# *Objective Quality Assessment of 3D Synthesized Views*

*Luís Miguel Domingos Nunes*
*Electrical and Computer Engineering Department*
Instituto Superior Técnico
Lisbon, Portugal
luismdnunes@tecnico.ulisboa.pt

***Abstract* —***With the explosion of digital technologies new models of visual representation have also emerged, from stereo to multiview video and for several applications such as 3D sports broadcasting, movie post-processing, etc. An interesting and promising representation format is the multiview plus depth where the distance from the camera to the objects can be acquired (referred as depth map) besides the texture for several views. This format allows to synthesize views at the decoder, avoiding the need to acquire, code and transmit a large number of views. At the decoder, the view synthesis techniques create novel views from a set of neighbouring views. Unfortunately, this process can create several types of geometric distortions in the synthesized view. Thus, to monitor the quality of experience for the end-user or even drive optimization of the encoding and transmission process (e.g. decide which views must be sent), automatic objective quality metrics need to be pursued. In this paper, a full-reference objective video quality metric is presented. This metric works on the spatiotemporal domain and address the flickering and 2D spatial distortions. In order to assess the metrics performance, a relevant synthesized video quality database was used.*

*The proposed video quality assessment metric outperforms relevant state-of-the-art 2D and 3D objective quality metrics. Also, it was shown that using a pixelwise just-noticeable difference model for flickering distortion has improved the metric linear correlation, but it lowered the monotonic correlation.*

***Keywords— Quality, Quality of Experience, 3D, 3D-HEVC, Synthesized Views, Synthesized Views Quality.***

## I. INTRODUCTION

The development of communication technologies has always played a key role in Human evolution due to its capability to exchange experiences amongst individuals, societies and cultures. However, the real world is not 2D but rather 3D, and thus visual representation naturally evolved towards 3D data. To provide richer 3D experiences, autostereoscopic displays use a larger number of scene views allowing the viewer to see different views and thus slightly different perspectives of the scene, as it takes different positions in front of the screen, thus offering also motion parallax. To avoid the acquisition and transmission of all views while still offering smooth motion parallax, a new representation approach was developed, known as view synthesis. View synthesis opens the possibility to interpolate as many as required virtual views at the receiver from one (or more) decoded views. The increased number of views and the increasing resolution of each view required the development of highly efficient compression and decompression tools, so-called codecs. Codecs can be segregated into two categories: lossless or lossy; meaning that either they exploit only data redundancy, or also exploit data perceivable irrelevancy. The key to the high performance of 3D codecs is the exploitation of the temporal and inter-view correlations with efficient tools that are able to capture the similarities in time and space together. However, whatever the coding and synthesis solutions, at the end of the day, quality has to be assessed to validate the developed solutions. In fact, quality assessment is paramount in evaluating the performance of video capture, compression and transmission steps. An accurate video quality assessment will provide information to optimize the overall system performance, and may lead to an optimization of the Quality of Experience (QoE) for the end users, which means that perceivable quality can be enhanced while spending the same bitrate.

This papers' goal is to present a solution to this paradigm of 3D synthesized views video quality assessment, while reviewing some of the most relevant objective quality assessment metrics, depth estimation and view synthesis processes. To be able to evaluate this solutions' performance, a 3D video synthesized views database was used; therefore, an overview of 3D image and video synthesized views databases had been added, as well as an overview of the correlation metrics used in this process.

## II. VIEW SYNTHESIS

Image-based modelling and rendering (IBMR) is a field of study in both computer graphics and computer vision areas. IBMR over the past recent years began to gain considerable attention in the scientific community, as emerging techniques allowed to generate far realistic synthesized images. View synthesis can be found as a part of the studies provided in this field, notably in the techniques which involve the synthesis of a nth view by only exploiting data from distinct perspectives of the same scene.

### *A. Depth Estimation*

The majority of state-of-the-art depth estimation techniques use 2D Markov random fields as a solution to derive depth maps from texture data, where each node is defined by all possible disparities and conforming probabilities. These probabilities are represented as log-probabilities and are scored by a sum of two functions, namely similar cost and smooth cost. In this way, the

disparities are computed and consequently a depth map is generated, notably an estimated (not acquired) depth map.

### B. *View Synthesis Reference Software (1-D Mode)*

This type of method is often called DIBR since it uses texture and depth maps to synthesize additional texture views. The synthesized views are derived out of correlation between the different perspectives by means of interpolation tools. As illustrated in Figure 1, the synthesis process includes the following steps:

1) *Depth mapping to target viewpoint*: First, depth maps are mapped into the desired viewpoint using appropriate camera information;

2) *Depth filtering by median filter*: The mapped depth maps are processed through an arithmetic mean filter to fill the smaller holes caused by rounding operations;

3) *Texture mapping*: Textures from left and right views are warped into corresponding left and right targeted viewpoints using the remapped depth maps;

4) *Hole filling*: Any holes in a new right mapped view are filled with information from the left mapped view, and vice-versa;

5) *Blending*: The desired texture view is then obtained by blending the two previously mapped texture views, depending on a factor related to their distances to the desired viewpoint position;

6) *Inpainting*: Finally, the holes within the synthesized frame are filled by an inpainting algorithm.

