Subjective and Objective Video Quality Assessment of 3D Synthesized Views With Texture/Depth Compression Distortion

Xiangkai Liu, Yun Zhang, *Member, IEEE*, Sudeng Hu, Sam Kwong, *Fellow, IEEE*, C.-C. Jay Kuo, *Fellow, IEEE*, and Qiang Peng

***Abstract*— The quality assessment for synthesized video with texture/depth compression distortion is important for the design, optimization, and evaluation of the multi-view video plus depth (MVD)-based 3D video system. In this paper, the subjective and objective studies for synthesized view assessment are both conducted. First, a synthesized video quality database with texture/depth compression distortion is presented with subjective scores given by 56 subjects. The 140 videos are synthesized from ten MVD sequences with different texture/depth quantization combinations. Second, a full reference objective video quality assessment (VQA) method is proposed concerning about the annoying temporal flicker distortion and the change of spatio-temporal activity in the synthesized video. The proposed VQA algorithm has a good performance evaluated on the entire synthesized video quality database, and is particularly prominent on the subsets which have significant temporal flicker distortion induced by depth compression and view synthesis process.**

***Index Terms*— Video quality assessment, synthesized video quality database, temporal flicker distortion, multi-view video plus depth, view synthesis, 3D video.**

I. Introduction

THE Video Quality Assessment (VQA) plays an important role in evaluating the performances of video processing systems in different applications such as video capture, compression, transmission and so on. A reliable VQA method will provide guidance for optimizing the overall system performance and be beneficial to improve the Quality of Experience (QoE) of the end-users.

Recently, the 3D video is becoming more popular. The advanced 3D video systems are mostly based on Multi-view Video plus Depth (MVD) format [1], which has been adopted by the Moving Picture Experts Group (MPEG) as the recommended 3D video format [2], [3]. In MVD based 3D video system, synthesized views will be generated from the transmitted texture views and their associated depth maps at the receiver using Depth-Image-Based Rendering (DIBR) technology [4]. The visual quality of the synthesized view will affect the perceived picture quality, depth perception and visual comfort of the 3D video systems, which are denoted as the three basic factors determining the QoE of a 3D video system [5]. A VQA algorithm for synthesized view is urgently needed to maintain the 3D QoE requirements.

The appropriate choice of a VQA method for the synthesized view also plays an important role in the performance evaluation of different 3D video processing systems, such as depth generation [6], depth pre/post-processing [7], [8], texture/depth compression [9]–[12], texture/depth bit allocation [13], [14] and so on. The optimization objective functions of these algorithms are all to minimize the distortion of the synthesized view, so the quality of the synthesized view is considered as the most significant evaluation criterion for the whole 3D video processing system.

The most common distortion of the synthesized view comes from texture/depth lossy compression of the MVD data. 3D-AVC [2], [15] and 3D-HEVC [3], [16] are the ongoing standards for MVD data compression which are being developed by the JCT-3V, the Joint Collaborative Team on 3D Video Coding Extensions Development which are composed of experts from MPEG and ITU-T Video Coding Experts Group (VCEG). Compression noise of the texture/depth view pairs will affect the synthesized video through the view synthesis process. In this paper, we mainly focus on the video quality assessment of the synthesized view with texture/depth compression distortion. This study will be valuable for the design, optimization and evaluation of MVD based 3D video systems, e.g., help adjust the distortion metrics of different 3D video coding strategies, to improve the perceptual quality of the synthesized video with less bit cost.

Manuscript received January 11, 2015; revised June 9, 2015 and July 21, 2015; accepted August 7, 2015. Date of publication August 17, 2015; date of current version September 18, 2015. This work was supported in part by the National Natural Science Foundation of China under Grant 61471348, Grant 61102088, and Grant 61272289, in part by the Shenzhen Overseas High-Caliber Personnel Innovation and Entrepreneurship Project under Grant KQCX20140520154115027, and in part by the Guangdong Special Support Program for Youth Science and Technology Innovation Talents under Grant 2014TQ01X345. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Ivana Tosic. *(Corresponding author: Yun Zhang.)*

X. Liu is with the School of Information Science and Technology, Southwest Jiaotong University, Chengdu 610031, China, and also with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China (e-mail: xk.liu@siat.ac.cn).

Y. Zhang is with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China (e-mail: yun.zhang@siat.ac.cn).

S. Hu and C.-C. J. Kuo are with the Ming Hsieh Department of Electrical Engineering, University of Southern California, Los Angeles, CA 90089 USA (e-mail: sudenghu@usc.edu; cckuo@sipi.usc.edu).

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

Q. Peng is with the School of Information Science and Technology, Southwest Jiaotong University, Chengdu 610031, China (e-mail: qpeng@home.swjtu.edu.cn).

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

Digital Object Identifier 10.1109/TIP.2015.2469140

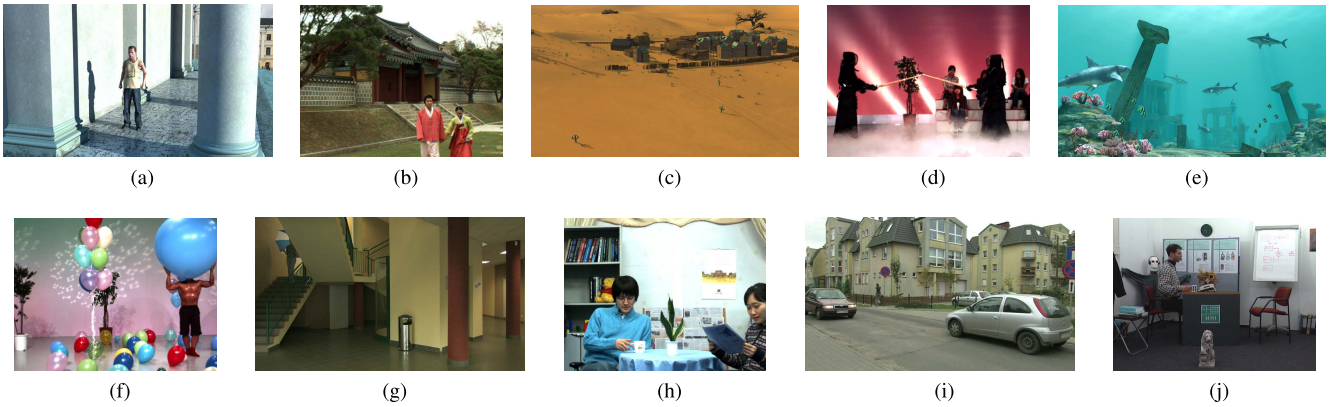

Fig. 1. Ten MVD sequences used in SIAT database. (a) Dancer. (b) Lovebird1. (c) GTFly. (d) Kendo. (e) Shark. (f) Balloons. (g) PoznanHall2. (h) Newspaper. (i) PoznanStreet. (j) BookArrival.

II. Related Works

There are two categories of VQA studies: subjective and objective studies. Subjective method is performed by asking human subjects to rate the quality of the video that they are viewing. ITU-R Rec.BT.500 [17] and ITU-T Rec.P.910 [18] described some subjective test procedures. Objective VQA algorithm is used to predict human's judgement on the video quality [19]. Subjective method is more accurate but it is time consuming and impractical for realtime video processing systems. The subjective experiments are usually constructed to provide the quality scores which can be used to evaluate the performances of different objective VQA methods.

A. Subjective Study

For traditional 2D video subjective quality assessment, there are two famous subjective studies conducted by the Video Quality Experts Group (VQEG) [20] and the Laboratory for Image and Video Engineering (LIVE) [21]. The video quality databases with the associated subjective scores of VQEG FRTV phase 1 and LIVE are publicly available and widely used. With the rapid development of 3D video technology, the 3D QoE about depth perception and visual comfort has attracted a lot of attentions. Lebreton *et al*. [22] conducted subjective experiments on the depth perception quality of different 3D video content. A new 3D QoE subjective assessment procedure is proposed in [23] to measure the visual discomfort. However, the stereo view pairs are expected to be conducted with the synthesized views in MVD based 3D video systems, in which case the picture quality affecting 3D QoE will be determined by the quality of the synthesized view.

For the subjective quality assessment of the synthesized view, the only publicly available database is the IRCCyN/IVC DIBR Videos Quality Database [24] proposed by Bosc *et al*. [25]. The IRCCyN database contains three MVD sequences and 102 videos in 1024×768 resolution. Quality scores obtained from 32 human subjects in an ACR-HR [18] experiment are provided. Only one single reference view was used to generate the synthesized video in the IRCCyN database, and the distortions in this database are mainly related to the hole filling strategies of different DIBR algorithms. However, in the MVD based 3D video systems, two reference views are usually used to generate the synthesized video which can significantly reduce the distortions at hole filling stages. Moreover, the significant distortions due to texture/depth compression are not included in the IRCCyN database. In [26], the quality of the synthesized view with depth compression distortion was studied. Using the decoded depth maps, 50 intermediate synthesized images were generated in-between two reference views and were used to simulate a smooth camera motion from left to right. The synthesized images were rated by 27 subjects. However, there is no temporal distortion as the test sequences in [26] are composed by still images at the same time instance.

In this paper, we develop a synthesized video quality database which includes ten different MVD sequences and 140 synthesized videos in 1024×768 and 1920×1088 resolution, as shown in Fig.1. For each sequence, 14 different texture/depth quantization combinations were used to generate the texture/depth view pairs with compression distortion. A total of 56 subjects participated in the experiment. Each synthesized sequence was rated by 40 subjects using single stimulus paradigm with continuous score. Individual votes and the Difference Mean Opinion Scores (DMOS) are provided. This subjective study and the synthesized videos, named as Shenzhen Institutes of Advanced Technology (SIAT) Synthesized Video Quality Database, are expected to supplement the IRCCyN database. The SIAT database with associated subjective data are available for download [27].

B. Objective Study

A simplified objective method for evaluating the video quality is to calculate the image distortion using an Image Quality Assessment (IQA) method, e.g., Peak Signal to Noise Ratio (PSNR) and Structural SIMilarity (SSIM) [28], and take the average of the image scores as the sequence score. For 3D stereoscopic IQA, a method based on binocular perception is proposed in [29], which divides the image into several binocular regions and evaluates each region independently. Lin and Wu [30] proposed a binocular frequency integration method to measure the quality of stereoscopic 3D images. Compared with IQA, the crucial temporal distortions are

concerned in VQA methods. For conventional 2D video quality assessment, the MOtion-based Video Integrity Evaluation (MOVIE) index [31] is proposed to measure the temporal distortions using the responses of spatio-temporal Gabor filters which are tuned to the movement in the reference video. In case of 3D stereoscopic video, a full reference VQA model is trained on a subjective database of compressed stereoscopic videos in [32].

For the objective VQA of the synthesized view, new distortion types have appeared due to new distortion sources. The depth compression distortions and the view synthesis process usually produces temporal flickering and spatial object edge distortions. However, the distortion of the synthesized view is simply computed using the Sum of Squared Differences (SSD) in the reference software of 3D-AVC [33]. There are two major reasons that the conventional 2D VQA methods are not sufficient enough to assess the visual quality of the 3D synthesized views. First, the conventional VQA metrics will underestimate some dominant distortions of the synthesized view such as flickering, inconsistent object shifting and edge distortion which are annoying and noticeable for subjects. Second, the distortions such as tiny geometric distortion, consistent object shifting, inter view camera noise and illumination difference which can hardly be perceived by human subjects might be overestimated by traditional VQA metrics [25].

