Quality Assessment of Free-Viewpoint Videos by Quantifying the Elastic Changes of Multi-Scale Motion Trajectories

Suiyi Ling, Jing Li, Zhaohui Che, Xiongkuo Min, Guangtao Zhai, Patrick Le Callet

Quality Assessment of Free-viewpoint Videos by Quantifying the Elastic Changes of Multi-Scale Motion Trajectories

Suiyi Ling, *Student Member, IEEE,* Jing Li, Zhaohui Che, Xiongkuo Min, Guangtao Zhai, *Member, IEEE,* and Patrick Le Callet, *Fellow, IEEE*

***Abstract*—Virtual viewpoints synthesis is an essential process for many immersive applications including Free-viewpoint TV (FTV). A widely used technique for viewpoints synthesis is Depth-Image-Based-Rendering (DIBR) technique. However, such technique may introduce challenging non-uniform spatial-temporal structure-related distortions. Most of the existing state-of-the-art quality metrics fail to handle these distortions, especially the temporal structure inconsistencies observed during the switch of different viewpoints. To tackle this problem, an elastic metric and multi-scale trajectory based video quality metric (EM-VQM) is proposed in this paper. Dense motion trajectory is first used as a proxy for selecting temporal sensitive regions, where local geometric distortions might significantly diminish the perceived quality. Afterwards, the amount of temporal structure inconsistencies and unsmooth viewpoints transitions are quantified by calculating 1) the amount of motion trajectory deformations with elastic metric and, 2) the spatial-temporal structural dissimilarity. According to the comprehensive experimental results on two FTV video datasets, the proposed metric outperforms the state-of-the-art metrics designed for free-viewpoint videos significantly and achieves a gain of 12.86% and 16.75% in terms of median Pearson linear correlation coefficient values on the two datasets compared to the best one, respectively.**

***Index Terms*—Free-viewpoint video, free-viewpoint TV, elastic metric, dense motion trajectory, video quality assessment.**

I. Introduction

With the rise of more advanced 3D displays, head-mounted displays and other advanced equipment, immersive media applications such as Free-viewpoint TV (FTV), 3DTV, and Virtual Reality (VR) have become hot topics for media ecosystems. FTV, which provides user with the 'flying in the view' feeling by letting them navigate freely among different viewpoints, is one of the most popular scenarios in the area. In the FTV system, normally, only a limited set of input views are expected to be available and transmitted among all possible viewing angles that end user could select. As presented contents are usually synthesized using Depth-Image-Based Rendering technology (DIBR) [1], [2], in addition to compression and smooth transition between views, reliable synthesis algorithms that are robust to sparser camera arrangements are critical factors with respect to the rendered quality. DIBR based algorithms have the tendency to introduce local non-uniform structure-related distortions. Following are some detailed introductions of the challenging spacial/temporal structure-related distortions that could be introduced by the FTV systems.

Spatial structure-related distortions within FTV system have the following characteristics: 1) non-uniform and locally distributed: unlike traditional global uniform artifact, *e.g.*, blocking artifacts, in most of the cases, the dominant spatial distortions of synthesized videos are the local non-uniform distortions and they distribute mostly around dis-occluded regions, which could be seen in the reference views but are occluded in the virtual views [3]. These dis-occluded regions are commonly located at the boundaries of objects, *i.e.*, 'regions of interest', and thus are more disturbing because local poor quality regions are with greater possibility to be perceived by observers than the global acceptable ones [4]; 2) structure-related local noncontinuous distortions are normally geometric distortions, which modify/deform the shape of the objects; 3) acceptable global shifting: DIBR based algorithms could also introduce global continuous shifting of objects. Observers are normally more tolerant to this type of distortion than the local serious one. Nevertheless, this type of distortion would be over-penalized by pixel to pixel metric like PSNR.

Temporal structure-related distortions within FTV system could be categorized into two types. 1) temporal structure-related distortions within one viewpoint: considering each individual viewpoint, the spatial geometric distortion aroused by DIBR process will lead to temporal structure inconsistencies. Therefore, special temporal flickering in a form of structure (*e.g.*, object boundaries) fluctuation could be observed within videos at a certain viewpoint location. For example, Fig. 1 shows the change of the shape of a static object along temporal axis, *i.e.*, from the first frame f_1 to f_5. Different degrees of local structure-related distortions could be introduced differently among different viewpoints with different contents. 2) temporal structure-related distortions among viewpoints/unsmooth transition among viewpoints: considering the scenario of navigating among different viewpoints, local artifacts around dis-occluded regions, *e.g.*, geometric distortions or inpainting related distortions, would incur structure inconsistencies from one view to another. The larger the baseline distance is used

Suiyi Ling and Jing Li make equal contributions. Jing Li is the corresponding author.

Suiyi Ling, Jing Li, and Patrick Le Callet are with the Équipe Image, Perception et Interaction, Laboratoire des Sciences du Numérique de Nantes, Université de Nantes, France (e-mail: suiyi.ling@univ-nantes.fr; jing.li.univ@gmail.com; patrick.lecallet@univ-nantes.fr).

Zhaohui Che, Xiongkuo Min, and Guangtao Zhai are with the Department of Electronic Engineering, Shanghai Jiao Tong University, China (e-mail: {chezhaohui, minxiongkuo, zhaiguangta}@sjtu.edu.cn)

Manuscript submitted March, 2019.

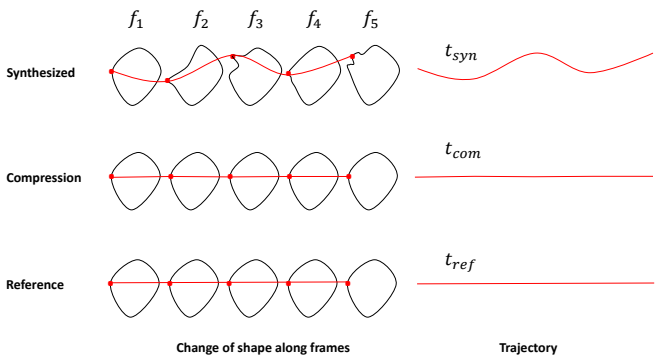


Fig. 1. Example of temporal trajectory deformation caused by spatial geometric distortion. t_{syn}: trajectory from synthesized video; t_{com}: trajectory from video contain transitional compression artifacts; t_{ref}: trajectory from reference video.

for synthesis, the more obvious the abrupt/sudden structural change could be observed when users switch their viewpoints from one to another. This unsmooth transition among different viewpoints could be considered as temporal flickering among viewpoints.

In extreme cases, the entire viewing experience of one Free Viewpoint Video (FVV) can be ruined by only one severely distorted region in one synthesized view [5], [6]. Quality assessment is vital for ensuring the quality of the entire system. Nevertheless, as most of existing image/video quality metrics have been tuned and designed to handle other types of distortions, *e.g.*, traditional uniform compression distortions including blocking artifact, blurriness *etc.*, they are mostly not suitable for FTV systems. New image/video quality assessment tools that can deal with these spatial and temporal structure-related distortions mentioned above are required. To evaluate the quality of free-viewpoint videos, some image and video quality assessment metrics have been proposed recently as introduced in the next section. However, their performances are limited, there is still a big room to improve.

II. Related Work

A. Limitations of Metrics Designed for Synthesized Views

Image quality assessment metric designed for synthesized views: in order to estimate the quality of synthesized views, there are many full reference (FR) metrics are proposed. The very first full reference approach that designed for evaluating the quality of synthesized images is proposed by Bosc *et al.* [7] by applying some prior knowledge acquired through subjective tests (*e.g.*, the common localization of view-synthesis artifacts along contours) to SSIM. Following this idea, Conze *et al.* [8] propose the view synthesis quality assessment (VSQA) metric, which improves SSIM with three visibility maps that characterizes the complexity of the images. Later, the '3D synthesized view image quality metric' (3DswIM) is proposed by Battisti *et al.* [9]. This metric is based on statistical features of wavelet sub-bands. In addition, Tsai and Hang [10] propose a metric based on compensating the shifts of the objects that appear in synthesized views by calculating the noise around them. Considering the fact that using multi-resolution approaches could increase the performance of image quality metrics, Sandić-Stanković *et al.* develop the 'Morphological Wavelet PSNR' (MW-PSNR) using a morphological wavelet decomposition [11]. Later they extend the work by using a multi-scale decomposition based on morphological pyramids, which is called 'Morphological Pyramid PSNR' (MP-PSNR) [12]. Recently, Stanković *et.al.* [13] point out that PSNR is more consistent with human judgment when it is calculated at higher morphological decomposition scales. They thus proposed reduced versions of the morphological multi-scale measures called reduced MP-PSNR and reduced MW-PSNR correspondingly (denoted as MP-PSNR$_r$ and MW-PSNR$_r$). According to their experimental results, the reduced versions (*i.e.*, MP-PSNR$_r$ and MW-PSNR$_r$) outperform the full versions. Li *et al.* [14] propose LOGs by considering both the geometric distortions as well as the sharpness of the images. NIQSV+ [15] is proposed based on a strong hypothesis that high-quality images are consist of flat areas separated by edges. Another state-of-the-art image quality metric is the EM-IQM (*i.e.*, EM$_{spa}$ in this paper) in [16] that quantifies the spatial structure deformations using elastic metric.

All the image metrics mentioned above suffers from at least one of the drawbacks mentioned below: 1) The human visual system is sensitive to severe local artifacts [17], [4]. The most upsetting artifacts in synthesized images are the inconsistent local geometric distortions instead of the consistent global uniform distortions. However, most of the existing metrics process the entire image equally and thus fail to locate and quantify local geometric distortions properly. Sensitive region selection should be considered as a pre-process module to select regions with structure-related distortions. 2) Global shifting within certain limits is acceptable for human observers but is punished severely by point-to-point based metrics. Due to equal-weighted pooling and point-wise comparison, some image quality assessment metrics mistakenly emphasize the consistent global shifting artifacts. 3) All of these metrics are not capable of quantifying the amount of temporal structure-related distortions.

