Prediction of the Influence of Navigation Scan-Path on Perceived Quality of Free-Viewpoint Videos

Suiyi Ling, *Student Member, IEEE*, Jesús Gutiérrez, Ke Gu, *Member, IEEE*, and Patrick Le Callet, *Fellow, IEEE*

***Abstract*— Free-viewpoint video (FVV) systems allow the viewers to freely change the viewpoints of the scene. In such systems, view synthesis and compression are the two main sources of artifacts influencing the perceived quality. To assess this influence, quality evaluation studies are often carried out using conventional displays and generating predefined navigation trajectories mimicking the possible movement of the viewers when exploring the content. Nevertheless, as different trajectories may lead to different conclusions in terms of visual quality when benchmarking the performance of the systems, methods to identify critical trajectories are needed. This paper aims at exploring the impact of exploration trajectories [defined as hypothetical rendering trajectories (HRT)] on the perceived quality of FVV subjectively and objectively, providing two main contributions. First, a subjective assessment test including different HRTs was carried out and analyzed. The results demonstrate and quantify the influence of HRT in the perceived quality. Second, we propose a new full-reference objective video quality assessment measure to objectively predict the impact of HRT. This measure, based on sketch-token representation, models how the categories of the contours change spatially and temporally from a higher semantic level. Performance in comparison with existing quality metrics for the FVV highlight promising results for the automatic detection of most critical HRTs for the benchmark of immersive systems.**

***Index Terms*— Free-viewpoint video, super multi-view, database, view-synthesis, subjective quality evaluation, objective quality metric, mid-level contour descriptor.**

I. Introduction

AS IMMERSIVE multimedia has developed in leaps and bounds along with the emergence of more advanced technologies for capturing, processing and rendering, applications like Free-viewpoint TV (FTV), 3DTV, Virtual Reality (VR) and Augmented Reality (AR) have engaged a lot of users and have become the novel hot topic in the multimedia field. In this sense, systems based on Free-Viewpoint Video (FVV) content, such as FTV [1], allow the users to immerse themselves into a scene by freely switching the viewpoints as they do in the real world. FTV enables Super Multi-View (SMV) and Free Navigation (FN) applications. On one hand, for SMV, a horizontal set of more than 80 views (linearly or angularly arranged) is needed to provide users a 3D viewing experience with wide-viewing horizontal parallax, and smooth transition between adjacent views [2]. On the other hand, for FN, only a limited set of input views is required, coming from sparse camera arrangements in large baseline setup conditions. In both cases, to deal with such a huge amount of data for delivery and storage, efficient compression techniques are essential, together with robust view-synthesis algorithms, such as Depth-Image-Based Rendering (DIBR) technology, which allows reconstructing the FVV content from a limited set of input views. These processes, as any other within the whole multimedia processing chain, can introduce effects that may influence the Quality of Experience (QoE) perceived by the end users. Thus, quality evaluation becomes essential for a successful development of the technology guaranteeing a satisfactory user experience. On one side, subjective assessment tests may help to understand the users' experience with FVV, while on the other side, objective Video Quality Metrics (VQM) may estimate the perceived quality by the users. In this sense, several aspects should be addressed, especially in comparison with other multimedia technologies that do not offer the possibility of exploring the content as desired by the user.

A. Importance and Difficulties of Video Quality Assessment in FVV

While hardware developments are leading the advances for capturing and rendering FVV, compression techniques and view synthesis algorithms are main focus of research, as reflected by the ongoing standardization activities within MPEG [3], [4]. This is mainly due to their importance on the perceived quality, and thus, on the success of the related applications and services [5].

Aside from the well-known compression artifacts, view synthesis techniques (e.g., DIBR) have to deal with disoccluded regions [6]. This is due to the reappearance of the sheltered regions, which are not shown in the reference views but are

Manuscript received July 30, 2018; revised December 2, 2018; accepted December 18, 2018. Date of publication January 21, 2019; date of current version March 11, 2019. The work of J. Gutiérrez was supported by the People Programme (Marie Curie Actions) of the European Union's 7th Framework Programme (FP7/2007-2013) through the PRESTIGE Programme coordinated by Campus France under REA Grant Agreement under Grant PCOFUND-GA-2013-609102. This paper was recommended by Guest Editor Mathias Wien. *(Corresponding author: Suiyi Ling.)*

S. Ling, J. Gutiérrez, and P. Le Callet are with the Équipe Image, Perception et Interaction, Laboratoire des Sciences du Numérique de Nantes, Université de Nantes, 44300 Nantes, France (e-mail: suiyi.ling@univ-nantes.fr; jesús.gutiérrez@univ-nantes.fr; patrick.lecallet@univ-nantes.fr).

K. Gu is with the Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China (e-mail: guke.doctor@gmail.com).

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/JETCAS.2019.2893484

made visible later in the generated ones. Techniques to recover disoccluded regions often introduce geometric distortion and ghosting artifacts. These synthesized-related artifacts are different in nature to compression artifacts, since they mostly appear locally along the disoccluded areas, while compression artifacts are usually spread over the entire video. In addition, view-synthesis artifacts increase with the baseline distance (i.e., number of synthesized views between two real views) until a point which they may be dominant over compression artifacts [4]. Thus, it is very unlikely that VQM proposed for compression-related distortions would be efficient for predicting the quality of sequences produced using synthesized views.

B. Impact of Navigation Scan-Path on Perceived Quality: Free Navigation vs. Predefined Trajectories

Immersive media technologies offer the users more freedom to explore the content allowing more interactive experiences than with traditional media. These new possibilities introduce the observers' behavior as an important factor for the perceived quality [4].

Given the fact that each observer can explore the content differently, two approaches can be adopted to practically study this factor: 1) let the observers navigate the content freely; 2) let the observer watch the sequences in a form of certain pre-defined navigation trajectories. By employing the first approach, one could obtain a common trajectory according to all the observers' data. However, this common trajectory does not necessarily represent the critical one that will stress the system to the worse case. Moreover, if observers are allowed to navigate freely during the test, it will become a new factor that increases the variability of the mean opinion score (MOS), despite observer's variability in forming a quality judgment. As a result, more observers are likely to be required to obtain MOS that can distinguish one system from another with statistical significance. The second approach (predefined trajectories) is not affected by this trajectory-source of variability but comes with the challenge of selecting the 'right' trajectory. In case of system benchmark, one could define 'right' trajectory as the most critical one or the weakest link, e.g. the one leading to the lowest perceived quality. Nevertheless, there is a good chance that this trajectory-effect is highly dependent on content, some being more sensitive than some others to the choice of trajectory. Identifying the impact of navigation trajectory among different viewpoints on perceived quality for a given content is then of particular interest. Thus, it may be useful to know how navigation affects the visual experience and which are the 'worst' trajectories for the system, to carry out quality evaluations of the performance of the system under study in the most stressful cases. Consequently, the availability of computational tools to select the critical trajectories would be extremely useful.

C. Contribution

Based on the discussion above, there are two main research questions in this paper, including 1) what are the impacts of navigation trajectories on perceived quality; 2) if trajectory affects quality, how to develop an objective metric to indicate the 'worst' trajectory. To answer these two questions, the contribution of this paper is twofold. Firstly, a subjective test is conducted to study the impact of the exploration trajectory on perceived quality in FVV application scenarios, containing compression and view-synthesis artifacts. In this sense, the concept of Hypothetical Rendering Trajectory (HRT) is introduced. Also, the annotated database obtained from this test is released for research purposes in the field. Secondly, a full-reference Sketch-Token-based Video Quality Assessment Metric (ST-VQM) is proposed by quantifying to what extent the classes of contours change due to view synthesis. This metric is capable of predicting if sequences based on a given trajectory are of higher/lower quality than sequences based on other trajectories, with respect to subjective scores.

The remainder of the paper is organized as follows. In Section II, an overview of the state-of-the-art in terms of subjective and objective quality evaluation in relation with FVV scenarios is presented and discussed. In Section III, the details of the subjective experiment are described, while Section V introduces the proposed VQA metric based on mid-level descriptor. The experimental results from the subjective experiment and the performance evaluation of the proposed objective metric are presented in Section IV and Section VI. Finally, conclusions are given in Section VII.

II. Related Work

A. Subjective Studies

Although the development of technical aspects related to FTV has been addressed already for some years, the subjective evaluation of the QoE of such systems is still an open issue [7]. As previously mentioned, the majority of the existing studies have been carried out using conventional screens and limiting the interactivity of the users by showing some representative content or predefined trajectories simulating the movement of the observers [8]. In FVV systems, this is especially the case given the limited access to SMV or light-field displays, since only a few prototypes are already available. Nevertheless, it is worth noting the preliminary subjective study that Dricot *et al.* [9] carried out a considering coding and view-synthesis artifacts using a light-field display.

