J. Vis. Commun. Image R. 21 (2010) 498–512

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

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

Depth perceptual region-of-interest based multiview video coding

Yun Zhang [a,b,c], Gangyi Jiang [a,b,*], Mei Yu [a], You Yang [b], Zongju Peng [a], Ken Chen [a]

[a] Faculty of Information Science and Engineering, Ningbo University, Ningbo 315211, China
[b] Institute of Computing Technology, Chinese Academic of Sciences, Beijing 100080, China
[c] Graduate School of the Chinese Academic of Sciences, Beijing 100080, China

ARTICLE INFO

Article history:
Received 1 July 2009
Accepted 3 March 2010
Available online 15 March 2010

Keywords:
Three-dimensional television
Multiview video coding
Multiview video plus depth
Depth perceptual region-of-interest
Bit allocation
Inter-view correlation
Hierarchical B picture
Human visual system

ABSTRACT

MultiView Video (MVV) has attracted considerable attention recently since it is capable of providing users with three-dimensional perception and interactive functionalities. However, these MVV data require large mount of storage and bandwidth in network transmission. In this paper, we present a novel Depth Perceptual Region-Of-Interest (DP-ROI) based Multiview Video Coding (RMVC) scheme to extensively improve data compression efficiency by exploiting redundancies in depth perception. Firstly, we define DP-ROI according to the three-dimensional depth sensation of human visual system. Then, a framework of RMVC is developed to improve compression efficiency by properly segmenting the MVV into different macroblock wise DP-ROIs and encoding them separately. And then, we propose three fast depth based DP-ROI extraction and tracking algorithms by jointly using motion, texture, depth as well as previous extracted DP-ROIs. Finally, on the basis of the extracted DP-ROI, bit allocation optimization model is proposed to allocate more bits on DP-ROIs for high image quality and fewer bits on background regions for high compression ratio. Experimental results show that the presented RMVC scheme achieves significant coding gains at high rate while comparing with original joint multiview video model. To be specific, up to 14.22–23.32% bit-rate are saved while 0.16–0.68 dB coding gains are achieved in DP-ROIs at the cost of the image quality degradation in background.

1. Introduction

MultiView Video (MVV) is attracting a lot of interests as one of the new video types because it provides viewers with Three-Dimensional (3D) scene and allows viewers freely to change their viewpoints as if they were there. With these features, it will be used for many new multimedia applications, such as photorealistic rendering of 3D scenes, Free Viewpoint Video (FVV) [1] and 3D TeleVision (3DTV) communications [2,3]. Multiview Video plus Depth (MVD) data format is proposed as the main 3D representation format that supports 3DTV and FVV [4,5]. It consists of MVV and multiple associated depth videos which indicate the distance between the captured scene and cameras. This MVD data format fulfills the 3D video system's requirements and supports wide angle of 3D displays and auto-stereoscopic displays [6]. Moreover, it also allows rendering a continuum of output views with high image quality and low-complexity [7]. However, since MVV is generated by simultaneous recording of a moving scene with multiple cameras located at different positions, it is with huge amount of data volume and we need to develop Multiview Video Coding (MVC) schemes to encode these MVV data efficiently.

MVC has been studied in relation to several video coding standards. MPEG-2 multiview profile is proposed for stereoscopic video coding. The MPEG-4 multiple auxiliary components are also related to MVC. In addition, H.263 and H.264 have also been tried for MVC. Since ISO/IEC JTC1/SC29/WG11 Moving Picture Experts Group (MPEG) has recognized the importance of MVC technologies, an Ad-Hoc Group (AHG) on 3D Audio and Visual (3DAV) was established. MPEG has surveyed some of MVC schemes, such as 'Group-of-GOP prediction (GoGOP)', 'sequential view prediction', 'checkerboard decomposition', and so on [8]. Oka et al. proposed MVC scheme using multi-directional picture for ray-space data [9]. Kitahara et al. proposed GoGOP prediction structure to improve the random accessibility of MVV system by adopting multiple intra frames [10]. Yea et al. proposed view synthesis prediction based MVC scheme to improve inter-view compression efficiency [11]. Zhang et al. developed an efficient MVC algorithm which adaptively selects optimal prediction structure according to the spatio-temporal correlation of MVV sequence [12]. Merkle et al. proposed a MVC scheme based on H.264/AVC using Hierarchical B Pictures (MVC-HBP) with superior compression efficiency and temporal scalability [13]. MVC-HBP has been adopted into MVC standardization draft by Joint Video Team (JVT) and used in

* Corresponding author at: Faculty of Information Science and Engineering, Ningbo University, Ningbo 315211, China. Fax: +86 574 87600940.
E-mail address: jianggangyi@126.com (G. Jiang).

1047-3203/$ - see front matter Crown Copyright © 2010 Published by Elsevier Inc. All rights reserved.
doi:10.1016/j.jvcir.2010.03.002

the Joint Multiview Video Model (JMVM), which was developed as an extension of H.264/MPEG-4 AVC.

In previous MVC schemes [8–13], intra, inter and inter-view prediction compensation technologies are adopted to eliminate spatial, temporal and inter-view redundancies, respectively. Additionally, YUV color space transform, integer transform and quantization technologies are utilized to explore visual redundancies including chroma redundancies and high frequency redundancies. However, these schemes have not taken regional selective attention and 3D depth perception of Human Visual System (HVS) into consideration. It has been revealed that HVS is more sensitive to the distortion in Region-Of-Interest (ROI) than background regions [14]. This fact indicates the visual redundancy coming from regional selective interests exists in MVV. Accordingly, ROI based approaches can benefit compression efficiency of video coding, quality of virtual image rendering and reliability of network transmission for MVV system [15]. Yang and Kaminsky et al. proposed ROI based bit allocation and dynamical computational power allocation to improve coding efficiency [16,17]. However, these bit allocation schemes were proposed for single-view video coding and cannot be directly applied to MVC because inter-view prediction is also adopted in MVC.

Research on ROI extraction is one of the most challenging topics for content based video processing. Usually, in single-view video, ROI extraction algorithm adopts color, illumination, contour and motion as key features [18]. Additionally, contrast, visual attention [19] and face detection [16] are also used for semantic ROI extraction. However, they are complex and hard to separate foreground from background due to lacking of depth information. Fortunately, these problems can be solved in MVV because depth information and multiple-channel videos are available. In [20], initial object segmentation is obtained by merging neighboring sampling positions with disparity vectors of similar size and direction. Starting from this initial segmentation, true object borders are then detected. Similarly, Marugame et al. proposed an object extraction method by utilizing disparity estimation and object contours [21]. In the existing ROI extraction schemes, previously extracted ROIs have not efficiently utilized for ROI extraction for time consecutive frames or neighboring views. Furthermore, ROI in MVV is different from that of conventional single-view video because MVV provides 3D depth perception, which makes people more likely interested in regions with small depth value and depth discontinuous regions. It means the contents close to the viewers and the regions provide strong depth perception shall be given a higher priority. Furthermore, the existing object tracking algorithms [22–24] were proposed for object tracking in time dimension but not suitable for object tracking in view dimension for MVV.

In this paper, we present a novel Depth Perceptual Region-Of-Interest (DP-ROI) based Multiview Video Coding (RMVC) scheme to improve video compression efficiency by extensively exploiting redundancies in depth perception. The rest of this paper is organized as follows. We define a DP-ROI for MVV and present a framework of RMVC in Section 2. Then, we discuss four ROI extraction schemes and propose related low-complexity DP-ROI extraction and tracking algorithms in Section 3. And then, Section 4 proposes regional selective bit allocation optimization based on extracted DP-ROIs. Section 5 presents ROI based image quality metrics. DP-ROI extraction, bit allocation and coding performance of RMVC are experimentally analyzed in Section 6. Finally, conclusions are given.

2. Framework of the RMVC

Generally, depth perception is based on various depth cues such as illumination, relative size, motion, occlusion, texture gradient, geometric perspective, disparity and so on. However, the most effective depth perception sensation is obtained by viewing a scene from slightly different viewing positions, i.e. disparity. In single-view video, ROI is often related to moving regions and textural regions. However, DP-ROI in MVV is extensively related to 3D perception in following two aspects. One is the regions with small depth values, i.e. large disparity, because they are closer to viewer and sometimes pop-out from video screen. The other is depth discontinuous regions which draw much attention because they provide relative strong 3D depth sensation.

Fig. 1 shows a framework of 3D video system in which RMVC is proposed for high compression efficiency and high rendering quality. The pink part of the Fig. 1 shows the core of RMVC. It mainly includes DP-ROI extraction module, MVC encoder using MVC-HBP prediction structure [13], and DP-ROI based bit allocation optimization. First, N synchronized color videos are captured by parallel or arc arranged video capture system. Then, N depth videos synchronized with color videos are captured by depth cameras or generated by disparity matching based depth creation algorithms. By using depth video and multiview texture video, the DP-ROI extraction module quickly and efficiently extracts macroblock (MB) wise DP-ROI mask for block-based video coding. By taking the advantage of the extracted DP-ROIs, MVC encoder and depth encoder based on advanced video coding standard are optimized for low-complexity, bit-rate saving in background regions and better quality in DP-ROIs. The MB-wise DP-ROI mask is not necessary transmitted to the client, thus it will not put burden on the network bandwidth. Thus, RMVC scheme is compatible with current block-based MVC standard and needs no high-level syntax modification. Finally, compressed color and depth bitstream are multiplexed, synchronized and transmitted.

At the client, the color and depth bitstream is de-multiplexed and decoded by MVC decoder and depth decoder, respectively. For multiview imaging, a scene is usually captured by limited number of cameras. It is therefore necessary to generate user's requested views from the limited captured views. With the decoded MVV, depth videos as well as the transferred video cameras' calibration data, arbitrary view generation module can efficiently render a continuum of output views, $N'(N' > N)$, through depth image based rendering [7]. According to different types of displayer, e.g. HDTV, stereoscopic displayer or multiview displayer, different number of views are rendered and transmitted. Finally, view selection signal from interactors, e.g. mouse, joystick or header tracking devices, feedback to view generation model for interactive viewing.

3. DP-ROI extraction for MVV

3.1. DP-ROI extraction and tracking schemes for MVV

Let 2D-GOP denote a 2D picture array, in which each row holds temporally successive pictures of one view, and each column consists of spatially neighboring views captured at the same time, as shown in Fig. 2. Obviously, the simplest approach is extracting ROI independently for all the frames in a 2D-GOP, as shown in Fig. 2(a). In the figure, each white rectangle represents one frame extracting ROI without using temporally or inter-view adjacent ROIs' information. As MVV data is originated from the same scene with different viewpoints, the inherent dependencies include inter-view ones among neighboring camera views and temporal ones among temporally successive images of each view. These dependencies can be used to joint ROI extraction in view and temporal dimensions. So the independent scheme, shown in Fig. 2(a), is highly time-consuming because previously extracted ROIs have not been efficiently utilized.

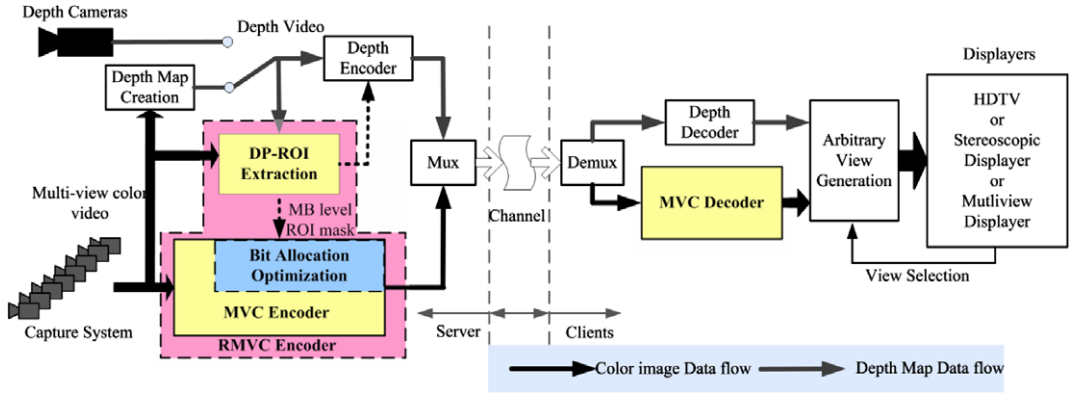

Fig. 1. Framework of 3D video system based on RMVC.

