A Novel Rate Control Technique for Multiview Video Plus Depth Based 3D Video Coding

Yanwei Liu, Qingming Huang, *Senior Member, IEEE*, Siwei Ma, Debin Zhao, Wen Gao, *Fellow, IEEE*, Song Ci, *Senior Member, IEEE*, and Hui Tang

***Abstract*—This paper presents a novel rate control technique for multiview video plus depth (MVD) based 3D video coding. In the proposed rate control technique, an image-stitching method is first utilized to simultaneously encode video and depth, and then a joint rate control algorithm for MVD is presented. The joint rate control algorithm is performed on three levels, namely view level, video/depth level and frame level. In the view level, different proportions of rates are allocated for different types of views according to the pre-statistical rate allocation. In the video/depth level, the target rates for video and depth are discriminatorily assigned to guarantee the high quality of video for the backward-compatible display. In the frame level, the hierarchical rate allocation is used to regulate the target bits for each frame. In addition to the above mentioned rate control strategies, according to the special characteristics of multiview hypothetical reference decoder (HRD), the buffer-related rate control is also considered to prevent the decoder buffer from overflow or underflow even outputting multiple views. Experimental results show that the proposed rate control technique can accurately control the bit-rate to satisfy the requirements of 3D video systems.**

***Index Terms*—Multiview video coding, rate control, video plus depth representation, 3D video coding.**

I. Introduction

WITH today's rapid development of multiview imaging and 3D display technologies, 3D video is increasingly becoming popular. As its depth enhanced viewing experience, 3D video has been recognized as one of the essential parts of the next-generation visual media. With the feature of depth perception, 3D video can enable 3DTV and accordingly change the consumer content experience. Recently, the industry society formed an alliance to speed the commercialization of 3D into homes [1], and 3D video contents are gradually moving into homes by 3D broadcast, 3D-DVD/Blu-ray disk, Internet, etc [2]. Moreover, with the popularization of mobile phone supporting the switchable 2D/3D stereoscopic display, 3D video services being brought into mobile is also becoming a reality [3].

3D video provides the third dimension information by rendering the real world scene. Because of the diversity of the real world, there exist various scene representations to build 3D video, such as model-based representation [4] and point sample based representation [5]. Among these representations, it is widely accepted that the representation of multiview color video plus its associated depth can provide both good rendering quality and flexible processing capability [6], [7]. Moreover, depth map or disparity map can efficiently describe the actual 3D information [8]–[10], while providing the backward compatibility with the existent 2D video. Therefore, MVD is possible to be used in the future 3D video applications. Recently, the exploration experiments of MVD based 3D video are in process in MPEG [11].

Due to the enormous amount of data, MVD generally needs being compressed. Multiview color video can be separately compressed using H.264/AVC or jointly compressed using the emerging multiview video coding (MVC) by exploiting the redundancy between views [12], [13]. The depth map is often compressed to save the channel bandwidth or storage space. Since the depth map sequence is often treated as gray scale image sequence, which is similar to the luminance component of color video sequence, it can also be compressed by H.264/AVC or H.264/MVC [14].

For auto-stereoscopic display, the advanced MVD-based 3D video system [2] is shown in Fig. 1. Suppose the system includes nine views for auto-stereoscopic display. To reduce the transmitted data amount, only the videos and depth data of three views (V1, V5 and V9) need to be compressed and transmitted. The other views are synthesized at the receiver side using depth-image-based rendering (DIBR) technologies. At the display end, the user can select two arbitrary adjacent views, which correspond to the user's left and right eyes respectively, to perceive the 3D impression.

In Fig. 1, three original views are encoded by MVC encoder, and the three views are sequentially encoded with one encoder. The coding structure is shown in Fig. 2. P-view is encoded

Manuscript received July 11, 2010; revised December 28, 2010; accepted January 04, 2011. Date of publication January 31, 2011; date of current version May 25, 2011. This work is supported in part by the National Natural Science Foundation Research Program of China under Grant 60736043, Grant 61025011, and Grant 60972083, and by the National Basic Research Program of China (973 Program, 2009CB320905).

Y. Liu and H. Tang are with the Institute of Acoustics, Chinese Academy of Sciences, Beijing 100190, China (e-mail: ywliu@jdl.ac.cn, liuyw@hpnl.ac.cn; tangh@hpnl.ac.cn).

Q. Huang is with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China, and also with the Graduate University of the Chinese Academy of Sciences, Beijing 100039, China (e-mail: qmhuang@jdl.ac.cn).

S. Ma and W. Gao are with the National Engineering Laboratory for Video Technology, Peking University, Beijing 100871, China (e-mail: swma@jdl.ac.cn, wgao@jdl.ac.cn).

D. Zhao is with the Department of Computer Science, Harbin Institute of Technology, Harbin 150001, China (e-mail: dbzhao@jdl.ac.cn).

S. Ci is with the High Performance Network Lab, Institute of Acoustics, Chinese Academy of Sciences, Beijing 100039, China, and also with the Intelligent Ubiquitous Computing Lab, University of Nebraska-Lincoln, Lincoln 68588 USA (e-mail: cisong@ieee.org).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TBC.2011.2105652

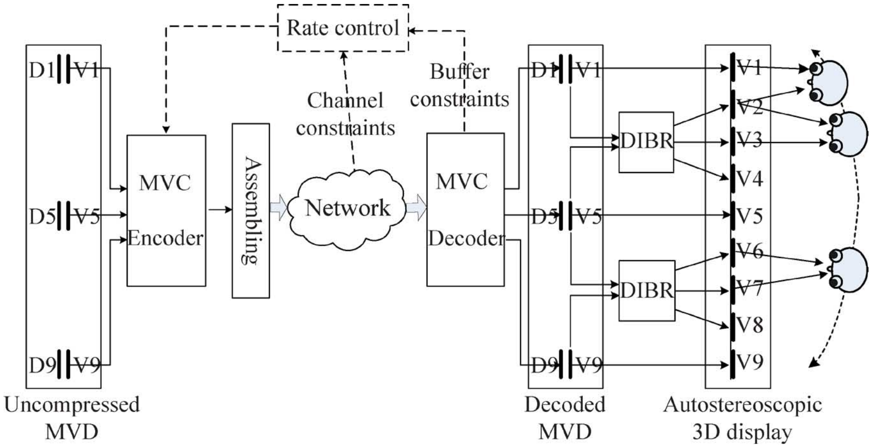


Fig. 1. Advanced 3D video system with multiview video plus depth representation (V: view/image; D: depth).

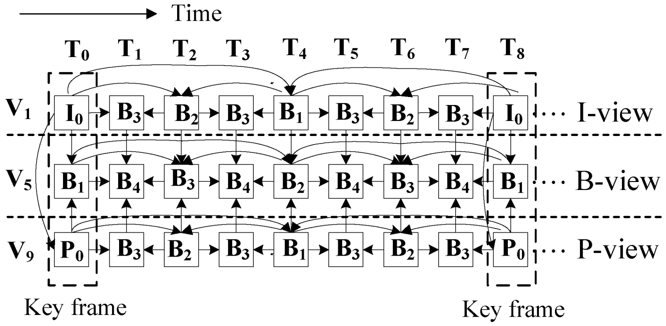

Time
T0 T1 T2 T3 T4 T5 T6 T7 T8
V1 I0 B3 B2 B3 B1 B3 B2 B3 I0 ··· I-view
V5 B1 B4 B3 B4 B2 B4 B3 B4 B1 ··· B-view
V9 P0 B3 B2 B3 B1 B3 B2 B3 P0 ··· P-view
Key frame
Key frame

Fig. 2. MVC coding structure with hierarchical B pictures.

with unidirectional inter-view prediction from the reconstructed I-view. Likewise, B-view is encoded with bidirectional inter-view prediction from the reconstructed I-view and P-view. After encoding, one stream, which is assembled from the three compressed bit-streams, is transmitted to the receiver. At the receiver side, the assembled stream is directly decoded and three reconstructed video sequences are simultaneously outputted.

