High-Efficiency 3D Depth Coding Based on Perceptual Quality of Synthesized Video

Yun Zhang, *Senior Member, IEEE*, Xiaoxiang Yang, Xiangkai Liu, Yongbing Zhang, Gangyi Jiang, *Member, IEEE*, and Sam Kwong, *Fellow, IEEE*

***Abstract*— In 3D video systems, imperfect depth images often induce annoying temporal noise, e.g., flickering, to the synthesized video. However, the quality of synthesized view is usually measured with peak signal-to-noise ratio or mean squared error, which mainly focuses on pixelwise frame-by-frame distortion regardless of the obvious temporal artifacts. In this paper, a novel full reference synthesized video quality metric (SVQM) is proposed to measure the perceptual quality of the synthesized video in 3D video systems. Based on the proposed SVQM, an improved rate-distortion optimization (RDO) algorithm is developed with the target of minimizing the perceptual distortion of synthesized view at given bit rate. Then, the improved RDO algorithm is incorporated into the 3D High Efficiency Video Coding (3D-HEVC) software to improve the 3D depth video coding efficiency. Experimental results show that the proposed SVQM metric has better consistency with human perception on evaluating the synthesized view compared with the state-of-the-art image/video quality assessment algorithms. Meanwhile, this SVQM metric maintains low complexity and easy integration to the current video codec. In addition, the proposed SVQM-based depth coding scheme can achieve approximately 15.27% and 17.63% overall bit rate reduction or 0.42- and 0.46-dB gain in terms of SVQM quality score on average as compared with the latest 3D-HEVC reference model and the state-of-the-art depth coding algorithm, respectively.**

***Index Terms*— 3D video, synthesized video quality, depth video coding, video quality assessment, temporal distortion.**

Manuscript received January 8, 2016; revised July 16, 2016; accepted September 26, 2016. Date of publication October 5, 2016; date of current version October 25, 2016.This work was supported in part by the National Natural Science Foundation of China under Grant 61471348, Grant 61272289, and Grant U1301257, in part by Guangdong Natural Science Funds for Distinguished Young Scholar under Grant 2016A030306022, in part by the Project for Shenzhen Science and Technology Development under Grant JSGG20160229202345378, in part by Shenzhen Overseas High-Caliber Personnel Innovation and Entrepreneurship Project under Grant KQCX20140520154115027, and in part by the National High Technology Research and Development Program of China under Grant 2014AA01A302. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Catarina Brites.

Y. Zhang, X. Yang, and X. Liu are with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China (e-mail: yun.zhang@siat.ac.cn; xx.yang@siat.ac.cn; xk.liu@siat.ac.cn).

Y. Zhang is with the Shenzhen Key Laboratory of Broadband Network and Multimedia, Graduate School at Shenzhen, Tsinghua University, Shenzhen 518055, China (e-mail: zhang.yongbing@sz.tsinghua.edu.cn).

G. Jiang is with the Faculty of Information Science and Engineering, Ningbo University, Ningbo 315211, China (e-mail: jianggangyi@nbu.edu.cn).

S. Kwong is with the Department of Computer Science, City University of Hong Kong, Hong Kong, and also with the City University of Hong Kong Shenzhen Research Institute, Shenzhen 5180057, China (e-mail: cssamk@cityu.edu.hk).

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/TIP.2016.2615290

I. Introduction

RECENTLY, a promising prospect has been expected for the Three-Dimensional (3D) video system, such as 3D Television (3DTV) and Free-viewpoint Television (FTV), as it can provide viewers with the impressive visual experience of depth perception and interactivity. Over the past decade, Multiview Video plus Depth (MVD) [1] has become the most popular format of representing 3D scene. However, the data volume of multiple views of color and depth videos is many times larger than traditional 2D video. To compress the MVD data more efficiently, the 3D extension of the state-of-the-art High Efficiency Video Coding (HEVC) [2]–[4] standard, known as 3D-HEVC [5], has been developed by the Joint Collaborative Team on 3D Video (JCT-3V) of ITU-T Video Coding Experts Group (VCEG) and ISO/IEC Moving Picture Experts Group (MPEG) since July 2012. As important components of MVD data, the multiview depth images can provide the 3D distance information from an object to the cameras, which are typically utilized to generate arbitrary intermediate views for the 3D video applications by means of Depth Image Based Rendering (DIBR) technique [6]. Therefore, depth video coding has recently attracted more and more attention [7]–[14].

Traditionally, the depth video is regarded as the illumination component of color video and is encoded with the traditional color video encoder [1]. However, since the depth data is not used for display but regarded as supplementary data for view synthesis, the depth video coding performance should be evaluated with the quality of the synthesized intermediate views. Therefore, using the conventional color video codec [2]–[4] to encode the depth video directly is not so efficient since the depth video has different view-spatial-temporal correlations [15], visual properties and functionalities from those of color video [10], [11]. Moreover, synthesized distortion and the corresponding effects on perceptual quality should be taken into account in the coding and optimization processes.

The Peak Signal-to-Noise Ratio (PSNR) or Mean Squared Error (MSE) based methods for estimating video quality of synthesized videos have been investigated in [7]–[14]. Yuan *et al.* [7] analyzed the distortion models of synthesized view and its relationship to color and depth coding bits, and proposed a coarse to fine bit allocation among color and depth channels for high efficiency 3D coding. Fang *et al.* [8] proposed an analytical model to establish the mathematical relationship between synthesized distortion and depth

distortion by analyzing color video properties and the rendering process. Oh and Oh [9] presented a View Synthesis Distortion (VSD) function with the corresponding texture information and applied it to Rate-Distortion Optimization (RDO) for HEVC-compatible 3D video coding. Zhang *et al.* [10] found that depth distortion in color textural area has more severe impacts on the synthesized image quality than that of smooth area. Then, regional bit allocation algorithm is exploited to reasonably assign quantization errors and thus improve the depth video coding efficiency. In [11], due to the many-to-one mapping relation between depth distortion and rendering position errors, an Allowable Depth Distortion (ADD) model was presented to exploit the allowable depth distortion in view synthesis and minimize VSD at a given bit rate. Significantly depth bit reduction was achieved as the ADD model was incorporated in mode decision and motion estimation modules. Meanwhile, fast mode decision and reference selection algorithms exploiting ADD are proposed to lower the depth coding complexity [12]. Kim *et al.* [13] proposed a synthesized view distortion estimation method based on the local video characteristics and derived a new Lagrange multiplier for the RDO of depth video coding. In addition, Synthesized View Distortion Change (SVDC) [5] was proposed for the View Synthesis Optimization (VSO) [14] in 3D-HEVC. However, the above mentioned methods compute frame-by-frame distortion by means of PSNR values in evaluating the quality of the synthesized views, which has not sufficiently considered the effects of synthesized temporal distortion on the perceptual quality of Human Visual Systems (HVS).

Currently, a number of quality assessment metrics and HVS based coding algorithms [16]–[22] have been developed. Wang *et al.* [16] proposed a perceptual divisive normalization video coding scheme, where the distortion was measured with the Structural Similarity Index Metric (SSIM) [23]. Zhao *et al.* [17] proposed a coarse-grain Scalable Video Coding (SVC) approach by using SSIM as the visual quality criterion, which aims to maximize the overall coding performance of the scalable layers. In [18], an MSE based full reference Video Quality Assessment (VQA) metric was designed at first, and then an efficiency rate control algorithm was developed based on this VQA metric. Luo *et al.* [19] proposed a perceptual video coding method based on the Just Noticeable Distortion (JND) model. The quantization component was modified with a JND-based model and the Lagrange multiplier for the RDO process was also derived in terms of the equivalent distortion. Su *et al.* [20] developed new bivariate/correlation Natural Scene Statistical (NSS) models and a convergent cyclopean image model, and applied them in a non-reference Image Quality Assessment (IQA) for stereo-pair images. Shao *et al.* [21] proposed a full-reference stereoscopic images quality assessment method by learning binocular receptive field properties with sparse features. Silva *et al.* [22] developed a training based stereoscopic video quality metric which jointly considers structural distortions, asymmetric blur, spatial and temporal content complexity. However, these metrics are mainly designed for monocular videos or stereo videos. They have not considered the characteristics of 3D synthesized video and can hardly be applied for the depth video coding.

Since the depth video is used to provide scene geometry information in view synthesis instead of being viewed directly, measuring the distortion of depth image only can hardly reflect the perceptual quality of the synthesized video. To handle the quality assessment of synthesized videos, Bosc *et al.* [24] proposed a synthesized image quality assessment metric by considering the specific artifacts located around the disoccluded areas and the contours consistency on object edge. However, it is an IQA method for still image and the temporal distortion of synthesized video was not considered. Hewage and Martini [25] proposed a Reduced-Reference (RR) quality metric for color plus depth 3D video by comparing the edge or contour information of the color and depth images. Jung proposed a Just Noticeable Depth Difference (JNDD) model [26] to evaluate depth sensation, which was then applied to depth image enhancement. A full reference Peak-Signal-to-Perceptible- Temporal-Noise Ratio (PSPTNR) metric was proposed in [27], which evaluated the synthesized view quality by measuring temporal noise. In our previous work [28], a full reference synthesized video quality assessment algorithm measuring the temporal flicker distortion and spatial activity distortion was proposed. However, this approach is too complex to be integrated into the current 3D depth coding optimization.

In this paper, a full reference Synthesized Video Quality Metric (SVQM) is proposed and has better consistency with human perception on evaluating the synthesized view in 3D video system. Then, a SVQM based depth coding scheme is proposed, achieving promising synthesized video quality and compression ratio gain. The remaining sections of the paper are organized as follows, motivations and analyses are presented in Section II. Section III presents the proposed 3D video coding system framework. Section IV and Section V present the proposed SVQM model and the SVQM based 3D depth coding optimization, respectively. In Section VI, the proposed SVQM and the depth coding algorithm are validated, and then their performances are compared with the state-of-the-art algorithms. Finally, conclusions are drawn in Section VII.

II. Motivations and Analyses

In 3D video system, the depth videos are used to synthesize the virtual views and the quality of synthesized view relies on the quality of both depth and texture videos. However, different from conventional coding distortions in 2D video, distortions in depth videos often result in spatial distortion (e.g. contour displacement, ring artifact) and temporal noise (e.g. flickering) in synthesized video. Fig.1 shows the temporal inconsistency or flickering distortion of the synthesized view video, where the top and bottom figures show the pixel value fluctuation at positions that supposed to be in a stationary region in the original and synthesized views, respectively. We can observe that the pixel value in the synthesized view fluctuates dramatically, which is so called temporal noise or flickering [28] in this paper.

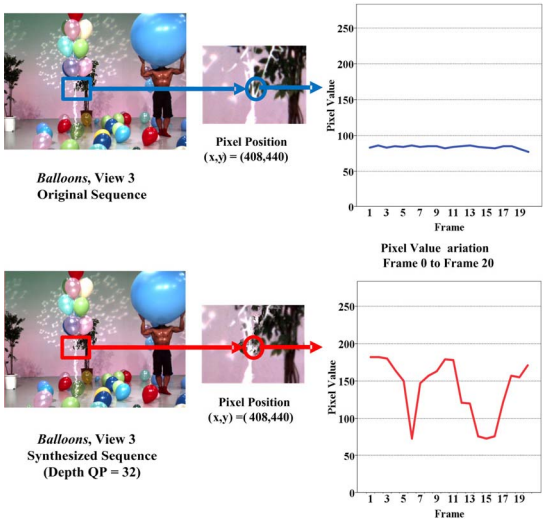


Fig. 1. Temporal inconsistency/distortion of the synthesized view video. [28].