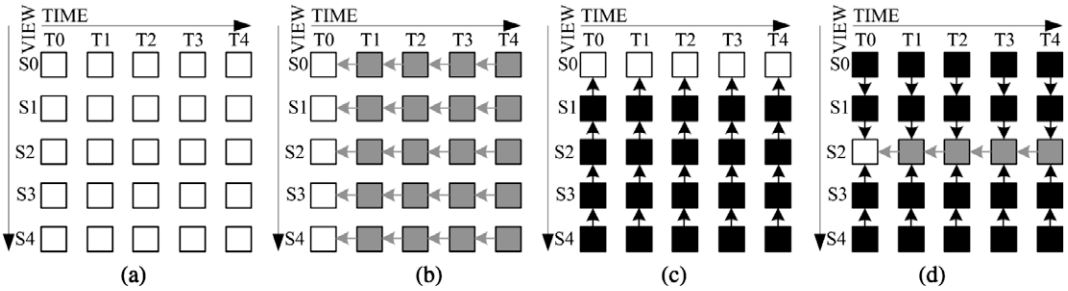

Fig. 2. ROI extraction and tracking schemes. (a) Independent scheme, (b) temporal extraction scheme, (c) inter-view extraction scheme and (d) joint extraction scheme.

Fig. 2(b) shows temporal extraction scheme, where gray rectangle, e.g. S0T 1 frame, represents a frame that extracts ROI by using ROIs of temporal preceding frames. The solid arrow represents tracking direction. ROIs of T 0 frames are extracted firstly. The rest frames of the 2D-GOP track the extracted ROIs of temporal preceding frames in order to reduce computational complexity. Though temporal extraction scheme utilizes temporal correlation to extract ROIs, it is simply expanded from conventional single-view video tracking and has not taken inter-view correlation into account. Fig. 2(c) shows inter-view extraction scheme, in which black rectangle, e.g. S1T 0, represents a frame that extracts ROIs by using neighboring extracted ROIs and inter-view dependencies of MVV. ROIs of frames in view S0 are extracted firstly. Then, the rest frames track the previously extracted ROIs of neighboring views. However, inter-view extraction scheme has not utilized temporal dependencies. Moreover, inter-view extraction scheme tracks view by view sequentially from S0. Error may be propagated due to long tracking path length.

Based on the above analyses, we propose a joint extraction scheme, shown in Fig. 2(d), in which previously extracted ROIs of temporal preceding or inter-view neighboring frames are utilized. This scheme can improve ROI consistency among inter-view/temporal frames and reduce computational complexity. For a 2D-GOP, there is only one frame of the center view, i.e. S2T 0 denoted by white rectangle, extracting ROI independently. That is ROIs in S2T 0 are extracted without prior ROIs' information. Then, the temporal successive frames in view S2, labeled as gray rectangles, track the ROIs of temporal preceding frames. Finally, neighboring views, i.e. S0, S1, S3 and S4, labeled as black rectangles, adopt the inter-view geometry dependencies with S2 to extract ROIs. The joint extraction scheme tracks ROIs from both temporal preceding frames and neighboring views, which is able to more efficiently reduce computational complexity and improve ROI consistency among inter-view/temporal frames In addition, the ROIs of the center view are tracked by neighboring views, which relieve error propagation and occlusion problems among views.

In the following subsections, Depth Based DP-ROI Extraction (DBDE), the ROI extraction algorithm indicated as the white rectangles in Fig. 2, is presented first. Then, we present inter-view and temporal tracking algorithms, the black and gray rectangles in Fig. 2. They adopts previous extracted DP-ROIs and inter-view/temporal correlations to facilitate DP-ROI extraction.

3.2. Depth based DP-ROI extraction

Depth data in MVD supports high quality rendering and low-complexity of rendering a continuum of output views. In addition, depth videos can also be utilized to facilitate semantic ROI extraction. In this study, DBDE is proposed to extract DP-ROI by jointly using motion, texture and depth information of MVD data.

Fig. 3 shows the flowchart of DBDE algorithm which includes the following four steps.

Step 1. Let vectors **F** and **D** be one frame of color and depth video, respectively. Motion mask $\mathbf{M}^m$ is extracted from the differences among temporally successive frames. We segment foreground regions $\mathbf{M}^f$ from the background regions by using a threshold, and the background regions are set as non-interested regions. Then, contours mask of color video, $\mathbf{M}^c$, and depth discontinuous regions, $\mathbf{M}^d$, are acquired by using edge detection algorithm.

Step 2. Because moving object and depth discontinuous regions are usually ROIs, we construct these two regions as a characteristic region, $\mathbf{M}^f \cap [\mathbf{M}^m \cup \mathbf{M}^d]$, which will be used as seeds of determining ROI depth planes. On the basis of the histogram value of the depth in $\mathbf{M}^f \cap [\mathbf{M}^m \cup \mathbf{M}^d]$ regions, the depth image **D** is divided into

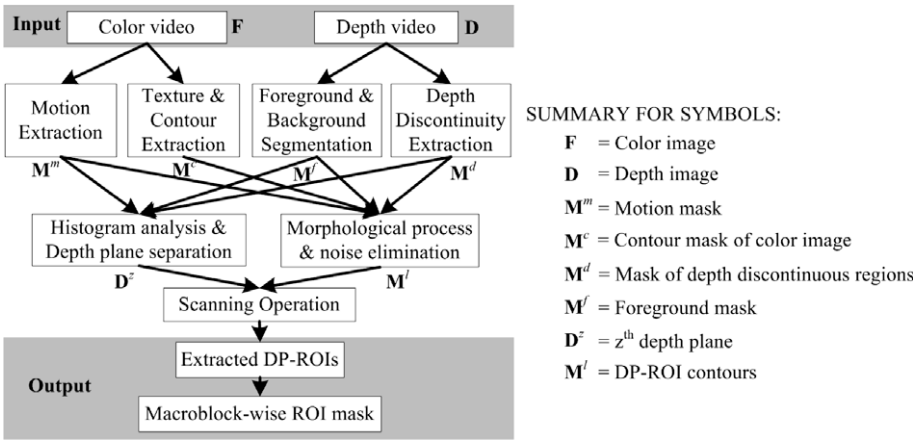

Fig. 3. Flowchart of the proposed DBDE algorithm.

several depth planes $\mathbf{D}^z$ according to the mean and variance of the histogram, where z is the ordinal number of the depth plane.

Step 3. The DP-ROI contours are constructed by integrating foreground motion region, depth contour and color contour, i.e. $\mathbf{M}^f \cap [\mathbf{M}^m \cup \mathbf{M}^d \cup \mathbf{M}^c]$. Then, morphological process, contour recovery and noise elimination operations are performed on $\mathbf{M}^f \cap [\mathbf{M}^m \cup \mathbf{M}^d \cup \mathbf{M}^c]$ to build a closed and more reliable DP-ROI contours, $\mathbf{M}^l$.

Step 4. To exclude the background regions in $\mathbf{D}^z$, a boundary scanning process guided by $\mathbf{M}^l$ is conducted on depth planes $\mathbf{D}^z$ by supposing image boundaries are background. Bit allocation and dynamic computational power allocation are performed on each MB for block-based MVV encoder. Therefore, a MB-wise DP-ROI mask is generated based on the extracted DP-ROI. The extraction process in DBDE is 8×8 block-based and most operations are simple logic operations for low-complexity.

3.3. Inter-view DP-ROI extractions

MVV sequences are captured from the same scene by a camera array in which cameras are in different positions. There are intrinsic inter-view dependencies among views but not simply a global shift. Fig. 4 shows comparisons on pictures in different views of Ballet sequence, captured by 1D-arc camera arrangement. Generally, two pictures in ith view and jth view are similar and there are high inter-view correlations among the two pictures. After DP-ROIs of one view, e.g. ith view, have been extracted, we can take advantage of them to extract corresponding DP-ROIs in another neighboring view, e.g. jth view. However, the relative distance between the girl and the wallpaper are obviously different in two views, shown as the circle marked as '1' in Fig. 4. It is the relative position displacements between views. In addition, there are occlusion and disclosing problems in the pictures, shown as the circle marked as '2' and '3' in Fig. 4. The simplest approach is adopting global disparity to calculate the global offsets and find the corresponding DP-ROI. It cannot tackle MVV sequences captured by 1D-arc camera arrangement and multiple DP-ROIs with different depth values. To tackle the above mentioned situations and extract DP-ROIs in neighboring views efficiently, we propose inter-view ROI tracking method based on depth information.

Let $\mathbf{Q} = (X, Y, Z)$ be a point in the world coordinate system, $\mathbf{q}_a = (x_a, y_a)$ be coordinate of a pixel, projected from $\mathbf{Q}$, on image plane a, $\mathbf{q}_b = (x_b, y_b)$ be coordinate of a projected pixel on the image plane b. Let Ω_{ROI} be a set of DP-ROIs' coordinates in the world coordinate and $\Omega^{\Psi}_{\mathrm{ROI}}$ be a set of DP-ROIs' coordinates in image plane Ψ, $\Psi \in \{a, b\}$. Let $\bar{\mathbf{Q}} = (X, Y, Z, \mathbf{I})$, $\bar{\mathbf{q}}_a = (x_a, y_a, \mathbf{I}_a)$ and $\bar{\mathbf{q}}_b = (x_b, y_b, \mathbf{I}_b)$ be

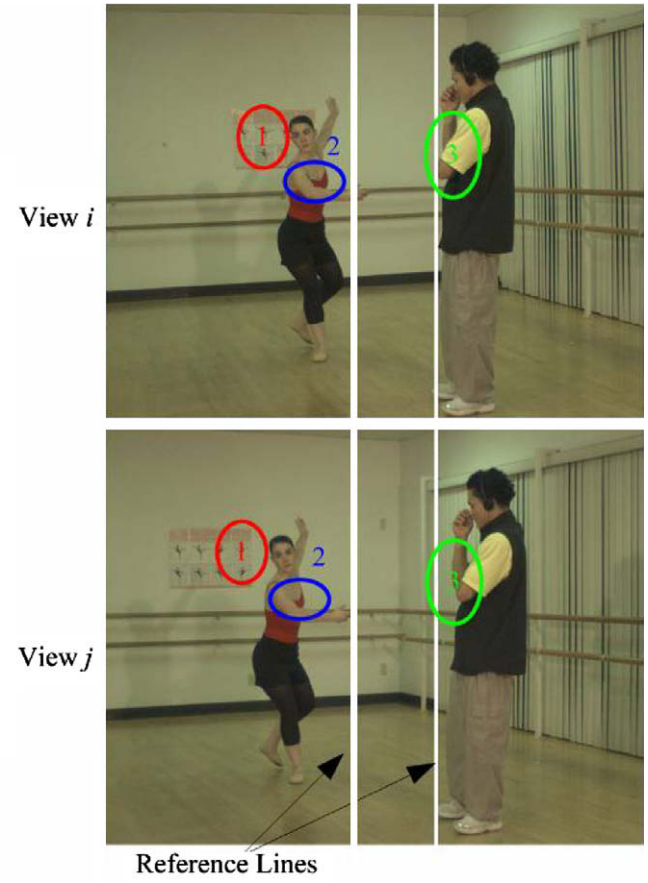

Fig. 4. Comparisons on pictures of different views.

augmented vectors of $\mathbf{Q}$, $\mathbf{q}_a$ and $\mathbf{q}_b$, i.e. $\mathbf{I}_a$ and $\mathbf{I}_b$ are the pixel value projected from $\mathbf{I}$. Then the projection equations result to

$$\begin{cases} s_a \bar{\mathbf{q}}_a = \mathbf{A}_a \cdot \mathbf{P} \cdot \begin{bmatrix} \mathbf{R}_a & \mathbf{t}_a \\ 0 & 1 \end{bmatrix} \cdot \bar{\mathbf{Q}} \\ s_b \bar{\mathbf{q}}_b = \mathbf{A}_b \cdot \mathbf{P} \cdot \begin{bmatrix} \mathbf{R}_b & \mathbf{t}_b \\ 0 & 1 \end{bmatrix} \cdot \bar{\mathbf{Q}} \end{cases} \quad (1)$$

where $\mathbf{Q} \in \Omega_{\mathrm{ROI}}$, $\mathbf{P}$ is a 3×4 normalized perspective projection matrix and s_a and s_b are scalars. The rotations, $\mathbf{R}_a$ and $\mathbf{R}_b$, and the translations, $\mathbf{t}_a$ and $\mathbf{t}_b$, form a 4×4 matrix that transforms a 3D point from the world coordinate into the camera coordinate of the image

plane. $\mathbf{A}_a$ and $\mathbf{A}_b$ are matrixes that specify the intrinsic parameters of the a and b camera, respectively.