Since MVD-based 3D video system is usually deployed on the network, the 3D video coding must satisfy the channel bandwidth constraint and the decoder buffer constraint. In this regard, the rate control in 3D video coding is the essential component to meet these constraints. Various rate control algorithms have been proposed for 2D video coding, such as MPEG-2 TM5 [15], H.263 TMN8 [16] and H.264 rate control algorithm [17], [18]. To the best of our knowledge, there are very few rate control works for 3D video coding. Lim et al. first proposed a 3D multiview rate control algorithm based on human visual system [19]. For video plus depth representation, the off-line rate control methods using evolution strategy [20] or Lagrangian optimization [21] are proposed. However, these 3D video rate control algorithms are not used for MVD-based 3D video coding, and they also do not aim at 3D video coding with the new H.264/MVC standard.

In this paper, we propose a novel 3D video rate control technique, which simultaneously controls the bit-rates of video and depth. The proposed rate control technique is performed on three levels, namely view level, video/depth level and frame level. In the view level, the target rates are allocated according to different types of views with a pre-statistical rate allocation proportion. In the video/depth level, the target rates for video and depth are discriminatorily assigned to guarantee the high quality of video. In the frame level, the rate control algorithm for hierarchical B pictures based MVC is designed. In addition to the channel rate constraint, multiview HRD requirement is also considered to greatly reduce the risks of the decoding buffer overflow or underflow even when the decoder simultaneously outputs the multiview data.

The rest of the paper is organized as follows. Section II provides the joint video/depth coding method for 3D video. The detailed 3D video rate control technique is described in Section III. Section IV shows the experimental results. Finally, the conclusions are given in Section V.

II. Joint Video Plus Depth Based 3D Video Coding

Multiview video and depth data are usually obtained at the receiver for directly displaying or DIBR. In order to avoid the display or rendering delay, the video and depth are often required to be synchronously obtained. In the current multiview video and depth coding, the video and depth are separately compressed using two independent encoders. The bit-rates of video and depth are also separately controlled. Though this method can maintain the total bit-rate of video and depth within the constraint, it cannot simultaneously control the bit-rates of video and depth. In the 3D video applications, the rate allocation needs maintaining an imbalanced proportion between video and depth to keep higher quality of video. Since the bit-rate fluctuations will occur during the rate control, the separate controlling of video and depth bit-rates does not guarantee the consistent rate proportion between video and depth at any time instance. Accordingly, the rendered virtual view cannot maintain the smooth visual quality in the temporal domain.

To facilitate the simultaneous controlling of the bit-rates of video and depth, we propose to encode the video and depth using a single MVC encoder. Specifically, the video image and depth image are stitched into one combined image to be encoded, as shown in Fig. 3. This kind of jointly coding method can simultaneously control the video bit-rate and depth bit-rate, and further provide the consistent rate proportion between video and depth

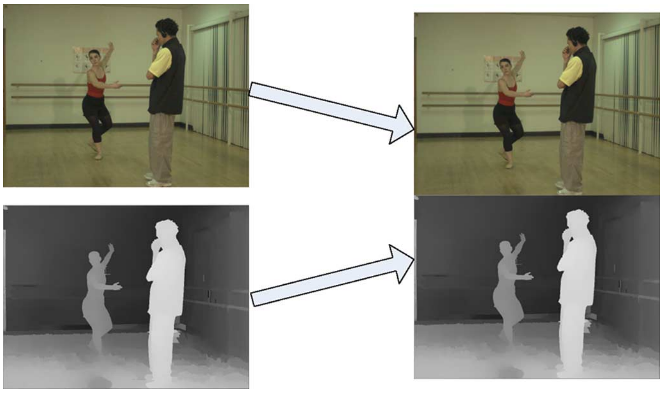


Fig. 3. Image stitching process of video and depth for the *Ballet* sequence.

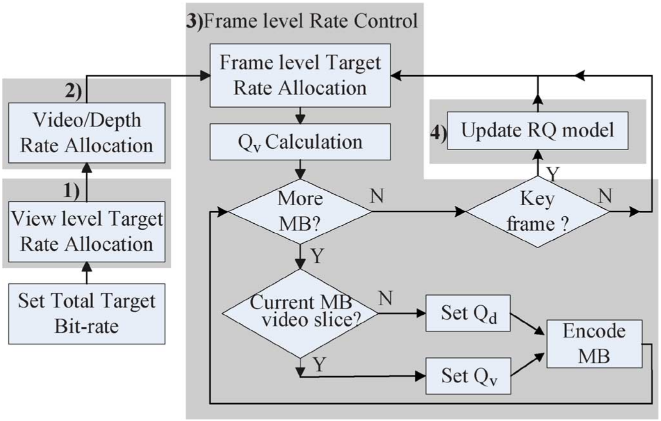
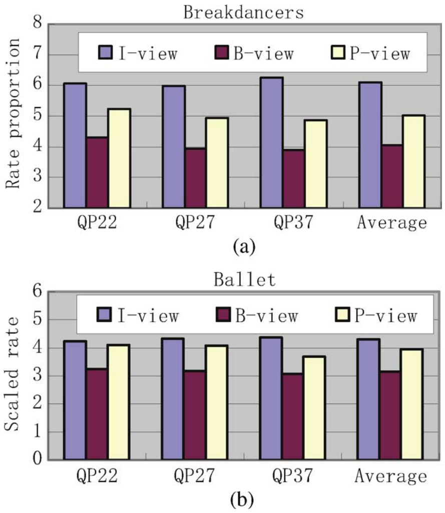

3)Frame level Rate Control
Frame level Target Rate Allocation
2)
Video/Depth Rate Allocation
Q_v Calculation
4) Update RQ model
1)
View level Target Rate Allocation
More MB?
N
Key frame ?
N
Y
Y
Set Total Target Bit-rate
Current MB video slice?
N
Set Qd
Encode MB
Y
Set Qv

Fig. 4. Block diagram of the proposed 3D video rate control algorithm.

TABLE I
SUMMARY OF NOTATIONS AND ACRONYMS

Variables	definition
RD	Rate-distortion
RQ	Rate-quantization
RCGOP	A rate control GOP which includes several GOPs
RC key frame	First key frame in a RCGOP
CPB	Coded picture buffer
DPB	Decoded picture buffer
R_v	Target bit-rates of video
R_d	Target bit-rates of depth
Q_v	Quantization parameters for video
Q_d	Quantization parameters for depth
$Q_{v,\ step}$	Quantization stepsize for video sequence
$Q_{d,\ step}$	Quantization stepsize for depth sequence
W	Available channel bandwidth
f	Frame rate



(a)

(b)

Fig. 5. Rate allocation proportion among the three views for 2D video coding.

for all frames in the entire sequence. Though it consumes the double encoding buffers due to the enlarged resolution of the combined image, the joint coding method does not increase the total encoding complexity of video and depth.

