Multimed Tools Appl (2018) 77:19051–19069
https://doi.org/10.1007/s11042-017-5327-0

# SSIM-based joint-bit allocation for 3D video coding

**Harshalatha Y**[1] · **Prabir Kumar Biswas**[1]

Received: 6 January 2017 / Revised: 5 September 2017 / Accepted: 19 October 2017 /
Published online: 4 November 2017


**Abstract** The quality of a 3D video display depends on virtual view synthesis process which is affected by the bit allocation criterion. The performance of a bit allocation algorithm is dependent on various encoding parameters like quantization parameter, motion vector, mode selection, and so on. Rate-distortion optimization (RDO) is used to efficiently allocate bits with minimum distortion. In 3D video, rate-distortion (RD) property of synthesized view is used to assign bits between texture video and depth map. Existing literature on bit allocation methods use mean square error (MSE) as distortion metric which is not suitable for measuring perceptual quality. In this paper, we propose structural similarity (SSIM)-based joint bit allocation scheme to enhance visual quality of 3D video. Perceptual quality of a synthesized view depends on texture and depth map quality. Thus, SSIM-based RDO is performed on both texture and depth map where SSIM is used as distortion metric in mode decision and motion estimation. SSIM-based distortion model for synthesized view is determined experimentally. As SSIM cannot be related to quantization step, SSIM-MSE relation is used to convert distortion model in terms of MSE. The Lagrange multiplier method is used to solve the bit allocation problem. The proposed algorithm is implemented using 3DV-ATM as well as HEVC. RD curves show reduction in bitrate with an improvement in SSIM of synthesized view.

**Keywords** 3D video · Virtual view synthesis · Bit allocation · Perceptual quality · SSIM

✉ Harshalatha Y
harshalatha.y@ece.iitkgp.ernet.in

Prabir Kumar Biswas
pkb@ece.iitkgp.ernet.in

[1] Department of Electronics and Electrical Communication Engineering,
Indian Institute of Technology Kharagpur, Kharagpur, 721302, India

## 1 Introduction

Traditional 2D video is being replaced by 3D video which is an emerging visual media. Research in the field of 3D video technology is getting more attention and importance with increased demand for consumer products. 3D television (3DTV) and free-viewpoint television (FTV) are the two main applications of 3D video along with sports, medical field, education, and so on. 3DTV provides visually realistic scene and FTV gives flexibility of changing view angle to the viewers. Starting with stereoscopic display technology which requires special glasses, now 3D technology has reached autostereoscopy where viewer can enjoy the essence of real scene without glasses and gives wide angle of view [16, 24]. Autostereoscopic display requires multiple views to be acquired, coded and transmitted that increases the complexity of the whole system. Multiview video plus depth (MVD) format is used to reduce the number of views and virtual view synthesis [6, 7] is carried out at the decoder to render intermediate views. Compression, depth accuracy, and rendering algorithm cause distortion in the virtual view.

Virtual view synthesis process uses depth-image-based rendering (DIBR) algorithm. In addition to compression, another important factor that affects the quality of a virtual view is the bit allocation between texture video and depth map. Though depth map is not displayed, its importance lies in the view synthesis process as it contains the geometric data of every pixel in the frame. There is no exact approach to allocate bits between texture and depth map. As a first attempt, fixed 5:1 ratio was used between texture videos and depth map [7]. Computationally complex full search algorithm was developed by Morvan and Farin [15] assuming that real view exists at synthesis position. Liu et al. [13] modeled view synthesis distortion without considering real view. Yaun et al. [30, 31] proposed a model-based optimal bit allocation strategy. They modeled bit allocation as convex optimization problem, and Lagrange multiplier method is used to find optimal solution. A quadratic model between texture video quantization step and depth map quantization step is described in [32]. Shao et al. [21] reported a distortion model between bit-rate and view synthesis distortion and bit-rate ratio is computed through optimization. They performed rate control at view, texture/depth, and frame level. A fast bit allocation method without pre-encoding was proposed by Oh et al. [18]. Adaptive bit allocation was proposed by Yang et al. [28] in which bit rate is adjusted between the views and texture/depth depending on variations in virtual view quality. All these methods use mean square error (MSE) to measure view synthesis distortion. However, human visual system is highly adapted to acquire structural information and structural similarity (SSIM) index is the quality metric that gives better approximation to perceived image quality. In this paper, we proposed a SSIM-based joint bit allocation method for 3D video.

Distortion metrics such as sum-of-squared error(SSE) or sum-of-absolute differences (SAD) used in rate-distortion optimization (RDO) do not contribute to perceptual quality. In conventional 2D video, many SSIM-based RDO schemes have been proposed to improve perceptual quality [4, 5, 11, 12, 14, 19, 29]. In this paper, SSIM is used as distortion metric in mode decision and motion estimation to improve perceptual quality of texture video and depth map and thus synthesized view. Here, we used Lagrangian multiplier as derived in [29] and scaled it using an additional empirical factor to enhance SSIM of a virtual view. Further, for bit allocation, we experimentally derived a relation between view synthesis distortion with texture distortion and depth distortion in terms of SSIM. We converted SSIM-based distortion model to MSE-based model using the relation given in [29]. Lagrangian optimization is used to find expressions for quantization steps of texture video and depth map. Proposed RDO and bit allocation algorithm are implemented using

H.264/AVC based 3DV-ATM reference software as well as HEVC based HM reference software.

SSIM-based 3DV RDO is explained in Section 2. SSIM-based distortion model is derived in Section 3. In Section 4, concept of joint bit allocation is described and expressions for texture and depth map quantization steps are derived. Experimental results are given in Section 5 in which, 3DV-ATM encoder results are discussed in Section 5.1, HEVC encoder results in Section 5.2 and Section 5.3 gives a comparison of performance of proposed algorithm using 3DV-ATM and HEVC. Section 6 concludes the paper.

## 2 SSIM-based RDO for 3D video

The process of RDO aims to achieve trade-off between the bitrate required and the distortion in a reconstructed video for a given rate constraint $R_c$. A classical approach, Lagrangian optimization, combines rate and distortion using Lagrange multiplier $\lambda$ to form a Lagrangian rate-distortion function J (1a). Optimum values of rate and distortion are obtained by minimizing the cost function J.

$$\begin{aligned} &\min D \\ &s.t. R \leq R_c \\ J &= D + \lambda R \qquad (1a) \\ \frac{dJ}{dR} &= \frac{dD}{dR} + \lambda = 0 \\ \frac{dD}{dR} &= -\lambda \qquad (1b) \end{aligned}$$

where $\lambda = 0.85 \times 2^{\frac{(QP-12)}{3}}$ and $QP$ is the quantization parameter.