From Eq. (1), if DP-ROI of the ath view has been extracted, the DP-ROI in the bth view can be expressed as

$$Z_b\bar{\mathbf{q}}_b = Z_a\mathbf{A}_b\mathbf{R}_b\mathbf{R}_a^{-1}\mathbf{A}_a^{-1}\bar{\mathbf{q}}_a - \mathbf{A}_b\mathbf{R}_b\mathbf{R}_a^{-1}\mathbf{t}_a + \mathbf{A}_k\mathbf{t}_a \quad \text{if } \mathbf{q}_a \in \Omega_{\text{ROI}}^a, \tag{2}$$

where Z_a and Z_b are the corresponding depth values of $\mathbf{q}_a$ and $\mathbf{q}_b$, $\mathbf{q}_b \in \Omega_{\text{ROI}}^b$. According to Eq. (2), the coordinate $\mathbf{q}_b$ is determined by $\mathbf{q}_a$ and Z_a, and it is rewritten as $\bar{\mathbf{q}}_b = G(\bar{\mathbf{q}}_a, Z_a)$ for short, where G is the mapping function represented as Eq. (2).

Let $\bar{\mathbf{q}}_a^+ = (x_a + e_x, y_a + e_y, \mathbf{I}_a^+)$ be a neighboring point of $\bar{\mathbf{q}}_a$, where e_x and e_y are offsets in x-axis and y-axis, $\mathbf{I}_a^+$ is pixel value at position $(x_a + e_x, y_a + e_y)$ in plane a, $\mathbf{I}_b^+$ is pixel value at position $(x_b + e_x, y_b + e_y)$ in plane b. Let Z_a^+ be depth value at position $(x_a + e_x, y_a + e_y)$.

If $\bar{\mathbf{q}}_a^+ \in \Omega_{\text{ROI}}^a$, we can get the pixel corresponding to $\bar{\mathbf{q}}_a^+$ on image plane b, $\bar{\mathbf{q}}_b^+ \in \Omega_{\text{ROI}}^b$, and $\bar{\mathbf{q}}_b^+$ is calculated as

$$\bar{\mathbf{q}}_b^+ = G(\bar{\mathbf{q}}_a^+, Z_a^+). \tag{3}$$

Let $\bar{\mathbf{q}}_b^+ = (x_b + e_x, y_b + e_y, \mathbf{I}_b^+)$ be a neighboring point of $\bar{\mathbf{q}}_b$. Because depth image is relatively smooth and with high spatial correlation among neighboring pixels in the interior region of DP-ROI, $\bar{\mathbf{q}}_b^+$ is approximate to $\bar{\mathbf{q}}_b^+$ when $|e_x|$ and $|e_y|$ are smaller than T_{ex} and T_{ey}.

$$\begin{cases} \bar{\mathbf{q}}_b^+ \approx \bar{\mathbf{q}}_b^+ \\ s.t. |e_x| \leqslant T_{ex}, |e_y| \leqslant T_{ey} \end{cases}. \tag{4}$$

It is especially true when e_x and e_y are small, e.g. 0 or ±1. Only part of DP-ROIs' pixels in the image plane b needs the projection as Eq. (2), other pixels corresponding to $\bar{\mathbf{q}}_a^+ \in \Omega_{\text{ROI}}^a$ are directly calculated by Eq. (4) instead of Eq. (3). Thus, the computational complexity of DP-ROI extraction in view dimension can be efficiently reduced by jointly using Eqs. (2) and (4). Finally, small holes are filled by applying averaging filter. DP-ROI is blocklized into MB, and MB-wise DP-ROI masks are generated for block-based MVC.

3.4. DP-ROI extraction based on temporal prediction

In content based video coding, it is essential to develop a fast ROI extraction algorithm as a pre-processing stage of video coding. We present a fast DP-ROI extraction approach for temporal successive frames in this subsection. In terms of the size and locations of temporal preceding extracted DP-ROIs, we can determine predictive windows of DP-ROIs in current frame. Then, DBDE algorithm is performed to refine DP-ROIs within the predictive windows. The areas out of the predictive windows are directly set as background. As a result, more processing time can be saved as the initial windows are precisely predicted.

Let $W_{k,t}$ ($x_{k,t}$, $y_{k,t}$, $w_{k,t}$, $h_{k,t}$) be a rectangle window of the kth DP-ROI in the frame at time t, where $(x_{k,t}, y_{k,t})$ is coordinate of the centroid of $W_{k,t}$, and $w_{k,t}$ and $h_{k,t}$ are the width and height of $W_{k,t}$. Let $W'_{k,t}$ ($x'_{k,t}$, $y'_{k,t}$, $w'_{k,t}$, $h'_{k,t}$) be a predictive window of the kth DP-ROI in the frame at time t. $(x'_{k,t}, y'_{k,t})$ is coordinate of the centroid of $W'_{k,t}$, $w'_{k,t}$ and $h'_{k,t}$ are the width and height of $W'_{k,t}$. We predict $W'_{k,t}$ from DP-ROI windows of the previous p frames. The center of $W'_{k,t}$ is calculated by

$$\begin{cases} x'_{k,t} = \sum_{i=1}^{p} \xi_{k,t-i}(x_{k,t-i} - x_{k,t-i-1}) + x_{k,t-1} \\ y'_{k,t} = \sum_{i=1}^{p} \varsigma_{k,t-i}(y_{k,t-i} - y_{k,t-i-1}) + y_{k,t-1} \end{cases}, \tag{5}$$

The width and height of $W'_{k,t}$ are predicted as

$$\begin{cases} w'_{k,t} = \lambda_w \cdot \sum_{i=1}^{p} \alpha_{k,t-i} w_{k,t-i} \\ h'_{k,t} = \lambda_h \cdot \sum_{i=1}^{p} \beta_{k,t-i} h_{k,t-i} \end{cases}, \tag{6}$$

where $\alpha_{k,t-i}$, $\beta_{k,t-i}$, $\xi_{k,t-i}$ and $\zeta_{k,t-i}$ are weighting coefficients, λ_ϕ is window size scaling coefficient correlated with motion magnitude and it is calculated as

$$\lambda_\phi = 1 + \max(0, \theta \times (\phi_{k,t-1} - \phi_{k,t-2}))/\phi_{k,t-1}, \quad \phi \in \{w, h\} \tag{7}$$

where θ is a scaling coefficient. In DP-ROI prediction, motions are divided into two categories, simple and complex. For simple motion, such as slow motion or shifting, DP-ROI size and location of $t+1$ can be efficiently and precisely predicted, and λ_ϕ is set as 1.0. For complex fast motion, such as fast non-rigid motion or irregular motion, λ_ϕ is enlarged in terms of window size variance to guarantee DP-ROI is located within $W'_{k,t}$.

4. DP-ROI based bit allocation optimization for MVC-HBP prediction structure

In HVS, more attentions will be paid to ROIs then background. Thus, the distortion in ROIs is more perceptible than that of background even though they have the same distortion values. Therefore, more bits should be allocated in ROIs to improve image quality, and fewer bits are allocated to background region for high compression ratio. Chi et al. proposed a ROI video coding scheme based on rate and distortion variations, in which blurring matrices are applied to reduce the high frequency information of background region, and then the required bits for background region also decreased by using fuzzy logical rate controller [25]. Similar bit allocation strategy is also presented in [16] for ROI based single-view video coding. However, these bit allocation schemes were proposed for single-view video coding and cannot be directly applied to MVC because inter-view prediction is also adopted in MVC.

In the presented RMVC scheme, an inter-view and temporal prediction hybrid prediction structure, MVC-HBP prediction structure [13], is adopted for high compression efficiency. Then, different quantization parameters (QPs) are set for DP-ROIs, background regions and transitional regions, where transitional regions are one or two MB wide regions between DP-ROIs and background regions. These regions are designed to make image quality changes smoothly. DP-ROIs are coded with QP, QP_{ROI}^l, determined by

$$QP_{\text{ROI}}^l = \begin{cases} bQP + 3 & \text{if } l = 1 \\ QP_{\text{ROI}}^{l-1} + 1 & \text{if } l > 1 \end{cases}, \tag{8}$$

where bQP is the basis QP of MVC-HBP prediction structure, l is hierarchical level of the picture in a GOP. QPs of background and transitional regions in the lth hierarchical level picture, QP_{BG}^l and QP_T^l, are set as

$$\begin{cases} QP_{\text{BG}}^l = QP_{\text{ROI}}^l + \Delta QP \\ QP_T^l = QP_{\text{ROI}}^l + \lfloor \Delta QP/\eta \rfloor \end{cases}, \tag{9}$$

where '$\lfloor\cdot\rfloor$' is floor operation, η is a parameter bigger than 1.0, ΔQP is a QP difference between background region and DP-ROI.

RMVC scheme is on the purpose of maximizing compression ratio and image quality in ROIs while at the cost of the image quality in background. So, we need to determine the optimal ΔQP which indicates the relative amount of bits allocated between ROIs and background regions. To this end, we use two quality indices, average Bit-rate Saving Ratio (BSR), R_{BSR}, and image quality degradation, ΔD, to represent the coding performance of RMVC scheme. The BSR, which is denoted by $R_{\text{BSR}}(bQP, \Delta QP_{\text{ROI}}, \Delta QP_{\text{BG}})$, is calculated as

$$R_{\mathrm{BSR}}(bQP, \Delta QP_{\mathrm{ROI}}, \Delta QP_{\mathrm{BG}}) = \frac{1}{N \times M} \sum_{j=1}^{N} \sum_{i=1}^{M} \frac{EB^{i,j}(QP^{l}_{\mathrm{ROI}}, QP^{l}_{\mathrm{ROI}}) - EB^{i,j}(QP^{l}_{\mathrm{ROI}} + \Delta QP_{\mathrm{ROI}}, QP^{l}_{\mathrm{ROI}} + \Delta QP_{\mathrm{BG}})}{EB^{i,j}(QP^{l}_{\mathrm{ROI}}, QP^{l}_{\mathrm{ROI}})}, \quad (10)$$

where N and M are the numbers of views and time instants in one GOP, and i and j are temporal and inter-view position. $EB^{i,j}(QP_1, QP_2)$ stands for the encoding bits of frame (i, j), whose ROIs are coded with QP_1 and background regions are coded with QP_2. ΔQP_{ROI} and ΔQP_{BG} denotes an additive value to QP in ROI and background regions, respectively. Fig. 5 shows the relationship between $R_{\mathrm{BSR}}(bQP, \Delta QP, \Delta QP)$ and ΔQP. $R_{\mathrm{BSR}}(bQP, \Delta QP, \Delta QP)$ is approximate to exponential decaying function as ΔQP increases. Thus, $R_{\mathrm{BSR}}(bQP, \Delta QP, \Delta QP)$ can be predicted as

$$R_{\mathrm{BSR}}(bQP, \Delta QP, \Delta QP) = A_0 e^{-\frac{\Delta QP}{T}} + y_0, \quad (11)$$

where A_0 and T are coefficients, which are correlated with bQP but independent to the contents of MVV. y_0 is the maximum BSR. Because ROI and background regions are not mutually exclusive, we can get

$$R_{\mathrm{BSR}}(bQP, \Delta QP, \Delta QP) = R_{\mathrm{BSR}}(bQP, 0, \Delta QP) + R_{\mathrm{BSR}}(bQP, \Delta QP, 0). \quad (12)$$

Once, ROI and background regions are segmented for MVV sequence, the BSR of ROI is approximately direct proportional to that of background region while ΔQP increases. The proportion is represented by

$$\rho = \frac{R_{\mathrm{BSR}}(bQP, 0, \Delta QP)}{R_{\mathrm{BSR}}(bQP, \Delta QP, 0)}, \quad (13)$$

where ρ is independent to ΔQP. Hence, applying Eqs. (12) and (13) to Eq. (11), we can obtain

$$R_{\mathrm{BSR}}(bQP, 0, \Delta QP) = Ae^{-\frac{\Delta QP}{T}} + y, \quad (14)$$

where $A = \frac{1}{1+\rho}A_0$ and $y = \frac{1}{1+\rho}y_0$. $|A|$ indicates amplitude of bit-rate saving. T indicates a ΔQP period of BSR becoming saturated.