When encoding the combined sequence of video and depth, the video and depth in one image will be partitioned into two slices. Hence, the video and depth can be packaged into different packets so that they can be independently transmitted as well as jointly transmitted. In order to improve the coding efficiency, hierarchical B pictures with inter-view prediction is utilized in the 3D video coding due to its high temporal prediction efficiency, as shown in Fig. 2. The multiple views of V1, V5 and V9 are coded as I-view, B-view and P-view, respectively. In each view, the first frame of each group of pictures (GOP) is called the key frame. For P-view, only the key frame is inter-view predicted. For B-view, all frames in a GOP are inter-view predicted.

III. 3D Video Rate Control Algorithm

According to the joint video/depth coding, we propose a joint video/depth rate control technique. The block diagram of the proposed 3D video rate control technique is shown in Fig. 4. It consists of four stages: 1) view level rate allocation stage; 2) video/depth rate allocation stage; 3) frame level rate control stage; 4) RQ model updating stage. In the proposed 3D video rate control, the first stage is off-line processed and the latter three stages are online processed. To facilitate understanding, the summary of some notations and acronyms is shown in Table I.

A. View Level Rate Allocation

In the 3D video systems, multiview video and depth are encoded by MVC method, as shown in Fig. 2. In MVC, different inter-view predictions can lead to different RD performances, e.g., I-view will consume a higher bit-rate than P-view or B-view at the same visual quality. According to the statistical observation, the average rate proportion among the three views, I-view, B-view and P-view, is approximately equal to 6: 4: 5 for *Breakdancers* and 4.3: 3.1: 3.9 for *Ballet*, respectively. Fig. 5 shows the statistical rate proportion among the three views. This rate proportion highly depends on the inter-view coding structure and image resolution. For the multiview sequences with specific resolution and the known inter-view prediction structure, we can obtain the approximate rate proportion among views through the statistical method after pre-encoding several frames in each view. The current MVC encodes each view separately and when the view is encoded with inter-view prediction, the inter-view reference sequences are fed into the encoding buffer. For such coding mode, the multiview sequences are not simultaneously encoded. Thus, under the total rate constraint, the view level rate allocation is needed to be offline performed before the actual encoding.

B. Video/Depth Rate Allocation

In 3D video systems, the virtual views and captured views are formed into many view-pairs for auto-stereoscopic display. The compression of multiview video and depth has a great effect on the virtual view rendering quality [14]. Under the constraint of total bit-rate for video and depth of each view, the different rate allocations between video and depth can lead to different rendering qualities of the virtual view.

In practical 3D video applications, depth map is only a kind of side information for virtual view rendering. In contrast, the 2D video usually needs to maintain higher quality for the purpose of being compatible with 2D video display. Hence, compared with 2D video, the depth sequence can be compressed at very low bit-rate to meet the constraint of channel bandwidth. In our 3D video rate control, we define

$$R_d = \alpha R_v \tag{1}$$

where α is a value within [0,1], which represents the rate proportion between depth and video.

According to the RD characteristics of H.264/AVC, there are two main models to characterize the RQ relation, linear RQ model [22] and quadratic RQ model [17], [23]. The linear model can derive a simpler relation between Q_d and Q_v than the quadratic model. For simplicity, the linear RQ model is used in the video/depth rate allocation stage and the quadratic RQ model is used in the frame level rate allocation. The linear relationship between R_v and $1/Q_{v,step}$ is expressed as

$$R_v = \frac{K_v}{Q_{v,\, step}} + C_v \tag{2}$$

where K_v and C_v are constants. Since MVC is an extension of H.264/AVC, the linear relationship of R_v and $1/Q_{v,step}$ is also suitable for hierarchical B pictures based MVC.

Since depth map is a kind of range data that represents the scene information, it has very similar structure with the color video. In addition, the temporal motion in the color video is likewise embodied in the depth map sequence. Based on this similarity, the relationship between R_d and $1/Q_{d,step}$ is also linear, which can be written as

$$R_d = \frac{K_d}{Q_{d,\, step}} + C_d \tag{3}$$

where R_d and K_d are constants. The relationship between R_d and $1/Q_{d,\, step}$ for *Breakdancers* is illustrated in Fig. 6.

According to (1)–(3), we have

$$\frac{K_d}{Q_{d,\, step}} + C_d = \alpha\left(\frac{K_v}{Q_{v,\, step}} + C_v\right) \tag{4}$$

Further, because $Q_{d,step} = 2^{(Q_d-4)/6}$ and $Q_{v,step} = 2^{(Q_v-4)/6}$, we can get

$$Q_d = Q_v + \left\lceil 6\log_2\left(\frac{K_d}{\alpha K_v + (\alpha C_v - C_d)\cdot 2^{(Q_v-4)/6}}\right)\right\rceil \tag{5}$$

where $\lceil x \rceil$ denotes the smallest integer not less than x. In order to limit Q_d within [1, 51], we further have

$$Q_d = \max\{1,\ \min\{51,\ Q_d\}\} \tag{6}$$

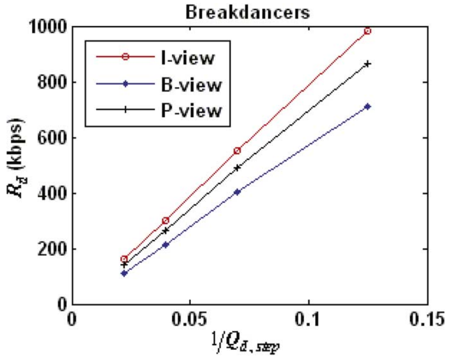


Fig. 6. Relationship between R_d and $1/Q_{d,\, step}$ for the multiview depth.

In (5), K_v, C_v, K_d and C_d are sequence-dependent constants. They are initialized after pre-encoding some frames in view level rate allocation stage and then updated once every other key frame by a linear regressive way [23] in the process of rate control.

C. Frame Level Rate Control

For the general IBBP coding structure, the quadratic RQ rate control algorithm has been implemented in H.264/AVC [17], [18], and for hierarchical B pictures coding, the rate control is further studied in [24]. However, the existing rate control cannot efficiently work for the hierarchical B pictures based MVC. One reason is that the rate allocations of B frames are not carefully considered in the rate control algorithm. Another reason is that MVC is a bit different from the pure hierarchical B pictures coding due to the inter-view prediction. Especially in P-view and B-view of MVC, the key frames are not I frames, they are inter-view predicted B frames or P frames. Based on the different frame types and prediction structures, we propose the following frame level MVC rate control algorithm.

In Fig. 2, after performing the view level rate allocation, each view is encoded with independent rate control. For each view, the hierarchical target bits allocation is first performed and then the target bits for each frame are regulated in terms of the coding complexity and buffer constraints. According to the quadratic RQ model, the quantization parameter for each frame is finally computed.

The key frames have different frame types in different views and their bits regulations take a key role in the total rate control. In order to facilitate the rate regulations of key frames, several GOPs compose one RCGOP, which is a rate control unit between the GOP-level and the sequence-level. With the definition of RCGOP, the rate allocation accuracy for one sequence will be increased due to the refined control unit segmentation. Assuming that the target rate for RC key frame is R_{RCK}, the target rate of each other key frame in the RCGOP is R_K, and the target rate for B frame in temporal hierarchical level l is R_B^l, $R_{RCK} = \omega_{RCK} R_K$ and $R_B^l = \omega_B^l R_K$, where ω_{RCK} and ω_B^l are scaling factors.