In recent years, several algorithms have been proposed to evaluate the quality of the 3D synthesized view [34]–[39], Zhao and Yu [34] proposed a Peak Signal to Perceptible Temporal Noise Ratio (PSPTNR) metric to compute the visible temporal noise in the background area of a synthesized sequence. Ekmekcioglu *et al*. [35] designed a synthesized VQA framework considering distortions in depth nearer regions and motionless regions. The limitation of the metrics in [34] and [35] is that if camera movement exists in the test video, these methods may not be applicable. Bosc *et al*. [25], [36] proposed to measure the wrong displacement of the object edge in synthesized view, which is induced by the DIBR process. In [37] and [38], a wavelet decomposition based synthesized view quality assessment method was presented. Before quality evaluation, a registration procedure in wavelet domain is used to align the content of the synthesized image and the reference image. This process is under the assumption that imperceptible object shift introduced by the synthesis process will not affect the visual quality. Tsai and Hang [39] proposed to compensate the object shift effect with the similar assumption to [37] and [38] and computed the noise at object boundaries of a synthesized image with depth distortion. However, the methods in [36]–[39] are all IQA methods not involving the studies about the significant temporal distortion. Actually, in case of watching the synthesized video, flicker distortion in temporal domain becomes the dominant noise in overall quality assessment.

In this paper, we propose a full reference objective VQA algorithm for the synthesized video quality evaluation. The proposed VQA is based on two quality features which are extracted from both spatial and temporal domains of the synthesized sequence. The first feature focuses on capturing the temporal flicker distortion induced by depth compression and view synthesis process, and the second feature is used to measure the change of the spatio-temporal activity in the synthesized sequence due to the blurring and blockiness distortion caused by texture compression.

In general, our contributions are two-fold. First, a synthesized video quality database with texture/depth compression distortion and the associated subjective scores are released. Second, a full reference objective VQA algorithm primarily focusing on the temporal flicker distortions of the synthesized video is proposed. The remainder of this paper is organized as follows. In Section III, we analyze the typical distortions in synthesized video. Subjective study of the synthesized video is described in Section IV. The objective VQA method is presented in Section V. In Section VI, we present the experimental results. Finally, Section VII gives the conclusions.

III. Distortion Analysis of Synthesized View

As new sources of distortions, e.g., depth compression and synthesis process, have been induced, there are new types of noise in synthesized video compared with the conventional 2D video. Quality score of the synthesized videos evaluated by the traditional 2D quality metrics may not match the perception results of human subjects. As seen in Fig. 2b and Fig. 2d, Y-PSNR of the synthesized image containing almost no perceptible noise is only 32.03 dB, less than the Y-PSNR of the JM [40] compressed image (32.04 dB, seen in Fig. 2a and Fig. 2c). However, the perceived quality of the synthesized image is much better than that of the compressed one. Actually, the synthesis process will not influence the value of the pixels but change their position. For synthesized view, the conventional 2D quality metrics might overestimate some distortions while underestimate the others. In this section, we will discuss these two kinds of distortions in synthesized view.

A. Overestimated Distortions

Fig. 2e shows the per-pixel comparison between the synthesized image and the reference image. If the absolute value of the difference between pixels at the same position in two images is more than the comparison threshold, the pixel value is set to 255 (white), otherwise it is set to 0 (black). It can be observed that there are many unmatched pixels in the synthesized image although this image is synthesized from the original texture/depth pairs. Inter view *camera noise* and *luminance difference* commonly exist in multi-view video capturing stage and will induce objective distortions all over the synthesized image. A large number of *tiny geometric distortions* are caused by the depth inaccuracy and the numerical rounding operation of pixel positions. Incorrect depth information also induces *object shifting* on the synthesized images as the distorted books shown in Fig. 2g. These distortions discussed above can be detected by the pixel-wise metrics, e.g., PSNR. However, most of these distortions may not be perceived by human subjects. The subjective quality of consistently shifted objects is still high as they are temporally constant. The degradation of the visual quality has been overestimated.

B. Underestimated Distortions

Fig. 2f and Fig. 2g show the per-pixel Squared Differences (SD) between the compressed/synthesized image

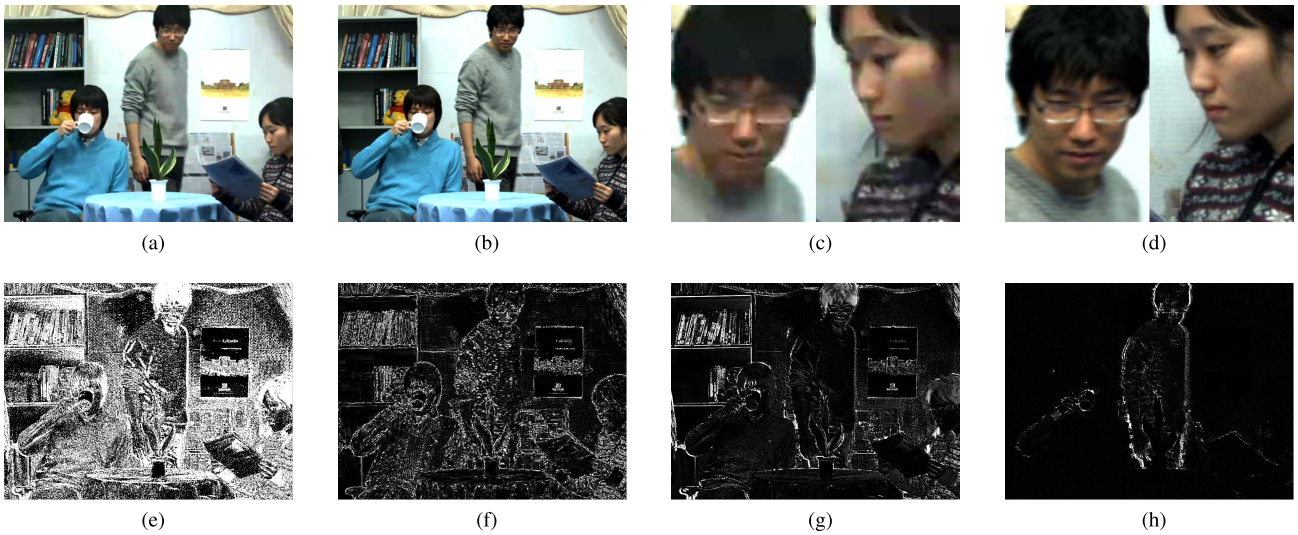

Fig. 2. Distortion analysis in the synthesized view. (a) The frame 100 of view 3 in *Newspaper* sequences compressed by JM18.0 as P frame with QP 42 (Y-PSNR: 32.03dB). (b) The frame 100 of view 3 in *Newspaper* sequences synthesized from view 2 and 4 with original texture/depth pair (Y-PSNR: 32.04dB). (c) Details of image (a) at human face region. (d) Details of image (b) at human face region. (e) Per-pixel comparison (threshold is three) between image (b) and the reference (the original frame 100 in view 3). (f) Squared differences between image (a) and the reference. (g) Squared differences between image (b) and the reference. (h) Temporal squared differences between image (b) and the reference. Note that (e)-(h) are all computed in luminance component domain.

and the reference image, respectively, which is computed as

$$SD_{i,n} = (P_{i,n} - R_{i,n})^2, \tag{1}$$

where (i, n) represents the ith pixel in the nth frame; P and R are the processed sequence (referring to the compressed or synthesized sequence) and the reference sequence. Distortions due to lossy compression are distributed on the most of regions in an image (Fig. 2f), while view synthesis induced noise mainly appear along the object edge (Fig. 2g). For conventional distortion types such as quantization distortion and additive noise, the Human Visual System (HVS) sensitivity in plain area is higher than that in the edge region [41]. However, for the view synthesis induced geometry distortion, errors in edge regions will be more obvious. When moving foreground object exists, the synthesis distortions around the border areas between the object and the background will be more noticeable. Conventional quality metrics usually underestimate the effect of *edge damage* and treat every region in an image equally [42]. For synthesized view quality assessment, distortion in edge regions should be considered seriously.

Some local statistics based image quality metrics, such as SSIM, may not be suitable for measuring the geometric distortion due to depth errors. As we increase the distortion intensity in the synthesized view by compressing the depth maps with larger quantization step size, the SSIM value only changes a little. This is because the distortions caused by the synthesis process and depth distortion are mainly geometric displacement distortion, which will not largely change the average luminance and the variance of a local image patch.

Fig. 2h shows the Temporal Squared Differences (TSD) between the synthesized and the original sequences, which is defined as

$$TSD_{i,n} = ((P_{i,n} - P_{i,n-1}) - (R_{i,n} - R_{i,n-1}))^2. \tag{2}$$

Actually, the temporal variation of TSD might induce the *flicker distortion* which will cause an obviously visual quality degradation. For the synthesized sequences, flickering is the main distortion affecting human perception. However, the influence of the temporal *flicker distortion* in quality evaluation is underestimated by most of the traditional 2D VQA algorithms, which is verified by the subjective experimental results in Section IV-D. More detailed discussions about the *flicker distortion* is presented at Section V-A.

IV. Subjective Study of the Synthesized Video

A. Generation of Synthesized Sequences

Ten MVD sequences, as shown in Fig. 1, provided by MPEG for the 3D video coding contest [43], [44] are used in our experiment. The information about the test sequences is given in Table I. The *Dancer*, *GT Fly* and *Shark* sequences are computer animation videos, and the others are natural scenes. There exist camera movement in sequences *Balloons*, *Kendo*, *Dancer*, *GT Fly*, *Shark* and *PoznanHall2*, while the cameras in *BookArrival*, *Lovebird1*, *Newspaper*, *PoznanStreet* are still.

