Signal Processing: *Image Communication* 24 (2009) 666–681

Contents lists available at ScienceDirect

Signal Processing: *Image Communication*

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

Joint video/depth rate allocation for 3D video coding based on view synthesis distortion model

Yanwei Liu [a,b], Qingming Huang [a,b,*], Siwei Ma [c], Debin Zhao [d], Wen Gao [c]

[a] *Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China*
[b] *Graduate University of Chinese Academy of Sciences, Beijing, China*
[c] *Institute of Digital Media, Peking University, Beijing, China*
[d] *Department of Computer Science, Harbin Institute of Technology, Harbin, China*

ARTICLE INFO

Article history:
Received 16 September 2008
Received in revised form
5 June 2009
Accepted 10 June 2009

Keywords:
View synthesis
Distortion model
Rate allocation
3D video coding

ABSTRACT

Joint video/depth rate allocation is an important optimization problem in 3D video coding. To address this problem, this paper proposes a distortion model to evaluate the synthesized view without access to the captured original view. The proposed distortion model is an additive model that accounts for the video-coding-induced distortion and the depth-quantization-induced distortion, as well as the inherent geometry distortion. Depth-quantization-induced distortion not only considers the warping error distortion, which is described by a piecewise linear model with the video power spectral property, but also takes into account the warping error correlation distortion between two sources reference views. Geometry distortion is approximated from that of the adjacent view synthesis. Based on the proposed distortion model, a joint rate allocation method is proposed to seek the optimal trade-off between video bit-rate and depth bit-rate for maximizing the view synthesis quality. Experimental results show that the proposed distortion model is capable of approximately estimating the actual distortion for the synthesized view, and that the proposed rate allocation method can almost achieve the identical rate allocation performance as the full-search method at less computational cost. Moreover, the proposed rate allocation method consumes less computational cost than the hierarchical-search method at high bit-rates while providing almost the equivalent rate allocation performance.

1. Introduction

3D video is an emerging new media for rendering dynamic real-world scenes. Compared with traditional 2D video, 3D video is the natural extension in the spatio-temporal domain as it provides the depth impression of the observed scenery. Besides the 3D sensation, 3D video also allows for an interactive selection of viewpoint and view direction within the captured range [1]. With these features, 3D video will revolutionize visual media by enabling 3D-TV and free viewpoint TV (FTV) for 3D display applications [2].

The acquisition of 3D video typically involves recording the synchronous multiview video streams. A variety of multiview video processing technologies from signal processing, computer vision and computer graphics are used in 3D video processing. 3D video lives on the convergence of these disciplines so that there are diverse and various scene representations for it [3]. In [1], the authors utilize point samples to represent the 3D information of scene. Theobalt et al. [4] propose a model-based representation to facilitate the processing for acquisition and rendering of 3D video. Similarly, a polyhedral visual hulls representation [5] provides a

* Corresponding author at: Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China.
E-mail addresses: ywliu@jdl.ac.cn (Y. Liu), qmhuang@jdl.ac.cn (Q. Huang).

0923-5965/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.image.2009.06.002

view-independent rendering with epipolar geometry. These representations can fulfill the requirements of 3D video and achieve a good rendering quality. However, they cannot perfectly provide the backward compatibility with the existent processing chain of traditional 2D video.

Recently, the 2D video plus depth format has emerged as an efficient data representation for 3D video because it allows the rendering of new views with very low processing costs. With accurate multiview geometry, 2D video and the associated per-pixel depth can provide high-quality rendering of new views [6,7]. Though the generation of depth maps is somewhat complicated, the video plus depth format is compatible with the current legacy devices and existent delivery infrastructure of 2D video. Moreover, this representation is apt to be coded using the popular standards, such as MPEG-2 and H.264/AVC [8]. To spur stereoscopic applications, the ISO/IEC 23002-3 standard had specified the video plus depth format [9]. Recently, MPEG group has started the exploration work on depth estimation and view synthesis for developing the 3D video standard [10].

The *N*-video plus *N*-depth is the proper delivery format for 3D video [11]. To reduce the great bandwidth consumption in 3D video communication, a lot of multi-view video coding (MVC) algorithms are proposed to exploit the inter-view redundancy between the multiple videos [12–18]. Compared with independent coding of views with the same video quality, these algorithms can save 20–50% of the overall bit-rate by [12]. The depth map can be separately encoded with conventional coding techniques, such as H.264/AVC, or jointly coded with MVC [19,20], and it also can be coded with some new coding methods that consider the special characteristics of depth maps, such as the method by exploiting depth smooth properties [21] and the method by jointly optimizing depth estimation and depth coding in the wavelet domain [22].

The compression of multiview video and depth has a great effect on the view synthesis quality [19,23]. The video and depth compression with different bit-rate overheads can lead to different synthesis qualities of the virtual view. For video/depth rate allocation, Daribo et al. [24] propose a rate-distortion optimized bit allocation strategy. In [25], Morvan et al. propose a joint bit allocation method for multiview video and depth coding to guarantee the optimal view synthesis. Given the bit-rate constraint, the optimal trade-off between the video bit-rate and depth bit-rate is exhaustively searched by full-search and hierarchical-search methods. To measure the quality of the synthesized view, Morvan et al. assume that the captured original view at the synthesis position exists and the view synthesis quality is evaluated by mean-squared error (MSE) between the synthesized image and the original image. Such a full-reference assessment [26] can accurately evaluate the quality of view synthesis. However, this assumption is not appropriate for many practical 3D video applications because the original view at the synthesized position is not always available. In practical 3D video systems, due to the enormous information involved, it is impossible to capture a scene's entire information with cameras. Usually, the scenes are captured with a wide-baseline camera setup. Especially in free-viewpoint video systems, for smooth browsing in view dimension, most of the views need to be interpolated by adjacent views.

To solve the no-reference evaluation problem for the synthesized view in joint video/depth rate allocation, this paper proposes a distortion model to characterize the view synthesis quality without requiring the original reference image. Three kinds of distortions are mainly concerned in the proposed additive view synthesis distortion model: the video-coding-induced distortion, the depth-quantization-induced distortion, and the inherent geometry distortion. Based on the analysis of the complex behavior of depth-quantization-induced warping errors, depth-quantization-induced warping error distortion is characterized by a piecewise linear model according to the video power spectra property. According to the multiview geometry, the geometry distortion is approximated from that of the adjacent view synthesis. Based on the proposed distortion model, we further propose a joint rate allocation method to find the optimal quantization setting for video and depth coding. The proposed rate allocation method first considers the relationship among the video's rate-distortion, the depth's rate-distortion and the virtual view synthesis quality, and then utilizes the view synthesis distortion model to optimize the video/depth rate allocation.

The rest of the paper is organized as follows. The distortion model is derived in Section 2. In this section, we first describe the depth-image-based view synthesis and then derive the distortion model for the synthesized view. In the model derivation, we emphasize on analyzing how the depth loss leads to the view synthesis distortion. In Section 3, based on the derived distortion model, the proposed joint video/depth rate allocation algorithm is described in detail. Section 4 presents the experimental results. Finally, the conclusion is provided in Section 5.

2. View synthesis distortion model

2.1. Depth-image-based view synthesis

Depth-image-based view synthesis is usually performed as 3D warping [27]. The virtual view synthesis at the middle of two captured views is shown in Fig. 1. According to the camera parameters, homography matrices from the adjacent source reference view to the virtual view can be obtained at different depth values and further the pixels in virtual view image can be warped from those in adjacent views. The homography matrix is usually computed by matching the correspondence points between two view images integrated with the pinhole camera model [28]. In Fig. 1, $\mathbf{H}_{A\to V}[Z]$ and $\mathbf{H}_{B\to V}[Z]$ are homography matrices at depth Z from view A and from view B to the virtual view, respectively. The relation amongst (u_V, v_V), (u_A, v_A) and (u_B, v_B) is described by

$$(u_V, v_V, 1) = \mathbf{H}_{A\to V}[Z](u_A, v_A, 1)^T = \mathbf{H}_{B\to V}[Z](u_B, v_B, 1)^T. \quad (1)$$

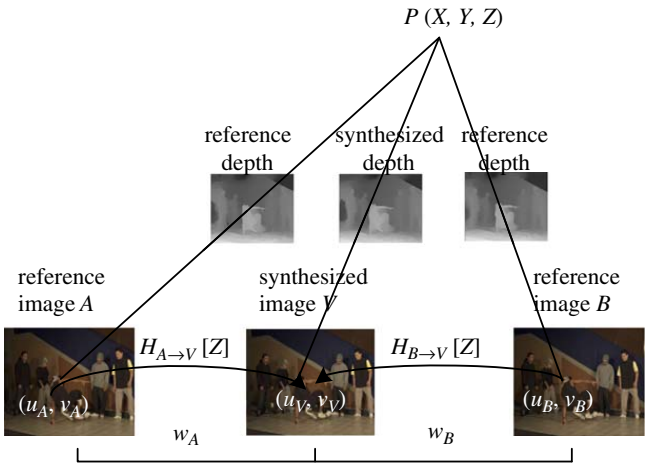

Fig. 1. Depth-image-based view synthesis.

Taking into account the occlusion effect, the virtual view synthesis is expressed as

$$I_V(u_V,v_V)=\begin{cases} w_A I_A(u_A,v_A)+w_B I_B(u_B,v_B), \\ \quad \text{if } (u_V,v_V) \text{ is both visible in view } A \text{ and } B \\ I_A(u_A,v_A), \\ \quad \text{if } (u_V,v_V) \text{ is only visible in view } A \\ I_B(u_B,v_B), \\ \quad \text{if } (u_V,v_V) \text{ is only visible in view } B \\ 0, \\ \quad \text{otherwise} \end{cases}, \tag{2}$$

where $I_V(u_V, v_V)$, $I_A(u_A, v_A)$ and $I_B(u_B, v_B)$ are the pixel values of the matching points in different views, and w_A and w_B are the distance-dependent blending weights with $w_A+w_B=1$. In one adjacent view, if there exists one point corresponding to (u_V, v_V) and the depth value at this point also exists, (u_V, v_V) is taken as visible in this adjacent view.

Because of the occlusion and pixel mapping uncertainty, some pixels in the virtual view have no matching points in the sources view A and B. They will be inpainted by the adjacent pixels that have been warped from the sources reference views. In the pixel mapping, the mapped pixel sometimes does not locate at an integer position, and it will be bilinearly interpolated or rounded to the nearest integer position.

2.2. Modeling of the virtual view synthesis distortion

In 3D video applications, the virtual view is generally synthesized by compressed video and depth. The compression brings the corresponding view synthesis distortion. Based on the specific analysis on the view synthesis behavior, we derive the view synthesis distortion model as follows.

Assume that S_V denotes the captured original image at the synthesis position and $\hat{S}_V$ represents the image synthesized by the compressed images of adjacent views. According to multiview geometry, the ideal S_V should include partial signal of the left view S_A, partial signal of the right view S_B and partial signal in the occluded areas that cannot be observed from the adjacent views. The signal in the occluded area is expressed as S_O. Let $\bar{S}_V$ represent the image synthesized by the original images of adjacent views. In ideal cases, S_V should be equal to $\bar{S}_V$. At the present time, however, the geometry information is not very perfect due to the noise effects when capturing video and computing depth. Hence there exist some differences between S_V and $\bar{S}_V$.

Let $\hat{S}_A$ and $\hat{S}_B$ be the compressed images of the left view and the right view, respectively. In the view synthesis, the occluded area is inpainted by adjacent pixels and the synthesized signal in the occluded area is $\hat{S}_O$. Therefore, the distortion for the synthesized virtual view, in terms of MSE, can be expressed as

$$\begin{aligned} D_V &= E\{[S_V-\hat{S}_V]^2\} \\ &= E\{[S_V-\bar{S}_V]^2\}+E\{[\bar{S}_V-\hat{S}_V]^2\} \\ &\quad +2E\{[S_V-\bar{S}_V][\bar{S}_V-\hat{S}_V]\}, \end{aligned} \tag{3}$$

where $E\{\cdot\}$ denote the expectation taken over all pixels in one image.