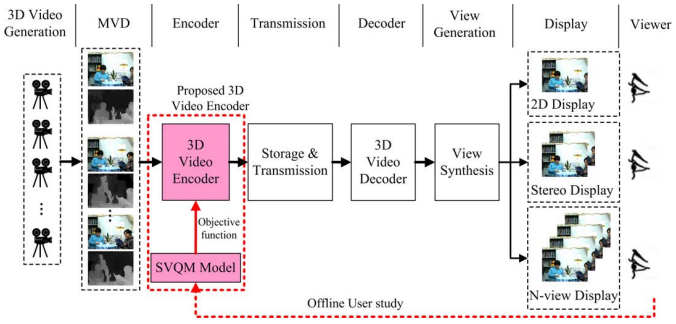


Fig. 2. Framework of the proposed perceptual 3D video coding system.

Compared with the spatial distortion, the dynamical intensity changing and video flickering are more noticeable and more annoying in subjective visual perception [28]. However, the conventional IQA metrics mainly utilize frame-by-frame calculation, such as PSNR and SSIM, and the obtained video quality is an average of the successive images of a sequence. The influence of temporal distortion on visual perception is not considered or underestimated. Meanwhile, those VQA metrics, such as VQM [29] and MOVIE [30], are focusing on the traditional monocular video distortion and the characteristics of synthesized view was not considered. Therefore, it is necessary to develop an efficient video quality metric of evaluating the synthesized view quality, and apply it in 3D video coding to improve the synthesized video quality with less bit rate cost.

III. Proposed 3D Video Coding System Framework

Fig.2 shows framework of the proposed perceptual 3D video coding system, where the red rectangles are a perceptual quality guided 3D video coding module consisting of SVQM model and an optimized 3D video encoder. In this system, MVD data consists of multiview texture videos and associated depth videos, which are jointly encoded by the perceptual 3D video encoder. 3D video bit stream is stored and/or transmitted to the client through network. At the client, arbitrary intermediate virtual views can be generated using the decoded color and depth videos with DIBR technique. The decoded real views or the synthesized ones are then displayed to viewers as demanded by a conventional 2D display, a stereo display and/or an N-view auto-stereoscopic display. It is observed that the multiview depth videos are only supplementary data for view synthesis instead of being watched directly. Moreover, off-line user study involving subjective and objective quality metrics is implemented at the client and feedback the perceptual quality information to the SVQM model, which then is used to optimize the 3D video encoder. The user study is off-line and it is not required any more once the 3D encoder has been developed.

The 3D video encoder consists of the multiview color video encoder and depth video encoder. In this work, the color video encoder is the original one and the depth video encoder is optimized based on the proposed SVQM model.

IV. The Proposed SVQM

As analyzed in Section II, it is found that both spatial distortion and temporal distortion shall be involved in evaluating the perceptual quality of the synthesized view. Especially, the temporal flickering of the synthesized video is different from the conventional 2D video distortion and shall be considered. In addition, since the synthesized video quality metric will be applied to the 3D video encoder as a distortion criteria during the coding parameter/mode decision process, it is necessary to maintain low computational complexity and easy integration into current video coding systems.

First of all, we keep the MSE based spatial distortion computing of synthesized view unchanged. Let $I_P(i,j,n)$ be the processed synthesized image pixel value at (i,j) in frame n generated by the original texture video and compressed depth video, and $I_R(i,j,n)$ be the reference image pixel value at (i,j) in frame n. The spatial distortion D_S of synthesized image $I_P(n)$ is calculated as

$$D_S(n) = \frac{1}{W \times H} \sum_{i=0}^{W-1} \sum_{j=0}^{H-1} [I_P(i,j,n) - I_R(i,j,n)]^2, \quad (1)$$

where W and H are the width and height of the processed images, respectively. Note that the reference image R is the rendered image generated by the original color video and the original depth video [9].

As for measuring the temporal distortion, we develop a new metric which computes the temporal gradient $\nabla I_\phi(i,j,n)$ between two consecutive frames and the temporal distortion $D_T(n)$ of the synthesized image $I_P(n)$ is calculated as

$$D_T(n) = \frac{1}{W \times H} \sum_{i=0}^{W-1} \sum_{j=0}^{H-1} [\nabla I_P(i,j,n) - \nabla I_R(i,j,n)]^2, \quad (2)$$

$$\nabla I_\varphi(i,j,n) = \frac{1}{K} \sum_{k=1}^{K} \left[I_\varphi(i,j,n) - I_\varphi(i,j,n-k) \right], \quad \varphi \in \{P,R\}, \quad (3)$$

where $\nabla I_P(i,j,n)$ and $\nabla I_R(i,j,n)$ represent the temporal gradient of $I_P(i,j,n)$ and $I_R(i,j,n)$, respectively, n is the n^{th} frame of the synthesized/reference video, $K \geq 1$ is the number of frames that are involved in the temporal gradient calculation.

In this paper, K is set as 1 for the sake of low complexity and simplicity. This temporal distortion measurement is motivated by PSPTNR [27]. However, it is worth mentioning that the JND model in [27] is not used in this paper. The reasons are twofold. On one hand, the adopted JND model is originally developed for measuring the perceptible distortion of conventional 2D video which is not suitable for synthesized video. On the other hand, the JND calculation significantly increases the complexity while it is applied to the RDO in the coding process. In addition, since the rendered image generated by the original texture video and the original depth video may also contain the temporal distortions, e.g. flickering, due to the depth temporal inconsistency, we use the captured real view image at the virtual viewpoint as the reference image in calculating the term $D_T(n)$.

Since the spatial and temporal distortions have joint impacts to HVS, we combine the spatial and temporal distortions of the synthesized view with weighted summation [29]. Then, the distortion metric for synthesized view (D_{SVQM}) is computed as

$$D_{SVQM}(n) = (1-\omega) \times D_S(n) + \omega \times D_T(n), \quad (4)$$

where ω is a weighting factor ranging from 0 to 1. Since different weighing factors may lead to different results, extensive experiments will be carried out to investigate the optimal weighting factor in Section VI.B.

Then, the D_{SVQM} value of the synthesized view is converted into a quality value Q_{SVQM} using a logarithmic function, which is similar to PSNR. Thus, the SVQM value of the n^{th} frame, $Q_{SVQM}(n)$, is defined as

$$Q_{SVQM}(n) = 10 \times \log_{10}\left(\frac{255^2}{D_{SVQM}(n)}\right). \quad (5)$$

The overall SVQM value for a synthesized video is the average of the remaining N-1 single frame values,

$$\overline{Q}_{SVQM} = \frac{1}{N-1} \times \sum_{n=2}^{N} Q_{SVQM}(n), \quad (6)$$

where N is the number of frames in a synthesized video sequence. Note that the temporal noise in the first frame is not calculated and ω is zero for the first frame. The performance of the proposed SVQM video metric is validated in Section VI.C.

V. SVQM-Based 3D Depth Coding Optimization

In the 3D-HEVC reference model [31], VSO algorithm is adopted to maximize synthesized image quality at a given bit rate. However, the distortion term in the VSO is still measured with the block based Sum of Squared Difference (SSD) or Sum of Absolute Difference (SAD) within a frame, which does not consider the temporal distortion in the synthesized video. In this section, we apply the SVQM metric to measure the synthesized distortions caused by depth compression, and optimize the depth encoder with the target of maximizing the perceptual quality of the synthesized view at a given bit rate as shown in the framework in Fig.2. We integrate the proposed SVQM into the RDO process of the 3D video encoder and then a novel Lagrange multiplier is derived. Additionally, some details regarding the implementations are also presented.

A. SVQM Based RDO and Lagrange Multiplier Adaptation

In the latest 3D-HEVC reference model, the synthesized image quality has been considered in the objective function of the depth video coding process. The RDO target of video coding is minimizing the Rate Distortion (RD) cost J, which is written as

$$\min\{J\}, \quad J = D + l_S \cdot \lambda_{mode} \cdot R_D, \quad (7)$$

where D is represented as $D = \eta_S D_S + \eta_d D_d$,$D_d$ is the depth distortion computed between the original and the reconstructed depth videos; D_S is the synthesized distortion defined in Eq.1, which can be calculated or estimated by the SVDC and VSD based VSO technique in 3D-HEVC [31]; η_S and η_d are weighting factors; l_S denotes a scaling factor depending on the quality of the corresponding color video, which is determined using a look-up table [31], and R_D is the depth bit rate. λ_{mode} indicates the HEVC Lagrange multiplier for mode decision, which is

$$\lambda_{mode} = \beta \cdot \kappa \cdot 2^{((QP-12)/3.0)}, \quad (8)$$

where κ is a weighting factor relying on coding configuration and Quantization Parameter (QP) offset, and β is a constant related to reference images.

Take the derivative of Eq. (7) with respect to R_D and set it to zero, we can obtain

$$\frac{\partial D}{\partial R_D} = \frac{\partial(\eta_S D_S + \eta_d D_d)}{\partial R_D} = -l_S \cdot \lambda_{mode}. \quad (9)$$

Through the DIBR process, the depth distortion leads to the rendering position displacement and VSD. To avoid the view synthesis being involved in RD cost calculation and lower the complexity, an estimation of the spatial distortion of view synthesis $D_S(i,j,n)$ is adopted in 3D-HEVC [9], which is

$$\begin{aligned} D_S(i,j,n) &= I_P(i,j,n) - I_R(i,j,n) \\ &= \frac{1}{2}\alpha \cdot D_d(i,j,n) \cdot G(i,j,n), \end{aligned} \quad (10)$$

where $D_d(i,j,n)$ indicates the distortion at pixel position (i,j) of the *n-th* encoding depth image, $G(i,j,n)$ denotes the gradient value calculated by

$$\begin{aligned} G(i,j,n) = &\left|\hat{I}_T(i,j,n) - \hat{I}_T(i-1,j,n)\right| \\ &+ \left|\hat{I}_T(i,j,n) - \hat{I}_T(i+1,j,n)\right|, \end{aligned} \quad (11)$$

where $\hat{I}_T(i,j,n)$ is the corresponding reconstructed texture pixel value at (i,j) of the current view and frame n[9], and α is the scene geometry information of view rendering calculated by

$$\alpha = \frac{f \cdot L}{255} \cdot \left(\frac{1}{Z_{near}} - \frac{1}{Z_{far}}\right), \quad (12)$$

where f is the focal length, L indicates the baseline between the reference and the rendered views. Z_{near} and Z_{far} represent the values of the nearest and farthest depth of a scene, respectively.

According to the Law of Large Number (LLN) [32], the average value of a large number of samples is approximated to the mathematical expectation value of these samples.

