A Modified Inter-view Prediction Scheme for Multiview Video Coding to Improve View's Interactivity

Ayman M. Hamadan, Hussein A. Aly and Mohamed M. Fouad
Department of Computer Engineering, Military Technical College, Cairo, Egypt

Keywords: Multiview Video Coding, Video Compression, Interactivity.

Abstract: In this paper, we present a modified inter-view prediction multiview video coding (MVC) scheme form the perspective of viewer interactivity. This latter requires a high transmission bite-rate, when a viewer requests some view(s). Conventional solutions to such a problem assume that the requested views are already encoded by a MVC scheme. In this paper, we modify the *MVC-HBP* standard scheme yielding reducing the average transmission bite-rate required to retrieve the requested views (*i.e.*, interactivity). With real data sequences, clear improvements are shown using the proposed MVC scheme compared to competing MVC schemes in terms of rate-distortion (RD) and average transmission bite-rate.

1 INTRODUCTION

Multiview video (MV) refers to the simultaneous capturing of multiple videos of the same scene by a large array of closely spaced cameras from different viewpoints(Kubota et al., 2007). The MV has wide applications in education, entertainment, surveillance, and health care. MVC has encoded the MV signals using various proposed schemes including both temporal and inter-view predicted frames (*i.e.*, frames are predicted not only from the temporally neighboring frames, but also from the corresponding frames in adjacent views)(Vetro et al., 2008). Both analysis of MVC schemes and real-time human-computer interaction are essential steps prior to developing any future IMVS system(Kalva et al., 2006). In an interactive multiview video streaming (IMVS) system(Vetro et al., 2011), the user only needs to retrieve the requested views (*i.e.*, not the whole set of frames) to play back the captured MV content. Thus, users may exploit cognitive views to generate additionally virtual views that are created at the decoder to increase the navigation capacities and the look-around effect(Mùandller et al., 2011).

MVC typically focuses on the RD performance of compressing all frames of all views, as presented in(Merkle et al., 2007; Kurutepe et al., 2007; Cheung et al., 2009; Yang et al., 2010; Lee et al., 2010; Cheung et al., 2011; Maugey et al., 2012). In (Merkle et al., 2007), an experimental analysis of MVC for various temporal and inter-view prediction structures is presented. The compression method used is based on a multiple-reference picture technique in the H.264/AVC video coding standard. In (Kurutepe et al., 2007), authors introduce a view-selective streaming strategy for streaming MV for single-user interactive 3DTV applications. In (Cheung et al., 2009), authors address the problem of designing a pre-encoded frame structure for a streaming server to enable a new interactive multiview switching. A streaming client can, then, send requests periodically to a server to switch to different views, while continuing uninterrupted temporal playback of the streaming video. In (Yang et al., 2010), authors modify the joint multiview video model (JMVM) with a proposed interactivity factor to evaluate the performance of different MVC schemes in corporation with RD performance. In (Lee et al., 2010), authors describe geometric prediction structure for multi-view video coding. By exploiting the geometric relations between each camera pose, to make prediction pair which maximizes the spatial correlation of each view. In (Cheung et al., 2011), authors propose a redundant representation of *I*-, *P*-, and merged frames, where each original picture can be encoded into multiple versions to facilitate view switching. In (Maugey et al., 2012), a new multiview representation is proposed to complete the classical color and depth information streams. An auxiliary information, related to those streams, roughly describes the occlusion region(s) in order to help the synthesis algorithm at the receiver.

Although the inter-view predictions of the afore-

Hamdan A., A. Aly H. and M. Fouad M..
A Modified Inter-view Prediction Scheme for Multiview Video Coding to Improve View's Interactivity.
DOI: 10.5220/0004198600350040
In *Proceedings of the International Conference on Computer Vision Theory and Applications* (VISAPP-2013), pages 35-40
ISBN: 978-989-8565-47-1

mentioned MVC schemes improve the RD performance, the corresponding bit-rate is increased when individually sending views. In this paper, we present a modified inter-view prediction MVC scheme form the perspective of viewer's interactivity. The proposed MVC scheme reduces the bite-rate required to retrieve the requested views (*i.e.*, interactivity) with comparable RD performance.