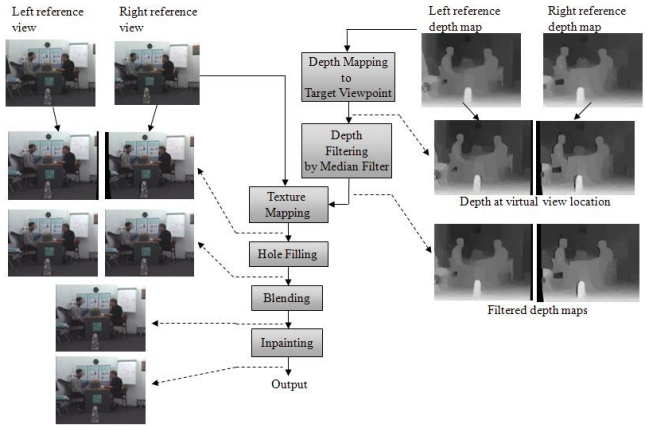



Figure 1 –*Simplified view synthesis architecture*.

## III. IMAGE AND VIDEO QUALITY ASSESSMENT

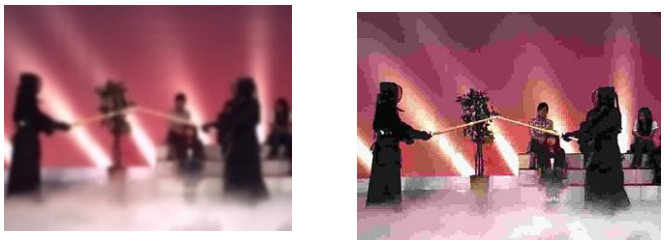
There are two main types of image and video quality assessment methodologies, notably subjective protocols and objective metrics. Subjective protocols consist in experimental processes that follow some recommended procedure, e.g. as defined in ITU-R Rec.BT.500 [1] and ITU-T Rec.P.910 [2]. These experimental processes are performed by asking human subjects to score the quality of some image and video content after its visualization. Subjective experiments are more reliable than objective metrics as they correspond to direct quality assessment based on Human quality perception. However, subjective experiments are far more resource demanding and slow, as a number of detailed procedures and requirements must be accurately met to produce trustworthy and statistically relevant results. Thus, reliable objective quality assessment metrics are of paramount importance to reduce the complexity and speed of the quality assessment process; naturally, to be useful, objective quality scores must correlate as much as possible with the corresponding subjective scores. In general, objective video quality assessment metrics may be classified into three major types, namely:

- *Full Reference*: This type of metrics directly compares the processed/decoded/synthesized view, also known as test sequence, with the corresponding original view which is taken as reference.
- *Reduced Reference*: This type of metrics compares features from the test sequence with features from the corresponding original view which are taken as reference.
- *No Reference*: This type of metrics directly assesses the test sequence without a reference, as no reference is available to make any comparisons.

### A. *Image and Video Distortions (2D and 3D)*

2D image-based distortions have been extensively studied throughout the years, leading to some widely commonly known distortion artefacts, notably:

- *Blurring Effect*: Refers to the blurry look that an image may display, see Figure 2 a); this look may result from over-filtering the texture samples and may be found in several applications; for example, it is rather common to find the blurring effect in JPEG 2000 overly compressed images.
- *Blocking Effect*: Refers to the blocking aspect that an image may exhibit, see Figure 2 b); it typically results from a highly lossy compression with block-based coding algorithms.



a) b)

Figure 2 – *Common 2D image-based distortions: a) blurring effect; b) blocking effect* [3].

The view synthesis process renders a synthesized view by exploiting its correlation with its neighboring perspectives by using appropriate interpolation tools; depending on the specific DIBR technique involved, irregularities and errors are produced, thus resulting into new distortion types. DIBR distortions might be originated from different sources, from the synthesis process itself to the texture and depth compression. As views synthesis involves new distortion types, the available 2D image-based objective quality metrics may not perform well to evaluate the quality of the synthesized views. In addition to the typical 2D distortions, synthesized views include a brand-new variety of artefacts, notably:

- *Geometric Distortions*: As shown in Figure 3 a), geometric distortions are usually found in the object

edges and are usually caused by depth map incongruences, e.g. resulting from high compression.

- *Ghosting Distortions*: As illustrated in Figure 3 b), ghosting occurs after the blending process, usually due to inaccurate camera parameters or errors in the edge matching process; it is typically caused by overfiltering and high texture compression.
- *Cracks*: As shown in Figure 3 c), they refer to perceivable image cracks. Cracks normally appear as a result of rounding operations.
- *Occluded Areas*: Occur when the reference views to be used for the synthesis process have occluded areas that are not anymore occluded in the synthesized viewpoint, large holes that need to be filled may appear, as shown in Figure 3 d).
- *Flicker Distortions*: Refer to the flickering effect that may only take place in video sequences. This distortion may be caused by depth map inaccuracies, which project pixels to wrong positions during short periods of time, thus creating some flickering effects.

