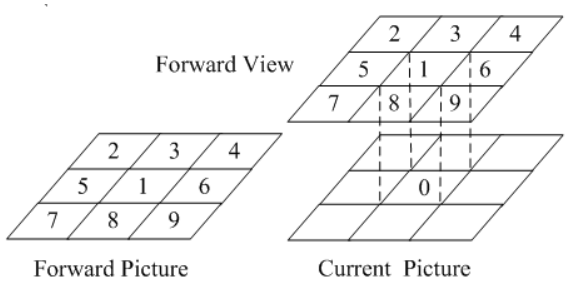

Fig. 5. Illustration of the reference frames in the auxiliary view

The proposed early termination algorithm firstly records the minimal RD cost of Inter 16×16 prediction mode for the previous coded MBs as shown in Figure 5. Then the adaptive threshold can be acquired by equation (3), which is employed to determine if the Direct mode should be skipped by exploiting the average value of the RD cost of the previously coded co-located MB and its neighboring MBs in the corresponding temporal and inter-view pictures, plus the minimal RD cost value of the previously coded MBs for Inter16×16 mode. After obtaining this adaptive threshold T, if the RD cost of the Direct mode of the current MB, J_{Direct}, is smaller than the adaptive threshold T, and the optimal mode of the previously coded MBs is Direct mode, early termination method will be carried out and Direct mode is selected as the optimal mode and the encoding mode decision can be omitted.

3.2. Proposed fast algorithm

In order to optimize RD performance for MVC, an appropriate prediction structure is very important to adaptive to vary with multi-view video sequences. Since GOP is a basic encoding unit in MVC, we select a GOP as the basic prediction structure in this paper. As Section 2 mentioned, multi-view video can be divided into four classes. If the frame distance of temporal reference frame prediction using ME for a GOP in given multi-view video sequence, D_T, is larger than or equal to D_1, this means temporal correlation is much stronger than inter-view correlation.

These sequences are usually composed of motionless or slow motion objects, and the improvement of inter-view prediction to RD performance is almost negligible. If D_T is less than D_1 and larger than D_2, only the nearest two reference frames (namely, $D = 3$ or $D = 4$) is negligible, and only those frames with $D = 1$ or $D = 2$ are used to perform inter-view prediction. If D_T is less than D_2 and larger than or equal to D_3, the nearest neighbor forward reference frame of the current frame (i. e., $D = 4$) is scarcely effect. If D_T is less than D_3, the inter-view correlation of these sequences is dominant. Hence, the inter-view prediction by using DE is indispensable for this type of sequences. The thresholds D_1 =0.9, D_2 =0.8 and D_3= 0.6 are empirically decided from extensive experimental results by using various multi-view video sequences.

Based on above-mentioned analysis, the proposed early Direct mode decision and adaptive multi-reference frame are summarized and given as follows.

Step 1: For each GOP of the input multi-view video, encode all frames in the base view and anchor frames in non-base views.

Step 2 Calculate the RD cost of Direct mode, J_{Direct}, and the adaptive threshold T according to Equation (3).

Step 3: If $J_{Direct} < T$ and the optimal mode of the previously coded MBs is Direct mode, perform an early Direct mode termination, go to step 7. Otherwise, go to step 4.

Step 4: Computer the frame distance of temporal reference frame prediction using ME for a GOP in given multi-view video sequence, D_T.

Step 5: Decide the most proper prediction reference frames based on D_T in Section 3.2.

Step 6 Encode the remaining frames in current GOP based on the selected prediction reference frame structure utilizing the presented Early Direct Mode Termination method. Then go to step 1 and proceed with next GOP.

Step 7: Direct mode is selected as the optimal mode and the mode decision process is early terminated, then go to step 2 and proceed with next MB.

4. Experimental results

The proposed fast algorithm for multi-view video coding is performed on JMVC 8.3 on various multi-view video sequences. Each sequence selects 8 views for experiment. V0 is selected as the base view, and V2, V4 and V6 are chosen as the main views, and V1, V3, V5 and V7 are used as the auxiliary views. Experimental setup is described as below: 1) HBP prediction structure with GOP = 12; 2) QP = 20, 24, 28, 32, 36, respectively; 3) RDO and CABAC are enabled; 4) the search range of the ME and DE is ±96.