Generally, one virtual view image comprises of the un-occluded area and the occluded area. Let n_U denote the number of pixels in the un-occluded area in $\bar{S}_V$, n_O the number of pixels in the occluded area in $\bar{S}_V$, $\hat{n}_U$ the number of pixels in the un-occluded area in $\hat{S}_V$, $\hat{n}_O$ the number of pixels in the occluded area in $\hat{S}_V$, n_S the image spatial resolution, $E_U\{\cdot\}$ the expectation taken over the pixels in the un-occluded area and $E_O\{\cdot\}$ the expectation taken over the pixels in the occluded area. Because of the effect of compression, we generally have $\hat{n}_O > n_O$ and $\hat{n}_U < n_U$. Thus,

$$E\{[\bar{S}_V-\hat{S}_V]^2\}=\frac{\hat{n}_U}{n_S}\cdot E_U\{[\bar{S}_U-\hat{S}_U]^2\}+\frac{\hat{n}_O}{n_S}\cdot E_O\{[\bar{S}_O-\hat{S}_O]^2\}, \tag{4}$$

where $\bar{S}_U$ denotes the un-occluded signal synthesized by the original images of the adjacent views and $\hat{S}_U$ denotes the un-occluded signal synthesized by the compressed images of the adjacent views. According to practical 3D warping, we have

$$\begin{aligned} &E_U\{[\bar{S}_U-\hat{S}_U]^2\} \\ &\quad \approx w_A^2(E\{[H_{A->V}(S_A)-H_{A->V}(\hat{S}_A)]^2\} \\ &\quad +E\{[H_{A->V}(\hat{S}_A)-\hat{H}_{A->V}(\hat{S}_A)]^2\}) \\ &\quad +w_B^2(E\{[H_{B->V}(S_B)-H_{B->V}(\hat{S}_B)]^2\} \\ &\quad +E\{[H_{B->V}(\hat{S}_B)-\hat{H}_{B->V}(\hat{S}_B)]^2\}) \\ &\quad +2w_Aw_BE\{[H_{A->V}(S_A)-H_{A->V}(\hat{S}_A)] \\ &\quad \times[H_{B->V}(S_B)-H_{B->V}(\hat{S}_B)]\} \\ &\quad +2w_Aw_BE\{[H_{A->V}(\hat{S}_A)-\hat{H}_{A->V}(\hat{S}_A)] \\ &\quad \times[H_{B->V}(\hat{S}_B)-\hat{H}_{B->V}(\hat{S}_B)]\}, \end{aligned} \tag{5}$$

where $H_{A\to V}()$ and $H_{B\to V}()$ denote the mapping transforms without any depth loss from view A and view B to the synthesized view, respectively. $\hat{H}_{A\to V}()$ and $\hat{H}_{B\to V}()$ denote the mapping transforms with quantization-induced depth loss from view A and view B to the synthesized view, respectively. Eq. (5) is a reasonable approxim-

ation for the practical view synthesis (see Appendix A). In the view synthesis, since $\hat{n}_U/n_S \approx 1$, as analyzed in Appendix A, we have

$$D_V \approx w_A^2(D_A + \Delta D_A|_{depth\_A}) + w_B^2(D_B + \Delta D_B|_{depth\_B}) + 2 \cdot w_A w_B(D_{AB} + \Delta D_{AB}|_{depth\_AB}) + D_G \quad (6)$$

with

$$\begin{aligned} D_A &= E\{[S_A - \hat{S}_A]^2\}, \\ D_B &= E\{[S_B - \hat{S}_B]^2\}, \\ D_{AB} &= E\{[S_A - \hat{S}_A] \cdot [S_B - \hat{S}_B]\}, \\ \Delta D_A|_{depth\_A} &= E\{[H_{A\to V}(\hat{S}_A) - \hat{H}_{A\to V}(\hat{S}_A)]^2\}, \\ \Delta D_B|_{depth\_B} &= E\{[H_{B\to V}(\hat{S}_B) - \hat{H}_{B\to V}(\hat{S}_B)]^2\}, \\ \Delta D_{AB}|_{depth\_AB} &= E\{[H_{A\to V}(\hat{S}_A) - \hat{H}_{A\to V}(\hat{S}_A)] \cdot [H_{B\to V}(\hat{S}_B) - \hat{H}_{B\to V}(\hat{S}_B)]\} \end{aligned}$$

and

$$D_G = E\{[S_V - \bar{S}_V]^2\} + 2E\{[S_V - \bar{S}_V][\bar{S}_V - \hat{S}_V]\} + \frac{\hat{n}_O}{n_S} \cdot E_O\{[\bar{S}_O - \hat{S}_O]^2\}.$$

In Eq. (6), there are three types of distortions. First, the video-coding-induced distortion includes D_A, D_B and D_{AB}. D_A and D_B are 2D video coding distortions. D_{AB} denotes the correlation distortion between the pixel intensity errors of two sources reference views. Second, depth-coding-induced distortion includes $\Delta D_A|_{depth\_A}$, $\Delta D_B|_{depth\_B}$ and $\Delta D_{AB}|_{depth\_AB}$. $\Delta D_A|_{depth\_A}$ and $\Delta D_B|_{depth\_B}$ are depth-quantization-induced warping error distortions for view A and view B, respectively. $\Delta D_{AB}|_{depth\_AB}$ denotes the depth-quantization-induced warping error correlation distortion between two sources reference views. Third, the geometry distortion includes D_{OG}, D_{CG} and D_O, where $D_{OG} = E\{[S_V - \bar{S}_V]^2\}$, $D_{CG} = E\{[S_V - \bar{S}_V][\bar{S}_V - \hat{S}_V]\}$ and $D_O = (\hat{n}_O/n_S) \cdot E_O\{[\bar{S}_O - \hat{S}_O]^2\}$. D_{OG} reflects the distortion caused by the inherent geometry errors and pixel-position rounding errors in 3D warping. The possible geometry errors include depth estimation errors, the depth quantization errors in the conversion from depth data to depth map, and inaccurate camera parameters. D_{CG} denotes the correlations between the inherent geometry errors and compression-induced geometry errors. In the synthesized view, geometry occlusions and pixel mapping ambiguities cause some holes and D_O is the distortion introduced by inpainting these holes.

2.3. View synthesis distortion estimation

Since the captured original image is not available, several parts in Eq. (6) must be estimated for video/depth rate allocation application. The video-coding-induced distortion, including D_A, D_B and D_{AB}, can be directly computed during video coding. As for depth-quantization-induced distortion and geometry distortion, we provide the corresponding estimation methods in the following subsections.

2.3.1. Depth-quantization-induced distortion estimation

Depth-quantization-induced distortion, including $\Delta D_A|_{depth\_A}$, $\Delta D_B|_{depth\_B}$ and $\Delta D_{AB}|_{depth\_AB}$, can be computed by twice warping using the original depth and reconstructed depth. However, this method usually needs encoding the depth map and synthesizing the view for many times during the rate allocation. To reduce the coding times, we propose an estimation method for $\Delta D_A|_{depth\_A}$ and $\Delta D_B|_{depth\_B}$.

In [29], the distortion due to motion warping error is characterized by a linear model. It is expressed as

$$D \approx ||\mathbf{\Delta n}||^2 \psi_x, \quad (7)$$

where $\mathbf{\Delta n}$ is the motion warping error and ψ_x represents the motion sensitivity factor, which is computed as

$$\psi_x = \frac{1}{2 \cdot (2\pi)^2} \cdot \iint_{(-\pi,\pi]} S_x(\omega_1, \omega_2) \cdot (\omega_1^2 + \omega_2^2) d\omega_1 d\omega_2 \quad (8)$$

In Eq. (8), $S_x(\omega_1, \omega_2)$ denotes the energy density of the warping reference frame and (ω_1, ω_2) is the two-dimensional frequency vector. Since the motion warping error corresponds to linear phase shift in frequency domain, the distortion introduced by the motion warping error can be computed as Eq. (7). In depth-based image warping, depth quantization introduces warping errors in the warped image. The depth-based image warping is very similar to the motion warping and hence the depth-quantization-induced warping error distortion can also be described by Eq. (7), which is based on the assumption that the motion error is constant. However, for non-constant motion errors in one frame, it is also reasonable that the frame distortion is characterized by Eq. (7) for the average magnitude of mean-squared motion errors, $\|\mathbf{\Delta n}\|^2$, over all samples in the frame [29]. For self-contained purpose we provide the brief derivation for Eq. (7) in Appendix B.

In depth coding, the quantization brings the depth loss. The depth loss further results in the warping error in the synthesized view image. Fig. 2 shows the relationship between depth loss and warping error. The left is the source reference view and the right is the virtual view to be synthesized. The pixel (u_A, v_A) in view A corresponds to

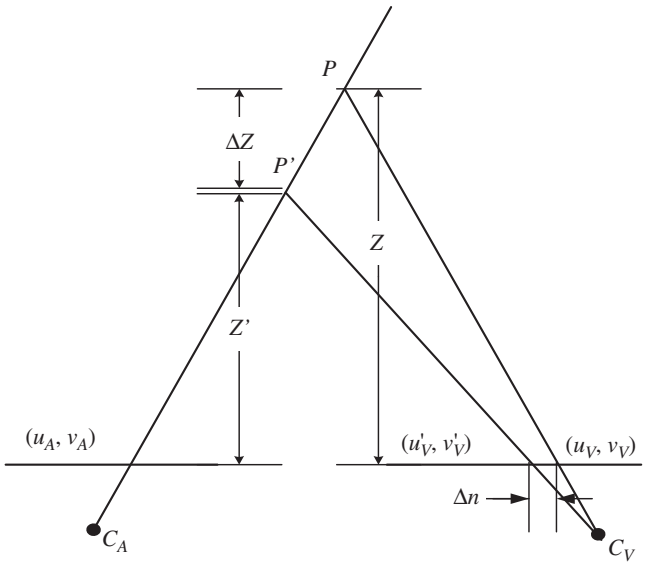

Fig. 2. Relationship between warping error and depth loss (parallel camera setup).

the 3D world point P with depth Z, and it is re-projected to the pixel (u_V, v_V) in the virtual view according to $(u_V, v_V, 1)^T = \mathbf{H}_{A\to V}[Z]\mathbf{m}$, where $\mathbf{m} = (u_A, v_A, 1)^T$ and $\mathbf{H}_{A\to V}[Z]$ is the homography matrix at depth Z from the left view to the virtual view. Due to the quantization of depth map, P loses ΔZ and changes into P'. P' is re-projected to (u'_V, v'_V) by $(u'_V, v'_V, 1)^T = \mathbf{H}_{A\to V}[Z]\mathbf{m}$. Then the warping error $\mathbf{\Delta n}$ is computed by $(\mathbf{\Delta n}, 1)^T = \mathbf{H}_{A\to V}[Z]-\mathbf{H}_{A\to V}[Z-\Delta Z]\mathbf{m}$.

Actually, the warping errors in the virtual view inversely reflect on the sources reference views. Fig. 3 gives the actual warping process with compressed video and depth. The points P_1, P_2 and P_3 are the actual points in surface S. Both the pixel at position **A** of view A and the pixel at position **B** of view B correspond to the 3D world point P_2 with depth Z. The point P_2 is projected to the pixel position V_1 in virtual view V. Hence, the pixel at position $\mathbf{V_1}$ is interpolated by the pixels at **A** and **B**.

In the actual warping, due to the effects of the compression, the 3D world point P_3 with depth $Z_{\mathbf{A_1}}$, which corresponds to pixel at $\mathbf{A_1}$ in view A, loses depth $\Delta Z_{\mathbf{A_1}}$ and then changes into P_4. The point P_4 with depth $Z'_{\mathbf{A_1}}$ will project to the position $\mathbf{V_1}$ in the virtual view. Assume that $\mathbf{m}_V = (u_V, v_V, 1)$, where (u_V, v_V) is the pixel at V_1. Then $\mathbf{\Delta n_A} = (\mathbf{H}_{V\to A}[Z] - \mathbf{H}_{V\to A}[Z-\Delta Z_{\mathbf{A_1}}])\mathbf{m}_V$, where $\mathbf{H}_{V\to A}[Z]$ is the homography matrix at depth Z from the virtual view to the left view A. Likewise, the 3D world point P_1, which corresponds to the pixel at $\mathbf{B_1}$ in view B, loses depth $\Delta Z_{\mathbf{B_1}}$ and then changes into P_5. The point P_5 with depth $Z'_{\mathbf{B_1}}$ will also project to the position $\mathbf{V_1}$ in the virtual view and $\mathbf{\Delta n_B} = (\mathbf{H}_{V\to B}[Z] - \mathbf{H}_{V\to B}[Z-\Delta Z_{\mathbf{B_1}}])\mathbf{m}_V$, where $\mathbf{H}_{V\to B}[Z]$ is the homography matrix at depth Z from the virtual view to the right view B.