This paper is structured as follows. Section 2 presents the proposed *MVC* scheme showing its relation to the views'interactivity. Data sets description, implementation setup, experiments and results are presented in Section 3. Finally, conclusions are given in Section 4.

2 THE PROPOSED MVC SCHEME

In this section, we first give a brief background on the standard inter-view prediction MVC scheme (Vetro et al., 2008). Then, we extend the mathematical relationship between the standard MVC scheme and views'interactivity in the IMVS system to finally present the proposed MVC scheme.

The *MVC-HBP* standard scheme has coding efficiency advantages compared to other schemes (Merkle et al., 2007) in terms of RD performance. An example, with eight linearly arranged cameras and a group of pictures (GOP) length of 8 (for simplicity), is shown in Fig. 1(a). This scheme first uses inter-view prediction to provide *P* pictures for even camera views (S_2, S_4 and S_6) at T_0 and T_8 of each GOP from the base view S_0. Rest of the pictures in the even camera views are predicted with hierarchical *B* pictures in the temporal direction (Schwarz et al., 2006). Whereas, odd camera views (S_1, S_3 and S_5) are obtained by combining an inter-view prediction from two adjacent even views and an hierarchical *B* coding structure in the temporal direction. For an even number of views, the last view represents a specific case for prediction. S_7 is coded as shown in Fig. 1(a), starting with an inter-view predicted *P* frame, followed by hierarchical *B*-frames, which are also inter-views predicted from the previous view. This coding scheme can be applied to any multiview with more than two views.

In IMVS systems, an user only needs to receive the requested views. The IMVS server, then, extracts the requested views from the whole set of frames with reference frames from other views. In turn, the IMVS server sends those frames as a subset in one bit-stream to the user via the network. To retrieve an i^{th} view, V_i,the required extracted frames (EF) for one GOP can be generally formulated as,

$$EF = I + n \times P + m \times B, \quad (1)$$

where *I*, *P*, and *B* are frames related to that specific view, *n* and *m* denote the number of the *P*-frames and *B*-frames, respectively, depending on the view's location at the whole set. To determine *n* and *m* for extracting the V_i in the *MVC-HBP* scheme for one GOP, given base view as S_0 as shown in Fig. 1(a), (1) can be rewritten as,

$$EF_i = I + \alpha_i P + [(2k+1)\beta_i - (k+1)\delta_i + k]B, \quad (2)$$

where $i \in \{0,1,2,\ldots,N-1\}$ denotes the view number, N denotes the number of the views, $[\cdot]$ denotes the number of *B*-frames. α_i that denotes the number of the *P*-frames, β_i that determines whether V_s is an odd or even, and δ_i that determines whether V_s is an edge view or not, can be obtained as,

$$\alpha_i = \lceil i/2 \rceil, \forall\, i \in \{0,1,2,\ldots,N-1\}, \quad (3)$$
$$\beta_i = i \bmod 2, \forall\, i \in \{0,1,2,\ldots,N-1\}, \quad (4)$$
$$\delta_i = i \bmod 2, \forall\, i \in \{0,N-1\}. \quad (5)$$

The base view isn't necessary to be set to S_0 as presented in the standard *MVC-HBP* scheme (Merkle et al., 2007). So, the symbol *i* in (3), (4) and (5) can be replaced with $\mid B_v - i \mid$, where $B_v \in \{0,1,2,\ldots,N-1\}$ denotes the number of the base view. Thus, (3), (4) and (5) can be rewritten as,

$$\alpha_x = \lceil (|x|)/2 \rceil, \forall\, i \in \{0,1,2,\ldots,N-1\}, \quad (6)$$
$$\beta_x = (|x|) \bmod 2, \forall\, B_v, i \in \{0,1,2,\ldots,N-1\}, \quad (7)$$
$$\delta_x = (|x|) \bmod 2, \forall\, B_v, i \in \{0,N-1\}, \quad (8)$$

where, $x = B_v - i$. Accordingly, (2) can be rewritten as,

$$EF_{(B_v-i)} = I + \alpha_{(B_v-i)} P + [(2k+1)\beta_{(B_v-i)} - (k+1)\delta_{(B_v-i)} + k]B. \quad (9)$$