Moreover, image quality degradation will caused by allocating fewer bits on background regions. Let $\Delta D(bQP, 0, \Delta QP)$ be the image quality degradation and it can be calculated by

$$\Delta D(bQP, \Delta QP_{\mathrm{ROI}}, \Delta QP_{\mathrm{BG}}) = \frac{1}{N \times M} \sum_{j=1}^{N} \sum_{i=1}^{M} \times \left[Q^{i,j}(QP^{l}_{\mathrm{ROI}} + \Delta QP_{\mathrm{ROI}}, QP^{l}_{\mathrm{ROI}} + \Delta QP_{\mathrm{BG}}) - Q^{i,j}(QP^{l}_{\mathrm{ROI}}, QP^{l}_{\mathrm{ROI}}) \right], \quad (15)$$

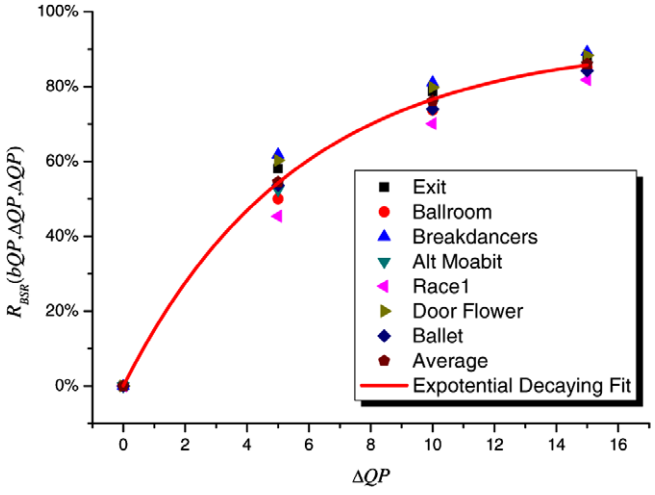

Fig. 5. The relationship between $R_{\mathrm{BSR}}(bQP, 0, \Delta QP)$ and ΔQP.

where $Q^{i,j}(QP_1, QP_2)$ stands for 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. ΔQP_{ROI} and ΔQP_{BG} denote QP changes in ROI and background regions, respectively. Because the relationship between distortion, such as peak signal-to-noise ratio (PSNR), and quantization factor is approximately linear [26], we define the image quality degradation of bit allocation, $\Delta D(bQP, 0, \Delta QP)$, as

$$\Delta D(bQP, 0, \Delta QP) = b_1 \cdot \Delta QP + a_1, \quad (16)$$

where a_1 is coefficient independent to ΔQP but correlated with bQP. $\Delta D(bQP, 0, \Delta QP)$ is negative and the value will decrease as ΔQP increases.

To achieve high compression ratio and maintain high image quality, we shall find the optimal ΔQP to maximize BSR R_{BSR} subject to a unnoticeable image quality degradation, T_D. It is mathematically expressed as

$$\begin{cases} \arg\max\{R_{\mathrm{BSR}}(bQP, 0, \Delta QP)\} \\ s.t. |\Delta D(bQP, 0, \Delta QP)| < T_D \end{cases}. \quad (17)$$

Instead of solving the constrained problem in Eq. (17), an unconstrained formulation is employed. That is

$$\arg\max_{\Delta QP \in Z^+} \{R_{\mathrm{BSR}}(bQP, 0, \Delta QP) + \mu \cdot \Delta D(bQP, 0, \Delta QP)\}, \quad (18)$$

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

$$\frac{\partial(R_{\mathrm{BSR}}(bQP, 0, \Delta QP) + \mu \cdot \Delta D(bQP, 0, \Delta QP))}{\partial \Delta QP} = 0. \quad (19)$$

Solving the Eq. (19), we calculate the optimal integer ΔQP as

$$\Delta QP = \lfloor T \ln \frac{A}{\mu \cdot T \cdot b_1} + 0.5 \rfloor. \quad (20)$$

where '$\lfloor \cdot \rfloor$' is floor operation. Meanwhile, ΔQP is clipped to 0 if ΔQP is smaller than 0. Coefficients A, T and b_1 relies on bQP and will be experimentally determined by coding experiments in Section 6.2.

5. ROI based objective image quality assessment metric

Pixel-wise image quality assessment metric, such as PSNR and Structural SIMilarity (SSIM), is not designed to match with human visual perception. It is originally designed for quality assessment of a whole distorted image $\boldsymbol{I}_D$ as compared to a whole reference image $\boldsymbol{I}_R$ without notice different human visual sensitivity between ROI and background regions. Engelke et al. proposed region-selective objective image quality metrics which is able to be combined with normalized hybrid image quality metric, reduced reference image quality assessment technique, SSIM or PSNR measures [27].

Both SSIM and PSNR have been adopted in current advanced video coding standard, H.264/AVC [28]. Thus, we use the region-selective SSIM and PSNR metrics to evaluate image quality of reconstructed frames. SSIM index between two images is computed as

$$\mathrm{SSIM} = \frac{(2\mu_R\mu_D + C_1)(2\sigma_{RD} + C_2)}{(\mu_R^2 + \mu_D^2 + C_1)(\sigma_R^2 + \sigma_D^2 + C_2)}, \quad (21)$$

Table 1
Properties of test MVV sequences.

MVVs	Provider	Size	Frame rate/ baseline/ camera array	Features
Ballet	MSR	1024 × 768	15 fps/20 cm/ 1D-arc	Both fast motion and slow motion
Breakdancers		1024 × 768	15 fps/20 cm/ 1D-arc	Very fast motion
Dog	Nagoya University	1280 × 960	30 fps/5 cm/1D	Slow motion, large image size
Doorflowers	HHI	1024 × 768	16.7 fps/ 6.5 cm/1D	Slow motion, complex texture
Alt Moabit		1024 × 768	16.7 fps/ 6.5 cm/1D	Outdoor scene, fast motion

where R and D are two nonnegative image signals to be compared, μ_R and μ_D are mean of images R and D, respectively, σ_R and σ_D are standard deviation of images R and D, respectively, and σ_{RD} is covariance of images R and D. The PSNR of illumination component, denoted by 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

$$\begin{cases} \text{PSNR\_Y} = 10\log\frac{255^2}{\text{MSE}} \\ \text{MSE} = \frac{1}{MN}\sum_{x=1}^{M}\sum_{y=1}^{N}[I_R(x,y) - I_D(x,y)]^2 \end{cases}. \tag{22}$$

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 to a region-selective metric, and the final Predictive Mean Opinion Score (PMOS) is computed as

$$\begin{cases} \Phi(\omega,\kappa,v) = [\omega\cdot\Phi_{\text{ROI}}^{\kappa} + (1-\omega)\Phi_{BG}^{\kappa}]^{\frac{1}{v}} \\ \text{PMOS}_{\Phi(\omega,\kappa,v)} = a\cdot e^{b\cdot\Phi(\omega,\kappa,v)} \\ \Phi \in \{\text{SSIM}, \text{PSNR\_}Y\} \\ \omega \in [0,1], \kappa, v \in Z^+ \end{cases}, \tag{23}$$

where ω, κ, v, a and b are derived from the subjective quality evaluation experiments in [27]. 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, performances of DP-ROI extraction algorithms and RMVC scheme are evaluated. The experiments have three subparts: (1) DP-ROI extraction experiments, (2) bit allocation optimization experiments and (3) MVC video coding experiments. Five MVV sequences, Ballet, Breakdancers, Dog, Doorflowers and Alt Moabit provided by Microsoft Research, Nagoya University and HHI, are selected for both ROI extraction experiments and MVC coding experiment. Table 1 shows the properties of test MVVs and Fig. 6 shows the eight views of the MVV sequences. They have different motion features, camera baselines, capturing frame rates, resolutions and indoor/outdoor video contents.

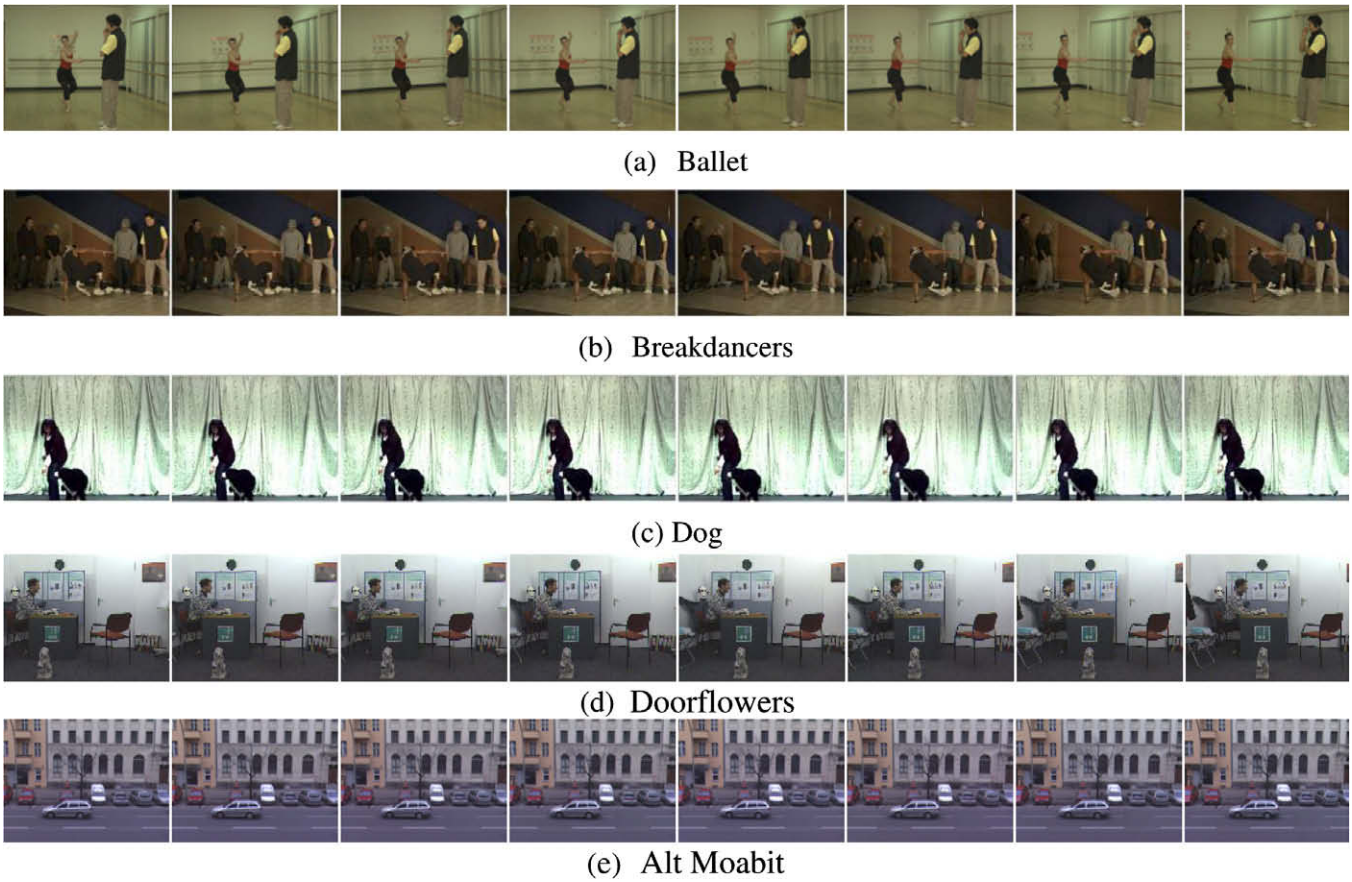

Fig. 6. Eight views of test MVV sequences.

6.1. DP-ROI extraction experiments

The DP-ROI extraction experiments include three phases, (1) DP-ROI in the center view is extracted by using the DBDE method. (2) DP-ROIs in neighboring views are extracted by using inter-view tracking and previous extracted view neighboring DP-ROIs; (3) DP-ROIs extraction for temporal successive frames.

6.1.1. DP-ROI extractions by DBDE

Fig. 7 illustrates the extracted DP-ROIs of the 4th view and 10th frame (denoted as the white rectangle in Fig. 2(d)) of Ballet sequence by using the DBDE method. Fig. 7(a) is the original image of Ballet sequence, and Fig. 7(b) shows the mask $\mathbf{M}^l$ generated by integration of motion, contour and depth discontinuities in foreground. It is the block-based contour of DP-ROI. In block-based video coding process, MB is the minimal unit of bit allocation. Thus, only MB-wise accuracy is required in the DP-ROI extraction. Additionally, to lower the complexity of DP-ROI extraction, 8×8 block-based operations are performed in the extraction processing for low-complexity. Fig. 7(c and d) shows two depth planes, $\mathbf{D}^z$, at which DP-ROI may locate. Fig. 7(e) is the finally extracted DP-ROIs. Generally, the extracted DP-ROIs cover almost all defined DP-ROIs and exclude most background regions. Additionally, the moving shadow of the female dancer is excluded by DBDE. Fig. 7(f) shows the MB-wise mask of DP-ROI corresponding to Fig. 7(e). The black blocks are DP-ROI blocks, gray blocks are the transitional regions and white regions are background. Figs. 8 illustrate other extracted results in the 4th view (25th, 40th, 55th and 70th frame) and all of the extraction results are satisfying and with high accuracy. Similar results can be obtained for other four test MVV sequences.

