Scalable Bit Allocation Between Texture and Depth Views for 3-D Video Streaming Over Heterogeneous Networks

Jimin Xiao, Miska M. Hannuksela, *Member, IEEE*, Tammam Tillo, *Senior Member, IEEE*, Moncef Gabbouj, *Fellow, IEEE*, Ce Zhu, *Senior Member, IEEE*, and Yao Zhao, *Senior Member, IEEE*

***Abstract*— In the multiview video plus depth (MVD) coding format, both texture and depth views are jointly compressed to represent the 3-D video content. The MVD format enables synthesis of virtual views through depth-image-based rendering; hence, distortion in the texture and depth views affects the quality of the synthesized virtual views. Bit allocation between texture and depth views has been studied with some promising results. However, to the best of our knowledge, most of the existing bit-allocation methods attempt to allocate a fixed amount of total bit rate between texture and depth views; that is, to select appropriate pair of quantization parameters for texture and depth views to maximize the synthesized view quality subject to a fixed total bit rate. In this paper we propose a scalable bit-allocation scheme, where a single ordering of texture and depth packets is derived and used to obtain optimal bit allocation between texture and depth views for any total target rates. In the proposed scheme, both texture and depth views are encoded using the quality scalable coding method; that is, medium grain scalable (MGS) coding of the Scalable Video Coding (SVC) extension of the Advanced Video Coding (H.264/AVC) standard. For varying target total bit rates, optimal bit truncation points for both texture and depth views can be obtained using the proposed scheme. Moreover, we propose to order the enhancement layer packets of the H.264/SVC MGS encoded depth view according to their contribution to the reduction of the synthesized view distortion. On one hand, this improves the depth view packet ordering when considered the rate-distortion performance of synthesized views, which is demonstrated by the experimental results. On the other hand, the information obtained in this step is used to facilitate optimal bit allocation between texture and depth views. Experimental results demonstrate the effectiveness of the proposed scalable bit-allocation scheme for texture and depth views.**

***Index Terms*— 3-D, 3-D scalability, 3-D streaming, bit allocation, depth view, heterogeneous network, medium grain scalable (MGS), quality scalable, synthesized view distortion, texture view.**

I. Introduction

THREE-DIMENSIONAL (3-D) video has drawn a lot of attention both from industry and academia during the last decade. Stereoscopic video is the basic form of 3-D video, and is prevalent in today's 3-D content and services. Autostereoscopic 3-D displays [1], [2] enable viewing 3-D content from different angles without the use of special headgear or glasses. As many different viewpoints of the same video content are required for autostereoscopic 3-D system, compared with the traditional stereo video, the required bit rate increases tremendously.

To further reduce the redundancy between different viewpoints of the 3-D video, besides the conventional temporal prediction which is commonly used in video compression, inter-view prediction [3] is also introduced in the Multiview Video Coding (MVC) [4] extension of the Advanced Video Coding (H.264/AVC) standard [5]. Though MVC has enormously improved the compression performance of multiview video, it still requires a bit rate that is proportional to the number of views [4]. The multiview video plus depth (MVD) format is a promising way to represent 3-D video content, and extensions supporting for the MVD format have been finished recently [6], [7]. With the MVD format, only a small number of texture views associated with their depth views are required. At the decoder or display side, depth-image-based rendering (DIBR) [8], [9] is used to synthesize additional viewpoint video.

This paper focuses on adaptive streaming of multiview video plus depth, where we have two main options [10].

1) *Simulcast Encoding:* encode each view and/or depth view independently using a scalable or nonscalable monocular video codec, which enables streaming each view over separate channels; and clients can request as many views as their 3-D displays require without worrying about inter-view dependencies.

Manuscript received February 1, 2014; revised April 26, 2014; accepted June 23, 2014. Date of publication June 27, 2014; date of current version January 5, 2015. This work was supported in part by the 973 Program under Grant 2011CB302204; in part by the National Natural Science Foundation of China under Grants 60972085, 61025013, 61125106, and 61228102; in part by the Program for Changjiang Scholars and Innovative Research Team in University under Grant IRT 201206; and in part by the Ph.D. Programs Foundation, Ministry of Education of China, under Grant 20130182110010. This paper was recommended by Associate Editor S. Shirani. *(Corresponding author: Ce Zhu.)*

J. Xiao and M. Gabbouj are with Tampere University of Technology, Tampere 33720, Finland (e-mail: xiaojimin1981@gmail.com; moncef.gabbouj@tut.fi).

M. M. Hannuksela is with Nokia Research Center, Tampere 33720, Finland (e-mail: miska.hannuksela@nokia.com).

T. Tillo is with the Department of Electrical and Electronic Engineering, Xi'an Jiaotong-Liverpool University, Suzhou 215123, China (e-mail: tammam.tillo@xjtlu.edu.cn).

C. Zhu is with the School of Electronic Engineering, University of Electronic Science and Technology of China, Chengdu 610051, China (e-mail: eczhu@uestc.edu.cn).

Y. Zhao is with Institute of Information Science, Beijing Jiaotong University, Beijing 100055, China (e-mail: yzhao@bjtu.edu.cn).

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

Digital Object Identifier 10.1109/TCSVT.2014.2334011

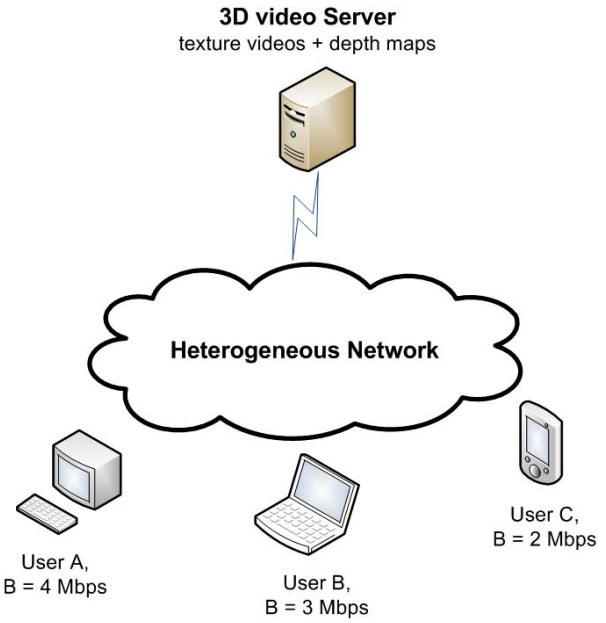


Fig. 1. 3-D video streaming over heterogeneous networks, different links have different bandwidth.

2) *Dependent Encoding:* encode views using MVC to decrease the overall bit rate by exploiting the inter-view redundancies. Simulcast encoding can be regarded as a more flexible approach than dependent encoding. For example, simulcast encoding enables a client to select the viewpoint dynamically out of a large number of views available in the server, whereas inter-view dependencies would either cause limitations on viewpoint selection or streaming of a larger number of views than necessary.

Simulcast coding using the H.264 Scalable Video Coding (SVC) [11] standard can produce scalable 3-D video, where each view is encoded independently. Here, two approaches can be followed for scalability: 1) either all views can be coded scalably or 2) some views can be coded scalably using SVC and others can be coded nonscalably using H.264/AVC. On the other hand, encoding views using MVC decreases the overall bit rate for the multiview video plus depth format. However, MVC offers only temporal and view scalability, but no quality or resolution scalability [10].

In the MVD format, the depth views are always used together with the associated texture views, and both the texture and depth views quality will affect the synthesized view quality. Thus, the allocation of bit rate between the texture and depth views is an important issue. During the last decade, many bit-allocation methods have been proposed [12]–[19]. However, these methods work well when the total target bit rate for the texture and depth views is fixed. However, for the 3-D video server, which is serving for many users with different link bandwidth in heterogeneous networks, the existing bit-allocation methods cannot work well. For example, in the application scenario shown in Fig. 1, to have proper allocation between texture and depth views, the 3-D video server has to allocate bit rate for the target rate 2, 3, and 4 Mb/s; the resulting three different versions of the encoded texture and depth views should be all stored in the server. It is worth noticing that, for practical applications, the link bandwidth over heterogeneous networks is varying; hence, more than 3 versions of texture and depth pairs should be stored in the server. This will cause two fundamental problems: one is that the bit-allocation algorithm should be carried out many times for varying target rates, which requires huge computational resources; another is that many different encoded versions for varying target rates should be stored in the 3-D video server, so the requirement for storage is tremendous.

In this paper, we propose a scalable bit-allocation scheme between texture and depth views, which can solve the preceding two fundamental problems. In this scheme, both texture and depth views are encoded using the quality Scalable Video Coding method; that is, H.264/SVC medium grain scalable (MGS) coding. For varying target rates, the optimal bit truncation points for both texture and depth views can be obtained using the proposed scheme. The contribution of this paper is manifold. Firstly, we propose to order the MGS enhancement layer packets (NAL units) based on their contribution to the reduction of the synthesized view distortion. Secondly, the information obtained in the depth view enhancement layer packet ordering step, that is, the synthesized view distortion reduction of each MGS enhancement layer packet, is used to facilitate optimal bit allocation between texture and depth views. Therefore, the optimal bit allocation can be obtained using one simple formula for varying total target rates, with negligible computational complexity. To the best of authors' knowledge, this is the first time to propose such a scalable bit-allocation scheme between texture and depth views. In addition, we studied how the number of synthesized views affects the bit allocation between texture and depth views by analytical and experimental methods; and it is interesting to find that with the increase of the number of synthesized views, higher ratio of bit should be allocated for the texture views.

The rest of this paper is organized as follows. Related works are reviewed in Section II. A brief description of the synthesized view distortion is provided in Section III. In Section IV, the proposed scalable bit-allocation scheme is presented in details. In Section V, some experimental results validating the proposed approach are given. Finally, the conclusion is drawn in Section VI.

II. Related Work

A. Depth Coding and Bit Allocation

During the last decade, many depth compression techniques have been proposed to improve coding performance by exploiting the depth characteristics. One important technique uses the synthesized view distortion metric in the rate-distortion optimization process. In [20], motivated by the fact that the depth views are not perceived by the viewers but only supplement data for view synthesis, instead of using the distortion of the depth view itself, the authors proposed to use the distortion of the synthesized view in the rate-distortion optimized depth coding mode selection step; later in [21], the

synthesized view distortion was modeled at pixel-level and in a more accurate way, which eventually led to better overall rate-distortion performance than that of [20].