Since the key frames including both the RC key frames and the other key frames in the same view have almost the same position in a prediction/reference sense, we set $\omega_{RCK} = 1.0$. In our experiment, to obtain the other scaling factors, we do a

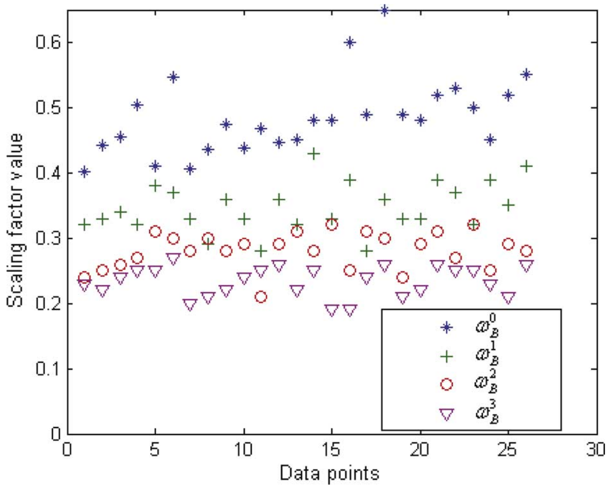


Fig. 7. Scaling factors for different frame types.

statistical pre-analysis for multiple sequences including multiview videos and depths, as shown in Fig. 7. According to the pre-analysis, we empirically set

$$\omega_B^l = \begin{cases} 0.5, & \text{if } l = 0 \\ 0.3, & \text{if } l = 1 \\ 0.3, & \text{if } l = 2 \\ 0.2, & \text{if } l = 3. \end{cases} \tag{7}$$

These empirical scaling factors work well for most sequences though they are sequence-dependent. The optimal scaling factors can be derived from the hierarchical prediction filters to increase the accuracy of rate allocation.

To perform a reasonable target rate allocation for different frame types in hierarchical coding, we first compute the rate of the key frame, and then obtain the rates of other types of frames using the corresponding scaling factors. According to the total rate for a RCGOP, we can get the number of the key frames which includes the number translated from other types of frames in rates, and then obtain the key frame rate allocation equation. Specifically, given the target bits per frame R, the number of total GOPs N_G in the current RCGOP, and the temporal level number L between two adjacent key frames, we have

$$R_K = \frac{R \times N_G \times (N_B + 1)}{\left(\sum_{l=0}^{L} n_B^l \omega_B^l + 1\right) \times N_G + \omega_{RCK} - 1} \tag{8}$$

where n_B^l is the number of B frames at temporal level l, and N_B is the total number of B frames between two adjacent key frames.

In the process of rate control, the target rate allocation must be regulated in real time. Assume we have encoded several frames in a RCGOP, and then we regulate the following target rate allocation for the each remaining frame. Considering the remaining bits for the total remaining frames, we compute a new ratio to scale the original pre-allocated target rate for different types of frames to get a more accurate target rate allocation than before. The scaling ratio can be computed as the remaining bits devised by the pre-allocated bits for all remaining frames. According to the above analysis, the first candidate target bits for the j^{th} frame in the i^{th} RCGOP are calculated as

$$\hat{T}_i(j) = \frac{R_{curr} \times B_i(j)}{N_{RCK,r} R_{RCK} + N_{k,r} R_k + \sum_{l=0}^{L} N_{B,r}^l R_B^l} \tag{9}$$

where $N_{RCK,r}$, $N_{k,r}$ and $N_{B,r}^l$ are the numbers of the remaining RC key frames, other key frames and B frames at hierarchical level l, respectively. $B_i(j)$ denotes the total remaining bits for encoding the current frame and the rest frames. R_{curr} is set to R_{RCK}, R_K and R_B^l depending on the type of the current frame (RC key frame, key frame and B frame).

Considering the buffer constraints, the second candidate target bits for the j^{th} frame in the i^{th} RCGOP are calculated as

$$\tilde{T}_i(j) = \frac{W}{f} + \gamma \times (CB_i(j) - TB_i(j)) \tag{10}$$

where γ is usually 0.5 [18], $CB_i(j)$ denotes the current buffer fullness, and $TB_i(j)$ denotes the target buffer level.

The target frame bits are finally a weighted combination of $\tilde{T}_i(j)$ and $\hat{T}_i(j)$ by

$$T_i(j) = \beta \times \hat{T}_i(j) + (1 - \beta) \times \tilde{T}_i(j) \tag{11}$$

where β is often set to 0.5 [18]. In the equation above it is assumed that the current frame is the key frame. Since (11) is based on the quadratic RQ model derived from the key frame statistics, if the current frame is not a key frame, empirically, the target bits can be scaled from (11) by

$$T_i(j) = \begin{cases} T_i(j)/\theta_1, & \text{if the frame is B frame} \\ T_i(j)/\theta_2, & \text{if the frame is RC key frame.} \end{cases} \tag{12}$$

For simplicity, the above scaling factors θ_1 and θ_2 are empirically set to 0.4 and 4.0, respectively [24]. The perfect solution is establishing different RQ models for different frame types and then the frames with the same frame type can share their own RQ model.

Based on the allocated target bits, the quantization stepsize $Q_{step,i}(j)$ can be computed by the quadratic RQ model [23]

$$T_i(j) = a_1 \times \frac{MAD_i(j)}{Q_{step,\,i}(j)} + a_2 \times \frac{MAD_i(j)}{Q_{step,\,i}^2(j)} + H_i(j) \tag{13}$$

where a_1 and a_2 are the model coefficients, $H_i(j)$ is the sum of header bits and motion bits, $MAD_i(j)$ is a prediction of the j^{th} mean absolute difference (MAD) between the original image data and its prediction data in the i^{th} RCGOP by means of a linear model,

$$MAD_i(j) = c_1 \times MAD_i(j - 1 - m) + c_2 \tag{14}$$

where c_1 and c_2 are model parameters, and their initial values can be set to 1 and 0, respectively. In (14), $MAD_i(j - 1 - m)$ is the actual MAD between the original $(j - 1 - m)^{th}$ frame and its prediction image data, and m is the number of B frames between two successive key frames.

D. RQ Model Updating

The proposed rate control algorithm is based on the quadratic RQ model. To accurately characterize the RD relationship, the model coefficients a_1 and a_2 in (13), and c_1 and c_2 in (14) need to be updated once every other key frame using a linear regressive technique [23]. In the RQ model updating stage, only the RQ model of video part is updated. The rate and distortion of the depth part are indirectly reflected on the RQ model of the video part in terms of the relationship between Q_v and Q_d.

E. Multiview HRD Consideration

Due to the huge amount of data in 3D video applications, the decoder must prevent the resource buffer from overflow or underflow. Especially, when the set-top box is adopted in the receiver side, the decoding buffer needs to be more carefully controlled because of the limited storage resources. Hence, the 3D video rate control must take into account the buffer requirement.