When the spatial distortion of the synthesized view is measured with MSE in Eq. (1), $D_S(n)$ is the mathematical expectation of squared $D_S(i,j,n)$, which is

$$D_S(n) \approx E\left(D_S(i,j,n)^2\right) = \frac{1}{4}\alpha^2 \phi_{G(n)} D_d(n), \tag{13}$$

where $E()$ is the mathematical expectation function, and $D_d(n)$ and φ_{Gn} indicate the mean squared distortion of depth image and the mean squared gradient value of current reconstructed color image, which are

$$\begin{cases} D_d(n) = E(D_d(i,j,n)) = \dfrac{1}{W \times H}\displaystyle\sum_{i=0}^{W-1}\sum_{j=0}^{H-1}\left[D_d(i,j,n)\right]^2 \\ \phi_{G(n)} = E(G(i,j,n)) = \dfrac{1}{W \times H}\displaystyle\sum_{i=0}^{W-1}\sum_{j=0}^{H-1}\left[G(i,j,n)\right]^2. \end{cases} \tag{14}$$

where $G(i,j,n)$ indicates the gradient value of current reconstructed texture image n at (i,j) defined in Eq.(11). Therefore, applying Eq. (13) to Eq. (9), we obtain

$$\frac{\partial D_d}{\partial R_D} = -\frac{l_S \cdot \lambda_{mode}}{\left(\frac{1}{4}\eta_S \alpha^2 \phi_{G(n)} + \eta_d\right)}. \tag{15}$$

Since the SVQM considers both spatial and temporal distortions and is supposed to be more consistent with human perception than the traditional D_S in Eq. (7) only while measuring the synthesized video quality, the SVQM, instead of using D_S, shall be integrated into the RDO to improve the 3D depth coding in perceptual aspect. Accordingly, an improved objective function for mode decision in depth coding is employed by replacing the D_S in Eq. (9) with D_{SVQM}, which is

$$\min\left\{J_{SVQM}\right\}, \quad J_{SVQM} = \eta_d D_d + \eta_S D_{SVQM} + \lambda_{SVQM} \cdot R_D, \tag{16}$$

where λ_{SVQM} represents a new Lagrange multiplier while adopting the SVQM. To calculate the SVQM based Lagrange multiplier, we take the derivative of J_{SVQM} with respect to R_{D} and set it to zero. Thus, we obtain

$$\begin{aligned} \lambda_{SVQM} &= -\frac{\partial\left(\eta_d D_d + \eta_S D_{SVQM}\right)}{\partial R_D} \\ &= -\frac{\partial\left[\eta_d D_d + \eta_S (1-\omega) D_S + \eta_S \omega D_T\right]}{\partial R_D}. \end{aligned} \tag{17}$$

According to Eqs. (2), (3) and (10), the pixel-wise view synthesis temporal distortion estimation $D_T(i,j,n)$ is defined as

$$\begin{aligned} &D_T(i,j,n) \\ &= \left(I_P(i,j,n) - I_P(i,j,n-1)\right) - \left(I_R(i,j,n) - I_R(i,j,n-1)\right) \\ &= \frac{1}{2}\alpha\left[D_d(i,j,n)\, G(i,j,n) - D_d(i,j,n-1)\, G(i,j,n-1)\right], \end{aligned} \tag{18}$$

where $D_d(i,j,n)$ and $D_d(i,j,n-1)$ indicate the depth distortion of the current n^{th} depth image and the pervious $(n-1)^{\mathrm{th}}$ depth image at (i,j), respectively. $G(i,j,n-1)$ indicates the gradient value of pervious reconstructed color image at (i,j). Similarly, the temporal distortion $D_T(n)$ measured with MSE in Eq. (2) can be rewritten as

$$\begin{aligned} D_T(n) &\approx E\left(D_T(i,j,n)^2\right) \\ &= \frac{1}{4}\cdot\alpha^2\cdot\begin{bmatrix} E\left((D_d(i,j,n))^2\right) \\ \times E\left((G(i,j,n))^2\right) \\ +E\left((D_d(i,j,n-1))^2\right) E\left((G(i,j,n-1))^2\right) \\ -2E\left(D_d(i,j,n)\, D_d(i,j,n-1)\right) \\ \times E\left(G(i,j,n)\, G(i,j,n-1)\right) \end{bmatrix}, \end{aligned} \tag{19}$$

where the gradient operations $G()$ of color image and depth distortion D_d can be recognized as two independent variables in the depth coding process. As presented in [33], the distortion $D_d(i,j,n)$ can be modeled as zero mean Generalized Gaussian Density (GGD) function, thus, the mean and variance of the depth image n and $n-1$ are the same, i.e. $\sigma^2_{D_d(i,j,n)} = \sigma^2_{D_d(i,j,n-1)}$ and $E\left(D_d(i,j,n)\right) = E\left(D_d(i,j,n-1)\right) = 0$. Therefore, the correlation coefficient ρ between $D_d(i,j,n)$ and $D_d(i,j,n-1)$ can be expressed as

$$\begin{aligned} \rho &= \frac{cov\left(D_d(i,j,n), D_d(i,j,n-1)\right)}{\sqrt{\sigma^2_{D_d(i,j,n)}}\cdot\sqrt{\sigma^2_{D_d(i,j,n-1)}}} \\ &= \frac{E\left(D_d(i,j,n)\, D_d(i,j,n-1)\right)}{E\left((D_d(i,j,n))^2\right)}. \end{aligned} \tag{20}$$

Because of the zero mean condition, we can remove the mean without changing the results of MSE in Eq. (20). Applying Eq. (20) into Eq. (19), the D_T is

$$D_T(n) = \frac{1}{4}\alpha^2 D_d(n)\left[\phi_{G(n)} - 2\rho\mu_{G(n)\cdot G(n-1)} + \phi_{G(n-1)}\right], \tag{21}$$

where $\mu_{G(n)\cdot G(n-1)}$ is the mean of $G(n)\cdot G(n-1)$ calculated by

$$\mu_{G(n)\cdot G(n-1)} = \frac{1}{W \times H}\sum_{i=0}^{W-1}\sum_{j=0}^{H-1}\left[G(i,j,n)\, G(i,j,n-1)\right]. \tag{22}$$

Applying Eqs. (13), (15), (21) into (17) and solving (17), we can obtain an improved Lagrange multiplier λ_{SVQM} for the mode decisions as

$$\lambda_{SVQM} = \left[1 - \omega + \omega \times \frac{\eta_d + \frac{1}{4}\eta_S\alpha^2\left[\phi_{G(n)} - 2\rho\mu_{G(n)\cdot G(n-1)} + \phi_{G(n-1)}\right]}{\frac{1}{4}\eta_S\alpha^2\phi_{G(n)} + \eta_d}\right] l_S \lambda_{mode}. \tag{23}$$

In 3D-HEVC reference software, the weight $\eta_S{:}\eta_d$ is set as 10:1 when the VSO is enabled, and they are squared, i.e. 100:1 [34], while the distortion is measured MSE or SSD in the mode decision. Therefore, η_d is much smaller than the items $\frac{1}{4}\eta_S\alpha^2\phi_{G(n)}$ and $\frac{1}{4}\eta_S\alpha^2[\phi_{G(n)} - 2\cdot\rho\cdot\mu_{G(n)\cdot G(n-1)} + \phi_{G(n-1)}]$, thus, can be omitted

for simplicity. Then, the Lagrange multiplier λ_{SVQM} can be approximated as

$$\lambda_{SVQM} \approx \left[1-\omega+\omega \times \frac{\left[\phi_{G(n)}-2\rho\mu_{G(n)\cdot G(n-1)}+\phi_{G(n-1)}\right]}{\phi_{G(n)}}\right] l_S\lambda_{mode}. \tag{24}$$

It means that we shall add the weight in the bracket in Eq. (24) to the traditional HEVC Lagrange multiplier $l_S\lambda_{\text{mode}}$.

B. Technical Implementations and Settings

1) Applicability of the SVQM Based RDO: The proposed method is implemented on the 3D-HEVC Test Model (3D-HTM) reference software. The VSO is enabled while its distortion metric is replaced by the SVQM method. The proposed SVQM based RDO is used in Coding Unit (CU) partitioning, merging, prediction unit decision, reference frame selection, intra mode pre-determination and residual quad-tree partitioning, while the conventional SSD/SAD based distortion metric is used in Motion Estimation (ME) and Rate-Distortion Optimized Quantization (RDOQ) for low complexity.

2) Reference Video Selection: Either the original captured video or synthesized video can be used as reference video to compute the SVQM. However, since some depth videos generated by frame-by-frame stereo matching based estimation methods are more or less imperfect, flickering areas still prevalent on synthesized videos rendered from such depth maps. Thus, synthesized video generated by the original texture video and the original depth video may also have such annoying temporal flickering artifacts, the original video is thus more reliable in temporal distortion measurement. Therefore, in Eq. (2), we use the original captured video if provided as the reference video to calculate the temporal distortion at the encoder. In terms of the spatial distortion, the SVDC is used. Meanwhile, we use synthesized video rendered from color video and original depth video as the reference video for the spatial distortion computing in Eq. (1), which is identically the same as 3D-HTM. Since the SVQM based RDO is only used in the optimal mode/parameter selection process at the encoder, the original captured video is not required in the decoder side.

3) Calculation for Correlation Coefficient ρ: Before performing the RDO, λ_{SVQM} should be determined first, which means the correlation coefficient ρ in Eq. (24) shall be obtained first. However, the distortion of the current whole depth image $D_d(n)$ used to compute ρ is unavailable, which is a dilemma since $D_d(n)$ can be obtained only after the whole frame has been coded. To solve this dilemma, ρ of the current frame is estimated from that of the previous frame for Inter frames, while it is set as 1 for Intra frames while calculating λ_{SVQM}. This strategy is presented as

$$\hat{\rho}_n = \begin{cases} 1 & INTRA \\ \rho_{n-1} & INTER, \end{cases} \tag{25}$$

where $\hat{\rho}_n$ is an estimation of ρ for the current frame n and ρ_{n-1} denotes the ρ of the previous frame. Applying $\hat{\rho}_n$ into Eq. (24) and the λ_{SVQM} for the current frame can be calculated. Actually, for the first frame, the temporal distortion doesn't exist and the RDO is just the same as that of the original 3D-HTM. After encoding the current frame, the ρ is then updated with $\Delta D_d(n)$.

VI. Experimental Results and Analyses

The experiment validation is divided into two phases. First, the proposed SVQM metric is validated and compared against the state-of-the-art video quality assessment metrics. Second, the proposed SVQM based 3D video coding is implemented on 3D-HEVC, and its performance is evaluated and compared with the original HTM and the depth coding method in [13].

A. Video Database

In our previous work [28], we have released a video database for subjective quality evaluation of the synthesized videos. The distorted synthesized videos in the database can be divided into three categories, which are 1) compressed texture and uncompressed depth (C_TU_D), 2) uncompressed texture and compressed depth (U_TC_D), and 3) compressed texture and compressed depth (C_TC_D). In this paper, the synthesized videos of both U_TC_D and C_TC_D dataset were used to evaluate the performance of the proposed SVQM metric.

Eight 3D sequences (multiview color videos and their corresponding depth videos) including *Balloons*, *Kendo*, *Newspaper*, *PoznanHall2*, *UndoDancer, PoznanStreet, GT_Fly* and *Shark,* with various resolutions, motion features, camera properties and scenes, were employed. Example frames of these MVD sequences are shown in Fig.3. These sequences were also used as the standard sequences in the 3DV core experiments of JCT-3V [35]. The descriptions of the sequences are shown in Table I. For each sequence, 200 frames were encoded at five different QPs and used to generate synthesized videos using View Synthesis Reference Software (VSRS) 1D fast [36]. For each intermediate virtual view, 40 test stimuli were generated. The five different QPs were chosen non-uniformly such that the resulting distribution of subjective quality scores were approximately uniform over the entire range. Note that 0 indicates the original depth video without compression, while other QPs denote the depth or color videos were compressed with such QP values. According to ITU-R Rec. BT.500 [37], Mean Opinion Score (MOS) was obtained and then normalized and scaled into the range of [0, 1]. More details of the database can be referred to [28].