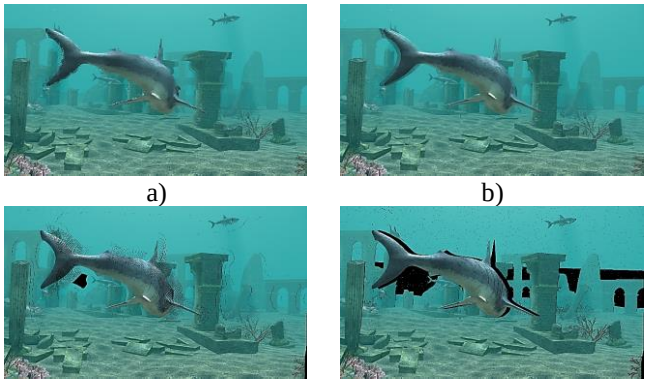



Figure 3 – *Common 3D image-based distortions: a) geometric distortion; b) ghosting distortion; c) cracks; d) occluded areas.*

### *B. 2D Objective Quality Metrics*

Image quality assessment metrics provide an objective, mathematical way to compute the perceivable quality without directly using human observers. That follow four broadly related although distinct assessment paradigms:

- *Signal-based approach*: The quality assessment of a sequence is based on the signal fidelity, meaning that by measuring the signal fidelity of a test sequence against the reference signal, a score will be created.
- *Perceptual-based approach*: The quality assessment based on a perceptual-based approach considers that perceptually-like quality estimators have a better understanding of human visual system perception.
- *Structural-based approach*: Structural-based assessment methods are also included in the perceptual-based approach, as these methods have their theory based on the human visual structural perception as human visual perception is based on the extraction of structural information.
- *Human visual system (HVS)-based approach*: This approach is also related to the perceptual-based approach, as it is a more detailed application of the perceptually-like quality estimators. However, this HVS-based approach relies on HVS modelling from psychophysical experiments.

1) *PSNR:* Peak Signal-to-Noise Ratio

Peak Signal-to-Noise Ratio (PSNR) and the related Mean Square Error (MSE) are the most broadly used metrics to assess lossy compressed signals.

MSE is a mathematical error function used to assess the quality of an estimation; in the context of FR image quality assessment, it essentially measures the difference between a distorted visual signal and its reference (ground truth) in a pixelwise way as follows:

$$MSE = \frac{1}{I_L I_W} \sum_{j=1}^{I_L} \sum_{i=1}^{I_W} [I_R(i,j) - I_D(i,j)]^2 \qquad (1)$$

where $I_L$ and $I_W$ represent the number of pixels in the image horizontal and vertical dimensions, $I_R$ and $I_D$ are the reference and distorted image, and $i$ and $j$ are the indexes used as pixel coordinates in the MSE computation process.

The PSNR is a quality metric, and not anymore a distortion metric as the MSE, and it is computed as the ratio between the image maximum dynamic range and the MSE expressed in base 10 logarithmic scale as:

$$PSNR = 10 \log_{10} \left( \frac{255}{MSE} \right) \qquad (2)$$

2) *SSIM:* Structural Similarity Index

The Structural Similarity (SSIM) Index was first introduced by Wang and Bovik in 2004 [4], and brought a new image quality assessment paradigm to the scientific community. SSIM organizes the similarity measurement into three comparisons between the reference ($R$) and the distorted ($D$) images, namely: luminance ($l(\cdot)$), contrast ($c(\cdot)$) and structure ($s(\cdot)$), as follows:

$$SSIM(R,D) = \left[ l(R,D)^{\alpha} \cdot c(R,D)^{\beta} \cdot s(R,D)^{\gamma} \right] \qquad (3)$$

3) *VIF:* Visual Information Fidelity

In 2015, Sheikh and Bovik proposed a quality metric that is based on Shannon's information theory featuring statistical models named Visual Information Fidelity (VIF) [5]. This solution assumes that human beings developed their visual system over the years to best perceive natural scenes, which correspond to a small subset of all possible signals/images.

4) *VQM:* Video Quality Metric

The Video Quality Metric (VQM) was proposed in late 2003 by Wolf and Pinson from the National Telecommunications and Information Administration (NTIA) as a model to estimate video quality [6]. The VQM consists in a four-staged process, notably: i) Calibration (spatial and temporal misalignment handling, identifying valid regions and correcting the gain and level values); ii) Extraction and Perception-based Features Processing; iii) Video Quality Parameters Computation; iv) Final Score (linear combination of the parameters that were computed in the previous stage).

5) *MOVIE:* Motion-based Video Integrity Evaluation

In early 2009, Seshadrinathan and Bovik released a new full reference video quality assessment metric that integrates both the spatial and temporal aspects of distortion assessment, names

as MOtion-based Video Integrity Evaluation (MOVIE) [7]. MOVIE metric uses a 3D spatio-temporal Gabor filter bank to extract multiscale spatial and temporal coefficients from the reference and test video sequences; these coefficients are selected according to the motion estimated from the reference video sequence in the form of an optical flow field.

### C. 3D Quality Metrics

As mentioned above, view synthesis techniques use new processing tools that generate new forms of distortions. This means that the previously available quality assessment metrics do not typically perform well in the assessment of synthesized views. Thus, new appropriate quality assessment metrics have to be developed. The new 3D quality assessment metrics typically follow a perceptual-based approach, some of which are:

#### 1) *3DSwIM:* 3D Synthesized view Image Quality Metric

The 3D Synthesized view Image quality Metric (3DSwIM) is a full reference video quality assessment metric for 3D synthesized views [8] that relies on two main assumptions: i) the visual quality of synthesized images is not greatly affected by displacement differences, which is relevant as synthesized views will typically display shifted objects regarding other views; and ii) distortion effects around human subjects are far more visually impacting. 3DSwIM performs quality assessment in a block-based structure, each block is then processed by analyzing the statistics between the reference and synthesized views, while adding a weighting factor, in this case related to skin detection.