Firstly, the input texture views and their associated depth maps of each sequences were compressed using different quantization parameters (QPs) spanned from QP = 20 to QP = 50. Then, for each sequence, 14 combinations of texture/depth view pairs after compression were selected manually to be used for view synthesis. Finally, 14 synthesized videos were generated at the target viewpoint for each sequence using the selected texture/depth view pairs. The 14 synthesized videos can be divided into 4 distortion categories:

- Uncompressed Texture and Uncompressed Depth ($U_T U_D$), one synthesized video is generated using the original texture/depth view pair.
- Uncompressed Texture and Compressed Depth ($U_T C_D$), the depth views are compressed with four different QPs,

TABLE I
SEQUENCE INFORMATION AND TEXTURE/DEPTH COMPRESSION QP PAIRS

Sequence	Resolution	Frame Rate	Synthesized Frames	Input View Pair	Output View	Dep. QP ($U_T C_D$)	Tex. QP ($C_T U_D$)	(Tex.,Dep.) QP Pair ($C_T C_D$)
BookArrival	1024 × 768	16.67fps	100	6 - 10	8	28,36,40,44	28,34,38,42	(22,26),(28,32),(34,36),(38,40),(42,44)
Balloons	1024 × 768	30fps	200	1 - 5	3	32,36,40,46	24,32,38,42	(24,32),(28,36),(32,40),(40,42),(42,46)
Kendo	1024 × 768	30fps	200	1 - 5	3	32,38,44,48	24,28,32,40	(24,32),(32,34),(36,38),(40,44),(44,46)
Lovebird1	1024 × 768	30fps	200	4 - 6	5	36,38,40,48	28,30,34,38	(28,36),(30,40),(34,44),(38,48),(42,50)
Newspaper	1024 × 768	30fps	200	2 - 4	3	28,36,44,50	24,30,34,38	(28,32),(32,40),(38,44),(36,50),(42,48)
Dancer	1920 × 1088	25fps	200	1 - 9	5	24,28,40,45	28,32,40,44	(24,20),(30,24),(32,28),(32,40),(44,35)
PoznanHall2	1920 × 1088	25fps	200	5 - 7	6	28,32,40,46	24,28,32,38	(24,28),(26,32),(34,36),(32,40),(40,42)
PoznanStreet	1920 × 1088	25fps	200	3 - 5	4	32,38,44,48	26,30,38,42	(22,28),(26,40),(30,44),(34,48),(42,35)
GT Fly	1920 × 1088	25fps	200	1 - 9	5	28,36,44,48	24,36,40,44	(24,28),(32,36),(34,38),(40,44),(44,48)
Shark	1920 × 1088	25fps	200	1 - 9	5	28,36,40,44	24,32,36,40	(24,28),(32,36),(36,40),(40,36),(42,48)

and four synthesized videos are generated using the original texture and the compressed depth views.

- Compressed Texture and Uncompressed Depth ($C_T U_D$), the texture views are compressed with four different QPs, and four synthesized videos are generated using the compressed texture and the original depth views.
- Compressed Texture and Compressed Depth ($C_T C_D$), the texture/depth view pairs are compressed with five different texture/depth QP pairs, and five synthesized videos are generated using the compressed texture/depth view pairs.

The input texture/depth view pairs were encoded with 3DV-ATM v10.0 [45], which is the 3D-AVC [2] based reference software for MVD coding. The view synthesis algorithm used is the VSRS-1D-Fast software [46] implemented in the 3D-HEVC [3] reference software HTM [47].

There are two criteria on selecting the compressed texture/depth view pairs. First, the subjective quality of the synthesized videos should span a large range, as recommend by the VQEG in [20]. Second, the quality gap between each test sequences should not be too small to avoid difficulties in discrimination. With these two criteria, a large set of synthesized videos were generated and viewed by a group of subjects, and only a subset of these videos were chosen to be included in our SIAT Synthesized Video Quality Database. We believe that a carefully selected database is more useful than fixing the compression rates across sequences. The final chosen texture/depth compression QP pairs of each sequence are given in Table I. The aim of compiling this database is to evaluate the performance of different objective VQA for 3D synthesized video in the context of MVD compression.

B. Subjective Experiment Design

We use a single stimulus continuous quality scale method as it is in [21] to obtain subjective quality scores for the synthesized sequences. MSU Perceptual Video Quality Tool [48], a software for subjective video quality evaluation provided by the Graphics and Media Lab of Moscow State University (MSU), is adopted as the user interface in our

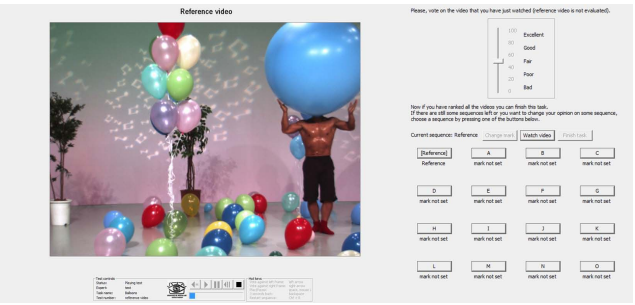

Fig. 3. MSU Perceptual Video Quality Tool [48]. For the convenience of exhibition, the scroll bar for video quality rating has been put together with the display interface in the figure.

subjective experiment, as shown in Fig. 3. During the test, subjects rated the video quality with continuous scores ranging from 0 to 100. The tags beside the scroll bar indicating five quality levels of Excellent, Good, Fair, Poor, Bad, which are same as the ITU-R ACR, would assist subjects making decisions. The reference videos of each test sequences were also viewed by the subjects. All sequences (including the reference videos) are hidden under letters (as shown in Fig. 3) and the play order for each subject is random.

The videos were viewed by the subjects on a LG 42LH30FR monitor with 1920 × 1080 resolution. The HP Z800 workstation with the nVIDIA Quadro 2000 graphics card and 1GB video memory is used to play out the test videos at the right frame rate without any delay. The environment of the subjective experiment is conducted as recommended by ITU-R Rec.BT.500.

The experiment is conducted in two sessions. 84 videos were evaluated in the first session and 56 videos were evaluated in the second session. There were a total of 56 non-expert subjects (38 males and 18 females with an average age of 24) participated in the experiment and each test video was evaluated by 40 subjects. The quality scores assigned by each subject should be normalized per person before computing DMOS. Because there were 24 subjects participated in both the first and the second sessions, to avoid inconsistent scoring criteria of the same subject between

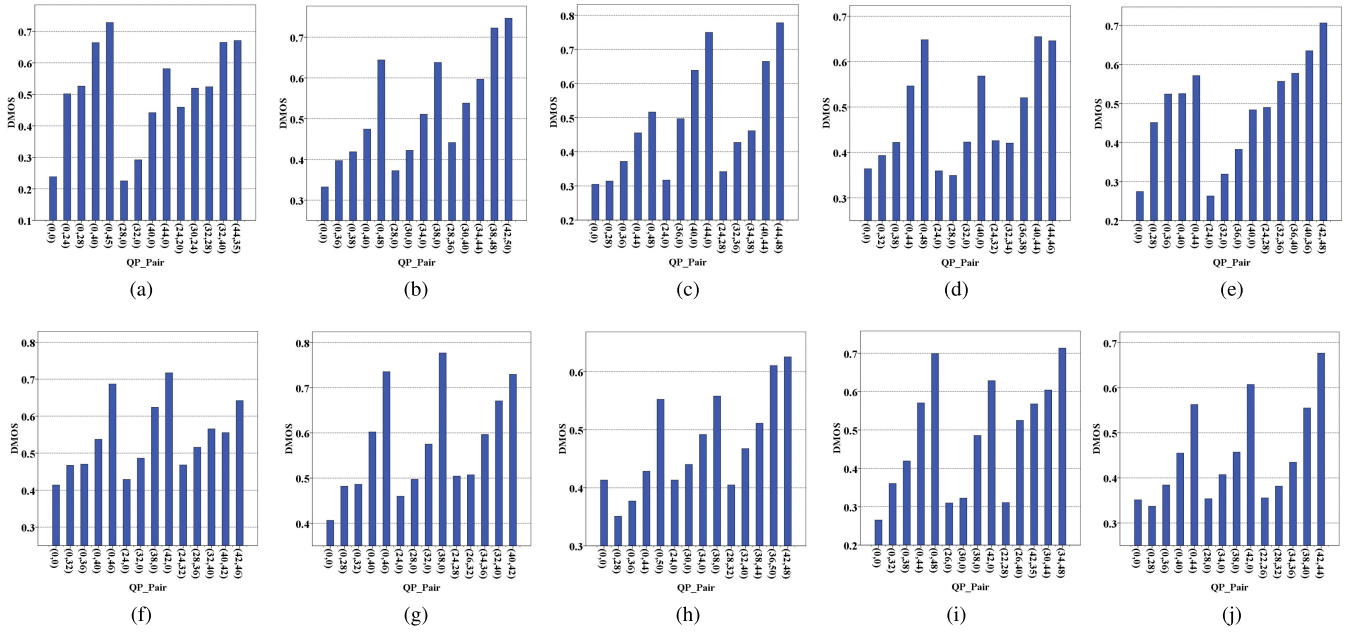

Fig. 4. DMOS of the synthesized sequences with different texture/depth compression QP pairs in the SIAT Synthesized Video Quality Database. The pair of values on the horizontal axis represent the (texture, depth) QP. The values on the vertical axis represent the DMOS value. (a) Dancer. (b) Lovebird1. (c) GT Fly. (d) Kendo. (e) Shark. (f) Balloons. (g) PoznanHall2. (h) Newspaper. (i) PoznanStreet. (j) BookArrival.

two sessions, all the reference videos were included in both sessions using hidden reference removal procedure. The DMOS score for each test video was computed *per session* using the quality score assigned to the corresponding reference video *in that session*. All subjects taking part in the experiment are graduate students. Every subject had been told the procedure of the test and had watched a training sequence before starting the experiment.

C. Processing of Subjective Scores

After finishing the rating process, subjects with extreme scores should be rejected as recommended in ITU-R.BT500 [17]. In this experiment, only 1 subject (male) in the first session is eliminated, and there are 55 remaining subjects in our final database. Then, difference scores are obtained by subtracting the score of the test sequence from the score of the reference sequence. Both scores are given by the same subject *in the same session*. Let $s_{i,j}$ denotes the score assigned by subject i to sequence j, and j_{ref} is the reference video of sequence j. The difference score $d_{i,j}$ is computed as

$$d_{i,j} = s_{i,j_{ref}} - s_{i,j}. \tag{3}$$

4.13% (228 of 5516) of the difference scores are negative, i.e., the score of the reference video is less than the score of the tested video. These negative difference scores might be from the special preference of some subjects, but not from experiment mistakes. We keep these scores negative to ensure the original diversity of the subjective data.

The difference scores are then normalized to z-scores per person and *per session*. To make the data more intuitive, the normalized z-scores are scaled to a mean of 0.5 with a standard deviation of 1/6. Finally, we compute DMOS for each synthesized video as

$$DMOS_j = \frac{1}{M}\sum_{i=1}^{M} z_{scaled}(i, j), \tag{4}$$

where $z_{scaled}(i, j)$ is the re-scaled z-score assigned by subject i to video j, which has been evaluated by M subjects.

D. Subjective Data Analysis

1) Subjective Scores Analysis: Fig. 4 shows the DMOS of the synthesized sequences with different texture/depth QP pairs in the SIAT database. Three conclusions can be drawn from the distribution of the DMOS data. First, the quality of the synthesized sequences span a wide range, e.g. from excellent to bad. Second, obvious discrimination exists between most of the test videos, which can avoid the inaccurate rating scores due to indistinguishable quality. Third, the videos with flicker distortion caused by depth distortion have worse perceptual quality than the videos with texture compression distortion. For example, in *Lovebird1* sequence, the Y-PSNR of the videos synthesized with texture/depth QP pair of (0, 40) and (28, 0) are 32.59 dB and 32.23 dB, respectively. However, their DMOS values are 0.474 and 0.373, which has proved the assumptions in Section III-B that the flicker distortion has been underestimated by traditional 2D VQA.