Prior to the compression technique of the depth view itself, one more fundamental problem that needs to be addressed is how to allocate bit rate between the texture and depth views. A heuristic approach with fixed ratio (5:1) bit allocation between texture and depth views was used in [9]. Later, Morvan *et al.* [12] proposed a Full-search algorithm to find the optimal quantization parameter (QP) pair for texture and depth views. This algorithm assumes that a real view exists at the synthesized viewpoint, which is not always true; its tremendous computational complexity is another major problem. Liu *et al.* [13] proposed a distortion model to estimate the distortion of the synthesized views without the need of comparing the synthesized view with its corresponding real view. A fast bit-allocation algorithm was proposed in [14] to reduce the complexity, where the allocation performance is comparable with that of [13]. In [15], a region-based view synthesis distortion estimation approach and a general R-D property estimation model was proposed. The reported results in [15] show that it can provide better R-D performance than [13] with lower computational cost. Recent advances on texture and depth bit-allocation algorithms were presented in [16]–[19]. In [16] and [17], the concept of rate control was introduced into optimal bit-allocation paradigm, whereas in [19] the allocation was carried out at macroblock granularity, which led to better rate-distortion performance than full-search algorithm [12].

B. Scalable Coding

The H.264/SVC [11], which is an annex of the H.264/AVC standard [5], provides spatial, temporal, and quality scalability. SVC provides temporal scalability through the usage of hierarchical prediction structures, whereas spatial and quality scalability are supported by multilayer coding. Quality scalability is supported in two modes: coarse-grained scalability (CGS) and MGS. When CGS is used, rate adaptation has to be performed on complete layer basis. However, MGS concept allows any enhancement layer network abstraction layer (NAL) unit to be discarded from a quality scalable bit stream in decreasing quality_id order, providing packet-based scalability. MGS is particularly useful when the server wishes to match the transmitted bit rate with the currently prevailing network throughput for a client to minimize the end-to-end delay. Conventionally, when the transmitted bit rate is not accurately matched with the prevailing throughput, clients have to buffer more data to compensate situations where the reception rate is temporarily lower than the media playout rate.

MGS splits a given enhancement layer of a given video frame into up to 16 MGS layers (also referred to as quality layers) [40]. In particular, MGS divides the transform coefficients, obtained through transform coding of a given block, into multiple groups. Each group is assigned to one MGS layer. For example, let us consider a 4×4 block, and use w_i to denote the number of transform coefficients belonging to MGS layer i within an enhancement layer, with $\sum_{i=1}^{16} w_i = 16$. The number of transform coefficients is also referred to as the weight of MGS layer. An MGS encoding can be represented by giving the weights in the vector form $W = [w_1, w_2, w_3, \ldots, w_{16}]$, whereby $w_i = 0$ if it is not specified. Fig. 2 shows the splitting of the transform coefficients of a 4×4 block into three MGS layers with the weights $W = [3, 3, 10]$, that is, $w_1 = 3$, $w_2 = 3$ and $w_3 = 10$, while other weights being 0.

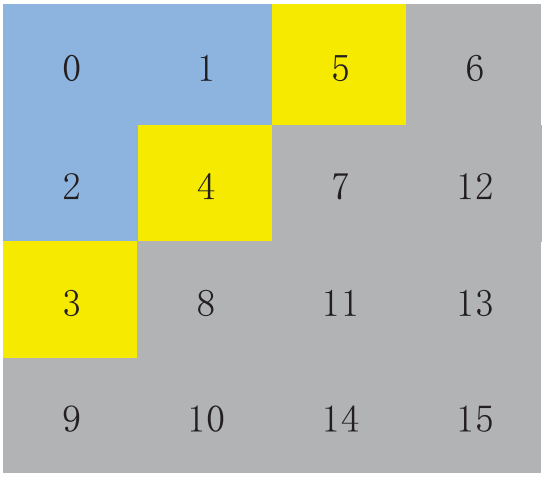


Fig. 2. Allocation of transform coefficients of a 4×4 block to MGS layers for weight vector $W = [3, 3, 10]$. MGS layer 1 includes coefficients with indices 0–2, MGS layer 2 includes coefficients with indices 3–5 and MGS layer 3 includes coefficients with indices 6–15.

Some pioneer works on adaptive scalable 3-D video coding are [22] and [23]. Chakareski *et al.* [22] proposed an optimization framework for joint view and rate scalable coding of multiview video content represented in the texture plus depth format. However, the view and rate embedded bitstream can only be constructed for a discrete set of transmission rates. In [23], the authors proposed a novel compression strategy for depth views that incorporates geometry information while achieving the goals of scalability and embedded representation. In this paper, the texture views were not jointly considered for 3-D video scalability.

C. 3-D Hole Filling and Quality Assessment

A critical problem in the DIBR system is how to deal with the hole regions after 3-D projection. Generally, the holes are generated because the occluded regions in the original view become visible in the virtual view after 3-D projection. There are many solutions to address the disocclusion problem. One type of solutions preprocesses the depth views before DIBR, aiming to reduce the depth value difference along the boundary between the foreground and background, so that no disocclusion appears in the virtual view. Solutions of this type include using a symmetric Gaussian filter [24] and an asymmetric filter [25] to smooth the depth view. Another type of solutions is using background information for hole filling. In [26], a background sprite is generated using the original texture and synthesized images from the temporally previous frames for disocclusion filling. In our recent work [27], the Gaussian mixture model is used to generate a stable

background for hole filling. One commonly used hole filling solution is view synthesis using the MVD video format. This approach exploits the fact that the invisible background part in the left view may be visible in the right view [28], and then the disoccluded regions in the virtual view warped from the left view can be filled with the background information from the right view, and vice versa. In the proposed scheme, MVD video format is used, so most of the disoccluded regions can be filled in complementary way using both the left and right views.

2-D image/video quality metrics have been widely researched [29], and many quality assessment metrics, such as peak signal-to-noise ratio (PSNR), Structural SIMilarity [30], visual information fidelity (VIF) [31], have been proposed. Most of the 3-D video quality assessment work in the literature is based on applying 2-D video quality measures on the depth view as well as the stereoscopic views and then finding the combination of these measures that best correlates with the subjective scores [33]–[35]. In [36], 3VQM was proposed for objectively evaluating the quality of stereoscopic 3-D videos generated by DIBR. In this method, the authors tried to derive an ideal depth estimation. Three measures, namely temporal outliers (TO), temporal inconsistencies (TI), and spatial outliers (SO) are used to constitute a vision-based quality measure for 3-D DIBR-based videos. Subjective quality assessment and objective quality measurement of 2-D video are mature fields. However, subjective assessment of 3-D video quality is still facing many problems to solve before the performance of 3-D video models can be properly evaluated to capture the essential QoE involved by such media [32]. Moreover, to the best of our knowledge, none of the 3-D quality measures is commonly recognized to be superior from others. Thus, JCT-3V common test conditions [37] still use PSNR as the only metric to evaluate the synthesized view quality in the 3-D video coding standardization process. Hence, based on this, in this paper we will use PSNR for our measures.

III. Distortion Model for Synthesized View

In the proposed scheme, for both depth coding and scalable bit allocation between the texture and depth views, the distortion model for the synthesized view is required. The synthesized view distortion will be estimated without comparing the virtual view with its corresponding real view, because in practical applications the existence of the real view is not guaranteed. The synthesized view distortion model presented in [21], which will be used in this paper, is reviewed in Section III-A. In this model, the distortion is modeled at pixel level, and it mimics the view synthesizing process with subpixel interpolation. It is worth noting that the model in [21] has been adopted in JCT-3V. However, this model is based on the assumption that the virtual view is generated using one reference view. So in Section III-B, this model is extended for MVD video format, where the virtual view is merged from wrapped views of both left and right reference views, and more than one virtual view is generated. Throughout this paper, subscript t and d indicate texture and depth information, respectively; l and r represent the left and right view, respectively.

A. Synthesized View Distortion Using One Reference View

In this section, our analysis will be based on the assumption that the virtual view is generated using one reference view. The distortion of the synthesized view will be the sum of squared distance (SSD) between two versions of the synthesized view. The first version, denoted by $V_{x',y'}$, is synthesized from the original texture and depth views, whereas the other one is generated from the compressed version of the decoded texture and depth views, denoted by $\tilde{V}_{x',y'}$. The SSD in this case is

$$\begin{aligned} SSD &= \sum_{(x',y')} |V_{x',y'} - \tilde{V}_{x',y'}|^2 \\ &= \sum_{(x,y)} |f_w(C, D_{x,y}) - f_w(\tilde{C}, \tilde{D}_{x,y})|^2 \end{aligned} \tag{1}$$

where C and D are the original texture view and depth view, respectively, $\tilde{C}$ and $\tilde{D}$ are the decoded texture and depth view, respectively, (x', y') is yhe warped pixel position for the synthesized view V corresponding to (x, y) in C and D by the predefined warping function, f_w, and (x, y) is the pixel inside the current nonsynthesized block. As in [21], (1) can be further simplified as

$$\text{SSD} = E_t + E_d(s) \tag{2}$$

with $E_t = \sum_{(x,y)} |f_w(C, D_{x,y}) - f_w(\tilde{C}, D_{x,y})|^2$, is the distortion caused by the compression of the texture view, and $E_d(s) = \sum_{(x,y)} |f_w(\tilde{C}, D_{x,y}) - f_w(\tilde{C}, \tilde{D}_{x,y})|^2$, is the distortion caused by the compression of the depth view. The parameter s in $E_d(s)$ is the distance between the current and the rendered view which will affect the value of $E_d(s)$.