There are more than 13 kinds of MB modes, including DIRECT, MOTION_SKIP, INTER_16 × 16, INTER_16 × 8, INTER_8 × 16, INTER_8 × 8, INTER_8 × 8Frext, INTER_4 × 8, INTER_8 × 4, INTER_4 × 4, INTRA_16 × 16, INTRA_8 × 8 and INTRA_4 × 4 etc., are adopted in JMVM. The mode with minimal rate distortion cost will be selected as the best mode to code MVV sequence for each MB. Because of high inter-view and temporal correlations of the MVV sequence, most MBs (over 60% in [29]) are coded with DIRECT mode, in which no residual are coded and transmitted. That means these MBs coded with DIRECT mode are independent to bit allocation as QP changes. Therefore, segmented regions that will be coded with DIRECT mode in RMVC will not have impacts on coding efficiency.

6.1.2. DP-ROI extractions in view dimension

Only partial pixels of DP-ROIs (i.e. $1/[(T_{ex}+1)\times(T_{ey}+1)]$) in the 5th view are projected from DP-ROIs of the 4th view point by point. The rest of pixels are calculated as Eq. (4). Then, the extracted DP-ROIs are blocklized into MBs. Fig. 9 shows the cases that DP-ROIs of 5th view tracks from the DP-ROIs of the 4th view with different T_{ex} and T_{ey}. As T_{ex} and T_{ey} increase, the DP-ROI extraction accuracy decreases. Meanwhile, the computational complexity decreases. We can see that almost identical MB-wise DP-ROI masks can be generated when T_{ex} and T_{ey} are smaller than 3. In the extraction experiments, T_{ex} and T_{ey} are set as 1 to reduce 75% complexity and sustain the DP-ROI extraction accuracy.

Fig. 10 shows DP-ROIs of another four neighboring views (the 2nd, 3rd, 5th and 6th view) at the 10th time instants. DP-ROIs with different depth projected to neighboring views are with different relative positions which can be precisely calculated by the proposed method. Though there are some holes near the edges of objects due to the occlusion among the views, average filtering and blocklized operation can fill small holes and relieve the occlusion problem. For large holes, which only occur in the MVV with large occlusion, they can be filled by using bi-directional inter-view tracking and smoothing operation at cost of doubled computational complexity. Here, we do not adopt bi-directional inter-view tracking for low-complexity. Meanwhile, the extracted DP-ROIs are accurate enough and fulfill the accuracy requirement of MB based bit allocation optimization. In summary, the DP-ROI extraction results have shown that the proposed inter-view DP-ROI extraction method extracts satisfying DP-ROIs for neighboring views.

6.1.3. DP-ROI extraction based on temporal prediction

In the experiments, p is set as 3, which means previous DP-ROIs in three temporal preceding frames are adopted to predict the locations of the DP-ROI in current frame. Other parameters are empirically set as $\alpha_{k,t-1}=0.6$, $\alpha_{k,t-2}=0.2$, $\alpha_{k,t-3}=0.2$, $\beta_{k,t-1}=0.6$, $\beta_{k,t-2}=0.2$, $\beta_{k,t-3}=0.2$, $\xi_{k,t-i}=0.33$ and $\xi_{k,t-i}=0.33$, θ is set as 2. Fig. 11 shows DP-ROI extraction results of temporal four successive frames, the 11th to 14th frames, in the 4th view of the Ballet sequence. As for Ballet sequence, the female dancer rotates and moves fast while the man moves very slowly. Low capture frame

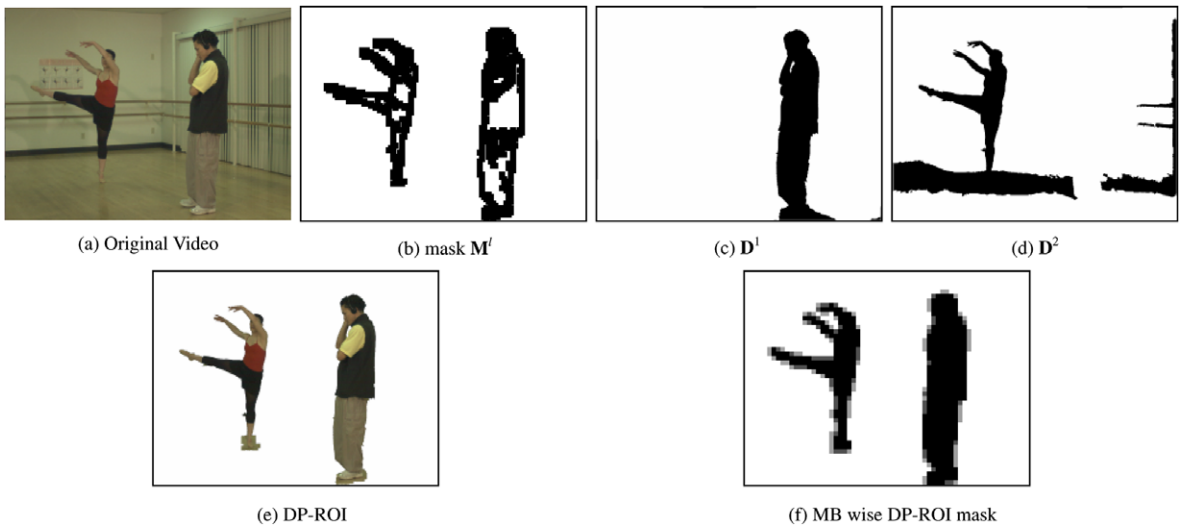

Fig. 7. Extracted results by using DBDE method (Ballet sequence, 4th view, 10th frame).

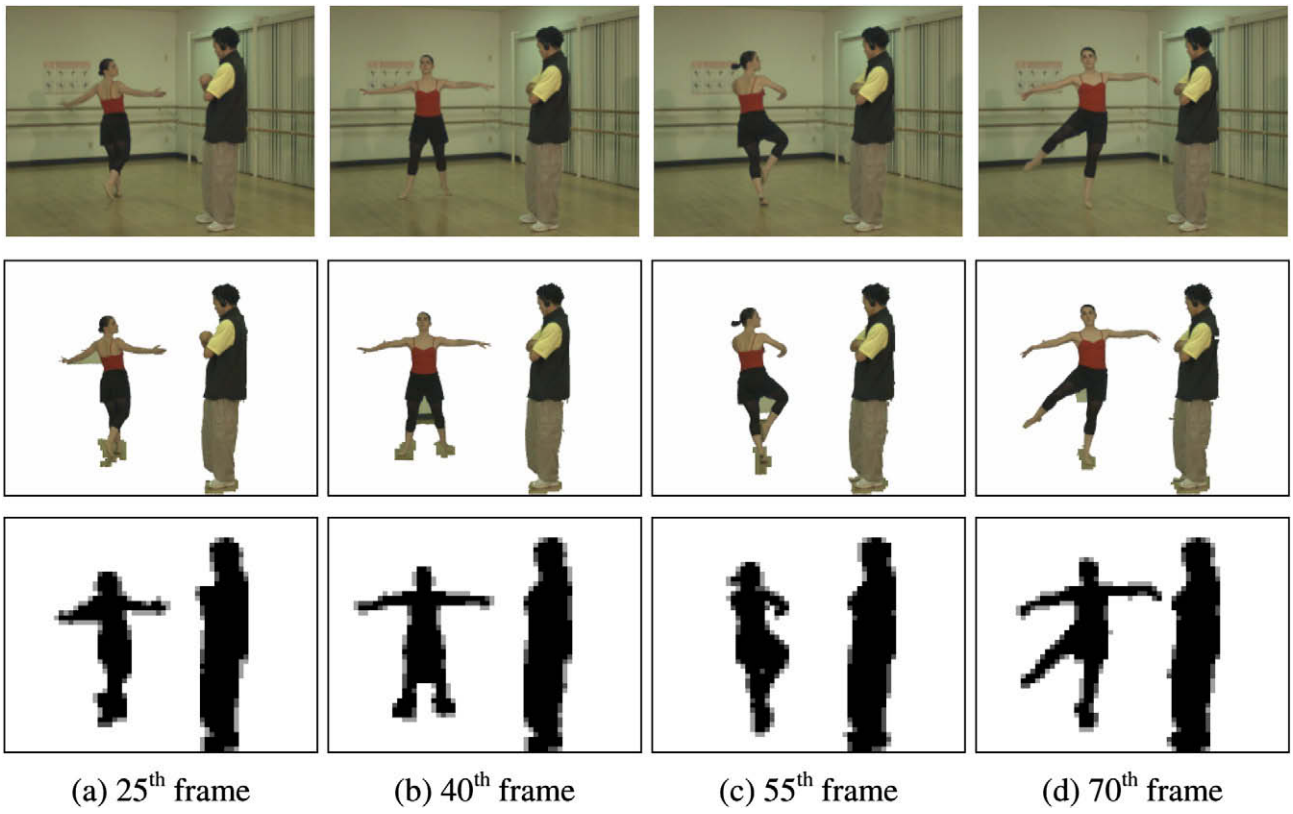

Fig. 8. Extracted results by using DBDE method (Ballet sequence, 4th view, 25th, 40th, 55th and 70th frame).

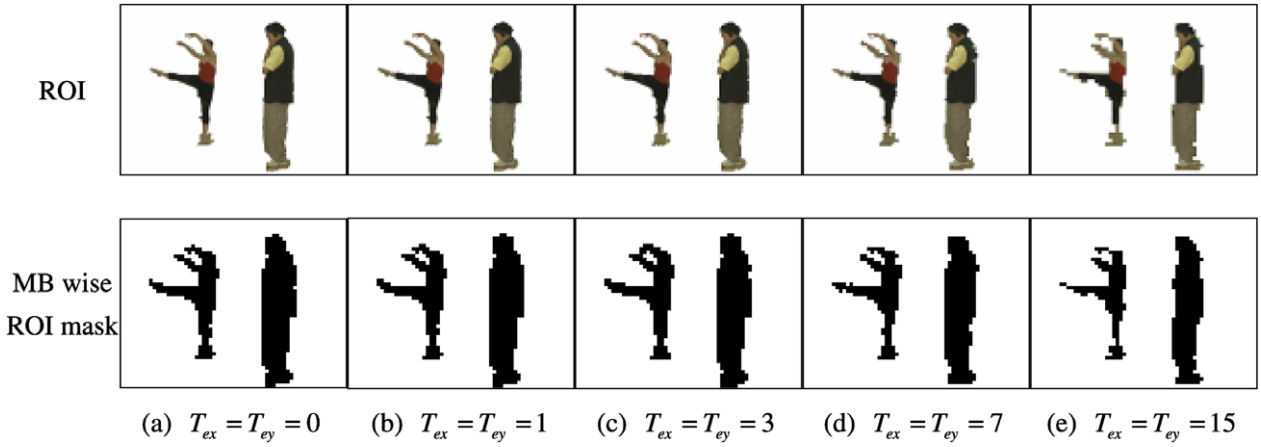

Fig. 9. Inter-view DP-ROI tracking and extraction results with different T_{ex} and T_{ey}.

rate (15 frames per second) leads to high relative motion among successive frames and it makes the temporal ROI tracking challenging. The experimental results show that, by using the proposed ROI extraction scheme, the DP-ROIs are tracked and extracted well for both slow moving object (the man) and fast moving object (the dancing girl). Furthermore, it reduces the extraction complexity and makes temporal DP-ROIs consistent by implementing DBDE within predicted windows instead of whole images. Three phases DP-ROI extraction experiments are also tested for other MVV sequences, including Breakdancers, Dog, Doorflowers and Alt Moabit sequences and similar DP-ROI results are found.

6.2. DP-ROI based bit allocation optimization for MVC

To determinate the optimal ΔQP used in the RMVC scheme, coding experiments are implemented on JMVM7.0 with MVC-HBP prediction structure, bQP and ΔQP are set as $bQP \in \{12, 17, 22, 27, 32, 37\}$ and $\Delta QP \in \{0, 2, 4, 6, 8, 10, 12\}$, respectively. η is empirically set as 3 and 6 for first and second level transitional areas. Moreover, region-selective objective image quality metrics [27] are adopted to evaluate image quality. All the encoding QPs for hierarchical B pictures with different levels are clipped from 0 to 51.