In addition to compression techniques, the evaluation and understanding of view-synthesis algorithms is crucial for the successful development of FTV applications and is still an open issue [4]. In this sense, some works that were carried out with previous technologies (e.g., multi-view video), should be taken into account in the study of the effects of view-synthesis in current FTV applications. Firstly, Bosc *et al.* [10]–[12] carried out subjective studies to evaluate the visual quality of synthesized views using DIBR. In these studies, the quality performance of view synthesis was evaluated through different ways, such as: a) the quality of synthesized still images [10], b) the quality of videos showing a synthesized view of Multi-View plus Depth (MVD) video sequences [11], and c) video sequences showing a smooth sweep across the different viewpoints of a static scene [12]. These different approaches are represented in Fig. 1, showing that the first approach only considers spatial DIBR-related artifacts,

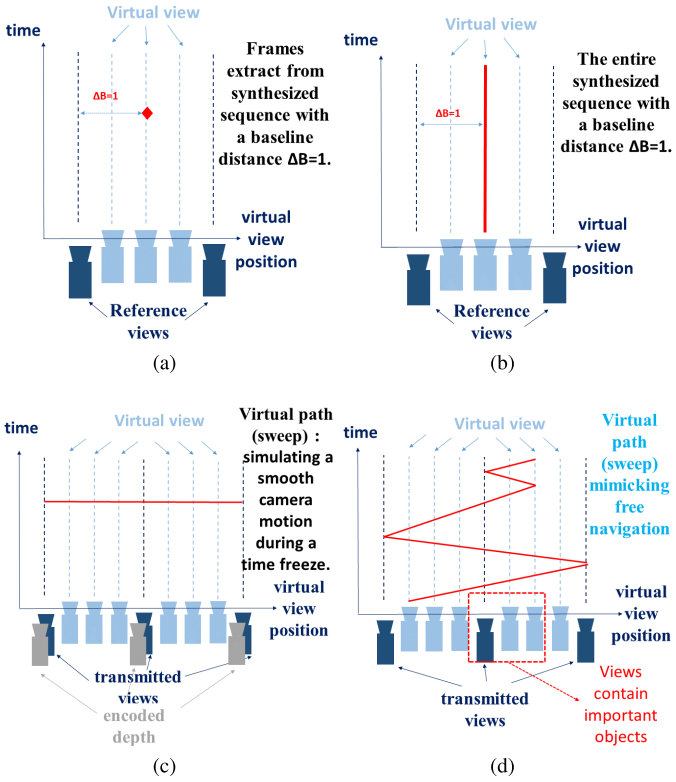

Fig. 1. Different possibilities to evaluate FTV content representing different degrees of navigation. (a) Synthesized image. (b) Video from a synthesized view (exploration along time). (c) Video containing a view sweep (exploration along views). (d) Video containing a view sweep from videos of various synthesized views (exploration along time and views)

the second approach considers also temporal distortions within the synthesized view, and the third approach considers spatial DIBR-related artifacts of all the views. To complete the evaluation, another use case should consider the use of view-sweep over the views in video sequences, as depicted in Fig. 1(d) (i.e., generating videos in which a sweep across the different viewpoints is shown, as if the observer was moving his head horizontally). This approach has been recently adopted in subjective studies with SMV [8], which were carried out to study different aspects of this technology, such as smoothness in view transitions and comfortable view-sweep speed [13], and the impact of coding artifacts [14]. MPEG has adopted this type of alternative for their current standardization activities regarding the evaluation of compression techniques for FTV [3].

Furthermore, as a result from subjective tests, the availability of appropriate datasets is a crucial aspect for the research on both subjective and objective quality. Especially for supporting the development of objective quality metrics, databases containing suitable stimuli (images/videos) annotated with results from subjective tests are essential. Some efforts have been already made to publish datasets containing free-viewpoint content [7] and some results of the aforementioned subjective tests [10]–[12], [15], [16]. Nevertheless, none of these datasets have considered the effect of content adapted trajectories in the "view-sweeping along time" scenario.

B. Objective Metrics

Some image quality metrics have been recently proposed especially designed to handle view-synthesis artifacts. For instance, Battisti *et al.* [17] proposed a metric based on statistical features of wavelet sub-bands. Furthermore, considering that using multi-resolution approaches could increase the performance of image quality metrics, Sandić-Stanković *et al.* [18], [19] proposed to use morphological wavelet decomposition, and multi-scale decomposition based on morphological pyramids. Later, the reduced version of these two metrics was presented in [20] claiming that PSNR is more consistent with human judgment when calculated at higher morphological decomposition scales.

All the aforementioned metrics are limited to quality assessment of synthesized static images, so they do not explicitly consider temporal distortions that may appear in videos containing synthesized views. Thus, some ad-hoc video metrics have been proposed. For instance, Zhao and Yu [21] proposed a measure which calculates temporal artifacts that can be perceived by observers in the background regions of the synthesized videos. Similarly, Ekmekcioglu *et al.* [22] proposed a video quality measure using depth and motion information to take into account where the degradations are located. Moreover, another video metric was recently introduced by Liu *et al.* [15] considering the spatio-temporal activity and the temporal flickering that appears in synthesized video sequences. However, the aforementioned video quality measures are able to predict the impact of view-synthesis degradations comparing videos corresponding with one single view (as represented in Fig. 1(b)). In other words, switching among views (resulting from the possible movement of the viewers) and related effects (e.g., inconsistencies among views, geometric flicker along time and view dimensions, etc. [7]) are not addressed. Hanhart *et al.* [5] evaluated the performance of state-of-the-art quality measures for 2D video in sequences generated by view-sweep [12] (as depicted in Fig. 1(c)), thus considering view-point changes, and reported low performance of all measures in predicting the perceptual quality. Therefore, an efficient objective video quality measure able to deal with the "view-sweeping along time" scenario is still needed.

III. Subjective Study of the Impact of Trajectory on Perceived Quality

As described in the introduction, the first research question of this paper is to identify the impact of the navigation trajectory among different viewpoints on the perceived quality taking contents into account. To this end, a subjective study was conducted by designing content related trajectories. A video quality database for FVV scenarios was built, including both compression and view-synthesis artifacts. It contains the scores from the subjective assessment test described in the following subsections. The videos in this database are generated by simulating exploring trajectories that the observers may use in real scenarios, which are set by the Hypothetical Rendering Trajectory (HRT), defined in the following subsection. This

database is named as 'Image, Perception and Interaction group Free-viewpoint Video Database' (IPI-FVV).[1]

A. Hypothetical Rendering Trajectory

A commonly used naming convention for subjective quality assessment studies was provided by the Video Quality Experts Group [23], including: SRC (i.e., source or original sequence), HRC (i.e., Hypothetical Reference Circuit or processing applied to the SRC to obtain the test sequences, such as compression techniques), PVS (i.e., Processed Video Sequence or the resulting test sequence from applying an HRC to a SRC). In the context of FN, one should reflect another dimension of the system under test related to the interactivity part (e.g., the use of exploration trajectories in quality evaluation of immersive media). Towards this goal, we introduce the term Hypothetical Rendering Trajectories (HRT), to reference the simulated exploration trajectory that is applied to a PVS for rendering. It is worth mentioning the generality of this term is applicable to all immersive media from multi-view video to VR, AR, point clouds, and light fields [24].

B. Test Material

Three different SMV sequences were used in our study: Champagne Tower (CT), Pantomime (P) [25], and Big Buck Bunny Flowers (BBBF) [26], which were selected among the only four available dense FVV sequences according to [26] (the remaining one has not been used due to its similarity with BBBF). These sequences were also selected as test materials in [3]. Unlike the contents selected in previous studies [11], [12], [27], [28], these contents have at least 80 original views, which make possible to generate reference sequences that mimic navigations among different views (by using the original videos, synthesized views could be then compared with actual camera view in the same location). Description of the three SMV sequences are summarized in Table I. For each of the 3 SRC sequences, 20 HRCs were selected, covering 5 baseline distances (B), and 4 rate-points (RP). In addition, 2 HRTs were also included to generate 120 PVSs. To select these test conditions, in addition to the restrictions of the duration of a subjective test, a pretest with seven expert viewers was carried out and details on this selection are given in the following subsections.