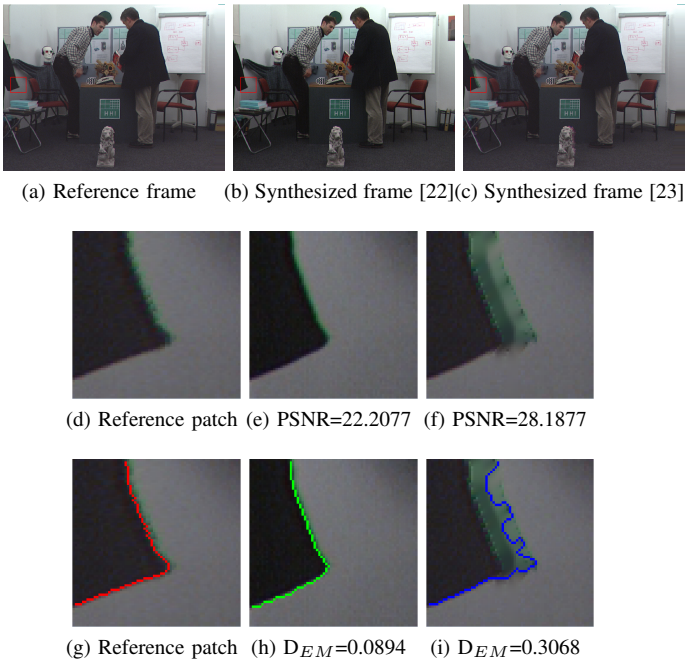
Video quality assessment metric designed for synthesized views: The 'Peak Signal to Perceptible Temporal Noise Ratio' (PSPTNR) metric, introduced by Zhao and Yu [18], quantifies temporal artifacts that can be perceived by observers in the background regions of the synthesized videos. Similarly, Ekmekcioglu *et al.* [19] propose a video quality metric by using depth and motion information to locate the degradations. The state-of-the-art video metric designed for free viewpoint videos is recently introduced by Liu *et al.* [20]. Their proposed metric (Liu-VQM) considers the spatio-temporal activity and the temporal flickering that appears in synthesized video sequences. However, none of the aforementioned video quality metrics is designed to quantify the unsmooth transition among views (temporal structure inconsistency observed during view switch). Compared to temporal distortions that could be observed at one viewpoint, during navigation, structure-related distortions could be amplified and new temporal structure deformation could be noticed. Therefore, temporal structure-related distortions are more challenging and should not be

underestimated.

In summary, a new video quality metric, which is able to quantify the aforementioned spatial-temporal structure-related distortions, is in urgent need.

B. So, How to Better Quantify the Spatial-Temporal Structure-related Distortions?

Elastic Metric (EM), which is first proposed in [21], is capable of quantifying the deformation between curves and thus could be the solution for quantifying the aforementioned specific distortions. It is first utilized in our previous work [16] to evaluate the quality of synthesized views spatially by calculating the amount of stretching or bending between curves/shapes in the reference and synthesized frames. Example of the advantage of using elastic metric compared to PSNR is shown in Fig. 2. By checking patches in Fig. 2 (e) and (f), it is obvious that the patch in Fig. 2 (e) with the slightly shifted object is of better visual quality than the one in Fig. 2 (f) with obvious structure-related distortions (*i.e.*, ghosting artifact). However, PSNR (the higher the score the better the quality) incorrectly indicates that the quality of (f) is better than (e). In the contrary, elastic metric (*i.e.*, D_{EM}, the higher the scores the the larger the amount of deformations between two compared curves) accurately points out that the blue curve in Fig. 2 (i) is more severely deformed compared to the red curve in Fig. 2 (g), indicating worse quality. After being able to quantify the spatial temporal structure-related distortions, then how to quantify the temporal ones, especially the ones observed during view switch? In this paper, multi-scale trajectory is employed along with elastic metric to handle this challenging problem.



(a) Reference frame (b) Synthesized frame [22](c) Synthesized frame [23]

(d) Reference patch (e) PSNR=22.2077 (f) PSNR=28.1877

(g) Reference patch (h) D_{EM}=0.0894 (i) D_{EM}=0.3068

Fig. 2. Examples of advantages of using elastic metric compared to pixel-to-pixel metric PSNR. Rows (from up to down): Reference/Synthesized images; Patches from images (a)-(c) bounded by red bounding boxes; Extracted contours of patches (d)-(f). Columns (from left to right): reference image, synthesized image with synthesis algorithm proposed in [22], synthesized image with synthesis algorithm proposed in [23].

Motion plays a vital role in visual perception of the contents and the perceived quality of sequence since 1) clues related to the objects' shapes are provided; 2) in most cases, visual attention is tend to be drawn on moving objects [24], [25], [26]. Human Visual System (HVS) tends to trace the salient moving objects when viewing a sequence [27], thus distortions around the moving objects may attract greater attention from the observers. There are already quality metrics designed based on this phenomenon [28], [29]. In the case of free viewpoint videos generated with synthesized views, apart from the special spatial distortions as described in [9], the synthesized videos suffer also from special temporal degradation related to motions of objects within one viewpoints or navigations among different viewpoints [20] as summarized in the previous section. Therefore, the success of one video metric dedicating to ensuring the quality of the entire FTV system relies on its capability of modeling and accounting both structure and motion perception in the HVS.

Motion trajectory, which traces moving objects, provides human observers with important spatial-temporal information for perceiving/detecting moving objects, *e.g.,*velocity, direction and even spatial information of the objects [30]. Since the detectability of a moving object could be impacted by the structural motion information (which could be represented by the motion trajectories), deformation of trajectory or changes of structure information along the trajectory that caused by distortions may affect the way how human detect objects and thus affect the perceived quality. Hence, on one hand, considering the characteristics of distortions produced by DIBR processes, synthesized sequences could be represented in sets of trajectories and their perceived quality could be evaluated with the trajectories and neighborhoods along them. For example, as shown in Fig. 1, due to the change of shape of the object, the trajectory of the synthesized sequence t_{syn} that traces one of the key point on the shape is deformed compared with the one of the reference sequence t_{ref} , while the one of the sequence that contain only common compression artifact remains almost unchanged. If one could quantify the amount of deformation of trajectories caused by related processes, the quality of the synthesized videos could be indicated. Since motion trajectories within sequence could be considered as open-curves, elastic metric is of potential to be used as a measure to quantify the deformation between the trajectory in a synthesized sequence and the one in the original sequence. On the other hand, as spatial-temporal distortions mainly happen around dis-occluded regions and distortions within the regions of moving objects are less tolerant for observers, the process of detecting meaningful moving trajectories could serve as a way to select severe distorted regions.

Multi-scale approaches, which transfer signals into a form of multi-scale representation, could be used to quantify structure loss caused by synthesized related artifacts. On one hand, the perceivability of videos' details is decided not only by the observer's visual system, but also by the viewing conditions (e.g., display resolution and viewing distance) [31]

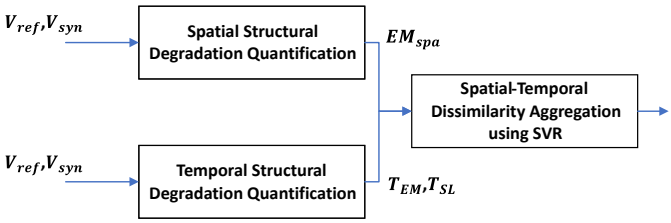


Fig. 3. Overall framework of the proposed EM-VQM metric.

Therefore, it is more reasonable to employ multi-scale strategy for quality assessment to take different subjective factors and configurations into account as claimed in [32]. On the another hand, as pointed out in [31], vision at a glance reflects high-level mechanism. Observers normally obtain the structure of the content first before looking into the details with scrutiny. In another word, structure of one content could be obtained from a lower resolution, while the details could be obtained through a higher resolution. The concept of multi-scale strategy is in line with human visual mechanism. Therefore, dissimilarity between the synthesis and reference videos calculated in different scales could be used differently depends on the subjective configurations/parameters to improve quality assessment metrics.

Based on the discussion above, in order to better evaluate the quality of free view-point videos by considering the characteristics of the spatial-temporal structure related distortions, an elastic metric and multi-scale trajectory based video quality assessment metric (EM-VQM) is proposed in this paper. The contributions of this paper are three-folds: 1) multi-scale motion trajectory is used as a proxy for temporal sensitive regions selection; 2) elastic metric is used to quantified the amount of motion trajectory deformations; 3) motion-structure-related descriptors are extracted along the multi-scale motion trajectories and used to quantify the spatial-temporal structural dissimilarities between the reference and synthesized videos.

The remainder of this paper is organized as follows. In Section III, the proposed model is introduced in detail. Then, the experimental results and analysis are presented in Section IV. Finally, conclusions are given in Section V.

III. The Proposed Model

The proposed elastic metric based video quality assessment metric is composed of two parts, including one part for quantifying the spatial structural degradation (Section III-A) and another part for quantifying temporal structural degradation (Section III-B). After computing the amount of spatial and temporal structure-related distortions at multi-scales separately, they are aggregated (Section III-C) to predict the overall quality score of one synthesized videos as illustrated in Fig. 3.

A. Quantify Spatial Structural Degradation

1) Spatial sensitive regions selection using key point matching: Unlike traditional distortions, which scatter over the entire frames, synthesized distortions distribute sparsely/locally

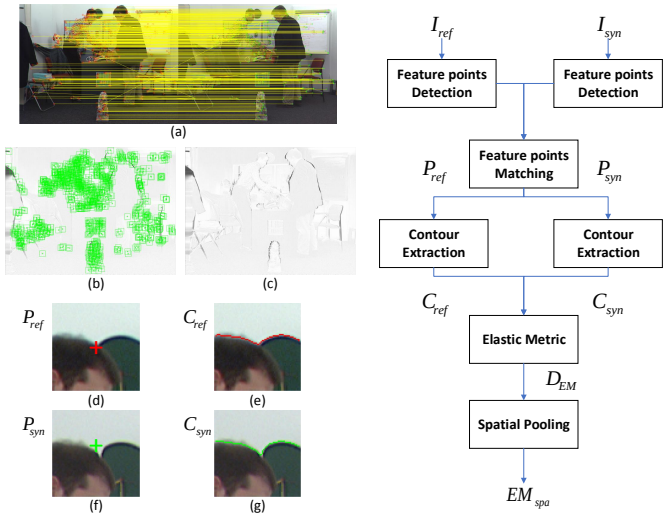


Fig. 4. Diagram of spatial structural degradation quantification. (a) Example of SURF feature points matching, where matched SURF feature points within reference and synthesized frames are connected with yellow color lines. (b) Selected sensitive patches, *i.e.*, patches centering at the matched SURF feature points bounded by green color boxes, plotted on the error map. (c) Error map obtained by comparing the reference and the synthesized frames (the darker the color the more severe the distortions within the regions). (d) Example of one sensitive region (patch) in the reference frame, whose center is one SURF feature point labeled with a red color cross. (e) Contour extracted from the patch. (f) Example of one sensitive region (patch) in the synthesized frame, whose center is one SURF feature point labeled with a green cross. (g) Contour extracted from the patch.

and thus need to be selected and dealt with particularly. Structural descriptor Speeded Up Robust Features (SURF) [33] is of the ability to detect important structure key points within images/videos. Since local synthesized artifacts are likely to appear around these key-points regions and draw greater attention from the human observers, SURF point detection and matching is used for sensitive regions selection, where geometric distortions are less acceptable for observers. It is worth mentioning that, SURF points matching could also compensate the global objects shifting artifacts and avoid over penalizing the acceptable uniform artifact as described and confirmed in our previous work [16].