#### 2) *SIQE:* Synthesized Image Quality Evaluator

The Synthesized Image Quality Evaluation (SIQE) metric is a RR video quality assessment metric [9] that is based upon the so-called *cyclopean eye theory*, which relates to the central reference midway point between the two eyes. The process to estimate a reference view begins by applying the divisive normalization transform to both views and then fuse them together. Regarding the synthesized view, the 3D video quality is then objectively scored through a statistical evaluation that compares the similarities between the synthesized and estimated reference cyclopean views. This RR video quality assessment is held in the Divisive Normalization transform, which has been endorsed as an effective nonlinear coding transform for natural sensory signals [10].

## IV. Spatio-Temporal Quality Assessment for Synthesized Views Metric

The VQA solution presented is largely inspired on the solution proposed in [11] and involves a full-reference VQA metric which follows a perceptive-based approach, notably for two types of distortions: i) conventional 2D spatial distortions; and ii) flickering distortions. According to this approach, the video sequence quality is assessed in the spatio-temporal domain along block-shaped, motion-coherent temporal tubes which are first detected in the reference view and after projected to the synthesized view.

As illustrated in Figure 4, this VQA metric includes by four main parts (red dashed boxes): i) Quality assessment structure definition; ii) Flickering distortion measurement; iii) Spatio-Temporal activity distortion measurement; and iv) Overall distortion measurement.

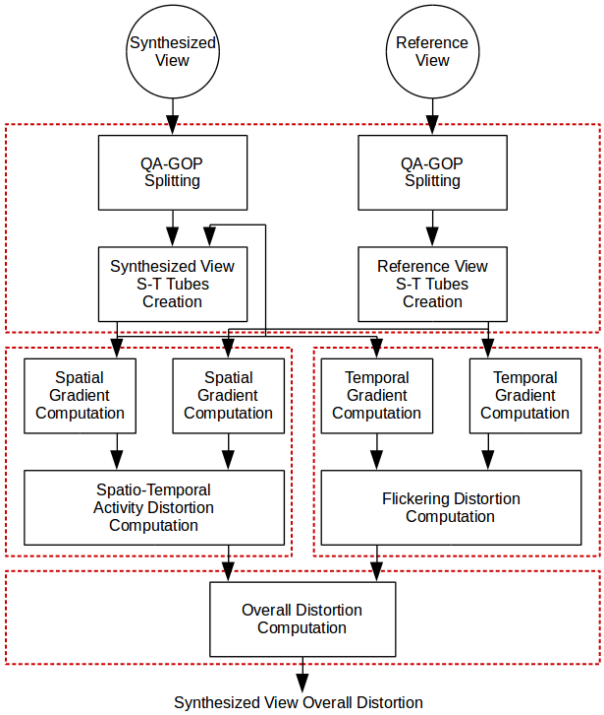



Figure 4 – *Processing architecture for the Spatio-Temporal Video Quality metric.*

The processing walkthrough for the Spatio-Temporal Video Quality metric involves the following main steps:

1) *QA-GOP Splitting*: Splits a sequence into groups of $2N + 1$ pictures here called as Quality Assessment-Group of Pictures (QA-GOP);

2) *Reference View Spatio-Temporal Tubes Creation*: Creates spatio-temporal (S-T) tubes for the reference view based on the motion estimated for each QA-GOP;

3) *Synthesized View Spatio-Temporal Creation*: Uses the spatio-temporal tubes defined structure created in Step 2) for the reference view to generate equivalent spatio-temporal tubes for the synthesized view video sequence;

4) *Spatial Gradient Computation*: Computes the spatial gradient for each block within each S-T tube;

5) *Temporal Gradient Computation*: Computes the temporal gradient for the blocks within each S-T tube;

6) *Spatio-Temporal Activity Distortion Computation*: Using the spatial gradient computed in Step 4), it computes the spatial-related distortion along time;

7) *Flickering Distortion Computation*: Using the temporal gradient computed in Step 5), it computes the so-called flickering distortion;

8) *Overall Distortion Computation*: Combines the activity distortion and flickering distortion scores to compute the final Spatio-Temporal Video Quality score.

This metric only processes the luminance component, the term ‘sample’ will always refer to ‘luminance sample’ in the following.

### A. Quality Assessment Group of Pictures Splitting

The Quality Assessment Group of Pictures (QA-GOP) Splitting module is responsible to divide the full set of pictures

in a video sequence into groups of pictures with a limited number of pictures, the so-called *quality assessment-group of pictures* (QA-GOP), each group composed by 2N+1 pictures.

### B. Reference View Spatio-Temporal Tubes Creation

The main objective of this Reference View Spatio-Temporal Tubes Creation module is to extract and define a temporal data structure called *spatio-temporal (S-T) tubes* for the reference view sequence by exploiting the motion in the scene, as locally represented in a QA-GOP.

S-T tubes creation architecture is represented in Figure 5. This module implements an iterative process that runs for all pictures in each QA-GOP, starting from its center and moving to both ends of the QA-GOP. The reason for the central frame to be used as the starting point is to minimize the error propagation along the successive motion estimation processes performed in this module.