In the 1-D parallel camera setting configuration, the 3-D configuration used in this paper, the synthesized view distortion caused by the depth view $E_d(s)$ can be further approximated as [21]

$$E_d(s) \approx \sum_{(x,y)} |\tilde{C}_{x,y} - \tilde{C}_{x-\Delta p(x,y,s),y}|^2 \tag{3}$$

where Δp is the translational horizontal rendering position error. It is already proven that it is proportional to depth view error

$$\Delta p(x, y, s) = \alpha(s) \cdot (D_{x,y} - \tilde{D}_{x,y}) \tag{4}$$

where $\alpha(s)$ is a proportional coefficient determined by

$$\alpha(s) = \frac{f \cdot s}{255} \left(\frac{1}{Z_{\text{near}}} - \frac{1}{Z_{\text{far}}} \right) \tag{5}$$

with f being the focal length, Z_{near} and Z_{far} being the values of the nearest and farthest depth of the scene, respectively. The value of $E_d(s)$ can be approximated as (6), as shown at the top of the page. Finally, (6) is modified to use the original texture when the reconstructed texture is unavailable as (7), as shown at the top of the next page. This is because in MVD video coding process, some of the depth views may be encoded before the associated texture views [38]. During encoding the depth views, the information of the

$$E_d(s) = \sum_{(x,y)} \left[\frac{|\Delta p(x,y,s)|}{2} \left(|\tilde{C}_{x,y} - \tilde{C}_{x-1,y}| + |\tilde{C}_{x,y} - \tilde{C}_{x+1,y}| \right) \right]^2$$
$$= \frac{s^2 f^2}{255^2} \left(\frac{1}{Z_{\text{near}}} - \frac{1}{Z_{\text{far}}} \right)^2 \sum_{(x,y)} \left[(D_{x,y} - \tilde{D}_{x,y}) \left(|\tilde{C}_{x,y} - \tilde{C}_{x-1,y}| + |\tilde{C}_{x,y} - \tilde{C}_{x+1,y}| \right) \right]^2 \quad (6)$$

$$E_d(s) \approx \frac{s^2 f^2}{255^2} \left(\frac{1}{Z_{\text{near}}} - \frac{1}{Z_{\text{far}}} \right)^2 \sum_{(x,y)} \left[(D_{x,y} - \tilde{D}_{x,y}) \left(|C_{x,y} - C_{x-1,y}| + |C_{x,y} - C_{x+1,y}| \right) \right]^2$$
$$= \frac{s^2 f^2}{255^2} \left(\frac{1}{Z_{\text{near}}} - \frac{1}{Z_{\text{far}}} \right)^2 \Psi. \quad (7)$$

reconstructed texture values may not be available. In (7), the value of

$$\sum_{(x,y)} \left[(D_{x,y} - \tilde{D}_{x,y}) \left(|C_{x,y} - C_{x-1,y}| + |C_{x,y} - C_{x+1,y}| \right) \right]^2$$

is denoted as Ψ for simplicity. From (2) and (7), it is important to note that the texture view distortion will be directly copied to the synthesized view distortion, while the texture view characteristics will also affect how the depth view distortion affects the synthesized view distortion.

B. Synthesized View Distortion for MVD Video

In the MVD video format, both the left and right views are used together to synthesize the virtual views, and the synthesized virtual view number could be more than one. In this paper, we extended the synthesized view distortion model [21] to MVD video, where the synthesized view is generated by merging the wrapped views from the left and right views, and the total distortion of more than one synthesized view is accounted in the framework. We assume that the final virtual view is generated by merging the wrapped views from the left and right views using linear combination. Hence, the merged virtual view can be represented as

$$\tilde{V}^m_{x,y} = \zeta \tilde{V}^l_{x,y} + (1-\zeta)\tilde{V}^r_{x,y} \quad (8)$$

where $\tilde{V}^l_{x,y}$ and $\tilde{V}^r_{x,y}$ are wrapped views from the left and right compressed texture and depth views, respectively, subscript l and r denote left and right, respectively. The weight ζ is determined by the distances between the virtual camera position and the left/right reference camera positions. Let us use d_l to denote the distance between the virtual view and the left view, d_r to denote the distance between the virtual view and the right view, in this case $\zeta = (d_r/d_l + d_r)$. The merged view distortion, denoted as $\text{SSD}^m_V(\zeta)$, could be evaluated as (9), as shown at the top of the next page, where $V^l_{x,y}$ and $V^r_{x,y}$ are the wrapped views from the left and right original texture and depth views, respectively. Normally, the compression distortion is regarded as white noise [39]. Thus, it is reasonable to assume that $(\tilde{V}^l_{x,y} - V^l_{x,y})$ and $(\tilde{V}^r_{x,y} - V^r_{x,y})$ are uncorrelated. This term in (10) can be ignored, and the merged view distortion could be evaluated as

$$\text{SSD}^m_V(\zeta) = \zeta^2 \left(\tilde{V}^l_{x,y} - V^l_{x,y} \right)^2 + (1-\zeta)^2 \left(\tilde{V}^r_{x,y} - V^r_{x,y} \right)^2$$
$$= \zeta^2 \text{SSD}^l_V + (1-\zeta)^2 \text{SSD}^r_V \quad (10)$$

where SSD^l_V and SSD^r_V are the distortion of the wrapped views from the left and right reference views, respectively, and they can be evaluated using (2).

At this stage, let us assume N total virtual views are generated between the left and right views. The N virtual views are evenly distributed, which means the distance between two neighboring virtual views is $(1/N+1)L$ (L is the distance between the left and right views). From left to right, the index of the virtual views are 1, 2, 3, ..., N. For the ith virtual view, $\zeta = (N+1-i/N+1)$; the left view caused distortion, $\text{SSD}^l_V(i)$, can be evaluated using (2) as (11), as shown at the top of the next page. Hence, for the N virtual views, the total distortion caused by the left view, SSD^l_{TV}, can be evaluated as (12), as shown at the top of the next page. Among the distortion SSD^l_{TV}, the left texture-view-caused distortion in N virtual views, E^l_t, could be evaluated as

$$E^l_t = \sum_{i=1}^{N} \left(\frac{N+1-i}{N+1} \right)^2 E_t. \quad (13)$$

The left depth view caused distortion in N virtual views, E^l_d, could be evaluated as

$$E^l_d = \sum_{i=1}^{N} \left(\frac{N+1-i}{N+1} \right)^2 \left(\frac{iL}{N+1} \right)^2 \frac{f^2}{255^2} \left(\frac{1}{Z_{\text{near}}} - \frac{1}{Z_{\text{far}}} \right)^2 \Psi. \quad (14)$$

Let us assume

$$\mu = \frac{\sum_{i=1}^{N} \left(\frac{N+1-i}{N+1} \right)^2}{\sum_{i=1}^{N} \left(\frac{N+1-i}{N+1} \right)^2 \left(\frac{i}{N+1} \right)^2}.$$

It is interesting to note that for $N = \{1, 3, 7\}$, μ will be $\{4, 6.58, 8.20\}$. This means that with the increase of the number of synthesized views, the ratio of texture-view-caused distortion over depth-caused distortion is increasing. This observation also indicates that with the increase of the number of synthesized views, more bit should be allocated for the texture views. This interesting observation is also supported in Section III. Similarly, for the N virtual views, the total distortion caused by the right view, SSD^r_{TV}, the right texture (depth) view caused distortion, E^r_t (E^r_d) can be evaluated with the same methods. If the left and right views are encoded using

$$\begin{aligned}\mathrm{SSD}_V^m(\zeta) &= \left(\zeta\,\tilde{V}_{x,y}^l + (1-\zeta)\,\tilde{V}_{x,y}^r - \left(\zeta\,V_{x,y}^l + (1-\zeta)\,V_{x,y}^r\right)\right)^2 \\ &= \zeta^2\left(\tilde{V}_{x,y}^l - V_{x,y}^l\right)^2 + (1-\zeta)^2\left(\tilde{V}_{x,y}^r - V_{x,y}^r\right)^2 + 2\zeta\,(1-\zeta)\left(\tilde{V}_{x,y}^l - V_{x,y}^l\right)\left(\tilde{V}_{x,y}^r - V_{x,y}^r\right) \quad (9)\end{aligned}$$

$$\begin{aligned}\mathrm{SSD}_V^l(i) &= \left(\frac{N+1-i}{N+1}\right)^2 \mathrm{SSD}_V^l = \left(\frac{N+1-i}{N+1}\right)^2\left(E_t + E_d\left(\frac{iL}{N+1}\right)\right) \\ &= \left(\frac{N+1-i}{N+1}\right)^2 E_t + \left(\frac{N+1-i}{N+1}\right)^2\left(\frac{iL}{N+1}\right)^2 \frac{f^2}{255^2}\left(\frac{1}{Z_{\mathrm{near}}} - \frac{1}{Z_{\mathrm{far}}}\right)^2 \Psi \quad (11)\end{aligned}$$

$$\begin{aligned}\mathrm{SSD}_{\mathrm{TV}}^l &= \sum_{i=1}^{N} \mathrm{SSD}_V^l(i) \\ &= \sum_{i=1}^{N}\left(\frac{N+1-i}{N+1}\right)^2 E_t + \sum_{i=1}^{N}\left(\frac{N+1-i}{N+1}\right)^2\left(\frac{iL}{N+1}\right)^2 \frac{f^2}{255^2}\left(\frac{1}{Z_{\mathrm{near}}} - \frac{1}{Z_{\mathrm{far}}}\right)^2 \Psi \quad (12)\end{aligned}$$

the same coding parameters, which is the coding scheme used in this paper, $E_t^r \approx E_t^l$, $E_d^r \approx E_d^l$ and $\mathrm{SSD}_{\mathrm{TV}}^r \approx \mathrm{SSD}_{\mathrm{TV}}^l$. This is because the two views typically share similar characteristics.

IV. Proposed Scalable Bit-Allocation Scheme

In the proposed scalable bit-allocation scheme, both texture and depth views are encoded using quality scalable coding method, H.264/SVC MGS [11] coding. In the current system, for simplicity, two viewpoints both including texture and depth views are used to represent 3-D video content. However, it is straightforward to extend the system to more than two views. The H.264/SVC MGS encoded texture and depth views are stored in the 3-D video server, including all the MGS base layer and enhancement layer packets [as shown in Fig. 1]. A user in the heterogeneous network can start 3-D video streaming service by informing the server its own link bandwidth, B. Meanwhile, some auxiliary information is also provided to the server, such as the number of virtual views that are going to be synthesized at the receiver side and the position of each virtual view. These auxiliary information will affect the bit rate allocation process. Upon receiving the video streaming request, the 3-D video server will decide the optimal MGS enhancement layer truncation points for both the texture and depth views, so that the total bit rate is within the user link bandwidth, B.

The procedure of the proposed scalable bit-allocation scheme works as follows.