The Lagrangian multiplier $\lambda$ plays a significant role in finding optimum values of rate and distortion. This can be illustrated using RD curve (Fig. 1) that shows a plot of distortion against changing rate. The $\lambda$ is the slope of the RD curve and defines the operating point. As $\lambda$ value changes, operating point also changes on RD curve. At operating point B as shown in Fig. 1, R and D will be optimum minimizing the value of J. RDO is applied for mode decision and motion estimation. Initially, motion estimation RDO is carried out as in (2a), where $D_{ME}$ is the prediction error measured using SSE and $R_{ME}$ is the number of bits required to represent motion vectors. This is followed by mode decision RDO as in (2b), where $D_{MD}$ is distortion between original and reconstructed block which is measured using SSE, and $R_{MD}$ is estimated bitrate of the associated mode.

$$J_{ME} = D_{ME} + \lambda_{ME} R_{ME} \qquad (2a)$$

$$J_{MD} = D_{MD} + \lambda_{MD} R_{MD} \qquad (2b)$$

The 3D video with texture plus depth format has to be efficiently coded to ensure better display quality that gets affected by view synthesis distortion. This is possible with proper RDO technique where encoder chooses the mode having reduced distortion with available bits. As our objective is to improve perceptual quality of synthesized view, distortion metrics like SSE and SAD need to be replaced by appropriate metric. So SSIM-based mode decision and motion estimation RDO is implied for 3D video [9]. SSIM index that considers the

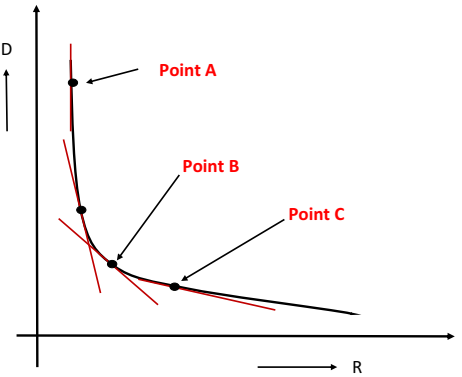


**Fig. 1** Lagrange multiplier on RD curve

similarities of local luminances, contrasts, and structures between two image blocks $x$ and $y$, is defined as in (3) [26].

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{3}$$

where $\mu_x$ and $\sigma_x$ are mean and standard deviation of block $x$ respectively, $\mu_y$ and $\sigma_y$ are mean and standard deviation of block $y$ respectively, and $\sigma_{xy}$ is cross-correlation between image blocks. $C_1$ and $C_2$ are the constants used to limit the range of SSIM values when mean and variance are close to zero. SSIM gives similarity measure and thus dSSIM is used as a distortion metric given by

$$dSSIM = \frac{1}{SSIM} \tag{4}$$

Yeo et al. [29] related dSSIM to MSE (5) and derived expression for Lagrange multiplier (6) that is used in motion estimation and mode decision RDO of every macroblock (MB). This replacement avoids the computation of SSIM for each block. We incorporated this method of SSIM-based RDO for 3D video and experimentally it is found that performance is degraded as slope is near to vertical axis (Point A in Fig. 1) on the RD curve. To improve the performance an empirical scaling factor $S_f$ is added (7). The $\lambda_{new}$ is used in motion estimation and mode decision for both texture and depth map to improve SSIM.

$$dSSIM \approx 1 + \frac{MSE}{2\sigma_x^2 + C_2} \tag{5}$$

$$\lambda_i = \frac{2\sigma_{x_i}^2 + C_2}{\exp\left(\frac{1}{M}\sum_{j=1}^{M}\log(2\sigma_{x_j}^2 + C_2)\right)}\lambda_{SSE} \tag{6}$$

$$\lambda_{new} = \frac{2\sigma_{x_i}^2 + C_2}{S_f\left(\exp\left(\frac{1}{M}\sum_{j=1}^{M}\log(2\sigma_{x_j}^2 + C_2)\right)\right)}\lambda_{SSE} \tag{7}$$

## 3 SSIM-based distortion model

The 3D video uses texture plus depth format to reduce the coding complexity by reducing the number of input videos. This introduces an additional process of virtual view synthesis to generate intermediate views at decoding end. A scheme of virtual view synthesis is shown in Fig. 2.

At the encoding side, three out of five (view 1, 3 and 5) texture videos and corresponding depth maps are coded and transmitted. These views are decoded and used to generate intermediate views 2 and 4. In Fig. 2, texture and depth views 1 and 3 are used to generate view 2, and similarly views 3 and 5 are used to generate view 4. Virtual view synthesis of a frame from the sequence *Balloons* is shown in Fig. 3.

The 3D warping and blending are the two stages in view synthesis. In 3D warping, a pixel in the reference view (existing view) is converted to 3D coordinate and then to virtual view (generated view). A reference view pixel $(u_r, v_r)$ is converted to 3D world point $(x_w, y_w, z_w)$ and then to target pixel $(u_v, v_v)$ as given in (8) and (9) respectively [23].

$$[x_w, y_w, z_w]^{\mathrm{T}} = R_{3X3,r}^{-1}\left(Z_{c,r}A_{3X3,r}^{-1}[u_r, v_r, 1]^{\mathrm{T}} - t_{3X1,r}\right) \tag{8}$$

$$Z_{c,v}[u_v, v_v, 1]^{\mathrm{T}} = A_{3X3,v}\left(R_{3X3,v}[x_w, y_w, z_w]^{\mathrm{T}} + t_{3X1,v}\right) \tag{9}$$

where $R$ is a rotation matrix, $t$ is the translation vector, $A$ is an intrinsic matrix of the camera, and $Z$ is the depth calculated from the depth maps. Since pixel mapping from 2D to 3D is not one to one, holes are created in the left image (Fig. 4b) if right image is taken as reference and vice versa. These holes can be filled with a process called blending which is given by (10).

$$I_v(x, y) = w_L I_L(x_L, y_L) + w_R I_R(x_R, y_R) \tag{10}$$

where $w_L = \frac{I_R}{(I_L+I_R)}$, $w_R = \frac{I_L}{(I_L+I_R)}$, $I_L$ is the baseline distance between left reference and virtual views, and $I_R$ is the baseline distance between right reference and virtual views. Warping can be efficient with an accurate depth map. Inaccuracy in the original depth maps or distortion in encoded depth maps due to lossy compression techniques cause distortions

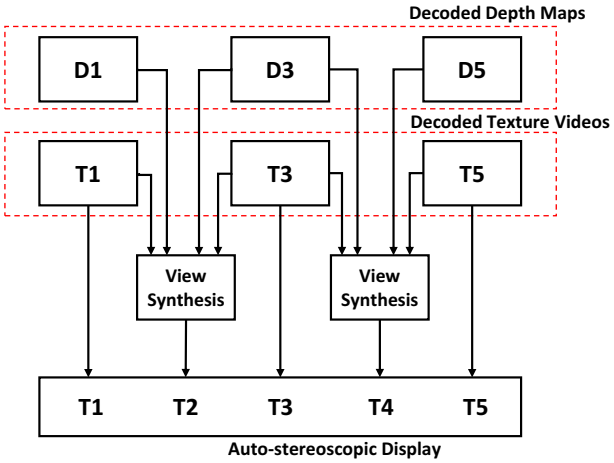


**Fig. 2** Decoding end of 3DV system that uses miltiview video plus depth format with three input texture videos and depth maps

