Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2010, Article ID 848713, 24 pages
doi:10.1155/2010/848713

Research Article

Stereoscopic Visual Attention-Based Regional Bit Allocation Optimization for Multiview Video Coding

Yun Zhang,[1,2] Gangyi Jiang,[1] Mei Yu,[1] Ken Chen,[1] and Qionghai Dai[3]

[1] *Faculty of Information Science and Engineering, Ningbo University, Ningbo 315211, China*
[2] *Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China*
[3] *Broadband Networks & Digital Media Lab, Tsinghua University, Beijing 100084, China*

Correspondence should be addressed to Gangyi Jiang, jianggangyi@126.com

Received 26 December 2009; Revised 2 May 2010; Accepted 18 June 2010

Academic Editor: Dimitrios Tzovaras

We propose a Stereoscopic Visual Attention- (SVA-) based regional bit allocation optimization for Multiview Video Coding (MVC) by the exploiting visual redundancies from human perceptions. We propose a novel SVA model, where multiple perceptual stimuli including depth, motion, intensity, color, and orientation contrast are utilized, to simulate the visual attention mechanisms of human visual system with stereoscopic perception. Then, a semantic region-of-interest (ROI) is extracted based on the saliency maps of SVA. Both objective and subjective evaluations of extracted ROIs indicated that the proposed SVA model based on ROI extraction scheme outperforms the schemes only using spatial or/and temporal visual attention clues. Finally, by using the extracted SVA-based ROIs, a regional bit allocation optimization scheme is presented to allocate more bits on SVA-based ROIs for high image quality and fewer bits on background regions for efficient compression purpose. Experimental results on MVC show that the proposed regional bit allocation algorithm can achieve over 20 ~ 30% bit-rate saving while maintaining the subjective image quality. Meanwhile, the image quality of ROIs is improved by 0.46 ~ 0.61 dB at the cost of insensitive image quality degradation of the background image.

1. Introduction

Three-Dimensional Video (3DV) provides Three-Dimensional (3D) depth impression and allows users to freely choose a view of a visual scene [1]. With these features, it would allow many multimedia applications, such as photorealistic rendering of 3D scenes, free viewpoint television [2], 3D television broadcasting, and 3D games, to introduce new and exciting features for users. Multiview video plus depth [3] supports high image quality and low complexity of rendering a continuum of output views. It has been the main representation of 3D scene and applied to many multiview multimedia applications. However, multiview video requires huge amount of storage and transmission bandwidth which are multiples of traditional monoview video. Thus, it is necessary to develop efficient Multiview Video Coding (MVC) algorithms for practical uses.

MVC had been studied on the basis of several video coding standards, including MPEG-2, MPEG-4, H.263, and H.264. Since the Moving Picture Experts Group (MPEG) had recognized the importance of MVC technologies, an Ad Hoc Group (AHG) on 3D Audio and Visual (3DAV) was established. The MPEG surveyed some MVC schemes, such as "Group-of-GOP prediction (GoGOP)", "sequential view prediction", and "checkerboard decomposition", [4]. Yea and Vetro proposed a view synthesis prediction-based MVC scheme for improving interview compression efficiency [5]. Yun et al. developed an efficient MVC algorithm which adaptively selects optimal prediction structure according to the spatiotemporal correlation of 3DV sequence [6]. Merkle et al. also proposed another MVC scheme using Hierarchical B Pictures (MVC-HBPs) and achieved superior compression efficiency and temporal scalability [7]. It has been adopted into MVC standardization draft by Joint Video Team (JVT) and used in the Joint Multiview Video Model (JMVM).

In many of the previous MVC schemes [4–7], intra, inter, and interview prediction compensation technologies are adopted to reduce spatial, temporal, and interview

redundancies. Additionally, YUV color space transform, integer transform, and quantization technologies are also utilized to explore visual redundancies including chroma redundancies and high frequency redundancies. According to the studies on visual psychology, the Human Visual System (HVS) in fact does not treat visual information equally from regions to regions of the video content [8]. It is mentioned that HVS is more sensitive to the distortion in the Region-Of-Interests (ROIs) or attention areas than those in background regions [9]. Those are visual redundancies coming from regional interests existing in 3DV. However, previous MVC schemes have not taken the regional selective property and 3D depth perception of HVS into consideration. Applying the concept of ROI to video coding is regarded as a promising way to improve coding efficiency by exploiting regional visual redundancies. However, there are two major problems to be tackled, they are ROI detection and the ROI-based bits allocation.

For unsupervised ROI extraction, visual attention has been introduced as one of the key technologies in video/image system [10, 11]. Accordingly, many efforts have been devoted to researches on visual attention model [11–16] so as to simulate the visual attention mechanism of HVS accurately. Itti and Koch developed a bottom-up visual attention model [12] for still images based on Treisman's stimulus integration theory [13]. It generates saliency map with the integration of perceptual stimuli from intensity contrast, colour contrast, and orientation contrast. Zhai et al. used the low-level features as well as cognitive features, such as skin colour and captions, in their visual attention model [14]. Motion is another important cue for visual attention detection in video, thus, a bottom-up spatiotemporal visual attention model is proposed for video sequences in [15]. Wang et al. proposed segment-based video attention detection method [16]. Ma et al. also proposed a bottom-up and top-down combined visual attention model by integrating multiple features, including contrast in image, motion, face detection, audition, and text [11]. However, all these visual attention models were proposed either for static image or single view video and did not take stereoscopic or depth perception into account. On the other hand, stereoscopic parallax is not available in the single-view video.

From the video coding point of view, many bit allocation algorithms [17–24] are proposed for improving compression efficiency. Kaminsky et al. proposed a complexity-rate-distortion model to dynamically allocate bits with both complexity and distortion constraints [17]. Lu et al. proposed a Group-Of-Picture (GOP-)level bit allocation [18] scheme and Shen et al. proposed another frame-level bit allocation method which decreases the average standard deviation of video quality [19]. Özbek and Tekalp proposed a bit allocation among views for scalable multiview video coding [20]. All these bit allocation schemes improve the average Peak Signal-to-Noise Ratio (PSNR) but did not take the regional selective properties of HVS into account. Chen and Wang et al. proposed a bit allocation scheme that allocated more bits on ROI for MPEG-4 standard [21, 22]. These two schemes require very high ROI extraction accuracy. Chi et al. proposed an ROI video coding based on H.263+ for low bit-rate multimedia communications [23]. In the scheme, the ROI was extracted according to skin-color clue and a fuzzy logic controller was designed adaptively to adjust the quantization parameters for each macroblock (MB). Tang et al. proposed a bit allocation scheme for 2D video coding which is guided by visual sensitivity considering motion and texture structures [24]. However, these bit allocation schemes were proposed for single-view video coding and can not be directly applied to MVC because interview prediction is adopted in MVC.

In this paper, we propose a Stereoscopic Visual Attention-(SVA-) based regional bit allocation for improving MVC coding efficiency. We firstly present a framework of MVC in Section 2. In Section 3, we propose an SVA model to simulate visual attention mechanism of HVS. And then, SVA-based bit allocation optimization algorithm is proposed for MVC in Section 4. Section 5 presents the regional selective image quality metrics which are adopted in the coding performance evaluation. In Section 6, SVA-based ROI extraction and multiview video coding experiments are performed and evaluated with various multiview video test sequences. Finally, Section 7 gives conclusions.

2. Framework of Multiview Video System Using Regional Bit Allocation Optimization

Figure 1 shows a framework of MVC with regional bit allocation optimization. Firstly, N channels synchronized color videos are captured by parallel or arc arranged video capture system. Then, N synchronized depth videos, the same resolution as color image, are captured by depth camera array or generated by depth creation algorithms. By using depth video and multiview texture video, the SVA-based ROI extraction module efficiently extracts the semantic ROI mask for MVC codec. With the automatically extracted ROIs, MVC encoder is optimized for bit-rate saving in background region and better quality in ROI using regional bit allocation optimization. Finally, the compressed color and depth video bitstream are multiplexed and transmitted/stored. In the framework, the MB-wise ROI mask may not necessary to be transmitted to the client. Moreover, the framework is compatible with current block-based video coding standard and rate control, and low-level, such as macroblock level, syntax modification is not needed.

At the client side, the color and depth bitstream is demultiplexed and decoded by the MVC decoder. With the decoded multiview color videos, depth videos as well as the transferred video cameras' parameters, view generation module renders a continuum of output views, $N'(N' > N)$, through depth image-based rendering [25]. According to different types of display device, for example, HDTV, stereoscopic display, or multiview display, different number of views is displayed.

3. Stereoscopic Visual Attention-Based ROI Extraction

3.1. Framework of SVA Model. Three-dimensional video provides the most effective stereoscopic perception obtained

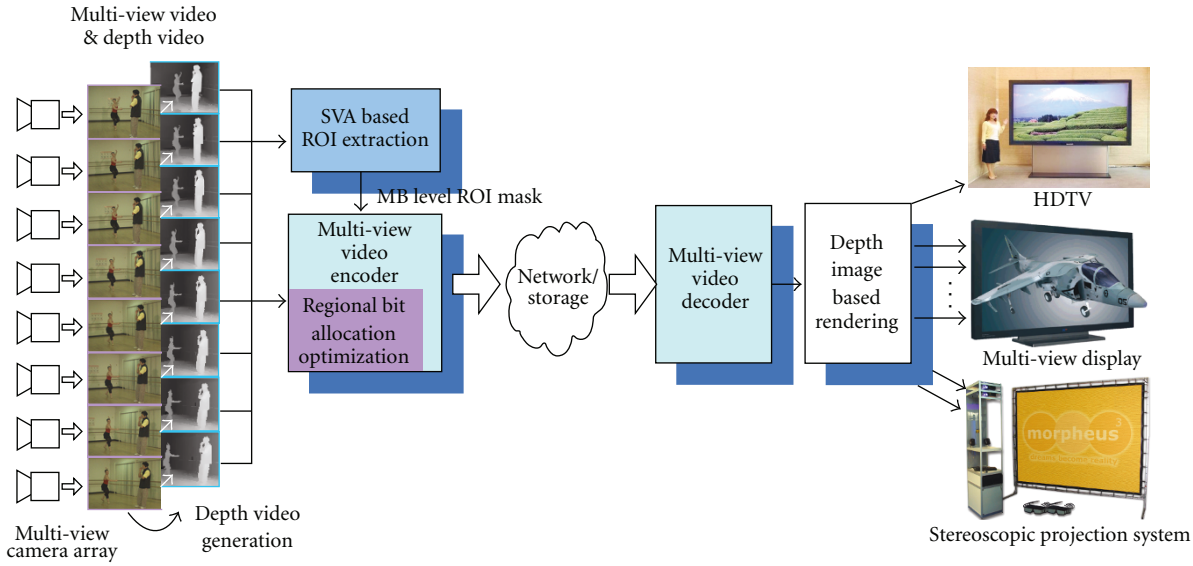


Figure 1: Framework of multiview video system using regional bit allocation optimization.

by viewing a scene from slightly different viewing positions. The depth perception makes the scene more vivid, and it is another important factor that affects human visual attention just like what motion and texture contrasts do in traditional two-dimensional (2D) video. For example, people are often interested in the regions popping out from video screen and the interesting ratio of attention regions decreases as they are getting far away. In our previous work, we presented an SVA model [26] in which depth map was directly adopted as depth visual saliency. In this work, the SVA model is further improved in the following two aspects. Firstly, depth saliency is detected via a depth attention algorithm instead of using the depth map directly and a new fusion algorithm is presented. Secondly, in Section 6.2, a subjective evaluation is performed to testify the effectiveness of the improved SVA model. Each SVA object is modeled by combining the four attributes with low-level features, including depth, depth saliency, image saliency, and motion saliency. The SVA model is defined as

$$S_{\mathrm{SVA}} = \{D, S_D, S_M, S_I\}, \tag{1}$$

where S_{SVA} is SVA saliency map, D is the intensity of depth maps which indicates the distance between video content and imaging camera/viewer, S_I, S_M, and S_D are image saliency, motion saliency and depth saliency, respectively.

Figure 2 presented the architecture of our proposed SVA-based ROI extraction. Image and motion saliency are detected from the multiview color video. Depth saliency is also detected from the multiview depth video. Afterward, a novel dynamic model fusion method is used to integrate the obtained pixelwise image saliency map, motion saliency, and depth saliency. The proposed SVA model does not incorporate any top-down, volitional component because it relies on the cognitive knowledge and differs from person to person. Finally, the MB level ROIs are extracted by threshold and block operation.

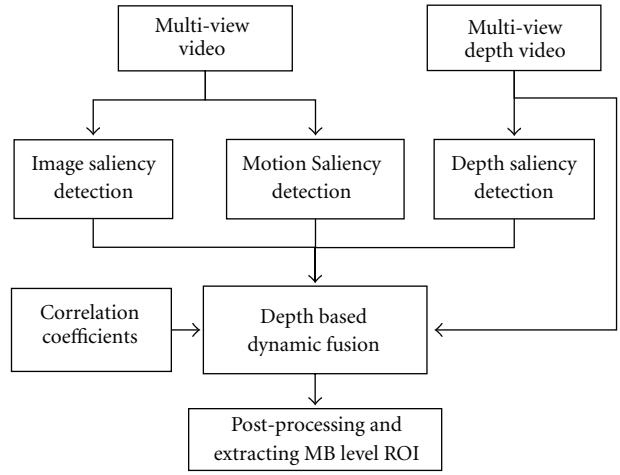


Figure 2: Flowchart of the proposed SVA-based ROI extraction.

3.2. Spatial Attention Detection for Static Image. We adopted Itti's bottom-up attention model [12, 27] for our spatial visual attention model. The seven neuronal features implemented are sensitive to color contrast (red/green and blue/yellow), intensity contrast, and four orientations (0°, 45°, 90°, and 135°) for static images. Centre and surround scales are obtained using dyadic Gaussian pyramids with nine levels. Then, Centre-Surround Differences (CSD) [27] are computed as the pointwise differences across pyramid levels; and then, six feature maps for CSD network are computed for each of the seven features, yielding a total of 42 feature maps. Finally, all feature maps are integrated into the unique scalar image saliency S_I.