2) Curve extraction based on patch segmentation: In order to check the magnitudes/amount of spatial-structural deformations after synthesis using elastic metric, curves need to be first extracted in an efficient way. After the process of local regions selection and matching, the SLIC superpixel approach [34] is then used to segment the matched patches P_{ori} and P_{syn} centering at matched key points. Then, the boundaries of the segmented super-pixels set SP_{ori} and SP_{syn} are extracted as the closed curves, which will be proceeded for latter comparison. Afterwards, the fast superpixels matching algorithm proposed in our previous work [16] is used to further obtain the set matched closed curves (C_{ori}, C_{syn}).

3) Spatial curves comparison using elastic metric: The amount of spatial structural deformation is computed by comparing each matched closed curves $(c^i_{ori}, c^j_{syn}) \subset (C_{ori}, C_{syn})$ obtained in the previous subsection using the elastic metric proposed in [21], [35]. A curve is first parameterized as c

with $k \in K$ as the parameter. It is defined as

$$c : K \rightarrow (x, y) \in \mathbb{R}^n, \tag{1}$$

where (x, y) represents the the coordinates of each point of curve. In general, $K = [0, 1]$. For closed curves, $K = \mathbb{S}^1$. The parameterized curve could be then represented using the Square-Root Velocity (SRV) function defined as $q : K \rightarrow (x, y) \in \mathbb{R}^n$, where

$$q(k) \equiv F(\dot{c}(k)) = \dot{c}(k)/\sqrt{\|\dot{c}(k)\|}, \tag{2}$$

where $\|\cdot\|$ is the Euclidean 2-norm in $\mathbb{R}^n$ and $\dot{c} = \frac{dc}{dk}$. The original curve could be derived reversely with the following equation

$$c(k) = \int_0^k q(s)\|q(s)\|ds. \tag{3}$$

Later, $\phi : K \rightarrow \mathbb{R}$, with $\phi(k) = ln(\|\dot{c}(k)\|)$ and $\theta : K \rightarrow \mathbb{S}^{n-1}$, with $\theta(k) = \dot{c}(k)/\|\dot{c}(k)\|$ are further defined by Srivastava *et al.* in [35] to quantify curve deformations. Based on these definition, a riemannian metric named 'Elastic Metric' D_{EM} on the tangent space τ of $\Phi \times \Theta$ could be then defined based on the computation of the following inner product:

$$\begin{aligned} D_{EM} &= \langle (u_1, v_1), (u_2, v_2) \rangle_{(\phi,\theta)} \\ &= a^2 \int_D u_1(k)u_2(k)e^{\phi(k)}dk + b^2 \int_D v_1(k)v_2(k)e^{\phi(k)}dk, \end{aligned} \tag{4}$$

where $\langle\cdot\rangle$ denotes the standard dot product in $\mathbb{R}^n$ and (u_1, v_1), $(u_2, v_2) \in \tau_{\phi,\theta}(\Phi \times \Theta)$. As explained in [21], [35], u_1 and u_2 in the first integral are variations of the log speed ϕ of the curves, while v_1 and v_2 in the second integral are the variations of the direction θ of the curves. The first and second integrals could be interpreted to measure the amount of 'stretching' and 'bending' correspondingly and a^2,b^2 are two parameters chosen to penalize these two types of deformations. To calculate Eq. (4) more efficiently, the SRV formulationin Eq. (2) are used and adjusted in terms of (ϕ, θ) by defining $q(k) = e^{\frac{1}{2}\phi(k)}\theta(k)$. Afterwards, the tangent vectors to $\mathbb{L}^2(K, \mathbb{R}^n)$ at q is obtained with $r = \frac{1}{2}e^{\frac{1}{2}\phi}u\theta + e^{\frac{1}{2}\phi}v$. For two elements r_1 and r_2 of $\tau_{\phi,\theta}(\Phi \times \Theta)$, computing the $\mathbb{L}^2$-metric (elastic metric) of them yields

$$\begin{aligned} D_{EM}(c^i_{ori}, c^j_{syn}) &= \langle r_1, r_2 \rangle \\ &= \int_K \langle \frac{1}{2}e^{\frac{1}{2}\phi}u_1\theta + e^{\frac{1}{2}\phi}v_1, \frac{1}{2}e^{\frac{1}{2}\phi}u_1\theta + e^{\frac{1}{2}\phi}v_2 \rangle dk \\ &= \int_K (\frac{1}{4}e^{\theta}u_1u_2 + e^{\theta}\langle v_1, v_2 \rangle)dk. \end{aligned} \tag{5}$$

4) Spatial Pooling: The D_{EM} calculates local elastic dissimilarity between each pair of matched closed curves from the reference and synthesized images based on region selection and elastic metric described in the previous section. As discussed in previous sections, human observers tend to perceive 'poor' regions than the 'good' ones within an image. For DIBR based synthesized images, the sensitive disoccluded regions are the 'poor' regions and should be penalized during the quality assessment.

As the curves are only extracted from the selected regions, where the annoying local distortions mainly appear, there is no need to apply other specific pooling strategies for pooling the elastic dissimilarity scores. Moreover, due to local regions selection, artifacts in local important disoccluded regions are penalized sufficiently, and at the same time, the global consistent artifacts are not over penalized. Hence, the final objective score is calculated by simply summing out all the elastic dissimilarities values, which is defined as

$$EM_{spa} = \sum D_{EM}(c^i_{ori}, c^j_{syn}), \tag{6}$$

where $(c^i_{ori}, c^j_{syn}) \subset (C_{ori}, C_{syn})$.

B. Quantify Temporal Structural Degradation

In this section, details of how to quantify the non-uniform temporal structure-related distortions are given. The framework is summarized in Fig. 5. As motion trajectory reveals important structural-motion information, distortions along motion trajectory are thus easier to be noticed by observers. Based on this fact, multi-scale trajectory representation is exploited in this work to quantify local structure related distortions that affect the quality of the synthesized sequence. In the proposed scheme, given one synthesized video V_{syn} and its reference video V_{ref}, they are firstly represented as a set of multi-scale trajectories T^s_{syn} and V^s_{ref} respectively (*i.e.* trajectory at different scales), where s indicates a certain scale. Considering the characteristics of DIBR based synthesis techniques, the neighborhoods around the trajectories could be considered as the candidates regions, where local non-uniform distortion may appear and severely degrade the quality of the entire sequence. With the multi-scale trajectory representation, spatial-temporal structure-related features, in the form of histograms, *i.e.*, H^s_{syn} and H^s_{ref}, along the trajectories are extracted. Finally, deformations of the object structures in the form of deformations of trajectories, *i.e.* T_{EM}, could be quantified using elastic metric with T^s_{syn} and T^s_{ref}, while the structural losses along trajectories, *i.e.* T_{SL}, could be quantified with the temporal-structure features/histograms H^s_{syn} and H^s_{ref}.

To check the performance of the proposed temporal structural distortion estimator, the obtained T_{EM}, T_{SL} are combined as a new video quality assessment metric, which is denoted as EM_{tem}. It has to be emphasized that, the final proposed EM-vQM is obtained by integrating T_{EM}, T_{SL}, and EM_{spa} using SVR as illustrated in Fig. 3 instead of combing EM_{spa} and EM_{tmp}. Details of the computation of spatial-temporal structural dissimilarity between a synthesized sequence and its reference, *i.e.*, T_{EM}, T_{SL}, are given below.

1) Multi-scale motion trajectory representation as spatial-temporal distortion regions selection: Dense motions trajectory, which is first proposed in [36] by Wang *et al.*, is first utilized in this work to represent free-viewpoint videos. It is a spatial-temporal representation for video with multi-scale dense trajectories and descriptors of structural-motion boundary along the trajectories.

After generating the multi-scales version of the video V with S spatial scales, feature points are sampled on each spatial scale $s \in \{1, ..., S\}$ with a sampling step W. In this

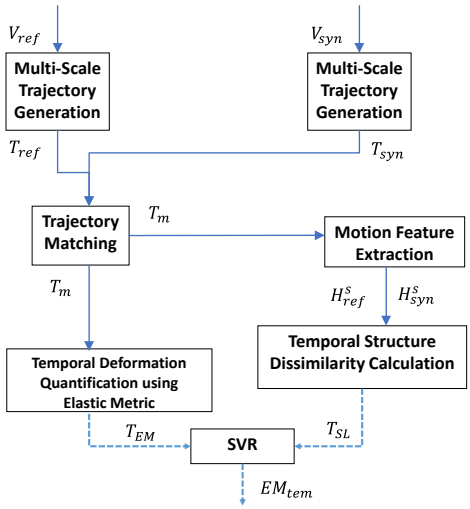


Fig. 5. Framework of temporal structural degradation quantification.

work, seven scales are considered. Each spatial scale increases with a factor of $1/\sqrt{2}$ as done in [36]. Considering that most of the local disturbing geometric distortions are commonly located around the boundaries of the objects (*i.e.*, the disoccluded regions) instead of homogeneous texture regions, points within regions that do not contain any structure are removed. Later, sampled points on each spatial scales are then tracked by using Large Displacement Optical Flow algorithm (LDOF) proposed in [37]. Then, each trajectory t_s obtained in a certain scale s could be represented as a sequence of points $(p_1, ..., p_{f_i}, ..., p_l)$ with a length of l (l is set as 15 in this paper). In t_s, p_{f_i} is a feature point at frame f_i, which is spatially-temporally related to features points in previous and latter frames, *i.e.*, $p_{f_{i-1}}$ and $p_{f_{i+1}}$, according to the calculated optical components. As human observers are more sensitive to moving structural regions, *e.g.*, moving objects, static trajectories that do not contain any motion are pruned. It is worth mentioning that the process of generating trajectories could be served as a proxy to select candidate sensitive regions, as local synthesized distortions that distribute around the moving objects attract most of the attention from the observers. Example is shown in Fig. 6, where it can be observed that most of the error regions have been covered by the detected motion trajectories.