For each pixel in the synthesized view image, the warping error reflected on one source reference image is expressed as $\mathbf{\Delta n}_i = (\Delta x_i, \Delta y_i)^T$, where Δx_i is the horizontal error and Δy_i the vertical error. For one image with $M \times N$ resolution, the average magnitude of the mean-squared warping errors is computed as

$$||\mathbf{\Delta n}||^2 \approx \frac{\sum_{i<M\cdot N}(\Delta x_i)^2 + (\Delta y_i)^2}{M \cdot N}. \tag{9}$$

As shown in Appendix B, Eq. (7) is approximated by first-order Taylor expansion for $|1-e^{-j\boldsymbol{\omega}\mathbf{\Delta n}}|^2$ so that it often overestimates the warping error induced distortion for large values of $\mathbf{\Delta n}$. By comparison, the linear model and quadratic model are both reasonable approximations to the actual distortion at low values of $\mathbf{\Delta n}$ [29]. However, they provide larger estimation errors for large values of $\mathbf{\Delta n}$, and the estimation error increases monotonously with the increasing of $\mathbf{\Delta n}$. For such a situation, there are two ways to achieve the further accurate estimation. One is incorporating more terms of Taylor series expansion for $|1-e^{-j\boldsymbol{\omega}\mathbf{\Delta n}}|^2$. This method results in high-order polynomials of $\|\mathbf{\Delta n}\|^2$ in the estimation model, so that the estimation becomes much more complex. The other method is introducing a piecewise linear model for large $\mathbf{\Delta n}$ in terms of the monotonous increment property of estimation error with increasing of $\mathbf{\Delta n}$. Since this method is simple but very effective, we adopt it and establish the following piecewise linear approximation:

$$|1-e^{-j\boldsymbol{\omega}\mathbf{\Delta n}}|^2 \approx \begin{cases} ||\boldsymbol{\omega}\mathbf{\Delta n}||^2, & \text{if } ||\mathbf{\Delta n}||^2 < 2 \\ (a_1 \cdot ||\boldsymbol{\omega}\mathbf{\Delta n}||^2 + b_1), & \text{if } 2 \le ||\mathbf{\Delta n}||^2 < 8, \\ (a_2 \cdot ||\boldsymbol{\omega}\mathbf{\Delta n}||^2 + b_2), & \text{otherwise} \end{cases} \tag{10}$$

where a_1 and a_2 are linear slopes limited in $(0, 1]$, and b_1 and b_2 are constants. These parameters are computed at the piecewise linear ends of $\|\mathbf{\Delta n}\|^2$ for small $(\boldsymbol{\omega}\mathbf{\Delta n})^2$. Quantization usually eliminates the high-frequency components, and therefore $(\boldsymbol{\omega}\mathbf{\Delta n})^2$ is large only at higher values of $\mathbf{\Delta n}$. Large values of $\mathbf{\Delta n}$ have the major effect on $|1-e^{-j\boldsymbol{\omega}\mathbf{\Delta n}}|^2$.

The comparison of linear approximation, piecewise linear approximation, quadratic approximation and $|1-e^{-j\boldsymbol{\omega}\mathbf{\Delta n}}|^2$ with increasing $\|\mathbf{\Delta n}\|^2$ is shown in Fig. 4. When $||\boldsymbol{\omega}||_2 = 0.78$, a_1, a_2, b_1, and b_2 are set to 0.65, 0.5, $\boldsymbol{\omega}^2$, and $2\boldsymbol{\omega}^2$, respectively. These parameters are obtained by linear fitting with the values of $|1-e^{-j\boldsymbol{\omega}\mathbf{\Delta n}}|^2$ at $\|\mathbf{\Delta n}\|^2 = 2$ and $\|\mathbf{\Delta n}\|^2 = 8$. Because $\|\mathbf{\Delta n}\|^2$ is caused by depth quantization and it is generally less than 15, the piecewise linear approximation can work well.

According to Eq. (B1) in Appendix B and Eq. (10), the distortion caused by depth loss for one source reference image can be rewritten as

$$\Delta D|_{depth} \approx \begin{cases} ||\mathbf{\Delta n}||^2 \psi_x, & \text{if } ||\mathbf{\Delta n}||^2 < 2 \\ (a_1 \cdot ||\mathbf{\Delta n}||^2)\psi_x + c_1, & \text{if } 2 \le ||\mathbf{\Delta n}||^2 < 8 \\ (a_2 \cdot ||\mathbf{\Delta n}||^2)\psi_x + c_2, & \text{otherwise} \end{cases} \tag{11}$$

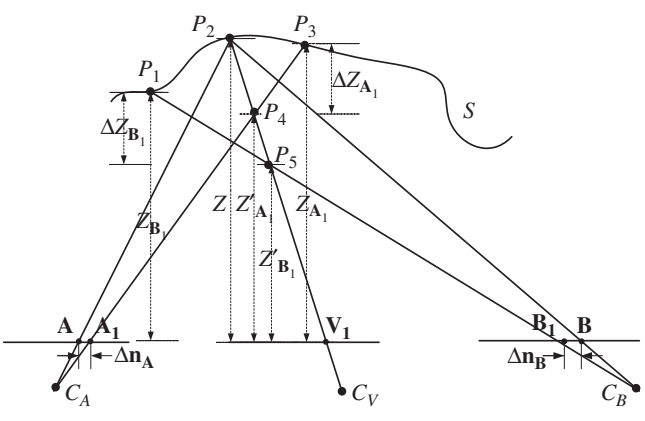

Fig. 3. The pixel mapping with compressed video and depth (parallel camera setup).

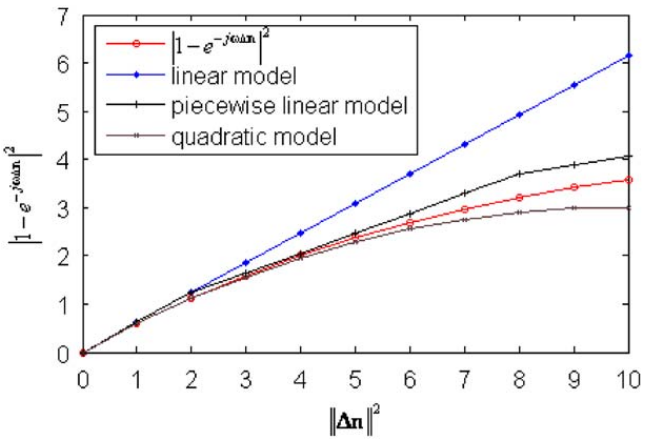

Fig. 4. The comparison of several approximations for $|1-e^{-j\boldsymbol{\omega}\mathbf{\Delta n}}|^2$.

with $c_1 = (1/(2\pi)^2)\int\int_{(-\pi,\ \pi]} S_x(\omega_1,\ \omega_2)\cdot b_1 d\omega_1 d\omega_2$ and $c_2 = (1/(2\pi)^2)\int\int_{(-\pi,\ \pi]} S_x(\omega_1,\ \omega_2)\cdot b_2 d\omega_1 d\omega_2$. In Eq. (11), $\Delta D|_{depth}$ is the depth-quantization-induced warping error distortion for one source reference view. Consequently, $\Delta D_A|_{depth\_A}$ and $\Delta D_B|_{depth\_B}$ can be obtained with Eqs. (9) and (11).

The effect of $\|\mathbf{\Delta n}\|^2$ on warping error distortion for *Breakdancers* is shown in Fig. 5. The experiment is performed with the warping from view0 to the virtual view1. The video is encoded at 200 kb/s bit-rate. The "measured" legend denotes that the distortion is the MSE between the warped frame using original depth and the one using compressed depth. The other legends denote the depth-quantization-induced warping error distortions with the corresponding models, respectively. As shown in Fig. 5, the depth-quantization-induced warping error distortion monotonously increases with the increase of $\|\mathbf{\Delta n}\|^2$. From the comparison, it can be seen that the piecewise linear model can work better than the other models for larger values of $\|\mathbf{\Delta n}\|^2$.

In the course of computing $\|\mathbf{\Delta n}\|^2$, twice warping using original depth and compressed depth are performed. At the same time of twice warping, the depth-quantization-caused warping error correlation distortion between the two sources reference views, $\Delta D_{AB}|_{depth\_AB}$, can be obtained.

2.3.2. Geometry distortion estimation

In Eq. (6), the geometry distortion includes three parts, namely D_{OG}, D_{CG} and D_O. D_{OG} mainly reflects the inherent geometry error effects and the pixel-position rounding effects. Without any priori knowledge, it is very hard to be accurately estimated. Fortunately, it is independent of the compression effect on view synthesis. In multiview camera setup, the same scene is captured by different cameras and hence the behaviors of geometry noise and pixel-position rounding for different view syntheses are very closely related to their baselines. For example, the view synthesis with wide-baseline has larger value of D_{OG} than the view synthesis with small baseline.

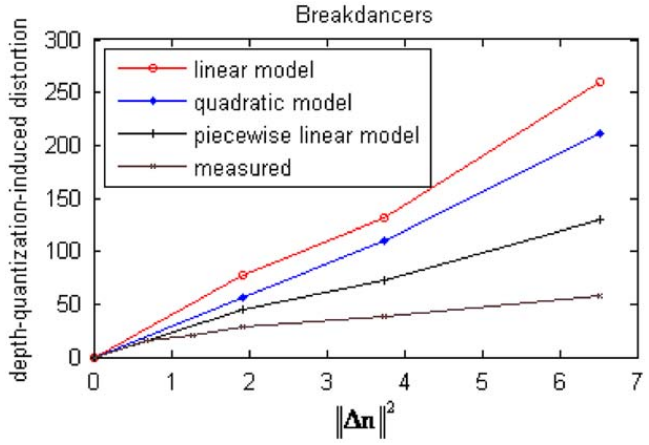

Fig. 5. Depth-quantization-induced distortion varies with mean-squared warping error.

Assume that the captured view0 (view9), view2 (view11), and view4 (view13) exist, and view1 (view10) is the virtual view, as shown in Fig. 6. Let D_{OG_view1} denote D_{OG} for view1 synthesized by view0 and view2, and D_{OG_view2} denotes D_{OG} for view2 synthesized by view0 and view4. For *Breakdancers* and *Ballet*, statistical results show that $D_{OG\_view1}/D_{OG\_view2}$ is approximately equal to 0.7, as shown in Table 1. As for Book *Arrival*, $D_{OG\_view1}/D_{OG\_view2}$ is approximately equal to 0.5. Thus, D_{OG_view1} can be approximately scaled by the already known D_{OG_view2}.

Though the assistant view synthesis can be found, the D_{OG} ratio between the current virtual view synthesis and the assistant view synthesis is not known in advance. Here, we set it with a constant value according to the specific relation between the current virtual view synthesis and the assistant view synthesis. This leads to a certain degree of estimation deviation. However, since D_{OG} is uncorrelated with the compression effect on view synthesis, it does not affect the identification of the optimal quantization pair in video/depth rate allocation.

D_{CG} is a part of compression-related distortion, which mainly reflects the correlations between the inherent geometry error and the compression-induced geometry error. Generally, it is a minus value. In small-baseline camera setup, with the same compression grade for all sources reference views, statistical results show that the values of D_{CG} for different view syntheses are proportional to their spatial relations. Fig. 7 shows the values of D_{CG} for

Table 1
Statistical results of D_{OG} for different view syntheses.

Sequence	D_{OG_view1}	D_{OG_view2}	$D_{OG\_view1}/D_{OG\_view2}$
Breakdancers	72.08	102.17	0.705
Ballet	93.02	132.51	0.701
Book Arrival	23.20	43.45	0.533

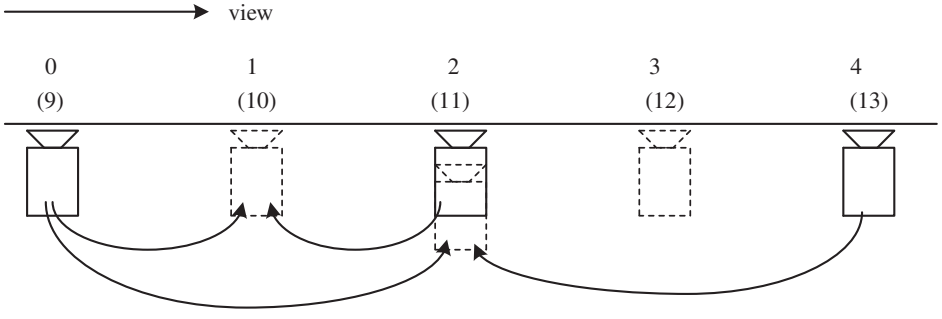

Fig. 6. The position relations among different views. In the experiments, the views for *Breakdancers* and *Ballet* correspond to view0–view4, and the views for *Book Arrival* correspond to view9–view13. The arrow denotes the 3D warping.