3.3. Temporal Attention Detection. Motion is one of the major stimuli on visual attention of dynamic scene. In this work, we adopt an optical flow algorithm based on block

matching method in [28] to estimate the motion of image objects between consecutive frames. Frame group $\mathbf{F}(v,t) = \{f_{v,t} \mid t = t+k, -w \le k \le w, k \in Z\}$, which consists of $2w+1$ temporal consecutive frames in view v, is employed to extract robust motion magnitude, where w is temporal window size. The horizontal and vertical motion channels of frame $f_{v,t}$ are determined by frame group $\mathbf{F}(v,t)$, then they are combined together as

$$M_{v,t}^{k} = \Theta_{m,n}\left[\left|P_{m,n}^{h}(f_{v,t}, f_{v,t+k})\right| + \left|P_{m,n}^{v}(f_{v,t}, f_{v,t+k})\right|\right], \tag{2}$$

where operators $P_{m,n}^{h}$ and $P_{m,n}^{v}$ denote the horizontal and the vertical optical flow operator with $m \times n$ block size, respectively. "$|\cdot|$" is the magnitude of motion velocity. Operator $\Theta_{m,n}$ performs upsampling operation of Gaussian pyramid decomposition with $m \times n$ times. Therefore, $M_{v,t}^{k}$ is with the same resolution as $f_{v,t}$. In this paper, a 4×4 ($m = n = 4$) block size is adopted to compute the optical flow because we found that it has a robust performance experimentally. Forward and backward motion is intersected so as to eliminate the background exposure phenomena, which refer to the background regions in current frame but attributed as motion regions by $M_{v,t}^{k}$ or $M_{v,t}^{-k}$, that is,

$$M_{v,t}(k) = \begin{cases} \dfrac{\left(M_{v,t}^{k} + M_{v,t}^{-k}\right)}{2}, & \text{if} \min\left(M_{v,t}^{k}, M_{v,t}^{-k}\right) > T_1, \\ 0, & \text{else}, \end{cases} \tag{3}$$

where T_1 is a trade-off between the sensitivity and error resilience of the motion detection, and it is set as 0 for sensitivity in this paper. Then, to reduce the error effects caused by noises, such as camera shaking and jitter of video sequences, several $M_{v,t}(k)$ are weighted combined to form a final motion map, M as follow:

$$M = \sum_{k=1}^{w} \varsigma_k \cdot M_{v,t}(k), \tag{4}$$

where ς_k are weighted coefficients satisfying $\sum_{k=1}^{w} \varsigma_k = 1$. In this paper, w is set to 3 since it has a good trade-off between complexity and error resilience, ς_k is $1/w$. Usually, motion attention level increases with relative motion. So motion saliency map is generated by using CSD network [27] and represented as

$$S_m = \mathcal{N}\left(\bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} \mathcal{N}\left|M(c) \ominus M(s)\right|\right), \tag{5}$$

where $\ominus$ denotes the across-level difference between two maps at the center (c) and the surround (s) levels of the respective feature pyramids, $s, c \in [0,8]$, $s = c + \delta$, $\delta = \{-3,-2,-1,1,2,3\}$; $\oplus$ is across-level addition; $\mathcal{N}(\cdot)$ is a normalization operator. There are also several normalization strategies available in [29], such as learning, iterative localized iteration. However, these normalization strategies are supervised or very time consuming. Therefore, we adopted the "Naive" strategy in [29] for its low complexity and unsupervised purpose, the normalization operator is $\mathcal{N}(i_{x,y}) = 255(i_{x,y} - \min_{(x,y)\in\mathbf{I}}(i_{x,y}))/(\max_{(x,y)\in\mathbf{I}}(i_{x,y}) - \min_{(x,y)\in\mathbf{I}}(i_{x,y}))$, which adjusts the saliency value to fixed rang 0~255 (value 255 indicates being most salient) for image $\mathbf{I}$.

3.4. Depth Impacts on SVA and Depth Attention Detection. The stereoscopic perception can also be represented by the 2D video and the depth map which indicates the relative distance between video object and the camera system. Hence, we use a depth map to analyze the differences between 3D video and traditional 2D video. Compared with traditional 2D video, the depth's effect on human SVA is listed as the following four aspects.

(1) When watching 3D video, people are usually more interested in the regions visually moving out of the screen, that is, pop-out regions, which are with small depth values or large disparities.

(2) As the distance between video object and viewer/camera increases, interesting ratio of the video object decreases.

(3) The out of Depth-Of-Field (DOF) objects of the camera system is usually not the attention areas, for example, defocusing blurred background object or foreground object.

(4) Depth discontinuous regions or depth contrast regions are usually the attention areas in the 3DV as they provide strong depth sensation, especially when view angles or view positions are switching.

Depth map is an 8-bit gray image that can be captured by depth camera or computed using multiview video. Each pixel in the depth map represents a relative distance between video object and camera. In this paper, we firstly estimate the disparity for each pixel in multiview video by using stereo matching method. Then, the disparity is converted into perceptive depth. Finally, intensity of each pixel in depth map is mapped to an irregular space with nonuniform quantization [30]. HVS perceptive depth, Z, is shown as

$$Z = \frac{B \cdot f}{d_c}, \tag{6}$$

where f is the focal length of the cameras, B is the baseline between the neighboring cameras, d_c is the physical disparity (measured by centimeter) between the corresponding points of the neighboring views. However, disparity estimated using stereo matching is measured by pixel. So we use a centimeter-to-pixel ratio, λ, that is, a ratio of CCD size to image resolution, to convert physical disparity to pixel disparity

$$d_p = d_c/\lambda. \tag{7}$$

Because close object is usually more important than far away object, the depth value Z which corresponds to the pixel (x, y) is transformed into the 8-bit intensity $d(x, y)$ with non-uniform quantization [30]

$$d(x,y) = \left\lfloor 255 \cdot \frac{z^n}{Z(x,y)} \cdot \frac{z^f - Z(x,y)}{z^f - z^n} + 0.5 \right\rfloor, \tag{8}$$

where "$\lfloor \cdot \rfloor$" is floor operation, z^f and z^n indicate the farthest and nearest depth, respectively, and $z^f = Bf/\lambda \min\{d_p\}$, $z^n = Bf/\lambda \max\{d_p\}$, f is the focal length, and B is the baseline between cameras. The space between z^f and z^n is divided into narrow spaces around the z^n plane and is divided into wide spaces around the z^f plane.

It is observed that the depth contrast and the depth orientation contrast are usually attention-catching regions. Thus, we obtain the depth orientation information, $O(\sigma, \theta)$, from depth intensity maps $\mathbf{D}$ using oriented Gabor filters, where $\sigma \in [0 \cdots 8]$ represents the scale of different pyramids level and $\theta \in \{0, \pi/4, \pi/2, 3\pi/4\}$. The orientation feature maps of the depth video $\mathbf{F}_O$ are obtained from absolute CSD network [27] between the depth orientation-selective channels

$$\mathbf{F}_O = \frac{1}{4} \sum_{\theta \in \{0,\pi/4,\pi/2,3\pi/4\}} \mathcal{N}\left(\bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} \mathcal{N}(|O(c,\theta) \ominus O(s,\theta)|) \right) \tag{9}$$

Additionally, intensity feature maps of the depth video $\mathbf{F}_D$ are obtained from absolute CSD network between the depth intensity channels

$$\mathbf{F}_D = \mathcal{N}\left(\bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} \mathcal{N}(|\mathbf{D}(c) \ominus \mathbf{D}(s)|) \right), \tag{10}$$

where $\mathbf{D}$ indicates the depth intensity map. Finally, the orientation feature map and the intensity feature map are normalized and combined to form a depth saliency map as

$$\mathbf{S}_D = \frac{1}{2}(\mathcal{N}(\mathbf{F}_O) + \mathcal{N}(\mathbf{F}_D)) \otimes \mathbf{G}, \tag{11}$$

where $\mathbf{G}$ is a boundary depress matrix. The symbol $\otimes$ is scalar multiplication that indicates that each element of $(\mathcal{N}(\mathbf{F}_O) + \mathcal{N}(\mathbf{F}_D))$ is multiplied by the scaling factor in the same position in matrix $\mathbf{G}$. In the current stereoscopic display, the regions near by image boundary almost can not provide or just provide a little depth perception. Also, people pay more attention to the center location [31]. Thus, the depth saliency of the image boundary is depressed using a boundary depress matrix $\mathbf{G}$. Each element at position (x, y) in $\mathbf{G}$ is $g(x, y) = 1 - (1/L) \sum_{i=1}^{L} \alpha_i(x, y)$, where L is the number of levels for image boundary depression,

$$\alpha_i(x, y) = \begin{cases} 1, & x \in (iw_x, W - iw_x), y \in \left(iw_y, H - iw_y\right), \\ 0, & \text{others}, \end{cases} \tag{12}$$

w_x and w_y are width and height for each boundary depression level, W and H are width and height of the stereoscopic video, respectively.

3.5. Depth-Based Fusion for SVA Model. Psychological studies reveal that HVS is more sensitive to motion contrast when compared to color, intensity, and orientation contrast in single-view video. If a strong motion contrast is presented in the sequence, temporal attention is dominant over the spatial attention. However, if the motion contrast is low in the sequence, the spatial attention is more dominant. In the 3DV, the depth sensation is provided and depth is another key factor for visual attention in stereoscopic video. Thus, depth, spatial and temporal information of 3DV are jointly combined to construct SVA saliency as

$$s_{\text{SVA}}(x, y) = \mathcal{N}\left(Q(d(x, y)) \left(\sum_{a \in \{D,M,I\}} K_a s_a(x, y) - \sum_{a,b \in \{D,M,I\}, a \neq b} C_{ab} \Theta_{ab}(x, y) \right) \right), \tag{13}$$

where K_D, K_M, and K_I are weighted coefficients for depth saliency, motion saliency, and image saliency, respectively, and they satisfy $\sum_{a \in \{D,M,I\}} K_a = 1$, $0 \leq K_D, K_M, K_I \leq 1$. Relative larger weighted coefficient value shall be given to more dominative saliency. $\Theta_{ab}(x, y)$ denotes correlation between saliency a and saliency b, $\Theta_{ab}(x, y) = \min(s_a(x, y), s_b(x, y))$, C_{ab} is a weighted coefficient for $\Theta_{ab}(x, y)$, $0 \leq C_{ab} < 1$, and $Q()$ is a scaling function for depth intensity video. If the depth video is not provided, then (13) will be considered as a spatiotemporal scheme which fuses motion and still image saliency.

Based on the SVA saliency map, MB $\mathbf{B}_{u,v}$ is labeled as ROI when average $s_{\text{SVA}}(x, y)$ energy of an MB is lager than average

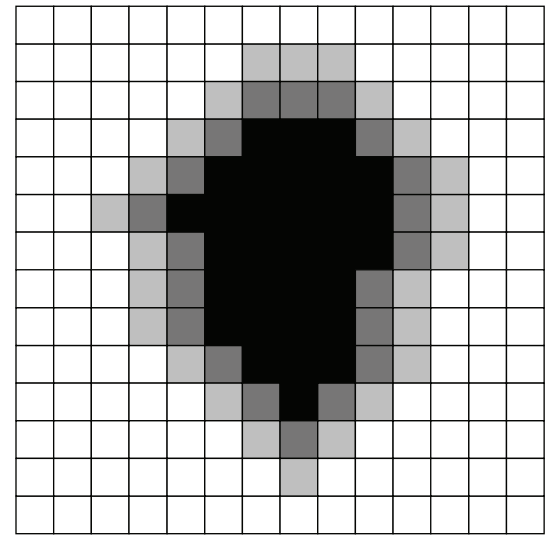


Figure 3: Sample of attention mask.

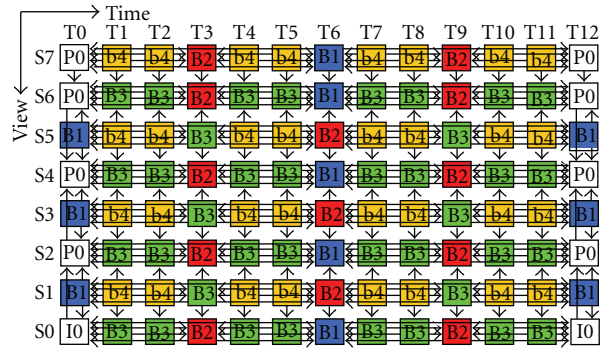


Figure 4: MVC-HBP prediction structure.

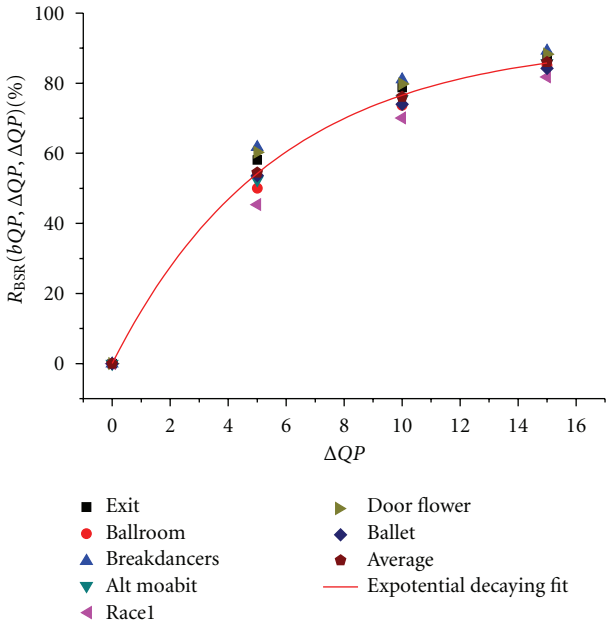


Figure 5: The relationship between $R_{BSR}(bQP, \Delta QP, \Delta QP)$ and ΔQP.

$s_{SVA}(x, y)$ energy of an image weighted by T_2, that is, $T_2(H_b \times W_b / H \times W) \sum_{y=1}^{H} \sum_{x=1}^{W} s_{SVA}(x, y)$; Otherwise, $\mathbf{B}_{u,v}$ is labeled as background, that is,

$$\mathbf{B}_{u,v} = \begin{cases} \text{ROI}, & \sum_{y=1}^{H_b} \sum_{x=1}^{W_b} s_{SVA}(x, y) \\ & \geq T_2 \frac{H_b \times W_b}{H \times W} \sum_{y=1}^{H} \sum_{x=1}^{W} s_{SVA}(x, y), \\ \text{Background}, & \text{else}, \end{cases} \tag{14}$$

where H_b and W_b are height and width of $\mathbf{B}_{u,v}$, W and H are width and height of the video, respectively. As threshold T_2 increases, $\mathbf{B}_{u,v}$ with lower SVA saliency will be determined as ROIs, and vice versa. In this paper, T_2 is set as 1.10. To transit image quality from ROI to background regions smoothly in MVC, two MB wide transitional regions between ROI and background MB are defined. A sample of ROI mask is shown in Figure 3. The black rectangles are ROI MBs, white rectangles are background MBs, and gray rectangles are transitional MB with different levels.