2) Temporal-structure-related trajectory descriptor: In order to better quantify the changes of spatial-temporal structural information along trajectories due to synthesis process, with respect to the reference, three motion-structure related descriptors [37] are extracted for each trajectory. More specifically, they are Histogram of Oriented gradient (HOG) [38], Histogram of Optical Flow (HOF) [39] and Motion Boundary Histogram (MBH) [40] extracted within a spatial-temporal volume that is aligned with one trajectory T_s as illustrated in Fig. 7. MBH is computed with the the derivatives of both the horizontal and vertical elements of one optical flow, which further ends up into two histograms for each component as MBH_h and MBH_v normalized with L_2 norm. Therefore, for each trajectory at a scale s, four spatial-temporal structural histograms, including H^s_{HOG}, H^s_{HOF}, H^s_{MBHx} and H^s_{MBHy}, are obtained after feature extraction procedure.

Multi-scale motion trajectory representation as spatialtemporal distortion regions selection:

3) Temporal structure dissimilarity: After getting the trajectory representations along with the extracted features, trajectories at each scale in the synthesized and reference sequences are first matched according to the averaged horizontal and vertical coordinates of the trajectories. Only the matched trajectory pairs (t^s_{ori}, t^s_{syn}) in the matched trajectory set T_m would be maintained for latter deformation quantification and structure loss computation. To quantify temporal degradation by considering the two typical temporal distortions mentioned in Section I, two main aspects are taken into consideration.

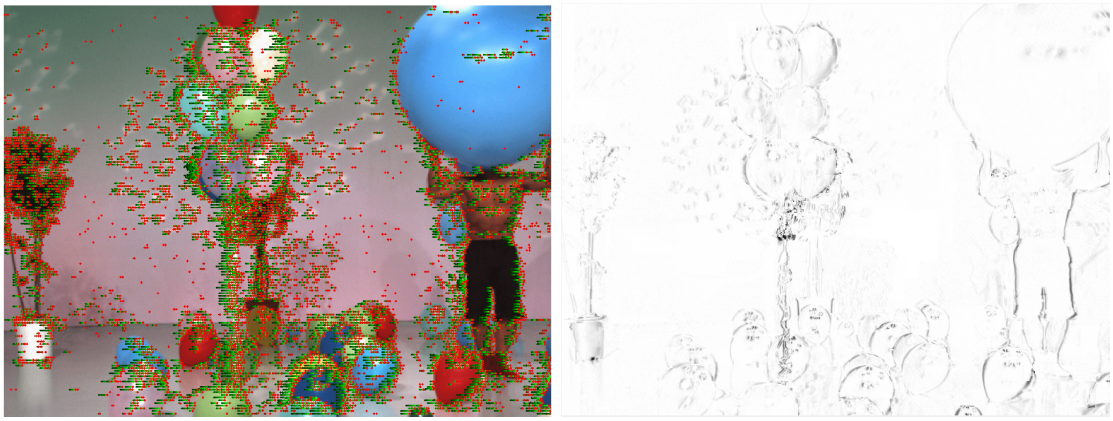
First, since temporal evolution of spatial local structure-related distortions might result in deformation of motion trajectories within the sequences, *e.g.*, the motion trajectory distributed along boundaries of foreground objects might fluctuate and result in changes of the shape of the trajectory. These changes of trajectories in term of global motion trajectory deformations could be quantified by using elastic metric described in Section III-A3. More specifically, the entire deformable changes of trajectories between the synthesized and their reference sequences on all the scales $T_{EM}(T_m)$ is defined by accumulating all the elastic errors between the trajectories:

$$T^s_{EM}(T_m) = \frac{\sum_{(t^s_{ori}, t^s_{syn}) \in T_m}^{N^s_t} D_{EM}(t^s_{ori}, t^s_{syn})}{N^s_t}, \quad (7)$$

where N^s_t is the number of matched trajectory pairs $(t^s_{ori}, t^s_{syn}) \in T_m$ at scale s. Since $T_{EM}(\cdot)$ compute the amount of deformations between trajectories, ideally, it is able to capture not only the temporal structure-related distortions within viewpoints (at one viewpoint position) but also the one among the viewpoints (smoothness of the transition among viewpoints).

As it has been pointed out in previous sections, structure-related distortions along the motion trajectories are the most disturbing temporal degradation, which could cause inconsistent transition of frames within and among viewpoints. Therefore, similar to the spatial elastic pooling stage (Section III-A4), here the temporal deformation errors between each pair of matched trajectories are simply summed up to get the temporal deformation score T_{EM} with Eq.(7). By doing so, temporal structure-related severe deformation could be well captured, while the global uniform distortions would not be over-penalized.

Second, to further quantify the non-contentiousness of transition of structure from one frame to another, structural statistical dissimilarities along trajectories are computed with the four extracted motion descriptors. More specifically, the temporal structural statistical loss T_{SL} is defined as the structural dissimilarity calculated based on computing distance between matched extracted features vectors set $(H^{i,s}_{ref}, H^{i,s}_{syn}) \in H_m$



(a) (b)

Fig. 6. Example of temporal sensitive regions selection: (a) Example of dense motion trajectory, where red points represents the key points in the current frame and the green lines connect the key points at the current frames with the ones in the previous frame. (b) Error map between frames extracted from the reference and synthesized views, where the darker color the more severe the distortions are within the regions.

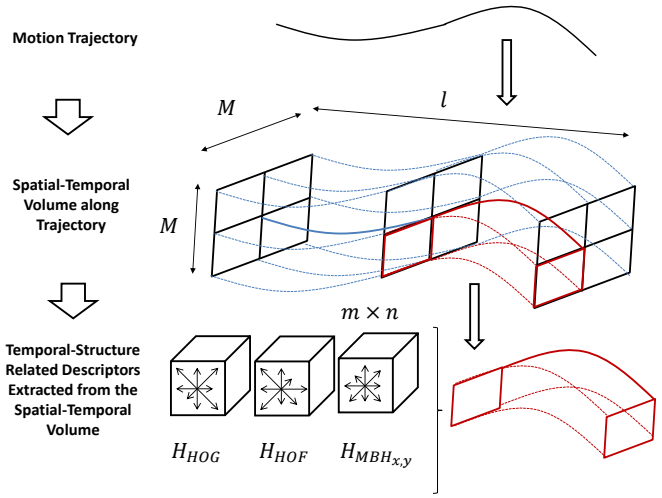


Fig. 7. Illustration of temporal-structure related trajectory descriptors extraction along motion trajectories.

from the matched trajectory set T_m at scale s:

$$T_{SL}^{ij}(H_m) = \frac{\sum_{(H_{ref}^{i,s}, H_{syn}^{i,s}) \in H_m}^{N_t^s} D_j(H_{ref}^{i,s}, H_{syn}^{i,s})}{N_t^s}, \quad (8)$$

where $H^{1,s} = H_{HOG}^s$, $H^{2,s} = H_{HOF}^s$, $H^{3,s} = H_{MBHx}^s$ and $H^{4,s} = H_{MBHy}^s$ indicates the four motion descriptors and D_j denotes one type of distance measures. In this paper, mainly four distance measures including D_1 using Jensen-Shannon divergence (JSD), D_2 = Euclidean distance, D_3 = Cosine distance and D_4 = Minkowski Summation, are considered for calculating the temporal structural dissimilarity along matched trajectory base on the corresponding feature $(H_{syn}^{i,s}, H_{ori}^{i,s})$.

C. Spatial-Temporal Scores Aggregation

Finally, in order to predict the final objective score, the Support Vector Machine Regression (SVR) is utilized to aggregate the calculated spatial elastic error EM_{spa}, temporal elastic error T_{EM} and the 16 temporal structural errors $T_{SL}^{ij}, i, j = 1, ..., 4$ at all scales with a linear kernel. As totally seven scales are considered in this paper, the final dimension of vector representing each free-viewpoint video is 120, *i.e.*, 7 scales $\times$ (16 dimension for T_{SL} + 1 dimension for T_{EM}) + 1 dimension for EM_{spa}. The SVR model training process is done according to [41], [42], [43] by employing a 1000-fold cross-validation. More specifically, for each dataset, it is randomly divided into 80% of the videos for training and 20% for testing, without overlap between them.

IV. Experimental Result

A. Datasets

The performance of the proposed metrics are evaluated on two datasets, including the IRCCyN/IVC DIBR Videos [3] and the Free-Viewpoint Synthesized Video [44] dataset. In general, the first dataset contains synthesized sequences at a certain viewpoint, while the second dataset contains sequences that mimic a time-free navigation among different viewpoints. The virtual scan paths of the sequences in the two datasets are illustrated in Fig. 8. These two datasets contain two types of synthesized temporal structure related distortionsas mentioned in Section I, *i.e.*, 1) temporal structure inconsistencies at one viewpoint position and 2) unsmooth structure transition among different viewpoints, respectively. Therefore, they are selected together to benchmark the quality metrics designed for synthesized views. Detailed introductions of the two datasets are given below.