B. Weighting Factor Determination for the Proposed SVQM

To determine the weighing factor ω in Eq. (4), we tested the performance of the proposed SVQM with different ω value ranging from 0 to 1 with 0.05 increasing step. Before evaluating the performance, we use a logistic function to convert the computed quality score to the predicted MOS (MOS_p), which is recommended by Video Quality Expert Group (VQEG) [38]. The MOS_p is calculated by

$$MOS_p(m) = \frac{\theta_1}{1+e^{-\theta_2\cdot(m-\theta_3)}}, \tag{26}$$

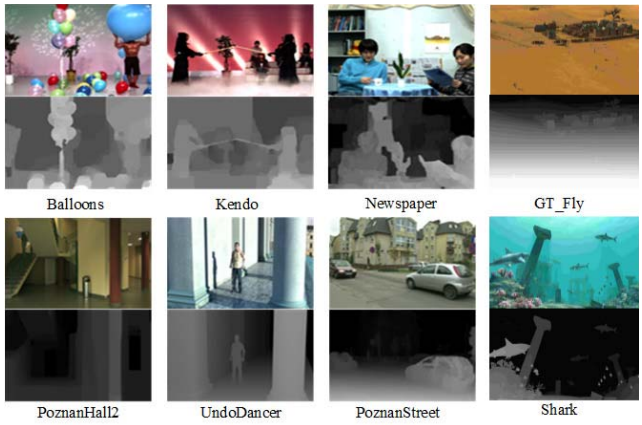


Fig. 3. Test multiview color videos and the corresponding depth videos.

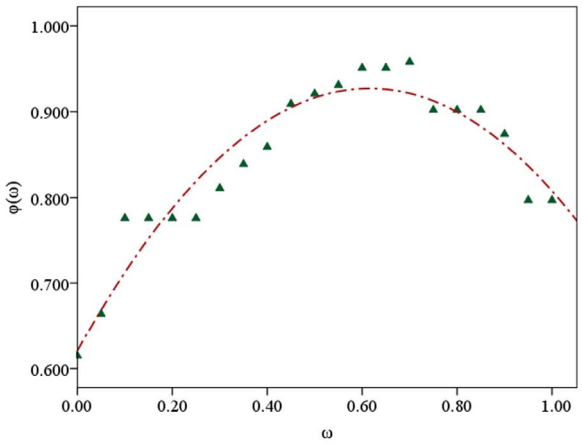


Fig. 4. Curve fitting for different ω values in terms of SROCC.

TABLE I
PROPERTIES OF THE TEST 3D VIDEO SEQUENCES

3DV Sequences	Resolution	Frame Rate /Spacing (fps/cm)	Input Views (Rendered View)	Compressed QP*
Balloons	1024x768	30/5	1-5 (3)	0, 32, 36, 38, 42
Kendo	1024x768	30/5	1-5 (3)	0, 32, 38, 44, 48
Newspaper	1024x768	30/5	2-4 (3)	0, 28, 36, 44, 50
PoznanHall2	1920x1088	25/13.75	5-7 (6)	0, 28, 32, 40, 46
PoznanStreet	1920x1088	25/13.75	3-5 (4)	0, 32, 38, 44, 48
UndoDancer	1920x1088	25/13.75	1-9 (5)	0, 24, 28, 40, 45
GT_Fly	1920x1088	25/13.75	1-9 (5)	0, 28, 36, 44, 48
Shark	1920x1088	25/13.75	1-9 (5)	0, 28, 36, 40 44

**QP* 0 indicates the depth or color video is original and uncompressed.

TABLE II
PERFORMANCES OF SVQM WITH DIFFERENT ω VALUES IN TERMS OF SROCC, LCC AND RMSE

ω	SROCC	LCC	RMSE	ω	SROCC	LCC	RMSE
0.00	0.615	0.676	0.1212	0.55	0.931	0.922	0.0568
0.05	0.664	0.717	0.1142	0.60	0.951	**0.925**	**0.0560**
0.10	0.776	0.753	0.1071	0.65	0.951	0.921	0.0602
0.15	0.776	0.786	0.0999	0.70	**0.958**	0.905	0.0651
0.20	0.776	0.816	0.0927	0.75	0.902	0.89	0.0722
0.25	0.776	0.842	0.0854	0.80	0.902	0.872	0.0792
0.30	0.811	0.865	0.0783	0.85	0.902	0.850	0.0868
0.35	0.839	0.884	0.0717	0.90	0.874	0.824	0.0944
0.40	0.859	0.900	0.0658	0.95	0.797	0.796	0.1016
0.45	0.909	0.912	0.0610	1.00	0.797	0.764	0.1086
0.50	0.921	0.919	0.0579				

where m is the score computed with the proposed metric, θ_1, θ_2 and θ_3 are parameters of the logistic function, respectively.

Three performance indicators, including Linear Correlation Coefficient (LCC), Spearman Rank Order Correlation Coefficient (SROCC) and Root Mean Square Error (RMSE) [38], were utilized to evaluate the consistency between the subjective and objective video quality scores. SROCC and LCC are the monotonicity and prediction accuracy between the objective and subjective scores. Larger value SROCC or LCC means better performance. Smaller RMSE values reflect higher accuracy and better consistency with the subjective scores.

Table II shows the performances of SVQM with different ω values in terms of the SROCC, LCC and RMSE values. To make sure that the proposed SVQM is generally suitable for most sequences and without losing the generality, three test sequences in the database, including *PoznanHall2, Kendo* and *PoznanStreet*, were used as a training set to find out the optimal weighing factor ω. We can observe that the SROCC, LCC and RMSE values have their peak values when weighing factor ranges from 0.60 to 0.70. For higher fidelity and simplicity, the optimal parameter ω was further refined and decided according to SROCC. As shown in Fig. 4, the triangle dots are the SROCC values with different ω and the red curve is the fitting curve. We fitted the SROCC with different ω by using a quadratic function as

$$\varphi(\omega) = a \times \omega^2 + b \times \omega + c, \tag{27}$$

where $a = -0.808$, $b = 0.995$ and $c = 0.621$ and the fitting accuracy (R^2) is 0.915. To find the optimal ω, we take the derivative of $\phi(\omega)$ with respect to ω and set it to zero. Then, we solve it and obtain the optimal ω as 0.616. This value is finally adopted in the SVQM and also applied to the SVQM based coding optimization. In addition, ω>0.5 also indicates temporal term has severer impacts than the spatial term in Eq. (4).

C. Performance Evaluation of the Proposed SVQM

We tested the proposed SVQM using both U_TC_D and C_TC_D databases [28]. Moreover, we compared the performance of the proposed SVQM with the state-of-the-art objective image and video quality metrics, including PSNR, SSIM [23], VQM [29], MOVIE [30] and PSPTNR [27], as well as the VQA for the synthesized video in [28] (denoted by 'VQA_SIAT').

Table III shows the comparison of the performance of video quality metrics in terms of LCC, SROCC and RMSE. The subscript 1 means the original video is used as reference to compute the quality, and the subscript 2 means the synthesized video generated by the original color and depth videos is used as reference. For the VQA_SIAT method, only the original captured color videos are used as reference [28]. Note that, for the SVQM method, the original color video is used for computing temporal distortion and the synthesized color video is used for computing spatial distortion. For comparison, the

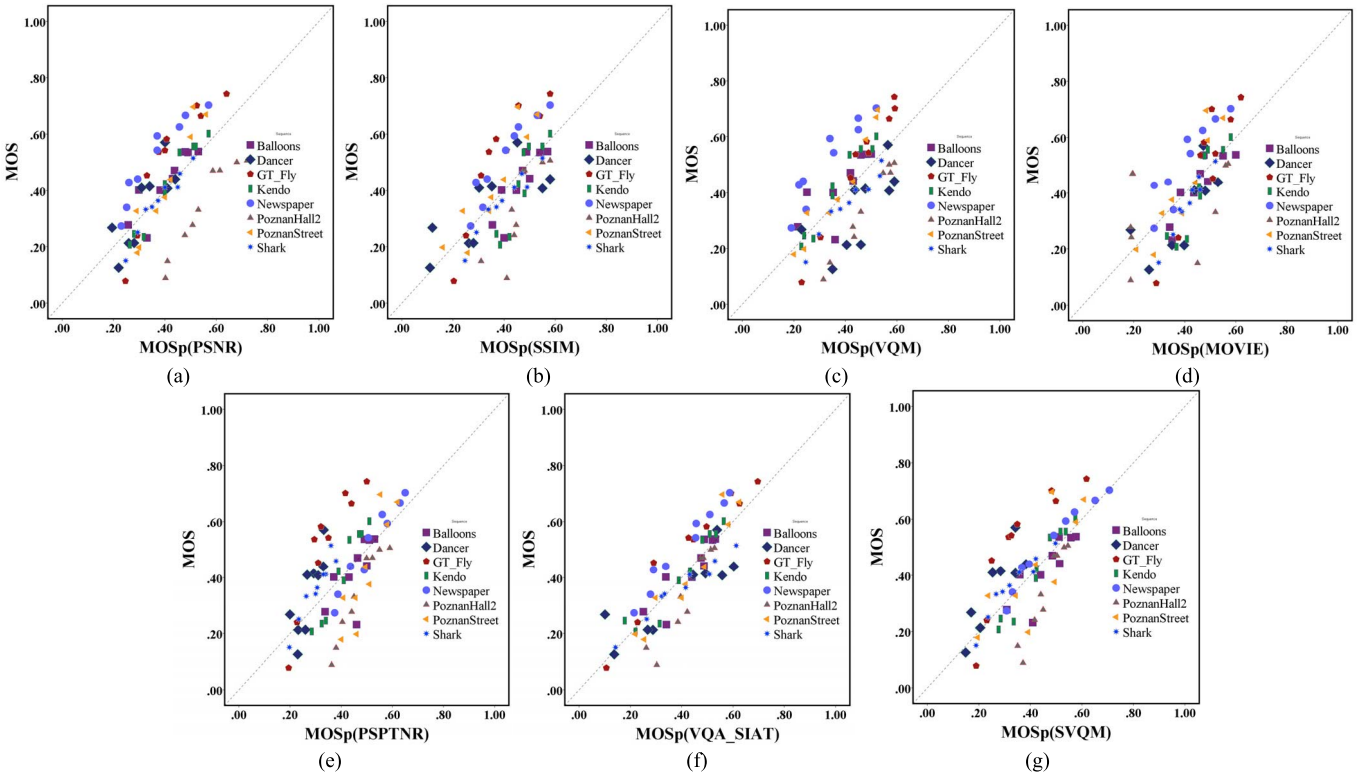


Fig. 5. Scatter plots comparison of different VQA metrics on the synthesized video database. Each sample point denotes one test stimuli. The red dashed line represents the linear fit between MOS_p and MOS. (a) PSNR_2, (b) SSIM_2, (c) VQM_2, (d) MOVIE_2, (e) PSPTNR_2, (f) VQA_SIAT, (g) SVQM.

TABLE III

COMPARISON OF THE PERFORMANCE OF QUALITY METRICS IN TERMS OF LCC, SROCC AND RMSE

Metrics	LCC	SROCC	RMSE
MOS_p(PSNR_1)	0.599	0.583	0.1306
MOS_p(SSIM_1) [23]	0.637	0.616	0.1255
MOS_p(VQM_1) [29]	0.570	0.566	0.1339
MOS_p(MOVIE_1) [30]	0.666	0.676	0.1213
MOS_p(PSPTNR_1) [27]	0.598	0.560	0.1308
MOS_p(PSNR_2)	0.679	0.710	0.1198
MOS_p(SSIM_2) [23]	0.705	0.708	0.1154
MOS_p(VQM_2) [29]	0.691	0.707	0.1177
MOS_p(MOVIE_2) [30]	0.715	0.749	0.1137
MOS_p(PSPTNR_2) [27]	0.647	0.626	0.1243
MOS_p(VQA_SIAT) [28]	**0.866**	**0.866**	**0.0814**
MOS_p(SVQM_synth)	0.735	0.721	0.1107
MOS_p(SVQM)	**0.763**	**0.748**	**0.1061**

Note: The subscript 1 means the original video is used as reference to compute the distortion. The subscript 2 means the synthesized video generated by the original texture and depth videos is used as reference.

performance of the SVQM_synth method which only uses the synthesized video as reference is also given.