1) Both texture and depth views are encoded using H.264/SVC MGS with two layers, that is, MGS base layer and enhancement layer. The base layer and enhancement layer QP pairs are $\{QP_b^t, QP_e^t\}$, $\{QP_b^d, QP_e^d\}$ for texture and depth views, respectively. Here subscript b and e denote base and enhancement layer, respectively. To improve the depth coding performance, in the rate-distortion optimized coding mode selection step, the synthesized view distortion metric (14) is used for both the MGS base layer and enhancement layer coding.
2) For both the texture and depth views, the MGS enhancement layer packets (NAL units) are ordered based on their contribution to the synthesized view distortion reduction to the whole GOP owing to the drift distortion [41], [42]. For the texture and depth view, each enhancement layer packet's contribution to the synthesized view distortion reduction to the current frame needs to be evaluated using (13) and (14), respectively. The synthesized view distortion reduction to the whole GOP is evaluated using the distortion drift model implemented in JSVM [41], [42]. According to (13), the synthesized view distortion is proportional to the texture view distortion. Thus, the enhancement layer packets of the texture view are ordered using existing method in H.264/SVC reference software, Joint Scalable video Model (JSVM) [43]. While for the depth view, the synthesized view distortion is not proportional to the depth view distortion. Thus, the enhancement layer packets of depth view should be ordered based on their contribution to the reduction of the synthesized view distortion to the GOP. The enhancement packet ordering information can be conveyed in the NAL unit header, through the syntax element priority_id, or using an optional supplemental enhancement information (SEI) message. Meanwhile, each texture (depth) view enhancement packet's size and its contribution to the synthesized view distortion reduction to the whole GOP is recorded as $\{R_t^i, D_t^i\}$ $\left(\{R_d^i, D_d^i\}\right)$ $(i \geq 1)$ for the ith enhancement packet, where the enhancement packet index i is obtained after the MGS enhancement layer packet ordering. The use of $\{R_t^i, D_t^i\}$ and $\{R_d^i, D_d^i\}$ will be described in the following steps.
3) The 3-D video server will decide the optimal MGS enhancement layer truncation points for both the texture and depth views.
4) The truncated texture and depth views, which suits for the user's link bandwidth, will be sent to the user. The user equipment can synthesize virtual views.

A. Optimal Enhancement Layer Truncation Point Selection

It is important to select proper MGS enhancement layer truncation points for both texture and depth views, which

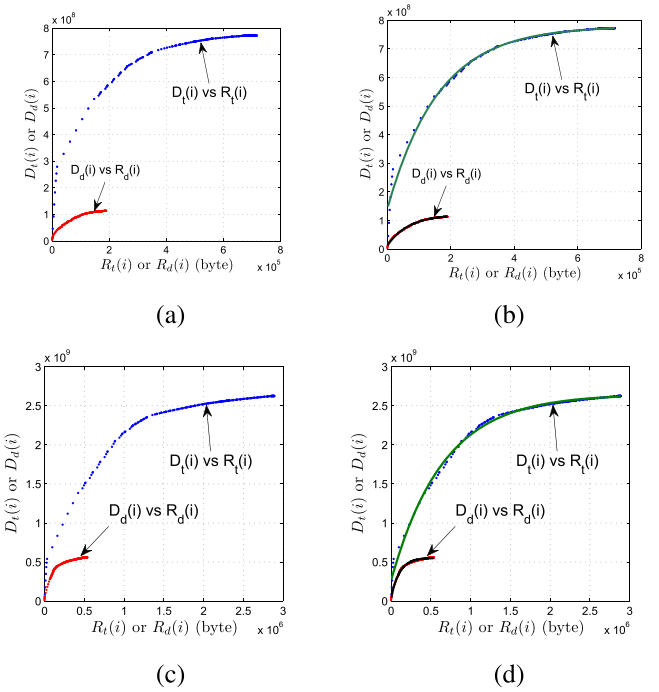


Fig. 3. Synthesized view distortion reduction versus enhancement layer packet rate for *Balloons* and *Poznan_Street* sequence. (a) and (c) $D_t(i)$ versus $R_t(i)$ and $D_d(i)$ versus $R_d(i)$. (b) and (d) Curve fitting results (solid curves) using exponential functions (15) and (16). (a) *Balloons*. (b) *Balloons*. (c) *Poznan_Street*. (d) *Poznan_Street*.

affects the bit-allocation performance. The detailed algorithm of selecting optimal truncation points is described as follows.

In the previous steps, we have got each enhancement packet's rate and contribution to the reduction of the synthesized view distortion to the GOP: $\{R_t^i, D_t^i\}$ for the texture view and $\{R_d^i, D_d^i\}$ for the depth view. For the texture view, let us use $R_t(i) = \sum_{n=1}^{i} R_t^n$ and $D_t(i) = \sum_{n=1}^{i} D_t^n$ to denote the accumulated MGS enhancement rate and synthesized view distortion reduction, respectively. Similarly, for the depth views, let us use $R_d(i) = \sum_{n=1}^{i} R_d^n$ and $D_d(i) = \sum_{n=1}^{i} D_d^n$, to denote the accumulated MGS enhancement rate and synthesized view distortion reduction, respectively. To have a better understanding of the values, in Fig. 3(a) and (c), we have reported $D_t(i)$ versus $R_t(i)$, and $D_d(i)$ versus $R_d(i)$ for the *Balloons* and *Poznan_Street* sequences. It is very interesting to note that, for the texture view, how $D_t(i)$ changes with $R_t(i)$ could be closely approximated by

$$D_t(R_t) = a_t \cdot \left(1 - e^{-\frac{R_t}{b_t}}\right) + c_t \tag{15}$$

with $a_t > 0$ and $b_t > 0$; here D_t and R_t are the short form of $D_t(i)$ and $R_t(i)$, respectively. Similarly, for the depth views, how $D_d(i)$ changes with $R_d(i)$ could be approximated by

$$D_d(R_d) = a_d \cdot \left(1 - e^{-\frac{R_d}{b_d}}\right) + c_d \tag{16}$$

with $a_d > 0$ and $b_d > 0$; D_d and R_d are the short form of $D_d(i)$ and $R_d(i)$, respectively. The parameters $\{a_t, b_t, c_t\}$ and $\{a_d, b_d, c_d\}$ could been obtained with least square curve fitting [44]. It is worth noticing that only the curve fitting results for *Balloons* and *Poznan_Street* sequences are reported herein. Nevertheless, for other sequences, for example, *Newspaper*, *Lovebird1*, *BookArrival*, *Poznan_Hall2*, similar results have been obtained. It is important to note that one synthesized virtual view is generated in Fig. 3. Nevertheless, the number of synthesized views, N, will not affect this finding, because according to (13) and (14) N will only change the scaling factor. To the best of our knowledge, this is the first time in literature to study the relationship between the reduction of the synthesized view distortion and the accumulated MGS enhancement rate of the texture and depth views.

At this stage, let us assume the user link bandwidth is B, and there is two views (left and right view) both including texture and depth views. It is reasonable to allocate the same amount of rate $(B/2)$ to both the left and right views, this is because the two views typically share similar characteristics. For simplicity, the following analysis will be based on one view. Let us use R_{t0} and R_{d0} to denote the base layer rate of one texture and depth view, respectively; so the maximum allocated rate for the MGS enhancement layer of one view should be $B_E = B/2 - R_{t0} - R_{d0}$.

In fact, based on (15) and (16) and the fact that the synthesized view distortion caused by texture and depth views is additive [21], the synthesized view distortion (2) could be rewritten as

$$\begin{aligned} \mathrm{SSD}(R_t, R_d) &= E_0 - D_t(R_t) - D_d(R_d) \\ &= E_0 - \left(a_t \cdot \left(1 - e^{-\frac{R_t}{b_t}}\right) + c_t\right) \\ &\quad - \left(a_d \cdot \left(1 - e^{-\frac{R_d}{b_d}}\right) + c_d\right) \end{aligned} \tag{17}$$

where E_0 is the synthesized view distortion when no MGS enhancement layer packet is used for both texture and depth views (only base layer packets used). Then the problem becomes to find the optimal $\{R_t, R_d\}$ that could minimize $\mathrm{SSD}(R_t, R_d)$, which could be mathematically written as

$$\begin{cases} \min \quad \mathrm{SSD}(R_t, R_d) \\ \text{s.t.} \quad R_t + R_d \le B_E. \end{cases} \tag{18}$$

It is worth mentioning, in (18) the mean squared error (MSE) is used as the quality metric. For more sophisticated quality metric, 3VQM [36] might be a good choice for the synthesized views quality assessment, which will be left for the future research work.

For the texture view based on the fact that $a_t > 0$, $b_t > 0$, we could have

$$\frac{\partial D_t(R_t)}{\partial R_t} = \frac{a_t}{b_t} e^{-\frac{R_t}{b_t}} > 0$$

and

$$\frac{\partial^2 D_t(R_t)}{\partial R_t^2} = -\frac{a_t}{b_t^2} e^{-\frac{R_t}{b_t}} < 0.$$

Similarly, for the depth view we could have

$$\frac{\partial D_d(R_d)}{\partial R_d} > 0, \quad \frac{\partial^2 D_d(R_d)}{\partial R_d^2} < 0.$$

TABLE I
EXPERIMENTAL ENVIRONMENTS FOR THE SIMULATIONS

Sequences	*BookArrival*	*Newspaper*	*Lovebird1*	*Balloons*	*Poznan_Street*	*Poznan_Hall2*
Resolution	1024×768	1024×768	1024×768	1024×768	1920×1088	1920×1088
GOP size	8	8	8	8	8	8
Frame	1-50	1-50	1-50	1-50	151-200	1-50
Intra period	8	8	8	8	8	8
View No.	$(8, 10) \rightarrow 9$	$(2, 4) \rightarrow 3$	$(6, 8) \rightarrow 7$	$(1, 3) \rightarrow 2$	$(3, 4) \rightarrow 3.5$	$(6, 7) \rightarrow 6.5$
Frame rate	16.7	30	30	30	25	25

At this stage, we can get the following properties for $\mathrm{SSD}(R_t, R_d)$:

$$\frac{\partial \mathrm{SSD}(R_t, R_d)}{\partial R_t} = -\frac{\partial D_t(R_t)}{\partial R_t} < 0 \tag{19}$$

$$\frac{\partial^2 \mathrm{SSD}(R_t, R_d)}{\partial R_t^2} = -\frac{\partial^2 D_t(R_t)}{\partial R_t^2} > 0 \tag{20}$$

$$\frac{\partial \mathrm{SSD}(R_t, R_d)}{\partial R_d} = -\frac{\partial D_d(R_d)}{\partial R_d} < 0 \tag{21}$$

$$\frac{\partial^2 \mathrm{SSD}(R_t, R_d)}{\partial R_d^2} = -\frac{\partial^2 D_d(R_d)}{\partial R_d^2} > 0. \tag{22}$$

With (19)-(22), $\mathrm{SSD}(R_t, R_d)$ is a concave function of both R_t and R_d, so the constrained optimization problem (18) can be solved by means of the standard Lagrangian optimization by minimizing the cost function

$$J = \mathrm{SSD}(R_t, R_d) + \lambda (R_t + R_d) \tag{23}$$

where λ is the Lagrangian multiplier. So by imposing $\nabla J = 0$ we get

$$\frac{\partial J}{\partial R_t} = -\frac{a_t}{b_t} e^{-\frac{R_t}{b_t}} + \lambda = 0 \tag{24}$$

$$\frac{\partial J}{\partial R_d} = -\frac{a_d}{b_d} e^{-\frac{R_d}{b_d}} + \lambda = 0. \tag{25}$$