4. SVA-Based Regional Bit Allocation Optimization for MVC

The MVC-HBP prediction structure [7], shown in Figure 4, is interview and temporal prediction hybrid. The even views are coded using motion prediction compensation, while the odd views are coded utilizing both interview prediction and temporal prediction. Since the MVC-HBP prediction structure is superior on both compression efficiency and temporal scalability, it is adopted by JVT and used in reference software JMVM. This superior coding performance is mainly owing to its novel quantization strategy. Given the basis Quantization Parameter (QP) of MVC-HBP prediction structure, bQP, the remaining QPs are determined as

$$\text{QP}^l = \begin{cases} \text{bQP} + 3, & \text{if } l = 1, \\ \text{QP}^{l-1} + 1, & \text{if } l > 1, \end{cases} \tag{15}$$

where l is hierarchical level of hierarchical B frame. In the proposed SVA-based MVC scheme, larger QPs are set for background regions and transitional regions for higher compression ratio. The QP of SVA-based ROI in level l is set as $\text{QP}^l_{\text{SVA}} = \text{QP}^l$. QPs of the background and the transitional regions in the lth hierarchical level picture, QP^l_{BG} and $\text{QP}^l_{T_i}$, are defined as

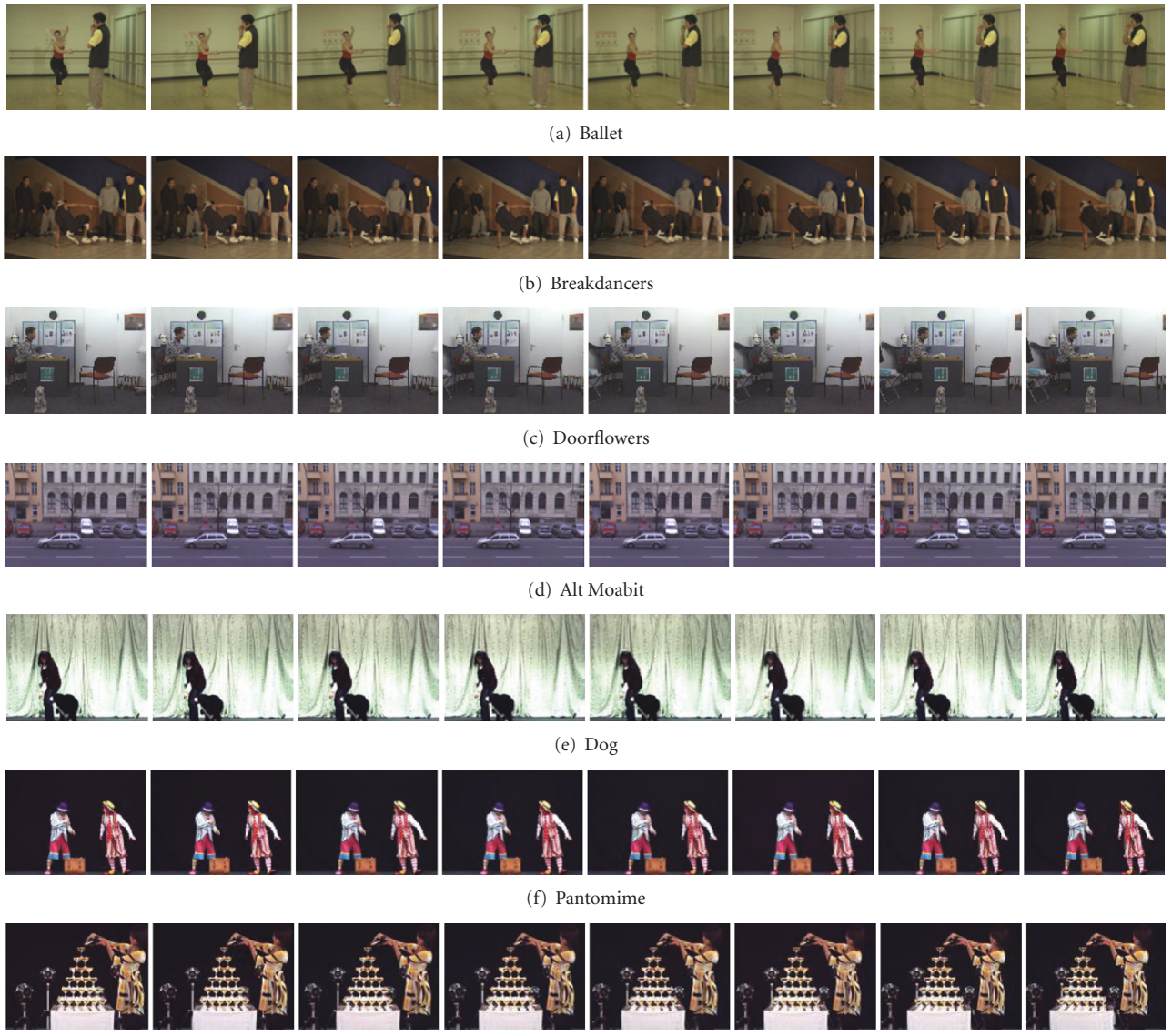
$$\begin{aligned} \text{QP}^l_{\text{BG}} &= \text{QP}^l_{\text{SVA}} + \Delta\text{QP}, \\ \text{QP}^l_{T_i} &= \text{QP}^l_{\text{SVA}} + \lfloor \Delta\text{QP}/\eta_i \rfloor, \end{aligned} \tag{16}$$

where "$\lfloor \cdot \rfloor$" is floor operation, η_i is a positive division parameter, and ΔQP is a QP difference between background region and ROI region and it indicates the relative amount of bits allocated between ROI and the background regions.

To exploit regional selective visual redundancies in 3DV, the SVA-based MVC scheme is used to maximize compression ratio while at the cost of imperceptible image quality loss in background. Therefore, we need to determine the optimal ΔQP. The bit allocation optimization scheme in [32] is adopted to determine bit allocation between SVA-based ROIs and background regions. Here, a short review on the bit allocation scheme is presented for better readability. Two indices, the average bit-rate saving ratio, R_{BSR}, and the image quality degradation, ΔD, are adopted to evaluate the coding performance of MVC scheme with different ΔQP. The bit-rate saving ratio, $R_{\text{BSR}}(b\text{QP}, \Delta\text{QP}_{\text{ROI}}, \Delta\text{QP}_{\text{BG}})$, is calculated as

$$\begin{aligned} &R_{\text{BSR}}(b\text{QP}, \Delta\text{QP}_{\text{SVA}}, \Delta\text{QP}_{\text{BG}}) \\ &= \frac{1}{V_{\text{GOP}} \times T_{\text{GOP}}} \sum_{j=1}^{V_{\text{GOP}}} \sum_{i=1}^{T_{\text{GOP}}} \frac{EB^{i,j}\left(\text{QP}^l_{\text{SVA}}, \text{QP}^l_{\text{SVA}}\right) - EB^{i,j}(\mathcal{A})}{EB^{i,j}\left(\text{QP}^l_{\text{SVA}}, \text{QP}^l_{\text{SVA}}\right)}, \end{aligned} \tag{17}$$

where $\mathcal{A} = \text{QP}^l_{\text{SVA}} + \Delta\text{QP}_{\text{SVA}}, \text{QP}^l_{\text{SVA}} + \Delta\text{QP}_{\text{BG}}$ and where V_{GOP} and T_{GOP} are the numbers of views and time instants



(a) Ballet

(b) Breakdancers

(c) Doorflowers

(d) Alt Moabit

(e) Dog

(f) Pantomime

(g) Champagne tower

FIGURE 6: Eight views of multiview video test sequences.

in one GOP, i and j are temporal and interview position, respectively. $EB^{i,j}(\mathrm{QP}_1, \mathrm{QP}_2)$ denotes the number of bits of encoding a frame at position (i, j) while its ROIs are coded with QP_1 and background regions are coded with QP_2 · $\Delta\mathrm{QP}_{\mathrm{SVA}}$ and $\Delta\mathrm{QP}_{\mathrm{BG}}$ denote the QP differences between the ROI and the background regions, respectively.

Figure 5 shows the relationship between $R_{\mathrm{BSR}}(\mathrm{bQP}, \Delta\mathrm{QP}, \Delta\mathrm{QP})$ and ΔQP in that one QP is used for both ROI and background regions. $R_{\mathrm{BSR}}(\mathrm{bQP}, \Delta\mathrm{QP}, \Delta\mathrm{QP})$ is subjected to the exponential decaying function as ΔQP increases. Thus, $R_{\mathrm{BSR}}(\mathrm{bQP}, \Delta\mathrm{QP}, \Delta\mathrm{QP})$ can be predicted as

$$R_{\mathrm{BSR}}(\mathrm{bQP}, \Delta\mathrm{QP}, \Delta\mathrm{QP}) = A_0 e^{-\Delta\mathrm{QP}/T} + y_0, \tag{18}$$

where A_0 and T are the coefficients of functions bQP and independent to the content of multiview video. y_0 is the maximum bit-rate saving ratio. Because ROI and background regions are mutual exclusive, we can obtain

$$\begin{aligned} R_{\mathrm{BSR}}(\mathrm{bQP}, \Delta\mathrm{QP}, \Delta\mathrm{QP}) = {} & R_{\mathrm{BSR}}(\mathrm{bQP}, 0, \Delta\mathrm{QP}) \\ & + R_{\mathrm{BSR}}(\mathrm{bQP}, \Delta\mathrm{QP}, 0). \end{aligned} \tag{19}$$

Once, ROI and background regions are segmented for 3DV sequence, the bit-rate saving ratio of ROI is approximately in direct proportion to that of background region while ΔQP increases. It is represented by

$$\rho = \frac{R_{\mathrm{BSR}}(\mathrm{bQP}, 0, \Delta\mathrm{QP})}{R_{\mathrm{BSR}}(\mathrm{bQP}, \Delta\mathrm{QP}, 0)}, \tag{20}$$

where ρ is independent of ΔQP. Hence, substituting (19) and (20) into (18), we can obtain

$$R_{\mathrm{BSR}}(\mathrm{bQP}, 0, \Delta\mathrm{QP}) = Ae^{-\Delta\mathrm{QP}/T} + y, \tag{21}$$

where $A = 1/(1+\rho)A_0$ and $y = 1/(1+\rho)y_0$. $|A|$ indicates amplitude of bit-rate saving. Parameter T indicates the period that R_{BSR} reaches the point of no more gain can be saved as ΔQP increases.

On the other hand, image quality degradation caused by allocating fewer bits on background regions $\Delta D(\mathrm{bQP}, \Delta\mathrm{QP}_{\mathrm{ROI}}, \Delta\mathrm{QP}_{\mathrm{BG}})$ is calculated as

$$\Delta D(\mathrm{bQP}, \Delta\mathrm{QP}_{\mathrm{SVA}}, \Delta\mathrm{QP}_{\mathrm{BG}}) = \frac{1}{V_{\mathrm{GOP}} \times T_{\mathrm{GOP}}} \sum_{j=1}^{V_{\mathrm{GOP}}} \sum_{i=1}^{T_{\mathrm{GOP}}} \left[Q^{i,j}(\mathcal{A}) - Q^{i,j}\left(\mathrm{QP}^l_{\mathrm{SVA}}, \mathrm{QP}^l_{\mathrm{SVA}}\right) \right], \tag{22}$$

where $\mathcal{A} = \mathrm{QP}^l_{\mathrm{SVA}} + \Delta\mathrm{QP}_{\mathrm{SVA}}, \mathrm{QP}^l_{\mathrm{SVA}} + \Delta\mathrm{QP}_{\mathrm{BG}}$ and where $Q^{i,j}(\mathrm{QP}_1, \mathrm{QP}_2)$ denotes the reconstructed image quality of a frame at position (i, j), while ROIs are coded with QP_1, and background regions are coded with QP_2. $\Delta\mathrm{QP}_{\mathrm{SVA}}$, and $\Delta\mathrm{QP}_{\mathrm{BG}}$ denote QP changes in ROI and background regions, respectively. Because the relationship between distortion, such as PSNR, and quantization factor in H.264 is approximately linear [33], we can define the image quality degradation of bit allocation, $\Delta D(\mathrm{bQP}, 0, \Delta\mathrm{QP})$, as

$$\Delta D(\mathrm{bQP}, 0, \Delta\mathrm{QP}) = b_1 \cdot \Delta\mathrm{QP} + a_1, \tag{23}$$

where a_1 is coefficient independent to ΔQP, and b_1 is a negative value which indicates the slope of image quality degradation. $\Delta D(\mathrm{bQP}, 0, \Delta\mathrm{QP})$ is a negative value and it will decrease as ΔQP increases to improve compression ratio.

To achieve a high compression ratio and also to maintain high image quality with bit allocation optimization, we ought to find the optimal ΔQP to maximize bit-rate saving ratio R_{BSR} subject to a unnoticeable image quality degradation, T_D. It is mathematically expressed as

$$\begin{aligned} &\arg\max\{R_{\mathrm{BSR}}(\mathrm{bQP}, 0, \Delta\mathrm{QP})\} \\ &\text{s.t.} |\Delta D(\mathrm{bQP}, 0, \Delta\mathrm{QP})| < T_D. \end{aligned} \tag{24}$$

Instead of solving the constrained problem in (24), an unconstrained formulation is employed. The optimal ΔQP^* is determined as

$$\Delta\mathrm{QP}^* = \arg\max_{\Delta\mathrm{QP} \in Z^+} \{R_{\mathrm{BSR}}(\mathrm{bQP}, 0, \Delta\mathrm{QP}) + \mu \cdot \Delta D(\mathrm{bQP}, 0, \Delta\mathrm{QP})\}, \tag{25}$$

where μ is a scaling constant putting R_{BSR} and ΔD in a same scale. We set the partial derivative of function $R_{\mathrm{BSR}}(\mathrm{bQP}, 0, \Delta\mathrm{QP}) + \mu\Delta D(\mathrm{bQP}, 0, \Delta\mathrm{QP})$ of ΔQP equal to 0, that is,

$$\frac{\partial\left(R_{\mathrm{BSR}}(\mathrm{bQP}, 0, \Delta\mathrm{QP}) + \mu \cdot \Delta D(\mathrm{bQP}, 0, \Delta\mathrm{QP})\right)}{\partial\Delta\mathrm{QP}} = 0. \tag{26}$$

By solving the (26), the optimal integer $\Delta\mathrm{QP}^*$ is obtained as

$$\Delta\mathrm{QP}^* = \left\lfloor T \ln \frac{A}{\mu \cdot T \cdot b_1} + 0.5 \right\rfloor. \tag{27}$$

where symbol "$\lfloor \cdot \rfloor$" is floor operation. Meanwhile, $\Delta\mathrm{QP}^*$ is truncated to 0 if $\Delta\mathrm{QP}^*$ is smaller than 0. Coefficients A, T, and b_1 are bQP dependent and will be modeled experimentally from MVC experiments presented in Section 6.3.