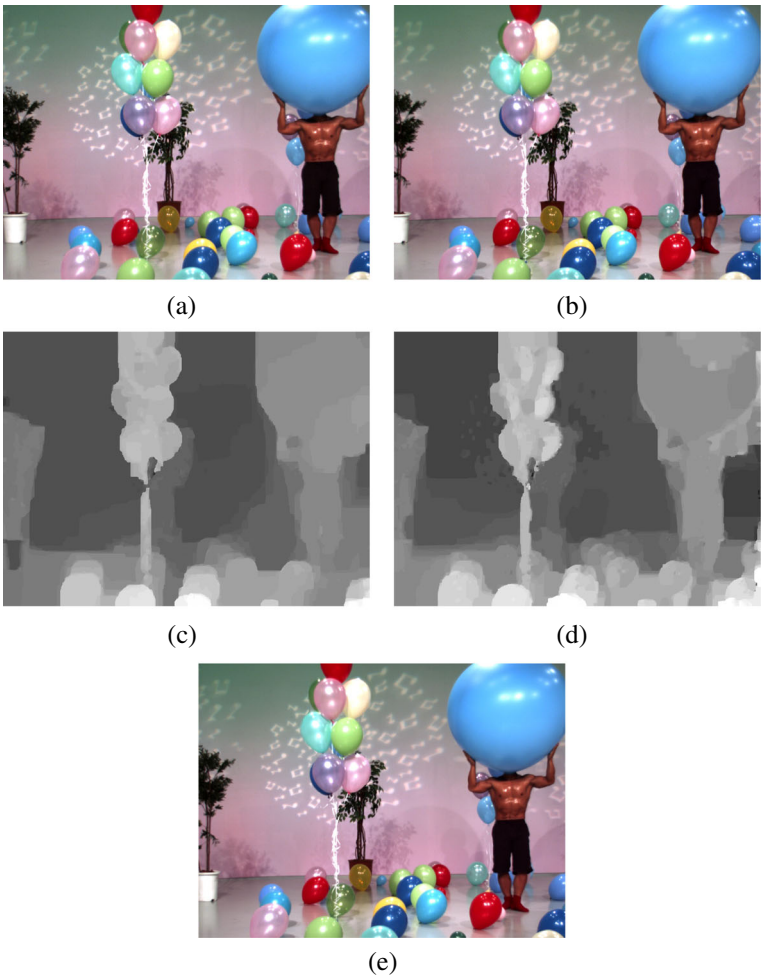


**Fig. 3** Virtual view synthesis **a** Texture frame 0 of view 1, **b** Texture frame 0 of view 3, **c** Depth map frame 0 of view 1, **d** Depth map frame 0 of view 3, and **e** Synthesized frame 0 of view 2

in virtual view. Also, texture distortion directly affects the quality of virtual view. Thus total virtual view distortion consists of texture distortion and depth distortion. If $S_v$ is the virtual synthesized image by original texture and depth video, $\overline{S}_v$ is the virtual image synthesized by original texture and compressed depth map, $\widetilde{S}_v$ is the virtual image synthesized by compressed texture and original depth, then view synthesis distortion is derived in [20] as in (11). Also, authors analyzed the effect of texture and depth distortion in [31] as in (12).

$$D_v \approx E[(S_v - \overline{S}_v)^2] + E[S_v - \widetilde{S}_v)^2] \tag{11}$$

$$D_v = AD_t + BD_d + C \tag{12}$$

where $D_v$ is the view synthesis distortion, $D_t$ is the texture distortion and $D_d$ is the depth distortion. $A$, $B$, and $C$ are the parameters that depend on the compression distortion.


**Fig. 4** Virtual view generated from **a** right view, **b** left view

For improving perceptual quality, earlier view synthesis distortion model (11) and (12) is modified to SSIM-based distortion model. The view synthesis distortion model was experimentally determined by encoding 3D video sequences. Texture videos and depth maps were encoded with quantization parameters (QPs) ranging from 20 to 44. View synthesis distortion ($dSSIM_v$) is computed between the original virtual view (generated by original texture and depth video) and the distorted virtual view (generated by compressed texture and depth video). Texture video distortion ($dSSIM_t$) is computed between the original and compressed texture video. Depth map distortion ($dSSIM_d$) is calculated between original and compressed depth map. Relationship between dSSIM of virtually synthesized view, texture video, and depth map is a planar model as shown in Fig. 5 and can be defined by (13).

$$dSSIM_v = a \cdot dSSIM_t + b \cdot dSSIM_d + c \tag{13}$$

where $a$, $b$, and $c$ are model parameters.

## 4 Joint bit allocation

Effective bit allocation to have good quality synthesized views is a challenging task. One of the methods for bit allocation is through finding optimal QP pairs. Thus, bit allocation tries

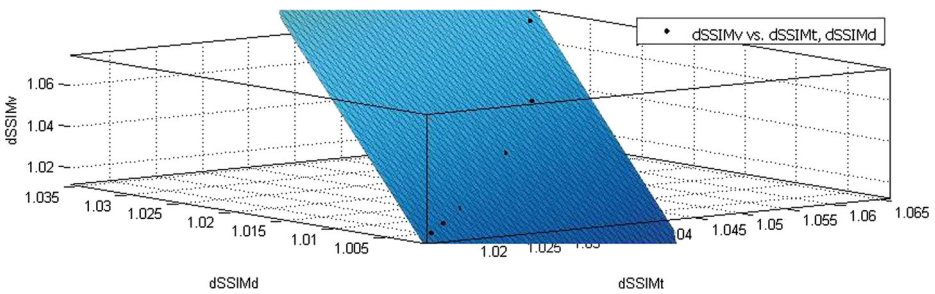


**Fig. 5** Relationship between dSSIM of virtual view, texture video and depth map

to find optimal bitrates (under total rate constraint $R_c$) for texture video and depth maps such that view synthesis distortion is minimized (14).

$$\min_{(R_t,R_d)} D_v \quad s.t. R_t + R_d \leq R_c \tag{14}$$

where $D_v$ is synthesis distortion (12), $R_t$ and $R_d$ are the bitrates of texture video and depth map respectively. In order to improve perceptual quality, bit allocation problem is stated as

$$\min_{(R_t,R_d)} dSSIM_v \quad s.t. R_t + R_d \leq R_c \tag{15}$$

where, $dSSIM_v$ is synthesis distortion measured using SSIM as a distortion metric. Solving bit allocation problem is nothing but finding optimum value of quantization step for both texture and depth video. As SSIM and quantization step cannot be related in closed form [3], using the SSIM-MSE relation (5) $dSSIM_v$ in terms of MSE is derived as

$$dSSIM_v = \frac{a}{2\sigma^2{}_{x_t} + C_2} D_t + \frac{b}{2\sigma^2{}_{x_d} + C_2} D_d + z \tag{16}$$

$$dSSIM_v = p_1 D_t + p_2 D_d + c \tag{17}$$

where $p_1 = \frac{a}{2\sigma^2{}_{x_t}+C_2}$ and $p_2 = \frac{b}{2\sigma^2{}_{x_d}+C_2}$.