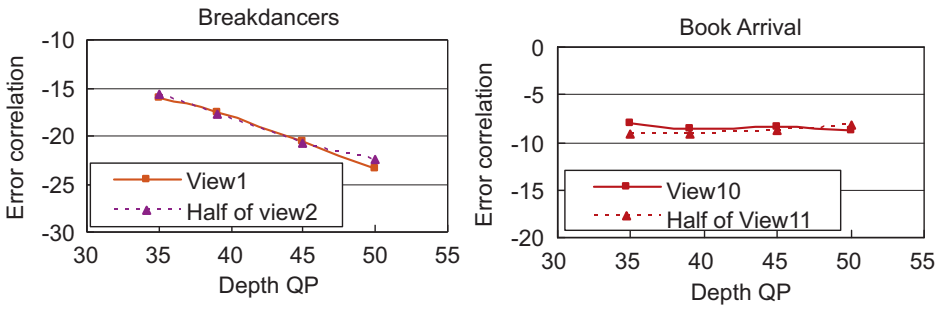

Fig. 7. The values of D_{CG} (error correlation denotes D_{CG} value).

Breakdancers and *Book Arrival*. In Fig. 7, the videos are compressed with $QP = 37$. It shows that the value of D_{CG} for view1 is approximately half of that of view2. This just matches the baseline relation between the view1 synthesis and the view2 synthesis. Using this approximate relation, D_{CG} for synthesizing view1 by compressed view0 and view2 can be estimated from that for synthesizing view2 by compressed view0 and view4.

In the view synthesis, inpainting the occluded areas also contributes a part to the total distortion. Though the holes by pixel mapping or occlusion are not very large, the distortion introduced by them cannot be neglected. In multiview camera array, the occlusion between the adjacent views is related to their spatial relations. Based on this property, occlusion-processing-induced distortion, D_O, for the virtual view can be obtained using the adjacent view synthesis. According to multiview geometry, the occlusion between two views is usually proportional to their baselines. Correspondingly, the occlusion-incurred distortion in the synthesized view1 is approximately a half of that of the synthesized view2. We first synthesize view2 using compressed view0 and view4, which are compressed with the same QP as the sources reference views of the virtual view1, to achieve the occlusion-processing-induced distortion for view2. Then the occlusion-processing-induced distortion for the virtual view1 is scaled by this achieved distortion. Since the sources reference views of the assistant view synthesis have the same compression grade as those of the current virtual view synthesis, a part of occlusion-processing-induced distortion, which is caused by depth-compression-induced ambiguous pixel mapping, is also approximately scaled.

All parts of geometry distortions are estimated from the already known assistant view syntheses. With the multiview camera setup, we are always able to find the assistant view syntheses, such as synthesizing view2 using uncompressed view0 and view4, and synthesizing view2 using compressed view0 and view4 in Fig. 6, to estimate the geometry distortion.

Geometry distortion highly depends on the specific camera setup. When the virtual view synthesis happens with a very small baseline, the occluded area is very tiny and the occlusion-processing-induced distortion can be neglected. The distance between the assistant view synthesis position and the virtual view position also has a great effect on the geometry distortion. At the present time, we restrict this distance to no more than twice the interval between two adjacent views. For *Breakdancers* and *Ballet* [6], the cameras are on an arc line with a small angle and the baseline between two adjacent views is generally less than 20 cm. For *Book Arrival* sequence [30], the cameras are parallel and the baseline is about 6.5 cm. Since the view synthesis quality becomes worse with the increase of baseline, our geometry distortion can get the efficient estimation when the baseline of view synthesis is no more than twice the interval between two adjacent views.

3. Joint video/depth rate allocation based on view synthesis distortion model

The video plus depth representation is able to provide the high quality view synthesis for 3D video and free-viewpoint video applications. As for the 3D broadcast application, the virtual viewpoint position is usually fixed and already known in advance at server side. In free-viewpoint video applications, the viewpoint position information can also be obtained in real time from the client side. Under the channel bandwidth constraint, either in 3D video application or in free-viewpoint video application, the rate allocation between 2D video and depth has a great influence on the view synthesis quality. Since the proposed distortion model can evaluate the view synthesis quality, it can be used to select the optimal combination of quantized video and depth to maximize the view synthesis quality.

The rate allocation problem for multiview video and depth coding is first addressed by Morvan et al. [25], which proposes to seek the optimal quantization parameter pair, $(q_v{}^{opt}, q_d{}^{opt})$, under the total bit-rate constraint R_c. Here, $q_v{}^{opt}$ and $q_d{}^{opt}$ are the optimal quantization parameters for video and depth. Let R_v, R_d, q_v, q_d denote the 2D video bit-rate, depth map bit-rate, 2D video quantization parameter, depth map quantization parameter, respectively. Then this problem is formulated as

$$(q_v^{opt}, q_d^{opt}) = \arg\min_{q_v, q_d \in Q} D_{virtual\_view}(q_v, q_d)$$
$$\text{subject to } R_v(q_v^{opt}) + R_d(q_d^{opt}) \le R_c, \tag{12}$$

where $D_{virtual\_view}(q_v, q_d)$ is the view synthesis distortion with q_v for video and q_d for depth, and Q is the candidate quantization parameter set.

According to Eq. (12), if the video has low quality, the final synthesis will have low quality no matter whether the depth is in low or high quality. Contrarily, if the depth map has low quality, the final synthesis will also have low quality no matter whether the video has low or high quality. In practical 3D video applications, depth map is only side information for view synthesis. However, 2D video is different from it and usually needs maintaining higher quality for the purpose of being compatible with 2D video display or reusing for other virtual view synthesis. Consequently, the joint video/depth rate allocation optimization problem can be re-defined as

$$(q_v^{opt}, q_d^{opt}) = \arg\min_{q_v, q_d \in Q} D_{virtual\_view}(q_v, q_d) \quad \text{subject to} \begin{cases} R_v(q_v^{opt}) + R_d(q_d^{opt}) \le R_c \\ R_v(q_v^{opt}) \ge R_{threshold} \end{cases}, \tag{13}$$

where $R_{threshold}$ is usually set to $R_c/2$ for the practical applications. Based on Eqs. (6) and (13), we propose a model-based rate allocation strategy for video/depth-based 3D video coding.

3.1. Two statistical relationships

In our proposed view synthesis distortion model, $\|\mathbf{\Delta n}\|^2$ and $\Delta D_{AB}|_{depth\_AB}$ can be computed in 3D warping using the original depth map and the compressed depth map. However, it involves 3D warping and depth coding before rate allocation. To avoid these complicated works, we propose two statistical relationships to estimate $\|\mathbf{\Delta n}\|^2$ and $\Delta D_{AB}|_{depth\_AB}$.

3.1.1. Relationship between $\|\mathbf{\Delta n}\|^2$ and q_d

In the experiment, we observe that the relationship between $\|\mathbf{\Delta n}\|^2$ and depth distortion D_d for the whole frame is approximated as a linear model. It is expressed as

$$||\mathbf{\Delta n}||^2 = \alpha \cdot D_d, \tag{14}$$

where α is a constant. At two already known depth coding points, $\|\mathbf{\Delta n}\|^2$ can be obtained by twice warping using the original depth and reconstructed depth, and correspondingly α can be computed. Fig. 8 shows the D_d–$\mathbf{\Delta n}$ model for *Breakdancers* with $\alpha = 0.17$ and *Book Arrival* with $\alpha = 0.018$. In the figure, 2D video is encoded with $QP = 37$.

The relationship between distortion and quantization parameter q (D–Q model) is described as $D_d = 255^2/10^{((k \cdot q+n)+10)}$ [31], where k and n are constants. Therefore, we have the following $\mathbf{\Delta n}$–q_d model:

$$||\mathbf{\Delta n}||^2 = \alpha \cdot \frac{255^2}{10^{((k \cdot q_d+n)+10)}}. \tag{15}$$

According to the $\mathbf{\Delta n}$–q_d model, $\|\mathbf{\Delta n}\|^2$ can be estimated with q_d.

3.1.2. Relationship between $\Delta D_{AB}|_{depth\_AB}$ and q_d

$\Delta D_{AB}|_{depth\_AB}$ is caused by the depth compression. When the video compression grade is fixed, the absolute value of $\Delta D_{AB}|_{depth\_AB}$ increases with the increase of q_d. Statistics show that the relationship between $\Delta D_{AB}|_{depth\_AB}$ and q_{d_step} can be taken as linear, as shown in Fig. 9. Here,

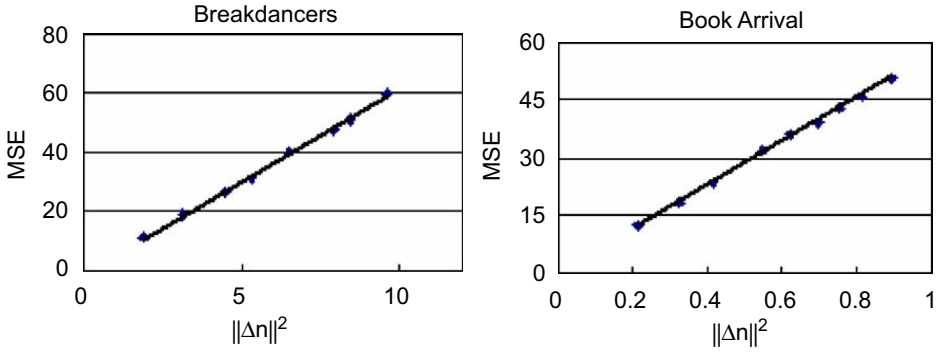

Fig. 8. D_d–$\mathbf{\Delta n}$ model.

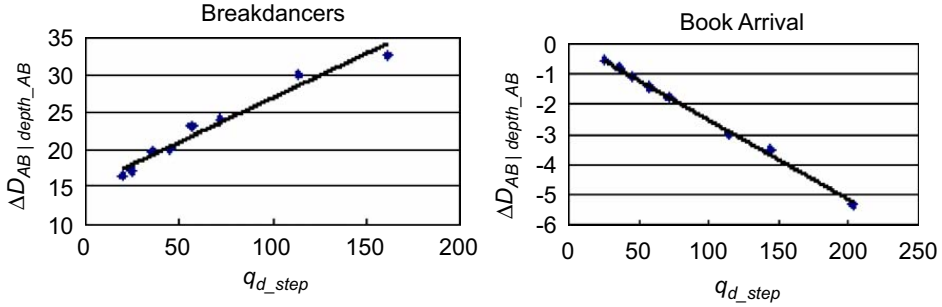

Fig. 9. The linear relationship between $\Delta D_{AB}|_{depth\_AB}$ and q_{d_step}.

q_{d_step} denotes the quantization step size. In H.264/AVC, the relation between q_d and q_{d_step} is that $q_{d\_step} = 2^{(q_d-4)/6}$. Hence, the relationship between $\Delta D_{AB}|_{depth\_AB}$ and q_d can be expressed as

$$\Delta D_{AB}|_{depth\_AB} = a_3 \cdot 2^{(q_d-4)/6} + b_3, \quad (16)$$

where a_3 and b_3 are constants. In Fig. 9, $a_3 = 0.145$ and $b_3 = 13.57$ for *Breakdancers*, and $a_3 = -0.025$ and $b_3 = 0.12$ for *Book Arrival*.

To guarantee the virtual view synthesis quality, the left source reference depth and the right source reference depth generally have the same quantization grade. Therefore, $\Delta D_{AB}|_{depth\_AB}$ can be estimated with q_d.

3.2. Model-based rate allocation algorithm

The proposed view synthesis distortion model can facilitate the rate allocation between video and depth. Hence, we propose a model-based rate allocation method, which can be performed as the following steps.

First, find the proper quantization parameter ranges for 2D video and depth map. Given the bit-rate constraint R_c, the possible bit-rate range of 2D video, $[R_{v\_min}(q_{v\_max}), R_{v\_max}(q_{v\_min})]$, is first found. Here, $R_{v\_min}(q_{v\_max})$ is set to $R_{threshold}$ and $R_{v\_max}(q_{v\_min})$ is set to an upper bound value. Because $R_{d\_min}(q_{d\_max}) = R_c - R_{v\_max}(q_{v\_min})$ and $R_{d\_max}(q_{d\_min}) = R_c - R_{v\_min}(q_{v\_max})$, the possible bit-rate range $[R_{d\_min}(q_{d\_max}), R_{d\_max}(q_{d\_min})]$ and q_d range $[q_{d\_min}, q_{d\_max}]$ for depth map can be obtained. After that, depth map is encoded at two bit-rate ends, q_{d_min} and q_{d_max}.

Second, estimate D_{OG}. With the virtual view position, the assistant view synthesis by uncompressed views can be determined. D_{OG} can be estimated using the method provided in Section 2.3.2.

Third, encode 2D videos at q_v, where $q_v \in [q_{v\_min}, q_{v\_max}]$ and then compute D_A, D_B, D_{AB} and the motion sensitivity factor ψ_x. After that, through 3D warping at two depth rate ends, two statistical relationships described in Section 3.1 can be established and consequently the model parameters α, a_3 and b_3 for the current 2D video bit-rate point can be obtained.