From (24) and (25), we can conclude that to minimize J, the following condition must be satisfied:

$$\frac{a_t}{b_t} e^{-\frac{R_t}{b_t}} = \frac{a_d}{b_d} e^{-\frac{R_d}{b_d}} = \lambda. \tag{26}$$

Hence, by jointly solving (26) and $R_t + R_d = B_E$, the optimal rate for the MGS enhancement layer of texture and depth views should be

$$R_t = \frac{1}{b_t + b_d} \left(b_t B_E + b_t b_d \ln \frac{a_t b_d}{b_t a_d} \right) \tag{27}$$

$$R_d = \frac{1}{b_t + b_d} \left(b_d B_E - b_t b_d \ln \frac{a_t b_d}{b_t a_d} \right). \tag{28}$$

By taking the fact that the enhancement layer rate should be in the range of $[0, B_E]$, the final allocated rate for the enhancement layer of texture view should be

$$R_t' = \begin{cases} R_t, & \text{if } 0 \le R_t \le B_E \\ 0, & \text{if } R_t < 0 \\ B_E, & \text{if } R_t > B_E. \end{cases} \tag{29}$$

Similarly, the final allocated rate for the enhancement layer of depth view should be

$$R_d' = \begin{cases} R_d, & \text{if } 0 \le R_d \le B_E \\ 0, & \text{if } R_d < 0 \\ B_E, & \text{if } R_d > B_E. \end{cases} \tag{30}$$

It is worth noticing that for (29) and (30), $R_t' + R_d' = B_E$ holds for all the cases.

V. EXPERIMENTAL RESULTS

In the experiments, 6 typical 3-D video sequences are used: *BookArrival* [45], *Newspaper* [46], *Lovebird1* [46], *Balloons* [47], *Poznan_Street* [48], and *Poznan_Hall2* [48]. The general experimental setting is listed in Table I unless otherwise noted. The proposed algorithm is implemented based on H.264/SVC reference software JSVM 9.19.15. A hierarchical prediction structure is used, with the GOP size being 8. The left view and right view are independently encoded using H.264/SVC MGS coding tool, and the virtual view is synthesized using View Synthesis Reference Software (VSRS 3.5) [49]. The default hole filling algorithm implemented in VSRS is used for the occluded regions. The transform coefficients of a macroblock are split into six MGS layers with the weights $W = [1, 2, 2, 3, 4, 4]$. We selected this weight vector because it was reported that these MGS weights led to competitive rate-distortion performance [40]. The baseline profile is used for the MGS base layer, while the scalable baseline profile is used for the enhancement layer, with 8×8 transform disabled at both layers. The enhancement layer is used in motion estimation and motion prediction for nonkey pictures in MGS layers.

A. Performance of Proposed Depth Coding and MGS Packet Ordering

In the first experiment, we tested the effects of the MGS base layer and enhancement layer QP pair $\{QP_b^t, QP_e^t\}$ and $\{QP_b^d, QP_e^d\}$ on the coding performance. The depth views of the *Newspaper* sequence are encoded using different QP pairs: $\{35, 25\}$, $\{35, 27\}$, and $\{35, 30\}$, with QP difference between the two layers being 10, 8 and 5, respectively. In this experiment, the texture view is not compressed. View 2 and 4 are used to synthesize virtual view 3. The synthesized view quality versus depth bit rate is shown in Fig. 4. It is noted that having large QP difference between the base layer and enhancement layer, that is, QP pair $\{35, 25\}$, leads to better depth coding performance. Meanwhile, with large QP difference, the MGS bit rate range is large, which can provide more flexibility for rate adaptation. It is interesting to note that when the MGS bitstream is truncated close to the base layer point, that is, rate lower than 1000 kb/s, QP pair $\{35, 30\}$ leads to the best performance, with QP pair $\{35, 25\}$ being the worst case. This is because the mismatch error is large when the QP difference is large, and the effects of

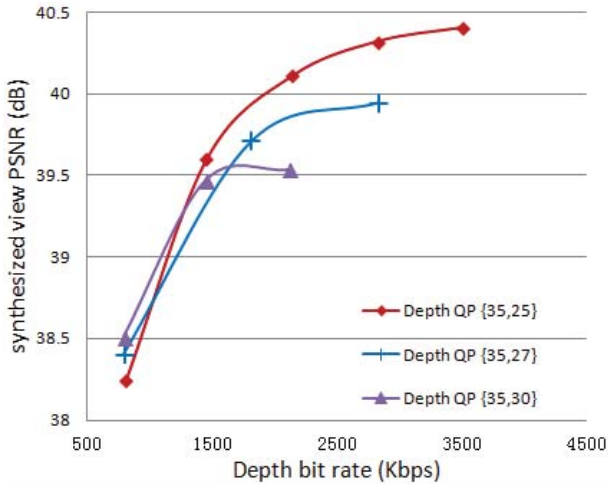


Fig. 4. Effects of having different QP pairs (MGS base layer and enhancement layer) for depth coding, *Newspaper* sequence is used; texture view is not compressed.

mismatch error is more obvious when the truncation point is close to the base layer. Nevertheless, with the whole bit rate range performance, generally, QP pair $\{35, 25\}$ outperforms other two cases. Similar results are obtained for the texture view. Meanwhile, we noticed that by using QP pair $\{35, 25\}$ for both the texture and depth views, $0 < R'_t < B_E$ and $0 < R'_d < B_E$ holds for various values of B_E, which also indicates the rationality of using this QP pair for the texture and depth views. Hence, in the following experiments, MGS base layer and enhancement layer QP pair $\{35, 25\}$ will be used for both the texture and depth views if not otherwise noted.

In the second set of experiments, the advantage of the proposed depth coding and MGS enhancement packet ordering methods are demonstrated. To do this, three different depth coding and bit extraction approaches using H.264/SVC MGS are compared. In the first approach, during the depth coding process, the synthesized view distortion metric is not used in the rate-distortion optimized coding mode selection and MGS enhancement layer packet ordering steps; whereas the traditional MSE is used [11], so this approach is named SVDC OFF+SVDO OFF. Here SVDC stands for synthesized view distortion metric based rate-distortion optimized coding mode selection, whereas *SVDO* stands for synthesized view distortion metric-based MGS enhancement layer packet ordering. In the second approach, the synthesized view distortion metric is applied in the rate-distortion optimized coding mode selection step, whereas the newly proposed synthesized view distortion-based MGS enhancement layer packet ordering is not used, so we call this approach as SVDC ON+SVDO OFF. These two approaches are used as the benchmarks for the proposed approach, SVDC ON+SVDO ON, where the synthesized view distortion metric is used for both coding mode selection and MGS enhancement layer packet ordering. MGS base layer and enhancement layer QP pair $\{35, 25\}$ is used for depth coding when SVDC is used, whereas for the case when SVDC is not used, other QP pairs that lead to similar bit rate are adopted. Texture views are not compressed. The rate-distortion performance comparison for the 6 video sequences is shown in Fig. 5. The proposed approach outperforms other two approaches for all the six video sequences. Comparing with the SVDC OFF+SVDO OFF, the PSNR gain of the proposed approach is around 1–7 dB. The BD-Rate and BD-PSNR [50] results compared with SVDC ON+SVDO OFF are listed in Table II, the average BD-PSNR gain is 0.31 dB, and the BD-Rate is -8.95%. All the results demonstrate the importance of using the synthesized view distortion metric for both rate-distortion optimized coding mode selection and MGS enhancement layer packet ordering. It is worth mentioning that for the start and end points shown in Fig. 5, the performance of the proposed approach and SVDC ON+SVDO OFF is the same, because no MGS enhancement packets and all MGS enhancement packets, respectively, are used, so MGS packet ordering algorithm has no effect for these cases.

TABLE II
BD-RATE AND BD-PSNR RESULTS BY COMPARING THE PROPOSED DEPTH CODING APPROACH WITH THE SVDC ON+SVDO OFF APPROACH

Sequences	BD-Rate	BD-PSNR
Balloons	-15.69%	0.42 dB
BookArrival	-25.35%	0.34 dB
Lovebird1	-0.59%	0.53 dB
Newspaper	-2.13%	0.16 dB
Poznan_Street	-8.77%	0.39 dB
Poznan_Hall2	-1.18%	0.07 dB
Average	-8.95%	0.31 dB

B. Performance of Proposed Scalable Bit Allocation

In Fig. 6, the performance of the proposed scalable bit allocation scheme is reported where we compared it with the fixed ratio (1:5) bit allocation between depth and texture views. To visualize the gain of each step, three reference curves are reported: SVDC ON+SVDO ON (1:5) means that for both the MGS depth coding and enhancement packet ordering, the synthesized view distortion metric is used; SVDC ON+SVDO OFF (1:5) means that the SVD metric is used for depth coding but not for packet ordering; SVDC OFF+SVDO OFF (1:5) means that the SVD metric is not used at either stage. For the curve *Full Search*, the allocated depth rates obtained using (30) are multiplied with a scaling factor of $\{0.7, 0.8, 0.9, 1.0, 1.1, 1.2, 1.3\}$, and this curve is generated by selecting the best points in terms of synthesized view PSNR among all the scaled depth view rates. This curve is used to determine an approximated full-search result [12], which is the upper bound for bit rate allocation performance. QP pair $\{35, 25\}$ is used for both texture and depth coding when SVDC is used, whereas for the case when SVDC is not used, the QP pairs that lead to similar bit rate is adopted for depth views. From Fig. 6, we can notice that the proposed scheme outperforms the three reference schemes using fix ratio allocation. SVDC ON+SVDO ON (1:5) performs best among the three fix ratio allocation schemes. It is also very interesting to note that, the curve of the proposed scheme is almost overlapping with that of