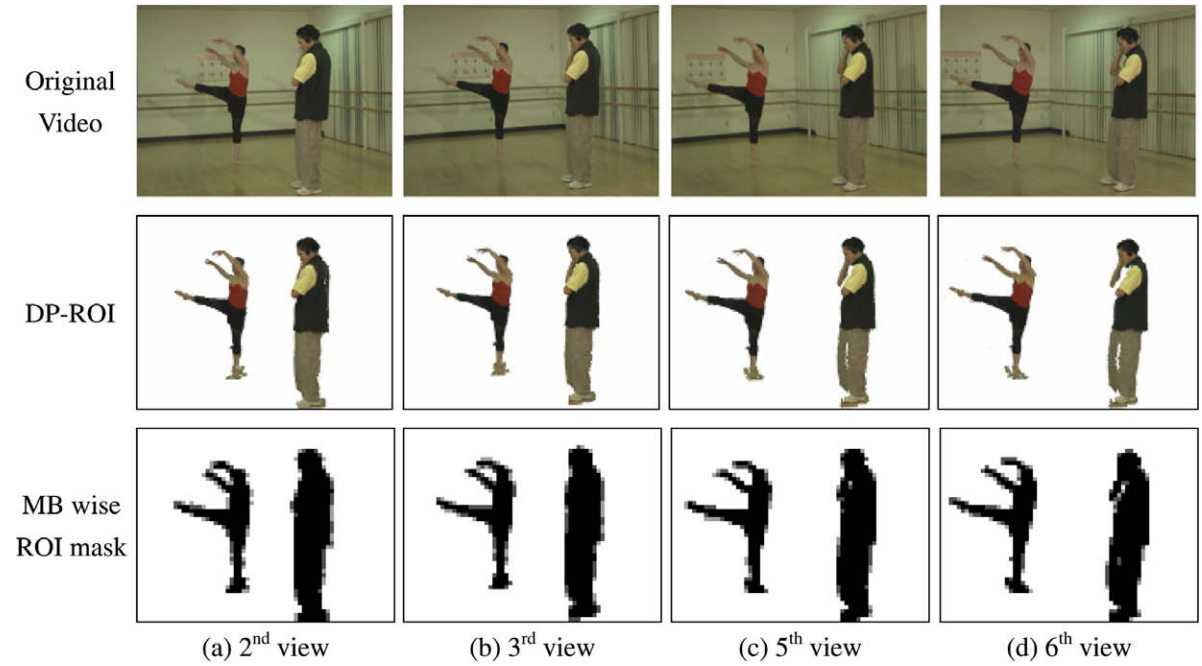

Fig. 10. Extracted ROI of the neighboring views of the 10th frame, Ballet.

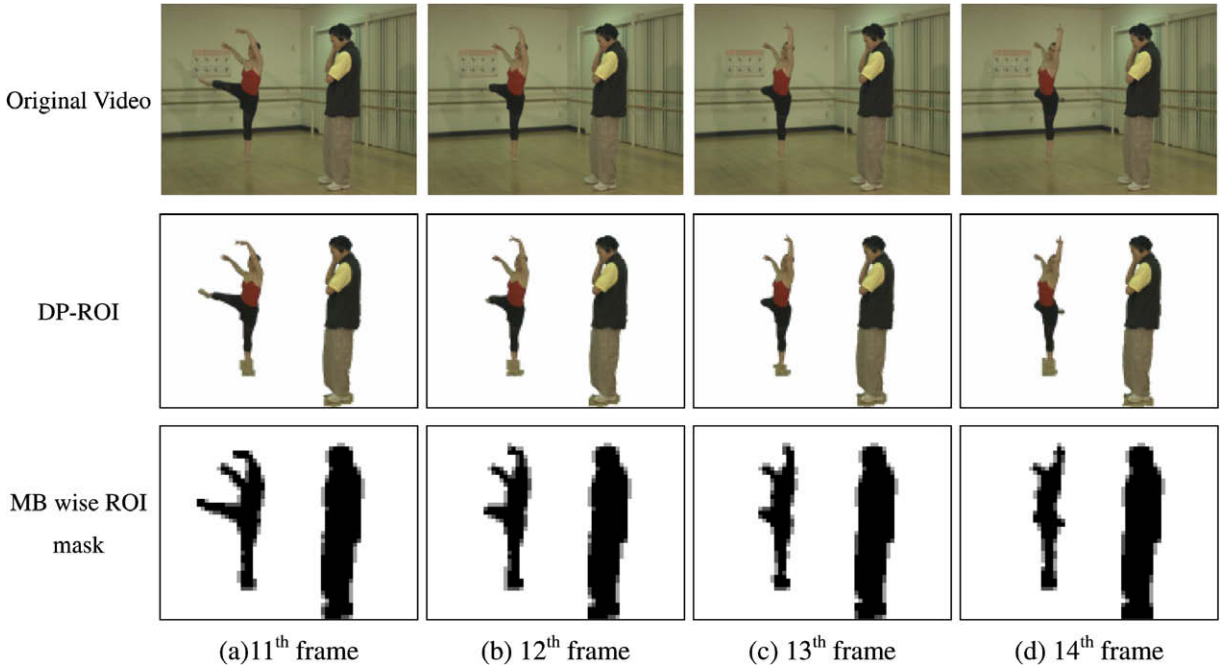

Fig. 11. DP-ROI of the time successive frames in the 4th view, Ballet.

Fig. 12 show the relation maps of $R_{BSR}(bQP, 0, \Delta QP)$ to ΔQP for Ballet and Breakdancers sequences. As shown in Fig. 12, more bit-rate is saved as ΔQP becomes larger. However, on one hand, the gradient of $R_{BSR}(bQP, 0, \Delta QP)$ decreases as ΔQP increases. The BSR satisfies the exponential decaying function described by Eq. (14) as ΔQP increases. On the other hand, the gradient and up-bound of BSR decreases as bQP increases. As bQP is larger than 27 and ΔQP is larger than 8, the BSR even decreases as ΔQP increases. It is because the encoding bits of ΔQP cannot be neglected.

Figs. 13 and 14 show the relationships between bQP and coefficients T and A. $|A|$ indicates amplitude of BSR and decreases as bQP increases. As bQP increases, the up-bound of BSR decreases to zero and almost no coding gain can be expected while bQP is bigger than 35. T indicates the ΔQP period of bit-rate saving becoming saturated. As bQP increases, the BSR's ΔQP period of being saturated is getting larger. Each point in the figures is the coefficient T fitted from each curve of Fig. 12 by using exponential function in Eq. (14). The red points are the coefficients fitted from

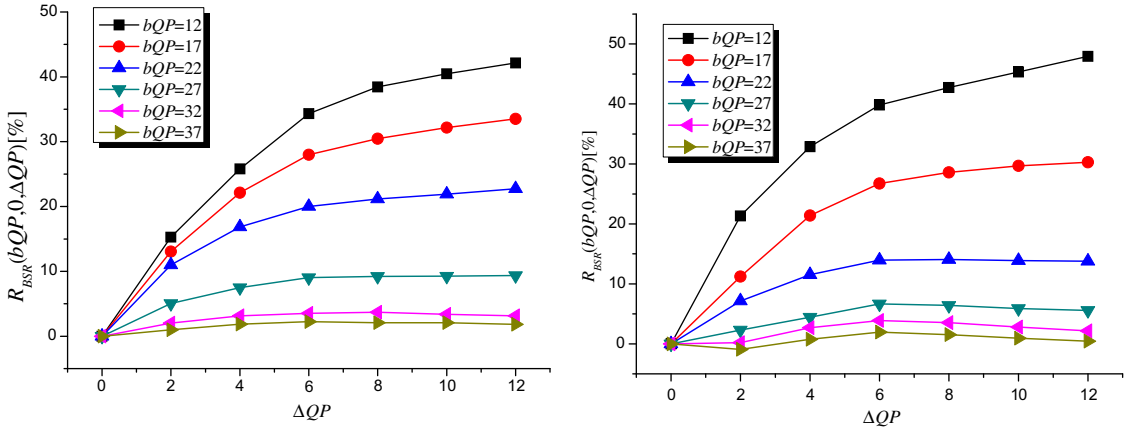

Fig. 12. The relation maps of $R_{BSR}(bQP, 0, \Delta QP)$ to ΔQP.

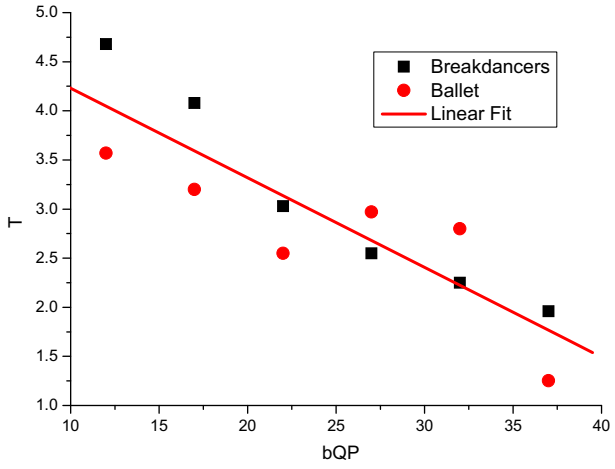

Fig. 13. Relationship between bQP and coefficient T.

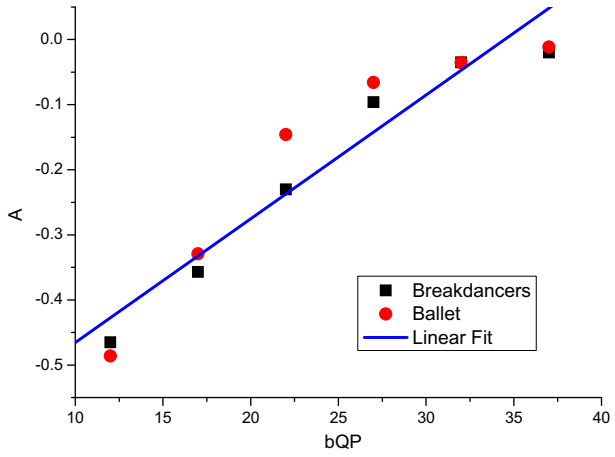

Fig. 14. Relationship between bQP and coefficient A.

Ballet sequence and the black points are fitted from Breakdancers sequence. We fit the obtained points in Figs. 13 and 14 using a linear function and obtain T and A as

$$\begin{cases} T = \alpha_1 + \beta_1 \cdot bQP \\ A = \alpha_2 + \beta_2 \cdot bQP \end{cases}, \tag{24}$$

where $\alpha_1 = 5.143$, $\beta_1 = -0.091$, $a_2 = -0.656$ and $\beta_2 = 0.019$.

In term of evaluating image quality of the reconstructed images, we use PMOS_PSNR, that is $Q^{i,j}$ in Eq. (15) is derived from Eq. (23) and Φ is PSNR_Y. Fig. 15 illustrates the relation maps of average PMOS_PSNR value to ΔQP. The y-axis is the average PMOS_PSNR and x-axis is ΔQP. Each line in the figure has the same bQP but different ΔQP. We can see that PMOS_PSNR linearly decreases as ΔQP increases for Breakdancers sequence. Additionally, the slope of image quality degradation is getting gentle as bQP increases. Similar results can also be found for Ballet sequence. Fig. 16 shows the relationship between bQP and coefficient b_1, which indicates slope of image quality degradation, $\Delta D(bQP, 0, \Delta QP)$. Each point in the figure is the coefficient b_1 fitted from$\Delta D(bQP, 0, \Delta QP)$ by using linear function in Eq. (16). The red points are the coefficients, b_1, fitted from Ballet sequence and the black points are b_1 fitted from Breakdancers sequence. We fit these b_1 points in Fig. 16 using exponential decaying function and obtain

$$b_1 = \alpha_3 + \beta_3 e^{-\frac{bQP}{\gamma_3}}, \tag{25}$$

where $\alpha_3 = -0.0876$, $\beta_3 = -31.911$ and $\gamma_3 = 2.044$.

Applying Eqs. (24) and (25) to Eq. (20), we can acquire the optimal and integer ΔQP for different μs, shown as Fig. 17. For different μs, the changing tendency of ΔQP is almost the same. However, the maximum value of ΔQP increases as μ decreases. We empirically set $\mu = 0.08$ for scaling BSR and image quality degradation in the same scale. Then, the final optimal ΔQPs are obtained while bQP ranges 12–33, as shown in Fig. 17. For low bQP, e.g. $bQP < 15$, significant bit-rate saving can be saved by choosing large ΔQP. However, the image quality is also degraded significantly. Thus, ΔQP is reasonable to be smaller than 8 so that a wise tradeoff between BSR and image quality degradation can be achieved. As for large bQP, e.g. $bQP > 33$, most MBs in background regions are already coded with DIRECT mode, in which no residuals are coded, and almost coding gain can be expected by enlarging ΔQP. In some cases, the BSR decreases as ΔQP increases because the encoding bits of ΔQP increase and they are not neglectable. Furthermore, larger ΔQP causes larger quantization error. Therefore, it is reasonable to limit ΔQP within the range from 0 to 2 at low bit-rate (large bQP).