Fourth, determine q_d for the current 2D video bit-rate point. According to the bit-rate constraint, we can obtain a proper bit-rate for depth map. Since the relationship between the bit-rate R and the reciprocal of quantization step $1/Q_{step}$ can be taken as linear in H.264/AVC [31], q_d for depth map with the known bit-rate can be estimated. The R–Q_{step} model is

$$R = \frac{d_1}{Q_{step}} + d_2, \quad (17)$$

where d_1 and d_2 are constants, which can be computed by the known R–Q points, $R_{d\_min}(q_{d\_max})$ and $R_{d\_max}(q_{d\_min})$.

Table 2
Encoding conditions.

Coding structure	Hierarchical B pictures
GOP size	15
Search range	48 (pixel)
Entropy coding	CABAC
ICMode & MotionSkipMode	Off

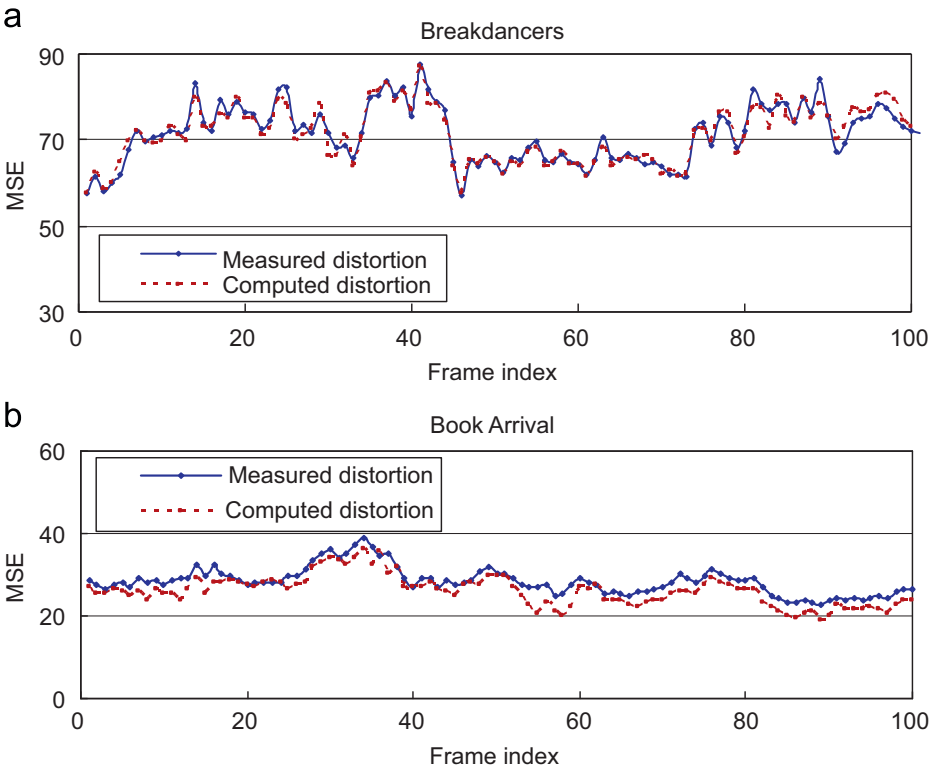

Fig. 10. The comparison between the actual distortion and the computed distortion using the distortion model: (a) the virtual view is view1 synthesized by view0 and view2 and (b) the virtual view is view10 synthesized by view9 and view11.

Fifth, estimate the virtual view synthesis distortion $D_{virtual\_view}(q_v, q_d)$ for the current video/depth quantization pair. Based on the assistant view synthesis with compressed views, D_{CG} and D_O can be first estimated for the current video/depth quantization pair. Using the statistical relationships established in the third step, $\|\mathbf{\Delta n}\|^2$ and $\Delta D_{AB}|_{depth\_AB}$ are then computed for the current video/depth quantization pair. Subsequently, based on Eq. (7), $\Delta D_A|_{depth\_A}$ and $\Delta D_B|_{depth\_B}$ can also be computed for the current video/depth quantization pair. With already obtained D_{OG}, D_A, D_B and D_{AB}, $D_{virtual\_view}(q_v, q_d)$ for the current video/depth quantization pair can be calculated.

Sixth, let $q_v = q_v+1$. If $q_v \le q_{v\_\max}$, go to the third step.

At last, select the optimal $(q_v{}^{opt}, q_d{}^{opt})$ with $(q_v^{opt}, q_d^{opt}) = \arg\min_{q_v,q_d \in Q} D_{virtual\_view}(q_v, q_d)$ from all quantization pairs.

Compared with the full-search method, we can summarize that the proposed algorithm has the following advantages. (1) It involves only one searching-iteration (2D video iteration) to find the optimal quantization setting pair. (2) It reduces the depth coding times from N to 2, where $N = q_{d\_\max} - q_{d\_\min}$. (3) It avoids the view synthesis for $M \times (N-3)-1$ times, where $M = q_{v\_\max} - q_{v\_\min}$. In model-based rate allocation, the assistant view syntheses using compressed videos and depth maps are performed for M times to obtain D_{CG}. In addition to that, one assistant view synthesis with uncompressed videos and depth maps for D_{CG}, and the view syntheses for $2M$ times in order to determine α, a_3 and b_3 are also needed. Thus, the total number of view syntheses is $3M+1$ in model-based rate allocation compared with $M \times N$ in full-search method. (4) It does not need the original viewpoint image as reference to guide the rate allocation.

Though model-based rate allocation method has the above merits, it introduces the estimation of the power spectrum density and consequently incurs additional computations. Also, its performance is related to the accuracy of the proposed distortion model. However, the introduced computational complexity is similar to that of the view synthesis. Even taking into account the additional assistant view syntheses, model-based method can

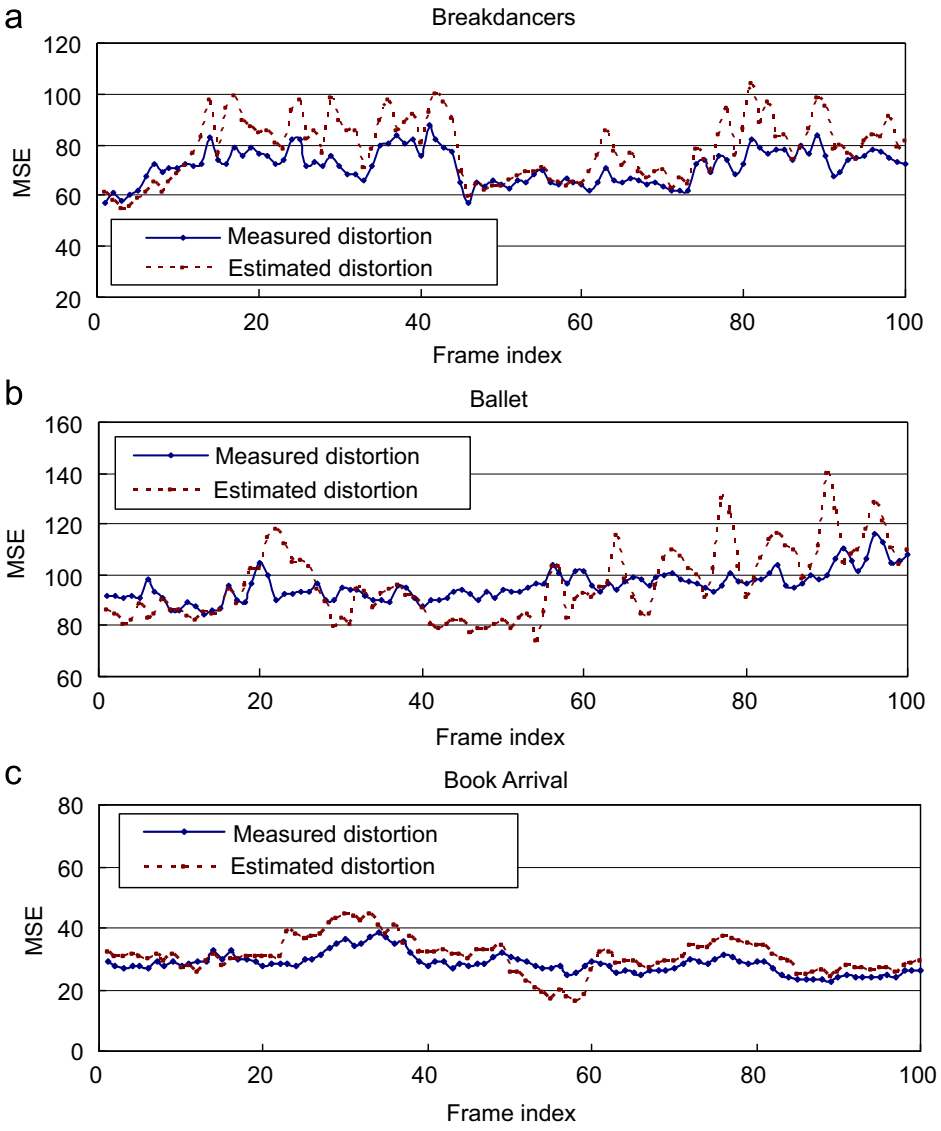

Fig. 11. Measured distortion versus estimated distortion.

also save much computational complexity compared with the full-search method.

4. Experimental results

This section provides the performance analyses of view synthesis distortion model and rate allocation. The 3D video sequences of *Breakdancers*, *Ballet* and *Book Arrival* (1024×768) are used in the experiments. As shown in Fig. 6, the views from 0 to 4 for *Breakdancers* and *Ballet*, and the views from 9 to 13 for *Book Arrival* are used. The multiview videos and depth maps are both coded with simulcast coding using MVC software JMVM6.0. The specific coding conditions are shown in Table 2.

4.1. View synthesis distortion model performance

4.1.1. Verification of distortion model

Eq. (6) has been derived as the mathematical description of virtual view synthesis distortion model in Section 2.2. It can accurately characterize the virtual view synthesis distortion, as illustrated in Fig. 10, which shows the comparisons between the actually measured distortion and the computed distortion for Breakdancers and Book Arrival. The computed distortion is obtained by Eq. (6). In Fig. 10, the captured original view at the virtual view position is assumed to be existed. The sources reference videos and depth maps are coded with $QP = 37$ and 35, respectively.

4.1.2. Distortion model estimation accuracy

Since several parts in the view synthesis distortion model must be estimated, this subsection verifies the estimation accuracy of the proposed view synthesis distortion model. Fig. 11 shows the comparison between the estimated distortion and the measured distortion for view1 (view10) synthesized by view0 (view9) and view2 (view11). The measured distortion is evaluated by MSE between the synthesized image and the original image. In Fig. 11, the virtual view is synthesized by the video compressed with $QP = 37$ and depth compressed with

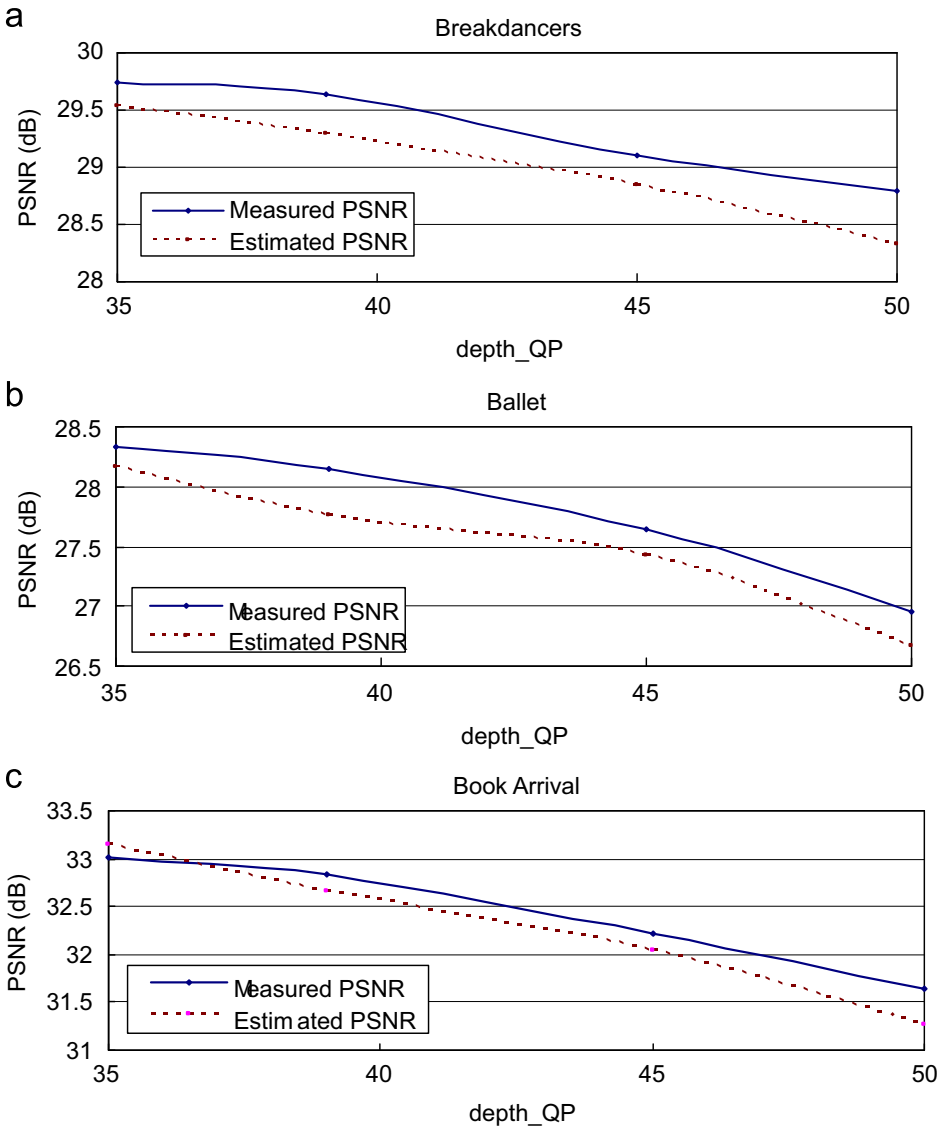

Fig. 12. Measured PSNR versus estimated PSNR with different *QP* of source views' depths.