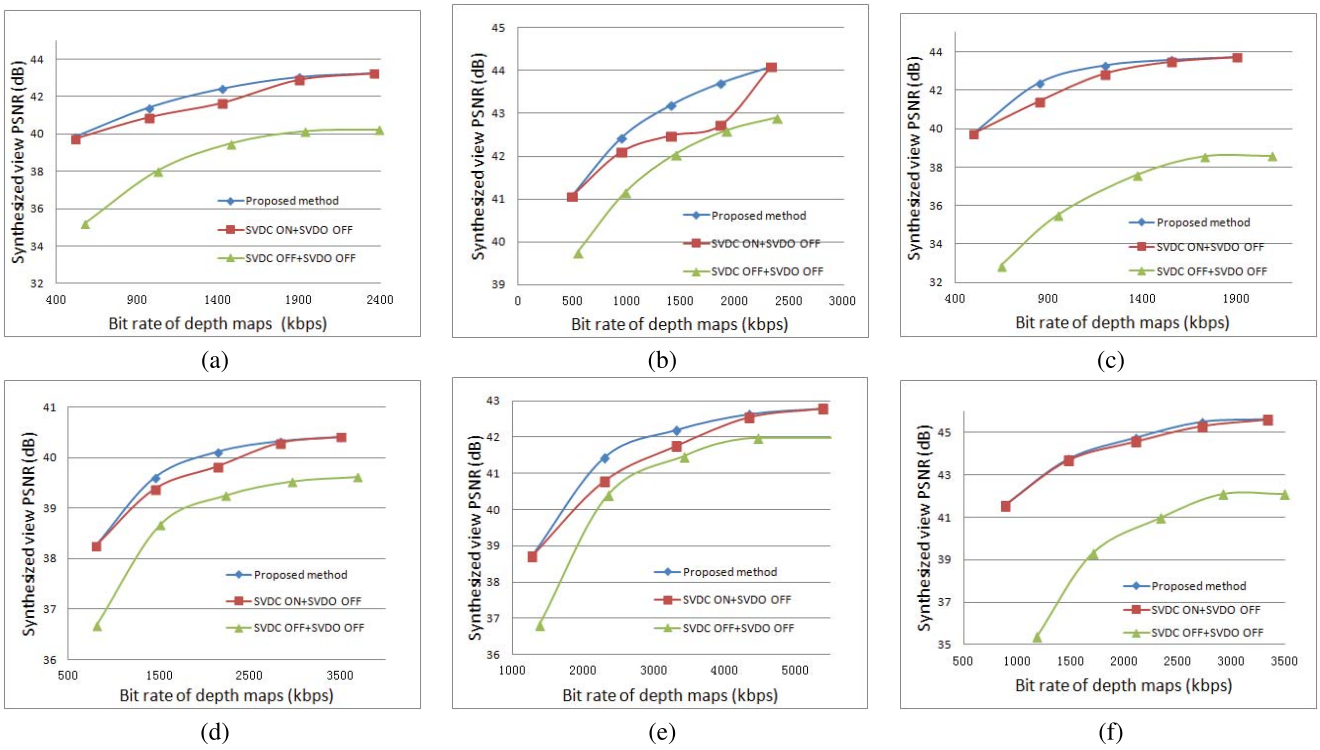


Fig. 5. Synthesized view PSNR versus bit rate of depth views for three different approaches. (a) *Balloons*. (b) *BookArrival*. (c) *Lovebird1*. (d) *Newspaper*. (e) *Poznan_Street*. (f) *Poznan_Hall2*.

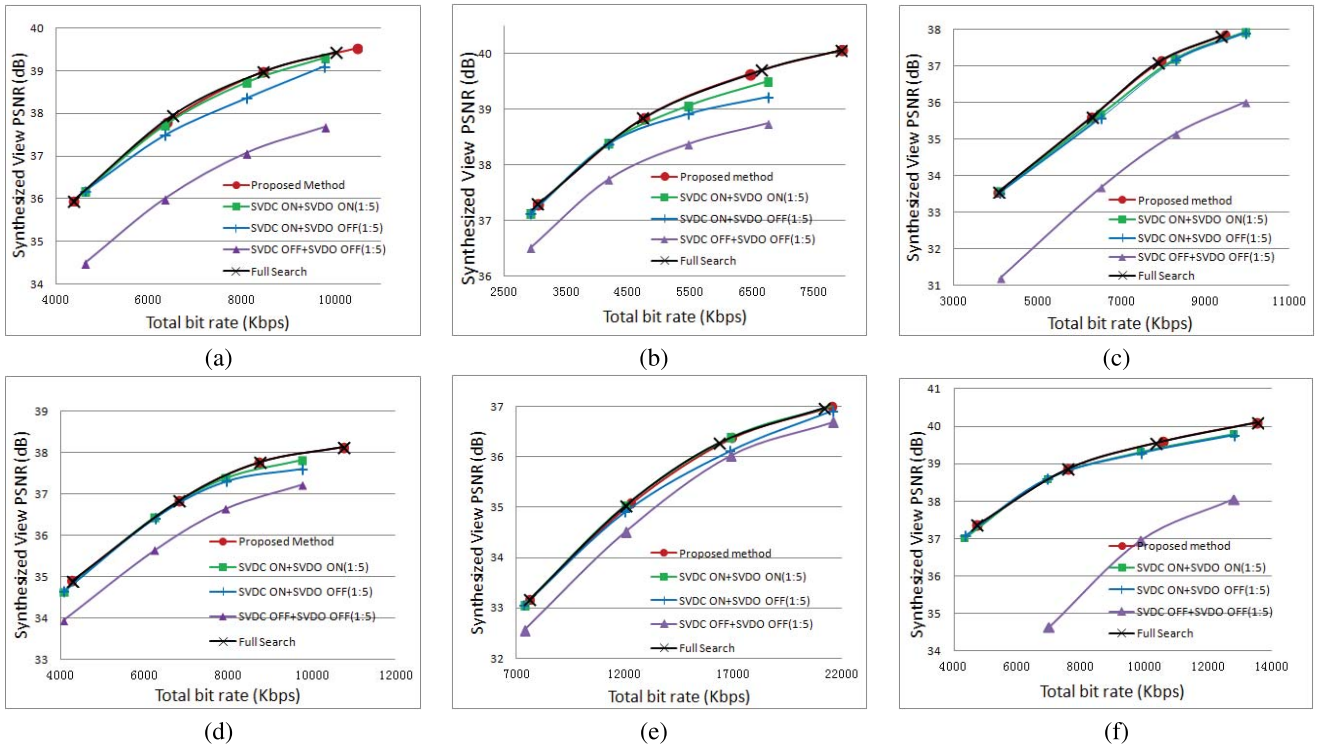


Fig. 6. Synthesized view PSNR versus total bit rate of texture and depth views. (a) *Balloons*. (b) *BookArrival*. (c) *Lovebird1*. (d) *Newspaper*. (e) *Poznan_Street*. (f) *Poznan_Hall2*.

the *Full Search*. The allocated bit rate ratio between depth view and texture view, and the BD-Rate and BD-PSNR gains for the proposed scheme over SVDC ON+SVDO ON (1:5) are listed in Table III. The average bit rate reduction of the proposed scheme over SVDC ON+SVDO ON (1:5) is 1.65%. It is noted that for some video sequences, foe example, *Lovebird1*, the allocated depth bit rate is less than 20% of the texture view; whereas for some video sequences, the

TABLE III
ALLOCATED BIT RATE RATIO BETWEEN DEPTH VIEW AND TEXTURE VIEW; AND BD-RATE AND BD-PSNR RESULTS BY COMPARING THE PROPOSED SCALABLE BIT-ALLOCATION SCHEME WITH SVDC ON+SVDO ON (1:5)

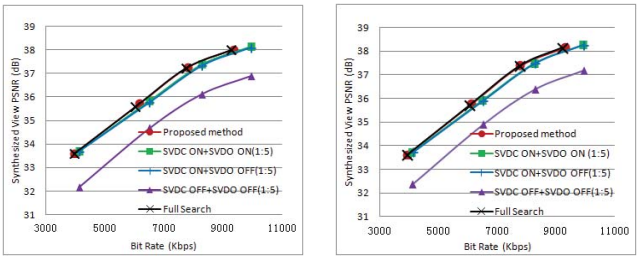
Sequence	Ratio	BD-Rate	BD-PSNR
Balloons	0.239	−0.75%	0.05 dB
BookArrival	0.380	−4.55%	0.12 dB
Lovebird1	0.157	−2.22%	0.11 dB
Newspaper	0.314	−0.74%	0.06 dB
Poznan_Street	0.209	0.60%	−0.02 dB
Poznan_Hall2	0.292	−2.25%	0.07 dB
Average	0.265	−1.65%	0.07 dB

TABLE IV
BD-RATE AND BD-PSNR RESULTS BY COMPARING THE PROPOSED SCHEME WITH FULL SEARCH

Sequence	BD-Rate	BD-PSNR
Balloons	0.99%	−0.03 dB
BookArrival	0.04%	−0.00 dB
Lovebird1	0.06%	0.00 dB
Newspaper	0.00%	0.00 dB
Poznan_Street	0.79%	−0.02 dB
Poznan_Hall2	0.03%	0.00 dB
Average	0.32%	−0.01 dB

allocated depth bit rate is more than that 20% of the texture. This observation serves to demonstrate the robustness and effectiveness of the proposed scalable bit-allocation scheme. The BD-Rate and BD-PSNR for the proposed scheme over *Full Search* are listed in Table IV, where the average BD-Rate of the proposed scheme over *Full Search* is only 0.32%. The precision of the proposed scalable bit-allocation scheme is demonstrated by the fact that performance of the proposed bit-allocation scheme is very close to the upper bound curve *Full Search*.

In Fig. 6, the reported results are for the case that only one synthesized view is generated. To demonstrate the effectiveness of the proposed bit-allocation method for more than one synthesized view, in Fig. 7 the bit-allocation performance is reported when 3 and 7 synthesized views are generated. The generated synthesized views are evenly distributed between the left and right views (i.e., view 6 and view 8) for the *Lovebird1* sequence. It is seen that with 3 and 7 synthesized views, the proposed scalable bit-allocation performance is still quite close to that of *Full Search*, and the BD-Rate over SVDC ON+SVDO ON (1:5) are −3.08% and −3.44% for 3 and 7 synthesized views, respectively, which are larger than the gain of one synthesized view (−2.22%). All these results indicate the accuracy of the proposed bit-allocation algorithm. It is also observed that the average bit rate ratios between depth view and texture view are {0.157, 0.137, 0.129} for generating {1, 3, 7} synthesized views. These results confirm our argument that with the increase of the number of synthesized views, more bit should be allocated for the texture views.



(a) (b)

Fig. 7. Average synthesized view PSNR versus total bit rate of texture and depth views; multiple synthesized views are generated for the *Lovebird1* sequence. (a) 3 synthesized views are generated. (b) 7 synthesized views are generated.