In H.264/AVC, HRD is used to constrain the variations of bitrate to meet the buffer requirement of decoder. A mathematical model, also known as leaky bucket, is employed to characterize HRD. It includes a CPB, a DPB and an instantaneous decoding process. Currently, the multiview HRD process in MVC JD 8.0 [25] allows checking the conformances of decoders when only one view is outputted. However, besides that, MVC also supports checking the conformances of decoders with outputting more than one view. When multiple views are simultaneously outputted, the operations related to CPB and DPB are repeatedly invoked for each view in view decoding order. In current 3D video applications, all views separately encoded are jointly decoded and simultaneously outputted. This situation requires both the encoder and decoder to synchronize reference picture buffer resources on both sides. Thus, the bit-rate must be constrained in the encoder to guarantee that DPB never overflow or underflow when not only outputting one single view, but also outputting multiple views.

For the constant bit-rate control, the bits allocated to the j^{th} frame in the i^{th} RCGOP must not surpass an upper bound $U_i(j)$ and also not be lower than a lower bound $L_i(j)$. $U_i(j)$ and $L_i(j)$ are defined as

$$L_i(j) = \begin{cases} B_{i-1}(N_{i-1}) + \frac{W}{f}, & \text{if } j = 1 \\ L_i(j-1) + \frac{W}{f} - b_i(j-1), & \text{otherwise.} \end{cases} \tag{15}$$

and

$$U_i(j) = \begin{cases} B_{i-1}(N_{i-1}) + 0.9 \times t_{r,1}(1), & \text{if } j = 1 \\ U_i(j-1) + 0.9 \times \left(\frac{W}{f} - b_i(j-1)\right), & \text{otherwise} \end{cases} \tag{16}$$

where $B_{i-1}(N_{i-1})$ denotes the total bits for the rest frames in the $(i-1)^{th}$ RCGOP when only the first frame is coded, $b_i(j-1)$ is the actual generated bits in the $(j-1)^{th}$ coded frame, $t_{r,1}(1)$ is the removal time of the first frame from the CPB. Therefore, the target bits in (11) or (12) can be revised as

$$T_i(j) = \max\{\min\{T_i(j), U_i(j)\}, L_i(j)\} \tag{17}$$

For single view coding, the bounds of (15) and (16) aim at preventing the CPB from overflow or underflow, and further guarantee that the DPB never overflow or underflow. In MVC, when encoding I-view, it is reasonable that the target bits are bounded by (17) because only one view can be outputted in the decoder. However, since the decoder is possible to output multiple views when encoding P-view and B-view, the lower bound and upper bound of the target bits need to be revised with considering the multiview DPB. Specifically, when encoding P-view and B-view, the target bits are not only constrained by the single view bounds, but also constrained by another set of bounds which prevents DPB from overflow or underflow when the decoder outputs multiple views.

The other set of bounds for multiview outputting is

$$L_{M,i}(j) = \begin{cases} B_{i-1}(N_{i-1}) + \frac{nW}{f}, & \text{if } j = 1 \\ L_{M,i}(j-1) + \frac{nW}{f} - b_{M,i}(j-1), & \text{otherwise} \end{cases} \tag{18}$$

and

$$U_{M,i}(j) = \begin{cases} B_{i-1}(N_{i-1}) + 0.9 \cdot t_{r,1}(1), & \text{if } j = 1 \\ U_{M,i}(j-1) + 0.9 \\ \quad \cdot \left(\frac{nW}{f} - b_{M,i}(j-1)\right), & \text{otherwise} \end{cases} \tag{19}$$

where $b_{M,i}(j-1) = b_i(j-1) + b_{r,i}(j-1)$, and $b_{r,i}(j-1)$ denotes the actual coded bits of the reference views pictures at the same temporal instance with the $(j-1)^{th}$ coded picture. In (18) and (19), $n = n_r + 1$, where n_r is the number of total reference views in the current view coding. Thus, the final target bits are

$$T_i(j) = \max\{\min\{T_i(j), U\}, L\} \tag{20}$$

where U and L are defined as

$$U = \min(U_i(j), U_{M,i}(j)) \tag{21}$$

and

$$L = \max(L_i(j), L_{M,i}(j)). \tag{22}$$

IV. Experimental Results

In order to evaluate the proposed 3D video rate control algorithm, several experiments are performed with 3D video sequences of *Breakdancers*, *Book Arrival*, *Alt Moabit*, *Mobile* and *Kendo*. The sequence of *Breakdancers* is captured by Microsoft Research [26], and the provided corresponding depth map is obtained by stereo vision method. *Book Arrival* and *Alt Moabit* (1024 × 768) are provided by Fraunhofer HHI, and their depth maps are created by depth estimation reference software, kindly provided by Nagoya University [27]. *Mobile* (720 × 528) is provided by Philips corporation. These three sequences are captured with the static camera setup. Comparably, the *Kendo* sequence is captured by Nagoya University with the moved camera array.

TABLE II
ACCURACY OF THE PROPOSED RATE CONTROL ALGORITHM

Sequence	Target (kbps)	Rate (kbps)	PSNR(dB)		Inaccuracy (%)	Rate ratio (depth/video) (rate control coding)
			video	depth		
GOP size = 8						
Breakdancers	465	479.62	34.92	35.77	3.14	0.46
	2884	2883.11	38.68	40.88	-0.03	0.26
Book Arrival	333	327.71	32.63	34.71	-1.58	0.37
	1516	1533.92	39.16	39.74	1.12	0.43
Alt Moabit	365	373.44	32.89	38.38	2.19	0.24
	1550	1533.70	39.44	43.56	-1.05	0.31
Kendo	446	466.10	34.42	34.64	4.41	0.24
	2300	2318.26	43.75	42.25	0.78	0.21
Mobile	583	593.34	36.14	51.34	1.72	0.26
	1316	1277.96	42.80	57.45	-2.96	0.24
GOP size=15						
Breakdancers	396	418.18	32.04	34.98	5.60	0.44
	1886	1891.45	36.91	40.22	0.28	0.31
Book Arrival	313	314.40	39.36	35.42	0.31	0.29
	1113	1102.11	39.30	34.58	-1.08	0.53
Alt Moabit	245	250.88	34.77	36.69	2.39	0.32
	1013	1002.51	38.82	41.92	-1.06	0.26
Kendo	400	421.57	32.68	33.81	5.25	0.32
	1886	1889.18	43.18	42.10	0.15	0.27
Mobile	470	494.88.10	37.85	48.00	5.29	0.22
	926	968.27	43.05	55.70	4.56	0.26

In the experiments, we use the revised MVC software JMVM6.0 to implement the rate control algorithm. The first three views (view0, view1, view2) for *Breakdancers*, the three middle views (view10, view11, view12) for *Book Arrival* and *Alt Moabit*, and the three middle views (view1, view3, view5) are used to simulate v1, v5, v9 in Fig. 1. In the rate control coding, the temporal GOP size is set to 8 or 15 and the search range is set to 48. Since the original length of the sequences (*Breakdancers*, *Book Arrival*, and *Alt Moabit*) is very short for evaluating the proposed rate control algorithm, each sequence is extended to 300 frames with three duplicate 100 frames. And also, the first 300 frames for the *Kendo* sequence, and the 200 frames for the *Mobile* sequence are used for experiments.

A. 3D Video Rate Control Accuracy

We first confirm the accuracy of the proposed 3D video rate control algorithm. Table II summarizes the matching accuracy between the controlled bit-rates and the target ones. In Table II, the target bit-rate and the actually controlled bit-rate is the average value (video + depth) of three views, and the PSNR value is also the average value of three views.