Some special results should also be noted. As seen in Fig. 4h, for the *Newspaper* sequence, the perceptual quality of the synthesized video with texture/depth QP pair of (0, 28), (0, 36) and (28, 32) is better than the quality of the video synthesized from the original texture/depth pairs. This is because the compression distortion in depth maps has induced a smoothing effect on the inaccurate edge region, which can reduce the noise around object edge in the synthesized view

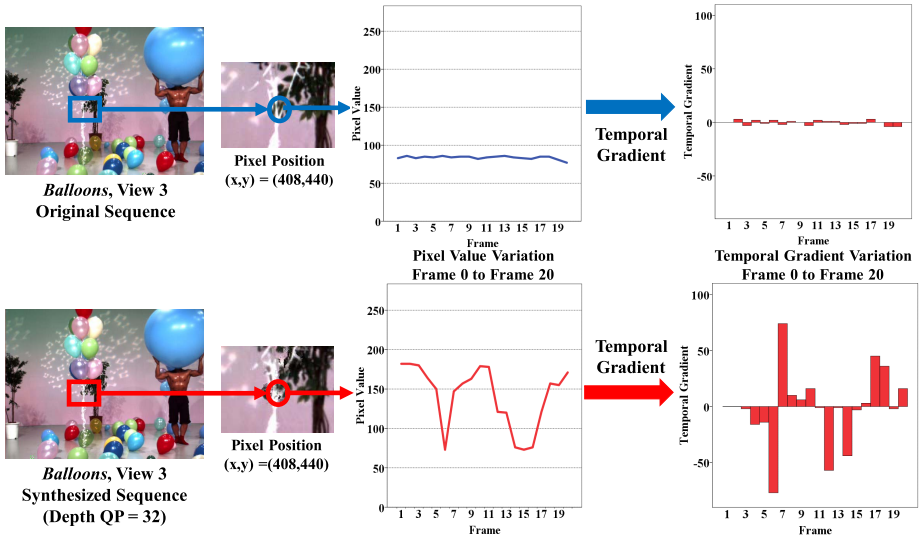

Fig. 5. Temporal Flickering Analysis.

TABLE II
ANOVA OF EACH SEQUENCE IN DIFFERENT DISTORTION CATEGORY

Sequence	$U_T C_D$	$C_T U_D$	$C_T C_D$
Dancer	2.31×10^{-21}	1.68×10^{-23}	4.34×10^{-21}
Lovebird1	5.93×10^{-19}	2.47×10^{-22}	3.80×10^{-33}
GT Fly	6.76×10^{-15}	1.33×10^{-41}	1.57×10^{-59}
Kendo	1.76×10^{-18}	8.62×10^{-15}	2.73×10^{-24}
Shark	6.53×10^{-6}	1.10×10^{-22}	1.38×10^{-18}
Balloons	5.53×10^{-13}	1.67×10^{-20}	2.73×10^{-10}
PoznanHall2	1.41×10^{-18}	1.17×10^{-25}	2.62×10^{-25}
Newspaper	2.50×10^{-12}	6.75×10^{-8}	1.19×10^{-22}
PoznanStreet	1.06×10^{-34}	3.30×10^{-35}	1.42×10^{-45}
BookArrival	5.71×10^{-22}	1.12×10^{-23}	4.45×10^{-39}

due to imperfect original depth maps. Another special case is shown in Fig. 4d, for the *Kendo* sequence, the quality of the video synthesized with texture/depth QP pair of (28, 0) is higher than the video generated with a QP pair of (24, 0). As the temporal distortion around the dis-occluded regions due to view synthesis process is quite strong in this sequence, a slight compression distortion in the texture views may mask the synthesis noise in some degree.

2) Significant Difference in Each Sequence of Different Subsets: There are four distortion levels in subsets of $U_T C_D$, $C_T U_D$, and five in $C_T C_D$. An Analysis of Variance (ANOVA) [49] are performed for the quality scores of each sequence in subsets $U_T C_D$, $C_T U_D$ and $C_T C_D$, respectively. The purpose of this analysis is to examine the significance of difference between each distortion level, since one of our criteria on conducting the database is to make the quality of each test video be easily distinguished. As seen in Table II, the results of ANOVA for all sequences and distortion categories are far less than 0.05 (significance level is 95%), which means that significant differences exist in each sequence of different subsets. To sum up, the resulting subjective data proves the effectiveness of the synthesized sequences selection and the subjective experiment design.

V. OBJECTIVE VIDEO QUALITY ASSESSMENT METHOD FOR SYNTHESIZED VIEW

The proposed objective VQA for synthesized view is based on two quality features. The first feature is used to measure the temporal flickering and the second is related to the spatio-temporal activity. The distortions computed with these two features are integrated to obtain the overall quality of the synthesized video.

A. Temporal Flickering Analysis

The most annoying temporal distortion in synthesized sequences is flickering. The inaccurate depth and the synthesis process will induce geometric distortions, e.g., pixels are projected to wrong positions [50]. However, not all the geometric distortions cause flickering. The inconsistent position errors around high contrast regions, which are always the dis-occluded area of the foreground objects edge, are prone to induce flickering.

Actually, flickering can be observed as significant and high frequency alternated variation between different luminance levels. For example, in the synthesized sequence of *Balloons* with depth compression distortion, an obvious flicker distortion exists around the balloon rope as shown in Fig. 5. The variation of the pixel luminance at position (408, 440) in the first 20 frames is shown. There are significant fluctuations of the pixel value in the synthesized view, but the value changes very little in the original view at the same pixel position. The minor changes of the pixel value in original view is caused by camera random noise, and the significant variation of the synthesized pixel value is due to inconsistent geometric distortion. Let $I(x, y, i)$ denotes the luminance value of the pixel at spatial position (x, y) in frame i.

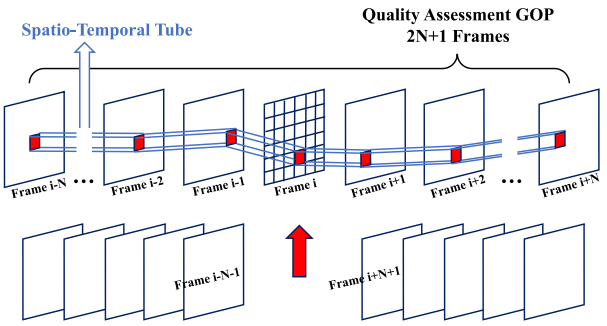

Fig. 6. Quality Assessment GOP and Spatio-Temporal Tube.

The temporal gradient vector $\vec{\nabla} I_{x,y,i}^{temporal}$ can be computed as

$$\vec{\nabla} I_{x,y,i}^{temporal} = I(x, y, i) - I(x', y', i-1), \tag{5}$$

where (x', y') is the coordinates corresponding to (x, y) along the motion trajectory in frame $i-1$. As shown in Fig. 5, the flickering can be seen as the high frequency direction variation $\vec{\nabla} I_{x,y,i}^{temporal} \times \vec{\nabla} I_{x',y',i-1}^{temporal} < 0$ and the significant magnitude change $\left| \vec{\nabla} I_{x,y,t}^{temporal} - \vec{\nabla} I_{x',y',i-1}^{temporal} \right|$ of the temporal gradient vector.

B. Spatio-Temporal Structure for Quality Assessment

Two structures called Quality Assessment Group of Pictures (QA-GoP) and Spatio-Temporal (S-T) tube are defined in our VQA algorithm. As shown in Fig. 6, the sequence can be divided into several QA-GoPs, which contain a group of pictures with $2N+1$ frames length. The structure of S-T tube is generated with a block based motion estimation process. First, The central picture in a QA-GoP is partitioned into a number of 8×8 blocks. Then, each block will be used as a reference to search a matching block in both forward and backward adjacent frames. The obtained matching blocks are used as reference blocks to continue the motion search process until both ends of the QA-GoP have been reached. Finally, $2N+1$ blocks along the motion trajectory starting from the central frame construct a S-T tube. Using the central frame as the starting point of the motion search process can reduce the error accumulation due to imperfect motion estimation. The motion estimation process is operated on the original video.

The purpose of introducing the structure of S-T tube is to compute the distortion along the motion trajectory. This is because the view synthesis distortion, e.g., the flickering, always locates on the edge of the foreground object which usually has a movement. Actually, even when the synthesis distortion locates around the background motionless object, there may still exist a camera movement induced global motion in the scene. Considering the efficiency and performance, the new three-step search algorithm [51] is used for block motion estimation. With the block based motion vector field, a six-parameter affine motion model is estimated for computing global motion [52]

$$\begin{pmatrix} x'' \\ y'' \end{pmatrix} = \begin{pmatrix} \theta_1 x + \theta_2 y + \theta_3 \\ \theta_4 x + \theta_5 y + \theta_6 \end{pmatrix}, \tag{6}$$

where (x, y) and (x'', y'') are coordinates in the current frame and the target frame, respectively. The vector $\Theta = [\theta_1, \ldots, \theta_6]$ represents the parameters of the affine motion model. Then, the global motion vector can be computed with subtracting (x, y) from (x'', y''). The affine model can describe common camera movement such as translation, rotation and zoom. The obtained global motion vector is used to detect the disappearance of the tracked object in which case the distortion calculated in the corresponding blocks should be excluded.

During the quality assessment process, distortions of each S-T tube will be computed at first. Then, the quality score of the QA-GoP is obtained from the S-T tube scores using a spatial pooling process. Finally, the distortion of the whole sequence is computed by averaging the QA-GoP scores.

C. Flicker Distortion Measurement

From a pixel position of (x_i, y_i) in the central frame i of a QA-GoP as in Fig. 6, the corresponding pixel coordinates along the motion trajectory in the S-T tube could be obtained as $[(x_{i-N}, y_{i-N}), \ldots, (x_i, y_i), \ldots, (x_{i+N}, y_{i+N})]$ as described in Section V-B. The flicker distortion DF_{x_i,y_i} at position (x_i, y_i) of the central frame i will be calculated along the motion trajectory as

$$DF_{x_i,y_i} = \sqrt{\frac{\sum_{n=i-N+1}^{i+N} \Phi(x_n, y_n, n) \cdot \Delta(x_n, y_n, n)}{2N}}, \tag{7}$$

where $2N+1$ is the length of a QA-GoP. The product of $\Phi(x, y, n) \cdot \Delta(x, y, n)$ denotes the flicker distortion at position (x, y) in frame n. Specifically, $\Phi(\cdot)$ is used to detect the potential sensible flicker distortion, and $\Delta(\cdot)$ is used to measure the strength of the flicker distortion. $\Phi(x, y, n)$ is defined as

$$\Phi(x, y, n) = \begin{cases} & \text{if } \vec{\nabla} I_{x,y,n}^{temporal} \times \vec{\nabla} \widetilde{I}_{x,y,n}^{temporal} \le 0 \\ 1, & \text{and } \vec{\nabla} \widetilde{I}_{x,y,n}^{temporal} \neq 0 \\ & \text{and } \left| I(x, y, n) - \widetilde{I}(x, y, n) \right| > \mu \\ 0, & \text{otherwise,} \end{cases} \tag{8}$$

where I and $\widetilde{I}$ denote the original and the synthesized pixels, respectively. $\vec{\nabla} I_{x,y,n}^{temporal} \times \vec{\nabla} \widetilde{I}_{x,y,n}^{temporal} \le 0$ and $\vec{\nabla} \widetilde{I}_{x,y,n}^{temporal} \neq 0$ mean that the directions of the temporal gradients of $\widetilde{I}(x, y, n)$ and $I(x, y, n)$ are different. $\vec{\nabla} \widetilde{I}_{x,y,n}^{temporal} \neq 0$ is used to exclude the case in which the synthesized pixel has no temporal variation. This detection function is used to find the wrong variation of pixel luminance, which may be a potential flicker distortion. The value μ is a perceptual threshold. When the absolute value of the difference between $I(x, y, n)$ and $\widetilde{I}(x, y, n)$ is less than μ, errors will be ignored.