$QP = 35$. The geometry distortion is estimated from that of the synthesized view2 (view11) by view0 (view9) and view4 (view13).

In the estimation, the geometry distortion is not sufficiently described so that the estimated distortion curve takes on a little jitter. However, the estimated curve shows the similar trend as the measured curve. Since two items of the geometry distortion, D_{OG} and D_{CG}, are obtained using the adjacent view synthesis, the multiview geometry noise effects and pixel-position rounding effects characterized by them are only approximately estimated. Actually, if we can get very perfect geometry information so that $S_V \approx \bar{S}_V$, D_{OG} and D_{CG} will disappear. Once these two parts of distortions are both removed, the accuracy of the proposed view synthesis distortion model can be greatly improved. Additionally, the current model does not consider the impacts of illumination and color inconsistencies among views on the view synthesis. These factors also affect the actual view synthesis and make the estimated distortion with a little deviation from the actual distortion.

Fig. 12 illustrates the estimated PSNR performances at different depth QP points of the sources reference views. The source video QP is unchanged with $QP = 37$. Fig. 12 shows that the average estimation error, in terms of PSNR, is less than 0.6 dB. From Fig. 12, it can also be observed that both the estimated distortion and the measured distortion monotonously decline with the increasing of depth QP. This monotonous changing property of the estimation model can assist the video/depth rate allocation.

Fig. 13 shows the comparison of the estimated PSNR and measured PSNR for different rate pairs at the fixed total rate point (here, $R_{threshold} = 0$). In Fig. 13, the distortion model overestimates the actual distortion so that the estimated PSNR curve is lower than the measured

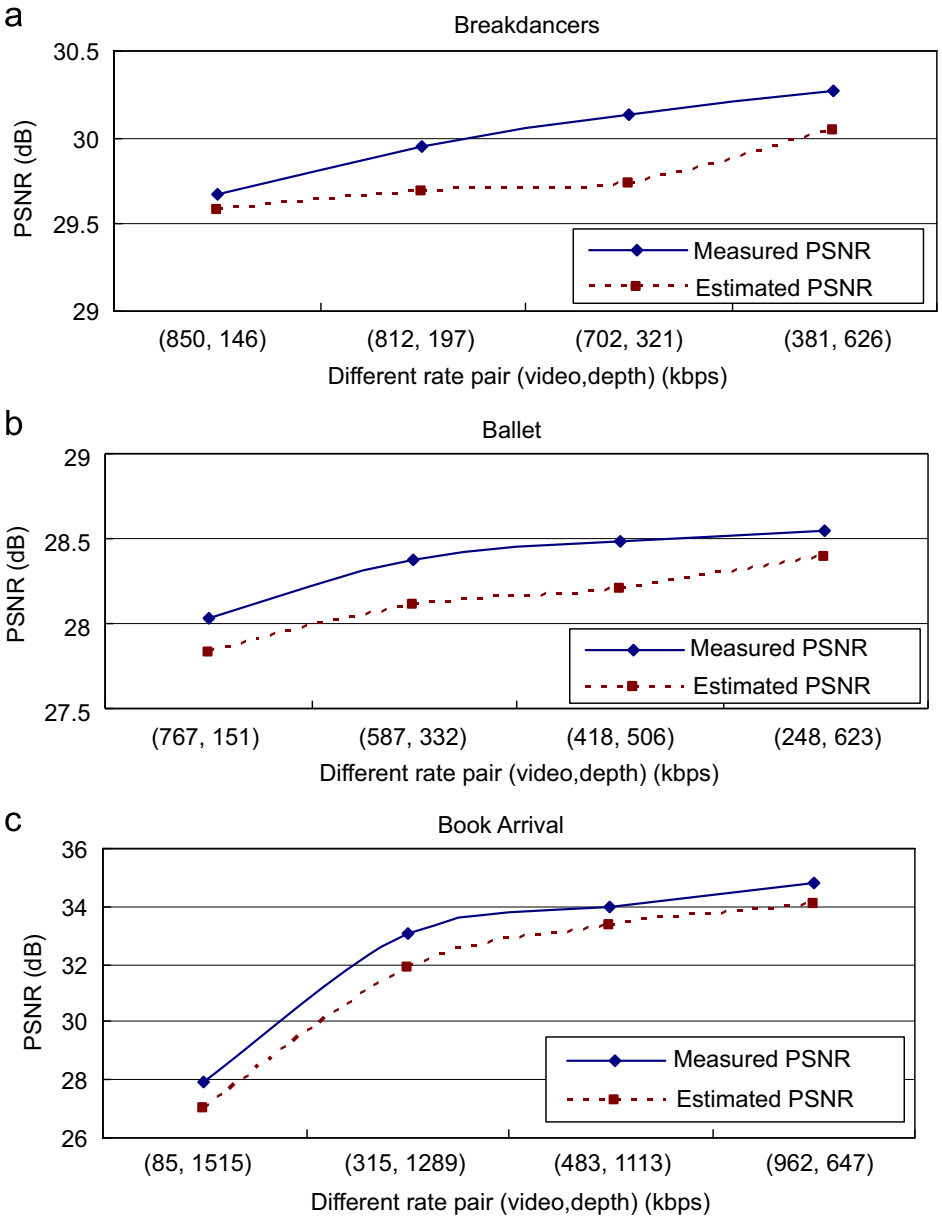

Fig. 13. Estimated PSNR versus measured PSNR for different rate pairs at the fixed total rate point: (a) at 1000 kbps, (b) at 900 kbps and (c) at 1600 kbps.

PSNR curve. Since our distortion estimation for rate allocation only involves identification of the optimum quantization parameter combination between video and depth, a certain degree of inaccuracy in the estimated distortion can be tolerated. It can be observed that the distortion model can differentiate different qualities for the synthesized view with different video/depth quantization combinations.

4.2. Rate allocation performance

In this section, we provide the performance analysis of the proposed model-based rate allocation. The sequence-level rate allocation performance is verified in the experiments. Three anchors are used in the experiments. The first is full-search rate allocation method. This method iteratively searches the possible video/depth quantization pairs to find the optimal rate allocation. The second is hierarchical-search rate allocation method. This method performs a coarse-to-fine searching in all candidate video/depth quantization pairs through a hierarchical-search pattern. Either full-search method or hierarchical-search method, the search-based method selects the optimal quantization pairs via constructing a joint video and depth rate-distortion surface. The third is constant rate allocation with the predefined ratio of 5:1 between 2D video bit-rate and depth bit-rate [11].

Fig. 14 shows the performance comparisons for different rate allocation methods. In Fig. 14, the PSNR represents the distortion of the synthesized view. Since the coding structure, the coded frame number, the total bit-rate description (x-axis in Fig. 14), and especially the view synthesis position, are different from those in Ref. [25], the RD curve in Fig. 14 has some differences from that in [25].

From Fig. 14, it can be seen that model-based method has almost the identical rate allocation performance as search-based (full-search and hierarchical-search) methods. In model-based rate allocation, the distortion model sometimes only finds the sub-optimal quantization pairs because the estimation is not accurate enough at those points and it hence has a little performance loss compared with the full-search method. The full-search method finds the optimal balance between video and depth bit-rates for each bit-rate point of the synthesized view, by traversing all quantization combinations of video and depth. Evidently, the full-search method is more robust.

For the three sequences, it can be seen that search-based or model-based method improves the compression performance about 0.3–1 dB over the fixed ratio 5:1 method. It illustrates that the joint video/depth rate allocation optimization can achieve a compression performance improvement over the constant video/depth ratio rate allocation. In Fig. 14(c), since the original estimated depth maps are not very perfect, the depth compression has a smaller effect on the virtual view synthesis quality compared to that of video compression. Hence, the fixed ratio 5:1 rate allocation method presents better compression performance.

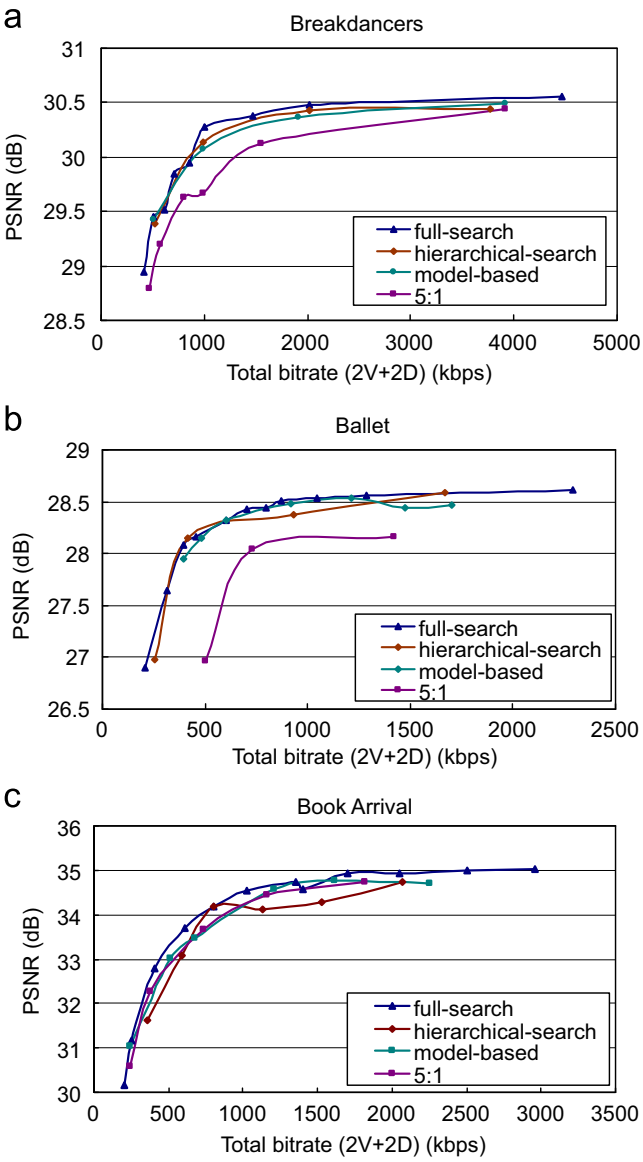

Fig. 14. The view synthesis quality comparison of several rate allocation methods. The virtual view is view1 synthesized by view0 and view2 in (a) and (b), and the virtual view is view10 synthesized by view9 and view11 in (c).

4.3. Computational complexity analysis

To compare the complexities of different rate allocation methods, we execute different methods on a PC with 3.2 GHz Single Core Pentium(R) CPU and 1 GB RAM. Fig. 15 shows the computational time comparisons between model-based method and search-based method. Here, 2D video and depth are offline encoded for 100 frames, and the computational time of rate allocation does not include the coding time.