In (9), you can notice that only decreasing the value of $\alpha_{(B_v-i)}$ minimizes the number of required extracted *P*-frames as $\beta_{(B_v-i)}$ and $\delta_{(B_v-i)}$ are binary. Thus, the bit-rate to retrieve a specific view V_i is reduced. To decrease the value of $\alpha_{(B_v-i)}$, we propose that the value of B_v should be set to $\lceil$ median $\{0,1,2,\ldots,N-1\}$ $\rceil$, as shown in Fig. 1(b), yielding improving the view's interactivity. Table 2 shows the number of extracted frames to retrieve a specific view of one GOP using different base views using (9). Setting B_v to 3 or 4 yields to a minimum number of extracted frames. Therefore, in the rest of this paper, the number of views, N is set to 8, and B_v is set to 4 (*i.e.*, the base view is set to S_4).

The proposed scheme, shown in Fig. 1(b), first uses the inter-view prediction to provide *P*-frames for

even camera views (S_2, S_0 and S_6) at T_0 and T_8 of each GOP from the base view S_4. In the even camera views, the rest of the frames are predicted with hierarchical *B*-frames in the temporal direction. Whereas, odd camera views (S_1, S_3 and S_5) are obtained by combining the inter-view prediction from two adjacent even views and the hierarchical *B*-frames in the temporal direction. For an even number of views, the last view is to be predicted in a specific manner. In our case, the last view, S_7, is predicted, as shown in Fig. 1(b), in two steps. First, an inter-view *P*-frame is predicted. Then, hierarchical *B*-frames are predicted, where the latter are also inter-views predicted from the previous view. This coding scheme can be applied to any multiview with more than two views.

For example, to extract necessary frames from MVC bit-stream for a specific view at one GOP, the following frames are required to be extracted in an hierarchical order:

- *I*-frame: $S_0/(T_0, T_8)$.
- *P*-frame: $S_2/(T_0, T_8)$, $S_4/(T_0, /T_8)$, $S_6/(T_0, T_8)$.
- *B*1-frame: S_4/T_4, S_5/T_0, S_5/T_8, S_6/T_4.
- *B*2-frame: S_5/T_4, S_4/T_6, S_6/T_6.
- *B*3-frame: S_4/T_7, S_5/T_6, S_6/T_7.

3 EXPERIMENTS & RESULTS

In this section, data sequences used are described in Section 3.1, the implementation setup of experiments is given in Section 3.2, and finally results are discussed in Section 3.3.

3.1 Data Sets Description

The data set used in the experiments includes two categories: MERL[1] and KDDI[2]. Their characteristics are provided in Table 1. MERL includes three sequences. The first sequence, **Ballroom**, shows a dynamic scene containing fast motion of the dancers and many overlapping objects. The second sequence, **Exit**, represents a static scene with few persons slowly moving from right to the middle of the scene. The third sequence, **Vassar**, has been captured in an ambient day light and contains no discernable motion blur on the boundaries of the moving objects. KDDI includes the **Race1** sequence that represents a scene captured by a cameras moving from left to right, then returning back from right stop at the middle of the scene. The latter sequence contains small moving cars on the race track in sunny day.