1) Camera Configuration: For each source sequence (SRC), five stereo baseline values, as summarized in Table I, were selected in the test including the setting S_{B_0} without using synthesized views. The baseline is measured based on the camera distances/gaps between left and right real views. Here, B_i represents the stereo baseline distances that were settled to generate the synthesized virtual views, where i is the number of synthesized views between two real views. For instance, for camera setting S_{B_4} in the upper part of Fig. 2, using each pair of views captured by original cameras (indicated by two closest black cameras in the figure) as left and right references,

[1]The dataset can be downloaded from: ftp:\\ ftp.ivc.polytech.univ-nantes.fr\LS2N_IPI_FVV\

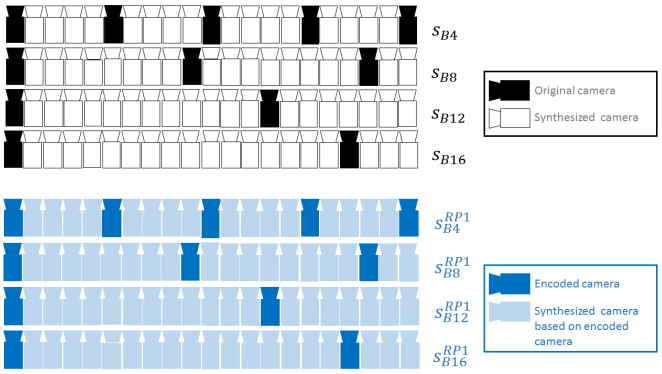

Fig. 2. Camera arrangements: 1) The upper part of the figure is the configuration designed in [8] and [13] where the black cameras represent the sequences taken with real original cameras while the white ones indicate the synthesized view using the original ones as reference; 2) The lower part of the figure is the camera configuration in our experiment, the deep blue camera represents the encoded/transmitted sequence taken from the corresponding original camera while the lighter blue ones indicate the synthesized ones using the encoded ones as reference.

four virtual views were synthesized in-between. In this case, the baseline distance is four, denoted as B_4. Fig. 2 illustrates the baseline setting for synthesized views generation in the subjective study. For example, in the lower part of Fig. 2, for $S^{RP_1}_{B_4}$, between each two transmitted encoded views, four virtual synthesized views were generated.

2) 3D-HEVC Configuration: In our experiment, HTM 13.0 in 3D High Efficiency Video Coding (3D-HEVC) mode was used to encode all the views of the three selected SMV sequences. However, as there are no ground truth depth maps for the selected contents [26], estimated depth maps were used instead of the encoded depth maps. These encoded views along with the selected original views are used as the reference views in the following synthesis process, which are also named as 'anchors'. The configuration of the 3D-HEVC encoder recommended in [3] was adopted in the experiment. Specifically, in this experiment, taking into account the contents and the limitations of the duration of subjective experiment tests, three rate-points (from the four rate-points selected by the MPEG community [29]) were considered for each SRC according to the results of the pretest to cover the full quality range. As summarized in Table I, for each content, the original sequences without compression are also included in the experiment (denoted as RP_0).

3) Depth Maps and Virtual Views Generation: In this paper, reference software tools were used for the preparation of the synthesized views, including Depth Estimation Reference Software (DERS) and View Synthesis Reference Software (VSRS), which have been developed throughout the MPEG-FTV video coding exploration and standardization activities. To generate virtual views with reference sequences taken by real cameras, depth maps and related camera parameters are needed. For sequences 'CT' and 'P', since original depth maps were not provided, DERS (version of 6.1) was used to generate depth map for each corresponding view. For synthesized views-generation, the version 4.1 of VSRS was applied. Relative parameters were set as recommended

TABLE I
INFORMATION OF THE SEQUENCES, INCLUDING PROPERTIES AND SELECTED CONFIGURATIONS (RATE-POINTS AND BASELINE DISTANCES)

Name	Views	Resolution	Fps	Seconds	Frames	QP values				Baseline Distance
						RP_1	RP_2	RP_3	RP_4	
BBBF	91	1280 x 768	24	5	121	35	-	45	50	B_0, B_2, B_5, B_9, B_{13}
CT	80	1280 x 960	29.4	10	300	37	43	-	50	B_0, B_4, B_8, B_{12}, B_{16}
P	80	1280 x 960	29.4	10	300	37	43	-	50	B_0, B_2, B_6, B_{12}, B_{16}

in [30] and [31] for each corresponding content. The baseline distances used in the test for each content are shown in Table I, which were selected after analyzing the results of the pretest (where ten baselines were considered for each content, from B_0 to B_{18} with steps of two), aiming at obtaining a good distribution of quality levels and taking into account the limitations on the duration of the subjective test sessions.

4) Navigation Trajectory Generation: One of the purposes of this subjective experiment was to check whether semantic contents of the videos and how the navigation trajectories among views will affect the perceived quality. Therefore, different HRTs were considered in this study, generating sweeps that focus more on important objects, since human visual system tends to attach greater interest on 'Regions of Interest' (ROI) that contain important objects [32]. Specifically, three different types of HRTs were considered initially and evaluated in the pretests: 1) first scanning from the left-most to the right-most views to observe the overall contents in the video, then scanning back to the views that contain the main objects and looked left and right around the central views that contains the object several times; 2) first scanning from the left-most to the right-most views to observe the overall contents in the video, then scanning back to the views that contain the main objects and finally stay in the central view that contains the main object; and 3) scanning from the left-most to the right-most views and scanning back to the leftmost until the sequence played completely (as in [3] and [8]). Each type of HRT was evaluated using two velocities (i.e., six configurations in total) to obtain appropriate view transitions in terms of smoothness and speed avoiding, undesired artifacts for the test (e.g., jerkiness). Then, two HRTs were finally selected for the test denoted with T_1 and T_2 as represented in Fig. 3: one of the first type aforementioned at 1 frame per view (fpv), and another of the second type at 2 fpv. The main reasons behind this selection are: 1) human observers may pay more attention and even stop navigating to observe targeted objects in the video; 2) according to the pretest results, sequences that are in form of T_1 and T_2 obtain the highest and lowest MOS scores correspondingly, which are in line with the goal of this study to verify the impact of trajectories on perceived quality so that the system could be better designed/improved for the worst cases.

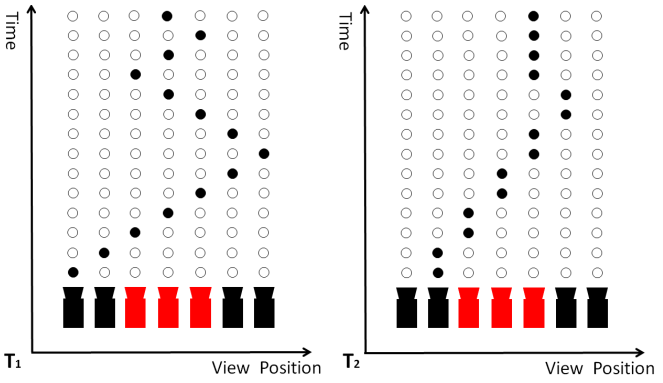

Fig. 3. Description of generated trajectories. In the figure, red cameras indicate views contain important objects while the black ones represent the one mainly contain background. **Left** T_1: Sweeps (navigation path) were constructed at a speed of one frame per view (as done in MPEG). **Right** T_2: Sweeps (navigation path) were constructed at a speed of two frames per view (the depiction of the sweep is reduced for the sake of the representation, e.g., the leftest/rightest views are actually navigated in T_2.).

C. Test Methodology

The methodology of Absolute Category Rating with hidden reference (ACR-HR) [33] was adopted for the subjective experiment. Thus, the observers watched sequentially the test videos, and after each one, they provided a score using the five-level quality scale. For this, an interface with adjectives representing the whole scale was shown until the score was provided, and then, the next text video was displayed. Also, it is worth noting that each test video was shown only once and the test videos were shown to each observer in different random orders. At the beginning of the test session, an initial explanation was given to the participants indicating the purpose and how to accomplish the test. Then, a set of training videos was shown to the observers to familiarize them with the test methodology and the quality range of the content. The entire session for each observer lasts for around 30 minutes.

D. Environment and Observers

The test sequences were displayed on a professional screen TVLogic LVM401W, using a high-performance computer. Observers were provided with a tablet connected to the displayed computer for voting. The test room was set up according to the ITU recommendation BT.500 [34], so the walls were covered by gray-color curtains and the lighting conditions were regulated accordingly to avoid annoying reflections. Also, a viewing distance of 3H (H being the height of the screen) was chosen.

There were totally 33 participants in the subjective test, including 21 females and 12 males, with ages varying from 19 to 42 (average age of 24). Before the test, the observers were screened for correct visual acuity and color vision using the Snellen chart and Ishihara test, respectively, and all of

TABLE II
THREE-WAY ANOVA RESULTS

Factor	SQ	DF	MS	F	p-value
B	11.5795	4	2.89487	41.7	0
T	0.3111	1	0.31115	4.48	0.0375
RP	29.3095	3	9.76983	140.75	0
B*T	0.169	6	0.02817	0.41	0.873
B*RP	9.526	22	0.433	6.24	0
T*RP	0.0841	4	0.02102	0.3	0.8752

them reported normal or corrected-to-normal vision. After the subjective test, the obtained scores were screened (to reject invalid or unreliable observers from further analysis) according to the procedure recommended by the ITU-R BT.500 [34] and the VQEG [23]. As a result after this screening, four observers were removed.