Compared with full-search method, model-based method does not need encoding all quantization pairs. The computational time of model-based method is far less than that of full-search method. It is worth noting that the time required by full-search method depends on the number of iterations. The more iterations are performed, the more view synthesis operations are involved. In the experiment, the iteration number of full-search method is

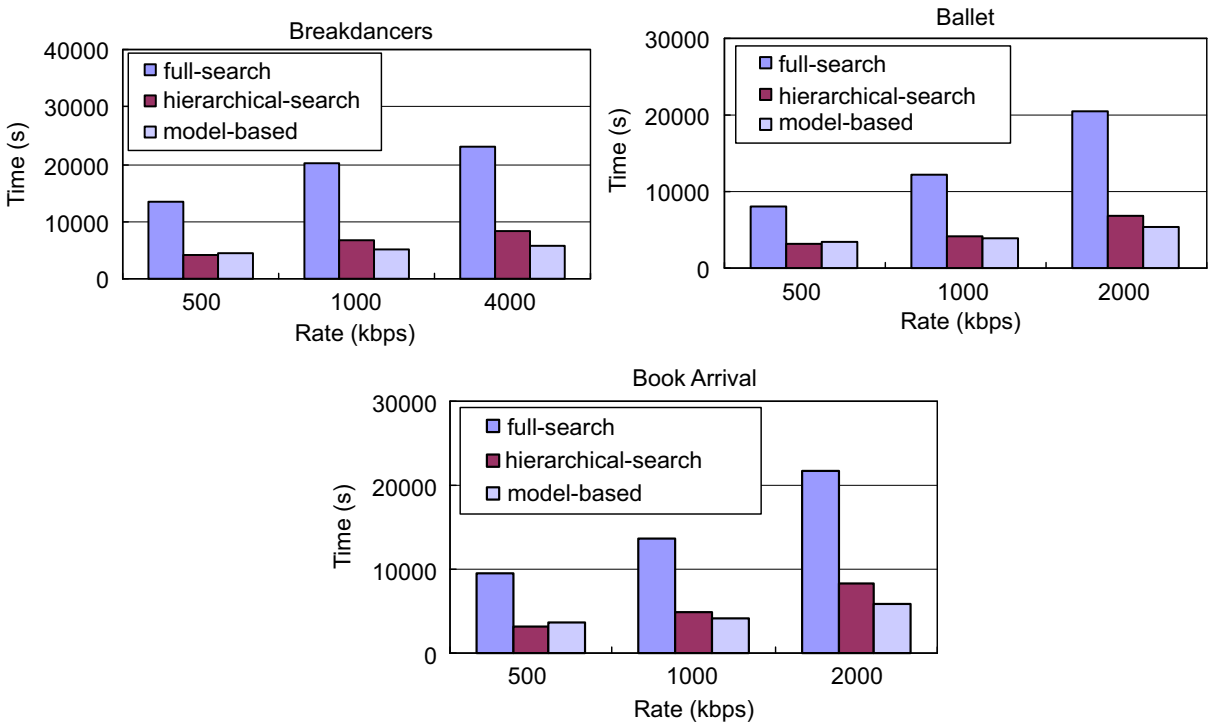

Fig. 15. Computational time comparisons for different methods.

$M \times N$, where $M = q_{v\_\max} - q_{v\_\min} = 28$ and $N = q_{d\_\max} - q_{d\_\min} = 28$. In Fig. 15, at low bit-rate, the bit-rate range is reduced due to the low bit-rate constraint so that the iteration number is reduced, and thus the timing cost at low bit-rate is lower than that at high bit-rate.

For model-based method, it only needs M iterations. For each iteration, it only needs power spectra density estimation and the assistant view synthesis. The video power spectra density estimation involves finding the Fourier transform of a windowed autocorrelation estimate [32] so that it is also a time-consuming work. In our experiment, the view synthesis has the similar complexity with the power spectra density estimation. The assistant view synthesis needs to be performed M+1 times. If we regard the power spectra density estimation as one view synthesis in the sense of complexity, the total number of view synthesis in model-based method is also far less than that of full-search method. As a result, model-based rate allocation can save much more computations than full-search method.

In the experiment, hierarchical-search method makes use of the 3×3 search pattern [25] to reduce the search points by a coarse-to-fine mode. However, it also involves two-dimensional recursive searching. For a full-search with $M \times N$ iterations, hierarchical-search usually performs about $M \times N/4$ iterations. By comparison, model-based method performs about M iterations. In hierarchical-search method, one view synthesis is needed in each iteration. Taking into account the cost of power spectra density estimation, the assistant view synthesis and the twice warping for estimating the statistical models, model-based method will consume about $3M$+1 view syntheses. At low bit-rate, since N is smaller than 12, the model-based method consumes more time than the hierarchical-search method. However, at high bit-rate, N is much larger than 12, and model-based method has less computational complexity than the hierarchical-search method, as illustrated in Fig. 15.

In sequence-level rate allocation experiments, though the coding iterations in model-based method are greatly reduced, the coding still consumes a huge amount of time compared to the rate allocation process. In model-based rate allocation, the ratio of coding time to rate allocation time is approximately equal to 14:1. Since our experiments aim at rate allocation complexity reduction, the coding is not optimized for speed. When fast video encoding with multi-core chips is adopted, the proposed rate allocation method will present the significant effect on the total complexity reduction.

5. Conclusion

This paper proposes a joint video/depth rate allocation method based on view synthesis distortion model for 3D video coding. The proposed view synthesis distortion model takes into account three types of dominating distortion contributions in the view synthesis, namely the video-coding-induced distortion, the depth-quantization-induced distortion and the geometry distortion. With the distortion model, the actual distortion for the intermediate view can be approximately estimated in absence of the original reference view. Given the channel rate constraints, the proposed rate allocation method can find the optimal trade-off between depth bit-rate and video bit-rate to maximize the view synthesis quality. Experimental results indicate that, compared with full-search method, the proposed method can optimize the joint video/depth rate allocation with less

computational complexity. At high bit-rate, the proposed method also consumes less computational cost than hierarchal-search method while providing almost the same rate allocation performance.

The proposed distortion model is sufficiently accurate for sequence-level rate allocation. However, it needs to further promote the estimation accuracy to increase the rate allocation reliability. For real-time 3D video application on variable bit-rate (VBR) channel, the GOP-level or frame-level rate control is very important for guaranteeing the service quality. In the future, we shall consider promoting the model accuracy with more precise geometry information and investigate the efficiency of the proposed method for 3D video GOP-level rate allocation on VBR channel.

Acknowledgements

This work was supported in part by the National Basic Research Program of China (973 Program) under Grant 2009CB320905, and National Natural Science Foundation of China under Grant 60736043 and 60833006. The authors would like to thank Microsoft Research for providing the 3D video sequences of *Ballet* and *Breakdancers*, and thank Fraunhofer HHI for providing the 3D video sequence of *Book Arrival*. The authors also would like to thank the editors and anonymous reviewers for their valuable comments.

Appendix A. Proof of Eq. (5)

To facilitate understanding, we first provide some nomenclatures for the virtual view synthesis in Fig. 1. $\hat{S}_A$ and $\hat{S}_B$ are the compressed images of the left view and the right view, respectively; S_{U_A} and S_{U_B} are the partial original signals come from the left view and the right view, respectively; $\hat{S}_{U_A}$ and $\hat{S}_{U_B}$ are the partial compressed signals come from the left view and the right view, respectively; $\hat{n}_U$ denotes the number of pixels in the un-occluded area in the virtual view image synthesized by the compressed images of the adjacent views, $\hat{n}_O$ denotes the number of pixels in the occluded area in the virtual view image synthesized by the compressed images of the adjacent views; n_S denotes the image spatial resolution; $E_U\{\cdot\}$ denotes the expectation taken over the pixels in the un-occluded area; $E\{\cdot\}$ denotes the expectation taken over all pixels in one image.

According to the specific view synthesis process, the un-occluded area in the virtual view is generally blended by the partial signals of sources reference views. Therefore, the compression-induced view synthesis distortion in the un-occluded area is

$$\begin{aligned}
&E_U\{[\bar{S}_U-\hat{S}_U]^2\}\\
&\quad=E_U\{[w_AH_{A\to V}(S_{U_A})+w_BH_{B\to V}(S_{U_B})\\
&\qquad-w_A\hat{H}_{A\to V}(\hat{S}_{U_A})-w_B\hat{H}_{B\to V}(\hat{S}_{U_B})]^2\}\\
&\quad=E_U\{[w_A(H_{A\to V}(S_{U_A})-\hat{H}_{A\to V}(\hat{S}_{U_A}))]^2\}\\
&\qquad+E_U\{[w_B(H_{B\to V}(S_{U_B})-\hat{H}_{B\to V}(\hat{S}_{U_B}))]^2\}\\
&\qquad+2w_Aw_BE_U\{[H_{A\to V}(S_{U_A})-\hat{H}_{A\to V}(\hat{S}_{U_A})]\\
&\qquad\times[H_{B\to V}(S_{U_B})-\hat{H}_{B\to V}(\hat{S}_{U_B})]\}\\
&\quad=w_A^2E_U\{[H_{A\to V}(S_{U_A})+H_{A\to V}(\hat{S}_{U_A})\\
&\qquad-H_{A\to V}(\hat{S}_{U_A})-\hat{H}_{A\to V}(\hat{S}_{U_A})]^2\}\\
&\qquad+w_B^2E_U\{[H_{B\to V}(S_{U_B})+H_{B\to V}(\hat{S}_{U_B})\\
&\qquad-H_{B\to V}(\hat{S}_{U_B})-\hat{H}_{B\to V}(\hat{S}_{U_B})]^2\}\\
&\qquad+2w_Aw_BE_U\{[H_{A\to V}(S_{U_A})+H_{A\to V}(\hat{S}_{U_A})\\
&\qquad-H_{A\to V}(\hat{S}_{U_A})-\hat{H}_{A\to V}(\hat{S}_{U_A})]\\
&\qquad\times[H_{B\to V}(S_{U_B})+H_{B\to V}(\hat{S}_{U_B})\\
&\qquad-H_{B\to V}(\hat{S}_{U_B})-\hat{H}_{B\to V}(\hat{S}_{U_B})]\},
\end{aligned}\tag{A1}$$

where, as defined in Section 2.2, $H_{A\to V}()$ and $H_{B\to V}()$ denote the mapping transforms without any depth loss from view A and view B to the synthesized view, respectively. $\hat{H}_{A\to V}()$ and $\hat{H}_{B\to V}()$ denote the mapping transforms with quantization-induced depth loss from view A and view B to the synthesized view, respectively.

In Eq. (A1), $H_{A\to V}(S_{U_A})-H_{A\to V}(\hat{S}_{U_A})$ and $H_{A\to V}(\hat{S}_{U_A})-\hat{H}_{A\to V}(\hat{S}_{U_A})$ are independent errors, and $H_{B\to V}(S_{U_B})-H_{B\to V}(\hat{S}_{U_B})$ and $H_{B\to V}(\hat{S}_{U_B})-\hat{H}_{B\to V}(\hat{S}_{U_B})$ are also uncorrelated. Also, $H_{A\to V}(S_{U_A})-H_{A\to V}(\hat{S}_{U_A})$ and $H_{B\to V}(\hat{S}_{U_B})-\hat{H}_{B\to V}(\hat{S}_{U_A})$ are uncorrelated. $H_{B\to V}(S_{U_B})-H_{B\to V}(\hat{S}_{U_B})$ and $H_{A\to V}(\hat{S}_{U_A})-\hat{H}_{A\to V}(\hat{S}_{U_A})$ are uncorrelated. As a result, Eq. (A1) can be rewritten as

$$\begin{aligned}
&E_U\{[\bar{S}_U-\hat{S}_U]^2\}\\
&\quad=w_A^2(E_U\{[H_{A\to V}(S_{U_A})-H_{A\to V}(\hat{S}_{U_A})]^2\}\\
&\qquad+E_U\{[H_{A\to V}(\hat{S}_{U_A})-\hat{H}_{A\to V}(\hat{S}_{U_A})]^2\})\\
&\qquad+w_B^2(E_U\{[H_{B\to V}(S_{U_B})-H_{B\to V}(\hat{S}_{U_B})]^2\}\\
&\qquad+E_U\{[H_{B\to V}(\hat{S}_{U_B})-\hat{H}_{B\to V}(\hat{S}_{U_B})]^2\})\\
&\qquad+2w_Aw_BE_U\{[H_{A\to V}(S_{U_A})-H_{A\to V}(\hat{S}_{U_A})]\\
&\qquad\times[H_{B\to V}(S_{U_B})-H_{B\to V}(\hat{S}_{U_B})]\}\\
&\qquad+2w_Aw_BE_U\{[H_{A\to V}(\hat{S}_{U_A})-\hat{H}_{A\to V}(\hat{S}_{U_A})]\\
&\qquad\times[H_{B\to V}(\hat{S}_{U_B})-\hat{H}_{B\to V}(\hat{S}_{U_B})]\}.
\end{aligned}\tag{A2}$$