C. Examples of Subjective Quality

Besides the objective results, some examples of subjective quality are also reported herein. Firstly, the results of the proposed depth view coding and MGS enhancement packet ordering are reported in Fig. 8. In this testing, texture views are not compressed, whereas depth view coding and bit extraction approaches include SVDC OFF+SVDO OFF, SVDC ON+SVDO OFF and SVDC ON+SVDO ON. The depth view bit rate is the same for the three approaches (1421 kb/s). To better visualize the difference of the three frames, some regions, that is, the edges of the balloons, are zoomed in for better display. It is clear that the subjective quality of SVDC ON+SVDO ON is the best, with the balloon edges well protected. Secondly, the subjective results of the proposed bit-allocation scheme are shown in Fig. 9. In this comparison, synthesized frames of *BookArrival* sequence are reported for SVDC OFF+SVDO OFF(1:5) , SVDC ON+SVDO ON(1:5), and *Proposed method*. To have fair comparison, for all the three approaches, the total bit rate is 6758 kb/s. It is noted that the proposed bit-allocation scheme leads to the best subjective quality, especially for the zoomed in regions.

D. Computational Complexity Analysis

The computational complexity of the proposed scalable bit allocation scheme is moderate. First, for synthesized view distortion based depth MGS enhancement packet ordering, the only difference with normal MGS enhancement packet ordering is using (14) to evaluate distortion instead of using MSE of depth view. Thus, the computational complexity of this step should be equivalent with MSE-based MGS packet ordering process. Regarding the computational complexity of normal MGS enhancement packet ordering, please refer to [41], [42]. Second, for the optimal bit-allocation step, least square fitting is used to get parameter value $\{a_t, b_t, c_t\}$ and $\{a_d, b_d, c_d\}$, which needs a computational complexity of $O(NM^2)$, where N is the number of samples, and M is the number of parameters for fitting. In our case, N equals to the number of frames in the video sequence multiply the MGS layer number (6 in our case); M equals to 3. The curve fitting although has complexity $O(NM^2)$, it is negligible in comparison with the video coding

(a) (b) (c)

Fig. 8. Synthesized view (view 2) of Frame 22 for *Balloons* sequence; depth view bit rate is 1421 kb/s; texture views are not compressed. (a) SVDC OFF+SVDO OFF, 40.28 dB. (b) SVDC ON+SVDO OFF, 42.00 dB. (c) SVDC ON+SVDO ON, 43.57 dB.

(a) (b) (c)

Fig. 9. Synthesized view (view 9) of Frame 6 for *BookArrival* sequence; total bit rate of texture and depth views is 6758 kb/s. (a) SVDC OFF+SVDO OFF(1:5), 39.62 dB. (b) SVDC ON+SVDO ON(1:5), 40.44 dB. (c) Proposed method, 40.57 dB.

complexity, because the complexity of coding is proportional to the pixel number in each frame in addition to the frame number.

VI. Conclusion

In this paper, a *scalable* bit-allocation scheme for texture and depth views has been proposed. In this scheme, both the texture and depth views are encoded using the MGS tool of H.264/SVC. The optimal truncation points for the texture and depth views can be found using this scheme. This kind of *scalable* bit allocation is very important for the 3-D video server, which provides 3-D video streaming service for users with different link bandwidth in the heterogeneous networks. Another merit of this scheme is that the information generated in the MGS enhancement packet ordering process is exploited during the bit-allocation stage, so the optimal truncation points for the texture and depth views can be obtained using one simple formula for varying total target rates. Experimental results has demonstrated the effectiveness of the proposed *scalable* bit-allocation scheme.

In this paper, to use the H.264/SVC MGS tool, different views are compressed independently. This is because the current multiview video coding standard, that is, MVC, does not support quality scalable coding. In the future, we are going to investigate the proposed framework in a quality-scalable multiview coding system where inter-view prediction is supported.

Acknowledgment

The authors would like to thank Associate Editor Prof. S. Shirani and the anonymous reviewers for their valuable comments and suggestions.

References

[1] J. Konrad and M. Halle, "3-D displays and signal processing," *IEEE Signal Process. Mag.*, vol. 24, no. 6, pp. 97–111, Nov. 2007.

[2] P. Benzie *et al.*, "A survey of 3DTV displays: Techniques and technologies," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 11, pp. 1647–1658, Nov. 2007.

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

[4] A. Vetro, T. Wiegand, and G. J. Sullivan, "Overview of the stereo and multiview video coding extensions of the H.264/MPEG-4 AVC standard," *Proc. IEEE*, vol. 99, no. 4, pp. 626–642, Apr. 2011.

[5] *Advanced Video Coding for Generic Audiovisual Services*, document ISO/IEC, ITU-T RECOMMENDATION, Apr. 2003.

[6] M. M. Hannuksela, Y. Chen, T. Suzuki, J.-R. Ohm, and G. Sullivan, Ed., *3D-AVC Draft Text 8*, document JCT-3V JCT3V-F1002, Nov. 2013.

[7] Y. Chen, M. M. Hannuksela, T. Suzuki, and S. Hattori. (2014, May). Overview of the MVC+D 3D video coding standard. *J. Vis. Commun. Image Represent.* [Online]. *25(4)*. pp. 679–688. Available: http://dx.doi.org/10.1016/j.jvcir.2013.03.013

[8] P. Merkle, A. Smolic, K. Muller, and T. Wiegand, "Multi-view video plus depth representation and coding," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2007, pp. I-201–I-204.

[9] C. Fehn, "Depth-image-based rendering (DIBR), compression, and transmission for a new approach on 3D-TV," *Proc. SPIE*, vol. 5291, pp. 93–104, May 2004.

[10] C. Gurler, B. Gorkemli, G. Saygili, and A. M. Tekalp, "Flexible transport of 3-D video over networks," *Proc. IEEE*, vol. 99, no. 4, pp. 694–707, Apr. 2011.

[11] H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the scalable video coding extension of the H.264/AVC standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 9, pp. 1103–1120, Sep. 2007.

[12] Y. Morvan, D. Farin, and P. H. N. de With, "Joint depth/texture bit-allocation for multi-view video compression," in *Proc. Picture Coding Symp. (PCS)*, 2007, pp. 265–268.

[13] Y. Liu, Q. Huang, S. Ma, D. Zhao, and W. Gao, "Joint video/depth rate allocation for 3D video coding based on view synthesis distortion model," *Signal Process., Image Commun.*, vol. 24, no. 8, pp. 666–681, 2009.

[14] H. Yuan, Y. Chang, J. Huo, F. Yang, and Z. Lu, "Model-based joint bit allocation between texture videos and depth maps for 3-D video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 4, pp. 485–497, Apr. 2011.

[15] Q. Wang, X. Ji, Q. Dai, and N. Zhang, "Free viewpoint video coding with rate-distortion analysis," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 6, pp. 875–889, Jun. 2012.

[16] F. Shao, G. Jiang, W. Lin, M. Yu, and Q. Dai, "Joint bit allocation and rate control for coding multi-view video plus depth based 3D video," *IEEE Trans. Multimedia*, vol. 15, no. 8, pp. 1843–1854, Dec. 2013.

[17] Y. Liu *et al.*, "A novel rate control technique for multiview video plus depth based 3D video coding," *IEEE Trans. Broadcast.*, vol. 57, no. 2, pp. 562–571, Jun. 2011.

[18] Y. Zhang, S. Kwong, L. Xu, S. Hu, C. Kuo, and G. Jiang, "Regional bit allocation and rate distortion optimization for multiview depth video coding with view synthesis distortion model," *IEEE Trans. Image Process.*, vol. 22, no. 9, pp. 3497–3512, Sep. 2013.

[19] J. Xiao, T. Tillo, H. Yuan, and Y. Zhao, "Macroblock level bits allocation for depth maps in 3-D video coding," *J. Signal Process. Syst.*, vol. 74, no. 1, pp. 127–135, 2013.

[20] W.-S. Kim, A. Ortega, P. Lai, D. Tian, and C. Gomila, "Depth map coding with distortion estimation of rendered view," *Proc. SPIE*, vol. 7543, p. 75430B, Jan. 2010.

[21] B. T. Oh, J. Lee, and D.-S. Park, "Depth map coding based on synthesized view distortion function," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 7, pp. 1344–1352, Nov. 2011.

[22] J. Chakareski, V. Velisavljevic, and V. Stankovic, "User-action-driven view and rate scalable multiview video coding," *IEEE Trans. Image Process.*, vol. 22, no. 9, pp. 3473–3484, Sep. 2013.

[23] R. Mathew, D. Taubman, and P. Zanuttigh, "Scalable coding of depth maps with RD optimized embedding," *IEEE Trans. Image Process.*, vol. 22, no. 5, pp. 1982–1995, May 2013.

[24] L. Zhang, W. J. Tam, and D. Wang, "Stereoscopic image generation based on depth images," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, vol. 5. Oct. 2004, pp. 2993–2996.

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

[26] P. Ndjiki-Nya *et al.*, "Depth image-based rendering with advanced texture synthesis for 3-D video," *IEEE Trans. Multimedia*, vol. 13, no. 3, pp. 453–465, Jun. 2011.

[27] C. Yao, T. Tillo, Y. Zhao, J. Xiao, H. Bai, and C. Lin, "Depth map driven hole filling algorithm exploiting temporal consistent information," *IEEE Trans. Broadcast.*, vol. 60, no. 2, pp. 394–404, Jun. 2014.

[28] Y. Zhao, C. Zhu, Z. Chen, D. Tian, and L. Yu, "Boundary artifact reduction in view synthesis of 3D video: From perspective of texture-depth alignment," *IEEE Trans. Broadcast.*, vol. 57, no. 2, pp. 510–522, Jun. 2011.

[29] W. Lin and C.-C. J. Kuo, "Perceptual visual quality metrics: A survey," *J. Vis. Commun. Image Represent.*, vol. 22, no. 4, pp. 297–312, May 2011.

[30] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[31] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 430–444, Feb. 2006.

[32] Q. Huynh-Thu, P. Le Callet, and M. Barkowsky, "Video quality assessment: From 2-D to 3-D challenges and future trends," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2010, pp. 4025–4028.

[33] H. Shao, X. Cao, and G. Er, "Objective quality assessment of depth image based rendering in 3DTV system," in *Proc. 3DTV Conf. True Vis., Capture, Transmiss. Display 3D Video*, May 2009, pp. 1–4.

[34] A. Tikanmaki, A. Gotchev, A. Smolic, and K. Miller, "Quality assessment of 3D video in rate allocation experiments," in *Proc. IEEE Int. Symp. Consumer Electron. (ISCE)*, Apr. 2008, pp. 1–4.