5. ROI-Based Objective Image Quality Assessment Metric

Pixelwise image quality assessment metric, such as PSNR, has been widely used for video quality evaluation. However, it does not match well with the human visual perception. Engelke et al. proposed a region-selective objective image quality metric [34] which is able to be combined with normalized hybrid image quality metric, reduced-reference image quality assessment technique, Structural SIMilarity (SSIM) [35], or PSNR measures. Since both SSIM and PSNR have been adopted in advanced video coding standard, H.264/AVC, we apply both the region-selective SSIM and PSNR metrics [34] to evaluate the proposed MVC scheme. The SSIM index [34] between two images is computed as

$$\mathrm{SSIM} = \frac{(2\mu_R\mu_D + C_1)(2\sigma_{RD} + C_2)}{(\mu_R^2 + \mu_D^2 + C_1)(\sigma_R^2 + \sigma_D^2 + C_2)}, \tag{28}$$

where R and D are two nonnegative image signals to be compared, μ_R and μ_D are the means of images R and D, σ_R and σ_D are standard deviation of images R and D, respectively, and σ_{RD} is covariance of images R and D, C_1 and C_2 are constants. The PSNR of illumination component (PSNR_Y) measures the fidelity difference of two image signals $I_R(x,y)$ and $I_D(x,y)$ on a pixel-by-pixel basis as

$$\mathrm{PSNR\_Y} = 10\log\frac{\Gamma^2}{1/(W\times H)\sum_{x=1}^{W}\sum_{y=1}^{H}\left[I_R(x,y) - I_D(x,y)\right]^2}, \tag{29}$$

where Γ is the maximum pixel value, here it is 255.

The objective image quality metrics have been used to independently assess the image quality of ROI and background region to enable region-selective quality metric design. An ROI quality metric Φ_{ROI} is calculated on ROI of reference and distorted images. Similarly, background regions of reference and distorted images are used to assess quality of the background region by computing Φ_{BG}. In a pooling stage, Φ_{ROI} and Φ_{BG} are combined with a region-selective metric, and the final Predictive Mean Opinion Score (PMOS) is computed as follows [34]:

$$\begin{aligned} \Phi(\omega, \kappa, \nu) &= \left[\omega \cdot \Phi_{\mathrm{ROI}}^{\kappa} + (1-\omega)\Phi_{\mathrm{BG}}^{\kappa}\right]^{1/\nu}, \\ \mathrm{PMOS}_{\Phi(\omega,\kappa,\nu)} &= a \cdot e^{b\cdot\Phi(\omega,\kappa,\nu)}, \end{aligned} \tag{30}$$

where $\omega \in [0,1]$, $\kappa, \nu \in Z^+$, $\Phi \in \{\mathrm{SSIM}, \mathrm{PSNR\_Y}\}\omega,\kappa,\nu,a$, and b are derived from the subjective quality evaluation

Table 1: Parameters and Features of the Test Multiview Videos.

3DV	Provider	Image size	Frame rate	Baseline/camera array	Feature	Depth
Ballet	MSR	1024 × 768	15 fps	20 cm/1D arc	Slow and fast motion	A
Breakdancers		1024 × 768	15 fps	20 cm/1D arc	Slow and very fast motion	A
Doorflowers	HHI	1024 × 768	16.7 fps	6.5 cm /1D	Complex indoor scene, slow motion	N/A
Alt Moabit		1024 × 768	16.7 fps	6.5 cm /1D	Outdoor scene and complex background	N/A
Dog	Nagoya Univ.	1280 × 960	29.4 fps	5 cm/1D	Large image size, slow motion	N/A
Pantomime		1280 × 960	29.4 fps	5 cm/1D	Simple background	N/A
Champagne tower		1280 × 960	29.4 fps	5 cm/1D	Simple background and very slow motion	N/A

experiments in [34]. In the following sections, PMOSs of PSNR_Y and SSIM are denoted by PMOS_PSNR and PMOS_SSIM, respectively.

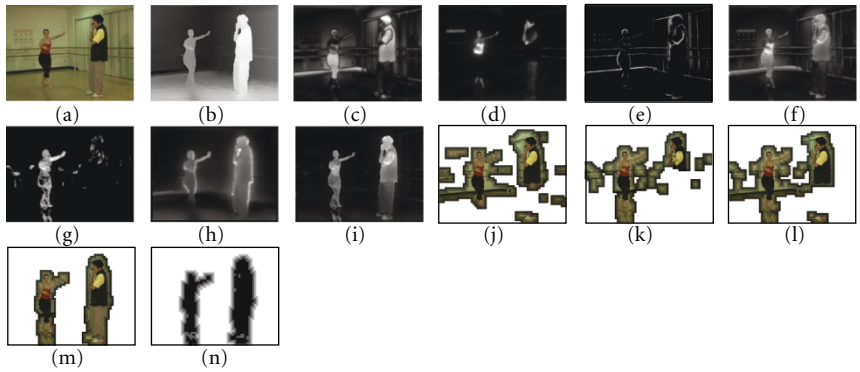
6. Experimental Results and Analyses

In this section, the performance of SVA-based ROI extraction algorithms and SVA-based MVC are evaluated. Experiments include three steps. First, SVA-based ROI extraction experiments are performed and evaluated with subjective experiments. Secondly, regional bit allocation optimization experiments are performed for allocating reasonable mounts of bits among ROI and background regions and optimal ΔQP is determined. Finally, MVC experiments are implemented to verify the efficiency of the SVA-based bit allocation optimization. In these experiments, we adopt seven typical multiview video sequences provided by Heinrich Hertz Institute (HHI) [36], Microsoft Research (MSR) [37], and Nagoya University [38]. These 3DV sequences are with different textures, motion properties, resolutions, capturing frame rates, and camera arrangements. Eight views of the test sequences are illustrated in Figure 6. Table 1 shows the properties of the test multiview video sequences. Depth maps of Breakdancers and Ballet test sequences, marked as "A" in last column, are available. The depth maps of the rest videos, marked as "N/A", are generated by Depth Estimation Reference Software (DERS) [39].

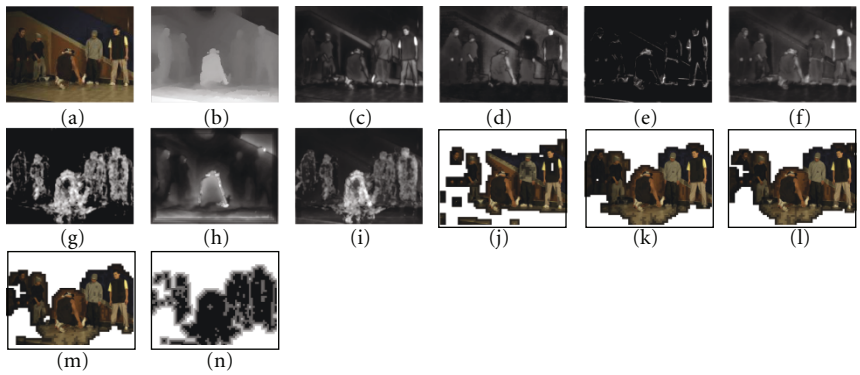
6.1. SVA-Based ROI Extraction. In the 3DV, motion saliency object is usually the most salient regions in the visual attentive area; next is the image saliency. Depth saliency is relatively less important and is given smaller weighted coefficient while comparing with motion saliency and image saliency except that the 3DV provides strong depth perception. So in the experiments, relative larger weighted coefficient value is given to dominative or more important motion saliency, and K_D, K_I, and K_M are empirically set as 0.2, 0.35, and 0.45 under the constraints $K_D < K_I < K_M$ and $K_D + K_I + K_M = 1$. On the other hand, in the Multiview video, image, motion, and depth saliencies are correlated with each other. The correlation between image and motion saliencies is higher than the other two correlations, that is, correlations between depth and image saliency, depth and motion saliency. It is because detected moving objects are likely textural objects. However, there are no explicit correlations between depth and image/motion saliency. Thus, the weighted coefficients C_{IM} are larger than C_{DM},C_{DI}, and they empirically are set as 0.6, 0.2, and 0.2, respectively. Actually, in order to accurately simulate the mechanism of human visual attention, values of parameters K_D, K_I, and K_M, and C_{IM}, C_{DM}, and, C_{DI} should be adjusted according to motion, textual, and depth characteristics of the multiview video sequences.

In the depth video, the z^f and z^n planes are mapped to 0 and 255, respectively, with the non-uniform quantization process in (8), which treats z^f plane as infinite far away and supposes that saliency in z^f plane is completely unimportant. However, z^f planes of the video sequences are usually not infinite. So, we use the scaling function $Q(x) = x + \gamma$, where γ is a positive constant, to map the $z^n \sim z^f$ plane to $\gamma \sim \gamma + 255$ and take the saliency in z^f plane into account. Usually, γ shows the importance of the saliency in z^f plane compared with that of z^n plane. It increases as z^f plane closes to z^n plane and decreases to 0 as z^f becomes infinite. In the SVA extraction experiments, γ is set to 50 because most of the test video sequences are indoor scene and their z^f planes close to z^n plane.

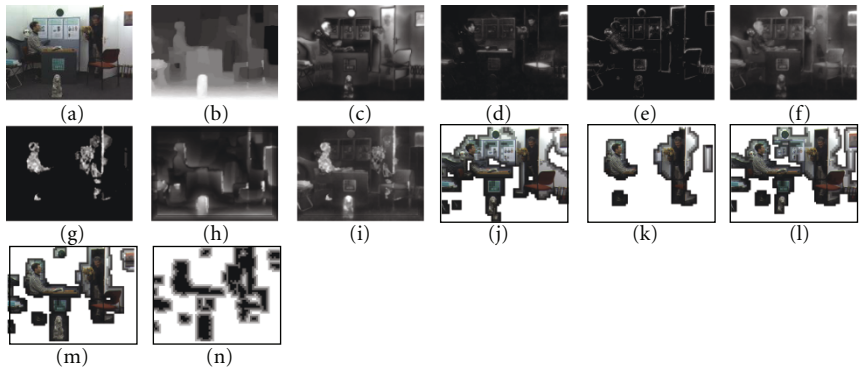
Figure 7 shows the SVA-based ROI extraction results for different multiview test sequences. Figure 7(a) renders one view of original multiview video. Figure 7(b) shows one view of multiview depth video in which large depth comes with small intensity and small depth with large intensity. Figures 7(c), 7(d), and 7(e) show feature maps of intensity, color, and orientation, respectively. In these feature maps and saliency maps followed, white pixel indicates a high saliency pixel and black pixel indicates a low saliency pixel in the multiview video. Figure 7(f) exhibits static image saliency combining feature maps of intensity, color, and orientation. Figure 7(j) shows the extracted ROI based on static image saliency only. The spatial attention model can simulate the visual attention mechanisms well for some sequences with simple background, such as Champagne tower and Pantomime. However, for the sequences with complex background, such as Ballet and Dog, the spatial attention model is not accurate enough. Other information, such as motion and depth, shall



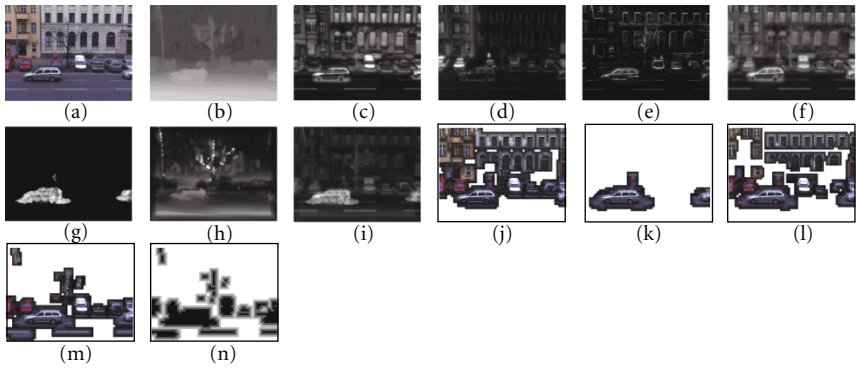
(a) Ballet



(b) Breakdancers



(c) Doorflowers



(d) Alt Moabit

Figure 7: Continued.

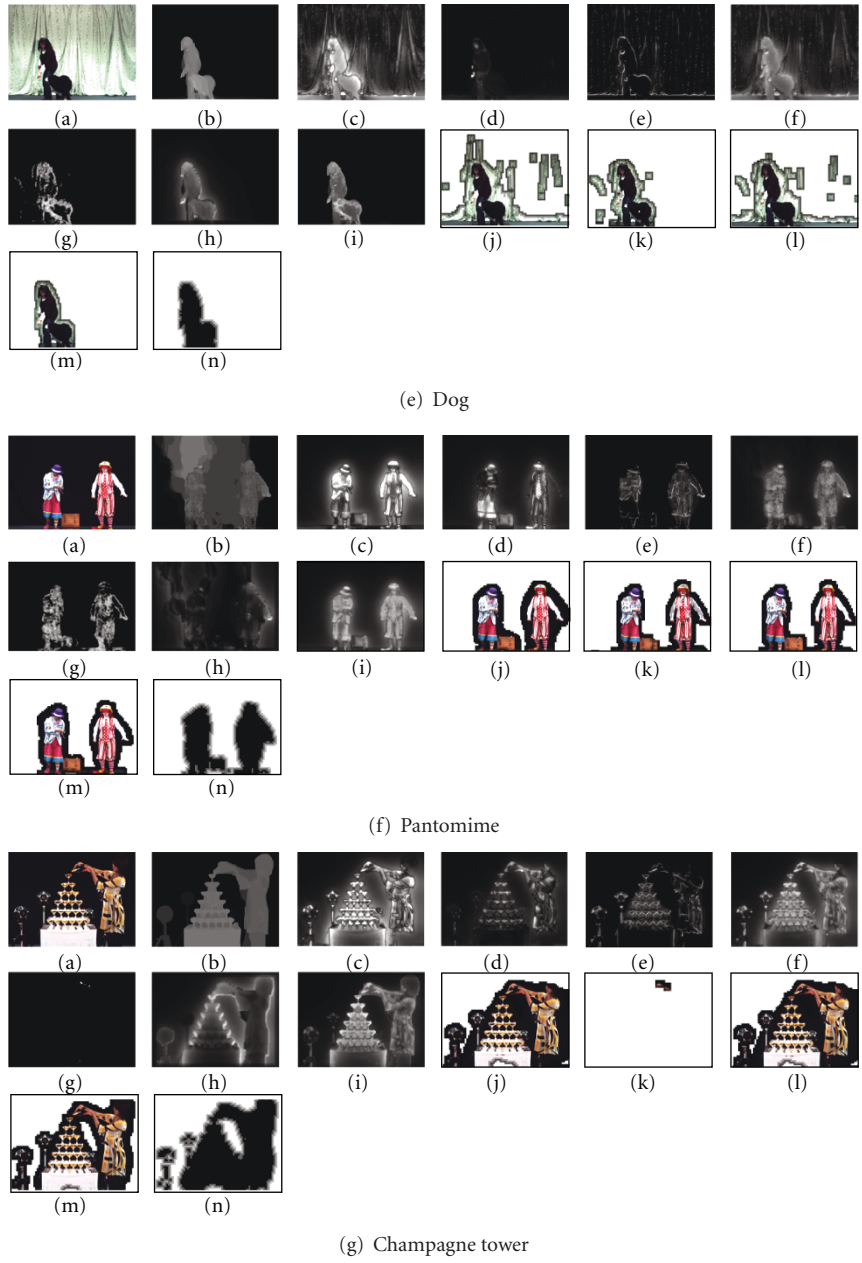


Figure 7: SVA-based ROI extraction results. (a) One view of original multiview video; (b) One view of multiview depth video; (c) Feature maps of intensity; (d) Feature maps of color; (e) Feature maps of orientation; (f) Static image saliency map; (g) Motion saliency map; (h) Depth saliency map; (i) Final SVA saliency map (proposed); (j) Extracted ROI using static image saliency (S-scheme); (k) Extracted ROI using motion saliency (T-scheme); (l) Extracted ROI using spatiotemporal saliency (ST-scheme); (m) Extracted ROI based on SVA model (proposed); (n) MB-level ROI mask (proposed).

be utilized to improve visual attention model for dynamic stereoscopic visual scenes.