In the actual view synthesis, for the general scene with little self-occluding objects, $\hat{n}_O/n_S$ generally less than 0.05 in small-baseline camera setup so that $\hat{n}_U/n_S\approx1$. Thus $E_U\{[H_{A\to V}(S_{U_A})-H_{A\to V}(\hat{S}_{U_A})]^2\}\approx E\{[S_A-\hat{S}_A]^2\}$, $E_U\{[H_{B\to V}(S_{U_B})-H_{B\to V}(\hat{S}_{U_B})]^2\}\approx E\{[S_B-\hat{S}_B]^2\}$, $E_U\{[H_{A\to V}(\hat{S}_{U_A})-\hat{H}_{A\to V}(\hat{S}_{U_A})]^2\}\approx E\{[H_{A\to V}(\hat{S}_A)-\hat{H}_{A\to V}(\hat{S}_A)]^2\}$ and $E_U\{[H_{B\to V}(\hat{S}_{U_B})-\hat{H}_{B\to V}(\hat{S}_{U_B})]^2\}\approx E\{[H_{B\to V}(\hat{S}_B)-\hat{H}_{B\to V}(\hat{S}_B)]^2\}$, we can get

$$\begin{aligned}
&E_U\{[\bar{S}_U-\hat{S}_U]^2\}\\
&\quad\approx w_A^2(E\{[H_{A\to V}(S_A)-H_{A\to V}(\hat{S}_A)]^2\}+E\{[H_{A\to V}(\hat{S}_A)-\hat{H}_{A\to V}(\hat{S}_A)]^2\})\\
&\quad+w_B^2(E\{[H_{B\to V}(S_B)-H_{B\to V}(\hat{S}_B)]^2\}+E\{[H_{B\to V}(\hat{S}_B)-\hat{H}_{B\to V}(\hat{S}_B)]^2\})\\
&\quad+2w_Aw_BE\{[H_{A\to V}(S_A)-H_{A\to V}(\hat{S}_A)]\cdot[H_{B\to V}(S_B)-H_{B\to V}(\hat{S}_B)]\}\\
&\quad+2w_Aw_BE\{[H_{A\to V}(\hat{S}_A)-\hat{H}_{A\to V}(\hat{S}_A)]\cdot[H_{B\to V}(\hat{S}_B)-\hat{H}_{B\to V}(\hat{S}_B)]\}.
\end{aligned}\tag{A3}$$

Appendix B. Proof of Eq. (7)

Assume that due to the constant motion error, the block signal f_x with Fourier spectrum $P_x(\boldsymbol{\omega})$ turns into f_y with Fourier spectrum $P_y(\boldsymbol{\omega})$. Since motion error

corresponds to linear phase shift in frequency domain, we have $P_y(\boldsymbol{\omega}) = P_x(\boldsymbol{\omega})e^{-j\boldsymbol{\omega}\Delta\mathbf{n}}$, where $\boldsymbol{\omega} = (\omega_1, \omega_2)$ is the frequency vector, and $\Delta\mathbf{n}$ the motion error vector. According to the Parseval's theorem, the total squared error D_S for f_x caused by motion error is given by

$$D_S = \frac{1}{(2\pi)^2} \iint_{(-\pi,\pi]} S_x(\boldsymbol{\omega}) \cdot |1 - e^{-j\boldsymbol{\omega}\Delta\mathbf{n}}|^2 d\omega_1 d\omega_2, \tag{B1}$$

where $S_x(\boldsymbol{\omega})$ is the energy density of f_x. Via the first-order Taylor series expansion approximation for $|1-e^{-j\boldsymbol{\omega}\Delta\mathbf{n}}|^2$ in $(\boldsymbol{\omega}\Delta\mathbf{n})^2$ yields

$$\begin{aligned} D_S &\approx \frac{1}{(2\pi)^2} \iint_{(-\pi,\pi]} S_x(\boldsymbol{\omega})(\boldsymbol{\omega}\Delta\mathbf{n})^2 d\omega_1 d\omega_2 \\ &\approx (\Delta x)^2\psi_1 + (\Delta y)^2\psi_2 + (\Delta x\Delta y)\psi_{1,2} \end{aligned} \tag{B2}$$

with $\psi_1 = (1/(2\pi)^2)\int\int_{(-\pi,\ \pi]} S_x(\omega_1, \omega_2)\cdot\omega_1^2 d\omega_1 d\omega_2$, $\psi_2 = (1/(2\pi)^2)\int\int_{(-\pi,\ \pi]} S_x(\omega_1, \omega_2)\cdot\omega_2^2 d\omega_1 d\omega_2$ and $\psi_{1,2} = (2/(2\pi)^2)\int\int_{(-\pi,\ \pi]} S_x(\omega_1, \omega_2)\cdot\omega_1\omega_2 d\omega_1 d\omega_2$.

Expressing $\Delta\mathbf{n}$ in polar coordinates, Eq. (B2) changes into $D_s = \|\Delta\mathbf{n}\|^2\psi_x(\theta_{\Delta\mathbf{n}})$, where $\psi_x(\theta_{\Delta\mathbf{n}}) = \psi_1\cos^2(\theta_{\Delta\mathbf{n}}) + \psi_2\sin^2(\theta_{\Delta\mathbf{n}}) + \psi_{1,2}\cos(\theta_{\Delta\mathbf{n}})\sin(\theta_{\Delta\mathbf{n}})$, and it represents the motion sensitivity at the orientation of $\theta_{\Delta\mathbf{n}}$. Since a natural image often exhibits isotropic power spectra, the sensitivity to motion error can be represented by the average sensitivity ψ_x over all motion error orientations. That is $\psi_x = (1/(2\pi))\int_{-\pi}^{\pi}\psi_x(\theta_{\Delta\mathbf{n}})d\theta_{\Delta\mathbf{n}} = (\psi_1+\psi_2)/2$. Therefore, for small and constant motion errors, we have

$$D_s = ||\Delta\mathbf{n}||^2\psi_x. \tag{B3}$$

References

[1] S. Würmlin, E. Lamboray, M. Gross, 3D video fragments: dynamic point samples for real-time free-viewpoint video, computers and graphics, Compression and Streaming Techniques for 3D and Multimedia Data (Special Issue on Coding) 28 (1) (2004) 3–14.

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

[3] A. Smolic, P. Kauff, Interactive 3-D video representation and coding technologies, Proceedings of the IEEE 93 (1) (2005) 98–110.

[4] C. Theobalt, G. Ziegler, M. Magnor, H.-P. Seidel, Model-based free-viewpoint video acquisition, rendering and encoding, In: Proceedings of Picture Coding Symposium, San Francisco, USA, December 2004, pp. 1–6.

[5] W. Matusik, C. Buehler, L. McMillan, Polyhedral visual hulls for real-time rendering, in: Proceedings of 12th Eurographics Workshop on Rendering, Eurographics Association, 2001, pp. 115–125.

[6] C.L. Zitnick, S.B. Kang, M. Uyttendaele, S. Winder, R. Szeliski, High-quality video view interpolation using a layered representation, In: ACM SIGGRAPH and ACM Transactions on Graphics, Los Angeles, CA, USA, August 2004.

[7] A. Smolic, K. Müller, K. Dix, P. Merkle, P. Kauff, T. Wiegand, Intermediate view interpolation based on multiview video plus depth for advanced 3D video systems, In: Proceedings of ICIP 2008, pp. 2448–2451.

[8] A. Vetro, S. Yea, A. Smolic, Towards a 3D video format for auto-stereoscopic displays, In: Proceedings of the SPIE: Applications of Digital Image Processing XXXI, San Diego, CA, USA, 2008.

[9] A. Bourge, C. Fehn, Committee Draft of ISO/IEC 23002–3 Auxiliary Video Data Representations. ISO/IEC JTC 1/SC 29/WG 11, Doc. N8038. Montreux, Switzerland, April 2006.

[10] Description of Exploration Experiments in 3D Video Coding. ISO/IEC JTC1/SC29/WG11, Doc. N9783, Archamps, France, May 2008.

[11] P. Kauff, N. Atzpadin, C. Fehn, M. Müller, O. Schreer, A. Smolic, R. Tanger, Depth map creation and image-based rendering for advanced 3DTV services providing interoperability and scalability, Signal Processing: Image Communication 22 (2) (2007) 217–234.

[12] P. Merkle, A. Smolic, K. Müller, T. Wiegand, Efficient prediction structures for multiview video coding, IEEE Transactions on Circuits and Systems for Video Technology, Special Issue on Multi-view Video Coding and 3DTV 17 (11) (2007) 1461–1473.

[13] S.-U. Yoon, Y.-S. Ho, Multiple color and depth video coding using a hierarchical representation, IEEE Transactions on Circuits and Systems for Video Technology, Special Issue on Multi-view Video Coding and 3DTV 17 (11) (2007) 1450–1460.

[14] J.H. Kim, P. Lai, J. Lopez, A. Ortega, Y. Su, P. Yin, C. Gomila, New coding tools for illumination and focus mismatch compensation in multiview video coding, IEEE Transactions on Circuits and Systems for Video Technology, Special Issue on Multi-view Video Coding and 3DTV 17 (11) (2007) 1519–1535.

[15] M. Zwicker, A. Vetro, S. Yea, W. Matusik, H. Pfister, Frédo Durand, Resampling, antialiasing, and compression in multiview 3-D displays, IEEE Signal Processing Magazine 24 (6) (2007) 88–96.

[16] X. Guo, Y. Lu, F. Wu, W. Gao, Inter-view direct mode for multiview video coding, IEEE Transactions on Circuits and Systems for Video Technology 16 (12) (2006) 527–1532.

[17] Y. Liu, Q. Huang, X. Ji, D. Zhao, W. Gao. Multi-view video coding with flexible view-temporal prediction structure for fast random access, In: Proceedings of Seventh Pacific-Rim Conference on Multimedia (PCM 2006), Hangzhou, China, pp. 564–571.

[18] N. Ozbek, M. Tekalp, Scalable multi-view video coding for interactive 3DTV, In: Proceedings of ICME 2006, pp. 213–216.

[19] P. Merkle, A. Smolic, K. Müller, T. Wiegand, Multi-view video plus depth representation and coding, In: Proceedings of ICIP 2007, pp. 201–204.

[20] S. Yea, A. Vetro, View synthesis prediction for multiview video coding, Signal Processing: Image Communication 24 (1) (2009) 89–100.

[21] Y. Morvan, D. Farin, P.H.N. de With, Depth-image compression based on an R-D optimized quadtree decomposition for the transmission of multiview images, In: Proceedings of ICIP 2007, pp. 105–108.

[22] M. Maitre, Y. Shinagawa, M.N. Do, Rate-distortion optimal depth maps in the wavelet domain for free-viewpoint rendering, In: Proceedings of ICIP 2007, pp. 125–128.

[23] P. Merkle, Y. Morvan, A. Smolic, D. Farin, K. Mueller, P.H.N. de With, T. Wiegand, The effect of depth compression on multiview rendering quality, Signal Processing: Image Communication 24 (1) (2009) 73–88.

[24] I. Daribo, C. Tillier, B. Pesquet-Popescu, Motion vector sharing and bit-rate allocation for 3D video-plus-depth coding, EURASIP Journal on Advances in Signal Processing 2009 (Special Issue on 3DTV) (2009) Article ID 258920.

[25] Y. Morvan, D. Farin, P.H.N. de With, Joint depth/texture bit-allocation for multi-view video compression, In: Proceedings of Picture Coding Symposium, Lisbon, Portugal, November 2007.

[26] ITU-R BT.1683, Objective perceptual video quality measurement techniques for standard definition digital broadcast television in the presence of a full reference. Geneva, June 2004.

[27] M. Tanimoto, T. Fujii, K. Suzuki, Experiment of view synthesis using multi-view depth. ISO/IEC JTC1/SC29/WG11, Doc. M14889, October 2007.

[28] R. Hartley, A. Zisserman, Multiple View Geometry in Computer Vision, Cambridge University Press, UK, 2003.

[29] A. Secker, D. Taubman, Highly scalable video compression with scalable motion coding, IEEE Transactions on Image Processing 13 (8) (2004) 1029–1041.

[30] I. Feldmann, M. Mueller, F. Zilly, R. Tanger, K. Muller, A. Smolic, P. Kauff, T. Wiegand, HHI Test Material for 3D Video. ISO/IEC JTC 1/SC 29/WG 11, Doc. N15413. Archamps, France, April 2008.

[31] S. Ma, W. Gao, Y. Lu, Rate-distortion analysis for H.264/AVC video coding and its application to rate control, IEEE Transactions on Circuits and Systems for Video Technology 15 (12) (2005) 1533–1544.

[32] J. Proakis, D. Manolakis, Digital Signal Processing: Principles, Algorithms and Applications, Prentice-Hall, Englewood Cliffs, NJ, 1995.