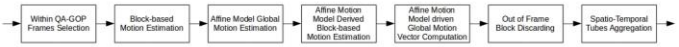



*Figure 5 – S-T tubes creation architecture.*

The various steps in this module are described in the following (this process is repeated for all the QA-GOPs in the sequence):

1) *Within QA-GOP Frames Selection:* This step is responsible to define the pairs of pictures to feed to the processing chain within this module.

2) *Block-based Motion Estimation:* This motion estimation step consists in applying a full-search block matching algorithm to estimate the motion vectors.

3) *Affine Model Global Motion Estimation:* Global motion is defined as the motion that applies to all parts of the frame, e.g. as generated by a camera horizontal displacement. The global motion is modeled by a six-parameter affine motion model, which parameters have to be estimated from the motion vectors field computed in the previous step.

4) *Affine Motion Model Derived Block-based Motion Estimation:* In this step, a motion vector is derived for each block in the source frame by computing the displacement between the source block under processing and the affine model warping of the same block. This will provide block-based motion information but now derived from the global motion information, this means the previously derived affine motion model, which characterizes the motion at the frame level.

5) *Affine Motion Model driven Global Motion Vector Computation:* The global motion vector is here computed as the average of the block-based motion vectors derived at the previous step.

6) *Out of Frame Blocks Discarding:* This step is responsible to identify and discard the blocks that might be leaving the visual scene, meaning that a motion compensated block in the DST frame falls out of it.

7) *Spatio-Temporal Tube Aggregation:* This final step is responsible to aggregate the set of motion compensated blocks into a specific S-T tube structure defined within a QA-GOP.

### C. Synthesized View Spatio-Temporal Tube Creation

The main objective of this module is to create the spatio-temporal (S-T) tubes for the sequence of frames in the synthesized view, notably using the S-T tubes already defined for the reference view.

### D. Spatial Gradient Computation

The main objective of this module is to compute the spatial gradient for each sample in the blocks contained in the previously defined S-T tubes. The horizontal and vertical gradient vectors are computed by a convolution operation using the defined horizontal and vertical kernels shown in Figure 6.

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

a)

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

b)

*Figure 6 – Spatial gradient kernels: a) horizontal; b) vertical.*

### E. Temporal Gradient Computation

The main objective of this module is to compute the temporal gradient for each block in the S-T tubes, both for the reference and synthesized views, based on the (sample) intensity changes (difference) along the tracked motion trajectory.

### F. Spatio-Temporal Activity Distortion Computation

The spatio-temporal activity distortion computation module is responsible to measure rather traditional 2D distortions, notably blurring and blocking artifacts; by computing the standard deviation of the spatial gradient for each S-T tube.

### G. Flickering Distortion Computation

The main objective of this module is to measure the flickering distortion in the synthesized video sequence. By studying the temporal gradient fluctuation for the samples along a specific S-T tube, as follow:

$$DF(x_t, y_t) = \sqrt{\frac{\sum_{t=2}^{2N+1} \Phi(x_t, y_t, t) \cdot \Delta(x_t, y_t, t)}{2N}} \qquad (IV)$$

where $\Phi$ and $\Delta$ are respectively the potential sensibility function and flickering distortion intensity. To measure the contribution of each sample in terms of flickering distortion intensity, equation (1) adopts a squared function that considers both the magnitude of the temporal gradient distortion (numerator), and the temporal masking effect (denominator). A $C$ constant term is added to avoid divisions by zero. The flickering distortion intensity at position $(x, y)$ is computed along the S-T tubes which capture the motion trajectory as follows:

$$\Delta(x, y, t) = \left( \frac{\vec{\nabla} I^{temporal}_{x,y,t} - \vec{\nabla} I^{temporal}_{x,y,t}}{\left| \vec{\nabla} I^{temporal}_{x,y,t} \right| + C} \right)^2 \qquad (1)$$

The potential sensibility function is computed as:

$$\Phi(x,y,t) = \begin{cases} 1, & \vec{\nabla} I^{temporal}_{x,y,t} \times \vec{\nabla} \tilde{I}^{temporal}_{x,y,t} \leq 0 \\ & \wedge \ \vec{\nabla} \tilde{I}^{temporal}_{x,y,t} \neq 0 \\ & \wedge \ \left| I(x,y,t) - \tilde{I}(x,y,t) \right| > \mu(x,y,t) \\ 0, & otherwise \end{cases} \quad (6)$$

where $I$ and $\tilde{I}$ denote the original and synthesized pixels respectively. $\vec{\nabla} I^{temporal}_{x,y,t} \times \vec{\nabla} \tilde{I}^{temporal}_{x,y,t} \leq 0$ and $\vec{\nabla} \tilde{I}^{temporal}_{x,y,t} \neq 0$ are used to exclude cases when reference and synthesized pixels have the same direction or synthesized pixels have no variation. The $\mu$ is the perceptual threshold that dictates the sensibility level from which an observer may notice a difference. This threshold is obtained from a Just-Noticeable Difference (JND) model computed as:

1) *JND Model Computation*

The JND model computation, illustrated in the left red-dashed part of Figure 7, relies upon two main factors: luminance adaptation (LA) and contrast masking (CM). The CM is computed as a sum of two evaluated masking effects, notably: edge masking (EM) and texture masking (TM).