IRCCyN/IVC DIBR Videos (IVC-DIBR) [3]: The IVC-DIBR database consists of 102 videos in resolution of 1024×768 generated with three multi-view plus depth contents. This database is designed for the evaluation of the reliability of DIBR algorithms by assessing the quality of the synthesized virtual views. Totally seven DIBR related algorithms, which are denoted as A1-A7 [2], [22], [45], [46], [47], [48], are used to obtain 4 new virtual viewpoints for each content. It contains only synthesis related spatial-temporal artifacts within viewpoints, as there are no navigation among different viewpoints to mimic free navigation. Apart from the 9 reference sequences

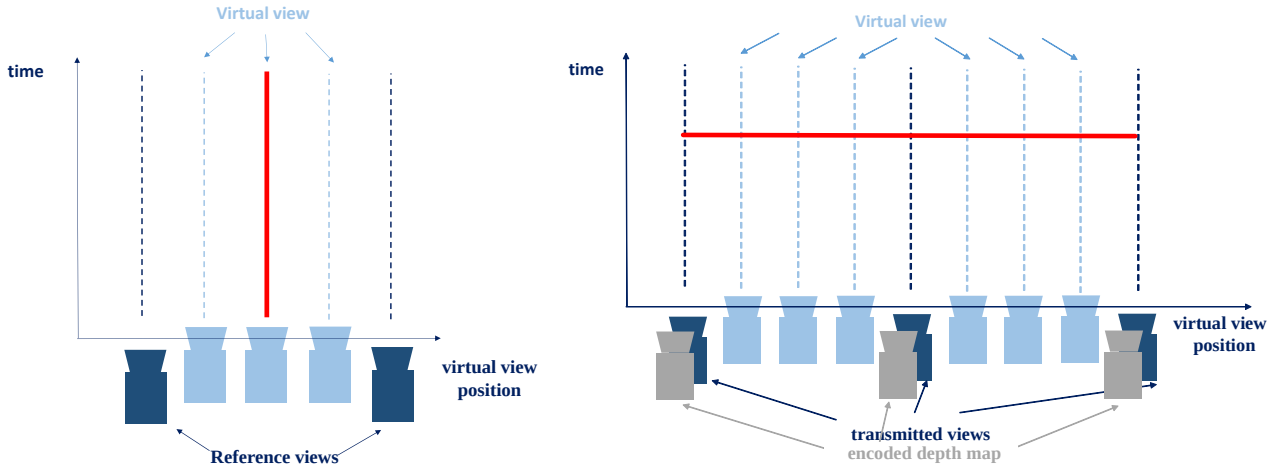


Fig. 8. Explanation of sequences' scan paths within the two datasets, where the red curves represent the virtual navigation scan-paths, the dark blue camera icons represent the reference views, the light blue camera icons represent the synthesized virtual views, and the gray camera icons represent the depth maps of the corresponding views. (a) IVC database: sequences contain temporal structure inconsistencies at one viewpoint position. (b) FVV database: sequences contain unsmooth structure transition among different viewpoint positions.

and the 84 synthesized virtual viewpoints, there are also 9 sequences that contain only traditional compression artifacts by encoding the texture (only videos, without compressing the depth maps) of the reference sequences. Since the purpose of this experiment is to verify the performance of metrics dedicated for capturing of related artifacts, these 9 sequences are excluded from the experiment to stress the capability of under-test metrics.

Free-Viewpoint Synthesized Video database (FVV) [44]: The FVV database is composed of 264 videos sequences in resolution of 1024×768 / 1920×1080 generated with six multi-view plus depth original sequences. The database is released to evaluate the impacts of depth map coding algorithms on the perceived quality of the synthesized views. Since depth maps are important during the DIBR based rendering process, seven codecs and three bitrates are adopted to encode the depth map for later synthesis process. These seven algorithms include (C1) 3D-HEVC [49], (C2) MVC [50], (C3) HM 6.1 [51], (C4) JPEG2000 [52], (C5) lossless-edge based codec [53], (C6) proposed in [54] using color frames' correlations, and (C7) Z-LAR-RP [55] using local information. After generating the synthesized viewpoints between the reference views with a certain configuration, sequences are then constructed with 100 key frames extracted from the synthesized viewpoints by navigating from one view to another from the left to the right and then turning around. As thus, sequences in this database contain only synthesis related spatial-temporal artifacts among viewpoints. Unlike [56], since blending is important in the process of DIBR based algorithm, the experiment is conducted on the entire database instead of excluding the one generated with blending mode.

B. Performance Evaluation Methodologies

It is emphasized in [11], [12], that commonly used image/video quality assessment metrics fail to quantify the synthesis related distortions. Therefore, the proposed metrics are only compared to the state of the art metrics designed for synthesized views in FTV scenarios as introduced in Section II-A, including: 1) 3DswIM [9], 2) MW-PSNR [11], 3) MP-PSNR [12], their reduced versions 4) MW-PSNR$_r$ [13], 5) MP-PSNR$_r$ [13], 6) the spatial-temporal activity distortion indicator (Liu-activity) proposed in [20], 7) the flicker distortion indicator (Liu-flicker) proposed in [20], 8) and the video quality metric (Liu-VQM) proposed in [20] that combines Liu-activity and Liu-flicker. The predicted score of IQM, *i.e.*, metrics 1)-5) and EM$_{spa}$, for one sequence are calculated by averaging the scores of all the frames.

Except for EM$_{tem}$ and EM-VQM (since SVR is already utilized), a non-linear logistic function recommended by [57] is employed to map the objective scores $OBJ(i)$ predicted by the i_{th} quality metric to the subjective quality scores before the performance evaluation. It is defined as

$$OBJ_{map}(i) = \frac{\beta_1}{1 + e^{-\beta_2 \times (OBJ(i) - \beta_3)}}. \qquad (9)$$

As described in Section III-C, support vector regression is employed to obtain the predicted quality scores of EM_{tem} and EM-VQM, and the performances are computed throughout a 1000-fold cross-validation as recommended in [41]. More specifically, the median and average Pearson linear Correlation Coefficient (PCC$_m$ and PCC$_a$), median and average Spearman rank order Correlation Coefficient (SCC$_m$ and SCC$_a$), as well as median and average Root Mean Squared Error (RMSE$_m$ and RMSE$_a$) between subjective and objective scores are reported across the 1000 runs for performance evaluation. For fair comparison, all the compared metrics are evaluated with the same 1000-fold cross validation, where PCC, SCC and RMSE are calculated with $OBJ_{map}(i)$ of the i_{th} compared metric computed on 20% test dataset for each fold.

Apart from the frequently used performance evaluation methodologies, in order to better evaluate the performance of different metrics the methodology proposed by Krasula *et al.* [58], [59] is also used. In their model, it is assumed that the capability of an objective metric depends its capabilities

TABLE I
PERFORMANCE COMPARISON OF THE PROPOSED METRIC WITH STATE-OF-THE-ART METRICS.

Dataset	IVC-DIBR						FVV					
Metric	**PCC**$_m$	**SCC**$_m$	**RMSE**$_m$	**PCC**$_a$	**SCC**$_a$	**RMSE**$_a$	**PCC**$_m$	**SCC**$_m$	**RMSE**$_m$	**PCC**$_a$	**SCC**$_a$	**RMSE**$_a$
Image Quality Assessment Metrics (IQM) Designed for Synthesized Views												
3Dswim	0.7188	0.6415	0.4394	0.6998	0.6224	0.4392	0.5587	0.5831	0.5289	0.5599	0.5786	0.5253
MW-PSNR	0.5586	0.4736	0.5085	0.5305	0.4342	0.5141	0.4762	0.3769	0.5620	0.4628	0.3633	0.5653
MW-PSNR$_r$	0.5801	0.5292	0.5019	0.5435	0.4726	0.5082	0.4751	0.3811	0.5620	0.4579	0.3684	0.5667
MP-PSNR	0.6148	0.5791	0.4807	0.5938	0.5359	0.4849	0.4932	0.3912	0.5549	0.4730	0.3696	0.5577
MP-PSNR$_r$	0.5693	0.5329	0.5116	0.5532	0.4907	0.5101	0.4750	0.3791	0.5661	0.4606	0.3667	0.5664
EM$_{spa}$	0.7200	0.6262	0.4409	0.6961	0.6059	0.4395	0.5589	0.5679	0.5345	0.5510	0.5635	0.5270
Video Quality Assessment Metrics (VQM) Designed for Synthesized Views												
Liu-activity	0.7595	0.6440	0.3978	0.7237	0.6190	0.4175	0.6413	0.6468	0.4891	0.6315	0.6159	0.4872
Liu-flicker	0.5561	0.4796	0.5192	0.5470	0.4670	0.5207	0.6465	0.6463	0.4871	0.6340	0.6297	0.4880
Liu-VQM	0.7316	0.6464	0.4188	0.6988	0.6308	0.4366	0.6676	0.6716	0.4843	0.6448	0.6233	0.4788
EM$_{tem}$	0.8201	0.8091	0.3021	0.7964	0.7836	0.3114	0.7756	0.7562	0.3868	0.7469	0.7464	0.4017
EM-VQM	**0.8257**	**0.8102**	**0.3008**	**0.8060**	**0.7914**	**0.3077**	**0.7794**	**0.7627**	**0.3778**	**0.7566**	**0.7545**	**0.3949**

TABLE II
STATISTIC SIGNIFICANCE RESULTS BASED ON THE 1000 TIMES CROSS PERFORMANCE EVALUATION. FOR SYMBOLS IN EACH ENTRY OF THE TABLE CORRESPOND TO IVC-DIBR AND FVV DATA SET IN ORDER, *i.e.*, IVC-DIBR\FVV. THE VALUE '1' INDICATES THE QUALITY METRIC IN THE ROW OUTPERFORM SIGNIFICANTLY THE ONE IN THE COLUMN, WHILE '-1' INDICATES THE OPPOSITE CASE, AND '0' INDICATES THAT THE TWO QUALITY METRICS PERFORM EQUIVALENTLY.