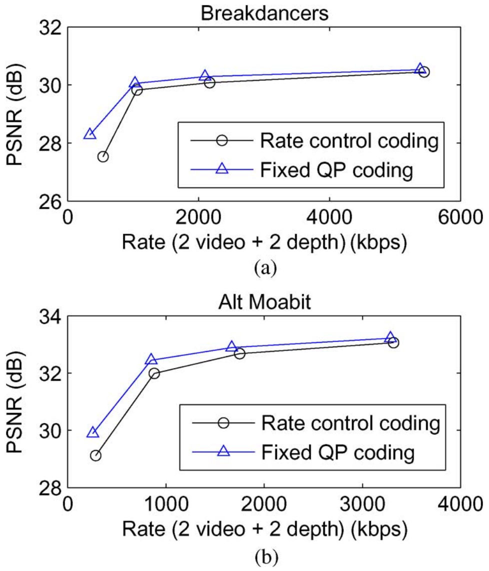
In the experiments, the target rate proportion between depth and video is set to 1:4 ($\alpha = 0.25$), and the actual rate ratio between depth and video after rate control is also shown in Table II. The actual rate ratio is obtained by gathering the rates for all frames of video and depth. It can be seen from the table that the rate control can perform the video/depth rate allocation towards the predefined rate ratio.

Table II indicates that the absolute inaccuracy of the proposed 3D video rate control algorithm is within 3.14% for GOP size of 8 and within 5.6% for GOP size of 15. The proposed rate control not only considers the target rate allocation for the different view levels, hierarchical levels of B frames, and the video/depth level, but also considers the relationship between the quadratic RQ models for different frame types. Through the simple scaling method, the rate allocation can approximately adapt to the RD characteristics of the encoded frame. Consequently, the proposed method can provide a certain degree of rate control accuracy for 3D video coding. Table II has illustrated it.

The current rate control only considers updating the RQ model of key frame, and the RD relations of B frames are scaled by the RD relation of key frame. Hence, the accuracy of the rate control has a certain dependency on the number of key frames in the whole sequence. Moreover, the current rate control only updates the RQ model of the video part, which also partly affects the accuracy of rate control. For the sequences with the slow motion, such as *Book Arrival* and *Alt Moabit*, the proposed rate control can achieve good accuracy and slightly reduce the effect of long GOP size on rate control accuracy.

B. View Synthesis Quality Comparison

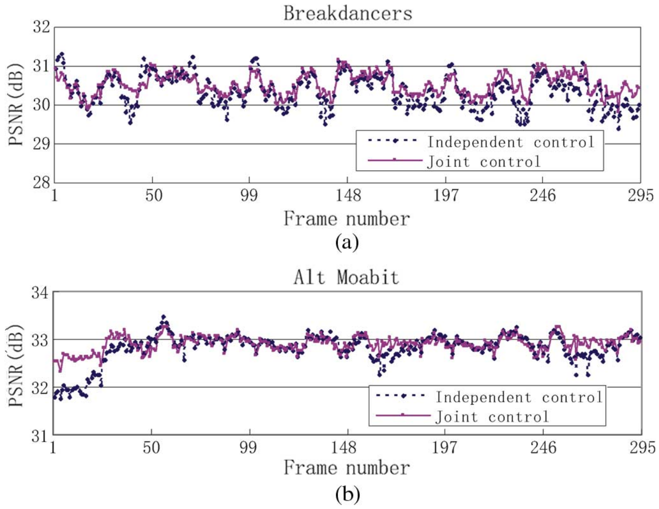
To guarantee the balanced visual qualities cross views, the fixed cascading QP coding [28] is generally used to encode multiview video. Fig. 8 shows the view synthesis RD performance comparison between the proposed 3D video rate control coding and the fixed cascading QP coding. Since the sequences captured by the same provider have the similar geometry characteristics, we select only two sequences from the two different providers to verify the RD performances of the proposed rate control algorithm. In Fig. 8, the virtual view is view1 synthesized by view0 and view2 for *Breakdancers* and the virtual view is view11 synthesized by view10 and view12 for *Alt Moabit*, the basis-QPs in the fixed QP coding are set to 22, 27, 32, 42 for video and 27, 32, 37, 47 for depth, respectively. In the rate control coding, the GOP size is set to 8 and the rate proportion between depth and video is set to 0.25. It can be seen that, compared with the fixed QP coding, the proposed rate control



(a)

(b)

Fig. 8. The view synthesis RD performance comparison between the proposed rate control coding and the fixed cascading QP coding.



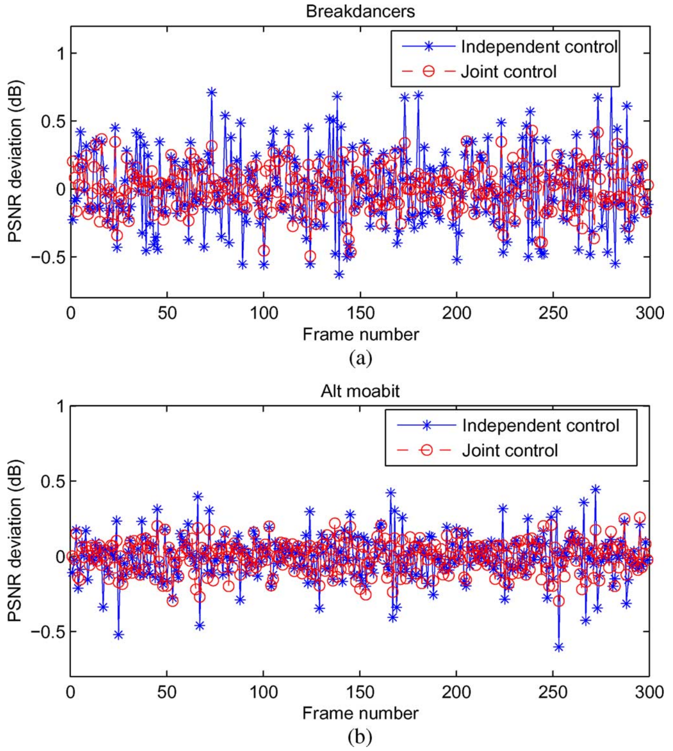
(a)

(b)

Fig. 9. Virtual view quality comparison between the proposed video/depth rate control and the independent video/depth rate control.

method can almost achieve the comparable virtual view synthesis RD performance at high bit-rate, a slightly inferior virtual view synthesis RD performance at low bit-rate.

Under the total rate constraints, the proposed method can control the video bit-rate and depth bit-rate to smooth the visual quality of the virtual view. Fig. 9 shows the virtual view quality comparison between the proposed video/depth rate control and independent rate control. For *Alt Moabit*, the target rate is about 3100 kbps of total rates (2 videos and 2 depths) for synthesizing view11 using view10 and view12, and for *Breakdancers*, the target rate is about 5450 kbps for synthesizing view1 using view0 and view2. It can be seen that the PSNR curve of the virtual view of joint rate control fluctuates lower than that of the independent video/depth rate control. Fig. 10 shows the PSNR differences of consecutive frames for the virtual view with the joint video/depth rate control and independent rate control. we can find that the joint video/depth rate control reduces the virtual view PSNR fluctuation between consecutive frames as compared to the independent video/depth rate control.