Figure 7(g) illustrates motion saliency maps and Figure 7(k) shows the ROI extracted on the basis of motion saliency only. Generally, large motion contrast areas are very likely to be potential attention areas. However, it is not always true. For example, for Ballet sequence, the shadow of the dancing girl exhibits high motion contrast, but it is not an attentive area. This kind of noise can be eliminated by combining the depth saliency and static image saliency.

Figure 7(h) shows the depth saliency extracted from depth video by using the proposed algorithm in Section 3.3. As we can see from the depth saliency map, the depth contrast regions are extracted as the most salient, which is coinciding with the discovery that people are in particularly interested in depth contrast regions because it provides more impressive stereoscopic perception. Besides, regions with small depth, that is, large intensity in depth map, are also extracted as salient region, which is also in accordance with the fact that people are usually more interested in an

Table 2: z-scores, mean opinion score and standard errors for ROI extraction schemes.

		Champagne tower	Dog	Door-flowers	Break-dancers	Alt Moabit	Pantomime	Ballet	Average
z-scores	S	−0.700	−0.446	−0.669	−0.597	−0.354	−0.467	−0.703	−0.541
	T	−1.786	−0.159	−0.661	0.167	−0.485	−0.718	−0.775	−0.574
	ST	−0.700	−0.286	−0.538	**0.401**	−0.066	**0.104**	−0.409	−0.193
	SVA	**0**	**0**	**0**	0	**0**	0	**0**	**0**
MOS	S	2.45	2.85	2.80	3.35	2.70	2.80	2.85	2.83
	T	3.70	2.40	2.80	2.25	2.90	3.15	2.95	2.88
	ST	2.35	2.6	2.6	**1.9**	2.25	**1.95**	2.40	2.29
	SVA	**1.50**	**2.15**	**1.80**	2.50	**2.15**	2.10	**1.80**	**2.00**
Std. errors for MOS	S	0.74	1.31	1.03	1.15	1.31	1.25	1.11	1.13
	T	0.90	0.66	1.36	0.99	1.30	1.06	0.92	1.03
	ST	0.79	0.80	0.86	0.77	0.70	0.74	0.73	0.77
	SVA	0.74	1.39	0.81	0.97	0.85	0.89	1.25	0.99

object close to them in a view than that far away from them. According to the extracted depth saliency of various test sequences, the proposed depth saliency detection algorithm is efficient and maintains high accuracy as the depth map is accurate. However, for inaccurate depth and the sequences with weak depth perception, only depth saliency turns out to be not sufficient to simulate visual attention. Such cases can be noted in Pantomime and Breakdancers.

Figure 7(i) shows the final SVA saliency map generated by the proposed SVA model. We can see that Figure 7(i) can simulate visual attention mechanism of HVS better for all sequences when compared with Figures 7(f)–7(h). Taking Ballet sequence as an example, the proposed SVA model can depress the noise in spatial saliency map (black region on the wall in color image), noise in motion saliency map (shadow of the dancing girl), and noise in depth (the foreground floor). Favorable saliency map and ROI are created. For Doorflowers sequence, multiple attention cues including motion (two men and the door), static image attention (clock, painting, and chair), and depth (the sculpture) are integrated together very well by the proposed model. Similar results can be found for other multiview video sequences. Therefore, it can be concluded that the proposed model detects the SVA accurately and simulates HVS well by fusing depth information, static image saliency, and motion saliency. Additionally, though there are noises in both the depth map and/or the image saliency, the proposed model still can obtain satisfactory SVA jointly using depth, motion, and texture information and depress noises in each channel. Thus, the proposed model is error resilient and with high robustness.

The ROI extraction results, as illustrated in Figures 7(j)–7(m), are generated by four schemes, that is, S-scheme, T-scheme, ST-scheme, and proposed SVA scheme. S-scheme denotes ROI extraction only using static image information. T-scheme denotes that ROI is extracted only using motion information. ST-scheme indicates ROI extraction using both static image information and motion information. SVA denotes ROI is extracted based on our proposed SVA model. Figure 7(m) shows the extracted MB level ROI based on SVA and Figure 7(n) is MB level ROI mask in which Black blocks are ROI MBs, gray blocks are transitional MBs, and white blocks are background MBs. Comparing Figures 7(j), 7(k), and 7(l) with Figure 7(m), we can see that extracted ROIs based on SVA model are similar to this ROI extraction based on static image saliency (S-scheme) for simple textural multiview video, such as Pantomime and Champagne tower. However, for complex textural multiview video, such as Dog, Ballet, Alt Moabit, and Doorflowers, the ROIs extracted based on the proposed SVA model are much better and more favorable than S-scheme, T-scheme, and ST-scheme because they lack of information from depth or motion channel.

6.2. Subjective Evaluation for SVA-Based ROI Extraction. Subjective evaluation of SVA-based ROIs extraction results has also been performed. Polarization-multiplexed display method is used for displaying stereo video and image. Stereoscopic images are played back through a stereoscopic dual projection system, where two BenQ P8265 DLP projectors are used to project left and right view images on a 150-inche silver screen. Viewers wear polarized glasses to watch the stereo video. Extracted ROI results are randomly ordered and displayed on a traditional monoview 21″ LCD display at the time when stereoscopic video is being displayed via the stereoscopic video system. The experiment is conducted in a special room with ambient illumination, color temperature, and ambient sound controlled according to the requirements in ITU-R Recommendation 500 [40]. There are 20 participants recruited in campus, age from 22 to 32, 7 females and 13 males, 2 participants are experts, 15 participants have some stereoscopic image processing knowledge, and the rest 3 participants do not have image processing knowledge. That is the 18 participants are nonexpert and they are not concerned with the visual attention and the ROI extraction in their normal work. All participants passed the color vision test and achieved the minimum criteria: acuity of 20 : 30 vision, stereoscopic visual acuity of 40 sec.arc.

Seven multiview video sequences illustrated in Figure 6 are adopted for ROI subjective evaluation. Sequences are displayed in the order of Champagne tower, Dog, Doorflowers, Breakdancers, Alt Moabit, Pantomime, and then Ballet. The ROI extraction results, as illustrated in Figure 7(j)–7(m),

(a) (b)

Figure 8: (a) Example of stereoscopic video (b) ROIs on monoview LCD display.

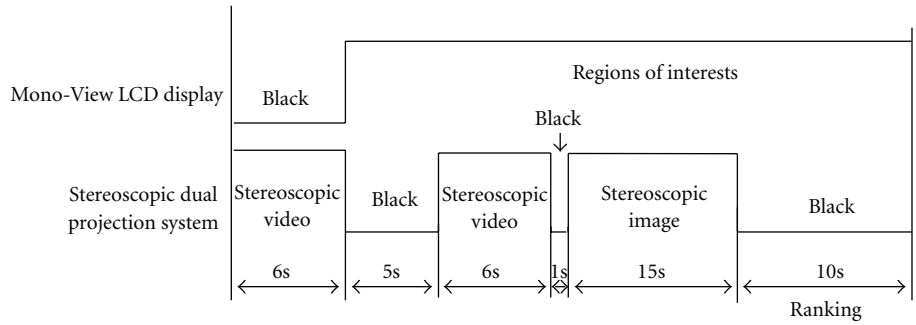


Figure 9: Displaying time interval of stereoscopic video and the extracted ROIs.

are generated by the four schemes, S-scheme, T-scheme, ST-scheme, and proposed SVA scheme, respectively. Example of stereoscopic video displaying is shown in Figure 8(a). ROIs by different schemes are randomly displayed on four areas of monoview LCD display and example of the demonstration is shown as Figure 8(b). The displaying time interval for each sequence is shown in Figure 9. Both stereoscopic video and stereoscopic image are displayed on the dual projection system in different time slot. Before the subjective experiment, participants had a try of the stereo vision system with several stereopair images from Middlebury Stereo Vision Page (http://vision.middlebury.edu/stereo/). All participants were informed of the stereo video and ROI images displaying procedure for each sequence, shown as Figure 9. And in the ranking stage after showing stereo video/images, they were asked to make a comparison on the four extracted ROIs and rank them from 1 to 4 for the ROIs shown on monoview LCD display based on their viewing experience of stereo video/images, where 1 indicates the best one (the ROI is most identical to their ROI impression) and 4 indicates the worst one (the ROI is least identical to their ROI impression).

With the ranking scores, the preference metrics of SVA scheme over other schemes are obtained. Then, Thurstone model and paired comparisons [41] have been adopted to analyze the performance of the four ROI extraction schemes. Table 2 shows the z-scores, Mean Opinion Score (MOS), and its standard errors for four ROI extraction schemes with different test 3DVs. The proposed SVA-based ROI extraction scheme is set as 0 for reference and proper identification for the z-scores. Higher z-score indicates better performance and the best performance scheme for each sequence is shown in yellow shadow.

As shown in Table 2, for the five sequences, including Champagne tower, Dog, Doorflowers, Alt Moabit, and Ballet sequences, ROIs generated by the proposed SVA-based ROI extraction scheme are of the highest z-scores which means these ROIs are most identical to people's preference. However, for Breakdancers sequence, the z-score of ST-scheme is 0.401 (better than the SVA scheme) because the sequence has dramatically high speed motion attracting more attentions. For Pantomime sequences, the proposed SVA scheme is ranked no. 2 because the sequence is with simple background and provides relatively weak stereoscopic perception. In addition, the extracted ROIs of the four schemes are quite similar and hard to be distinguished. Generally, according to the average z-scores, the proposed SVA extraction scheme achieves the best performance for the test 3DV. Then, the performance ST-scheme comes next. S-scheme and T-scheme have relatively low performance and low robust because they highly depend on the texture and motion properties of video sequences.

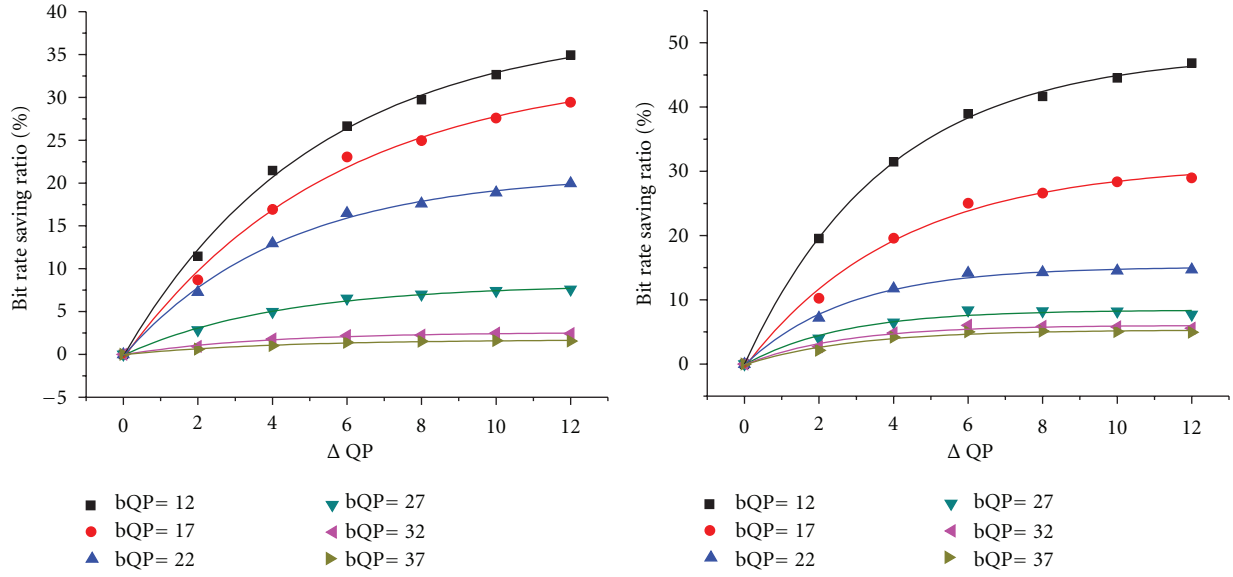
Table 3: Objective image quality and coding bits corresponding to Figure 18.