[35] C. T. E. R. Hewage, S. T. Worrall, S. Dogan, S. Villette, and A. M. Kondoz, "Quality evaluation of color plus depth map-based stereoscopic video," *IEEE J. Sel. Topics Signal Process.*, vol. 3, no. 2, pp. 304–318, Apr. 2009.

[36] M. Solh and G. AlRegib, "A no-reference quality measure for DIBR-based 3D videos," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2011, pp. 1–6.

[37] D. Rusanovskyy, K. Muller, and A. Vetro, "Common test conditions of 3DV core experiments," document ISO/IEC JTC1/SC29/WG11, M26349, Jul. 2012.

[38] M. M. Hannuksela *et al.*, "Multiview-video-plus-depth coding based on the Advanced Video Coding standard," *IEEE Trans. Image Process.*, vol. 22, no. 9, pp. 3449–3458, Sep. 2013.

[39] Y. Hui, L. Ju, X. Hongji, L. Zhibin, and L. Wei, "Coding distortion elimination of virtual view synthesis for 3D video system: Theoretical analyses and implementation," *IEEE Trans. Broadcast.*, vol. 58, no. 4, pp. 558–568, Dec. 2012.

[40] R. Gupta, A. Pulipaka, P. Seeling, L. J. Karam, and M. Reisslein, "H.264 coarse grain scalable (CGS) and medium grain scalable (MGS) encoded video: A trace based traffic and quality evaluation," *IEEE Trans. Broadcast.*, vol. 58, no. 3, pp. 428–439, Sep. 2012.

[41] I. Amonou, N. Cammas, S. Kervadec, and S. Pateux, "Optimized rate-distortion extraction with quality layers in the scalable extension of H.264/AVC," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 9, pp. 1186–1193, Sep. 2007.

[42] E. Maani and A. K. Katsaggelos, "Optimized bit extraction using distortion modeling in the scalable extension of H.264/AVC," *IEEE Trans. Image Process.*, vol. 18, no. 9, pp. 2022–2029, Sep. 2009.

[43] J. Reichel, H. Schwarz, and M. Wien, *Joint Scalable Video Model 11 (JSVM 11)*, document JVT-X202, 2007.

[44] Å. Björck, *Numerical Methods for Least Squares Problems*. Philadelphia, PA, USA: SIAM, 1996.

[45] (2013, Sep.). *3DV Sequences of HHI*. Fraunhofer Heinrich Hertz Institute, Berlin, Germany [Online]. Available FTP: ftp://ftp.hhi.de/HHIMPEG3DV

[46] (2008, Apr.). *3DV Sequences of ETRI and GIST*. Electronics and Telecommunications Research Institute and Gwangju Institute of Science and Technology, Gwangju, Korea [Online]. Available FTP: ftp://203.253.128.142/

[47] (2008, Mar.). *3DV Sequences of Nagoya University*. Nagoya University, Japan [Online]. Available: http://www.tanimoto.nuee.nagoya-u.ac.jp/mpeg/mpeg-ftv.html

[48] (2012, Nov.). *Poznan University 3DV Sequences Poznan University*, Poznan, Poland [Online]. Available FTP: ftp://multimedia.edu.pl/3DV/

[49] *MPEG 3-DV View Synthesis Reference Software* [Online]. Available: http://wg11.sc29.org/svn/repos/MPEG-4/test/trunk/3D/view.synthesis, accessed Mar. 2010.

[50] G. Bjontegaard, *Improvements of the BD-PSNR Model*, document ITU-T SG16 Q, 2008.

Jimin Xiao received the B.S. and M.E. degrees in telecommunication engineering from the Nanjing University of Posts and Telecommunications, Nanjing, China, in 2004 and 2007, respectively, and the dual Ph.D. degrees in electrical engineering and electronics from the University of Liverpool, Liverpool, U.K., and Xi'an Jiaotong-Liverpool University, Suzhou, China, in 2013.

He was a Software Engineer with Motorola Electronics Ltd., Nanjing, China, from 2007 to 2008, and a System Engineer with Realsil (Realtek) Semiconductor Corporation, Suzhou, China. In 2013, he served as a Visiting Researcher with Nanyang Technological University, Singapore. Since 2013, he has been a Senior Researcher with the Department of Signal Processing, Tampere University of Technology, Tampere, Finland, and an External Researcher with the Nokia Research Center, Tampere. His current research interests include video streaming, image and video compression, and multiview video coding.

Miska M. Hannuksela (M'03) received the M.Sc. degree in engineering and the Dr.Sc. degree in technology from the Tampere University of Technology, Tampere, Finland, in 1997 and 2010, respectively.

He has been with the Nokia Research Center, Tampere, since 1996, in different roles, including research manager/leader positions in the areas of video and image compression, end-to-end multimedia systems, and sensor signal processing and context extraction. He is currently a Distinguished Scientist in Multimedia Technologies with the Nokia Research Center. He has authored more than 100 conference papers, more than 25 journal papers, and hundreds of standardization contributions in JCT-VC, JCT-3V, JVT, MPEG, 3GPP, and DVB. He has granted patents from more than 70 patent families. His current research interests include video compression and multimedia communication systems.

Dr. Hannuksela has been an Associate Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS OF VIDEO TECHNOLOGY since 2010. He was a recipient of the Best Doctoral Thesis Award from the Tampere University of Technology in 2009 and the Scientific Achievement Award by the Centre of Excellence of Signal Processing, Tampere University of Technology, in 2010. He has authored the paper that received the Best Paper Award at the 2012 Visual Communications and Image Processing Conference.

Tammam Tillo (M'05–SM'12) received the Dipl.Ing. degree in electrical engineering from the University of Damascus, Damascus, Syria, in 1994, and the Ph.D. degree in electronics and communication engineering from Politecnico di Torino, Torino, Italy, in 2005.

He served as a Visiting Researcher with École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, in 2004, and was a Post-Doctoral Researcher with the Image Processing Laboratory, Politecnico di Torino, from 2005 to 2008. For a few months, he was an Invited Research Professor with the Digital Media Laboratory, SungkyunKwan University, Seoul, Korea, before joining Xi'an Jiaotong-Liverpool University (XJTLU), Suzhou, China, in 2008. He was promoted to Full Professor in 2012. From 2010 to 2013, he was the Head of the Department of Electrical and Electronic Engineering at XJTLU, and the Acting Head of the Department of Computer Science and Software Engineering from 2012 to 2013. He serves as an expert evaluator for several national-level research programs. His current research interests include robust transmission of multimedia data, image and video compression, and hyperspectral image compression.

Moncef Gabbouj (F'11) received the B.S. degree in electrical engineering from Oklahoma State University, Stillwater, OK, USA, in 1985, and the M.S. and Ph.D. degrees in electrical engineering from Purdue University, West Lafayette, IN, USA, in 1986 and 1989, respectively.

He is an Academy Professor with the Academy of Finland, Helsinki, Finland. He held several visiting professorships at different universities. He is a Professor of Signal Processing with the Department of Signal Processing, Tampere University of Technology, Tampere, Finland. He has published over 550 publications and has supervised 38 doctoral theses. His current research interests include multimedia content-based analysis, indexing and retrieval, machine learning, nonlinear signal and image processing and analysis, voice conversion, and video processing and coding.

Dr. Gabbouj is a member of the Finnish Academy of Science and Letters. He is currently the Chairman of the IEEE Circuits and Systems Society Technical Committee on Digital Signal Processing and a Committee Member of the IEEE Fourier Award for Signal Processing. He served as a Distinguished Lecturer for the IEEE Circuits and Systems Society. He served as an Associate Editor and the Guest Editor of many IEEE and international journals. He was a recipient of the 2012 Nokia Foundation Visiting Professor Award, the 2005 Nokia Foundation Recognition Award, and several Best Paper Awards.

Ce Zhu (M'03–SM'04) received the B.S. degree from Sichuan University, Chengdu, China, in 1989, and the M.Eng. and Ph.D. degrees from Southeast University, Nanjing, China, in 1992 and 1994, respectively, all in electronic and information engineering.

He is currently with the School of Electronic Engineering, University of Electronic Science and Technology of China, Chengdu. He was with Nanyang Technological University, Singapore, from 1998 to 2012, where he was promoted to Associate Professor in 2005. He was a Post-Doctoral Researcher with the Chinese University of Hong Kong, Hong Kong, in 1995, the City University of Hong Kong, Hong Kong, and the University of Melbourne, Melbourne, VIC, Australia, from 1996 to 1998. He has held visiting positions with Queen Mary, University of London, London, U.K., and Nagoya University, Nagoya, Japan. His current research interests include image/video coding, streaming and processing, 3-D video, joint source-channel coding, and multimedia systems and applications.

Dr. Zhu serves on the editorial boards of seven international journals, including as an Associate Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, the IEEE TRANSACTIONS ON BROADCASTING, and the IEEE SIGNAL PROCESSING LETTERS, an Editor of the IEEE *Communications Surveys and Tutorials*, and an Area Editor of *Signal Processing: Image Communication* (Elsevier). He has served on technical/program committees, organizing committees, and as a Track/Area/Session Chair for about 60 international conferences. He was a recipient of the 2010 Special Service Award from the IEEE Broadcast Technology Society, and is an IEEE BTS Distinguished Lecturer from 2012 to 2014.

Yao Zhao (M'06–SM'12) received the B.S. degree from the Department of Radio Engineering, Fuzhou University, Fuzhou, China, in 1989, and the M.E. degree from the Department of Radio Engineering, Southeast University, Nanjing, China, in 1992, and the Ph.D. degree from the Institute of Information Science, Beijing Jiaotong University (BJTU), Beijing, China, in 1996.

He became an Associate Professor at BJTU in 1998 and a Professor in 2001. From 2001 to 2002, he was a Senior Research Fellow with the Information and Communication Theory Group, Faculty of Information Technology and Systems, Delft University of Technology, Delft, The Netherlands. He is currently the Director of the Institute of Information Science at BJTU. He is currently leading several national research projects from the 973 Program, 863 Program, and the National Science Foundation of China. His current research interests include image/video coding, digital watermarking and forensics, and video analysis and understanding.

Dr. Zhao serves on the editorial boards of several international journals, including as an Associate Editor of the IEEE TRANSACTIONS ON CYBERNETICS and the IEEE SIGNAL PROCESSING LETTERS, an Area Editor of *Signal Processing: Image Communication* (Elsevier), and an Associate Editor of *Circuits, System, and Signal Processing* (Springer). He was named a Distinguished Young Scholar by the National Science Foundation of China in 2010, and was elected as a Chang Jiang Scholar of the Ministry of Education of China in 2013.