(a)

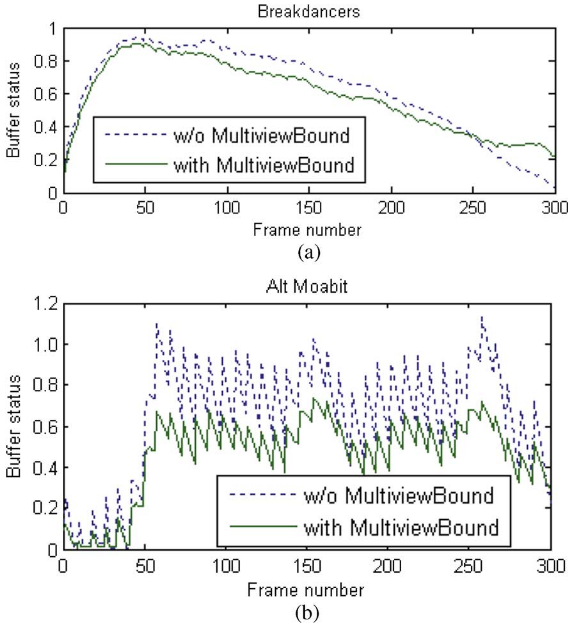
(b)

Fig. 10. Comparison of PSNR difference between the consecutive frames of the virtual view.

In the current research, the original depth is not very accurate and the compression also further deteriorates the depth accuracy. These two kinds of depth inaccuracies produce a lot of distortions in the synthesized view. Comparably, the original depth inaccuracy generates the most of the distortions in the synthesized view. The compression of video and depth holds a lesser effect on the view synthesis quality than the original depth inaccuracy, and therefore the proposed joint rate control regulates the rates of depth and video, and can further obtain smooth, but only a little smoother view synthesis quality than that of the independent video/depth rate control.

C. 3D Video Buffer Control

Besides the bandwidth constraint, the decoding buffer constraint is also taken into account in our rate control algorithm. Our rate control scheme can regulate the number of bits to reduce the risks of the encoder output buffer overflow or underflow. The proposed rate control algorithm allocates the target bits for one encoding frame with considering the current buffer fullness and the target buffer level. Moreover, it has limited the target bits of the encoding frame using additional upper bound and lower bound. The bounds have been considered with that the actual coding bits of the encoding frame are possibly higher than the target allocated bits. The actually used bounds have been zoomed using the scaling parameter, such as 0.9 in (16) or (19). Also, the rate-quantization model is gradually updated and it can accurately reflect the rate and quantization relation in a short time even if the frame is easy or complex to code. Hence,



(a)

(b)

Fig. 11. Buffer status for 3D video rate control coding.

the actual bits for the encoding frame are able to approach the target bits. Thus, the proposed algorithm can provide a certain degree of provision for avoiding buffer overflow or underflow.

The buffer size is set to 0.5 × bit-rate and the coding GOP size is 8. When multiple views are outputted in the decoder, the buffer statuses are simulated in the encoder. Fig. 11 shows the comparison between buffer statuses with and without the multiview bounds for outputting three views. It can be seen that the proposed rate control can maintain suitable buffer occupancy levels to prevent buffer from overflow and underflow for outputting multiple views compared with the independent buffer control of each view.

V. Conclusion

This paper presented a novel rate control technique for MVD based 3D video coding. The proposed rate control technique is performed on three levels, namely the view level rate allocation, video/depth level rate allocation, and frame level rate control. At the video/depth level and view level, the rates are discriminatorily allocated according to the special characteristics of 3D video coding. At the frame level, a new rate control algorithm with multiview HRD consideration for MVC structure is proposed. Experimental results show that the proposed rate control technique can accurately control the bit-rate to satisfy the channel and buffer requirements of 3D video systems.

Acknowledgment

The authors would like to thank Dr. Huifang Sun of MERL for his valuable suggestions, and thank the editors and anonymous reviewers for their valuable comments. The authors also would like to thank Microsoft Research for providing the 3D video sequences of *Breakdancers* and *Ballet*, Fraunhofer HHI for providing the 3D video sequences of *Book Arrival* and *Alt Moabit*, Philips for providing the 3D sequence of *Mobile* and Nagoya University for providing the 3D video sequence of *Kendo*.

References

[1] "3D@Home consortium," [Online]. Available: http://www.3dathome.org

[2] A. Vetro, S. Yea, and A. Smolic, "Towards a 3D video format for autostereoscopic displays," in *Proc. SPIE: Appl. Digital Image Process. XXXI*, San Diego, CA, 2008.

[3] A. Gotchev, A. Smolic, S. Jumisko-Pyykk, D. Strohmeier, G.-B. Akar, P. Merkle, and N. Daskalov, "Mobile 3D television: Development of core technological elements and user-centered evaluation methods toward an optimized system," in *Electro. Imaging Symp.*, San Jose, California, Jan. 2009.

[4] C. Theobalt, G. Ziegler, M. Magnor, and H.-P. Seidel, "Model-based free-viewpoint video acquisition, rendering and encoding," in *Proc. Picture Coding Symp.*, San Francisco, Dec. 2004, pp. 1–6.

[5] S. Wrmlin, E. Lamboray, and M. Gross, "3D video fragments: Dynamic point samples for real-time free-viewpoint video," *Comput. Graphics*, vol. 28, no. 1, pp. 3–14, 2004.

[6] Y. K. Park, K. Jung, Y. Oh, S. Lee, J. K. Kim, H. Lee, K. Yun, N. Hur, and J. Kim, "Depth-image-based rendering for 3DTV service over T-DMB," *Signal Process.: Image Commun.*, vol. 24, pp. 122–136, 2009.

[7] P. Kauff, N. Atzpadin, C. Fehn, K. Müller, O. Schreer, A. Smolic, and R. Tanger, "Depth map creation and image-based rendering for advanced 3DTV services providing interoperability and scalability," *Signal Process.: Image Commun.*, vol. 22, no. 2, pp. 217–234, 2007.

[8] L. Zhang and W. James Tam, "Stereoscopic image generation based on depth images for 3DTV," *IEEE Trans. Broadcast.*, vol. 51, no. 2, pp. 191–199, 2005.

[9] L. Zhang, "Fast stereo matching algorithm for intermediate view reconstruction of stereoscopic television images," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 10, pp. 1259–1270, Oct. 2006.

[10] D. Kim, D. Min, and K. Sohn, "A stereoscopic video generation method using stereoscopic display characterization and motion analysis," *IEEE Trans. Broadcast.*, vol. 54, no. 2, pp. 188–197, Jun. 2008.

[11] "Introduction to 3D video," Archamps, France, ISO/IEC JTC1/SC29/WG11, Doc.N9784, May 2008.

[12] P. Merkle, A. Smolic, K. Müller, and T. Wiegand, "Efficient prediction structures for multiview video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 11, pp. 1461–1473, Nov. 2007.

[13] L. Q. Shen, Z. Liu, S. X. Liu, Z. Y. Zhang, and P. An, "Selective disparity estimation and variable size motion estimation based on motion homogeneity for multi-view coding," *IEEE Trans. Broadcast.*, vol. 55, no. 4, pp. 761–766, Dec. 2009.