In addition to the flickering detection, $\Delta(x, y, n)$ in (7) is used to measure the strength of the flicker distortion, which is computed as

$$\Delta(x, y, n) = \left(\frac{\vec{\nabla} \widetilde{I}_{x,y,n}^{temporal} - \vec{\nabla} I_{x,y,n}^{temporal}}{\left| \vec{\nabla} I_{x,y,n}^{temporal} \right| + C} \right)^2, \tag{9}$$

where $\vec{\nabla} \widetilde{I}_{x,y,n}^{temporal} - \vec{\nabla} I_{x,y,n}^{temporal}$ reflects the magnitude of the temporal gradient distortion. Temporal visual masking effect

is considered by dividing the difference by $\left|\vec{\nabla} I_{x,y,n}^{temporal}\right| + C$. C is a constant to avoid the denominator being zero. In this paper, the constant C is chosen to be 1 which is of the same order of magnitude as the quantities in the denominators [31]. Because the directions of these two temporal gradients are different, the value of $\Delta(x, y, n)$ will be proportional to the strength of flicker distortion.

The flicker distortion in a S-T tube, DF^{tube}, is computed by averaging the distortion $DF_{x,y}$ in the block of the central frame as

$$DF^{tube} = \frac{\sum_{x=1}^{w}\sum_{y=1}^{h} DF_{x,y}}{w \times h}, \tag{10}$$

where w and h are the width and height of the blocks in the S-T tube, which are both 8. Because the most serious synthesis distortions often occupy a small part of the image instead of uniformly distributed in the whole picture, a worst-case spatial pooling method [53] is used to obtain the flicker distortion DF^{GoP} of the QA-GoP, which is computed as

$$DF^{GoP} = \frac{1}{N_{\mathbb{W}}}\sum_{k\in\mathbb{W}} DF_k^{tube}, \tag{11}$$

where $\mathbb{W}$ denotes the set with the worst $W\%$ DF^{tube} in the QA-GoP, and $N_{\mathbb{W}}$ is the number of S-T tubes in $\mathbb{W}$. For flicker distortion measurement, W is set to 1, i.e., only the worst 1% DF^{tube} are averaged to obtain the DF^{GoP} as the small regions with serious flicker distortion have the most dominant effect on the overall perceptual quality.

Finally, the flicker distortion DF^{Seq} of the sequence is computed by taking average of the DF^{GoP} as

$$DF^{Seq} = \frac{1}{K}\sum_{m=0}^{K} DF_m^{GoP}, \tag{12}$$

where K is the number of QA-GoPs in the sequence.

D. Distortion of Spatio-Temporal Activity

The compression of the texture view pairs will induce blurring and blockiness distortion to the synthesized view, which is very similar to the traditional 2D situation. In this paper, the spatio-temporal activity is defined based on the variance of the pixel gradients in a spatio-temporal region. Image blurring caused by texture compression will weaken the activity [54], but the synthesis noise and compression blocking artifacts may strengthen the activity. To compute the activity distortion, the same structures of QA-GoP and S-T tube as described in Section V-B are used.

First, as shown in Fig. 7, the spatial horizontal and vertical gradients of each pixel are calculated with the corresponding operators. Let $\vec{\nabla} I_{x,y,n}^{spatial\_h}$ and $\vec{\nabla} I_{x,y,n}^{spatial\_v}$ denote the horizontal and vertical gradient vectors of pixel $I(x, y)$ at frame n. The spatial gradient value of the pixel can be computed as

$$\nabla I_{x,y,n}^{spatial} = \sqrt{\left|\vec{\nabla} I_{x,y,n}^{spatial\_h}\right|^2 + \left|\vec{\nabla} I_{x,y,n}^{spatial\_v}\right|^2}. \tag{13}$$

The left-top pixel position of a S-T tube in the central frame i is denoted as (x_i, y_i). The corresponding pixel coordinates along the motion trajectory in the S-T tube are $[(x_{i-N}, y_{i-N}), \ldots, (x_i, y_i), \ldots, (x_{i+N}, y_{i+N})]$. For each S-T tube, the mean value $\overline{\nabla I_{tube}^{spatial}}$ and the standard deviation σ_{tube} of the spatial gradient are calculated along the motion trajectory as

$$\overline{\nabla I_{tube}^{spatial}} = \frac{\sum_{n=i-N}^{i+N}\sum_{y=y_n}^{y_n+h}\sum_{x=x_n}^{x_n+w}\nabla I_{x,y,n}^{spatial}}{w \times h \times (2N+1)}, \tag{14}$$

$$\sigma_{tube} = \sqrt{\frac{\sum_{n=i-N}^{i+N}\sum_{y=y_n}^{y_n+h}\sum_{x=x_n}^{x_n+w}(\nabla I_{x,y,n}^{spatial} - \overline{\nabla I_{tube}^{spatial}})^2}{w \times h \times (2N+1)}}, \tag{15}$$

where w and h are both 8, and $2N+1$ is equal to the QA-GoP length. The spatio-temporal activity of the S-T tube Γ_{tube} is defined as

$$\Gamma_{tube} = \begin{cases} \sigma_{tube}, & \text{if } \sigma_{tube} > \varepsilon \\ \varepsilon, & \text{otherwise,} \end{cases} \tag{16}$$

where ε is a perceptual threshold. The value of spatio-temporal activity Γ_{tube} is clipped by the threshold ε to avoid measuring imperceptible pixel gradient distortion.

The distortion of spatio-temporal activity in S-T tube is computed as

$$DA^{tube} = \left|log_{10}\left(\frac{\widetilde{\Gamma}_{tube}}{\Gamma_{tube}}\right)\right|, \tag{17}$$

where Γ_{tube} and $\widetilde{\Gamma}_{tube}$ are the activities of the S-T tube in original and synthesized views, respectively. The blurring distortion will lead to a smaller $\widetilde{\Gamma}_{tube}$, and the blocking artifacts and synthesis noise will result in a bigger $\widetilde{\Gamma}_{tube}$.

Similar to the spatial pooling procedure in Section V-C, the distortion of spatio-temporal activity of a QA-GoP can also be obtained with a worst-case method as

$$DA^{GoP} = \frac{1}{N_{\mathbb{W}}}\sum_{k\in\mathbb{W}} DA_k^{tube}, \tag{18}$$

where $\mathbb{W}$ denotes the set with the worst $W\%$ DA^{tube} in the QA-GoP, which is similar to (11). However, for activity distortion measurement, W is set to five. This is because the activity distortions always take over larger area than flicker distortion. Finally, the activity distortion of the sequence DA^{Seq} is computed by taking average of the DA^{GoP} as

$$DA^{Seq} = \frac{1}{K}\sum_{m=0}^{K} DA_m^{GoP}. \tag{19}$$

1	1	0	-1	-1
3	3	0	-3	-3
8	8	0	-8	-8
3	3	0	-3	-3
1	1	0	-1	-1

(a)

1	3	8	3	1
1	3	8	3	1
0	0	0	0	0
-1	-3	-8	-3	-1
-1	-3	-8	-3	-1

(b)

Fig. 7. Operators for computing gradient. (a) Horizontal. (b) Vertical.

E. Integration for Overall Distortion

After obtaining the flicker distortion DF and the activity distortion DA, how to integrate them together to measure the overall distortion D of the synthesized video remains a problem. A linear weighted integration strategy may not be suitable, because selecting the weighting factor used to adjust the relative importance of the two components is not easy since the distortion degree of the texture and depth video are varied. In order to make the DF and DA response equally to the overall distortion D in the synthesized video, the simple multiplication strategy is used as

$$D = DA \times log_{10}(1 + DF). \tag{20}$$

To make the relationship between the DF and the subjective scores more linearly, the logarithm process is used for scaling DF as $log_{10}(1 + DF)$.

F. Determination of the Parameter Values

There are still three parameters in our algorithms to be determined, i.e., N, μ and ε, where $2N + 1$ is the QA-GoP length, μ and ε are perception thresholds used in DF and DA, respectively.

For the QA-GoP length $2N + 1$, the selection criterion is to make the distortion computation period no more than the human fixation duration, which is about 200-400 ms according to the average duration of visual fixation [55], [56]. We selected N to be 2, i.e., the QA-GoP length is 5. As long as the video frame rate is higher than 12.5 fps, the QA-GoP length will not exceed the fixation duration.

The perception threshold μ in (8) is computed using an edge emphasized Just Noticeable Difference (JND) model. The classic JND model in pixel domain [57] does not pay much attention to distinguishing the edge masking and the texture masking [58]. However, for synthesized view, the annoying temporal flicker distortions mostly locate around the object edge. In this paper, the JND threshold in edge region is reduced with a simple and effective texture/edge distinguish operation. First, an edge map is generated using a Canny operator [59]. Then the map is divided into 8×8 blocks. If the number of edge pixels in a block is more than 48, these edge pixels will be marked as texture pixels. The JND value of the remaining edge pixels in the edge map will be multiplied by 0.1 to fit the facts that HVS are sensitive to the edge error in synthesized view. Note that the JND threshold and the edge map are all computed on the synthesized video.

The perception threshold ε in (16) is used to compute the activity distortion index DA. Actually, DA is mainly related to the texture video compression induced distortion such as blurring and blockiness. To obtain the optimal value of ε, we use the video compression distortion subsets of LIVE database [21] to train the threshold ε. The used subsets contain two types of video compression distortions, i.e., MPEG-2 and H.264 compression, and there are 80 videos with different compression degree. The pearson linear correlation coefficient between the original DMOS value and the value predicted by DA index as a function of ε achieves the highest point when ε is 180 (correlation coefficient is 0.741). This value is adopted in our algorithm.

VI. Experimental Results

A. Performance Comparison

The performance of the proposed objective VQA for synthesized video is evaluated on the SIAT Synthesized Video Quality Database. The following VQA algorithms are also tested for comparison.

- *PSNR*: The simple pixel-based objective quality metric used as a baseline for evaluation of the objective VQA.
- *WSNR*: The objective IQA method using contrast sensitivity function as the weighting function [60].
- *SSIM*: The objective IQA method computing the structure similarity between images [28].
- *MS-SSIM*: The objective IQA method computing the multi-scale structural similarity between images [61].
- *VQM*: The objective VQA method proposed by the National Telecommunications and Information Administration (NTIA) [54], which has been adopted by the American National Standards Institute (ANSI) and ITU-T J.144. The software is available for download at [62].
- *MOVIE*: The motion-based objective VQA method developed by LIVE [31]. The software is available for download at [63].
- *Spatial MOVIE*: The spatial version of MOVIE, which primarily focuses on the spatial distortion.
- *Temporal MOVIE*: The temporal version of MOVIE, which primarily focuses on the temporal distortion.
- *BoscPCS*: The objective IQA method for synthesized image proposed by Bosc *et al.* in [36].
- *PSPTNR*: The objective VQA method for synthesized video proposed in [34].