3.2 Implementation Setup

Our implementation runs on a 2.4 GHz Core i3, with 2GB of RAM. The proposed MVC scheme (*i.e.*, referred to as *proposed*, B_V=4) is compared to i) the MVC standard scheme (Vetro et al., 2008) (*i.e.*, referred to as *MVC-HBP*, B_V=0), ii) that of (Yang et al., 2010) (*i.e.*, referred to as *YANG*, B_V=4), and iii) simulcast scheme (Merkle et al., 2007) (*i.e.*, coding each view independently with no inter-view prediction, referred to as *simulcast*). We use the joint multiview video coding (JMVC) software (v.8.5) [3] for encoding/decoding the data sets. In this paper, the performance of competing approaches is evaluated by rate-distortion (in dB/Kbps) of MVC and average transmission bit-rate (in KBps) of interactive views. The quantization parameter (QP) is set to 24, 28, 32 and 36.

3.3 Results & Discussion

Fig. 2(a-d) show the RD performance using competing approaches with the data sequences in Section 3.1 at different QPs (24, 28, 32, 36). It can be shown that the proposed MVC scheme gives comparable RD performance as compared to the *MVC-HBP* scheme. Whereas, the proposed MVC scheme outperforms the *YANG* and simulcast schemes by 10% and 37%, respectively, in terms of RD performance. This improvement is accomplished as the *YANG* scheme uses less inter-view predictions and the simulcast scheme has no inter-view predictions.

Fig. 2(e-h) show the average transmission bit-rate with competing approaches with the data sequences in Section 3.1 at different QPs (24, 28, 32, 36). It can be shown that the proposed MVC scheme outperforms the *MVC-HBP* scheme by 15%. Whereas, the *YANG* and simulcast schemes outperform the *MVC-HBP* scheme by 27% and 53%, respectively, in terms of average transmission bit-rate. This improvement of the *YANG* and simulcast schemes is achieved at the cost of their RD performance shown above. Recall that the RD performance and the average transmission bit-rate are trade-offs.

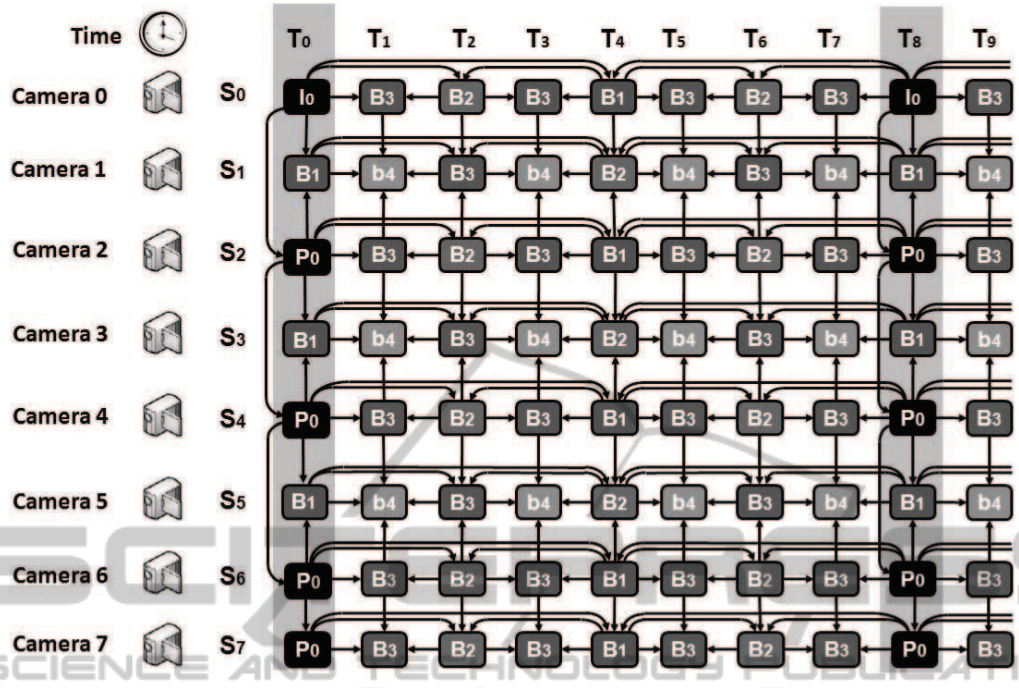
4 CONCLUSIONS

In this paper, we modify the inter-view prediction structure of the MVC standard scheme. This modi-