[14] P. Merkle, A. Smolic, K. Müller, and T. Wiegand, "Multi-view video plus depth representation and coding," in *Proc. IEEE Int. Conf. Image Process.*, San Antonio, Texas, Sep. 2007, pp. 201–204.

[15] "MPEG 2 Test Model 5," Rev. 2 ISO/IEC JTC1/SC29WG11, Apr. 1993.

[16] J. Ribas-Corbera and S. Lei, "Rate control for low-delay video communications," ITU Study Group 16, VCEG, Portland, Doc. Q15-A-20, 1997.

[17] Z. G. Li, F. Pan, K. P. Lim, G. Feng, X. Lin, and S. Rahardja, "Adaptive basic unit layer rate control for JVT," Pattaya II, Thailand, Doc. JVT-G012r1, 2003.

[18] S. Ma, F. Wu, and Z. Li, "Proposed draft of adaptive rate control," Geneva, Switzerland, Doc. JVT-H017r3, 2003.

[19] J. E. Lim, J. Kim, K.-N. Ngan, and K. Sohn, "Advanced rate control technologies for 3D-HDTV," *IEEE Trans. Consum. Electron.*, vol. 49, no. 4, pp. 1498–1507, Nov. 2003.

[20] S. L. P. Yasakethu, W. A. C. Fernando, and A. M. Kondoz, "Rate controlling in offline 3D video coding using evolution strategy," *IEEE Trans. Consum. Electron.*, vol. 55, no. 1, pp. 150–157, Feb. 2009.

[21] B. Kamolrat, W. A. C. Fernando, and M. Mrak, "Rate controlling for color and depth 3D video coding," in *Proc. SPIE: Appl. Digital Image Process. XXXI*, San Diego, CA, 2008.

[22] S. Ma, W. Gao, and Y. Lu, "Rate-distortion analysis for H.264/AVC video coding and its application to rate control," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 12, pp. 1533–1544, Dec. 2005.

[23] H.-J. Lee, T. Chiang, and Y.-Q. Zhang, "Scalable rate control for MPEG-4 video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 10, no. 6, pp. 878–894, Sep. 2000.

[24] A. Leontaris and A. M. Tourapis, "Rate control reorganization in the joint model (JM) reference software," San Jose, CA, Doc.JVT-W042, 2007.

[25] A. Vetro, P. Pandit, H. Kimata, A. Smolic, and Y.-K. Wang, "Joint draft 8.0 on multiview video coding Hannover, DE, Doc.JVT-AB204, 2008.

[26] C. L. Zitnick, S. B. Kang, M. Uyttendaele, S. Winder, and R. Szeliski, "High-quality video view interpolation using a layered representation," in *ACM SIGGRAPH*, Los Angeles, CA, Aug. 2004.

[27] M. Tanimoto, T. Fujii, K. Suzuki, N. Fukushima, and Y. Mori, "Reference softwares for depth estimation and view synthesis," Archamps, France, ISO/IEC JTC1/SC29/WG11 MPEG2008/M15377, Apr. 2008.

[28] Y. Su, A. Vetro, and A. Smolic, "Common test conditions for multiview video coding," Hangzhou, China, Doc. JVT-U211, 2006.

Yanwei Liu received the B.S. degree in applied geophysics from Jianghan Petroleum University, China, in 1998, the M.S. degree in computer science from China Petroleum University (Beijing) in 2004 and the Ph.D. degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences in 2010.

In 2010, he joined the Institute of Acoustics, Chinese Academy of Sciences as an assistant researcher. His research interests include digital image/video processing, multiview and 3D video coding, and wireless video communication.

Qingming Huang (A'04–M'04–SM–08) received the Ph.D. degree in computer science from the Harbin Institute of Technology, Harbin, China, in 1994.

He was a Postdoctoral Fellow in National University of Singapore from 1995 to 1996, and worked in Institute for Infocomm Research, Singapore as Member Research Staff from 1996 to 2002. Currently, he is a Professor in Graduate School of Chinese Academy of Sciences. His current research areas are image processing, video analysis, video coding, and pattern recognition. He has published over 100 scientific papers.

Siwei Ma received the B.S. degree from Shandong Normal University, Jinan, China, in 1999 and the Ph.D. degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, in 2005.

He was a Postdoctoral Fellow in Media Communications Lab, University of Southern California, from 2005 to 2007. Now he is an Associate Professor of Peking University. His research interests include image and video coding, video streaming, and transmission.

Debin Zhao received the B.S., M.S., and Ph.D. degrees in computer science from the Harbin Institute of Technology, China, in 1985, 1988, and 1998, respectively.

He was a Research Fellow in the Department of Computer Science, City University of Hong Kong, from 1989 to 1993. Currently, he is a Professor in the Department of Computer Science, Harbin Institute of Technology. His research interests include image and video coding, video processing, video streaming, and pattern recognition. He has published 2 books and over 200 scientific papers.

Wen Gao (S'87–M'88–SM'05–F'09) received the Ph.D. degree in electronics engineering from the University of Tokyo, Japan, in 1991.

He is a Professor of Computer Science at Peking University, China. Before joining Peking University, he was a Professor of Computer Science at Harbin Institute of Technology from 1991 to 1995, and a Professor at the Institute of Computing Technology of Chinese Academy of Sciences. He has published extensively including four books and over 600 technical articles in refereed journals and conference proceedings in the areas of image processing, video coding and communication, pattern recognition, multimedia information retrieval, multimodal interface, and bioinformatics. His research interests include pattern recognition, artificial intelligence, image understanding, data compression, hand gesture recognition, multimodal interface, and computer vision. He has published 4 books and over 600 scientific papers.

Dr. Gao served or serves on the editorial board for several journals, such as IEEE Trans. Circuits and Systems for Video Technology, IEEE Trans. Multimedia, IEEE Trans. Autonomous Mental Development, EURASIP Journal of Image Communications, Journal of Visual Communication and Image Representation. He chaired a number of prestigious international conferences on multimedia and video signal processing, such as IEEE ICME and ACM Multimedia, and also served on the advisory and technical committees of numerous professional organizations.

Song Ci (S'99–M'02–SM'06) is an Assistant Professor of computer and electronics engineering at the University of Nebraska-Lincoln. He is the Director of the Intelligent Ubiquitous Computing Lab (iUbiComp Lab) and holds a courtesy appointment of UNL Ph.D. in the Biomedical Engineering Program. He is also affiliated with Nebraska Biomechanics Core Facility at the University of Nebraska at Omaha and Center for Advanced Surgical Technology (CAST) at University of Nebraska Medical Center, Omaha, NE.

His research interests include dynamic complex system modeling and optimization, green computing and power management, content-aware quality-driven cross-layer optimized multimedia over wireless, cognitive network management and service-oriented architecture, and cyber-enable e-health care.

Hui Tang received the B.S. degree from Lanzhou University in 1992, the M.S. degree from the Institute of Computing Technology of the Chinese Academy of Sciences in 1995, and the Ph.D. degree from the Institute of Acoustics of the Chinese Academy of Sciences in 1998.

Since 2004, he has become the founding director of the High Performance Network Laboratory of the Institute of Acoustics of the Chinese Academy of Sciences, and his team has undertaken several key national projects. Since 2008, he has served on the executive committee of the national key project "The Next Generation Broadband Wireless Mobile Network." His research interests include next generation Internet, wireless multimedia technologies, Internet of things, mobile Internet and p2p technologies.