	Breakdancers			Ballet			Doorflowers		
	JMVM	Proposed	ΔΨ	JMVM	Proposed	ΔΨ	JMVM	Proposed	ΔΨ
QP_{SVA}/QP_{BG}	23/23	22/31	−/−	23/23	22/31	−/−	23/23	22/31	−/−
PSNR_Y_{SVA}	40.92 dB	41.50	**0.58** dB	40.81 dB	41.27 dB	**0.46** dB	41.42 dB	41.95 dB	**0.53** dB
PSNR_Y_{BG}	40.87 dB	39.75	−**1.12** dB	42.08 dB	40.97 dB	−**1.11** dB	41.73 dB	40.86 dB	−**0.87** dB
PSNR_Y	40.88 dB	40.16 dB	—	41.88 dB	41.01 dB	—	41.67 dB	41.05 dB	—
EB	306720 b	237744 b	**22.49**%	190104 b	150072 b	**21.06**%	218936 b	167176 b,	**23.64**%
PMOS_PSNR	82.06	81.51	−**0.55**	83.32	82.62,	−**0.70**	83.69	83.35,	−**0.34**
PMOS_SSIM	73.72	74.58	**0.86**	72.79	73.57	**0.78**	76.67	77.02	**0.35**
	Alt Moabit			Pantomime			Dog		
	JMVM	Proposed	ΔΨ	JMVM	Proposed	ΔΨ	JMVM	Proposed	ΔΨ
QP_{SVA}/QP_{BG}	23/23	22/31	−/−	22/22	22/31	−/−	23/23	22/31	−/−
PSNR_Y_{SVA}	41.19 dB	41.78 dB	**0.59** dB	43.59 dB	43.60 dB	**0.01** dB	43.08 dB	43.69 dB	**0.61** dB
PSNR_Y_{BG}	41.54 dB	39.73 dB	−**1.81** dB	46.74 dB	46.48 dB	−**0.26** dB	42.34 dB	40.65 dB	−**1.69** dB
PSNR_Y	41.50 dB	39.93 dB	—	45.64 dB	45.50 dB	—	42.50 dB	41.16 dB	—
EB	341920 b	224688 b,	**34.29**%	249640 b	229192 b	**8.19**%	376944 b	251256 b	**33.34**%
PMOS_PSNR	83.18	81.83,	−**1.35**	92.37	92.08,	−0.29	86.50	85.30,	−**1.20**
PMOS_SSIM	77.91	78.30	**0.39**	80.16	80.15,	−**0.01**	78.17	78.57,	**0.40**
	Champagne tower								
	JMVM	Proposed	ΔΨ						
QP_{SVA}/QP_{BG}	22/22	22/31	−/−						
PSNR_Y_{SVA}	43.34 dB	43.35 dB	**0.01** dB						
PSNR_Y_{BG}	46.20 dB	45.32 dB	−**0.88** dB						
PSNR_Y	44.83 dB	44.43 dB	—						
EB	447416 b	409024 b	**8.58**%						
PMOS_PSNR	91.41	90.37	−1.04						
PMOS_SSIM	79.69	79.70	**0.01**						

The middle four rows of the Table 2 show MOS of the ranking ROIs, in which smaller value indicates better performance. As far as MOS is concerned, similar results can be found. The proposed SVA-based ROI extraction scheme has the best performance as it has the lowest MOS for five test sequences and lowest average MOS. In the last four rows, standard errors for MOS are also illustrated. We can see that the deviation for SVA scheme (0.99 on average) is larger than ST-Scheme (0.77 on average). It is because the participants' depth sensations vary from person to person. While viewing the stereo video and images, some non-expert viewers seem to be more sensitive to depth perception. On the contrary, expert viewers pay more attentions on motion, textural, or semantic areas because they are already familiar with the depth sensation.

6.3. SVA-based Regional Bit Allocation Optimization for MVC. To determine the optimal ΔQP used in the MVC scheme, video coding experiments are implemented on JMVM7.0 [42] with MVC-HBP prediction structure, bQP and ΔQP are set as bQP $\in \{12, 17, 22, 27, 32, 37\}$ and ΔQP $\in \{0, 2, 4, 6, 8, 10, 12\}$. Multiview video sequences, Ballet and Breakdancers, are adopted because they have both slow and fast motion characteristic. Eight views and 91 frames in each view (6 GOPs while GOP length is 15) are encoded. Parameter η_1 and η_2 are empirically set as 3 and 6 for first and second level transitional areas.

Figure 10 shows the relation maps of $R_{BSR}(\text{bQP}, 0, \Delta\text{QP})$ to ΔQP for Ballet and Breakdancers sequences. More bit-rate can be saved as ΔQP becomes larger. However, the gradient of $R_{BSR}(\text{bQP}, 0, \Delta\text{QP})$ decreases as the ΔQP increases, The bit-rate saving ratio, $R_{BSR}(\text{bQP}, 0, \Delta\text{QP})$, obeys the exponential decaying function described in (21). Besides, the gradient and up-boundary of $R_{BSR}(\text{bQP}, 0, \Delta\text{QP})$ decreases as bQP increases. Figures 11 and 12 show the relationships between bQP and the coefficients, that is, T and A. Each point in the figures is fitted from each curve of Figure 10 adopting exponential function in (15). The red points are the coefficients fitted from Ballet sequence and the black points are fitted from Breakdancers sequence. $|A|$ indicates amplitude of bit-rate saving and decreases as bQP increases. As bQP increases, the up-boundary of bit-rate saving ratio decreases to zero and little coding gain can be expected as bQP is bigger than 35. T indicates the velocity of bit-rate saving



(a) Breakdancers (b) Ballet

Figure 10: The relation maps of bit saving ratio $R_{\text{BSR}}(b\text{QP},0,\Delta\text{QP})$ to ΔQP.

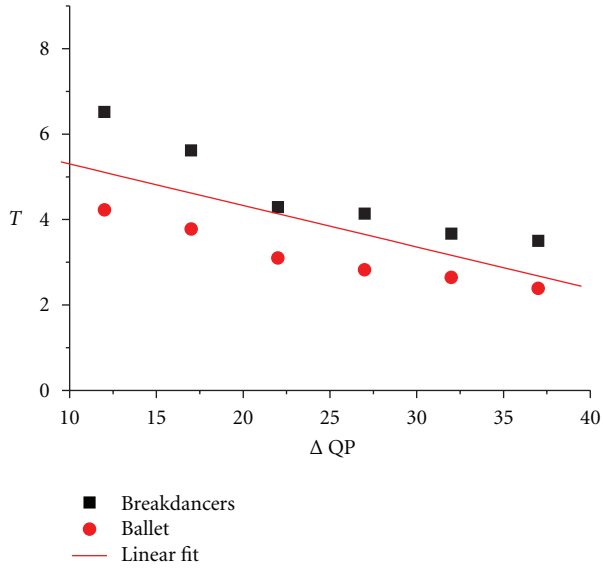


Figure 11: Relation map of bQP and coefficient T.

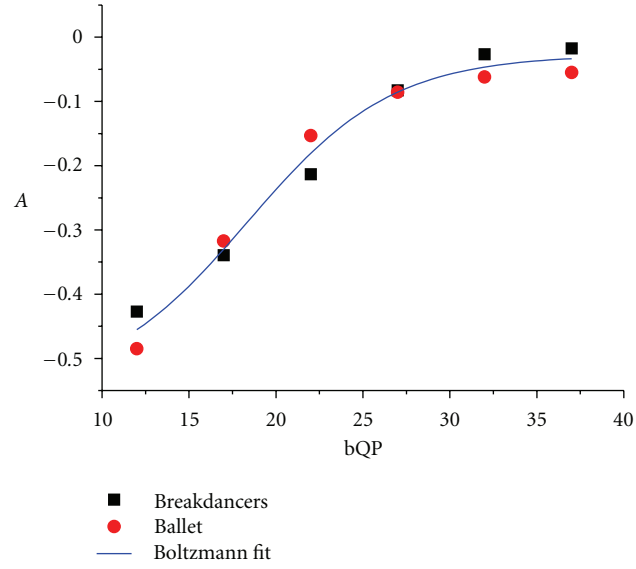


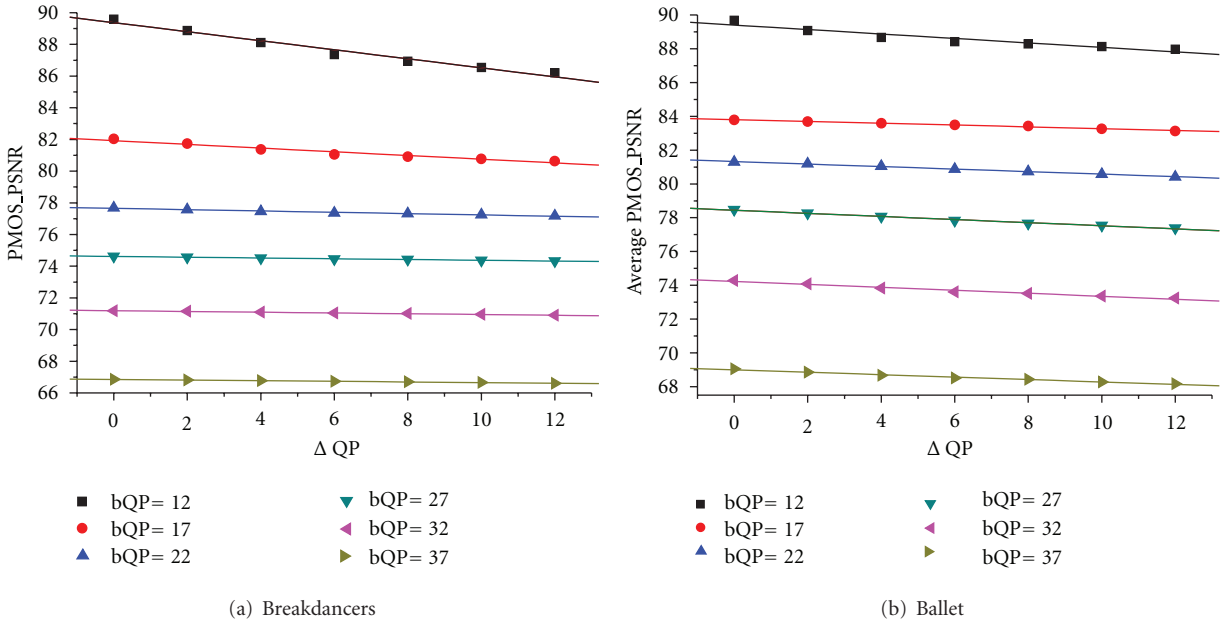
Figure 12: Relation map of bQP and coefficient A.

becoming saturated. As bQP increases, the velocity is getting faster. Then, we fit the obtained points in Figures 11 and 12 using a linear and Boltzman function. Parameters T and A are expressed as

$$T = \alpha_1 + \beta_1 \cdot \text{bQP},$$
$$A = \alpha_2 + \frac{\beta_2}{1 + e^{(\text{bQP}-r_2)/\omega_2}}, \tag{31}$$

where $\alpha_1 = 6.27$, $\beta_1 = -0.10$, $a_2 = -2.75$, $\beta_2 = -52.10$, $r_2 = 18.3$ and $\omega_2 = 4.17$.

We use the PMOS_PSNR index to evaluate the reconstructed image quality, that is, $Q^{i,j}$ in (22) is derived from (30), and Φ is PSNR_Y. The average PMOS_PSNR value to ΔQP is illustrated in Figure 13. Each line in the figure has one bQP but different ΔQP. We can see that the image quality evaluated by PMOS_PSNR linearly decreases as ΔQP increases. Besides, the slope of image quality degradation is getting flat as bQP increases. Figure 14 shows the relationship between bQP and the coefficient b_1, which indicates the slope of image quality degradation, $\Delta D(\text{bQP}, 0, \Delta\text{QP})$. Each point in the figure is the coefficient b_1 fitted from $\Delta D(\text{bQP}, 0, \Delta\text{QP})$ using linear function in (23). The red



(a) Breakdancers (b) Ballet

Figure 13: The relation maps of image quality to ΔQP.

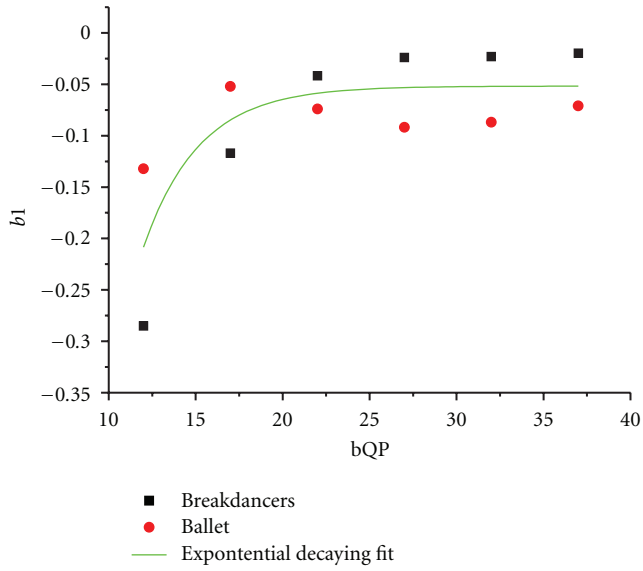


Figure 14: Relation map of bQP and b_1.