From Table III, we have several following observations, 1) The SVQM has better performance than SVQM_synth because the synthesized video which may contain temporal noise is not suitable to be used as reference for computing temporal distortion. 2) The metrics with subscript 2 have better performance and are more consistent with human perception than those with subscript 1, which indicates it is better to use the synthesized videos as the reference for traditional 2D image/video quality metrics. 3) LCC, SROCC and RMSE of MOVIE_2 are 0.715, 0.749 and 0.1137, respectively, which is the best one among traditional 2D metrics. However, the 2D video/image metrics are not good enough since they have not taken the characteristics of the synthesized video into consideration. 4) The PSPTNR metric computes the temporal distortion which is larger than a JND threshold for the synthesized video, and its performances are 0.647, 0.626 and 0.1243 respectively for the LCC, SROCC and RMSE. These results are worse than the traditional 2D video metrics (e.g. PSNR_2, SSIM_2 and MOVIE_2) since it underestimates the spatial distortion of the synthesized view for some sequences. 5) For the VQA_SIAT, the LCC, SROCC, and RMSE scores are 0.866, 0.866 and 0.0814, respectively, which is the best among all the comparison algorithms. 6) The LCC, SROCC and RMSE of the SVQM are 0.763, 0.748 and 0.1061, respectively. The SROCC of the SVQM and MOVIE are almost the same, but the LCC and RMSE of the proposed SVQM are better than those of MOVIE. Overall, the SVQM is in the second place among the benchmarks and is better than the conventional 2D image/video metrics and the PSPTNR in measuring the quality of synthesized videos.

Fig.5 illustrates the scatter plot comparisons of different metrics measured on the video database. For the PSNR, SSIM, VQM, MOVIE and PSPTNR metrics, only the results using the synthesized video as reference are shown, since they are better than those using the original video as reference.

TABLE IV
BDBR AND BDSVQM COMPARISON BETWEEN THE ORIGINAL HTM, KIM_TIP [13] AND THE PROPOSED SVQM_ENC

Sequences	Texture QP and Depth QP	HTM			Kim_TIP [13]			Proposed SVQM_ENC			Total bit&SVQM		Depth bit&SVQM	
		Depth Bitrate (kbps)	Total Bitrate (kbps)	SVQM (dB)	Depth Bitrate (kbps)	Total Bitrate (kbps)	SVQM (dB)	Depth Bitrate (kbps)	Total Bitrate (kbps)	SVQM (dB)	BDBR /BDSVQM v.s. HTM	BDBR /BDSVQM v.s. Kim_TIP	BDBR /BDSVQM v.s. HTM	BDBR /BDSVQM v.s. Kim_TIP
Balloons	25-34	471.22	2105.61	39.50	484.00	2120.37	39.42	287.5	1924.02	39.43	2.8% /-0.03dB	-4.0% /0.09dB	-17.80% /0.33 dB	-27.10% /0.5 dB
	30-39	182.71	1043.75	38.75	187.20	1047.30	38.65	127.35	988.08	38.64				
	35-42	81.79	559.90	37.57	85.62	564.40	37.41	63.35	541.99	37.40				
	40-45	37.17	317.42	35.88	39.16	318.88	35.81	31.77	311.32	35.72				
BookArrival	25-34	349.64	2299.69	39.13	362.24	2312.65	39.13	255.63	2207.39	39.40	-4.9% /0.20dB	-5.7% /0.22dB	-20.50% /0.57 dB	-24.40% /0.66 dB
	30-39	125.46	1053.23	37.87	129.44	1057.54	37.86	97.79	1025.99	37.99				
	35-42	55.07	535.46	36.09	58.04	537.04	36.07	46.59	527.21	36.10				
	40-45	26.12	286.10	33.98	27.68	287.56	33.96	24.06	284.68	33.93				
Kendo	25-34	520.37	2177.65	38.75	532.57	2190.85	38.72	361.18	2019.18	39.74	-16.3% /0.70dB	-18.1% /0.75dB	-33.00% /0.99 dB	-36.60% /1.06 dB
	30-39	196.44	1052.68	38.05	200.62	1056.50	38.02	148.98	1003.92	38.65				
	35-42	85.38	557.82	36.79	88.26	560.56	36.74	69.38	541.79	37.02				
	40-45	36.32	314.11	34.97	37.82	315.79	34.90	32.12	310.14	35.03				
UndoDancer	25-34	485.88	9476.47	35.13	498.01	9489.15	35.11	407.20	9401.74	35.68	-12.2% /0.39dB	-13.4% /0.41dB	-21.00% /0.69 dB	-25.40% /0.85 dB
	30-39	223.55	3620.26	33.13	234.72	3635.79	33.11	201.30	3598.57	33.44				
	35-42	116.13	1480.54	31.22	122.63	1490.48	31.20	96.85	1465.26	31.39				
	40-45	60.17	665.93	29.77	64.83	670.95	29.74	53.08	659.20	29.84				
GTFly	25-34	495.20	4490.38	38.27	480.30	4480.09	38.25	363.69	4368.92	38.49	-6.3% /0.18dB	-7.6% /0.21dB	-19.80% /0.55 dB	-22.60% /0.59 dB
	30-39	162.60	1694.62	36.67	163.65	1695.58	36.65	130.61	1663.56	36.79				
	35-42	65.64	705.63	34.91	68.70	710.07	34.89	58.62	700.50	34.98				
	40-45	30.51	322.60	32.96	32.55	325.08	32.94	28.10	320.83	33.02				
PoznanCarPark	25-34	2018.58	7810.39	35.47	2034.13	7821.70	35.42	860.04	6654.11	36.07	-19.4% /0.45dB	-22.2% /0.49dB	-46.50% /0.7 dB	-49.70% /0.75 dB
	30-39	573.27	2843.23	35.22	585.82	2854.56	35.18	312.27	2580.89	35.56				
	35-42	210.65	1276.40	34.55	214.92	1280.37	34.50	133.34	1199.61	34.69				
	40-45	75.90	583.22	33.42	79.11	586.68	33.39	56.63	564.34	33.38				
PoznanHall2	25-34	275.29	1843.19	35.36	287.06	1853.54	35.35	257.02	1828.96	37.35	-87.4% /1.77dB	-87.6% /1.77dB	-88.20% /1.77 dB	-88.90% /1.79 dB
	30-39	123.28	756.50	35.23	126.97	759.88	35.23	111.01	743.29	36.96				
	35-42	62.35	379.86	34.94	64.22	381.46	34.94	57.56	376.33	36.37				
	40-45	29.81	207.51	34.34	30.87	208.71	34.34	29.11	207.20	35.33				
PoznanStreet	25-34	856.82	6266.34	38.35	817.71	6226.83	38.27	620.17	6033.96	38.54	-4.2% 0.12dB	-6.4% /0.17dB	-18.50% /0.31 dB	-15.20% /0.23 dB
	30-39	261.70	2139.48	37.60	178.80	2005.36	37.53	204.88	2084.18	37.66				
	35-42	97.14	930.03	36.48	99.41	932.04	36.44	83.14	915.13	36.46				
	40-45	38.10	448.44	35.04	40.42	450.02	35.00	34.91	445.28	34.94				
Shark	25-34	1587.78	9552.32	38.20	1581.38	9557.33	38.15	1085.51	9056.36	38.44	-4.4% /0.21dB	-5.5% /0.25dB	-24.00% /0.92 dB	-25.80% /0.97 dB
	30-39	481.00	4218.53	35.53	481.33	4230.27	35.50	357.86	4104.57	35.61				
	35-42	174.49	1995.68	32.82	178.80	2005.36	32.80	143.39	1967.67	32.87				
	40-45	62.4	955.84	30.27	65.35	961.29	30.25	57.98	952.07	30.31				
Newspaper	25-34	455.56	2131.89	40.31	468.50	2141.95	40.27	306.8	1981.20	40.67	-0.39% /0.20dB	-5.8% /0.25dB	-21.70% /0.57 dB	-25.50% /0.66 dB
	30-39	166.01	990.11	39.32	170.34	993.94	39.27	123.06	948.17	39.38				
	35-42	71.73	508.76	37.82	74.05	510.94	37.79	57.75	494.61	37.75				
	40-45	30.50	281.90	35.90	31.60	282.69	35.88	27.26	278.85	35.75				
Average											**-15.27% /0.42dB**	**-17.63% /0.46dB**	**-31.10% /0.74 dB**	**-34.12% /0.81 dB**

Vertical and horizontal axes are subjective value (*MOS*) and predicted objective value (MOS_p), respectively. We can observe that for the proposed SVQM metric the scatter points surround the dash line closely and the objective MOS_p values are more consistent with the subjective MOS values. The performance of SVQM is comparable to the VQA_SIAT and better than rest of benchmarks in terms of the consistency between the subjective and objective scores. Furthermore, it is found that the performances of traditional 2D metrics vary with the content of the sequences. For example, PSNR underestimates the subjective MOS values of the sequence *Newspaper* and overestimates that of *PoznanHall2*. VQM overestimates the subjective MOS values of *UndoDancer* while PSPTNR underestimates the synthesized distortion. The VQA_SIAT and SVQM have better consistency with HVS since both spatial distortions and temporal flickering distortions are considered.

Moreover, we performed an experiment to compare the computational complexity among these video quality metrics. The experimental environment and setup are set as follows:

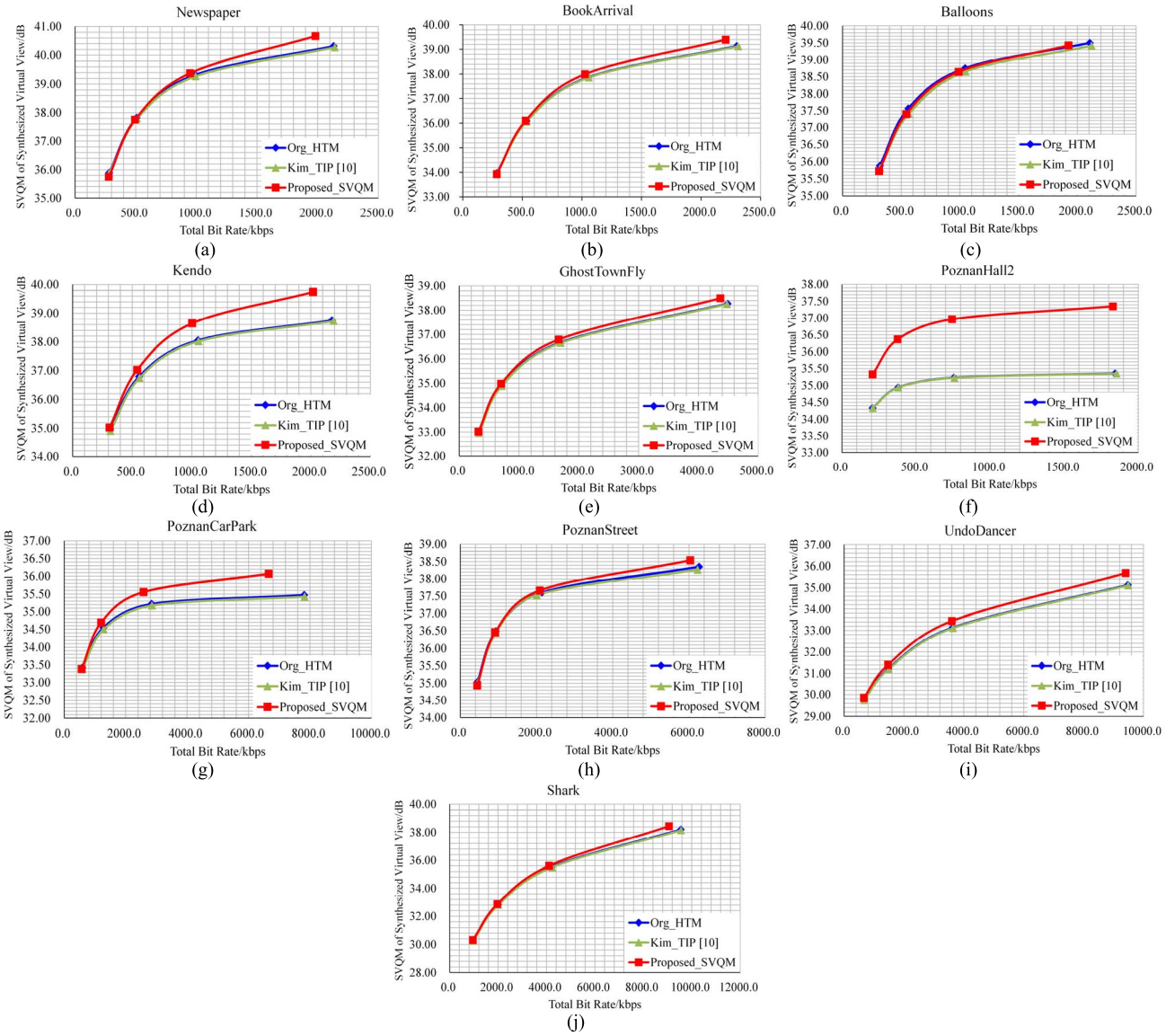


Fig. 6. RD curves of Org_HTM, Kim_TIP [13] and Proposed SVQM_ENC. (a) *Newspaper*, (b) *BookArriaval*, (c) *Balloons*, (d) *Kendo*, (e) *GhostTownFly*, (f) *PoznanHall2*, (g) *PoznanCarPark*, (h) *PoznanStreet*, (i) *UndoDancer*. (j) *Shark*.

The CPU is Intel Core i7 4790 with 3.6 GHz frequency. The RAM is 16 GB and the OS is Windows 8.1 64 bit. Eight sequences were used. Each sequence has five test videos with different degrees of distortion and the length of each video is 200 frames. To compare the computational complexity, each quality metric was used to compute the quality of the five distorted videos of each test sequence and the average time was recorded and calculated. We obtained that the average computation time of a test sequence using PSNR, SSIM, VQM, MOVIE, PSPTNR, VQM_SIAT and SVQM are 9.0s, 21.4s, 340.2s, 12045.0s, 58.3s, 114.6s, and 27.5s, respectively. The complexity of SVQM is three times to PSNR and 1.29 times to SSIM. Compared with the VQM, MOVIE, PSPTNR and VQA_SIAT metrics, the SVQM has much lower complexity. Overall, the SVQM has a good consistency with human HVS and meanwhile maintains low complexity as measuring the quality of the synthesized videos. Also, it is easier to be applied to the 3D video encoder.

D. Coding Performance Evaluation of the Proposed SVQM Based 3D Encoder

We implemented the proposed SVQM based depth video encoder on the 3D-HEVC reference software, HTM-11.0 [39] and made comparisons on the coding performances with the original 3D-HTM and another state-of-the-art depth coding method in [13]. The encoder configuration settings follow the Common Test Conditions (CTC) [35] recommended by the JCT-3V. The inter-view IPPP coding structure of two-view 3DV data was chosen and one intermediate view was rendered by VSRS 1D fast mode. Group-of-Picture (GOP) length was set as 4. Intra frame occurred every 16 frames. RDOQ and VSO were enabled while the VSD and early skip mode of VSO were disabled. One interpolated view was used in VSO configuration. The basis *QP*s for color video were set as 25, 30, 35, 40 and the corresponding depth QP values were 34, 39, 42 and 45. All the sequences in video database were available and encoded in the coding experiments.

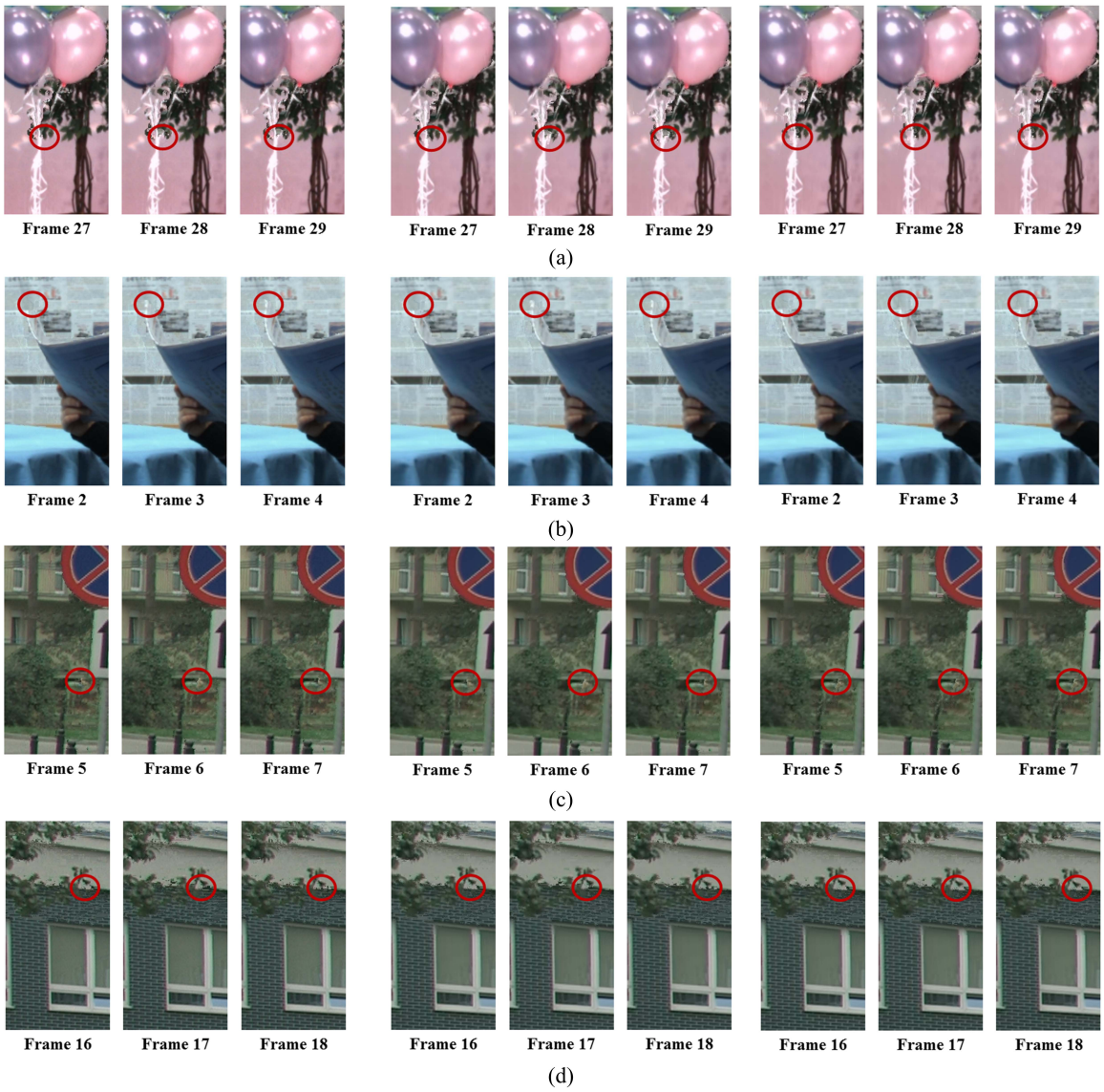


Fig. 7. Visual quality comparison of successive frames of synthesized videos. The left column are synthesized with the original texture and depth videos. The middle and the right columns are synthesized with the texture and depth videos compressed by HTM and the SVQM_ENC with a QP pair of (25, 34), respectively. The pixels in the red circle area have temporal noise in successive frames. (a) Balloons, (b) Newspaper, (c) PoznanStreet, (d) PoznanCarPark.

Furthermore, two additional sequences*BookArrival* and *PoznanCarPark* were added to make sure that the proposed method was still reliable and efficient for sequences out of the video database. We encoded these sequences with three schemes, the original HTM-11.0 (denoted by 'Org_HTM'), the depth coding method in [13] (denoted by 'Kim_TIP') and our proposed SVQM based 3D encoder (denoted by 'SVQM_ENC'), where the depth encoder was optimized and the color encoder was original. Average SVQM value of the synthesized video was used to measure the 3D video quality. Similar to Bjontegaard Delta Bit Rate (BDBR) and Bjontegaard Delta PSNR (BDPSNR) [40], Bjontegaard delta SVQM (BDSVQM) and BDBR were utilized to evaluate the coding performance.

Table IV shows the BDBR and BDSVQM comparisons between the SVQM_ENC, Org_HTM and Kim_TIP. In Table IV, total bits (depth plus color bits) and depth bits were counted in bit rate in BDBR and BDSVQM calculation, respectively. Negative BDBR and positive BDSVQM indicate bit rate reduction and quality gain when comparing the SVQM_ENC with Org_HTM and Kim_TIP scheme, respectively. As shown in Table IV, for sequences *Kendo, UndoDancer*, *PoznanHall2* and*PoznanCarPark*, the SVQM_ENC can achieve from 12.2% to 87.4% BDBR reduction, or from 0.39dB to 1.77dB visual quality gain in terms of SVQM, respectively, when compared with the Org_HTM. The BDBR reduction or BDSVQM gain is even greater as compared with the Kim_TIP. There are mainly two reasons: 1) as shown in Fig.5 (a) and Fig.5 (f), the temporal distortions of the sequences *Kendo* and *PoznanHall2* are seriously underestimated as measured by PSNR. Thus, it is a large improvement by using the perceptual SVQM_ENC; 2) the temporal distortion fluctuation areas of the sequences *Kendo*,*PoznanCarPark* and *PoznanHall2* are much larger than those of other sequences, which leads to larger coding gain by the proposed SVQM_ENC. For the rest sequences, the