Experimental results of the proposed fast algorithm are shown in Table 2. In Table 2, the ΔT (%), $\Delta PSNR$ (dB) and ΔB (%) are defined in equation (4), (5) and (6):

$$\Delta T = \frac{T_{proposed} - T_{JMVC}}{T_{JMVC}} \times 100\% \tag{4}$$

$$\Delta PSNR = PSNR_{proposed} - PSNR_{JMVC} \tag{5}$$

$$\Delta B = \frac{BR_{proposed} - BR_{JMVC}}{BR_{JMVC}} \times 100\% \tag{6}$$

where $T_{proposed}$, $PSNR_{proposed}$ and $BR_{proposed}$ are the encoding time, *PSNR* and bit rate resulted from the proposed fast algorithms, respectively. T_{JMVC}, $PSNR_{JMVC}$ and BR_{JMVC} are the encoding time, *PSNR* and bit rate resulted from the JMVC reference software, respectively.

Table 2. Experimental result comparison between Reference [4] (A) and our algorithm (B)

Sequences	Method	ΔPSNR (dB)	ΔB (%)	ΔT (%)
Ballroom	<A>	-0.258	5.87	-61.49
	<B>	-0.078	1.70	-75.14
Race1	<A>	-0.301	6.30	-46.77
	<B>	0.066	1.25	-78.39
Exit	<A>	-0.291	9.65	-64.16
	<B>	-0.075	2.25	-82.04
Ballet	<A>	-0.185	6.30	-65.65
	<B>	-0.025	0.84	-79.95
Doorflowers	<A>	-0.067	2.81	-73.57
	<B>	-0.112	4.03	-87.13
Breakdancers	<A>	-0.304	12.64	-52.21
	<B>	-0.044	1.67	-73.78
Uli	<A>	-0.242	5.54	-64.23
	<B>	-0.070	1.49	-75.31
Average	<A>	-0.235	7.01	-61.15
	<B>	-0.067	1.89	-78.82

One can see from Table 2 that our algorithm can greatly reduce the computational complexity while keeping almost the same coding efficiency, compared with the full mode decision in MVC reference software. It has reduced encoding time about 78.82% on average, with maximum of 87.13% in "Doorflowers". The loss of coding efficiency is negligible in our proposed algorithm: only 0.067 dB PSNR loss on average and 1.89% increment in the bit rate on average. Therefore, our method can significantly reduce encoding time while keeping a good RD performance.

Table 2 compares the proposed fast algorithm with that proposed in reference [4]. The proposed algorithm shows better comparison results. Compared with algorithm

presented in reference [4], a speed up of 17.67% on average can be acquired in our algorithm. Meanwhile, 0.168 dB PSNR improvement and 5.52% BDBR bit rate reduction are achieved by our proposed in comparison with reference [4]. In a word, the proposed method outperforms reference [4] in terms of both coding efficiency maintenance and computation complexity reduction.

For a better illustration, Figure 6 provides a comparison of the time-saving ratio between reference [4] and the proposed algorithm. We can easily observe that the proposed method can more efficiently reduce computational load. In a word, the proposed algorithm is superior to reference [4] in terms of reducing encoding time and RD performance.