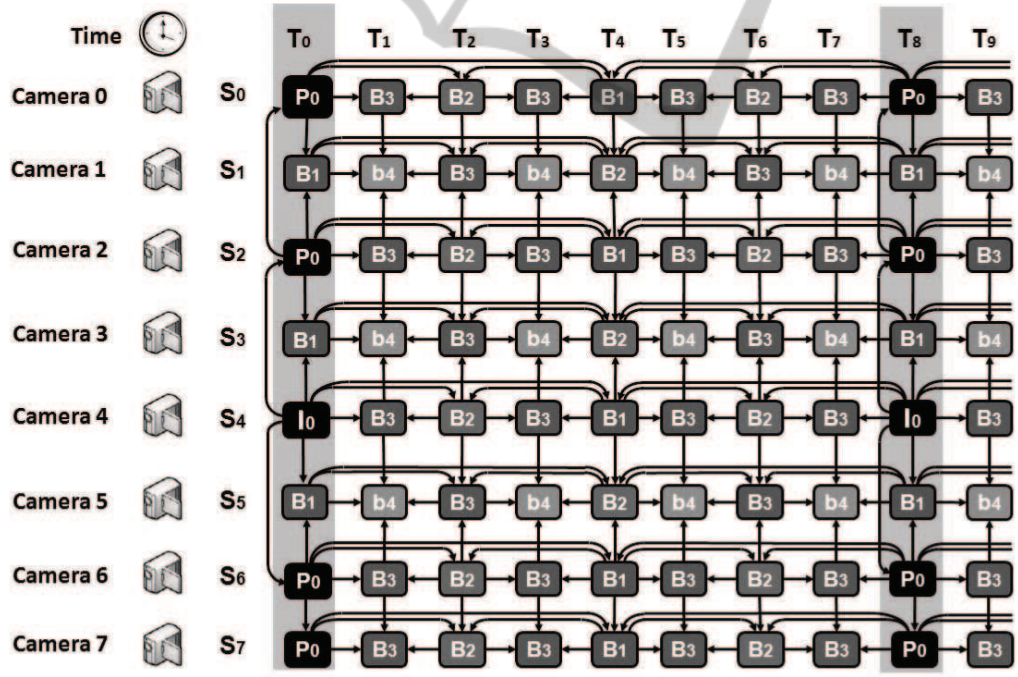
[1] ftp://ftp.merl.com/pub/avetro/mvc-testseq