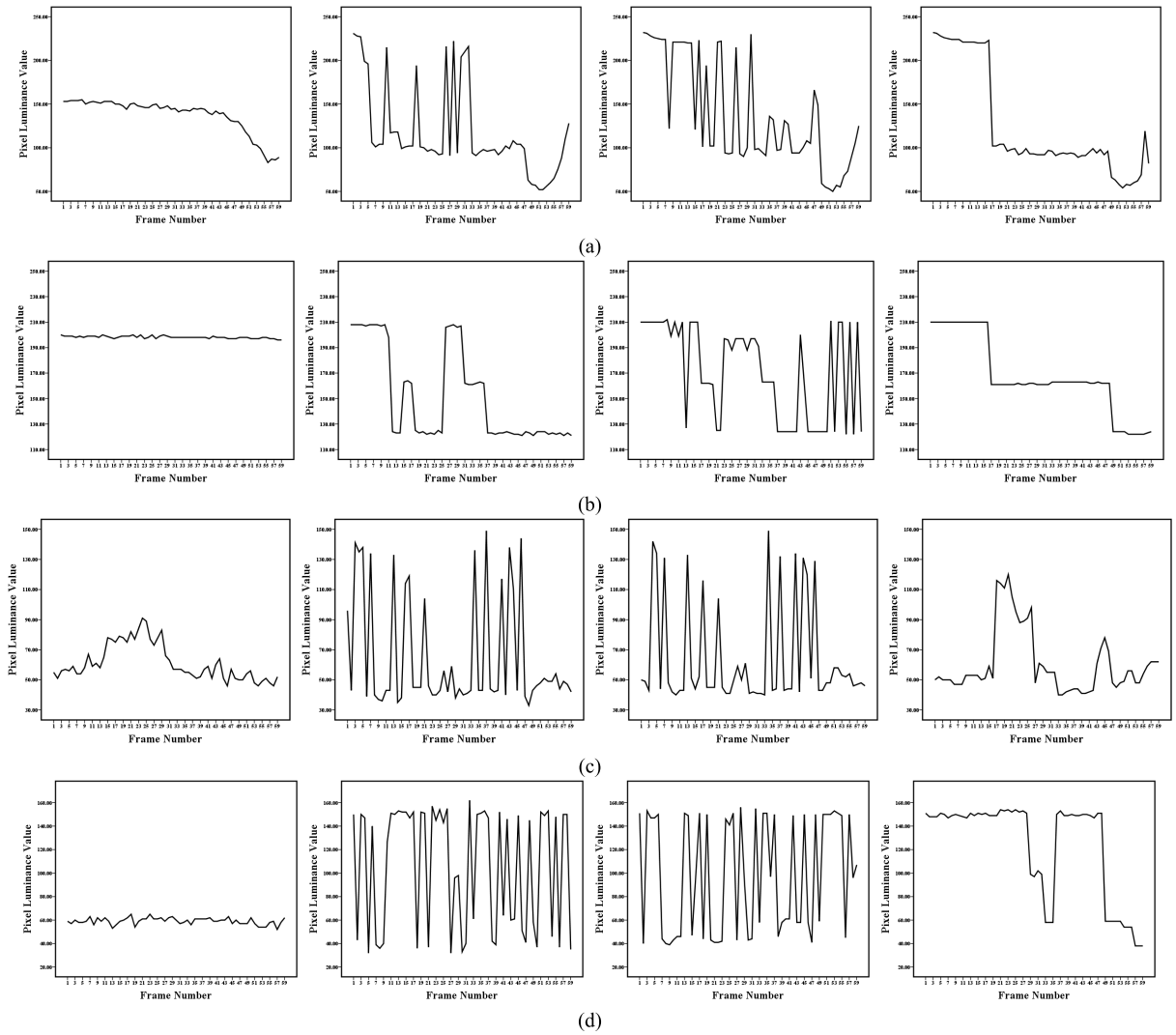


Fig. 8. The temporal distortion of synthesized videos. From left to right, the first column is the original video. The second column is the video synthesized with the original texture and depth videos. The third and the last columns are synthesized with the texture and depth videos compressed by HTM and the SVQM_ENC with a *QP* pair of (25, 34), respectively. (a) Pixel at position (414, 385) of *Balloons* sequence, (b) Pixel at position (609, 511) of *Newspaper* sequence, (c) Pixel at position (1860, 544) of *PoznanStreet* sequence, (d) Pixel at position (319, 126) of *PoznanCarPark* sequence.

proposed SVQM_ENC achieves less gain. This is mainly because the un-optimized color video bits account for much larger proportion compared to the depth video bits, which leads to smaller BDBR respect to the total bit rate. As shown in Table IV, if only depth bits were counted in bit rate and BDBR, the bit rate reduction or quality gain are more promising. Generally, when the total bits (depth and color bits) were counted, the SVQM_ENC can achieve 15.27% or 17.63% BDBR reduction and 0.42dB or 0.46dB visual quality gain in terms of BDSVQM on average as compared with the Org_HTM or Kim_TIP, respectively. When only the depth bits were counted, the SVQM_ENC can achieve 31.10% or 34.12% BDBR reduction and 0.74dB or 0.81dB BDSVQM gain on average as compared with the Org_HTM or Kim_TIP, respectively.

For better observation, Fig. 6 shows the RD curves of the Org_HTM, Kim_TIP and the SVQM_ENC. Vertical and horizontal axes are the SVQM values of the synthesized view and the total bit rate (depth plus color bits), respectively. From the figures, we can observe that the proposed SVQM_ENC reduces bit rate significantly for *Kendo*, *PoznanCarPark* and *PoznanHall2* sequences. For the sequences *Newspaper, BookArrival*, *GhostTownFly*, *PoznanStreet*, *Shark* and *Unodancer*, the proposed algorithm achieves better performance at high bit rate. While encoding the *Balloons* sequence, both the bit rate and the synthesized video quality are slightly reduced, and the RD performance is similar to that of original 3D-HTM. Overall, experimental results demonstrated that the proposed SVQM_ENC achieves much better RD performance for most sequences, especially at the high bit rate.

In addition, to evaluate the performance of the SVQM_ENC more completely, we also performed the experiments that uses VQA_SIAT [28], which is independent to the SVQM, as the coded video quality metric. The coding results and comparison are shown in Table V. According to the results, when the total bits (depth and color bits) were counted, the SVQM_ENC achieves 14.58% or 15.56% BDBR reduction and 0.48dB or 0.49dB visual quality gain in terms of Bjontegaard delta

TABLE V
BDBR AND BD-VQA_SIAT COMPARISON BETWEEN THE ORIGINAL HTM, KIM_TIP [13] AND THE PROPOSED SVQM_ENC

Sequences	Total bit & VQA_SIAT		Depth bit & VQA_SIAT	
	BDBR /BD-VQA_SIAT v.s. HTM	BDBR /BD-VQA_SIAT v.s. Kim_TIP	BDBR /BD-VQA_SIAT v.s. HTM	BDBR /BD-VQA_SIAT v.s. Kim_TIP
Balloons	-18.3%/0.45 dB	-22.6%/0.49 dB	-36.80%/0.65 dB	-43.10%/0.72 dB
BookArrival	-5.0%/0.34 dB	-5.0%/0.33 dB	-20.60%/0.87 dB	-23.80%/0.96 dB
Kendo	-17.6%/0.61 dB	-17.6%/0.61 dB	-34.10%/0.81 dB	-35.70%/0.83 dB
UndoDancer	-8.1%/0.52 dB	-9.0%/0.54 dB	-18.40%/0.97 dB	-22.80%/1.18 dB
GTFly	-2.4%/0.20 dB	-3.2%/0.23 dB	-16.40%/0.93 dB	-19.0%/0.98 dB
PoznanCarPark	-15.6%/0.23 dB	-18.3%/0.23 dB	-43.20%/0.42 dB	-46.70%/0.45 dB
PoznanHall2	-51.9%/1.37 dB	-52.6%/1.38 dB	-56.30%/1.39 dB	-58.20%/1.42 dB
PoznanStreet	-7.7%0.27 dB	-8.0%/0.27 dB	-20.90%/0.46 dB	-18.00%/0.35 dB
Shark	-2.9%/0.29 dB	-3.8%/0.33 dB	-22.40%/1.20 dB	-24.0%/1.25 dB
Newspaper	-16.3%/0.56 dB	-15.5%/0.53 dB	-31.80%/0.69 dB	-33.10%/0.68 dB
Average	**-14.58%/0.48 dB**	**-15.56%/0.49 dB**	**-30.09%/0. 84dB**	**-32.44%/0.88 dB**

VQA_SIAT (denoted as BD-VQA_SIAT) on average as compared with the Org_HTM or Kim_TIP, respectively. When only the depth bits were counted, the SVQM_ENC can achieve 30.09% or 32.44% BDBR reduction and 0.84dB or 0.88dB BD-VQA_SIAT gain on average compared with the Org_HTM or Kim_TIP, respectively. Overall RD comparisons using SVQM and VQA_SIAT metrics have proved the proposed SVQM_ENC achieves significant gains and it is effective in 3D depth video coding.

The visual quality of the synthesized videos were also evaluated, as shown in Fig.7 and Fig.8. The most important feature of the proposed SVQM metric is measuring the temporal distortion in the synthesized video, and while the SVQM is applied to the depth video coding, the temporal distortions of the synthesized videos from the depth videos compressed by the SVQM_ENC have been significantly reduced. As shown in Fig.7, the pixels in the red circle marked area have temporal distortions when the depth video were compressed with the Org_HTM. Moreover, even for the synthesized video generated from the original depth and color videos, temporal distortion still exists because of the inaccurately estimated depth data. Fortunately, the proposed SVQM_ENC can improve the depth quality and reduce the temporal distortion in the synthesized video. The variation of the synthesized pixel values shown in Fig. 8 can more clearly demonstrate the effect of visual improvement by the proposed SVQM_ENC. The pixel values of the original captured video are smooth along frames. However, the pixel values in the synthesized video from original and the Org_HTM compressed depth are suffered temporal flickering and fluctuate dramatically. As for the synthesized video from SVQM_ENC compressed depth, the curves of the pixel values are much smoother and closer to those of the original captured video.

In addition to the RD performance and visual quality, the coding complexity of the proposed encoder was also evaluated. Fig.9 shows the average encoding time of the Org_HTM and SVQM_ENC over four QPs. Compared with the Org_HTM, SVQM_ENC scheme increases the computational complexity from 7.07% to 38.00%, with 21.26% on average. It is because the new SVQM based distortion metric requires additional operations for temporal distortion calculation compared with the original MSE based distortion. Meanwhile, as being integrated in the depth coding processes, such as VSO, RDO and CU/PU mode decision *etc.*, the SVQM based distortion metric is high frequently called by RD cost calculation. The complexity varies over sequences because the complexity of the adopted fast or early termination algorithms in 3D-HTM varies over the sequences. Overall, the complexity increase of the proposed SVQM_ENC is acceptable.

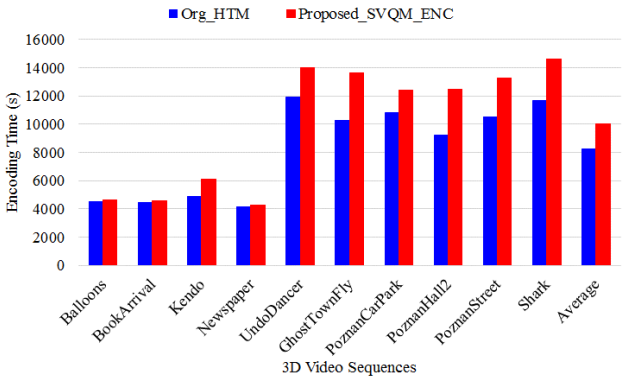


Fig. 9. Computational complexity of the SVQM_ENC.

VII. CONCLUSIONS

In this paper, we propose a full reference Synthesized Video Quality Metric (SVQM) metric to evaluate the video quality of synthesized view for 3D video system. Then, a perceptual 3D video coding framework is presented, in which SVQM-based rate-distortion (RD) optimization and view synthesis optimization methods are developed to improve the 3D depth video coding efficiency. Novel RD cost function and Lagrange multiplier have been investigated in the proposed 3D encoder. Experimental results show that the proposed SVQM metric is more consistent with the HVS when compared the conventional video quality metrics. Furthermore, it is demonstrated that the proposed SVQM based encoder is more efficient as compared with the original 3D-HTM and the state-of-the-art 3D depth coding method. Meanwhile, visual quality of the synthesized video from the proposed coding algorithm is also improved.

REFERENCES

[1] C. Bal and T. Q. Nguyen, "Multiview video plus depth coding with depth-based prediction mode," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 6, pp. 995–1005, Jun. 2014.