Only the luminance component is used to compute the quality score of each VQA method except for VQM. The scores of PSNR, WSNR, SSIM, MS-SSIM and BoscPCS for the whole sequence are obtained by averaging the scores of each frame.

Three criteria recommended by [20] are used to evaluate the performances of the objective VQA:

- *Pearson Linear Correlation Coefficient (r)*: Measure the prediction accuracy.
- *Spearman Rank Order Correlation Coefficient (ρ)*: Measure the prediction monotonicity.
- *Root Mean Square Error (RMSE)*: Measure the prediction residuals.

Before the performance evaluation, a non-linear logistic function suggested by [20] is used to transform the quality score Q obtained from each VQA to the predicted subjective score DMOS_p, which is defined as

$$DMOS_p = \frac{\beta_1}{1 + e^{-\beta_2 \times (Q - \beta_3)}}. \tag{21}$$

The r, ρ and RMSE are computed between the actual DMOS values and the predicted DMOS_p values.

The scatter plots of DMOS versus DMOS_p for each VQA algorithm are shown in Fig.8. The DMOS_p of MS-SSIM and

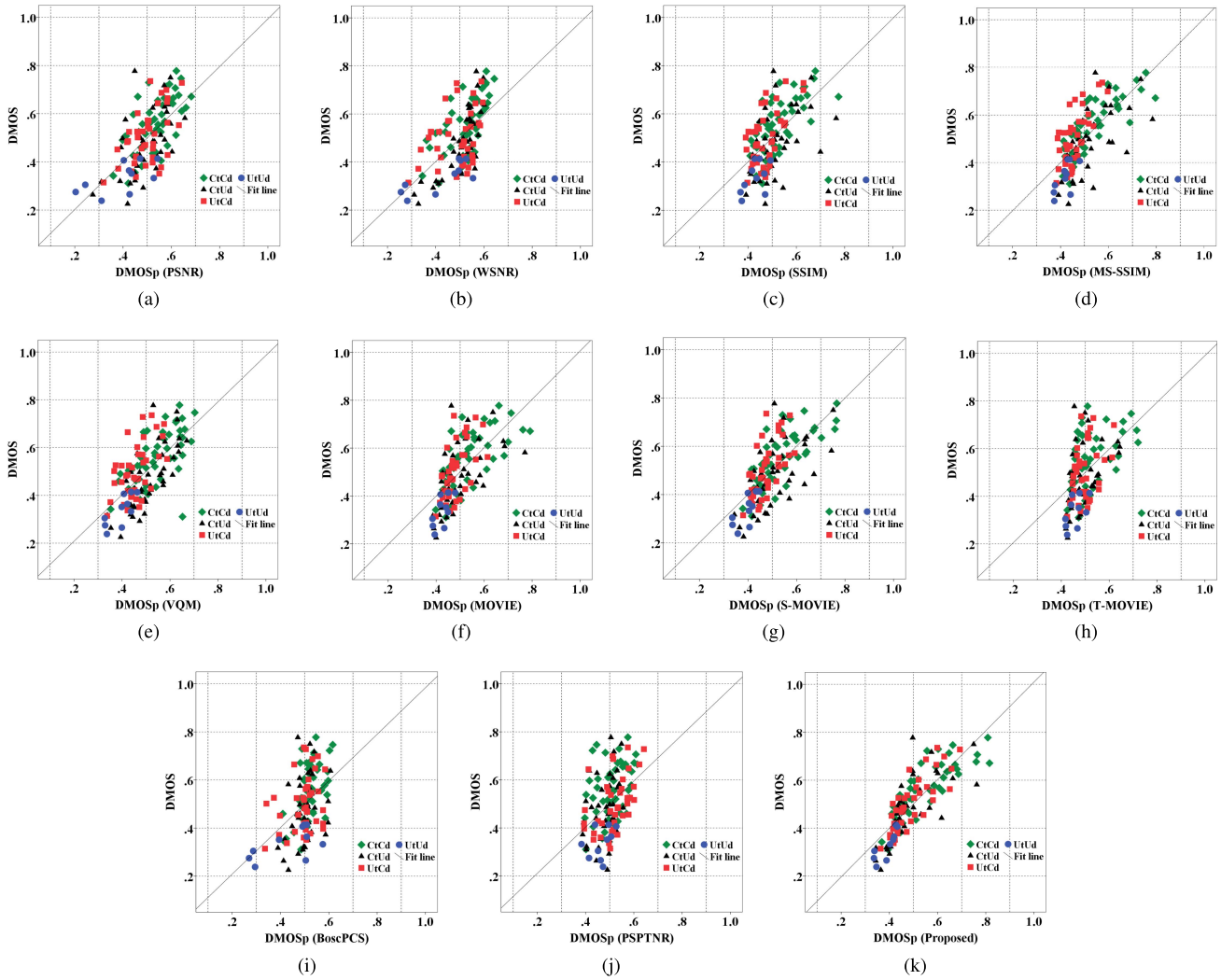

Fig. 8. DMOS versus DMOSp for different VQA models. The results in different subsets are marked with symbols of different shapes. (a) PSNR: $r^2 = 0.42$. (b) WSNR: $r^2 = 0.366$. (c) SSIM: $r^2 = 0.37$. (d) MS-SSIM: $r^2 = 0.495$. (e) VQM: $r^2 = 0.448$. (f) MOVIE: $r^2 = 0.418$. (g) S-MOVIE: $r^2 = 0.487$. (h) T-MOVIE: $r^2 = 0.237$. (i) BoscPCS: $r^2 = 0.205$. (j) PSPTNR: $r^2 = 0.187$. (k) Proposed: $r^2 = 0.665$.

Spatial MOVIE correlated well with the DMOS on the entire data set ($r^2 = 0.495$ and 0.487, respectively), and the proposed method has the highest linear correlation ($r^2 = 0.665$).

The results of r, ρ and RMSE are computed for the $U_T C_D$, $C_T U_D$, $C_T C_D$ subsets and the entire ALL DATA set, respectively, as shown in Table III. Note that the $U_T U_D$ subset which only contains ten synthesized videos is included in the ALL DATA set. Among all the comparison VQA methods, the MS-SSIM has the most satisfactory performance for each subset and the entire data set. This result reveals that the spatial distortion of the synthesized view can be effectively measured by the spatial features based 2D VQA method, although new distortion types, such as geometric distortions, have been induced. Besides MS-SSIM, the WSNR performs best for the $C_T U_D$ subset in terms of r. This is because the distortion type in $C_T U_D$ subset is very close to the conventional 2D compression distortion, such as blurring. The performance of Temporal MOVIE is not satisfying, which proves that the temporal distortion in synthesized video is quite different from that in traditional 2D sequences. The temporal flicker distortion is the most significant difference between the traditional 2D video and the synthesized video.

The proposed method has the best performance for the $U_T C_D$, $C_T C_D$ subsets and the ALL DATA set in terms of r, ρ and RMSE, and also performs best at the $C_T U_D$ subset in terms of ρ and RMSE. As seen in Fig. 8k, there are several outlier points which belong to the $C_T U_D$ subset. Two obvious outlier points are from *Dancer* sequence with QP pair of (40,0) and (44,0). For this sequence, even the texture video was seriously compressed, the subjects were still satisfied with the quality of the video. This may be explained by the significant luminance masking effect in *Dancer* sequence, e.g., the large area of the bright white wall. It is worth mentioning that, for the $U_T C_D$ subset, the values of r and ρ of the proposed method has an obvious advantage compared with the other methods. This is because our algorithm primarily focus on the flicker distortion caused by the depth distortion and the view synthesis process, which has not been considered in the conventional quality assessment metrics.

TABLE III
PERFORMANCE COMPARISON OF OBJECTIVE VQA. THE BEST RESULTS IN EACH COLUMN ARE MARKED IN BOLD

VQA	U_TC_D			C_TU_D			C_TC_D			ALL DATA		
	r	ρ	RMSE	r	ρ	RMSE	r	ρ	RMSE	r	ρ	RMSE
PSNR	0.544	0.481	0.093	0.570	0.566	0.111	0.659	0.666	0.088	0.648	0.627	0.097
WSNR [60]	0.311	0.295	0.114	**0.773**	0.778	0.095	0.609	0.645	0.093	0.605	0.589	0.102
SSIM [28]	0.576	0.465	0.095	0.499	0.534	0.122	0.686	0.704	0.086	0.608	0.598	0.101
MS-SSIM [61]	0.695	0.626	0.094	0.660	0.718	0.109	0.803	0.849	0.070	0.703	0.731	0.091
VQM [54]	0.548	0.526	0.102	0.751	0.755	0.100	0.630	0.643	0.089	0.669	0.655	0.095
MOVIE [31]	0.585	0.573	0.093	0.590	0.649	0.110	0.657	0.713	0.089	0.646	0.693	0.097
S-MOVIE [31]	0.643	0.653	0.090	0.632	0.664	0.109	0.726	0.759	0.080	0.698	0.731	0.091
T-MOVIE [31]	0.375	0.346	0.102	0.461	0.536	0.119	0.456	0.498	0.108	0.486	0.487	0.112
BoscPCS [36]	0.295	0.292	0.108	0.334	0.376	0.128	0.510	0.489	0.103	0.453	0.431	0.114
PSPTNR [34]	0.486	0.527	0.098	0.340	0.391	0.124	0.352	0.338	0.086	0.433	0.453	0.115
Proposed	**0.815**	**0.824**	**0.065**	0.732	**0.838**	**0.090**	**0.827**	**0.863**	**0.067**	**0.815**	**0.869**	**0.074**

TABLE IV
VARIANCE OF THE RESIDUALS BETWEEN INDIVIDUAL SUBJECTIVE SCORES AND VQA $DMOS_p$. THE BEST RESULTS IN EACH COLUMN ARE MARKED IN BOLD

Residual Variance	U_TC_D	C_TU_D	C_TC_D	ALL DATA
σ^2 (optimal model)	0.01169	0.01161	0.00975	0.01088
$\widetilde{\sigma}^2$ (PSNR)	0.02032	0.02346	0.01689	0.02040
$\widetilde{\sigma}^2$ (WSNR)	0.02453	0.01970	0.01769	0.02130
$\widetilde{\sigma}^2$ (SSIM)	0.01989	0.02503	0.01646	0.02123
$\widetilde{\sigma}^2$ (MS-SSIM)	0.01871	0.02151	0.01429	0.01917
$\widetilde{\sigma}^2$ (VQM)	0.02020	**0.01962**	0.01769	0.01996
$\widetilde{\sigma}^2$ (MOVIE)	0.01998	0.02310	0.01730	0.02045
$\widetilde{\sigma}^2$ (S-MOVIE)	0.01918	0.02228	0.01587	0.01931
$\widetilde{\sigma}^2$ (T-MOVIE)	0.02217	0.02547	0.02040	0.02346
$\widetilde{\sigma}^2$ (BoscPCS)	0.02346	0.02723	0.01923	0.02393
$\widetilde{\sigma}^2$ (PSPTNR)	0.02107	0.02716	0.02116	0.02427
$\widetilde{\sigma}^2$ (Proposed)	**0.01590**	0.01972	**0.01411**	**0.01638**
Number of samples	1572	1572	1979	5516
Threshold F-ratio	1.0865	1.0865	1.0767	1.0452