With (17), distortion of both texture and depth video can be expressed as a function of quantization step. To obtain Distortion-Quantization (D-Q) relation, texture and depth maps were pre-encoded with different quantization parameters. The D-Q models of texture video and depth map are given in (18a) and (18b) respectively. Rate-Quantization (R-Q) relation is assumed to be linear as in H.264/AVC. R-Q models for both texture and depth map are given in (18c) and (18d) respectively, and are verified experimentally.

$$D_t = \alpha_t Q_t + \beta_t \tag{18a}$$

$$D_d = \alpha_d Q_d + \beta_d \tag{18b}$$

$$R_t = a_t Q_t^{-1} + b_t \tag{18c}$$

$$R_d = a_d Q_d^{-1} + b_d \tag{18d}$$

where $a_t$, $a_d$, $b_t$, $b_d$, $\alpha_t$, $\alpha_d$, $\beta_t$, $\beta_d$ are the parameters calculated from pre-encoding the texture video and the depth sequences. $Q_t$ and $Q_d$ are quantization steps of texture video and depth map respectively. Bit allocation problem is framed (19) and minimized to get optimum values of $Q_t$ and $Q_d$ as

$$\min(p_1 D_t + p_2 D_d) \quad s.t. (a_t Q_t^{-1} + b_t + a_d Q_d^{-1} + b_d) \leq R_c \tag{19}$$

$$Q_t = \frac{a_t + \sqrt{\frac{K_1 a_t a_d}{K2}}}{R_c - b_t - b_d} \tag{20a}$$

$$Q_d = \sqrt{\frac{K_2 a_d}{K_1 a_t}} Q_t \tag{20b}$$

where $K_1 = p_1\alpha_t$ and $K_2 = p_2\alpha_d$

## 5 Experimental results

In this section, we discuss the performance of joint bit allocation algorithm proposed for 3D video to improve perceptual quality. 3D sequences Kendo, Balloons and Breakdancer [8, 33] of size 1024 × 768 are used. We evaluated the performance of SSIM-based RDO and bit allocation using H.264/AVC based 3DV-ATM encoder and HEVC based HTM encoder.

### 5.1 3DV-ATM results

3DV-ATM used for encoding the 3D sequences is based on H.264/AVC encoder and it accomplishes higher coding efficiency through RDO. Each MB is divided into sub-blocks with different sizes in intra as well as inter mode prediction. The block partition size is considered to be 4 × 4 and 16 × 16 in intra prediction and 16 × 8, 8 × 16, 8 × 8 for inter mode. An 8 × 8 block is subdivided into 8 × 4, 4 × 8, and 4 × 4. Every mode is coded, reconstructed and the best mode has to be selected considering the factors: (i) amount of bits used for coding (rate) and (ii) quality of reconstructed block.

In this section, we discuss the performance of SSIM-based RDO and joint bit allocation algorithm proposed for 3D video to improve perceptual quality. For encoding, we used Nokia's 3DV-ATMv5.lr2 reference software [1]. Virtual view synthesis is done using View Synthesis Reference Software 3.0 [25]. In both SSIM-based RDO and joint bit allocation, encoder parameters used are as given in Table 1.

#### *5.1.1 SSIM-based RDO in 3DV-ATM*

Performance of SSIM-based RDO is compared with SSE-based RDO of original 3DV-ATM encoder. In both methods, SSIM is calculated between the virtual views generated by original views and reconstructed views. RD curves are plotted with rate along horizontal axis and SSIM along vertical axis. Figure 6a shows the RD curves for Kendo sequence and both SSE and SSIM-based RDO have almost the same performances. Since depth maps are not displayed, we performed SSE-based RDO on depth map and SSIM-based RDO on texture video to improve perceptual quality of synthesized view (Fig. 6b). However, in Balloons (Fig. 7) and Breakdancer (Fig. 7a and b) sequences, SSIM-based RDO finds significant improvement. Performance of proposed method is also evaluated using BD-rate figures [2] as shown in Table 2. BD-rate figures compute average percentage of saving in bit-rate along with average gain in SSIM.

**Table 1** Encoder parameter setting

| Parameter | Setting |
|---|---|
| FrameRate | 30 Frames/s |
| ProfileIDC | 139 |
| QP | 20–40 |
| SearchRange | 96 |
| 3DVCodingOrder | T0D0D1T1 |
| SymbolMode | CABAC |

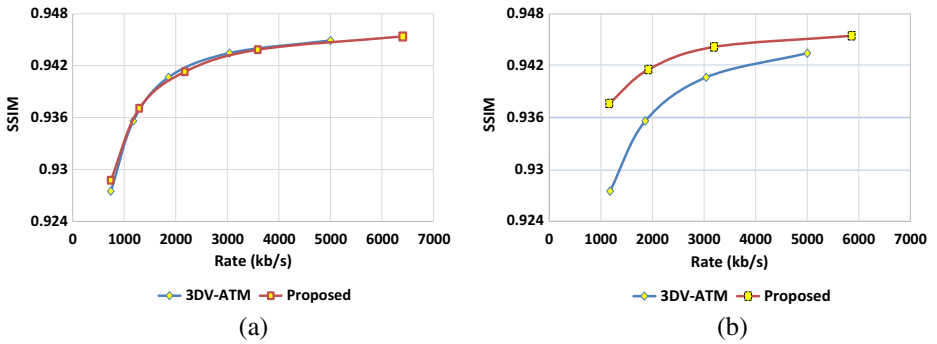


**Fig. 6** SSIM Vs Rate for Kendo sequence when view 2 synthesized from views 1 and 3, when SSIM is used as distortion metric for **a** both texture and depth, **b** only for texture

### *5.1.2 SSIM-based joint bit allocation in 3DV-ATM*

SSIM-based joint bit allocation algorithm is evaluated and compared with model-based joint bit allocation. Model-based joint bit allocation proposed in [31] is evaluated in two steps. First, texture and depth sequences are encoded and model parameters are determined. Using model parameters, quantization step of texture video and depth map are calculated. In the second step, using new optimized values of quantization step, texture and depth sequences are encoded. Here, bit allocation is done at sequence level.

SSIM-based joint bit allocation also requires pre-encoding and computation of model parameters. Quantization parameters are calculated for each MB. In calculating QP for texture MB, variance of corresponding depth MB is computed and vice versa. For both model-based and SSIM-based bit allocation, we evaluated SSIM of virtual views and plotted RD graph as shown in Fig. 8. BD-rate is calculated between the RD curves to get average gain in terms of SSIM ($\Delta$SSIM) and bit-rate reduction ($\Delta$Rate) as shown in Table 3.