	3DS wIM	MW-PS NR$_r$	MW-PS NR$_f$	MP-PS NR$_r$	MP-PS NR$_f$	Liu-act ivity	Liu-fli cker	Liu-final	EM$_{spa}$	EM$_{tem}$	EM -VQA
3DSwIM	-	1\1	1\1	1\1	1\1	-1\-1	1\-1	0\-1	0\0	-1\-1	-1\-1
MW-PSNR$_f$	-1\-1	-	0\0	-1\0	-1\0	-1\-1	0\-1	-1\-1	-1\-1	-1\-1	-1\-1
MW-PSNR$_r$	-1\-1	0\0	-	-1\-1	0\0	-1\-1	0\-1	-1\-1	-1\-1	-1\-1	-1\-1
MP-PSNR$_f$	-1\-1	1\0	1\1	-	1\1	-1\-1	1\-1	-1\-1	-1\-1	-1\-1	-1\-1
MP-PSNR$_r$	-1\-1	1\0	0\0	-1\-1	-	-1\-1	0\-1	-1\-1	-1\-1	-1\-1	-1\-1
Liu-activity	1\1	1\1	1\1	1\1	1\1	-	1\0	1\0	1\1	-1\-1	-1\-1
Liu-flicker	-1\1	0\1	0\1	-1\1	0\1	-1\0	-	-1\0	-1\1	-1\-1	-1\-1
Liu-VQM	0\1	1\1	1\1	1\1	1\1	-1\0	1\0	-	0\1	-1\-1	-1\-1
EM$_{spa}$	0\0	1\1	1\1	1\1	1\1	-1\-1	1\-1	0\-1	-	-1\-1	-1\-1
EM$_{tem}$	1\1	1\1	1\1	1\1	1\1	1\1	1\1	1\1	1\1	-	-1\0
EM-VQM	1\1	1\1	1\1	1\1	1\1	1\1	1\1	1\1	1\1	1\0	-

of making reliable decisions about 1) when comparing two stimuli, whether they are qualitatively different and 2) if the are, which of them is of higher quality . The 'Krasula' model is based on determining the classification capabilities of the objective models considering 'Better or Worse' and 'Different or Similar' scenarios.

More specifically, the capability of one objective metric to distinguish similar from significantly different pairs and the capability to indicate one stimulus is better/worse than another could determined by employing the receiver operating characteristic (ROC) analysis. Then, the performance of the metric can be verified with the area under the ROC curve (AUC) for both the 'Better or Worse' and 'Different or Similar' analysis, *i.e.*, AUC$_{DS}$ and AUC$_{BW}$, the higher the AUC values the better metric in categorizing significantly different pairs from the similar ones as well as telling one stimulus is better/worse compared to another. (Readers are recommended to refer to [58], [59] for more detailed information.)

C. Performance Comparison Results

1) Performance evaluation using commonly used evaluation methodologies: Comprehensive performance evaluations of the proposed metrics are reported in this subsection. The 1000-fold cross validations results of the metrics on both the IVC-DIBR and FVV datasets are summarized in TABLE I

In general, according to the table, the proposed EM-VQM achieves the best performance on both the IVC-DIBR and the FVV datasets. It has a gain of 12.86% in PCC$_m$ values on IVC-DIBR dataset and a gain of 16.75% in PCC$_m$ values on IVC-DIBR dataset compared to the state-of-the-art video quality metric designed for synthesized videos, *i.e.*, Liu-VQM. It could also be observed from the table that the objective scores predicted by most of the image quality metrics have poor correlations with subjective scores, specially on the FVV dataset. Synthesis related temporal distortions are difficult for image quality metrics to capture. Interestingly, 1) the performance of Liu-Flicker, which quantifies the amount of temporal

flicker perform much better on the FVV dataset; 2) the overall performances of the quality models on FVV datasets are worse than the ones on the IVC-DIBR dataset. It could be indicated from these two observations that the second type of structure related temporal distortions in FTV scenarios, *i.e.*, the unsmooth transition among viewpoints, are more challenging. By comparing the performance of EM_{tem} and EM-VAM on the two datasets, it could be noticed that integrating EM_{spa} does not improve the performance significantly. EM_{tem} play the most important role in predicting the perceived quality.

The scatter plots of the tested quality metrics are illustrated in Fig. 9 and Fig. 10 respectively. For EM_{tem} and EM-VQM, the model that yields the median PCC values was used to plot the figures.

In general, the quality scores predicted by both EM_{tem} and EM-VQM are more consistent with DMOS compared to other image/video quality models, as most of the points shown in Fig. 9 are compactly distributed along the diagonal. It could be observed from Fig. 9 (a)-(d) that, the four point-wise PSNR based metrics tend to predict the same quality scores for sequences generated using synthesis algorithms A1. Sequences obtained using this algorithms mainly contain acceptable global shifting distortions. However, it is obvious that MP-PSNR, MP-PSNR$_r$, MW-PSNR, and MW-PSNR$_r$ over-penalize these distortions. From Fig. 9 (g), most of the sequences in IVC-DIBR dataset are predicted with scores in a range of $[2, 3]$. That is because the magnitudes of flicker distortions within sequences are similar, the Liu-flicker metric could not well quantify the temporal structure inconsistencies caused by synthesis algorithms. This also explains why the performance of Liu-VQM does not outperform Liu-activity as shown in TABLE I (since Liu-VQM is composed of Liu-activity and Liu-flicker).

According to Fig. 10, the objective scores predicted by EM_tem and EM-VQM are better aligned with the DMOS on the FVV dataset compared to the other metrics. By observing Fig. 10 (a)-(i), it could be noticed that the points are gathered as a cluster in an objective score range of $[2, 3]$, and they are poorly in predicting sequences with high/bad quality.

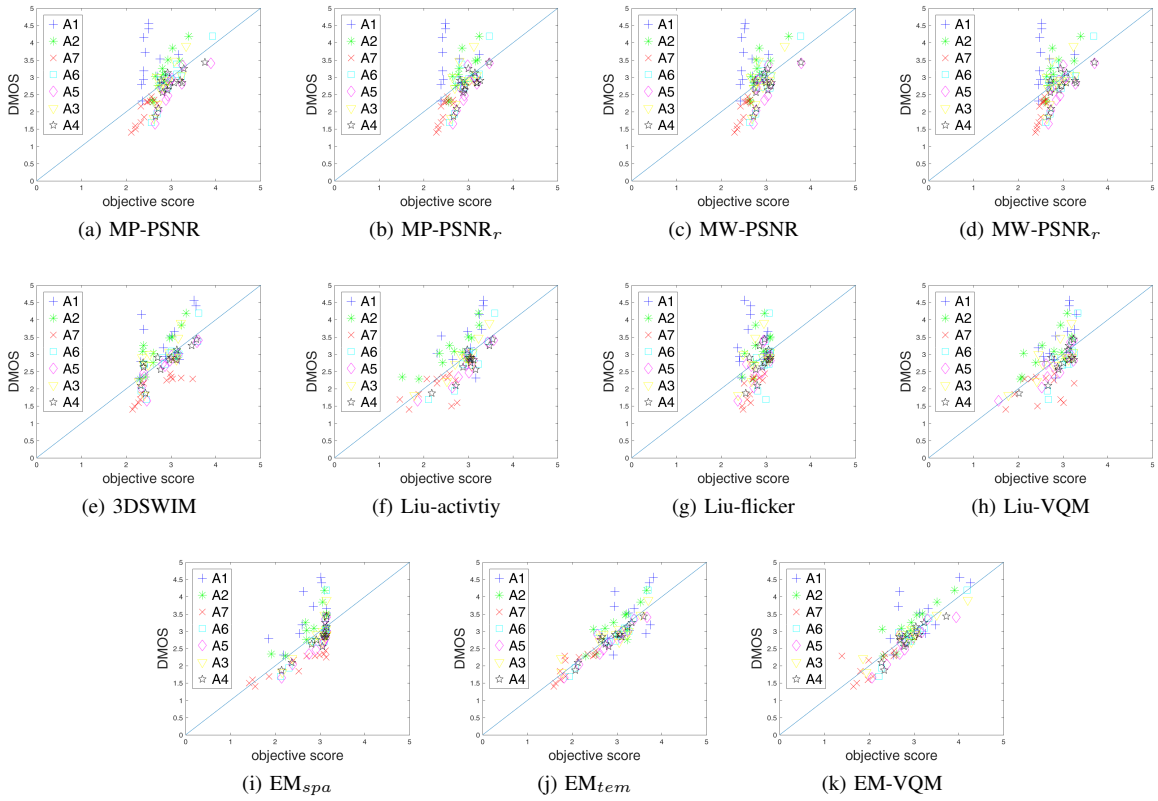
2) Statistical significant test: To examine the significance of the performances between each two tested quality metrics, student's t-test is conducted. More specifically, the 1000-fold PCC values obtained during the cross performance evaluation described above for each tested metric are used as input for t-test. The results are concluded in Table II with a significance level of 0.05, where '1' represents that the performance of the under-test metric in row outperforms the one in column significantly, '-1' represents the opposite situation and '0' represents that there is no significant difference. According to the table, both the proposed EM_{tem} and EM-VQM significantly outperform all the other metrics on the two datasets. In addition, EM-VQM is significantly superior to EM_{tem} on IVC-DIBR dataset, but not on the FVV dataset. It reveals the fact that EM_{spa} plays an important role in predicting the perceived quality score on IVC-DIBR dataset but not on the FVV dataset. Temporal distortions among viewpoints are more challenging for existing metrics, and model that considers temporal structure due to views' switch, *e.g.*, EM_{tem}, should be considered.

TABLE III
PERFORMANCE COMPARISON OF THE PROPOSED METRIC WITH STATE-OF-THE-ART METRICS USING THE KRASULA MODEL

Dataset	IVC-DIBR		FVV	
Metric	A_{DS}	A_{BW}	A_{DS}	A_{BW}
IQM				
3Dswim	0.548	0.830	0.541	0.775
MW-PSNR	0.537	0.704	0.498	0.641
MW-PSNR$_r$	0.522	0.717	0.499	0.642
MP-PSNR	0.531	0.754	0.499	0.648
MP-PSNR$_r$	0.521	0.739	0.501	0.648
EM_{spa}	0.494	0.830	0.497	0.761
VQM				
Liu-activity	0.547	0.701	0.523	0.784
Liu-flicker	0.473	0.677	0.514	0.793
Liu-VQM	0.507	0.704	0.525	0.795
EM_{tem}	0.593	0.934	0.542	0.919
EM-VQM	**0.602**	**0.946**	**0.543**	**0.921**

3) Performance evaluation using Krasula methodology: The performance results of the metrics using the evaluation methodologies proposed by Krasula *et al.* are reported in TABLE III. For EM_{tem} and EM-VQM, the SVR model that obtains the median PCC values during the 1000-folds cross validation is utilized for the calculation of AUC_{DS} and AUC_{BW} as introduced in Section IV-B. It could be observed from the table that the proposed EM-VQM obtains the best performances, in terms of AUC_{DS} and AUC_{BW} values among the compared metrics. Among the compared metrics, the EM-VQM is the best in 1) distinguishing pairs of stimuli that are of similar/significantly different quality; 2) indicating sequences are of better/worse quality than others.