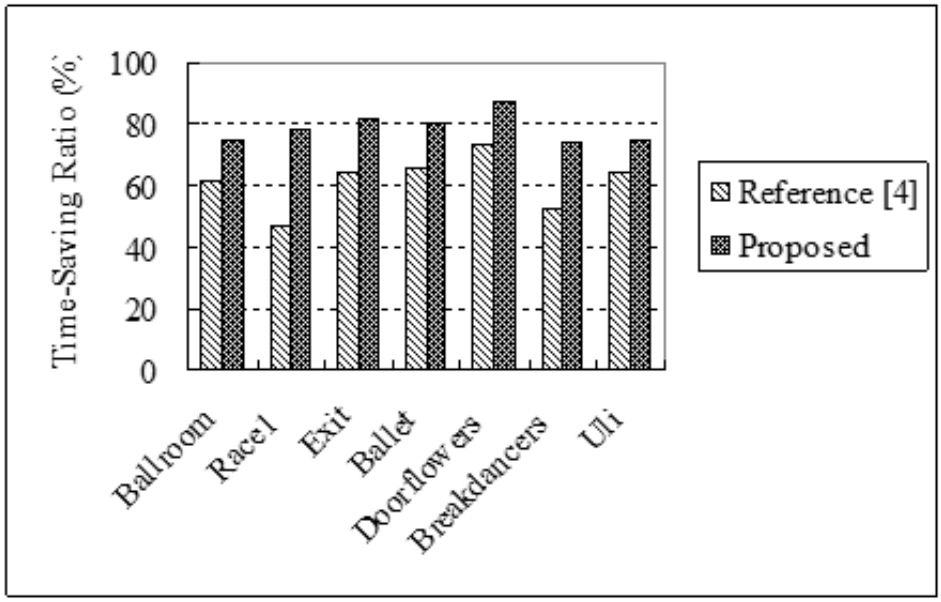

Fig. 6. Comparison of time-saving ratio between Reference [4] and the proposed algorithm

5. Conclusions

In this paper, an efficient MVC fast algorithm is proposed based on adaptive reference frame and early Direct mode decision. The proposed method provides an early Direct mode termination scheme for omitting the unnecessary time- consuming ME and DE. Then four prediction structures are divided to suitable for different video contents of the input multi-view video, and the most proper prediction reference frame is selected according to the frame distance of the current GOP. Experimental results show that the proposed algorithm the proposed algorithm acquires time saving by 78.82% on average, whereas the image quality only incurs 0.067 dB PSNR loss.

References

[1] K. Muller, P. Merkle, T. Wiegend: *3-D video representation using depth maps.* Proceedings of the IEEE *99* (2011), 643–656.

[2] M. Tanimoto, M. P. Tehrani, T. Fujii, T. Yendo: *FTV for 3-D spatial communication.* Proceedings of the IEEE *100* (2012), 905–917.

[3] A. Vetro, T. Wiegand, G. J. Sullivan: *Overview of the stereo and multiview video coding extensions of the H.264/MPEG-4 AVC standard.* Proceedings of the IEEE 2011 *99* (2011), 626–642.

[4] L. Q. Shen, Z. Liu, P An, R. Ma, Z. Y. Zhang: *Low-complexity mode decision for MVC.* IEEE Transactions on Circuits and Systems for Video Technology *21* (2011), 837–843.

[5] Y. Zhang, S. Kwong, G. Jiang, X. Wang, M. Yu: *Statistical early termination model for fast mode decision and reference frame selection in multiview video coding.* IEEE Transactions on Broadcasting *58* (2012), 10–23.

[6] T. Zhao, S. Kwong, H. L. Wang, Z. Wang, Z. Q. Pan, C. J. Kuo: *Multiview coding mode decision with hybrid optimal stopping model.* IEEE Transactions on Image Processing *22* (2013) 1598–1609.

[7] Y. Li, G. Yang, N. Chen, Y. Zhu, X. Ding: *Early DIRECT mode decision for MVC using MB mode homogeneity and RD cost correlation.* IEEE Transactions on Broadcasting *62* (2016), 700–708.

[8] Y. Zhang, S. Kwong, L. Xu, G. Y. Jiang: *DIRECT mode early decision optimization based on rate distortion cost property and inter-view correlation.* IEEE Transactions on Broadcasting *59* (2013), 390–398.

[9] S. Khattak, R. Hamzaoui, S. Ahmad, P. Frossard: *Fast encoding techniques for multiview video coding.* Signal Processing: Image Communication *28* (2013), No. 4, 569–580.

[10] L. Q. Shen, P. An, Z. Liu, Z. Y. Zhang: *Low complexity depth coding assisted by coding information from color video.* IEEE Transactions on Broadcasting *60*, (2014), 128–133.

[11] M. Li, M. Chen, C. Yeh, K. Tai: *Performance improvement of multi-view video coding based on geometric prediction and human visual system.* International Journal of Imaging Systems and Technology *25* (2015), 41–49.

[12] W. Zhu, X. Tian, F. Zhou, Y. W. Chen: *Fast disparity estimation using spatio-temporal correlation of disparity field for multiview video coding.* IEEE Transactions on Consumer Electronics *56* (2010), 957–964.

Received November 16, 2016