Using SSIM as distortion metric for 3D video, we performed RDO followed by bit allocation. Objective evaluation shows the improvement in perceptual quality at reduced rate. To make visual appearance better, structural information is to be maintained. To illustrate this, 50th and 60th frame of breakdancer sequence from decoded sequences of SSIM-based and model-based bit allocation algorithm are shown in Figs. 9 and 10. In 50th frame of Fig. 9a,

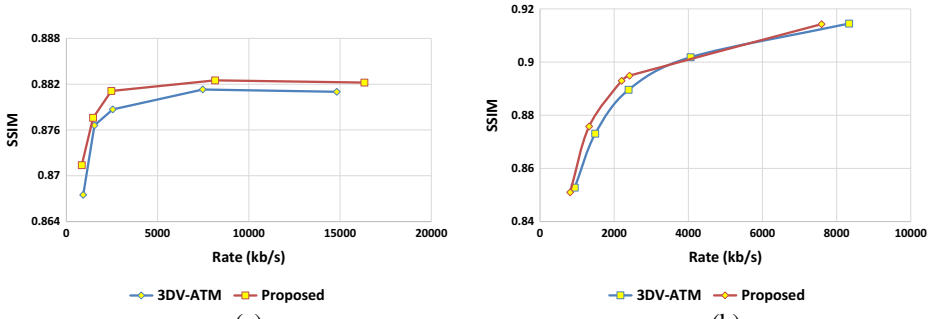


**Fig. 7** SSIM Vs Rate for **a** Balloons sequence and **b** Breakdancer sequence. In both cases view 2 is synthesized from views 1 and 3 with SSIM as distortion metric

**Table 2** Comparison of SSIM-based RDO with SSE-based RDO

| Sequence | ΔSSIM | ΔRate |
|---|---|---|
| Kendo | 0.0015 | −6.2971 |
| Balloons | 0.0017 | −7.0832 |
| Breakdancer | 0.0048 | −15.4161 |

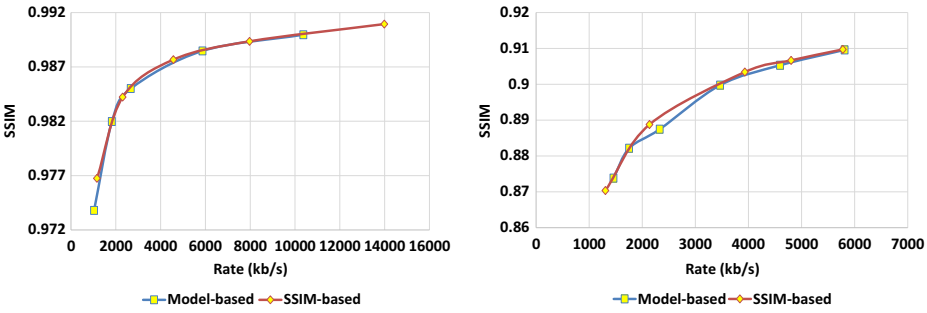


**Fig. 8** RD curve of bit allocation algorithm for **a** Kendo, **b** Breakdancer

**Table 3** Improved SSIM and Rate reduction of SSIM-based bit allocation

| Sequence | ΔSSIM | ΔRate |
|---|---|---|
| Kendo | 0.0002 | −2.5166 |
| Breakdancer | 0.0015 | −5.4418 |

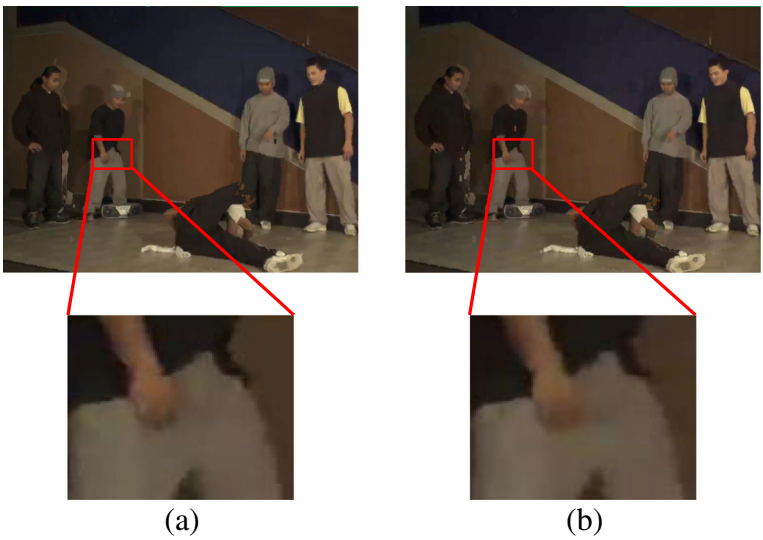


**Fig. 9** 50th frame from decoded sequence using **a** SSIM-based bit allocation, **b** Model-based bit allocation

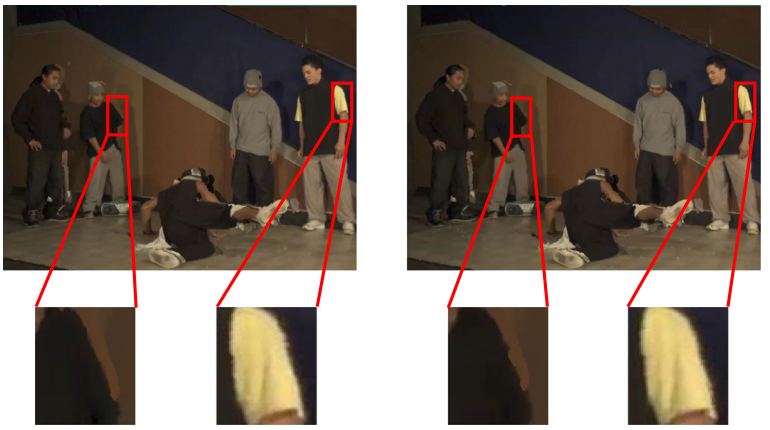


**Fig. 10** 60th frame from decoded sequence using **a** SSIM-based bit allocation, **b** Model-based bit allocation

edges are preserved and structure is retained (particularly man's finger) from SSIM-based bit allocation as compared to model-based bit allocation in Fig. 9b. A similar comparison is shown in Fig. 10.

### *5.1.3 Subjective evaluation*

Since the aim of our proposed algorithm is to improve visual quality, we conducted subjective evaluation. For the evaluation, the Kendo and Breakdancer sequences encoded with SSIM-based bit allocation and model-based bit allocation were used. Ten viewers evaluated the quality of the video in each case. Each pair of videos were shown five times. Participants were asked to rate the videos on the scale of one to five with the following notation: 5—Excellent, 4—Good, 3—Fair, 2—Poor, 1—Very poor. Mean opinion score (MOS) and standard deviation (SD) were obtained as shown in Table 4. In both the sequences, SSIM-based bit allocation obtained higher MOS and better standard deviation.

### *5.1.4 Temporal SSIM*

Objective quality metrics used to determine the video quality use the spatial information within a frame. For a video, quality depends both on spatial information and temporal information. Thus, it is necessary to use a quality metric that measures both spatial and temporal distortion. As we aim to improve perceptual quality in 3D video, we measured spatio-temporal SSIM for quality assessment of virtual view synthesized videos.