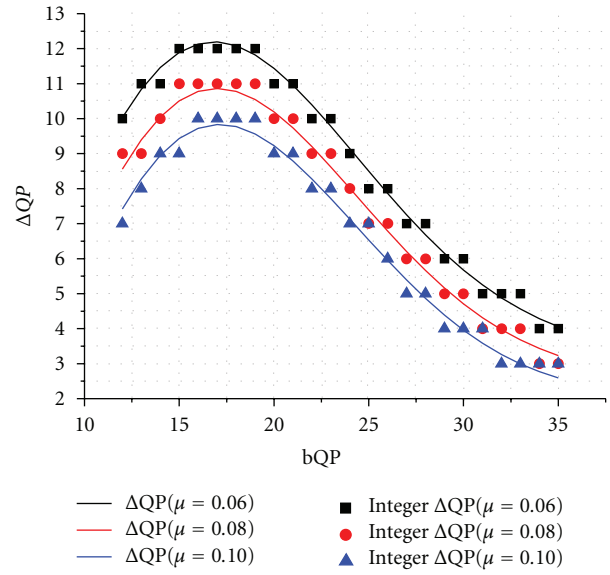


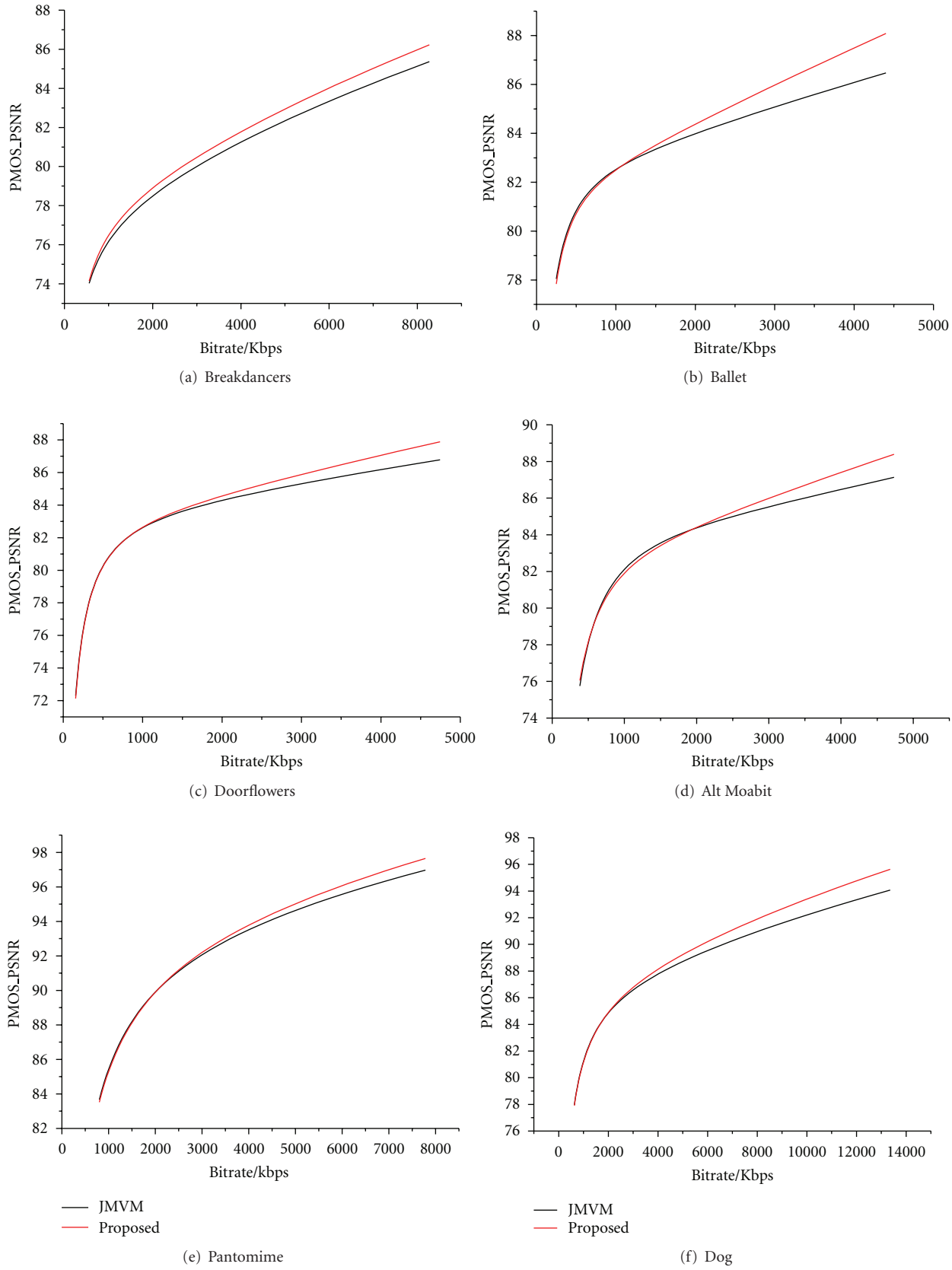
Figure 15: Optimal and integer ΔQP for different μs.

points are the coefficients, b_1, fitted from Ballet sequence and the black points are fitted from Breakdancers sequence. We fit these points in Figure 14 using exponential decaying function and obtain

$$b_1 = \alpha_3 + \beta_3 e^{-\mathrm{bQP}/\gamma_3}, \tag{32}$$

where $\alpha_3 = -0.05$, $\beta_3 = -6.57$, and $\gamma_3 = 3.21$.

Applying (31) and (32) to (27), optimal ΔQP for different μs is obtained, shown as Figure 15. The maximum point of ΔQP increases as μ decreases. However, the trends of the optimal ΔQP are similar for different scaling μs. We set μ as 0.08 to scale R_{BSR} and ΔD into a same scale according to the test sequence. Then, the final optimal ΔQP is obtained. For low bQP, for example, bQP $<$ 15, significant bit-rate saving can be achieved by selecting large ΔQP. However, the image quality is also degraded a lot. Thus, ΔQP is reasonable to be smaller than 8 so that a wise tradeoff between bit-rate saving ratio and image quality degradation can be achieved. As for large bQP, for example, bQP $>$ 33, most MBs in background regions are already coded with SKIP/DIRECT mode, in which no residuals are coded, and little coding gain can be expected by choosing large ΔQP. In some cases, the bit-rate saving ratio will not increase as ΔQP increases



(a) Breakdancers

(b) Ballet

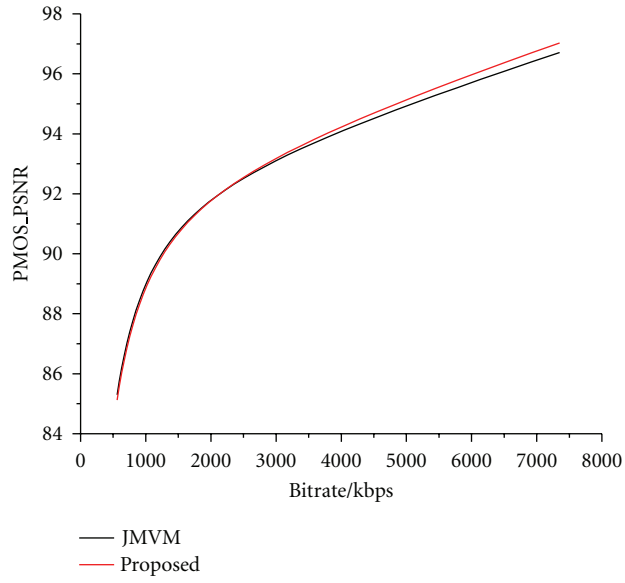
(c) Doorflowers

(d) Alt Moabit

(e) Pantomime

(f) Dog

Figure 16: Continued.



(g) Champagne tower

Figure 16: Rate-distortion performances comparisons between the proposed MVC and JMVM (distortions are measured with PMOS_PSNR).

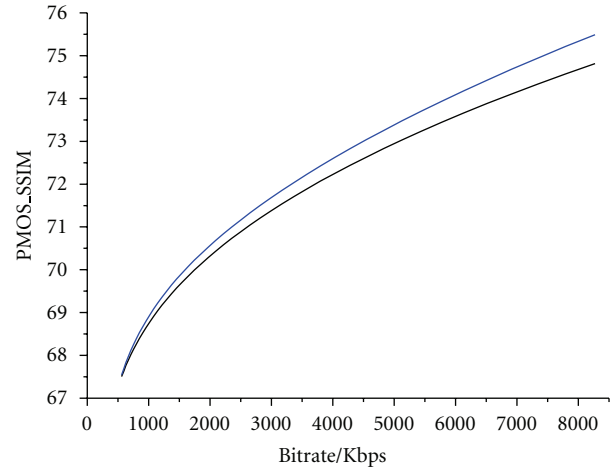
because the encoding bits of ΔQP will increase along with the increasing ΔQP. Therefore, it is reasonable to limit ΔQP within the range from 2 to 4 at low bit rate (large bQP).

6.4. MVC Experiments. SVA-based MVC experiments are implemented on the JMVM 7.0 reference software with seven multiview video sequences and their ROI masks, Ballet, Breakdancers, Doorflowers, Alt Moabit, Pantomime, Champagne tower, and Dog, to evaluate the effectiveness of the proposed SVA-based bit allocation. The MVC-HBP prediction structure is adopted for MVC simulation. Eight views and GOP Length are 15, fast motion/disparity estimation is enabled, and search range is 64. There are three kinds of picture in the MVC-HBP prediction structure: intracoded picture (I-picture), interpredicted picture (P-picture), and hierarchical bidirectional predicted picture (B-picture). In the coding experiment, all B- and P-pictures are coded with regional bit allocation optimization and I-pictures are coded with original MVC scheme without bit allocation optimization. The bQP is set as 12, 17, 22, 27, or 32, and the QPs of background and ROI are set according to (16) and obtain optimal ΔQP in Figure 15. PMOS_SSIM and PMOS_PSNR are adopted to evaluate image quality of the reconstructed video frames.

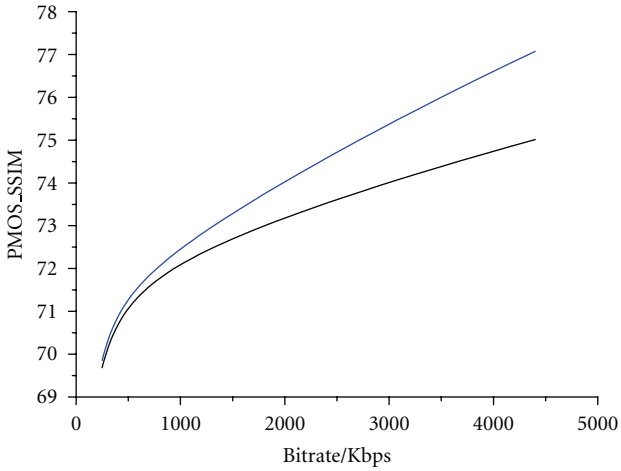
Figure 16 shows the rate-distortion curves of the proposed MVC and JMVM, where the distortions are measured with PMOS_PSNR. Figure 17 shows the rate-distortion curves of the proposed MVC and JMVM, where the distortions are measured with PMOS_SSIM. Curves in the figures are fitted with the algorithm provided in [43]. As we can see from Figure 16, more than 10% bit rate is saved while maintaining the same PMOS_PSNR for Breakdancers when bit rate is higher than 4 Mbps. For Ballet sequences, the proposed scheme attains the same coding performance at low bit rate, but improves coding performance significantly at high bit rate, that is, more than 20% bit rate is saved at high bit rate. Similar results can be found for Doorflowers, Alt Moabit, Pantomime, Champagne tower, and Dog sequences. Also, as we can see from Figure 17 in which distortion is measured with PMOS_SSIM, the proposed MVC scheme outperforms JMVM more distinctively, with 20% ∼ 40% bit rate saving while maintaining the same PMOS_SSIM, from low bit rate to high bit-rate for most of these test multiview video sequences.

Figure 18 shows images reconstructed with the proposed MVC scheme and the JMVM benchmark and Table 3 shows the objective image quality value and coding bits corresponding to Figure 18. They show the reconstructed images of the 15th frame of the 2nd view (i.e., S1T15 in Figure 4) of the test 3DV sequences. Encoding bits and another five image quality indices including PSNR_Y_{SVA}, PSNR_Y_{BG}, PSNR_Y, PMOS_SSIM, and PMOS_PSNR are compared for each sequence. PSNR_Y_{SVA}, PSNR_Y_{BG}, and PSNR_Y denote the PSNR of illumination component for SVA-based ROI regions, background region, and the entire picture, respectively. PMOS_PSNR and PMOS_SSIM represent the PMOS of PSNR_Y and SSIM, respectively. In addition, the differences of the bit-rate saving ratio and image quality indices are also given and they are computed using the following formulas:

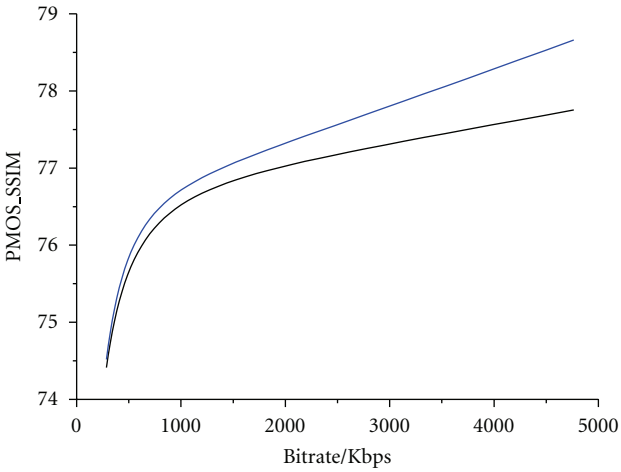
$$\Delta\Psi = \Psi_{\text{Proposed}} - \Psi_{\text{JMVM}},$$
$$\Delta EB^{i,j}[\%] = \frac{EB^{i,j}_{\text{JMVM}} - EB^{i,j}_{\text{Proposed}}}{EB^{i,j}_{\text{JMVM}}} \times 100[\%], \tag{33}$$



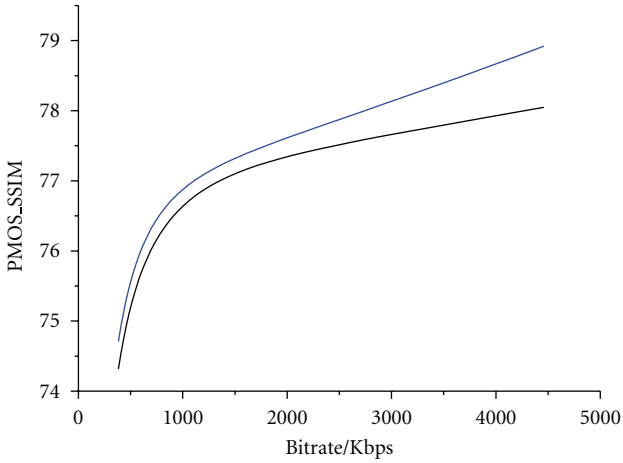
(a) Breakdancers



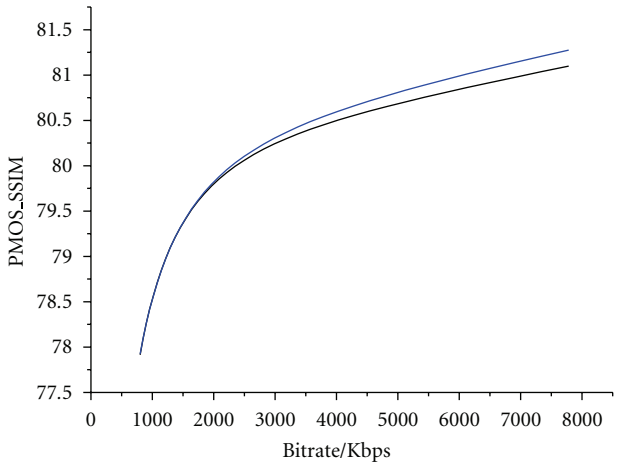
(b) Ballet



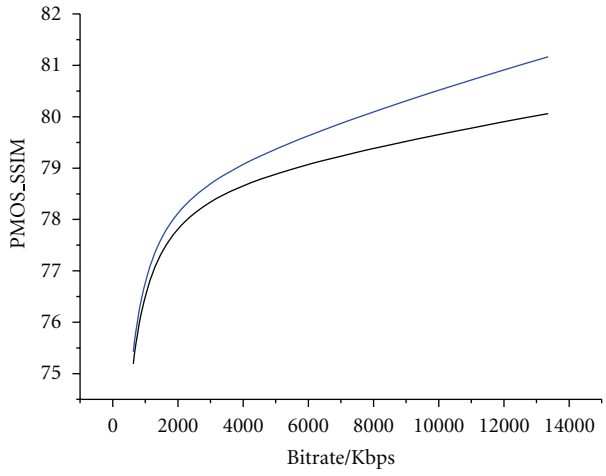
(c) Doorflowers



(d) Alt moabit

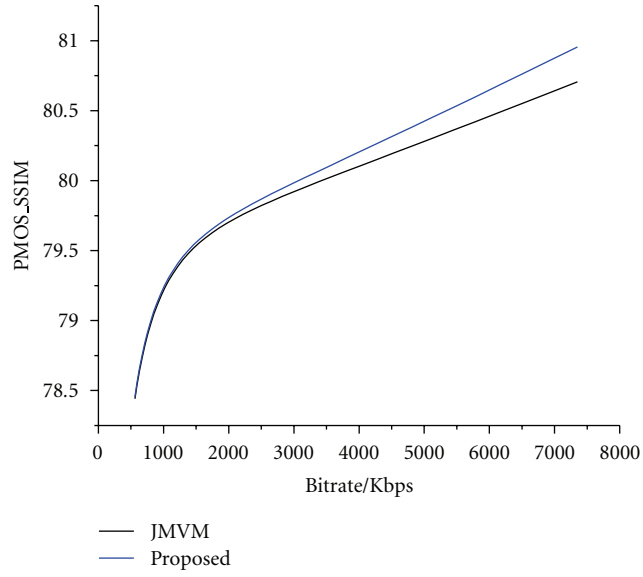


(e) Pantomime



(f) Dog

Figure 17: Continued.