IV. SUBJECTIVE EXPERIMENT RESULTS AND ANALYSIS

The results of the subjective tests are shown in Fig. 4, where each sub-graph represents the MOS values with confidence intervals [34] for each content. Apart from MOS, the differential mean opinion score (DMOS) is also provided along with the database, computed from the hidden references according to [33]. As required for a quality dataset, the MOS values are well distributed covering almost the whole rating scale. In order to verify whether different Baselines (B), Rate-Points (RP) and, specially, virtual Trajectories (T) have significant impacts on perceived quality, a three-way analysis of variance (ANOVA) [35] was performed, so it is possible to check the main effects and the interaction effects (i.e., whether the effect of one variable depends on the value of another variable). The results of the ANOVA analysis are summarized in Table II, where 'SQ' represents the sum of squares due to each factor, 'DF' represents the degrees of freedom associated with each source, 'MS' represents the Mean Squares for each factor (i.e., the ratio SQ/DF), 'F' represents the F-statistic (which is the ratio of the mean squares), and finally, the 'p-value' is the corresponding probability value, which allows to know if the analyzed effects are statistically significant. In particular, a significance level of 95% was considered, so statistical significance is shown when a given 'p-value' is lower than 0.05.

From the results of this test and the results shown in Fig. 4, the following main conclusions could be drawn:

- For the same configuration (i.e., baseline, rate-point and trajectory), the quality scores obtained with different contents are significantly different.
- The effects of view-synthesis and compression artifacts are obvious, as shown when considering how the perceived quality changes with only baseline (for a given RP), or with only bitrate (fixing the baseline). The accumulation of the effects can be also observed in the scores for the tests sequences with combined degradations.
- The three considered factors (baselines, rate-points, and trajectories), have a significant impact on the perceived quality ($p-value = 0$ for B and RP, and $p-value = 0.038$ for T).
- In terms of interaction between the considered factors, the interaction between baseline distance and coding quality has a significant effect on the MOS scores ($p-value = 0$), as expected.

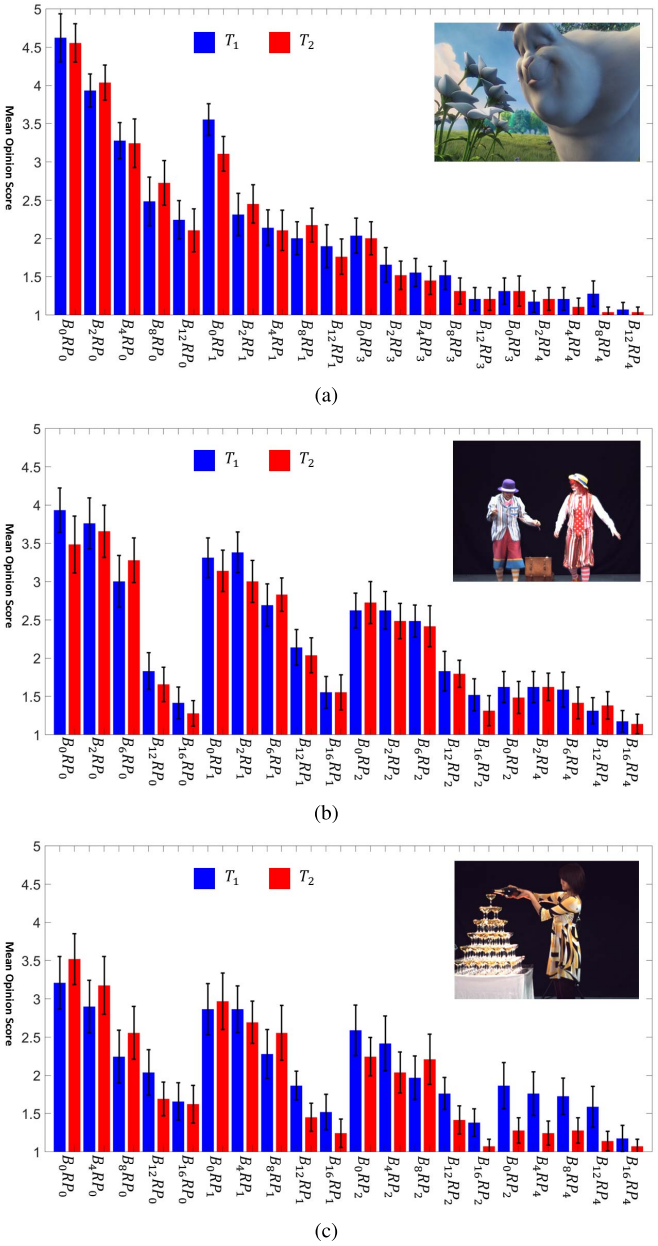

Fig. 4. MOS of the sweeping sequences with different RPs, Bs and Ts in the IPI-FVV Database, where blue bars and red bars represent sequences that are in the form of T_1 and T_2, respectively. Thumbnails of the three selected contents are shown in the top-right part of the each sub-figure.

In the following, a more detailed analysis of the impact of the trajectory on perceived quality is provided:

1) The averaged MOS values (averaged contents 'CT', 'P', 'BBBF' and conditions) of sequences in form of T_2 is smaller than the one of T_1. Apart from the ANOVA test, to further confirm the impact of the trajectory on the perceived quality, the database is divided into two

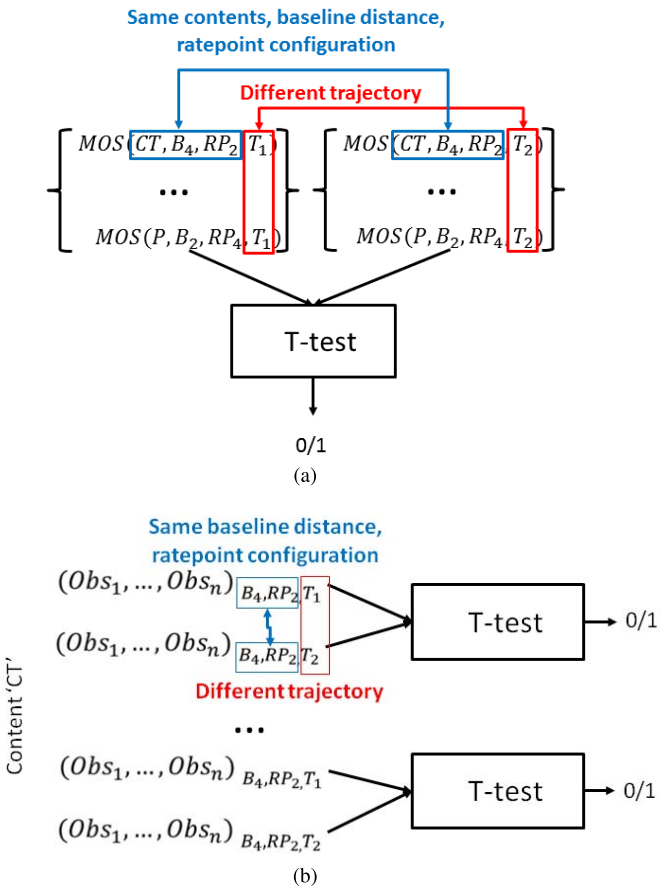

Fig. 5. Figure explaining how the t-test is conducted for further checking the impact of the trajectory on perceived quality. (a) $MOS(\cdot)$ denotes the mean opinion score of a sequence. (b) $Obs_i(\cdot)$ denotes the quality score given by the i_{th} observer.

sets according to in which trajectory the sequences are generated (i.e., sequence with T_1 and with T_2). A t-test was also conducted comparing two sets containing: 1) the MOS values for each configuration of baseline and rate-point for all contents in form of T_1, and 2) the corresponding MOS values in form of T_2. The t-test process is illustrated in Fig. 5 (a). According to the result, there is a significant difference between the quality of these two sets, thus between T_1 and T_2.