TABLE V
VARIANCE OF THE RESIDUALS BETWEEN SUBJECTIVE DMOS AND VQA $DMOS_p$. THE BEST RESULTS IN EACH COLUMN ARE MARKED IN BOLD

Residual Variance	U_TC_D	C_TU_D	C_TC_D	ALL DATA
$\widehat{\sigma}^2$ (PSNR)	0.00887	0.01219	0.00729	0.00961
$\widehat{\sigma}^2$ (WSNR)	0.01319	0.00830	0.00811	0.01051
$\widehat{\sigma}^2$ (SSIM)	0.00840	0.01382	0.00682	0.01044
$\widehat{\sigma}^2$ (MS-SSIM)	0.00721	0.01020	0.00462	0.00837
$\widehat{\sigma}^2$ (VQM)	0.00875	**0.00822**	0.00807	0.00915
$\widehat{\sigma}^2$ (MOVIE)	0.00851	0.01181	0.00767	0.00965
$\widehat{\sigma}^2$ (S-MOVIE)	0.00768	0.01137	0.00623	0.00850
$\widehat{\sigma}^2$ (T-MOVIE)	0.01075	0.01420	0.01081	0.01265
$\widehat{\sigma}^2$ (BoscPCS)	0.01213	0.01602	0.00964	0.01317
$\widehat{\sigma}^2$ (PSPTNR)	0.00961	0.01594	0.01161	0.01346
$\widehat{\sigma}^2$ (Proposed)	**0.00430**	0.00836	**0.00444**	**0.00555**
Number of samples	40	40	50	140
Threshold F-ratio	1.6927	1.6927	1.5994	1.3217

B. Statistical Significance Test

Two F-test procedures are used to examine the significance of difference between each objective VQA, which have been suggested for significance testing by the authors of the VQEG database [20] and the LIVE database [21].

The first F-test is to compare the performance of each objective VQA with the theoretical optimal model. For a subject i and a test sequence j, the individual score $z_{scaled}(i, j)$ and the average score $DMOS(j)$ for sequence j can be obtained as described in Section IV-C. The residual R between each individual score and the corresponding $DMOS(j)$ is computed as

$$R(i, j) = z_{scaled}(i, j) - DMOS(j). \tag{22}$$

Similarly, the residual between each individual score and the VQA predicted score $DMOS_p(j)$ can be computed as

$$\widetilde{R}(i, j) = z_{scaled}(i, j) - DMOS_p(j). \tag{23}$$

Let σ^2 and $\widetilde{\sigma}^2$ denote the variance of the $R(i, j)$ and $\widetilde{R}(i, j)$, respectively. The smaller variance means the better. The performance of the test VQA is determined to be significantly equivalent to the theoretical optimal upper bound if the ratio of $\widetilde{\sigma}^2$ and σ^2 is smaller than the F-ratio threshold as

$$\frac{\sigma^2(\text{VQA})}{\widetilde{\sigma}^2(\text{optimal model})} < \text{F-ratio}. \tag{24}$$

The F-ratio threshold can be obtained with look-up table. The value is related to the sample size and the significance level (95% in this paper). The variance of the theoretical optimal

TABLE VI

RESULTS OF THE F-TEST ON THE RESIDUALS BETWEEN $DMOS_p$ AND DMOS. FOUR SYMBOLS IN EACH ENTRY OF THE TABLE CORRESPOND TO "U_TC_D", "C_TU_D", "C_TC_D" AND "ALL DATA" DATA SET IN ORDER. THE VALUE "1" INDICATES THE VQA IN THE ROW IS SIGNIFICANTLY SUPERIOR TO THAT IN THE COLUMN. THE VALUE "0" INDICATES THE VQA IN THE ROW IS SIGNIFICANTLY INFERIOR TO THAT IN THE COLUMN. THE SYMBOL "−" INDICATES THE TWO VQA ARE SIGNIFICANTLY EQUIVALENT

VQA	PSNR	WSNR	SSIM	MSSSIM	VQM	MOVIE	SMOVIE	TMOVIE	BoscPCS	PSPTNR	Proposed
PSNR	−−−−	−−−−	−−−−	−−−−	−−−−	−−−−	−−−−	−−−−	−−−1	−−−1	0−00
WSNR	−−−−	−−−−	−−−−	0−0−	−−−−	−−−−	0−0−	−1−−	−1−−	−1−−	0−00
SSIM	−−−−	−−−−	−−−−	−−−−	−−−−	−−−−	−−−−	−−−−	−−−−	−−1−	0−−0
MSSSIM	−−−−	1−1−	−−−−	−−−−	−−1−	−−1−	−−−−	11−−	−−11	−−11	−−−0
VQM	−−−−	−−−−	−−−−	−−0−	−−−−	−−−−	−−−−	−1−1	−1−1	−1−1	0−00
MOVIE	−−−−	−−−−	−−−−	−−0−	−−−−	−−−−	−−−−	−−−−	−−−1	−−−1	0−00
SMOVIE	−−−−	1−1−	−−−−	−−−−	−−−−	−−−−	−−−−	−−11	−−−1	−−11	0−−0
TMOVIE	−−−−	−0−−	−−−−	−−00	−0−0	−−−−	−−00	−−−−	−−−−	−−−−	0000
BoscPCS	−−−0	−0−−	−−−−	−−00	−0−0	−−−0	−−−0	−−−−	−−−−	−−−−	0000
PSPTNR	−−−0	−0−−	−−0−	−−00	−0−0	−−−0	−−00	−−−−	−−−−	−−−−	0000
Proposed	1−11	1−11	1−−1	−−−1	1−11	1−11	1−−1	1111	1111	1111	−−−−

model and each test VQA model are presented in Table IV. It can be observed that none of the VQA can be significantly equivalent to the theoretical optimal model. VQM performs best for the C_TU_D subset. The proposed model has the best performance for the ALL DATA set and the U_TC_D, C_TC_D subsets, respectively.

The second F-test is based on the residual difference $\widehat{R}(j)$ between the predicted score $DMOS_p(j)$ and the actual score $DMOS(j)$, which can be computed as

$$\widehat{R}(j) = DMOS_p(j) - DMOS(j). \tag{25}$$

Let $\widehat{\sigma}^2$ denotes the variance of the residual $\widehat{R}(j)$ of each VQA, as it is shown in Table V. A smaller variance means a better performance. To test the significance difference between each VQA, the ratio between the variance $\widehat{\sigma}^2$ of different VQA is computed. For example, VQA_i is determined to be significant superior to VQA_j if their ratio is greater than the F-ratio threshold as

$$\frac{\widehat{\sigma}^2(\text{VQA}_j)}{\widehat{\sigma}^2(\text{VQA}_i)} > \text{F-ratio}. \tag{26}$$

The results of the significance test is presented in Table VI. It can be seen that the performance of the proposed VQA is significantly superior to the others for the ALL DATA set, and is also superior to the others except for MS-SSIM for the U_TC_D subset. For the C_TC_D subset, the proposed VQA is significantly superior to the others except for SSIM, MS-SSIM and Spatial MOVIE. However, for the C_TU_D subset, none of the VQA algorithms has the absolute advantage. There is still large development space to improve the performance of the objective VQA for synthesized video.

VII. CONCLUSIONS AND FUTURE WORKS

The studies of subjective and objective video quality assessment for synthesized views with texture/depth compression distortion are presented. The criterion for the database design is to ensure the discrimination and the wide range of different quality level between each synthesized sequences. To achieve this aim, the quantization level of each texture/depth view pairs are selected manually from a large number of candidates. We believe this is more suitable for evaluating the performance of different objective VQA methods compared with generating test sequences using fixed coding bit rate. A full reference objective VQA algorithm for synthesized video has also been introduced. The proposed method primarily focuses on the temporal flicker distortion due to depth compression distortion and the view synthesis process. The performance of the proposed algorithm is evaluated on the synthesized video quality database. The experimental results show that our VQA method has a good performance on the entire database compared with the state of the art objective VQA methods, and is particularly prominent on the subsets which has significant temporal flicker distortion caused by depth compression and view synthesis process. Our future works will concern on supplementing the synthesized video quality database with other distortion sources, such as video transmission distortion.

REFERENCES

[1] K. Muller, P. Merkle, and T. Wiegand, "3D video representation using depth maps," *Proc. IEEE*, vol. 99, no. 4, pp. 643–656, Apr. 2011.

[2] M. M. Hannuksela, Y. Chen, T. Suzuki, J. Ohm, and G. Sullivan, *3D-AVC Draft Text 7*, Joint Collaborative Team 3D Video Coding Extensions (JCT-3V), document JCT3V-E1002V2, Jan. 2013.

[3] G. Tech, K. Wegner, Y. Chen, and S. Yea, *3D-HEVC Draft Text 5*, Joint Collaborative Team 3D Video Coding Extensions (JCT-3V), document JCT3V-I1001V3, Jul. 2014.

[4] C. Fehn, R. Barre, and S. Pastoor, "Interactive 3-DTV-concepts and key technologies," *Proc. IEEE*, vol. 94, no. 3, pp. 524–538, Mar. 2006.

[5] *Subjective Methods Assessment Stereoscopic 3DTV Systems*, document ITU-R Rec. BT.2021, 2012.

[6] L.-H. Wang, X.-J. Huang, M. Xi, D.-X. Li, and M. Zhang, "An asymmetric edge adaptive filter for depth generation and hole filling in 3DTV," *IEEE Trans. Broadcast.*, vol. 56, no. 3, pp. 425–431, Sep. 2010.

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

[8] Y. Zhao, C. Zhu, Z. Chen, and L. Yu, "Depth no-synthesis-error model for view synthesis in 3-D video," *IEEE Trans. Image Process.*, vol. 20, no. 8, pp. 2221–2228, Aug. 2011.

[9] 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.

[10] 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.

[11] S. Hu, S. Kwong, Y. Zhang, and C.-C. J. Kuo, "Rate-distortion optimized rate control for depth map-based 3D video coding," *IEEE Trans. Image Process.*, vol. 22, no. 2, pp. 585–594, Feb. 2013.

[12] H. Yuan, S. Kwong, J. Liu, and J. Sun, "A novel distortion model and lagrangian multiplier for depth maps coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 3, pp. 443–451, Mar. 2014.

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

[14] T.-Y. Chung, J.-Y. Sim, and C.-S. Kim, "Bit allocation algorithm with novel view synthesis distortion model for multiview video plus depth coding," *IEEE Trans. Image Process.*, vol. 23, no. 8, pp. 3254–3267, Aug. 2014.

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

[16] K. Muller *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.

[17] *Methodology for the Subjective Assessment of the Quality of Television Pictures*, document ITU-R Rec. BT.500, 2002.