[2] ftp://ftp.ne.jp/KDDI/multiview

[3] http://mpeg.chiariglione.org/standards.php



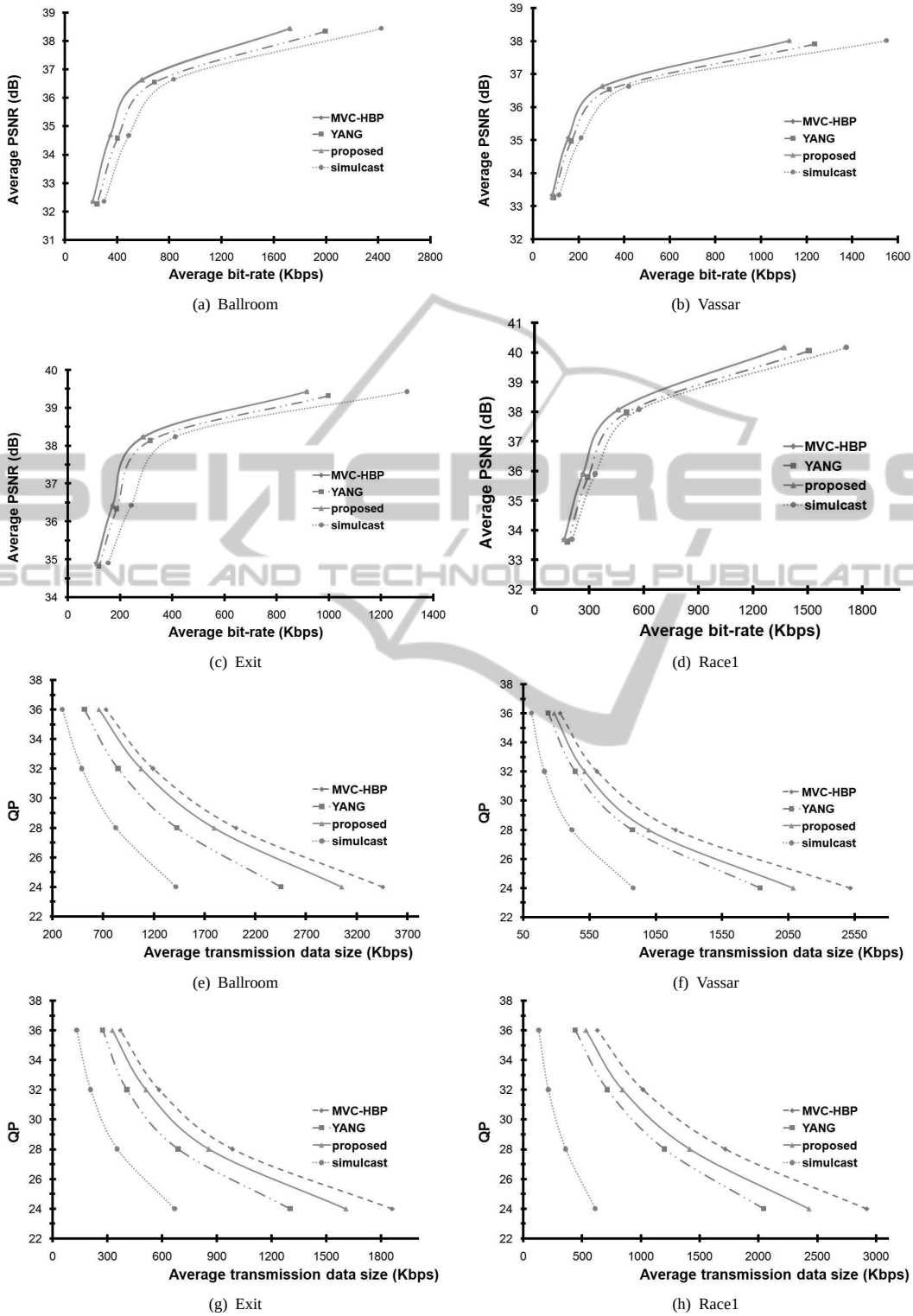
(a) The *MVC-HBP* standard scheme(Merkle et al., 2007)



(b) The proposed MVC scheme

Figure 1: Competing MVC schemes use 8 cameras with 8 frames in a GOP.

fication is based on setting the base view of the proposed MVC scheme to a median view. Clear improvement is shown using the proposed scheme in term of RD performance by an average improvement of 10% and 37% compared to the *YANG* and simulcast schemes, respectively. Whereas, the proposed MVC scheme provides a comparable RD performance compared to the MVC standard scheme. Unlike the *YANG* and simulcast schemes, the proposed MVC scheme decreases the average transmission bit-rate for view's



(a) Ballroom (b) Vassar

(c) Exit (d) Race1

(e) Ballroom (f) Vassar

(g) Exit (h) Race1

Figure 2: (a-d)RD performance, (e-h) The average transmission bit-rate, with competing approaches at different QPs (24, 28, 32, 36).

Table 1: Data sets description.