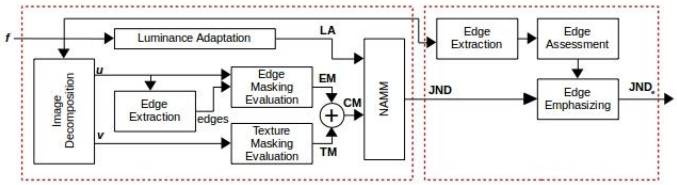



*Figure 7 – Edge emphasized JND model architecture.*

- *Luminance Adaptation*: Luminance adaptation, as expressed in equation (7) [12], attempts to model the HVS retina adaptation by reproducing relative changes described by the Webers' law [13], and ultimately addressing the desired luminance masking effect.

$$LA(x,y) = \begin{cases} 17 \times \left(1 + \sqrt{\frac{f(x,y)}{127}}\right) + 3 & \text{if } f(x,y) \leq 127 \\ 3 \times \frac{(f(x,y) - 127)}{128} + 3 & otherwise \end{cases} \quad (7)$$

- *Image decomposition*: Contrast masking is an effect associates to the fact that the HVS is less sensitive in areas with high spatial variation and more sensitive in large smooth areas. To address these distinct characteristics, the input frame is decomposed into two sub-images, the structural ($u$) and textural ($v$) sub-images, in a way that the sum of both sub-images constitutes the original image, i.e. $f$=$u$+$v$, as proposed in [14]. The structural image it's a coarse granular description of the image $f$, meaning that is piecewise smooth, while containing at the same time sharp edges. On the other hand, the textural image contains only the fine-scale details, usually with some oscillatory nature. The image decomposition is performed using the TV-L1 model [15], which offers a solution to the standard Total Variation (TV) denoising problem.

- *Edge extraction*: From the extracted structural image, an edge map is computed. This map is obtained by using the Canny edge detection algorithm with Otsu's method to improve its outcome.

Moreover, contrast masking is defined as the sum of two similar masking effects which occur in high textured areas and edges. However, textured areas present a more intense masking effect than edge areas. Therefore, these masking effects have different weighting parameters assigned; notably, as expressed in equation (2), edge masking is assigned a weight ($w_e$) of 1, while texture masking weight ($w_t$) is assigned 3. Both masking effects are computed using the same extent of variation ($C_s$) technique, for different sub-images; the constraint $\beta$ is set to 0.117 as in [12].

$$\begin{cases} EM^u(x,y) = C_s^u(x,y) \cdot \beta \cdot w_e \\ TM^v(x,y) = C_s^v(x,y) \cdot \beta \cdot w_t \end{cases} \quad (2)$$

Cs denotes the maximal weighted average of gradients, also known as mean filter, around a pixel. Gradients are computed independently for each sub-image, using four directional high-pass filters for texture detection ($g_k$ , $k \in [1,2,3,4]$) as illustrated in Figure 8.

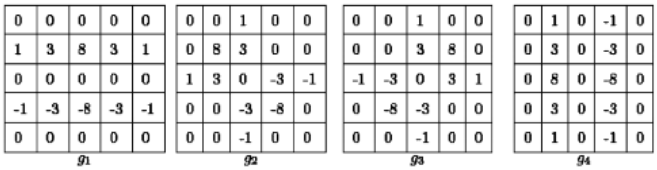



*Figure 8 – Directional high-pass filters for texture detection.*

- *Nonlinear Additivity Model for Masking*: To obtain the JND model, both LA and CM are joint using the nonlinear additivity model for masking (NAMM) [12]. NAMM, as expressed in equation (3), use a gain reduction factor ($C^{lc}$) which is related to overlapping effect in masking, and is also related to the viewing conditions (ambience lighting, display device, viewing distance, etc.). $C^{lc}$is set to 0.3 to adopt the same value as in [12].

$$JND = LA + CM - C^{lc} \times \min\{LA, CM\} \quad (3)$$

2) *Edge-based JND Emphasizing Post-Processing*

The resulting JND map is finally post-processed for edge emphasizing as a way to increase the sensibility around objects edges, illustrated as the red-dashed box on the right of the Figure 7. This method enhances the JND sensibility at pixel level for the edges in a given frame, by performing an edge extraction process and processing the resulting edges using an 8x8 block-based approach. At this stage, any block containing more than 48 edge pixels is classified as a texture block and the associated pixels are not emphasized. On the contrary, the remaining edges are used to emphasize the position corresponding previously computed JND model, by multiplying the JND values by 0.1.

### *H. Overall Distortion Computation*

The final module computes the overall distortion for the synthesized sequence, by combining both the spatio-temporal and the flickering distortions computed in the previous modules.

## V. QUALITY METRICS PERFORMANCE ASSESSMENT

The designed performance assessment workflow uses three main components: i) 3D Synthesized Views Video Quality Assessment Database; ii) Objective Quality Assessment Metric; iii) Quality Metrics Performance Assessment.

The 3D Synthesized Views Video Quality Assessment Database used in the performance assessment workflow comes from the Shenzhen Institute of Advanced Technology (SIAT) in China [16]. Views were synthesized using the VSRS-1D-Fast mode included in the 3D-HEVC reference software 3D-HTM v8.0 provided by the Fraunhofer Heinrich-Hertz-Institut [17].