2) Certain contents are more sensitive to certain trajectories. To further check whether the impact of certain trajectories depend on the content of the sequences, another t-test was conducted. More specifically, for each content, pairs of sequences generated with the same baseline and rate-point but different trajectory are first considered. Then, the t-test was conducted by taking the individual subjective scores (opinion scores from all the observers) of each pair of these sequences as input. The procedure of the T-test is depicted in Fig. 5. According to the t-test result, for content 'CT', 50% of the pairs are of significantly different perceived quality. However, for content 'CT' and 'BBBF', only around 10% of pairs are of significantly different quality. It is proven that the impact of the trajectory on quality is content dependent. In other words, the 'extreme trajectory' of videos with different contents may be different.
3) Whether the quality of a sequence in form of one trajectory is higher than another depends also on the quality range (in terms of baseline and rate-point setting). Results of the t-test taking individual subjective score of each trajectory pair (as illustrated in Fig. 5) as input also shows that, for content 'CT' videos in form of T_2 are of better quality than the ones in T_1 when quality is higher than a certain threshold (smaller baseline or smaller rate-point) and vice versa. For example, for content 'CT' with rate-points larger than RP_2, the sequences in form of T_1 is better than the one with T_2.

In conclusion, it is confirmed by the subjective study that there is an impact on the perceived quality from navigation trajectory. It is found that content related trajectory is able to stress the system one step further for a more extreme situation. Therefore, image/video objective metrics that are able to indicate that sequences in form of one trajectory are of better quality than others are required to better push the system to its limit according to the contents. To fill out this need, a video quality metric is introduced in the next section.

V. Video Quality Measure for Free Viewpoint Videos

An objective quality measure that could provide more robust indication of the quality for a given HRT is required. Towards this goal, a full-reference Sketch-Token-based Video Quality Measure (ST-VQM) is proposed to quantify the change of structure. 'Sketch-Token' (ST) [36] model is a bag-of-words approach training a dictionary for representing the contours with contour's categories. Considering the fact that: 1) content related trajectory is able to stress the system, 2) content is related to structure, and 3) geometric distortions are the most annoying degradations that interrupt structure introduced by view synthesis, the main idea of the proposed method is to assess the quality of the navigation videos by quantifying to what extent the classes of contours change due to view synthesis, compression and transition among views. It is an extended version of our previous work [37] (a quality measure for image) to cope with the FVV scenario. In this version, the complex registration stage is replaced by local regions selection, and an ST-based temporal estimator is incorporated to quantify temporal artifacts.

The improved video quality metric consists of two parts, including a spatial metric ST-IQM, as shown in Fig. 6, and a temporal metric ST-T, as shown in Fig. 7. Details of each part are given in the following subsections.

A. Sensitive Region Selection Based on Interest Points Matching and Registration

Sensitive region selection is important for the later evaluation of the quality of DIBR-based synthesized views mainly for the following reasons:

1) Instead of uniform distortions distributed equally throughout the entire frame, synthesized views contains mainly local nonuniform geometric distortion.

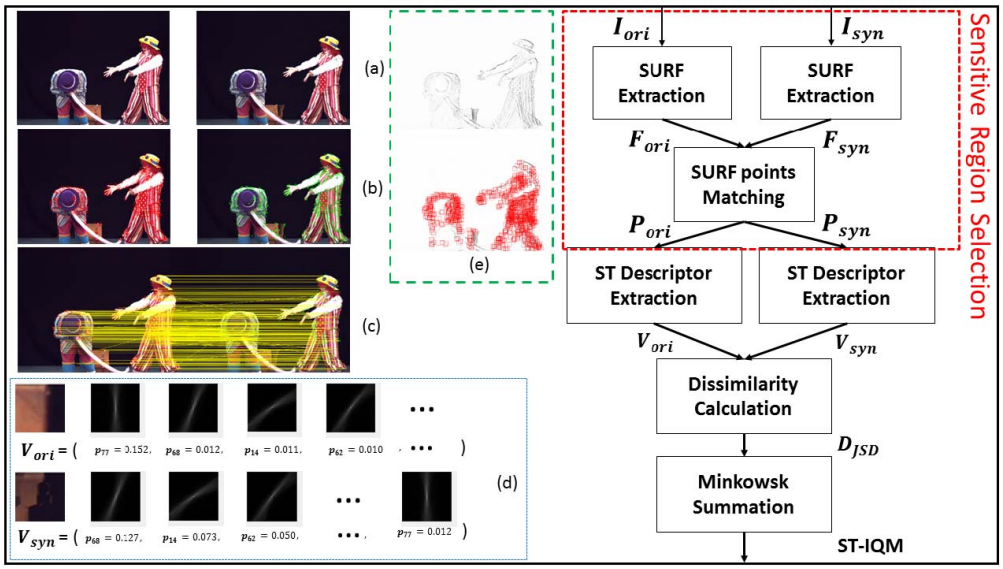

Fig. 6. Overall framework of the proposed objective metric: (a) Reference image (on the left) and synthesized image (on the right); (b) Extracted SURF key-points of the reference and synthesized images; (c) Matched key-points from the reference to the synthesized image (connected with yellow lines); (d) Extracted ST feature vector of the corresponding patches and its visualization of each contour category. (e) Error map between the reference and the synthesized frames (at the top), matched SURF points patches bounded with red color boxes, i.e., selected sensitive regions (at the bottom).

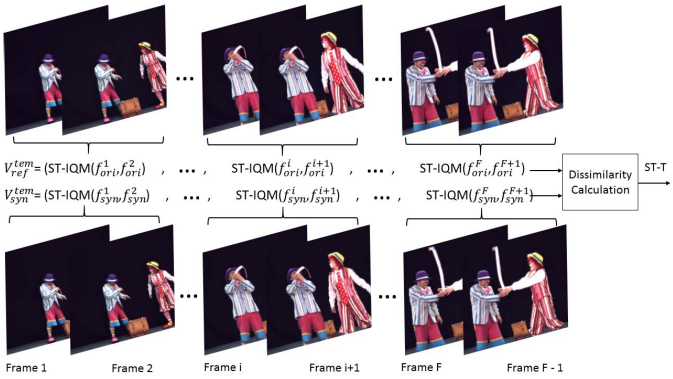

Fig. 7. Diagram of Sketch Token based temporal distortion computation, where F is the total frame number of the sequence.

2) Distortions distributed around the region of interest are less tolerable for human observers than a degradation located at an inconspicuous area [38]. Meanwhile, 'poor' regions are more likely to be perceived by humans in an image with more severity than the 'good' ones. Thus images with even a small number of 'poor' regions are penalized more gravely.
3) Global and local shifting of objects introduced by DIBR algorithms is a big challenge for point to point metrics like PSNR due to the mismatched correspondences.

Interest point-based descriptors,like Speeded Up Robust Features (SURF) [39], are local feature detectors frequently used in tasks like object detection. SURF reveals image's local properties and local shape information of objects are good candidates for selecting important local regions where DIBR local geometric artifacts could appear. Furthermore, later interest-point matching can also be useful to compensate for consistent 'Shift of Objects' artifacts which are, to some extent, acceptable for the human visual system.

The process of sensitive regions selection is summarized by the red dash bounding box in Fig. 6. First SURF F_{ori} and F_{syn} points are extracted in respectively both original I_{ori} and synthesized frames I_{syn}. Then, matching SURF points between the two frames are achieved by using an approximate feature-matching approach, based on priority search of multiple hierarchical clustering trees proposed in [40] (the original frame being considered as the reference for this matching process). Pairs of interest points that have significantly different x and y values are discarded, being considered as not plausible matched regions from the synthesis process. The patches P_{ori}, P_{syn} centered at the corresponding matched SURF points in synthesized and original images are then considered. The size of these patches is set as 35×35 to match ST formalism as introduced by [36] (see next section). The matching relation for all patches is encoded in a matching matrix $M_{match}(x_r, y_r) = (x_m, y_m)$, where (x_r, y_r) is the coordinate of the SURF point of the patch of the reference frame and (x_m, y_m) is the coordinate of its matched SURF point of the patch in the synthesized frame.

To illustrate the capability of SURF for selecting sensitive regions, one example is presented in Fig. 6 (e). The error maps are generated with the synthesized and the reference images as introduced in [18]. The darker the region the more distortions it contains, as shown in the top part of the dashed bounding green box in Fig. 6 (e). The red bounding box represents the sensitive regions as extracted by the proposed process. It can be observed that, as desired, the majority of regions containing severe local distortions are well identified by this process.

B. Sketch-Token Based Spatial Dissimilarity

Structures convey critical visual information and are beneficial to scene understanding, particularly the fine

structures (edge) and main structures (contour) [41], [42]. Considering the process for synthesizing virtual views by DIBR methods, the key target is to transfer the occluded regions (mainly occurred at the contour of the foreground objects) in the original view to be visible in the virtual view. Measuring the variations occurred at the contours is highly related to the degradation of image quality in that use case. Consequently, a method that correctly encodes contours would be a good candidate. The local edge-based mid-level features called 'Sketch Token' [36] has been proposed to capture and encode contour boundaries. It is based on the idea that the structure in an image patch can be described as a linear combination of 'contour' patches from a universal codebook (trained once for all).