6.3. DP-ROI based multiview video coding

MVC experiments are implemented on JMVM 7.0 reference software with the MVV sequences, Ballet, Breakdaners, Dog, Doorflowers and Alt Moabit, to evaluate the coding performance of the RMVC scheme. The MVC-HBP prediction structure is adopted for MVC simulation. GOP Length is 15, 8 views. There are three kinds

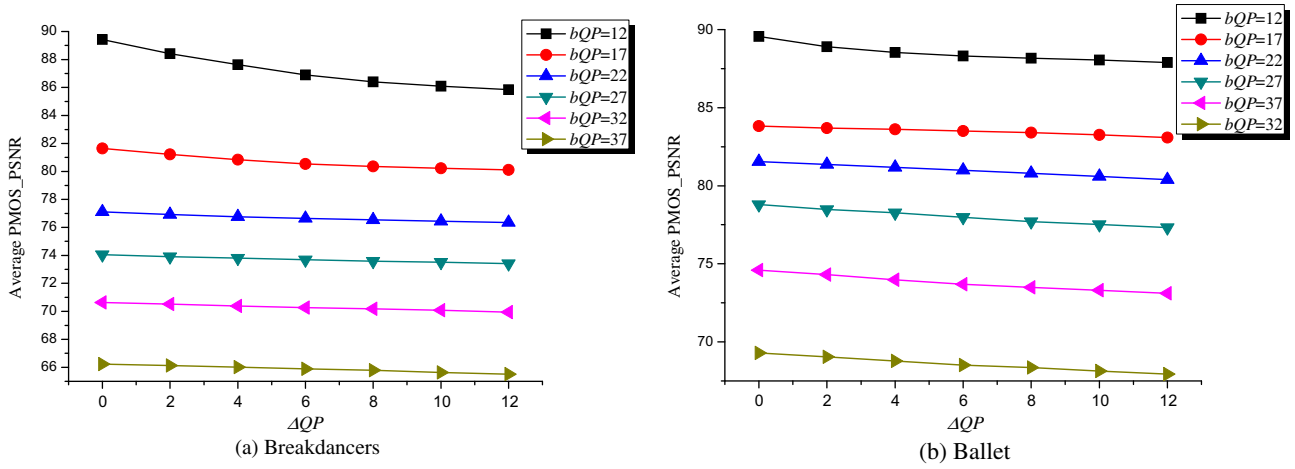

Fig. 15. The relation maps of average PMOS_PSNR value to ΔQP.

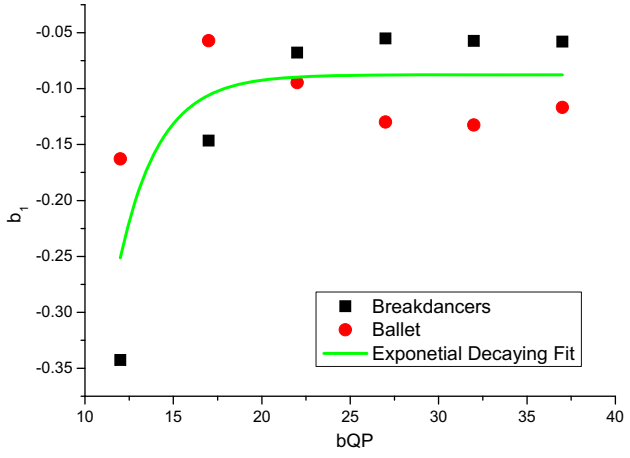

Fig. 16. Relation map of bQP and b_1.

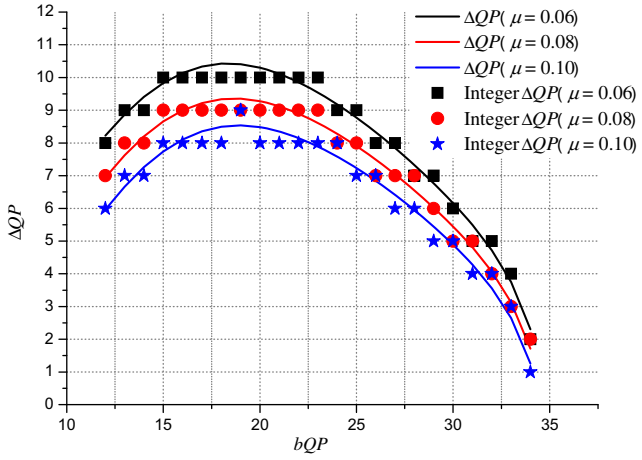

Fig. 17. Optimal and integer ΔQP for different μs.

of picture in MVC-HBP prediction structure: intra coded picture (I-picture), inter-predicted picture (P-picture) and hierarchical bi-directional predicted picture (B-picture). Here, we define two RMVC encoding modes denoted by "RMVC_IBP" and "RMVC_BP". In RMVC_IBP, all the three kinds of frames except the first I-picture are coded with the proposed RMVC scheme with the bit allocation optimization. In RMVC_BP, all B and P-pictures are coded with RMVC scheme with bit allocation optimization and I-pictures are coded with original MVC scheme without bit allocation optimization. The bQP are set as 17, 22, 27 or 32, and the QPs of background

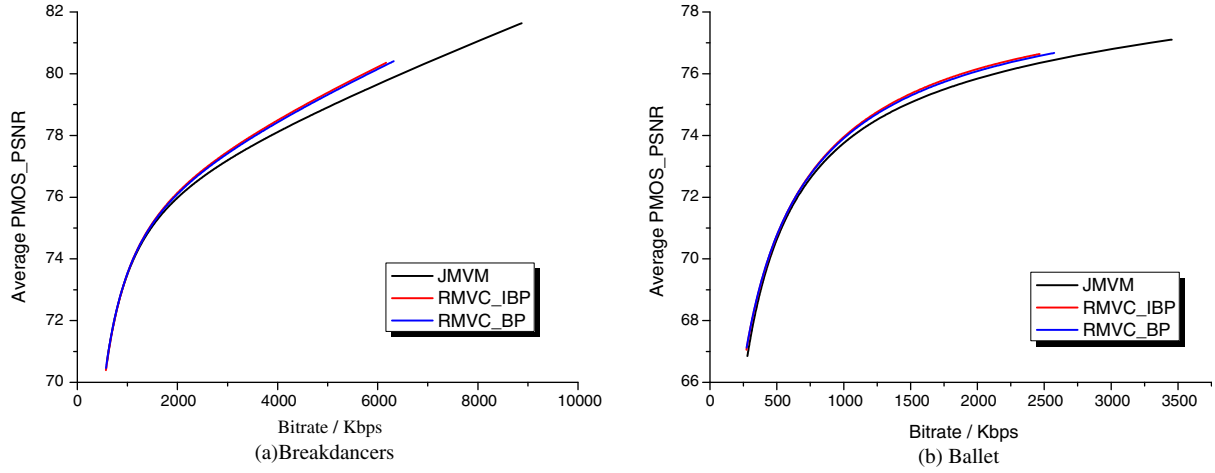

Fig. 18. Rate-distortion performances comparison in terms of PMOS_PSNR.

and ROI regions are set according to Eq. (9) and Fig. 17. Regional selective objective quality metrics in Section 5, PMOS_PSNR and PMOS_SSIM, are adopted to evaluate image quality of reconstructed frames.

Figs. 18 and 19 show coding performance comparison between the proposed RMVC and JMVM. Curves in the Figs. 18 and 19 are fitted with the algorithm provided in [30]. For Breakdancers sequence, the encoding performances of RMVC_IBP and RMVC_BP,

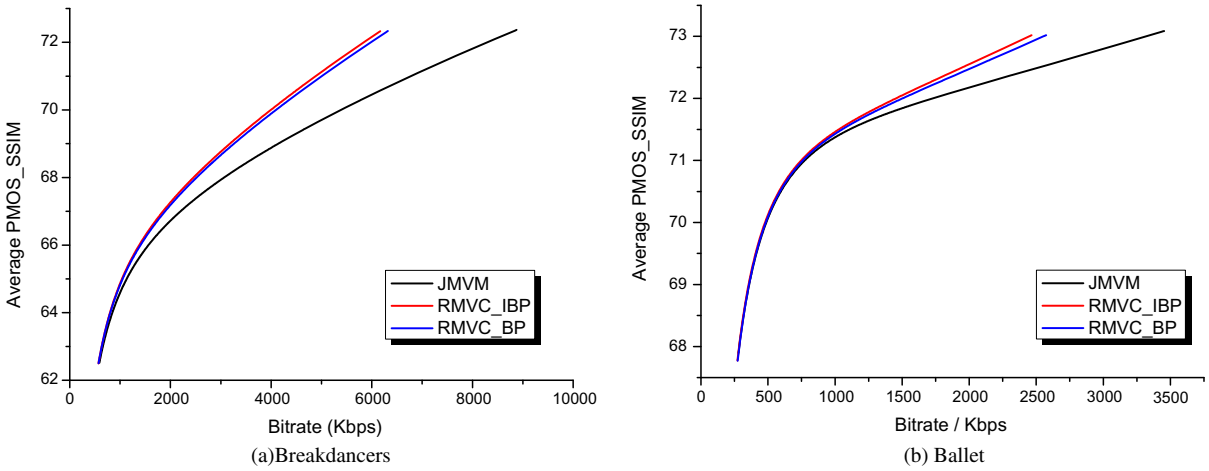

Fig. 19. Rate-distortion performances comparison in terms of PMOS_SSIM.

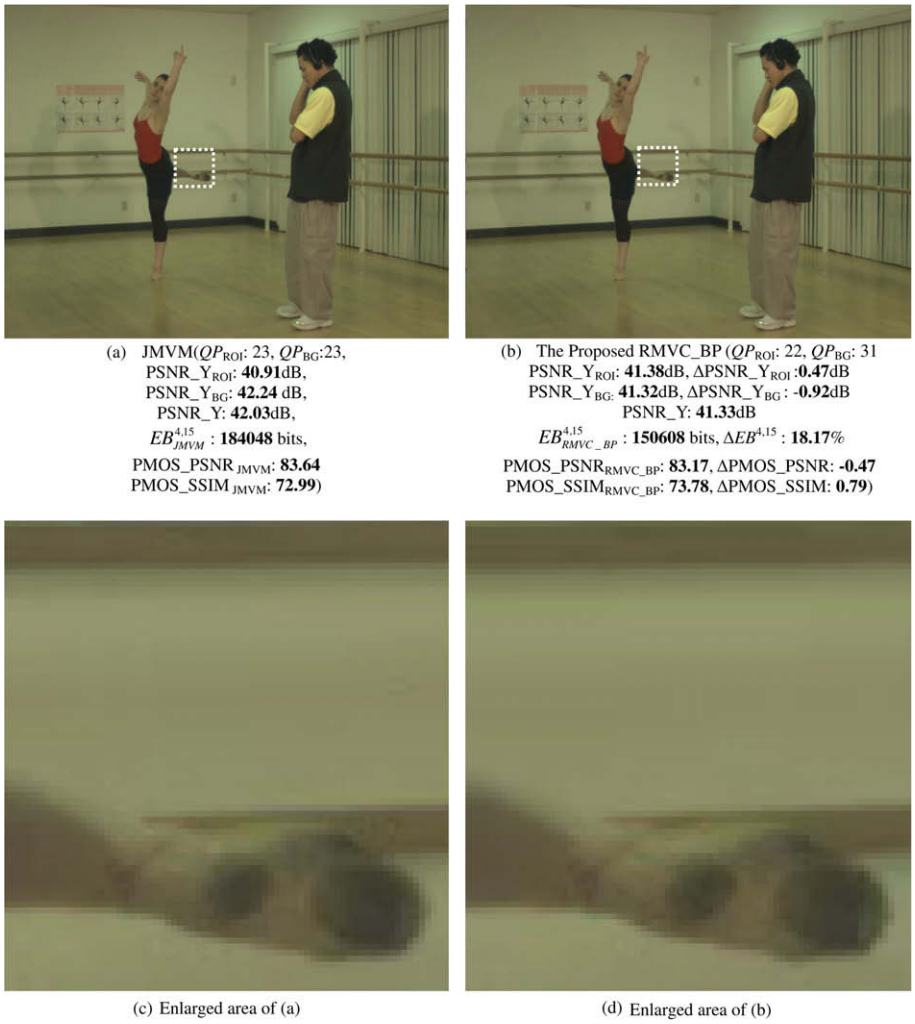

Fig. 20. Subjective and objective quality comparison of the reconstructed images (Ballet, 4th view, 15th frame).

evaluated by PMOS_PSNR, are better than JMVM at high bit-rate. As the reconstructed images are evaluated with PMOS_SSIM, significant coding gain is achieved by RMVC_IBP and RMVC_BP. In other words, up to 30% bit-rate can be saved at high rate while comparing with JMVM. Meanwhile, the average coding performance of RMVC_IBP is a little superior to that of RMVC_BP. Similar results are obtained for other sequences. For Ballet, Dog, Doorflowers and Alt Moabit sequences, while measured with PMOS_PSNR, RMVC outperforms JMVM at high bit-rate and maintains the same performance at low bit-rate. Additionally, as for the cases that distortion are measured by PMOS_SSIM, we can see that significant coding gains, over 20% bit-rate saving, are achieved by the proposed RMVC. Bit-rate saving mainly comes from moving area in background. However, in low bit-rate coding experiments, the most MBs in background regions are already coded with DIRECT mode or quantized with large *QP*s. In that case, enlarging *QP*, i.e. introducing larger ΔQP, can scarcely reduce encoding bits of quantized coefficients but increase number of bits of encoding ΔQP.