**Table 4** Mean opinion score and standard deviation

| Sequence | SSIM-based bit allocation | | Model-based bit allocation | |
|---|---|---|---|---|
| | MOS | SD | MOS | SD |
| Kendo | 4.24 | 0.82 | 4.0 | 0.81 |
| Breakdancer | 3.4 | 0.92 | 3.3 | 0.97 |

Wang et al. [27] proposed a quality metric to measure perceptual quality which takes care of spatio-temporal structural information in a video and we used this metric to evaluate both spatial and temporal quality of synthesized views. Gradient is computed using Sobel kernel in all the three directions x-y, x-t and y-t and spatio-temporal gradient magnitude is determined. A threshold is set and compared with the gradient magnitude to determine whether the pixel is salient or not. SSIM (21) is computed only on the patches surrounding the salient pixel at the center. We computed temporal SSIM for synthesized views resulted from SSIM-based bit allocation and model-based bit allocation for Kendo and Balloons sequences. SSIM-based bit allocation has better temporal SSIM as in Fig. 11.

$$SSIM_{xyt} = (SSIM_{xy} + SSIM_{xt} + SSIM_{yt})/3 \tag{21}$$

## 5.2 HEVC results

The High Efficiency Video Coding (HEVC) is the newly developed video coding standard that is replacing existing H.264/AVC standard. The development of hardware technology for acquisition resulted in better quality of video that further requires efficient coding algorithms. For example, HD videos and other advanced video formats need to be compressed with higher coding efficiency. Another concern of HEVC standard is to utilize parallel processing architectures [22]. HEVC is extended to encode 3D video where either stereo or multiview video plus depth format can be used as input [17].

HEVC uses advanced approach of quad tree coding where each picture is divided into coding tree units (CTUs). A CTU is equivalent to a macroblock in previous coding standards. Each CTU is divided into coding units (CUs). In HEVC, three different coding modes are used: intra coding mode, inter coding mode and merge mode. A CU can have variable block size ranging from $8 \times 8$ to $64 \times 64$. Coding mode is selected at CU level. A CU is again divided into prediction units (PUs). All PUs are coded in the same coding mode.

Intra coding mode has 35 different variations which includes one planar mode, one DC mode and remaining angular modes. The partition type includes 2Nx2N and NxN. Inter prediction has symmetric and asymmetric block partitions which include: 2Nx2N, NxN, Nx2N, 2NxN, 2NxnU, 2NxnD, nLx2N, nRx2N. For a CTU, RD cost is computed for all modes and the mode with minimum cost is considered to be optimal CTU.

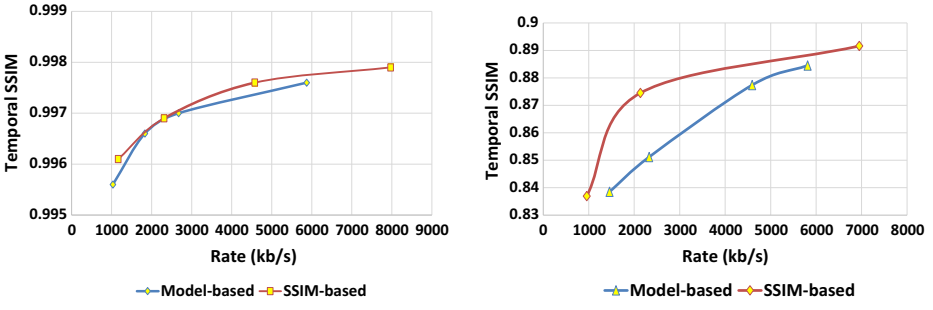


**Fig. 11** Comparison of temporal SSIM in **a** Kendo **b** Breakdancer

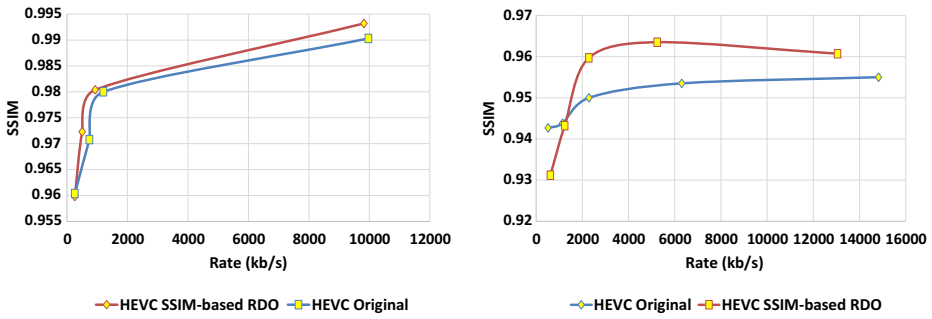


**Fig. 12** SSIM-based RDO **a** Kendo and **b** Balloons

In HEVC, distortion metrics used for prediction cost are SAD or SATD and SSE. For mode decision, SSE is used as distortion metric. RD cost functions used for computing prediction cost are given in (22a) to (22b).

$$J_{pred} = SAD + \lambda_{pred} \cdot B_{pred} \tag{22a}$$

$$J_{pred} = SATD + \lambda_{pred} \cdot B_{pred} \tag{22b}$$

where $B_{pred}$ is bit cost required for encoding the block. Similar to 3DV-ATM, we aim to replace traditional distortion metrics by SSIM to improve perceptual quality of 3D video which are encoded using HEVC. Qi et al. [19] used SSIM-MSE relation and new Lagrange multiplier derived in [29] to improve perceptual quality of 2D video using HEVC. Thus, we extended the same methodology including an additional empirical scaling factor to the modified Lagrange multiplier as in (7). This resulted in improvement of the perceptual quality in 3D video. Optimized bit allocation between texture and depthmap results in synthesized views of better quality. The objective is to minimize synthesis distortion while improving visual quality at available rate which includes texture rate and depth rate. This requires distortion model in terms of perceptual metric. We assumed the linear dSSIM distortion model as explained in Section 3 and used optimum value of quantization parameters as derived in Section 4.

### *5.2.1 SSIM-based RDO and bit allocation in HEVC*

SSIM-based RDO is implemented using HM-16.2 HEVC reference software [10]. Experiments were conducted using Kendo and Balloons sequences. View synthesis is performed using rendering algorithm available in HM reference software. Proposed SSIM-based RDO is compared with SSE-based original HEVC as shown in Fig. 12. SSIM-based RDO in HEVC performs better as compared to original HEVC. BD-rate comparison in Table 5 shows the improved SSIM at reduced rate.