The quality metrics performance assessment implements a non-linear regression process to allow the statistical study of the relationship between subjective and objective results. Following the ITU-T recommendations [18], the VQA metric scores ($Q$) are transformed (fitted) into predicted difference mean opinion score ($DMOS_p$) by applying the non-linear least squares regression analysis to the model defined in equation (10); notably, the $\beta$ parameters are defined by an optimal solution derived using the second-order optimization model [19] (Newton-Raphson) for a confidence interval greater than *95%*.

$$DMOS_p = \frac{\beta_1}{1 + e^{-\beta_2 \times (Q - \beta_3)}} \quad (10)$$

This step is necessary because the subjective and objective scores are scaled differently; thus, to perform an apples-to-apples comparison, objective scores need to be fitted into the subjective scale. Past this step, both scores are fit into the same scale, allowing the statistical study of all score pairs. As recommended by ITU-T [18], the performance assessment of a VQA metric is based in the following three-criteria analysis:

- *Pearson Linear Correlation Coefficient:* Pearson correlation coefficient is the most commonly used metric in the image-based QA field to measure the correlation between the subjective and objective scores.
- *Spearman Rank Correlation Coefficient:* Spearman correlation coefficient quantifies the monotonicity between the objective and the subjective scores.
- *Root Mean Square Error:* Root Mean Square Error (RMSE) is a widely-used statistical metric that quantifies the error between two variables, by aggregating the magnitude of errors between each variable-sample pair.

## A. VQA Metric Configuration Profiles

To study how the different approaches and techniques impact the metrics performance, the performance assessment of the objective VQA for synthesized views was evaluated using several configurations. The metric performance depends on three key aspects, notably: i) Motion vector estimation approach; ii) Flickering distortion perception threshold model; iii) Edge detection threshold approach.

1) *Motion Vector Estimation Approach (MVEA)*: Three different configurations were tested in terms of motion vectors estimation as shown in Table 1, where *w* and PF represent, respectively, the window size and motion vector amplitude penalty.

TABLE 1 – MOTION VECTOR ESTIMATION: CONFIGURATION CHARACTERISTICS.

| **Alg.** | **Block Matching** | | |
|---|---|---|---|
| **Config.** | BM_32 | BM_64 | BM_64_CF |
| **Details** | *w*: 32p; PF: MAD | *w*: 64p; PF: MAD | *w*: 64p; PF: MAD + $K \times \|\|mv\|\|_2$ where $K = 0.05$ |

2) *Configuration #2: Flickering Distortion Perception Threshold Model (FDPTM):* Two different hypotheses were considered as models of the perception threshold, notably: i) the JND model; ii) the edge emphasized JND model (JNDe).

3) *Configuration #3: Edge Detection Threshold Approach (EDTA)*: Two approaches were considered, one static and another adaptive. The static approach fixes the high and low thresholds used in the Canny edge detection algorithm as 200 and 100 respectively. The adaptive approach is based on the Otsu method, which is used to compute the high threshold value based on the content of the image *[20]*; then, the low threshold is found by multiplying the high threshold by 0.5 as in *[21]*.

## B. Performance Study of VQA Metric Configurations

As a baseline, all different parameters described in the previous Section are studied considering the three different score results, notably: i) Distortion Activity; ii) Flickering Distortion; iii) Overall Distortion. Table 2, Table 3 and Table 4 presents the performance assessment results of each distortion assessment score per adjusted parameter; hence, Motion Vector Estimation Approach, Perception Threshold Model and Edge Detection Threshold Approach, for respectively the distortion activity, flickering distortion and overall distortion.

The motion vector estimation technique has a direct impact over the S-T tubes creation module, which in turn has repercussion on the metrics performance. As it can be seen in Tables 2-4, changing the technical solution of the motion estimation has a considerable level of impact. The MVE has a higher impact over the flickering distortion, which is expected as it represents temporal analysis, meaning that it is highly sensible to the estimations quality.

TABLE 2 – PERFORMANCE ASSESSMENT ON RELEVANT METRIC PARAMETERS: DISTORTION ACTIVITY.

| Parameters | | | Distortion Activity | | |
|---|---|---|---|---|---|
| MVEA | FDPTM | EDTA | ρ | r | RMSE |
| BM_32 | – | – | **0.826** | **0.827** | **0.072** |
| BM_64 | – | – | 0.819 | 0.815 | 0.074 |
| BM_64_CF | – | – | 0.825 | 0.823 | 0.072 |

TABLE 3 – PERFORMANCE ASSESSMENT ON RELEVANT METRIC PARAMETERS: FLICKERING DISTORTION.

| Parameters | | | Flickering Distortion | | |
|---|---|---|---|---|---|
| MVEA | FDPTM | EDTA | ρ | r | RMSE |
| BM_32 | JND | Static | 0.703 | 0.689 | 0.091 |
| | | Adaptive | 0.697 | 0.685 | 0.092 |
| | JNDe | Static | 0.663 | 0.651 | 0.096 |
| | | Adaptive | 0.665 | 0.655 | 0.096 |
| BM_64 | JND | Static | **0.711** | 0.696 | **0.090** |
| | | Adaptive | 0.706 | 0.691 | 0.091 |
| | JNDe | Static | 0.674 | 0.662 | 0.095 |
| | | Adaptive | 0.674 | 0.664 | 0.095 |
| BM_64_CF | JND | Static | 0.705 | **0.701** | 0.091 |
| | | Adaptive | 0.698 | 0.688 | 0.092 |
| | JNDe | Static | 0.661 | 0.650 | 0.096 |
| | | Adaptive | 0.660 | 0.647 | 0.096 |