Figs. 20 and 21 show image quality comparison between the reconstructed images coded by RMVC_BP and original JMVM. It shows the reconstructed results of the 15th frame of the 4th view of the test MVV sequences. Encoding bits and five image quality indices including $PSNR\_Y_{ROI}$, $PSNR\_Y_{BG}$, PSNR_Y, PMOS_SSIM and PMOS_PSNR are compared for the coded picture of each sequence. $PSNR\_Y_{ROI}$, $PSNR\_Y_{BG}$ and PSNR_Y indicate the PSNR of illumination component in DP-ROIs, background and the entire picture, respectively. PMOS_PSNR and PMOS_SSIM represent the PMOS of PSNR_Y and SSIM, respectively. In addition, the difference of the image quality indices are also given and they are computed as

$$\begin{cases} \Delta\Theta = \Theta_{\text{Proposed}} - \Theta_{\text{JMVM}} \\ \Delta EB^{i,j}\ [\%] = \frac{EB^{i,j}_{\text{JMVM}} - EB^{i,j}_{\text{Proposed}}}{EB^{i,j}_{\text{JMVM}}} \times 100\ [\%] \end{cases}, \tag{26}$$

where $\Theta \in \{PSNR\_Y_{ROI}, PSNR\_Y_{BG}, PMOS\_PSNR, PMOS\_SSIM\}$, $\Delta EB^{i,j}$ is the BSR for the proposed RMVC scheme while encoding the picture at (i,j) position of a GOP. $EB^{i,j}_{\text{JMVM}}$ and $EB^{i,j}_{\text{RMVC\_BP}}$ denote encoding bits of the coded pictures by using JMVM and the proposed RMVC_BP method, respectively.

Because people pay less attention to background regions and more attention to DP-ROIs, HVS is less perceptible to distortion in background regions than that of DP-ROIs. That means people require high image quality in DP-ROIs. However, for Ballet sequence coded by JMVM, the PSNR of ROI and background are 40.91 and 42.24 dB, i.e. $PSNR_{ROI} < PSNR_{BG}$, which is not coincident with the requirement of HVS. With regard to the proposed RMVC_BP, the PSNR of ROI and background are 41.38 and 41.32 dB. $\Delta PSNR\_Y_{ROI}$ is 0.47 dB while $\Delta PSNR\_Y_{BG}$ is −0.92 dB. It means that the proposed RMVC_BP improves image quality of DP-ROIs up to 0.47dB; meanwhile, RMVC_BP allocates fewer bits on background regions for higher compression ratio at the cost of its $PSNR\_Y_{BG}$. Additionally, the image quality of DP-ROI is getting better than that of background region, i.e. $PSNR\_Y_{ROI} > PSNR\_Y_{BG}$, which meets the requirements of HVS. The quality of the reconstructed images is improved. As the reconstructed image is evaluated by the region-selective metrics, ΔPMOS_SSIM is 0.79 while ΔPMOS_PSNR is −0.47. It means a tiny and imperceptible difference between the image quality of reconstructed images coded by RMVC_BP and JMVM. However, the important and interesting thing is that $\Delta EB^{4,15}$ is 18.17%. That is

Fig. 21. Subjective and objective quality comparison of the reconstructed images (Breakdancers, 4th view, 15th frame).

18.17% fewer bits are utilized to encode the picture while comparing with JMVM. Fig. 20(c and d) shows the enlarged the white rectangle located area of Fig. 20(a and b). They show the subjective quality of parts of transitional area between DP-ROI and background regions. There is no block artifact and other artifacts in the transitional area by using the proposed bit allocation optimization comparing with JMVM.

For Breakdancers sequence, $\Delta PSNR\_Y_{ROI}$ is 0.68 dB and $\Delta PSNR\_Y_{BG}$ is −1.04 dB while $\Delta EB^{4,15}$ is 14.22%. It shows that 14.22% bit-rate is saved and image quality of DP-ROI is significantly improved at the cost of image quality degradation in the background regions only. PMOS_PSNR shows there is imperceptible difference between the reconstructed images of Breakdancers coded by RMVC_BP and JMVM. The other region-selective image quality metric, PMOS_SSIM, shows that the reconstructed image quality of Breakdancers coded by RMVC_BP is better than coded by JMVM. Also, up to 14.22% BSR is achieved. From the subjective image quality comparison of the reconstructed frames, it can be seen that the image quality coded by RMVC_BP is almost the same as JMVM. There is also no block artifact and other artifacts in the transitional area for Breakdaners sequence, as shown in Fig. 21(c and d). Similar results can be found for Doorflowers, Alt Moabit and Dog sequences. Up to 19.44–23.32% bit-rate is saved and quality of DP-ROI are improved 0.16–0.56 dB at the cost of image quality degradation in background.

In summary, the proposed RMVC scheme can achieve significant BSR, up to 14.22–23.32%, while the image quality of DP-ROI are improved 0.16–0.68 dB at the cost of the image quality degradation in background regions. Region-selective image quality metrics indicate that the proposed RMVC scheme can achieve significant BSR with imperceptible image quality degradation.

7. Conclusions

A framework of Depth Perceptual Region-Of-Interest (DP-ROI) based Multiview Video Coding (RMVC) has been proposed to improve compression efficiency significantly by properly segmenting the multiview video into different MB-wise DP-ROI and encoding them separately. Novel low-complexity DP-ROI extraction algorithms have also been proposed in this paper. The proposed depth based DP-ROI extraction algorithms maintains both low-complexity and high accuracy. Additionally, according to the extracted DP-ROI, a DP-ROI based bit allocation optimization algorithm is proposed for multiview video coding where inter-view and temporal predictions are jointly utilized for high compression efficiency. It is able to allocate more bits on DP-ROIs for maintaining high image quality and fewer bits on background and transitional regions for achieving high compression ratio. The proposed RMVC scheme achieves significant coding gain at the high rate while comparing with the joint multiview video model. To be specific, up to 14.22–23.32% bit-rate are saved while 0.16–0.68 dB coding gains are achieved in DP-ROIs at the cost of the image quality degradation in background.

Acknowledgments

Thanks for Microsoft Research, HHI and Nagoya University kindly providing us multiview video sequences. This work was supported by the Natural Science Foundation of China (Grants 60872094, 60832003), the Program for New Century Excellent Talents in University (NCET-06-0537) and the Innovation Fund Project for Graduate Student of Zhejiang Province (YK2008044). It was also sponsored by K.C. Wong Magna Fund in Ningbo University.

References

[1] M. Tanimoto, Overview of free viewpoint television, Signal Processing: Image Communication 21 (6) (2006) 454–461.
[2] A. Smolic, P. Kauff, Interactive 3-D Video representation and coding technologies, in: Proceedings of the IEEE, vol. 93 (No.1), Jan 2005, pp. 98–110.
[3] Y.S. Ho, K.J. Oh, Overview of multiview video coding, in: Proceeding of International Workshop on Systems, Signals and Image Processing (IWSSIP07), Jun 2007, pp. 5–12.
[4] A. Smolic, K. Müller, P. Merkle et al., Multiview video plus depth (MVD) format for advanced 3D video systems, MPEG and ITU-T SG16 Q.6, JVT-W100, San Jose, USA, April 2007.
[5] K. Müller, P. Merkle, T. Wiegand, Compressing time-varying visual content – generating 3-D scene representations, IEEE Signal Processing Magazine (2007) 58–65.
[6] A. Vetro, S. Yea, A. Smolic, Towards a 3D video format for auto-stereoscopic displays, in: SPIE Conference on Applications of Digital Image Processing XXXI, vol. 7073, Sep 2008, pp. 70730F–70730F-10.
[7] 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, Special Issue on 3D Video and TV, vol. 22 (No. 2), Feb 2007, pp. 217–234.
[8] Survey of Algorithms used for Multiview video coding (MVC), ISO/IEC JTC1/SC29/WG11, N6909, Hong Kong, China, Jan 2005.
[9] S. Oka, T. Endo, T. Fujii, Dynamic Ray-Space Coding using Multi-directional Picture, IEICE Technical Report, Dec 2004, pp.15–20.
[10] M. Kitahara, H. Kimata, S. Shimizu, et al., Multiview video coding using view interpolation and reference picture selection, in: Proceedings of IEEE Int'l Conf. Multimedia & Expo (ICME), Toronto, Canada, Jul 2006, pp. 97–100.
[11] S. Yea, A. Vetro, View synthesis prediction for multiview video coding, Image Communication 24 (1–2) (2009) 89–100.
[12] Y. Zhang, G.Y. Jiang, M. Yu, Y.S. Ho, Adaptive multiview video coding scheme based on spatiotemporal correlation analyses, ETRI Journal 31 (2) (2009) 151–161.
[13] P. Merkle, A. Smolic, K. Müller, T. Wiegand, Efficient prediction structures for multiview video coding, IEEE Transactions on Circuits and Systems for Video Technology 17 (11) (2007) 1461–1473.
[14] Z.K. Lu, W. Li, X.K. Yang, et al., Modeling visual attentions modulatory after effects on visual sensitivity and quality evaluation, IEEE Transactions on Image Processing 14 (11) (2005) 1928–1942.
[15] J.-R. Ohm, Encoding and reconstruction of multiview video objects, IEEE Signal Processing Magazine 16 (3) (1999) 47–54.
[16] L. Yang, G.L. Zheng, C.S. Yeng, Region-of-interest based resource allocation for conversational video communication of H.264/AVC, IEEE Transactions on Circuits and Systems for Video Technology 18 (1) (2008) 134–139.
[17] E. Kaminsky, D. Grois, O. Hadar, Dynamic computational complexity and bit allocation for optimizing H.264/AVC video compression, Journal of Visual Communication and Image Representation 19 (1) (2008) 56–74.
[18] Y. Wang, K.F. Loe, T. Tan, J.K. Wu, Spatiotemporal video segmentation based on graphical models, IEEE Transactions on Image Processing 14 (7) (2005) 937–947.
[19] J.W. Han, K.N. Ngan, M.J. Li, H.J. Zhang, Unsupervised extraction of visual attention objects in color images, IEEE Transactions on Circuits and Systems for Video Technology 16 (1) (2006) 141–145.
[20] E. Izquierdo M., Disparity/segmentation analysis: matching with an adaptive window and depth-driven segmentation, in: IEEE Transactions on Circuits and Systems for Video Technology, vol. 9 (No. 4), Jun 1999, pp. 589–607.
[21] A. Marugame, A. Yamada, M. Ohta, Focused object extraction with multiple cameras, IEEE Transactions on Circuits and Systems for Video Technology 10 (4) (2000) 530–540.
[22] A. Yilmaz, O. Javed, M. Shah, Object tracking: a survey, ACM Journal of Computing Surveys 38 (4) (2006) (Article 13).
[23] C. Stauffer, W. Grimson, Learning patterns of activity using real time tracking, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (8) (2000) 747–767.
[24] A. Yilmaz, X. Li, M. Shah, Contour-based object tracking with occlusion handling in video acquired using mobile cameras, IEEE Transactions on Pattern Analysis and Machine Intelligence 26 (11) (2004) 1531–1536.
[25] M.C. Chi, M.J. Chen, C.H. Yeh, J.A. Jhu, Region-of-interest video coding based on rate and distortion variations for H.263+, Signal Processing: Image Communication 23 (2) (2008) 127–142.
[26] K. Takagi, Y. Takishima, Y. Nakajima, A study on rate distortion optimization scheme for JVT coder, Visual Communications and Image Processing (VCIP) 5150 (2003) 914–923.
[27] U. Engelke, V.X. Nguyen, H.J. Zepernick, Regional attention to structural degradations for perceptual image quality metric design, in: International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Las Vegas, Nevada, USA, Mar 2008, pp. 869–873.
[28] Z. Wang et al., Image quality assessment: from error visibility to structural similarity, IEEE Transactions on Image Processing 13 (2004) 600–612.
[29] M. Yu, Z.J. Peng, G.Y. Jiang, "Statistical Analysis of Macroblock Mode Selection in JMVM", ISO/IEC JTC1/SC29/WG11 and ITU-T SG16 Q.6, JVT-Y026, Shenzhen, China, Oct 2007.
[30] G. Bjontegaard, Calculation of average PSNR differences between RD-curves, ITU-T VCEG. VCEG-M33, Apr 2001.