4) Benchmarking synthesis algorithms and depth map codecs: As one of the most important functionalities of an image/video quality metric is to benchmark the performance of the system considering relative techniques. In FTV system, depth maps codecs and synthesis algorithms are two of the most important techniques. More reliable synthesis algorithms and codecs could provide better free-viewpoint videos. From this point of view, the performances of using the objective quality metrics for benchmarking the synthesis algorithms, *i.e.*, A1-A7, in IVC-DIBR dataset and the seven depth map codecs, *i.e.*, C1-C7, in FVV dataset are also evaluated. More specifically, the DMOS values of sequences obtained using each synthesis algorithm/depth map codes are averaged to compute the ground truth ranking. Similarly, the predicted ranking of A1-A7/C1-C7 using the image/video quality assessment metrics are also obtained based on the mean predicted scores of each synthesis algorithm/depth map codec. The results are shown in TABLE IV. Comparing the ranking of A1-A7 predicted by the quality metrics with the one predicted by DMOS, the proposed EM_{tem} and EM-VQM achieve the most consistent rankings. For EM-VQM, only A6 is shifted two positions forward and the positions of A4, A5 are switched. For EM_{tem}, only A6 is shifted one position forward and the positions of A4, A5

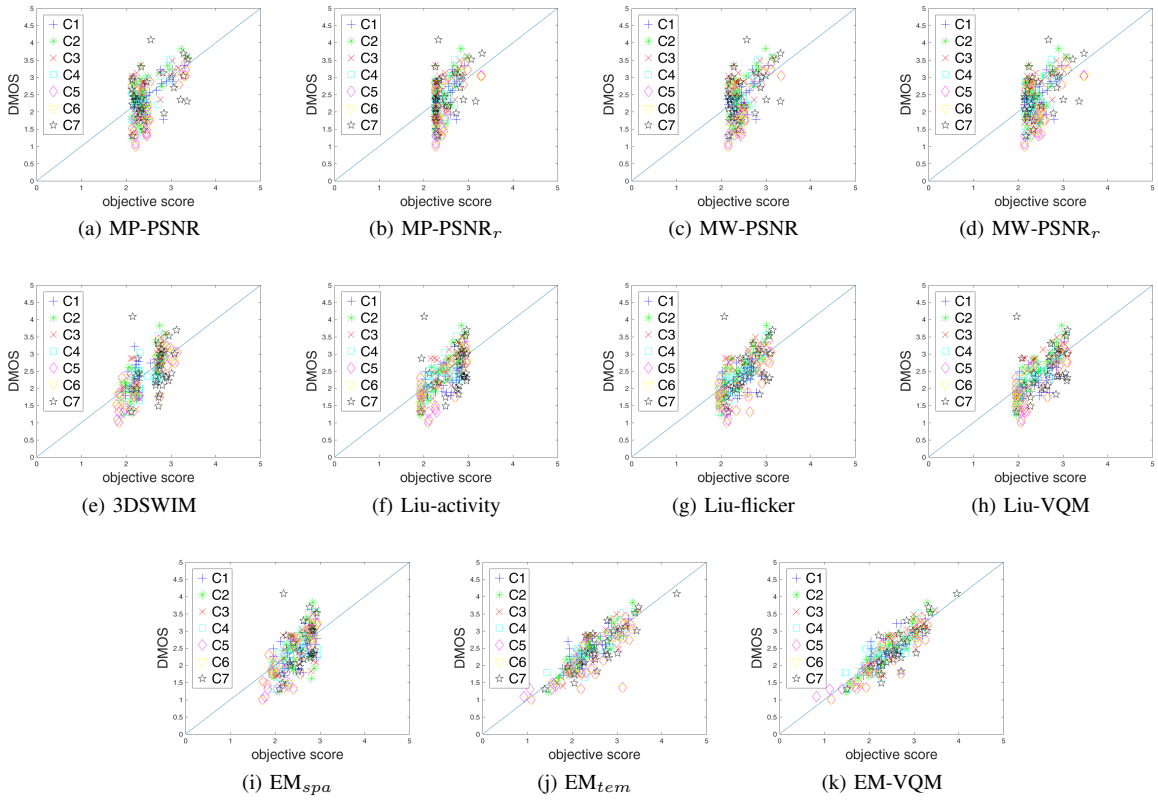


(a) MP-PSNR (b) MP-PSNR$_r$ (c) MW-PSNR (d) MW-PSNR$_r$

(e) 3DSWIM (f) Liu-activtiy (g) Liu-flicker (h) Liu-VQM

(i) EM$_{spa}$ (j) EM$_{tem}$ (k) EM-VQM

Fig. 9. Scatter plots of all quality metrics' scores versus DMOS on IVC-DIBR database [3]. Sequences that generated with different synthesis algorithms (*i.e.*, A1-A7) are labeled with different shapes and colors (better seen in color).



(a) MP-PSNR (b) MP-PSNR$_r$ (c) MW-PSNR (d) MW-PSNR$_r$

(e) 3DSWIM (f) Liu-activity (g) Liu-flicker (h) Liu-VQM

(i) EM$_{spa}$ (j) EM$_{tem}$ (k) EM-VQM

Fig. 10. Scatter plots of all quality metrics' scores versus DMOS on FVV database [44]. Sequences that generated with different depth coding algorithms (*i.e.*, C1-C7) are labeled with different shapes and colors (better seen in color).

TABLE IV
RANKING OF THE SEVEN SYNTHESIS ALGORITHMS IN IVC-DIBR DATASET AND THE SEVEN DEPTH MAP CODECS IN FVV DATASET RANKED WITH RESPECT TO THE DMOS VALUES AND PREDICTED OBJECTIVE SCORES CALCULATED WITH THE IMAGE/VIDEO QUALITY METRICS DESIGNED FOR SYNTHESIZED VIEWS. FROM LEFT TO RIGHT, THE RANKING OF THE SYNTHESIS ALGORITHMS/DEPTH MAP CODECS DECREASES.

Dataset	IVC-DIBR							FVV						
DMOS	A1	A2	A3	A6	A4	A5	A7	C7	C6	C3	C2	C1	C4	C5
3DSWIM	A6	A5	A1	A4	A2	A3	A7	C7	C3	C2	C6	C4	C5	C1
MW-PSNR	A4	A5	A6	A3	A2	A1	A7	C6	C7	C5	C1	C3	C2	C4
MW-PSNR$_r$	A4	A5	A6	A3	A2	A1	A7	C6	C7	C5	C1	C3	C2	C4
MP-PSNR	A4	A5	A6	A3	A2	A1	A7	C6	C7	C5	C1	C3	C2	C4
MP-PSNR$_r$	A4	A5	A6	A3	A2	A1	A7	C7	C6	C5	C1	C3	C2	C4
Liu-activity	A4	A6	A1	A5	A3	A2	A7	C7	C1	C3	C2	C6	C4	C5
Liu-flicker	A6	A4	A5	A2	A3	A7	A1	C7	C1	C6	C3	C5	C2	C4
Liu-VQM	A6	A4	A1	A5	A3	A2	A7	C7	C1	C3	C6	C2	C5	C4
EM$_{spa}$	A4	A3	A6	A5	A2	A1	A7	C7	C3	C1	C2	C4	C6	C5
EM$_{tem}$	A1	A2	A6	A4	A3	A5	A7	C7	C6	C3	C2	C1	C5	C4
EM-VQM	A1	A6	A2	A3	A5	A4	A7	C7	C3	C2	C6	C1	C4	C5

are switched. For most of the image quality metrics, the poor performing algorithms are ranked with higher positions, while the better performing ones are ranked lower. For the state-of-the-art DIBR-oriented video quality metric Liu-VQM, except for A7, the rest are all inconsistent with the ground truth. Comparing the ranking of C1-C7 indicated by the objective models with the ground truth, the two proposed metrics provide again the most consistent rankings. For EM$_{tem}$, only the positions of C4 and C5 are switched. For EM-VQM, the position of C6 is shifted slightly forward. Liu-activity also only incorrectly switches the position of C1 and C6, but the distance between these two codecs is far according to the ground truth. All the other compared metrics fail to provide a more correct ranking with less than three inconsistent rankings.

V. CONCLUSIONS

To evaluate the quality of FVV in FTV system, in this work, we present a multi-scale motion trajectory based video quality assessment metric by quantifying the elastic changes. Specifically, to quantify the two dominant temporal structure-related distortions contained in nowadays synthesized views, *i.e.*, the object deformation observed at a certain viewpoint and the structure inconsistencies observed during view switches, we calculated the amount of 1) deformation changes of spatial structure, 2) deformation changes of motion trajectories, and 3) the statistical change of motion-structure descriptors along the trajectories within reference and the synthesized videos. Then they are aggregated by SVR to predict the perceived quality. Experiments have been conducted on two databases which contain the two aforementioned temporal structure-related distortions. The results show that the proposed EM-VQM is superior to the state-of-the-art video quality metrics designed for FVV.

REFERENCES

[1] C. Fehn, R. De La Barre, and S. Pastoor, "Interactive 3-dtv-concepts and key technologies," *Proceedings of the IEEE*, vol. 94, no. 3, pp. 524–538, 2006.

[2] C. Fehn, "Depth-image-based rendering (dibr), compression, and transmission for a new approach on 3d-tv," in *Electronic Imaging 2004*. International Society for Optics and Photonics, 2004, pp. 93–104.

[3] E. Bosc, R. Pépion, P. Le Callet, M. Köppel, P. Ndjiki-Nya, L. Morin, and M. Pressigout, "Perceived quality of dibr-based synthesized views," in *Applications of Digital Image Processing XXXIV*, vol. 8135. International Society for Optics and Photonics, 2011, p. 81350I.