In Lim *et al.* work [36], to train the codebook of contour patches, human subjects were asked to draw sketches as structural contours for each image in a training set. A total of 151 classes of sketch token were formed by clustering 35×35 pixels patches from the labeled training set. After extracting a set of low-level features from the patches, random decision forests model was adopted to train 150 contour classifiers for the contours within patches. Each output of every trained contour classifier is the likeliness p_i of the existence of one correspondence contour i in the patch. The 151_{th} category is for patches that do not contain any structural contours (e.g. patches with only smooth texture). One can calculate p_{151} with $1 - \sum_{i \in (1,150)} p_i$, since $\sum_{i \in (1,151)} p_i = 1$. Finally, the output of these 151 classifiers are concatenated to form the ST vector so that with a given pixel (x, y), the corresponding patch can be represented as $V(x, y) = (p_1, p_2, ..., p_{151})$ and the set of classifiers as the universal codebook.

In our metric, we extract the ST vectors V_{ori} and V_{syn} for each patches P_{ori} and P_{syn} of the matched SURF points pairs in matching matrix M_{match}. The dissimilarity between each matched contour vectors V_{ori} and V_{syn} is then computed. As the vectors contain probability with the sum of all the p_i equals to 1, we propose to use Jensen–Shannon divergence as a dissimilarity measure which presents the advantages to be bounded as opposed to the original Kullback–Leibler divergence. The dissimilarity between the matched patches centering at (x_r, y_r) and (x_m, y_m) respectively is then calculated as

$$D_{JSD}(V_{ori}, V_{syn}) = \frac{1}{2} D_{KLD}(V_{ori}(x_r, y_r), A) + \frac{1}{2} D_{KLD}(V_{syn}(x_m, y_m), A) \quad (1)$$

Where $A = \frac{1}{2}(V_{ori}(x_r, y_r) + V_{syn}(x_m, y_m))$, and D_{KLD} is the Kullback–Leibler divergence defined as

$$D_{KLD}(V_{ori}, V_{syn}) = \sum_i V_{ori}(i) log \frac{V_{ori}(i)}{V_{syn}(i)} \quad (2)$$

In order to amplify error regions with larger dissimilarity, the Minkowski distance measure is used as pooling strategy across sensitive regions. The spatial part of the proposed metric ST-IQM is then defined as

$$ST\text{-}IQM(I_{ori}, I_{syn}) = \frac{[\sum_N D_{JSD}(V_{ori}(x_r, y_r), V_{syn}(x_m, y_m))^\beta]^{\frac{1}{\beta}}}{N} \quad (3)$$

Where N is the total number of matched SURF points in the frame and β is a parameter corresponds to the $\beta - norm$ defining the L^β vector space.

C. Sketch Token Based Temporal Dissimilarity

Sweeping between views introduces and amplifies specific temporal artifacts including flickering, temporal structure inconsistency and so on. Among them, temporal structure inconsistency is usually the most sensitive artifact for human observers since it is usually located around important moving objects and is more obvious to notice compared to other temporal artifacts.

To quantify temporal structure inconsistency, we further compute the dissimilarity score between each pair of continuous frames using the proposed Sketch-Token model introduced in section V-B. In the previous section, ST-IQM was used to quantify the difference of structure organization between two images (original purpose of this framework). It can also be used to encode and describe how structures are evolving from one frame to another along a given sequence. Temporal structure changes as observed in FVV should affect this description. This idea is exploited to refine the quality estimation in case of FVV in order to capture temporal inconsistency.

Fig. 7 is a diagram explaining how the Sketch Token based temporal distortion is calculated. More specifically, for each pair of continuous frames of a sequence S, f^i and f^{i+1}, one can compute $ST - IQM(f^i, f^{i+1})$ using equation (3). A vector V^{tem} can be formed considering all frames of the sequence (each component of the vector corresponding to $ST - IQM(f^i, f^{i+1})$). We define the Sketch Token based temporal dissimilarity (ST-T) between the original and the synthesized sequences as the Euclidean distance between the two temporal vectors of the original and the synthesized sequence:

$$ST\text{-}T(S_{ori}, S_{syn}) = ED(V^{tem}_{ori}, V^{tem}_{syn}) \quad (4)$$

where $ED(\cdot)$ is the Euclidean distance function.

D. Pooling

With the spatial Sketch Token based score (ST-IQM) and the temporal Sketch Token based score (ST-T), it is desirable to combine them to produce an overall score. The final quality score of a synthesized sequence is defined as:

$$ST\text{-}VQM = w_S \cdot ST\text{-}IQM + w_T \cdot ST\text{-}T + \gamma \quad (5)$$

where w_S, w_T are two parameters used to balance the relative contributions of the spatial and temporal scores with a bias term γ. The selection and the influence of the related parameters will be given in section VI.

TABLE III

PERFORMANCE COMPARISON OF THE PROPOSED MEASURE WITH STATE-OF-THE-ART

	PCC	**SCC**	**RMSE**
Image Quality Measures			
3DSwIM [17]	0.5230	0.5649	0.8640
MWPSNR [18]	0.5705	0.8192	0.8304
MWPSNRr [20]	0.5779	0.8295	0.8252
MPPSNR [19]	0.5706	0.8299	0.8304
MPPSNRr [20]	0.5603	0.8319	0.8377
ST-IQM	0.8805	0.8511	0.4793
Video Quality Measures			
Liu-VQM [15]	0.9286	0.9288	0.3753
ST-T	0.8336	0.8926	0.4837
ST-VQM	**0.9509**	**0.9420**	**0.3131**

TABLE IV

PERFORMANCE COMPARISON OF METRICS FOR DISTINGUISHING SEQUENCE IN DIFFERENT TRAJECTORIES

	AUC-DS	**AUC-BW**	**CC**
Image Quality Metrics			
3DSwIM [17]	0.4603	0.8311	**0.8667**
MWPSNR [18]	0.5571	0.6889	0.6000
MWPSNRr[20]	0.5317	0.6933	0.6667
MPPSNR [19]	0.5079	0.7022	0.6667
MPPSNRr[20]	0.5238	0.6933	0.6667
ST-IQM	0.5016	0.7244	0.6000
Video Quality Metrics			
Liu-VQM [15]	0.6270	0.8311	0.7333
ST-T	0.5857	0.8800	0.8000
ST-VQM	**0.6762**	**0.8933**	**0.8667**

VI. EXPERIMENT RESULTS OF THE PROPOSED ST-VQM

The IPI-FVV database described in section III is adopted for the evaluation of the objective measures' performance. For comparison, only image/video measures designed for quality evaluation of view-synthesis artifacts are tested, since commonly used metrics fail to quantify geometric distortions [15], [17], [18], [20]. To compare the performances of the proposed measure with the state of the art, we firstly used the common criteria of computing Pearson correlation coefficient (PCC), Spearman's rank order correlation coefficient (SCC) and root mean squared error (RMSE) between the subjective scores and the objective ones (after applying a non-linear mapping over the measures) [23]. In the case of image quality measures, their corresponding spatial objective scores are first calculated frame-wise, and the final object score is computed by averaging the spatial scores.

The overall results are summarized in Table III and the best performance values are marked in bold. As it can be observed from Table III, ST-VQM, Liu-VQM are the two best performing metrics, with PCC equal to 0.9509 and 0.9286, correspondingly. To analyze if the differences between those values are significant, the F-test based on the residual difference between the predicted objective scores and the subjective DMOS values as described in [15] is employed. More specifically, the residual $R_{ST-VQM} = ST\text{-}VQM - DMOS$ and $R_{Liu-VQM} = Liu\text{-}VQM - DMOS$ are first calculated. Then the variance of the two residuals are computed (i.e., $\hat{\sigma}^2(R_{Liu-VQM}) = 1.4846$ and $\hat{\sigma}^2(R_{ST-VQM}) = 0.3536$, where $\hat{\sigma}^2(\cdot)$ indicates the variance). The ratio $\frac{\hat{\sigma}^2(R_{Liu-VQM})}{\hat{\sigma}^2(R_{ST-VQM})} > F - ratio$ ($F\text{-}ratio$ is obtained according to the samples size and the significant level, i.e., 95% in this paper), which indicates that our proposed metric significantly outperform the second best performing metric (Liu-VQM). As it can be observed, the performance of the image metrics, including MW-PSNR and MP-PSNR, is very limited, which can be due to: 1) they over-penalize the consistent shifting artifacts, and 2) these measures do not take temporal distortions into account.