**Table 5** Comparison of SSIM-based RDO with SSE-based RDO in HEVC

| Sequence | ΔSSIM | ΔRate |
|---|---|---|
| Kendo | 0.0002 | −31.3885 |
| Balloons | 0.0053 | −24.0606 |

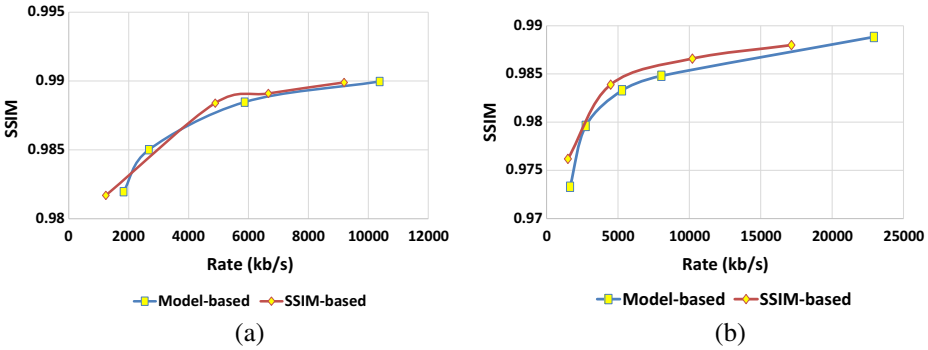


**Fig. 13** Bit allocation in HEVC **a** Kendo **b** Balloons

**Table 6** Comparison of SSIM-based bit allocation with model-based bit allocation

| Sequence | ΔSSIM | ΔRate |
|---|---|---|
| Kendo | 0.0016 | −42.0602 |
| Balloons | 0.0015 | −24.8328 |

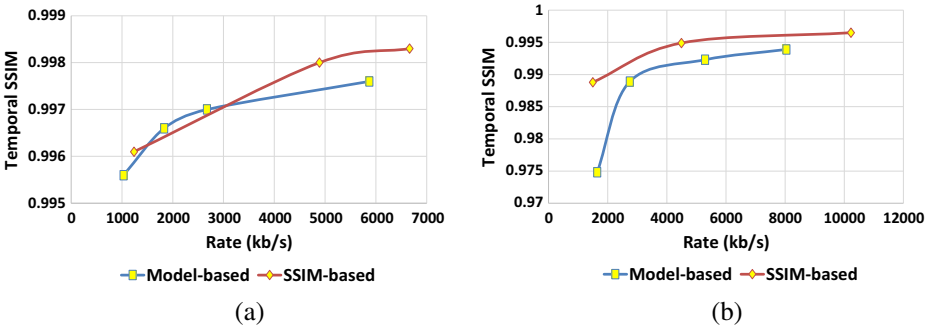


**Fig. 14** Comparison of temporal SSIM **a** Kendo **b** Balloons

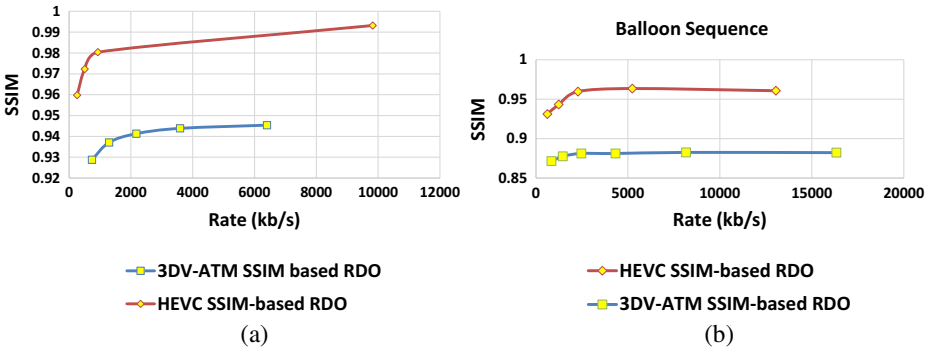


**Fig. 15** Bit allocation **a** Kendo **b** Balloons

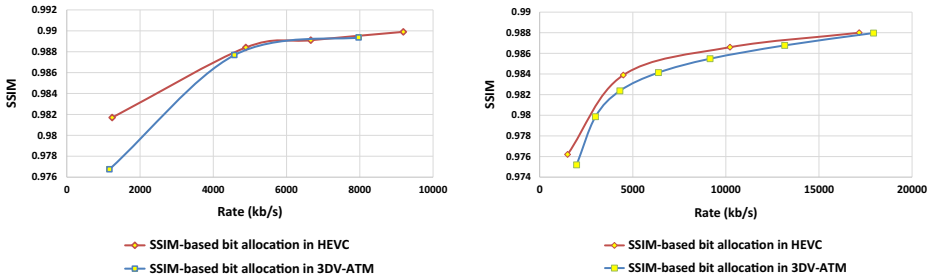


**Fig. 16** Bit allocation **a** Kendo **b** Balloons

Proposed SSIM-based bit allocation method results in improved visual quality of synthesized views. We pre-encoded the video sequences and calculated the parameters required for computation of quantization parameters. Sequences are then encoded using optimum value of quantization parameters. Results of proposed algorithm is compared with model-based bit allocation as shown in Fig. 13 along with BD-rate comparison in Table 6.

We computed temporal SSIM for both Kendo and Balloons sequence to verify the spatio-temporal quality of SSIM-based bit allocation. Results shown in Fig. 14 indicate improved perceptual quality with SSIM-based bit allocation.

### 5.3 Performance comparison of proposed algorithms on 3DV-ATM and HEVC encoders

Proposed SSIM-based RDO and bit allocation are implemented in two different encoders: H.264/AVC based 3DV-ATM encoder and HEVC encoder. We compare the performance of proposed algorithms on these encoders as shown in Figs. 15 and 16. As HEVC is designed for higher coding efficiency, it performs better compared to 3DV-ATM encoder. Thus, in both SSIM-based RDO and SSIM-based bit allocation, HEVC encoder achieves better SSIM at lower rates compared to 3DV-ATM due to its higher coding efficiency.

## 6 Conclusion

Bit allocation algorithm in 3D video ensures improvement in quality of the synthesized view. In this paper, we proposed SSIM-based bit allocation algorithm to enhance perceptual quality of 3D video using 3DV-ATM as well as HEVC. Initially, SSIM is used as distortion metric in mode decision and motion estimation of both texture and depth video where Lagrange multiplier is adjusted to improve SSIM and thus, visual quality. In SSIM-based joint bit allocation, we experimentally derived view synthesis distortion model using SSIM as distortion metric and converted it into MSE-based model. Further, the model is used to find optimal quantization parameters. Both objective and subjective evaluation show the improved perceptual quality with SSIM-based method.

SSIM-based RDO stage can be addressed more efficiently by automating the selection of empirical scaling factor. In bit allocation algorithm, we assumed that synthesis distortion

is linearly dependent on texture and depth distortion. However, influence of depth distortion on synthesis distortion varies depending upon texture details. Linear depth distortion can be replaced by nonlinear distortion model for more accurate results.