[4] 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 *Image Processing, 2007. ICIP 2007. IEEE International Conference on*, vol. 2. IEEE, 2007, pp. II–169.

[5] K. Gu, V. Jakhetiya, J.-F. Qiao, X. Li, W. Lin, and D. Thalmann, "Model-based referenceless quality metric of 3d synthesized images using local image description," *IEEE Transactions on Image Processing*, 2017.

[6] S. Ling, J. Gutiérrez, K. Gu, and P. Le Callet, "Prediction of the influence of navigation scan-path on perceived quality of free-viewpoint videos," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 2019.

[7] E. Bosc, R. Pepion, P. Le Callet, M. Koppel, P. Ndjiki-Nya, M. Pressigout, and L. Morin, "Towards a new quality metric for 3-d synthesized view assessment," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 7, pp. 1332–1343, 2011.

[8] P.-H. Conze, P. Robert, and L. Morin, "Objective view synthesis quality assessment," in *IS&T/SPIE Electronic Imaging*. International Society for Optics and Photonics, 2012, pp. 82 881M–82 881M.

[9] F. Battisti, E. Bosc, M. Carli, P. Le Callet, and S. Perugia, "Objective image quality assessment of 3d synthesized views," *Signal Processing: Image Communication*, vol. 30, pp. 78–88, 2015.

[10] C.-T. Tsai and H.-M. Hang, "Quality assessment of 3d synthesized views with depth map distortion," in *Visual Communications and Image Processing (VCIP), 2013*. IEEE, 2013, pp. 1–6.

[11] D. Sandić-Stanković, D. Kukolj, and P. Le Callet, "Dibr synthesized image quality assessment based on morphological wavelets," in *Quality of Multimedia Experience (QoMEX), 2015 Seventh International Workshop on*. IEEE, 2015, pp. 1–6.

[12] D. Sandic-Stankovic, D. Kukolj, and P. Le Callet, "Dibr synthesized image quality assessment based on morphological pyramids," in *2015 3DTV-Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON)*. IEEE, 2015, pp. 1–4.

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

[14] L. Li, Y. Zhou, K. Gu, W. Lin, and S. Wang, "Quality assessment of dibr-synthesized images by measuring local geometric distortions and global sharpness," *IEEE Transactions on Multimedia*, vol. 20, no. 4, pp. 914–926, 2018.

[15] S. Tian, L. Zhang, L. Morin, and O. Déforges, "Niqsv+: A no-reference synthesized view quality assessment metric," *IEEE Transactions on Image Processing*, vol. 27, no. 4, pp. 1652–1664, 2018.

[16] S. Ling and P. Le Callet, "Image quality assessment for dibr synthesized views using elastic metric," in *Proceedings of the 2017 ACM on Multimedia Conference*. ACM, 2017, pp. 1157–1163.

[17] A. K. Moorthy and A. C. Bovik, "Visual importance pooling for image quality assessment," *IEEE journal of selected topics in signal processing*, vol. 3, no. 2, pp. 193–201, 2009.

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

[19] E. Ekmekcioglu, S. Worrall, D. De Silva, A. Fernando, and A. M. Kondoz, "Depth based perceptual quality assessment for synthesised camera viewpoints," in *International Conference on User Centric Media*. Springer, 2010, pp. 76–83.

[20] 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 Transactions on Image Processing*, vol. 24, no. 12, pp. 4847–4861, 2015.

[21] W. Mio, A. Srivastava, and S. Joshi, "On shape of plane elastic curves," *International Journal of Computer Vision*, vol. 73, no. 3, pp. 307–324, 2007.

[22] A. Telea, "An image inpainting technique based on the fast marching method," *Journal of graphics tools*, vol. 9, no. 1, pp. 23–34, 2004.

[23] P. Ndjiki-Nya, M. Köppel, D. Doshkov, H. Lakshman, P. Merkle, K. Müller, and T. Wiegand, "Depth image based rendering with advanced texture synthesis," in *Multimedia and Expo (ICME), 2010 IEEE International Conference on*. IEEE, 2010, pp. 424–429.

[24] B. A. Wandell, *Foundations of vision*. sinauer Associates Sunderland, MA, 1995, vol. 8.

[25] L. Itti and C. Koch, "Computational modelling of visual attention," *Nature reviews neuroscience*, vol. 2, no. 3, p. 194, 2001.

[26] K. Seshadrinathan and A. C. Bovik, "Motion tuned spatio-temporal quality assessment of natural videos," *IEEE transactions on image processing*, vol. 19, no. 2, pp. 335–350, 2010.

[27] J. T. Todd, "Visual information about moving objects." *Journal of Experimental Psychology: Human Perception and Performance*, vol. 7, no. 4, p. 795, 1981.

[28] Z. Wang and Q. Li, "Video quality assessment using a statistical model of human visual speed perception," *JOSA A*, vol. 24, no. 12, pp. B61–B69, 2007.

[29] M. Barkowsky, J. Bialkowski, B. Eskofier, R. Bitto, and A. Kaup, "Temporal trajectory aware video quality measure," *IEEE Journal of Selected Topics in Signal Processing*, vol. 3, no. 2, pp. 266–279, 2009.

[30] B. Krekelberg and M. Lappe, "Temporal recruitment along the trajectory of moving objects and the perception of position," *Vision research*, vol. 39, no. 16, pp. 2669–2679, 1999.

[31] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Signals, Systems and Computers, 2004. Conference Record of the Thirty-Seventh Asilomar Conference on*, vol. 2. Ieee, 2003, pp. 1398–1402.

[32] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–8.

[33] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," in *European conference on computer vision*. Springer, 2006, pp. 404–417.

[34] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 11, pp. 2274–2282, 2012.

[35] A. Srivastava, E. Klassen, S. H. Joshi, and I. H. Jermyn, "Shape analysis of elastic curves in euclidean spaces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 7, pp. 1415–1428, 2011.

[36] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *International journal of computer vision*, vol. 103, no. 1, pp. 60–79, 2013.

[37] T. Brox and J. Malik, "Large displacement optical flow: descriptor matching in variational motion estimation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 3, pp. 500–513, 2011.

[38] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," in *BMVC 2009-British Machine Vision Conference*. BMVA Press, 2009, pp. 124–1.

[39] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–8.

[40] N. Dalal, B. Triggs, and C. Schmid, "Human detection using oriented histograms of flow and appearance," in *European conference on computer vision*. Springer, 2006, pp. 428–441.

[41] P. Gastaldo, R. Zunino, and J. Redi, "Supporting visual quality assessment with machine learning," *EURASIP Journal on Image and Video Processing*, vol. 2013, no. 1, p. 54, 2013.

[42] M. Narwaria, "Toward better statistical validation of machine learning-based multimedia quality estimators," *IEEE Transactions on Broadcasting*, 2018.

[43] E. Siahaan, A. Hanjalic, and J. A. Redi, "Augmenting blind image quality assessment using image semantics," in *Multimedia (ISM), 2016 IEEE International Symposium on*. IEEE, 2016, pp. 307–312.

[44] 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 *Quality of Multimedia Experience (QoMEX), 2013 Fifth International Workshop on*. IEEE, 2013, pp. 100–105.

[45] Y. Mori, N. Fukushima, T. Yendo, T. Fujii, and M. Tanimoto, "View generation with 3d warping using depth information for ftv," *Signal Processing: Image Communication*, vol. 24, no. 1, pp. 65–72, 2009.

[46] K. Mueller, A. Smolic, K. Dix, P. Merkle, P. Kauff, and T. Wiegand, "View synthesis for advanced 3d video systems," *EURASIP Journal on Image and Video Processing*, vol. 2008, no. 1, pp. 1–11, 2009.

[47] P. Ndjiki-Nya, M. Koppel, D. Doshkov, H. Lakshman, P. Merkle, K. Muller, and T. Wiegand, "Depth image-based rendering with advanced texture synthesis for 3-d video," *IEEE Transactions on Multimedia*, vol. 13, no. 3, pp. 453–465, 2011.

[48] M. Köppel, P. Ndjiki-Nya, D. Doshkov, H. Lakshman, P. Merkle, K. Müller, and T. Wiegand, "Temporally consistent handling of disocclusions with texture synthesis for depth-image-based rendering," in *2010 IEEE International Conference on Image Processing*. IEEE, 2010, pp. 1809–1812.

[49] "3d-htm," http://hevc.hhi.fraunhofer.de/.

[50] "Jm," http://iphome.hhi.de/suehring/tml/.

[51] "Hm," http://hevc.hhi.fraunhofer.de/.

[52] "Kakadu," http://www.kakadusoftware.com/.

[53] J. Gautier, O. Le Meur, and C. Guillemot, "Efficient depth map compression based on lossless edge coding and diffusion," in *Picture Coding Symposium (PCS), 2012*. IEEE, 2012, pp. 81–84.

[54] F. Pasteau, C. Strauss, M. Babel, O. Déforges, and L. Bédat, "Adaptive color decorrelation for predictive image codecs," in *Signal Processing Conference, 2011 19th European*. IEEE, 2011, pp. 1100–1104.

[55] E. Bosc, "Compression of multi-view-plus-depth (mvd) data: from perceived quality analysis to mvd coding tools designing," Ph.D. dissertation, INSA de Rennes, 2012.

[56] D. Sandić-Stanković, F. Battisti, D. Kukolj, P. Le Callet, and M. Carli, "Free viewpoint video quality assessment based on morphological multiscale metrics," in *Quality of Multimedia Experience (QoMEX), 2016 Eighth International Conference on*. IEEE, 2016, pp. 1–6.

[57] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, "Study of subjective and objective quality assessment of video," *IEEE transactions on image processing*, vol. 19, no. 6, pp. 1427–1441, 2010.

[58] 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 *Quality of Multimedia Experience (QoMEX), 2016 Eighth International Conference on*. IEEE, 2016, pp. 1–6.

[59] P. Hanhart, L. Krasula, P. Le Callet, and T. Ebrahimi, "How to benchmark objective quality metrics from paired comparison data?" in *Quality of Multimedia Experience (QoMEX), 2016 Eighth International Conference on*. Ieee, 2016, pp. 1–6.