As it has been verified in the subjective test results, navigation scan-paths affect the perceived quality. Therefore, it is important for an objective metric to point out whether the perceived quality using a given trajectory is better than using other trajectories. As thus, the metric can be used to evaluate the limit of the system in worse navigation situations. To this end, the Krasula performance criteria [43], [44] is used to assess the ability of objective measures to estimate whether one trajectory is better than another with the same rate-point and baseline configurations in terms of perceived quality. Pairs of sequences generated with the same configurations but in form of T_1 and T_2 in the dataset are selected to calculate the area under the ROC curve of the 'Better vs. Worse' categories (AUC-BW), area under the ROC curve of the 'Different vs. Similar' category (AUC-DS), and percentage of correct classification (CC) (see [43] and [44] for more details). More specifically, since pairs are collected in form of (T_1, T_2) with other parameters fixed, if one metric obtain higher AUC-BW, it shows more capability to indicate that sequences with certain trajectory are better/worse than with another. Similarly, if the metric obtains higher AUC-DS, then it can better tell whether the quality of sequences in form of one trajectory is different/similar to the ones in form of another trajectory. As it can be observed in the results are reported in Table IV, the proposed metric obtain the best performance in terms of the three evaluation measures. It is proven that the proposed ST-VQM is able to quantify temporal artifacts introduced by views switch. More importantly, ST-VQM is the most promising metric in telling sequence generated in form of which trajectory is of better quality than the others.

A. Selection of Parameters

It would be desirable that the performance of a VQM does not vary significantly with a slight change of the parameters. In this section, an analysis of the selection of the parameter of the proposed metric is presented. In order to properly select w_S, w_T and γ in equation (5), as well as to check the performance dependency of the parameters, a 1000 times cross-validation is conducted. More specifically, the entire database is separated into a training set (80%) and testing set (20%) 1000 times, and the most frequently occurred value will be selected for the corresponding parameter. Before the

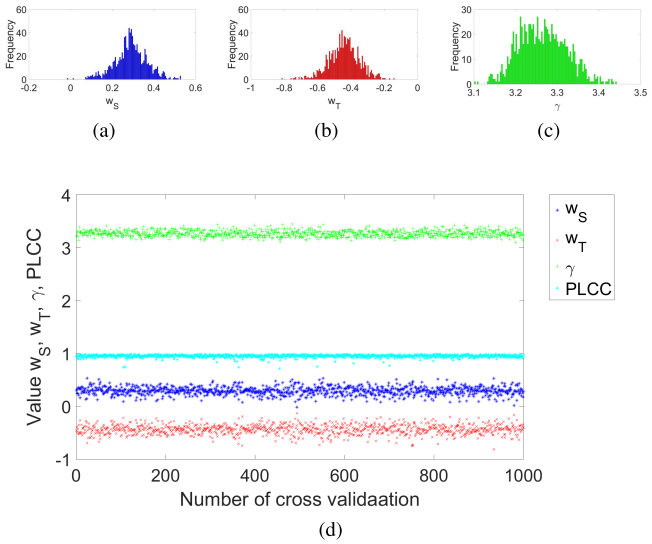

Fig. 8. (d) Values of w_S, w_T, γ and their corresponding PLCC across 1000 times cross-validation.

validation test, we first multiply $ST\text{-}IQM$ by 10^{10} and $ST\text{-}T$ by 10^5 so that the difference between the corresponding parameter w_S, w_T will be smaller making easier for latter visualization (it has to be pointed out that this operation does not change the performance). The values of the three parameters with the corresponding PCC value across of 1000 times cross-validation are shown in Fig. 8 (d). It can be observed that both the values of the three parameters and the performance do not change significantly throughout 1000 times, which verifies the fact that the performance of the metric does not change dramatically along with the modification of the parameters. Fig. 8 (a)-(b) depicts the histograms of frequencies of the three parameters' values relatively. As it can be seen that $w_S = 0.28, w_T = -0.43$ and $\gamma = 3.26$ are the three most frequent value among 1000 times. They are thus selected and fixed for reporting the final performance in Table III and Table IV. The mean value of PCC, SROCC, and RMSE of the proposed metric across the 1000 times is 0.9513, 0.9264 and 0.2895 correspondingly, which are close to the performance values reported in Table III with the selected configuration.

Subsequently, the performance dependency of the proposed algorithm on the exponent variable β in equation (3) and the distance approaches has been reported and examined in [37]. Therefore, in this paper, the same $\beta = 4$ and the Jensen Shannon divergence are selected.

VII. Conclusion

In this paper, aiming at better quantifying the specific distortions in sequences generated for FVV systems, both subjective and objective analyses have been conducted. On one side, in the subjective study, different configurations of compression and view-synthesis have been considered, which are the two main sources of degradations in FVV. In addition, following the approach of using simulating navigation trajectories that the users of immersive media may employ to explore the content, two different trajectories (referred as Hypothetical Rendering Trajectories) have been used to study their impact on the perceived quality. Knowing these possible effects may help on the identification of critical trajectories that may be more suitable to carry out quality evaluation studies related to the benchmark of systems in the worst cases. Also, it must be pointed out that the sweeps generated in this test focus more on views that contain regions of interest (e.g., moving objects) in videos since human observers are more interested in them and even stop navigating after these regions show up. By analyzing the subjective results, we find that the way of how the trajectories are generated does affect the perceived quality. Furthermore, the dataset generated for the subjective tests (called IPI-FVV), along with the obtained subjective scores is made available for the research community in the field. On the other side, in the objective study, a Sketch-Token-based VQA metric is proposed by checking how the classes of contours change between the reference and the degraded sequences spatially and temporally. The results of the experiments conducted on IPI-FVV database has shown that the performance of proposed ST-VQM is promising. More importantly, ST-VQM is the best performing metric in predicting if sequences based on a given trajectory are of higher/lower quality than sequences based on other trajectories, with respect to subjective scores. Finally, in the future, 1) related subjective and objective studies and datasets will be extended considering more contents and applications (e.g., more SMV contents, light field and virtual/augmented reality scenarios), 2) ST-VQM will be improved as no reference metric.

References

[1] M. Tanimoto, "FTV standardization in MPEG," in *Proc. 3DTV-Conf., True Vis.-Capture, Transmiss. Display 3D Video*, Budapest, Hungary, Jul. 2014, pp. 1–4.

[2] Y. Takaki, "Development of super multi-view displays," *ITE Trans. Media Technol. Appl.*, vol. 2, no. 1, pp. 8–14, Jan. 2014.

[3] *Call for Evidence on Free-Viewpoint Television: Super-Multiview and Free Navigation*, document ISO/IEC JTC1/SC29/WG11, N15348, 112th MPEG meeting, Warsaw, Poland, Jun. 2015.

[4] A. T. Hinds, D. Doyen, and P. Carballeira, "Toward the realization of six degrees-of-freedom with compressed light fields," in *Proc. IEEE Int. Conf. Multimedia Expo*, Hong Kong, Jul. 2017, pp. 1171–1176.

[5] P. Hanhart, E. Bosc, P. Le Callet, and T. Ebrahimi, "Free-viewpoint video sequences: A new challenge for objective quality metrics," in *Proc. IEEE 16th Int. Workshop Multimedia Signal Process. (MMSP)*, Sep. 2014, pp. 1–6.

[6] P. Merkle *et al.*, "The effects of multiview depth video compression on multiview rendering," *Signal Process., Image Commun.*, vol. 24, nos. 1–2, pp. 73–88, Jan. 2009.

[7] F. Battisti and P. Le Callet, "Quality assessment in the context of FTV: Challenges first answers and open issues," *IEEE COMSOC MMTC Commun.-Frontiers*, vol. 11, no. 2, pp. 22–27, 2016.

[8] P. Carballeira, J. Gutiérrez, F. Morán, J. Cabrera, and N. García, "Subjective evaluation of super multiview video in consumer 3D displays," in *Proc. Int. Workshop Quality Multimedia Exper.*, Costa Navarino, Greece, May 2015, pp. 1–6.

[9] A. Dricot *et al.*, "Subjective evaluation of Super Multi-View compressed contents on high-end light-field 3D displays," *Signal Process., Image Commun.*, vol. 39, pp. 369–385, Nov. 2015.

[10] 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.

[11] E. Bosc, P. Le Callet, L. Morin, and M. Pressigout, "Visual quality assessment of synthesized views in the context of 3D-TV," in *3D-TV System with Depth-Image-Based Rendering*. New York, NY, USA: Springer, 2013, pp. 439–473.

[12] 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.

[13] P. Carballeira, J. Gutiérrez, F. Morán, J. Cabrera, F. Jaureguizar, and N. García, "Multiview perceptual disparity model for super multiview video," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 1, pp. 113–124, Feb. 2017.

[14] R. Recio, P. Carballeira, J. Gutiérrez, and N. García, "Subjective assessment of super multiview video with coding artifacts," *IEEE Signal Process. Lett.*, vol. 24, no. 6, pp. 868–871, Jun. 2017.

[15] 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.