## References


1. 3DV-ATM Reference Software 3DV-ATMv5.lr2. Available at: http://mpeg3dv.nokiaresearch.com/svn/mpeg3dv/tags/3DV-ATMv5.1r2/, [Online; accessed on 06-January-2017]
2. Bjontegaard G Calculation of average PSNR differences between RD - curves. ITU-TQ.6/SG16 VCEG 13th Meeting, Available at: http://wftp3.itu.int/av-arch/video-site/0104_Aus/
3. Chen HH, Huang YH, Su PY, Ou TS (2010) Improving video coding quality by perceptual rate-distortion optimization. In: IEEE international conference on multimedia and expo (ICME), 2010. IEEE, pp 1287–1292
4. Chen Z, Lin W, Ngan KN (2010) Perceptual video coding: challenges and approaches. In: Proceedings of IEEE international conference on multimedia and expo (ICME), pp 784–789
5. Cui Z, Gan Z, Zhu X (2011) Structural similarity optimal MB layer rate control for H. 264. In: Proceedings of IEEE international conference on wireless communications and signal processing (WCSP), pp 1–5
6. Fehn C (2003) A 3D-TV approach using depth-Image-Based Rendering (DIBR). In: Proceedings of VIIP, vol 3
7. Fehn C (2004) Depth-image-based rendering (DIBR), compression, and transmission for a new approach on 3D-TV. In: Electronic imaging 2004. International Society for Optics and Photonics, pp 93–104
8. Fujii Laborotory, Nagoya University. Available at: http://www.fujii.nuee.nagoya-u.ac.jp/multiview-data/, [Online; accessed on 06-January-2017]
9. Harshalatha Y, Biswas PK (2016) Rate distortion optimization using SSIM for 3D video coding. In: International conference on pattern recognition (ICPR)
10. HM Reference Software HTM-16.2. Available at: https://hevc.hhi.fraunhofer.de/trac/3d-hevc/browser/3DVCSoftware/tags/HTM-16.2, [Online; accessed on 05-September-2017]. https://hevc.hhi.fraunhofer.de/trac/3d-hevc/browser3DVCSoftware/tags
11. Huang YH, Ou TS, Chen HH (2010) Perceptual-based coding mode decision. In: Proceedings of IEEE international symposium on circuits and systems (ISCAS), pp 393–396
12. Huang YH, Ou TS, Su PY, Chen HH (2010) Perceptual rate-distortion optimization using structural similarity index as quality metric. IEEE Trans Circuits Syst Video Technol 20(11):1614–1624
13. Liu Y, Huang Q, Ma S, Zhao D, Gao W (2009) Joint video/depth rate allocation for 3D video coding based on view synthesis distortion model. Signal Process Image Commun 24(8):666–681
14. Mai ZY, Yang CL, Po LM, Xie SL (2005) A new rate-distortion optimization using structural information in H. 264 I-Frame Encoder. In: Advanced concepts for intelligent vision systems. Springer, pp 435–441
15. Morvan Y, Farin D (2007) Joint depth/texture bit-allocation for multi-view video compression. In: Picture coding symposium (PCS)
16. Muller K, Merkle P, Wiegand T (2011) 3-D video representation using depth maps. Proc IEEE 99(4):643–656
17. Muller K, Schwarz H, Marpe D, Bartnik C, Bosse S, Brust H, Hinz T, Lakshman H, Merkle P, Rhee H et al (2013) 3D high efficiency video coding for multi-view video and depth data
18. Oh BT, Lee J, Park DS (2013) Fast joint bit-allocation between texture and depth maps for 3D video coding. In: IEEE international conference on consumer electronics (ICCE). IEEE, pp 193–194
19. Qi J, Li X, Su F, Tu Q, Men A (2013) Efficient rate-distortion optimization for HEVC using SSIM and motion homogeneity. In: Picture coding symposium (PCS), 2013. IEEE, pp 217–220
20. Shao F, Jiang GY, Yu M, Li FC (2011) View synthesis distortion model optimization for bit allocation in three-dimensional video coding. Opt Eng 50(12):120502–120502

21. Shao F, Jiang G, Lin W, Yu M, Dai Q (2013) Joint bit allocation and rate control for coding multi-view video plus depth based 3D video. IEEE Trans Multimed 15(8):1843–1854
22. Sullivan GJ, Ohm J, Han WJ, Wiegand T (2012) Overview of the high efficiency video coding (HEVC) standard. IEEE Trans Circuits Syst Video Technol 22(12):1649–1668
23. Tian D, Lai PL, Lopez P, Gomila C (2009) View synthesis techniques for 3D video. In: Proceedings of the SPIE applications of digital image processing XXXII, vol 7443, pp 74,430T–74,430T
24. Urey H, Chellappan KV, Erden E, Surman P (2011) State of the art in stereoscopic and autostereoscopic displays. Proc IEEE 99(4):540–555
25. View Synthesis Reference Software VSRS3.5. Available at: ftp://ftp.merl.com/pub/avetro/3dv-cfp/software/, [Online; accessed on 06-January-2017]
26. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP (2004) Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process 13(4):600–612
27. Wang Y, Jiang T, Ma S, Gao W (2012) Spatio-temporal ssim index for video quality assessment. In: Visual communications and image processing (VCIP), 2012 IEEE. IEEE, pp 1–6
28. Yang C, An P, Shen L (2016) Adaptive bit allocation for 3D video coding. In: Circuits, systems, and signal processing, pp 1–23
29. Yeo C, Tan HL, Tan YH (2013) On rate distortion optimization using SSIM. IEEE Trans Circuits Syst Video Technol 23(7):1170–1181
30. Yuan H, Chang Y, Li M, Yang F (2010) Model based bit allocation between texture images and depth maps. In: International conference on computer and communication technologies in agriculture engineering (CCTAE), vol 3. IEEE, pp 380–383
31. Yuan H, Chang Y, Huo J, Yang F, Lu Z (2011) Model-based joint bit allocation between texture videos and depth maps for 3-D video coding. IEEE Trans Circuits Syst Video Technol 21(4):485–497
32. Zhu G, Jiang G, Yu M, Li F, Shao F, Peng Z (2012) Joint video/depth bit allocation for 3D video coding based on distortion of synthesized view. In: IEEE international symposium on broadband multimedia systems and broadcasting (BMSB). IEEE, pp 1–6
33. Zitnick CL, Kang SB, Uyttendaele M, Winder S, Szeliski Rs (2004) High-quality video view interpolation using a layered representation. In: ACM transactions on graphics (TOG), vol 23. ACM, pp 600–608

**Harshalatha Y** received the B.E. degree in Electronics & Communication Engineering in the year 1998 from Malnad College of Engineering, Hassan, Karnataka, India and M.Tech degree in Digital Electronics from Visvesvaraya Technological University, Karnataka, India in 2009. She is presently pursuing her Ph.D. studies as a Research Scholar in the Department of Electronics & Electrical Communication Engineering, IIT Kharagpur, India. She is also a faculty member in the Department of Electronics and Communication Engineering at Siddaganga Institute of Technology, Tumkur, Karnataka, India since 2004. Her areas of interests include image and video processing and 3D video coding.

**Prabir Kumar Biswas** received the B.Tech. (Honors) degree in Electronics and Electrical Communication Engineering, the M.Tech. degree in Automation and Control engineering, and the Ph.D. degree in Computer Vision from the Department of Electronics and Electrical Communication Engineering, Indian Institute of Technology (IIT) Kharagpur, India, in 1985, 1989, and 1991, respectively. From 1985 to 1987, he was with Bharat Electronics Ltd., Ghaziabad, India, as a Deputy Engineer. Since 1991, he has been working as a Faculty Member in the Department of Electronics and Electrical Communication Engineering, IIT, where he is currently a Professor. He visited the University of Kaiserslautern, Germany, under Alexander von Humboldt Research Fellowship from March 2002 to February 2003. He has more than 100 research publications in international and national journals and conferences and has filed seven international patents. His areas of interest are image processing, pattern recognition, computer vision, video compression, parallel and distributed processing, and computer networks.