[2] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, "Overview of the high efficiency video coding (HEVC) standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1649–1668, Dec. 2012.

[3] Y. Zhang, S. Kwong, X. Wang, H. Yuan, Z. Pan, and L. Xu, "Machine learning based coding unit depth decisions for flexible complexity allocation in high efficiency video coding," *IEEE Trans. Image Process.*, vol. 24, no. 7, pp. 2225–2238, Jul. 2015.

[4] Z. Pan, J. Lei, Y. Zhang, X. Sun, and S. Kwong, "Fast motion estimation based on content property for low-complexity H.265/HEVC encoder," *IEEE Trans. Broadcast.*, vol. 62, no. 3, pp. 675–684, Sep. 2016.

[5] G. Tech, K. Wegner, Y. Chen, and S. Yea, *3D-HEVC Draft Text 6, Joint Collaborative Team on 3D Video Coding Extensions (JCT-3V)*, document JCT3V-J1001, 10th Meeting, Strasbourg, France, Oct. 2014.

[6] 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.

[7] H. Yuan, S. Kwong, C. Ge, X. Wang, and Y. Zhang, "Interview rate distortion analysis-based coarse to fine bit allocation algorithm for 3-D video coding," *IEEE Trans. Broadcast.*, vol. 60, no. 4, pp. 614–625, Dec. 2014.

[8] L. Fang, N.-M. Cheung, D. Tian, A. Vetro, H. Sun, and O. C. Au, "An analytical model for synthesis distortion estimation in 3D video," *IEEE Trans. Image Process.*, vol. 23, no. 1, pp. 185–199, Jan. 2014.

[9] B. T. Oh and K.-J. Oh, "View synthesis distortion estimation for AVC- and HEVC-compatible 3-D video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 6, pp. 1006–1015, Jun. 2014.

[10] Y. Zhang, S. Kwong, L. Xu, S. Hu, G. Jiang, and C.-C. J. Kuo, "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.

[11] Y. Zhang, S. Kwong, S. Hu, and C.-C. J. Kuo, "Efficient multiview depth coding optimization based on allowable depth distortion in view synthesis," *IEEE Trans. Image Process.*, vol. 23, no. 11, pp. 4879–4892, Nov. 2014.

[12] Y. Zhang, Z. Pan, Y. Zhou, and L. Zhu, "Allowable depth distortion based fast mode decision and reference frame selection for 3D depth coding," *Multimedia Tools Appl.*, pp. 1–20, Dec. 2015, doi: 10.1007/s11042-015-3109-0.

[13] W.-S. Kim, A. Ortega, P. L. Lai, and D. Tian, "Depth map coding optimization using rendered view distortion for 3D video coding," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 3534–3545, Nov. 2015.

[14] K. Müller *et al.*, "3D high-efficiency video coding for multi-view video and depth data," *IEEE Trans. Image Process.*, vol. 22, no. 9, pp. 3366–3378, Sep. 2013.

[15] Z. Pan, Y. Zhang, and S. Kwong, "Efficient motion and disparity estimation optimization for low complexity multiview video coding," *IEEE Trans. Broadcast.*, vol. 61, no. 2, pp. 166–176, Jun. 2015.

[16] S. Wang, A. Rehman, Z. Wang, S. Ma, and W. Gao, "Perceptual video coding based on SSIM-inspired divisive normalization," *IEEE Trans. Image Process.*, vol. 22, no. 4, pp. 1418–1429, Apr. 2013.

[17] T. Zhao, J. Wang, Z. Wang, and C. W. Chen, "SSIM-based coarse-grain scalable video coding," *IEEE Trans. Broadcast.*, vol. 61, no. 2, pp. 210–221, Jun. 2015.

[18] L. Xu, S. Li, K. N. Ngan, and L. Ma, "Consistent visual quality control in video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 6, pp. 975–989, Jun. 2013.

[19] Z. Luo, L. Song, S. Zheng, and N. Ling, "H.264/advanced video control perceptual optimization coding based on JND-directed coefficient suppression," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 6, pp. 935–948, Jun. 2013.

[20] C.-C. Su, L. K. Cormack, and A. C. Bovik, "Oriented correlation models of distorted natural images with application to natural Stereopair quality evaluation," *IEEE Trans. Image Process.*, vol. 24, no. 5, pp. 1685–1699, May 2015.

[21] F. Shao, K. Li, W. Lin, G. Jiang, M. Yu, and Q. Dai, "Full-reference quality assessment of stereoscopic images by learning binocular receptive field properties," *IEEE Trans. Image Process.*, vol. 24, no. 10, pp. 2971–2983, Oct. 2015.

[22] V. De Silva, H. K. Arachchi, E. Ekmekcioglu, and A. Kondoz, "Toward an impairment metric for stereoscopic video: A full-reference video quality metric to assess compressed stereoscopic video," *IEEE Trans. Image Process.*, vol. 22, no. 9, pp. 3392–3404, Sep. 2013.

[23] 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.

[24] E. Bosc *et al.*, "Towards a new quality metric for 3-D synthesized view assessment," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 7, pp. 1332–1343, Nov. 2011.

[25] C. T. E. R. Hewage and M. G. Martini, "Edge-based reduced-reference quality metric for 3-D video compression and transmission," *IEEE J. Sel. Topics Signal Process.*, vol. 6, no. 5, pp. 471–482, Sep. 2012.

[26] S.-W. Jung, "A modified model of the just noticeable depth difference and its application to depth sensation enhancement," *IEEE Trans. Image Process.*, vol. 22, no. 10, pp. 3892–3903, Oct. 2013.

[27] Y. Zhao and L. Yu, "A perceptual metric for evaluating quality of synthesized sequences in 3DV system," *Proc. SPIE*, vol. 7744, pp. 1–9, Jul. 2010.

[28] X. Liu, Y. Zhang, S. Hu, S. Kwong, C.-C. J. Kuo, and Q. Peng, "Subjective and objective video quality assessment of 3D synthesized views with texture/depth compression distortion," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 4847–4861, Dec. 2015.

[29] M. H. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," *IEEE Trans. Broadcast.*, vol. 50, no. 3, pp. 312–322, Sep. 2004.

[30] K. Seshadrinathan and A. C. Bovik, "Motion tuned spatio-temporal quality assessment of natural videos," *IEEE Trans. Image Process.*, vol. 19, no. 2, pp. 335–350, Feb. 2010.

[31] L. Zhang, G. Tech, K. Wegner, and S. Yea, *Joint Collaborative Team on 3D Video Coding Extensions (JCT-3V)*, document JCT3V-G1005, 7th Meeting, Test model 7 of 3D-HEVC and MV-HEVC, San Jose, CA, USA, Jan. 2014.

[32] W. Yang and W. Liu, "Strong law of large numbers and Shannon–McMillan theorem for Markov chain fields on trees," *IEEE Trans. Inf. Theory*, vol. 48, no. 1, pp. 313–318, Jan. 2002.

[33] M. A. Robertson and R. L. Stevenson, "DCT quantization noise in compressed images," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 1, pp. 27–38, Jan. 2005.

[34] J. Jung, S. Yea, S. Ryu, D. Kim, and K. Sohn, *Depth Distortion Metric With a Weighted Depth Fidelity Term*, document JCT2-A0116, JCT-3V of MPEG and VCEG, 1st Meeting, Stockholm, Sweden, Jul. 2012.

[35] K. Müller and A. Vetro, *Common Test Conditions of 3DV Core Experiments, Joint Collaborative Team on 3D Video Coding Extensions (JCT-3V)*, document JCT3V-G1100, 7th Meeting, San Jose, CA, USA, Jan. 2014.

[36] (2014). *VSRS-1D-Fast*. [Online]. Available: https://hevc.hhi.fraunhofer.de/svn/svn_3DVCSoftware

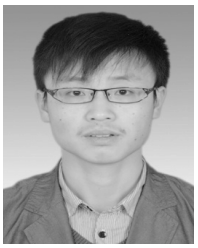
[37] *Methodology for the Subjective Assessment of the Quality of Television Pictures*, document ITU-R BT.500, Nov. 1993.

[38] (2003). *VQEG Final Report of FR-TV Phase II Validation Test*. [Online]. Available: http://www.itu.int/ITU-T/studygroups/com09/docs/tutorialopavc.pdf

[39] (2014). *HTM Reference Software*. [Online]. Available: https://hevc.hhi.fraunhofer.de/svn/svn_3DVCSoftware/tags/HTM-11.0/

[40] G. Bjøntegaard, *Calculation of Average PSNR Differences Between RD-Curves*, document VCEG-M33, 13th Video Coding Experts Group (VCEG) Meeting, Austin, TX, USA, 2001.

Yun Zhang (M'12–SM'16) received the B.S. and M.S. degrees in electrical engineering from Ningbo University, Ningbo, China, in 2004 and 2007, respectively, and the Ph.D. degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences (CAS), Beijing, China, in 2010. From 2009 to 2014, he was a Visiting Scholar with the Department of Computer Science, City University of Hong Kong, Hong Kong. In 2010, he joined the Shenzhen Institutes of Advanced Technology, CAS, as an Assistant Professor. Since 2012, he has been an Associate Professor. His research interests are 3-D video coding, high efficiency video coding, and perceptual video processing.



Xiaoxiang Yang received the B.S. and M.S. degrees in communication engineering from Ningbo University, Ningbo, China, in 2012 and 2015, respectively. From 2013 to 2015, he was a Visiting Student with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences. Since 2015, he has been an Architecture Video Engineer with the Platform Research and Development Department, UcPaas, Shenzhen, China. His research interests are 3-D depth video coding and video codec optimization.

Xiangkai Liu received the B.S. and Ph.D. degrees in computer science from Southwest Jiaotong University, Chengdu, China, in 2009 and 2016, respectively. From 2011 to 2012, he was a Research Associate with the Institute of Digital Media, Peking University, Beijing, China. From 2014 to 2015, he was a Visiting Student with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China. He is currently a Research and Development Engineer with ZTE, Shenzhen, China. His research interests include video coding and video quality assessment.

Yongbing Zhang received the B.A. degree in english and the M.S. and Ph.D degrees in computer science from the Harbin Institute of Technology, Harbin, China, in 2004, 2006, and 2010, respectively. He joined the Graduate School at Shenzhen, Tsinghua University, Shenzhen, China, in 2010, where he is currently an Associate Professor. He was a recipient of the Best Student Paper Award at the IEEE International Conference on Visual Communication and Image Processing in 2015. His current research interests include video processing, image and video coding, video streaming, and transmission.

Gangyi Jiang (M'10) received the M.S. degree in electronics engineering from Hangzhou University in 1992, and the Ph.D. degree in electronics engineering from Ajou University, South Korea, in 2000. He is currently a Professor with the Faculty of Information Science and Engineering, Ningbo University, China. His research interests mainly include video compression and multiview video coding. He has authored over 100 technical articles in refereed journals and proceedings in these fields.

Sam Kwong (M'93–SM'04–F'13) received the B.S. degree in electrical engineering from the State University of New York at Buffalo in 1983, the M.S. degree in electrical engineering from the University of Waterloo, Waterloo, ON, Canada, in 1985, and the Ph.D. degree from the University of Hagen, Germany, in 1996. From 1985 to 1987, he was a Diagnostic Engineer with Control Data Canada. He joined the Bell Northern Research, Canada, as a member of Scientific Staff. In 1990, he became a Lecturer with the Department of Electronic Engineering, City University of Hong Kong, where he is currently a Professor with the Department of Computer Science. His research interests are video and image coding and evolutionary algorithms.