[16] R. Song, H. Ko, and C. C. J. Kuo, "MCL-3D: A database for stereoscopic image quality assessment using 2D-image-plus-depth source," *J. Inf. Sci. Eng.*, vol. 31, no. 5, pp. 1593–1611, 2015.

[17] 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.

[18] D. Sandić-Stanković, D. Kukolj, and P. Le Callet, "DIBR synthesized image quality assessment based on morphological wavelets," in *Proc. Int. Workshop Quality Multimedia Exper.*, Costa Navarino, Greece, Jul. 2015, pp. 1–6.

[19] D. Sandić-Stanković, D. Kukolj, and P. Le Callet, "DIBR synthesized image quality assessment based on morphological pyramids," in *Proc. 3DTV-Conf., True Vis.-Capture, Transmiss. Display 3D Video*, Lisbon, Portugal, Jul. 2015, pp. 1–4.

[20] D. Sandić-Stanković, D. Kukolj, and P. Le Callet, "DIBR-synthesized image quality assessment based on morphological multi-scale approach," *EURASIP J. Image Video Process.*, vol. 2017, no. 1, p. 4, 2016.

[21] 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.

[22] E. Ekmekcioglu, S. Worrall, D. De Silva, A. Fernando, and A. M. Kondoz, "Depth based perceptual quality assessment for synthesised camera viewpoints," in *Proc. Int. Conf. User Centric Media*, Jan. 2010, pp. 76–83.

[23] "Report on the validation of video quality models for high definition video content," Univ. Krakow, Poland, U.K., Tech. Rep., Jun. 2010.

[24] I. Viola, M. Řeřábek, and T. Ebrahimi, "Impact of interactivity on the assessment of quality of experience for light field content," in *Proc. Int. Conf. Quality Multimedia Exper.*, Erfurt, Germany, Jun. 2017, pp. 1–6.

[25] *Nagoya University Sequences*. Accessed: Feb. 15, 2017. [Online]. Available: http://www.fujii.nuee.nagoya-u.ac.jp/multiview-data/

[26] *Overview of MPEG-I Visual Test Materials*, document ISO/IEC JTC1/SC29/WG11, Output N17718, 123th MPEG Meeting, Ljubljana, Slovenia, Jun. 2018.

[27] Y. Liu, S. Ma, Q. Huang, D. Zhao, W. Gao, and N. Zhang, "Compression-induced rendering distortion analysis for texture/depth rate allocation in 3d video compression," in *Proc. Data Compress. Conf. (DCC)*, Mar. 2009, pp. 352–361.

[28] J. Kilner, J. Starck, J.-Y. Guillemaut, and A. Hilton, "Objective quality assessment in free-viewpoint video production," *Signal Process., Image Commun.*, vol. 24, nos. 1–2, pp. 3–16, 2009.

[29] *Revised Summary of Call for Evidence on Free-Viewpoint Television: Super-Multiview and Free Navigation*, document ISO/IEC JTC1/SC29/WG11, N16523, 116th MPEG Meeting, Chengdu, China, Oct. 2016.

[30] *DERS Software Manual*, document ISO/IEC JTC1/SC29/WG11, M34302, 109th MPEG Meeting, Sapporo, Japan, Jul. 2014.

[31] *FTV Software Framework*, document ISO/IEC JTC1/SC29/WG11, N15349, 112th MPEG Meeting, Warsaw, Poland, Jun. 2015.

[32] U. Engelke and P. Le Callet, "Perceived interest and overt visual attention in natural images," *Signal Process., Image Commun.*, vol. 39, pp. 386–404, Nov. 2015.

[33] *Methods for the Subjective Assessment of Video Quality, Audio Quality and Audiovisual Quality of Internet Video and Distribution Quality Television in Any Environment*, document ITU-T Rec. P.913, 2014.

[34] *Methodology for the Subjective Assessment of the Quality of Television Pictures*, document Rec. ITU-R BT.500, 2012.

[35] L. St *et al.*, "Analysis of variance (ANOVA)," *Chemometrics Intell. Lab. Syst.*, vol. 6, no. 4, pp. 259–272, 1989.

[36] J. J. Lim, C. L. Zitnick, and P. Dollár, "Sketch tokens: A learned mid-level representation for contour and object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Portland, OR, USA, Jun. 2013, pp. 3158–3165.

[37] S. Ling and P. Le Callet, "Image quality assessment for free viewpoint video based on mid-level contours feature," in *Proc. IEEE Int. Conf. Multimedia Expo*, Hong Kong, Jul. 2017, pp. 79–84.

[38] A. Ninassi, O. Le Meur, P. Le Callet, and D. Barba, "Does where you gaze on an image affect your perception of quality? Applying visual attention to image quality metric," in *Proc. Int. Conf. Image Process.*, San Antonio, TX, USA, Sep. 2007, p. II-169.

[39] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (SURF)," *Comput. Vis. Image Understand.*, vol. 110, no. 3, pp. 346–359, 2008.

[40] M. Muja and D. G. Lowe, "Fast matching of binary features," in *Proc. Comput. Robot Vis.*, May 2012, pp. 404–410.

[41] K. Gu, L. Li, H. Lu, X. Min, and W. Lin, "A fast reliable image quality predictor by fusing micro- and macro-structures," *IEEE Trans. Ind. Electron.*, vol. 64, no. 5, pp. 3903–3912, May 2017.

[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] L. Krasula, K. Fliegel, P. Le Callet, and M. Klíma, "On the accuracy of objective image and video quality models: New methodology for performance evaluation," in *Proc. Int. Conf. Quality Multimedia Exper.*, Lisbon, Portugal, Jun. 2016, pp. 1–6.

[44] P. Hanhart, L. Krasula, P. Le Callet, and T. Ebrahimi, "How to benchmark objective quality metrics from paired comparison data?" in *Proc. Int. Conf. Quality Multimedia Exper.*, Lisbon, Portugal, Jun. 2016, pp. 1–6.

Suiyi Ling received the B.S. and M.S. degrees in computer science from the Guangdong University of Technology, France, and the M.S. degree in multimedia and big data management from the Université de Nantes, where he is currently pursuing the Ph.D. degree. His research interests include computer vision, machine learning, multimedia quality assessment, and perceptual image processing.

Jesús Gutiérrez received the degree in telecommunication engineering (five-year engineering program) from the Universidad Politécnica de Valencia, Spain, in 2008, and the master's degree in communications technologies and systems (two-year M.S. program) and the Ph.D. degree in telecommunication from the Universidad Politécnica de Madrid, Spain, in 2011 and 2016, respectively. Since 2016, he has been a Post-Doctoral Researcher with the IPI Team, LS2N Laboratory, Université de Nantes, France, where he was a Marie Curie Fellow with the PROVISION ITN and is currently a Marie Curie PRESTIGE Fellow. His research interests are in the areas of image and video processing, immersive media, the evaluation of multimedia quality of experience, human behavior, and visual perception.

Ke Gu received the B.S. and Ph.D. degrees in electronic engineering from Shanghai Jiao Tong University, Shanghai, China, in 2009 and 2015, respectively. He is currently a Professor with the Beijing University of Technology, Beijing, China. His research interests include environmental perception, image processing, quality assessment, and machine learning. He received the Best Paper Award from the IEEE TRANSACTIONS ON MULTIMEDIA, the Best Student Paper Award at the IEEE International Conference on Multimedia and Expo in 2016, and the Excellent Ph.D. Thesis Award from the Chinese Institute of Electronics in 2016. He was the Leading Special Session Organizer in the VCIP 2016 and the ICIP 2017. He serves as a Guest Editor for the *Digital Signal Processing Journal*. He is currently an Associate Editor of the IEEE ACCESS and the *IET Image Processing*. He is a reviewer for 20 top SCI journals.

Patrick Le Callet received the M.Sc. and Ph.D. degrees in image processing from the École Polytechnique de l'Université de Nantes. He was an Assistant Professor from 1997 to 1999 and a full-time Lecturer from 1999 to 2003 with the Department of Electrical Engineering, Technical Institute of the Université de Nantes. He led the Image and Video Communication Laboratory, CNRS IRCCyN, from 2006 to 2016, and was one of the five members of the Steering Board of CNRS from 2013 to 2016. Since 2015, he has been the Scientific Director of the Cluster Ouest Industries Creatives, a five-year program gathering over ten institutions (including three universities). Since 2017, he has been one of the seven members of the Steering Board of the CNRS LS2N Laboratory (450 researchers), as a Representative of the École Polytechnique de l'Université de Nantes, Université de Nantes. He is mostly involved in research dealing with the application of human vision modeling in image and video processing. His current research interests are the quality of experience assessment, visual attention modeling and applications, perceptual video coding, and immersive media processing.