(g) Champagne tower

Figure 17: Rate-distortion performances comparisons between the proposed MVC and JMVM (Distortions are measured with PMOS_SSIM).

where $\Psi \in \{\text{PSNR\_Y}_{\text{SVA}}, \text{PSNR\_Y}_{\text{BG}}, \text{PMOS\_PSNR}, \text{PMOS\_SSIM}\}$, $\Delta EB^{i,j}$ is the bit-rate saving ratio for the proposed MVC scheme with respect to the encoded picture at (i,j) position of a GOP. $EB^{i,j}_{\text{JMVM}}$ and $EB^{i,j}_{\text{proposed}}$ denote encoding bits of the coded pictures by using JMVM and the proposed MVC scheme, respectively.

Because people usually pay less attention to the background regions and more attention to ROIs, HVS is less perceptible to distortion in the background regions than that of ROIs. This implies that people are more sensitive to distortions in the ROIs than in the background region. As a result, high image quality is required in ROIs. For Ballet multiview video sequence, $\Delta\text{PSNR\_Y}_{\text{SVA}}$ is 0.46 dB while $\Delta\text{PSNR\_Y}_{\text{BG}}$ is -1.11 dB. It means that the proposed SVA-based MVC scheme improves image quality of ROI up to 0.46 dB; meanwhile, to improve compression ratio, the proposed SVA-based MVC scheme allocates fewer bits on the background regions and at the cost of its $\text{PSNR\_Y}_{\text{BG}}$. In the proposed MVC scheme, the image quality of ROIs is getting better than that of background region, that is, $\text{PSNR\_Y}_{\text{SVA}} > \text{PSNR\_Y}_{\text{BG}}$, which meets the requirements of HVS. Thus, the quality of the reconstructed images is improved. While evaluated by the regional selective image quality metrics, ΔPMOS_SSIM is 0.78 and ΔPMOS_PSNR is -0.70. It means the difference between the qualities of reconstructed images coded by the proposed MVC scheme and JMVM is tiny and imperceptible. However, the important and interesting fact is that $\Delta EB^{2,15}$ is 21.06%, which indicates that 21.06% bit rate saving is achieved by the proposed MVC scheme while comparing with JMVM benchmark. Similar results can also be found for Breakdancers, Doorflowers, Alt Moabit, and Dog sequence. For Pantomime and Champagne tower sequences, because the background regions are very flat and smooth, MBs in these regions are coded with SKIP/DIRECT mode and only very few bits are allocated by original JMVM, thus, a relative low saving ratio, 8.19% and 8.58%, is achieved by the proposed MVC.

In summary, the proposed MVC scheme achieves significant bit-rate saving ratio, up to 21.06% $\sim$ 34.29%; meanwhile, the ROIs' image quality is improved up to 0.46 $\sim$ 0.61 dB at the cost of imperceptible quality degradation at background regions. Additionally, PSNR_Y of ROI is better than that of background, which meets requirements of HVS. Moreover, the proposed MVC scheme can save over 20% bit rate with imperceptible image quality degradation according to the evaluation of region selective image quality metrics.

7. Conclusions

A stereoscopic visual attention- (SVA-) based regional bit allocation optimization scheme is proposed to improve the compression efficiency of MVC. We proposed a bottom-up SVA model to simulate the visual attention mechanisms of the human visual system with stereoscopic perception. This model adopts multiple low level perceptual stimuli, including color, intensity, orientation, motion, depth, and depth contrast. Then the semantic region-of-interest (ROI) is extracted based on the saliency maps of SVA. The proposed model is not only able to efficiently simulate stereoscopic visual attention of human eyes, but also can reduce noise in each stimulus channel. Based on the extracted semantic ROIs, a regional bit allocation optimization scheme is also proposed for high compression efficiency by exploiting visual redundancies. Experimental results on MVC showed that

(a) Breakdancers

(b) Ballet

(c) Doorflowers

(d) Alt moabit

Figure 18: Continued.

(e) Pantomime

(f) Dog

(g) Champagne tower

Figure 18: Subjective and objective quality comparison of the reconstructed images (Left: JMVM, Right: Proposed).

the proposed bit allocation algorithm can achieve over 21.06% ~ 34.29% bit-rate saving at high bit rate while maintaining the same objective image quality and subjective image qualities. Meanwhile, the image quality of ROIs is improved by 0.46 ~ 0.61 dB at the cost of indiscriminate image quality degradation in background regions, which is less conspicuous and sensitive to human visual system. It can be foreseen that the stereoscopic visual attention will play a more important role in the areas such as content-oriented three-dimensional video processing, video retrieval, and computer vision in future.

Acknowledgments

The Interactive Visual Media Group at Microsoft Research, HHI, and Nagoya University have kindly provided The authors with multiview video sequences and depth maps. Thanks are due to Dr. Sam Kwong for giving us many good suggestions and help. This work is supported by the Natural Science Foundation of China (Grant 60872094, 60832003), 863 Project of China (2009AA01Z327). It was also sponsored by K.C.Wong Magna Fund in Ningbo University.

References

[1] K. Muller, P. Merkle, and T. Wiegand, "Compressing time-varying visual content," *IEEE Signal Processing Magazine*, vol. 24, no. 6, pp. 58–65, 2007.

[2] M. Tanimoto, "Overview of free viewpoint television," *Signal Processing: Image Communication*, vol. 21, no. 6, pp. 454–461, 2006.

[3] P. Merkle, A. Smolic, K. Müller, and T. Wiegand, "Multi-view video plus depth representation and coding," in *Proceedings of the International Conference on Image Processing (ICIP '07)*, vol. 1, pp. 201–204, San Antonio, Tex, USA, 2007.

[4] "Survey of algorithms used for multi-view video coding (MVC)," ISO/IEC JTC1/ SC29/WG11, N6909, Hong Kong, China, January 2005.

[5] S. Yea and A. Vetro, "View synthesis prediction for multiview video coding," *Signal Processing: Image Communication*, vol. 24, no. 1-2, pp. 89–100, 2009.

[6] Z. Yun, G. Y. Jiang, Y. Mei, and S. H. Yo, "Adaptive multiview video coding scheme based on spatiotemporal correlation analyses," *ETRI Journal*, vol. 31, no. 2, pp. 151–161, 2009.

[7] P. Merkle and K. Müller, "Efficient prediction structures for multiview video coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 11, pp. 1461–1473, 2007.

[8] Z. Lu, W. Lin, X. Yang, E. Ong, and S. Yao, "Modeling visual attention's modulatory aftereffects on visual sensitivity and quality evaluation," *IEEE Transactions on Image Processing*, vol. 14, no. 11, pp. 1928–1942, 2005.

[9] J.-R. Ohm, "Encoding and reconstruction of multiview video objects," *IEEE Signal Processing Magazine*, vol. 16, no. 3, pp. 47–54, 1999.

[10] J. Han, K. N. Ngan, M. Li, and H.-J. Zhang, "Unsupervised extraction of visual attention objects in color images," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 16, no. 1, pp. 141–145, 2006.

[11] Y.-F. Ma, X.-S. Hua, L. Lu, and H.-J. Zhang, "A generic framework of user attention model and its application in video summarization," *IEEE Transactions on Multimedia*, vol. 7, no. 5, pp. 907–919, 2005.

[12] L. Itti and C. Koch, "Computational modelling of visual attention," *Nature Reviews Neuroscience*, vol. 2, no. 3, pp. 194–203, 2001.

[13] A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognitive Psychology*, vol. 12, no. 1, pp. 97–136, 1980.

[14] G. Zhai, Q. Chen, X. Yang, and W. Zhang, "Scalable visual sensitivity profile estimation," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '08)*, pp. 873–876, Las Vegas, Nev, USA, April 2008.

[15] Y. Zhai and M. Shah, "Visual attention detection in video sequences using spatiotemporal cues," in *Proceedings of the 14th Annual ACM International Conference on Multimedia (MM '06)*, pp. 815–824, Santa Barbara, Calif, USA, October 2006.

[16] P. P. Wang, W. Zhang, J. Li, and Y. Zhang, "Realtime detection of salient moving object: a multi-core solution," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '06)*, pp. 1481–1484, Las Vegas, Nev, USA, April 2008.

[17] E. Kaminsky, D. Grois, and O. Hadar, "Dynamic computational complexity and bit allocation for optimizing H.264/AVC video compression," *Journal of Visual Communication and Image Representation*, vol. 19, no. 1, pp. 56–74, 2008.

[18] Y. Lu, J. Xie, H. Li, and H. Cui, "GOP-level bit allocation using reverse dynamic programming," *Tsinghua Science and Technology*, vol. 14, no. 2, pp. 183–188, 2009.

[19] L. Shen, Z. Liu, Z. Zhang, and X. Shi, "Frame-level bit allocation based on incremental PID algorithm and frame complexity estimation," *Journal of Visual Communication and Image Representation*, vol. 20, no. 1, pp. 28–34, 2009.

[20] N. Özbek and A. M. Tekalp, "Content-aware bit allocation in scalable multi-view video coding," in *Proceedings of the Multimedia Content Representation, Classification and Security (MRCS '06)*, vol. 4105 of *Lecture Notes in Computer Sciences*, pp. 691–698, September 2006.

[21] Z. Chen, J. Han, and K. Ngi, "Dynamic bit allocation for multiple video object coding," *IEEE Transactions on Multimedia*, vol. 8, no. 6, pp. 1117–1124, 2006.

[22] H. Wang, G. M. Schuster, and A. K. Katsaggelos, "Rate-distortion optimal bit allocation for object-based video coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, no. 9, pp. 1113–1123, 2005.

[23] M.-C. Chi, M.-J. Chen, C.-H. Yeh, and J.-A. Jhu, "Region-of-interest video coding based on rate and distortion variations for H.263+," *Signal Processing: Image Communication*, vol. 23, no. 2, pp. 127–142, 2008.

[24] C.-W. Tang, C.-H. Chen, Y.-H. Yu, and C.-J. Tsai, "Visual sensitivity guided bit allocation for video coding," *IEEE Transactions on Multimedia*, vol. 8, no. 1, pp. 11–18, 2006.

[25] P. Kauff, N. Atzpadin, C. Fehn et al., "Depth map creation and image-based rendering for advanced 3DTV services providing interoperability and scalability," *Signal Processing: Image Communication*, vol. 22, no. 2, pp. 217–234, 2007.

[26] Y. Zhang, G. Jiang, M. Yu, and K. Chen, "Stereoscopic visual attention model for 3D video," in *Proceedings of the International Multimedia Modeling Conference (MMM '10)*, vol. 5916 of *Lecture Notes in Computer Sciences*, pp. 314–324, January 2010.

[27] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[28] J. L. Barron, D. J. Fleet, and S. S. Beauchemin, "Performance of optical flow techniques," *International Journal of Computer Vision*, vol. 12, no. 1, pp. 43–77, 1994.

[29] L. Itti and C. Koch, "Feature combination strategies for saliency-based visual attention systems," *Journal of Electronic Imaging*, vol. 10, no. 1, pp. 161–169, 2001.

[30] M. Tanimoto, T. Fujii, and K. Suzuki, "Improvement of depth map estimation and view synthesis," ISO/IEC JTC1/SC29/WG11, M15090, Antalya, Turkey, January 2008.

[31] F. Qi, J. J. Wu, and G. M. Shi, "Extracting regions of attention by imitating the human visual system," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '09)*, pp. 1905–1908, Taipei, Taiwan, April 2009.

[32] Y. Zhang, G. Y. Jiang, M. Yu, Y. Yang, Z. J. Peng, and K. Chen, "Depth perceptual region-of-interest based multiview video coding," *Journal of Visual Communication and Image Representation*, vol. 21, no. 5-6, pp. 498–512, 2010.

[33] K. Takagi, Y. Takishima, and Y. Nakajima, "A study on rate distortion optimization scheme for JVT coder," in *Visual Communication and Image Processing*, vol. 5150 of *Proceedings of SPIE*, pp. 914–923, Lugano, Switzerland, July 2003.

[34] U. Engelke, V. X. Nguyen, and H.-J. Zepernick, "Regional attention to structural degradations for perceptual image quality metric design," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '08)*, pp. 869–872, Las Vegas, Nev, USA, March 2008.

[35] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[36] I. Feldmann, M. Mueller, F. Zilly, et al., "HHI test material for 3D video," ISO/IEC JTC1/SC29/WG11, M15413, Archamps, France, April 2008.

[37] C. L. Zitnick, S. B. Kang, M. Uyttendaele, S. Winder, and R. Szeliski, "High-quality video view interpolation using a layered representation," in *Proceedings of ACM SIGGRAPH Transactions on Graphics*, pp. 600–608, ACM, Los Angeles, Calif, USA, August 2004.

[38] M. Tanimoto, T. Fujii, and N. Fukushima, "1D parallel test sequences for MPEG-FTV," ISO/IEC JTC1/SC29/WG11, M15378, Archamps, France, April 2008.

[39] O. Stankiewicz and K. Wegner, "Depth map estimation software version 2," ISO/IEC JTC1/SC29/WG11, M15338, Archamps, France, April 2008.

[40] ITU-R Recommendation BT.500-11, "Methodology for the subjective assessment of the quality of television pictures," 2002.

[41] R. Rajae-Joordens and J. Engel, "Paired comparisons in visual perception studies using small sample sizes," *Displays*, vol. 26, no. 1, pp. 1–7, 2005.

[42] A. Vetro, P. Pandit, H. Kimata, A. Smolic, and Y. K. Wang, "Joint multiview video model (JMVM) 7.0," Tech. Rep. JVT-Z207, Joint Video Team of ITU-T VCEG and ISO/IEC MPEG, Antalya, Turkey, January 2008.

[43] G. Bjontegaard, "Calculation of average PSNR differences between RD-curves," ITU-T VCEG, VCEG-M33, April 2001.