Data set	Sequences	Object Motion	Resolution	Format	Camera Arrangement	Total file size of 8 views for 10 sec.	Bit-rate (Mbps)
MERL [1]	Ballroom Vassar Exit	Medium Low High	640x480 (rectified) 25fps	4:2:0	8 cameras, 20cm inter-spacing, 1D/parallel	878(MB)	702.4
KDDI [2]	Race1	High	640x480 (non-rectified) 30fps	4:2:0	8 cameras, 20cm inter-spacing, 1D/parallel	1.02 (GB)	835.6

Table 2: Number of extracted frames at different base views to retrieve a specific view of one GOP. The less the number, the well choice the base view, the better the scheme performance.

View	Base view 0,7	Base view 1,6	Base view 2,5	Base view 3,4
S_0	I+7B	I+P+14B	I+P+7B	I+2P+7B
S_1	I+P+22B	I+7B	I+P+22B	I+2P+22B
S_2	I+P+7B	I+P+22B	I+7B	I+P+7B
S_3	I+2P+22B	I+P+7B	I+P+22B	I+P+22B
S_4	I+2P+7B	I+2P+22B	I+P+7B	I+7B
S_5	I+3P+22B	I+2P+7B	I+2P+22B	I+P+22B
S_6	I+3P+7B	I+3P+22B	I+2P+7B	I+P+7B
S_7	I+4P+14B	I+3P+7B	I+3P+14B	I+2P+14B
Total	8I+16P+108B	8I+13P+108B	8I+11P+108B	8I+10P+108B

interactivity by 15%, as opposed to the MVC standard scheme, which is a reasonable gain and not at the cost of the RD performance.

REFERENCES

Cheung, G., Cheung, N., and Ortega, A. (2009). Optimized frame structure using distributed source coding for interactive multiview video streaming. In *IEEE Intern. Conf. on Image Proc.*, pages 1389–1392.

Cheung, G., Ortega, A., and Cheung, N. (2011). Interactive streaming of stored multiview video using redundant frame structures. *IEEE Trans. on Image Proc.*, 20(3):744–761.

Kalva, H., Christodoulou, L., Mayron, L., Marques, O., and Furht, B. (2006). Challenges and opportunities in video coding for 3D TV. In *Proc. IEEE Inter. Conf. Multimedia & Expo*, pages 1689–1692.

Kubota, A., Smolic, A., Magnor, M., Tanimoto, M., Chen, T., and Zhang, C. (2007). Multiview imaging and 3DTV. *IEEESignal Processing Magazine*, 24(6):10–21.

Kurutepe, E., Civanlar, M., and Tekalp, A. (2007). Client-driven selective streaming of multiview video for interactive 3DTV. *IEEE Trans. Cir. and Sys. for Video Tech.*, 17(11):1558–1565.

Lee, S., Wey, H., Park, D., and Kim, D. (2010). Multi-view prediction structure for free viewpoint video. In *IEEE Intern. Conf. on Image Proc.*, pages 3409–3412.

Maugey, T., Frossard, P., and Cheung, G. (2012). Consistent view synthesis in interactive multiview imaging. In *IEEE Intern. Conf. on Image Proc.*, pages 2717–2720.

Merkle, P., Smolic, A., Muller, K., and Wiegand, T. (2007). Efficient prediction structures for multiview video coding. *IEEE Trans. on Cir. and Sys. for Video Tech.*, 17(11):1461–1473.

Mùandller, K., Merkle, P., and Wiegand, T. (2011). 3D video representation using depth maps. *Proceedings of the IEEE*, 99(4):643–656.

Schwarz, H., Marpe, D., and Wiegand, T. (2006). Analysis of hierarchical B pictures and MCTF. In *Proc. IEEE Inter. Conf. Multimedia & Expo*, volume 6, pages 1929–1932.

Vetro, A., Pandit, P., Kimata, H., Smolic, A., and Wang, Y. K. (2008). Joint draft 9.0 multi-iview video coding. Technical report, JVT Doc. JVT-AB204, Hannover, Germany.

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

Yang, Y., Dai, Q., Jiang, G., and Ho, Y. (2010). Comparative interactivity analysis in multiview video coding schemes. *ETRI Journal*, 32(4):566–576.