[18] *Subjective Video Quality Assessment Methods for Multimedia Applications*, document ITU-T Rec. P.910, 1999.

[19] S. Chikkerur, V. Sundaram, M. Reisslein, and L. J. Karam, "Objective video quality assessment methods: A classification, review, and performance comparison," *IEEE Trans. Broadcast.*, vol. 57, no. 2, pp. 165–182, Jun. 2011.

[20] *VQEG Final Report of FR-TV Phase II Validation*. [Online]. Available: http://www.itu.int/ITU-T/studygroups/com09/docs/tutorial_opavc.pdf, accessed May 2005.

[21] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, "Study of subjective and objective quality assessment of video," *IEEE Trans. Image Process.*, vol. 19, no. 6, pp. 1427–1441, Jun. 2010.

[22] P. Lebreton, A. Raake, M. Barkowsky, and P. Le Callet, "Evaluating depth perception of 3D stereoscopic videos," *IEEE J. Sel. Topics Signal Process.*, vol. 6, no. 6, pp. 710–720, Oct. 2012.

[23] T. Kim, J. Kang, S. Lee, and A. C. Bovik, "Multimodal interactive continuous scoring of subjective 3D video quality of experience," *IEEE Trans. Multimedia*, vol. 16, no. 2, pp. 387–402, Feb. 2014.

[24] (2011). *IRCCyN/IVC DIBR Videos Database*. [Online]. Available: http://ivc.univ-nantes.fr/en/databases/DIBR_Videos/

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

[26] E. Bosc, P. Hanhart, P. Le Callet, and T. Ebrahimi, "A quality assessment protocol for free-viewpoint video sequences synthesized from decompressed depth data," in *Proc. 5th Int. Workshop Quality Multimedia Exper. (QoMEX)*, Jul. 2013, pp. 100–105.

[27] *SIAT Synthesized Video Quality Database*. [Online]. Available: http://codec.siat.ac.cn/SIAT_Synthesized_Video_Quality_Database/index.html, accessed Aug. 2015.

[28] 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.

[29] F. Shao, W. Lin, S. Gu, G. Jiang, and T. Srikanthan, "Perceptual full-reference quality assessment of stereoscopic images by considering binocular visual characteristics," *IEEE Trans. Image Process.*, vol. 22, no. 5, pp. 1940–1953, May 2013.

[30] Y.-H. Lin and J.-L. Wu, "Quality assessment of stereoscopic 3D image compression by binocular integration behaviors," *IEEE Trans. Image Process.*, vol. 23, no. 4, pp. 1527–1542, Apr. 2014.

[31] 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.

[32] 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.

[33] D. Rusanovskyy, F.-C. Chen, L. Zhang, and T. Suzuki, *3D-AVC Test Model 8*, Joint Collaborative Team 3D Video Coding Extensions (JCT-3V), document JCT3V-F1003, Nov. 2013.

[34] Y. Zhao and L. Yu, "A perceptual metric for evaluating quality of synthesized sequences in 3DV system," *Proc. SPIE*, vol. 7744, p. 77440X, Aug. 2010.

[35] E. Ekmekcioglu, S. Worrall, D. De Silva, A. Fernando, and A. Kondoz, "Depth based perceptual quality assessment for synthesised camera viewpoints," in *User Centric Media* (Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering), vol. 60. Berlin, Germany: Springer-Verlag, 2012, pp. 76–83.

[36] E. Bosc, P. Le Callet, L. Morin, and M. Pressigout, "An edge-based structural distortion indicator for the quality assessment of 3D synthesized views," in *Proc. Picture Coding Symp. (PCS)*, May 2012, pp. 249–252.

[37] E. Bosc, F. Battisti, M. Carli, and P. Le Callet, "A wavelet-based image quality metric for the assessment of 3D synthesized views," *Proc. SPIE*, vol. 8648, p. 86481Z, Mar. 2013.

[38] F. Battisti, E. Bosc, M. Carli, P. Le Callet, and S. Perugia, "Objective image quality assessment of 3D synthesized views," *Signal Process., Image Commun.*, vol. 30, pp. 78–88, Jan. 2015.

[39] C.-T. Tsai and H.-M. Hang, "Quality assessment of 3D synthesized views with depth map distortion," in *Proc. Vis. Commun. Image Process. (VCIP)*, Nov. 2013, pp. 1–6.

[40] *H.264/AVC Reference Software JM 18*. [Online]. Available: http://iphome.hhi.de/suehring/tml/download/, accessed May 2015.

[41] Y. Jia, W. Lin, and A. A. Kassim, "Estimating just-noticeable distortion for video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 7, pp. 820–829, Jul. 2006.

[42] A. Liu, W. Lin, and M. Narwaria, "Image quality assessment based on gradient similarity," *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 1500–1512, Apr. 2012.

[43] *Draft Call for Proposals on 3D Video Coding Technology*, ISO/IEC JTC1/SC29/WG11, document MPEG2011/N11830, Jan. 2011.

[44] D. Rusanovskyy, K. Müller, and A. Vetro, *Common Test Conditions of 3DV Core Experiments*, Joint Collaborative Team 3D Video Coding Extensions (JCT-3V), document JCT3V-E1100, Aug. 2013.

[45] *Reference Software for 3D-AVC: 3DV-ATM V10.0*. [Online]. Available: http://mpeg3dv.nokiaresearch.com/svn/mpeg3dv/tags/, accessed Nov. 2013.

[46] *VSRS-1D-Fast*. [Online]. Available: https://hevc.hhi.fraunhofer.de/svn/svn_3DVCSoftware, accessed Aug. 2015.

[47] *Anchor Software for 3D-HEVC Experiments: 3DV-HTM V8.0*. [Online]. Available: https://hevc.hhi.fraunhofer.de/svn/svn_3DVCSoftware, accessed Aug. 2015.

[48] *MSU Perceptual Video Quality Tool*. [Online]. Available: http://www.compression.ru/video/quality_measure/perceptual_video_quality_tool_en.html, accessed Aug. 2015.

[49] R. Miller, *Beyond ANOVA: Basics of Applied Statistics* (Texts in Statistical Science Series). London, U.K.: Chapman & Hall, 1997.

[50] 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.

[51] R. Li, B. Zeng, and M. L. Liou, "A new three-step search algorithm for block motion estimation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 4, no. 4, pp. 438–442, Aug. 1994.

[52] Y. Su, M.-T. Sun, and V. Hsu, "Global motion estimation from coarsely sampled motion vector field and the applications," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 2, pp. 232–242, Feb. 2005.

[53] A. K. Moorthy and A. C. Bovik, "Visual importance pooling for image quality assessment," *IEEE J. Sel. Topics Signal Process.*, vol. 3, no. 2, pp. 193–201, Apr. 2009.

[54] 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.

[55] A. Ninassi, O. Le Meur, P. Le Callet, and D. Barba, "Considering temporal variations of spatial visual distortions in video quality assessment," *IEEE J. Sel. Topics Signal Process.*, vol. 3, no. 2, pp. 253–265, Apr. 2009.

[56] S. L. Cloherty, M. J. Mustari, M. G. Rosa, and M. R. Ibbotson, "Effects of saccades on visual processing in primate MSTd," *Vis. Res.*, vol. 50, no. 24, pp. 2683–2691, 2010.

[57] Z. Chen and C. Guillemot, "Perceptually-friendly H.264/AVC video coding based on foveated just-noticeable-distortion model," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 6, pp. 806–819, Jun. 2010.
[58] A. Liu, W. Lin, M. Paul, C. Deng, and F. Zhang, "Just noticeable difference for images with decomposition model for separating edge and textured regions," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 11, pp. 1648–1652, Nov. 2010.
[59] J. Canny, "A computational approach to edge detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-8, no. 6, pp. 679–698, Nov. 1986.
[60] N. Damera-Venkata, T. D. Kite, W. S. Geisler, B. L. Evans, and A. C. Bovik, "Image quality assessment based on a degradation model," *IEEE Trans. Image Process.*, vol. 9, no. 4, pp. 636–650, Apr. 2000.
[61] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Proc. Conf. Rec. 37th Asilomar Conf. Signals, Syst. Comput.*, vol. 2. Nov. 2003, pp. 1398–1402.
[62] *Video Quality Metric (VQM) Software*. [Online]. Available: http://www.its.bldrdoc.gov/resources/video-quality-research/software.aspx, accessed Aug. 2015.
[63] *LIVE Software Releases*. [Online]. Available: http://live.ece.utexas.edu/research/quality/, accessed Feb. 2015.

Xiangkai Liu received the B.S. degree in automation from Southwest Jiaotong University, Chengdu, China, in 2009. He is currently pursuing the Ph.D. degree from Southwest Jiaotong University. From 2011 to 2012, he was a Research Associate with the Institute of Digital Media, Peking University, Beijing, China. Since 2014, he has been with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China. His current research interests include 3D video quality assessment and 3D video coding.

Yun Zhang (M'12) 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 Post-Doctoral Researcher with the Department of Computer Science, The City University of Hong Kong, Hong Kong. In 2010, he became an Assistant Professor with the Shenzhen Institutes of Advanced Technology, CAS, where he has served as an Associate Professor since 2012. His research interests are video compression, 3D video processing, and visual perception.

Sudeng Hu received the B.Eng. degree from Zhejiang University, Hangzhou, China, in 2007, and the M.Phil. degree from the Department of Computer Science, The City University of Hong Kong, Hong Kong, in 2010. He is currently pursuing the Ph.D. degree with the Department of Electrical Engineering, University of Southern California, Los Angeles. From 2010 to 2011, he was a Research Associate with the Department of Computer Science, The City University of Hong Kong. In 2012, he took an internship with Mitsubishi Electric Research Laboratories, Cambridge, Massachusetts. His research interests include image and video compression, rate control, scalable video coding, and 3D video coding.

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 Bell Northern Research Canada, as a member of the Scientific Staff. In 1990, he became a Lecturer with the Department of Electronic Engineering, The 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.

C.-C. Jay Kuo (F'99) received the B.S. degree in electrical engineering from the National Taiwan University, Taipei, Taiwan, in 1980, and the M.S. and Ph.D. degrees in electrical engineering from the Massachusetts Institute of Technology, Cambridge, in 1985 and 1987, respectively. He is currently the Director of the Multimedia Communications Laboratory and a Professor of Electrical Engineering, Computer Science and Mathematics with the Ming-Hsieh Department of Electrical Engineering, University of Southern California, Los Angeles. He has co-authored about 200 journal papers, 850 conference papers, and 10 books. His research interests include digital image/video analysis and modeling, multimedia data compression, communication and networking, and biological signal/image processing. He is a fellow of The American Association for the Advancement of Science and The International Society for Optical Engineers.

Qiang Peng received the B.Eng. degree in automation control from Xi'an Jiaotong University, Xi'an, China, in 1984, and the M.Eng. degree in computer application and technology and the Ph.D. degree in traffic information and control engineering from Southwest Jiaotong University, Chengdu, China, in 1987 and 2004, respectively. He is currently a Professor with the School of Information Science and Technology, Southwest Jiaotong University. His research interests include digital video compression and transmission, image/graphics processing, traffic information detection and simulation, virtual reality technology, and multimedia systems and applications.