TABLE 4 – PERFORMANCE ASSESSMENT ON RELEVANT METRIC PARAMETERS: OVERALL DISTORTION.

| Parameters | | | Overall Distortion | | |
|---|---|---|---|---|---|
| MVEA | FDPTM | EDTA | ρ | r | RMSE |
| BM_32 | JND | Static | 0.815 | 0.802 | 0.074 |
| | | Adaptive | 0.811 | 0.800 | 0.075 |
| | JNDe | Static | 0.805 | 0.797 | 0.076 |
| | | Adaptive | 0.807 | 0.799 | 0.076 |
| BM_64 | JND | Static | 0.819 | 0.805 | 0.074 |
| | | Adaptive | 0.816 | 0.802 | 0.074 |
| | JNDe | Static | 0.812 | 0.799 | 0.075 |
| | | Adaptive | 0.814 | 0.801 | 0.075 |
| BM_64_CF | JND | Static | **0.820** | **0.809** | **0.074** |
| | | Adaptive | 0.814 | 0.804 | 0.075 |
| | JNDe | Static | 0.808 | 0.800 | 0.077 |
| | | Adaptive | 0.807 | 0.798 | 0.076 |

The second configuration, so-called flickering distortion perception threshold model, refers to the selection of one of the two available JND models. These models express a pixel-wise map of perceivable thresholds (μ(x,y,t)) that explicitly define the level from which the computed flickering effect is perceived by a human observer. The use of different models has impact over the flickering distortion sensibility function, that consequently impacts the flickering distortion and the overall distortion metric. As shown in Tables 2-4, the model selection is a factor that has a great impact on the flickering distortions' performance, exactly in the order of 4-5% for the Pearson and Spearman correlation coefficients for all cases.

The edge detection is only performed in the flickering distortion computation module, namely at the JND model computation and the edge emphasizing post-processing. As presented in Tables 2-4, the major difference of Pearson value is found for the JND flickering distortion perception threshold model; this means that the edge detection threshold approach has some impact (0.7%) over the flickering distortion performance when the JND is used as the FDPTM. For the Spearman correlation value, the difference between both EDTA in the JND flickering distortion perception threshold model is in the order of 1.3%, which indicate that the adaptive method for this case produce some jittery results.

To understand the impact that each profile has in the overall performance of the objective quality metric, a deeper analysis is performed for the motion vector estimation and edge detection threshold approaches.

1) *Motion Vector Estimation Approach Analysis:* The worst quality was obtained for the BM_64 approach. This effect is present when the block for which the motion is estimated is in a smooth area, but gets worse when areas are bigger. Additionally, BM_32 shown the same behaviour but in a smaller area due to the smaller window size. BM_64_CF was the best one, producing really consistent estimations. Therefore, it can be concluded that simply using the MAD criterion for motion vector estimation can lead to lower quality in comparison with better motion vector modelling criteria such as MAD plus the addition of a cost function that increase the penalization over great distances, or using an adaptive support-weight window penalizing function *[22]*; notably, MAD can create jittery S-T tubes and poor object tracking, which may also lead to inaccurate S-T tubes exclusion.

2) *Edge Detection Threshold Approach Analysis*: The adaptive threshold classifies more pixels as edges than the static approach. Considering these results with the performance results presented in Tables 2-4, it is possible to conclude that the number of edge marked pixels is inversely proportional to the performance results.

### C. *Performance Assessment Comparisson of the proposed video quality metric to 2D Objective Quality Metrics*

This section presents the performance assessment of the 2D objective quality metrics referred in Section III-B, but also adds the following: MSSIM [23]; UQI [24]; IFC [25]; NQM [26]; WSNR [27]; and SNR. The overall performance assessment results for the SIAT synthesized video quality database are shown in Table 5.

As shown in the Table 5 quality metrics which are a mixture of signal and perceptual principles, such as MSSIM and SNR, have shown to perform quite well. Perceptual-based 2D objective quality metrics have proven to perform worse, as some assumptions are made based on some specific 2D cases and do not consider the temporal-based distortions generated by the view synthesis process, e.g. flickering distortions.

TABLE 5 – PERFORMANCE COMPARISON OF OBJECTIVE VIDEO QUALITY ASSESSMENT: 2D VQA.

| VQA | ALL DATA | | |
|---|---|---|---|
| | *ρ* | *r* | *RMSE* |
| **MSE** | 0.653 | 0.631 | 0.097 |
| **PSNR** | 0.650 | 0.627 | 0.098 |
| **SSIM** | 0.581 | 0.546 | 0.104 |
| **MSSIM** | 0.748 | 0.736 | 0.085 |
| **VSNR** | 0.678 | 0.667 | 0.094 |
| **VIF** | 0.631 | 0.629 | 0.100 |
| **VIFP** | 0.658 | 0.630 | 0.097 |
| **UQI** | 0.477 | 0.459 | 0.113 |
| **IFC** | 0.477 | 0.459 | 0.113 |
| **NQM** | 0.554 | 0.515 | 0.107 |
| **WSNR** | 0.620 | 0.588 | 0.101 |
| **SNR** | 0.759 | 0.718 | 0.084 |
| **VQM** | 0.674 | 0.665 | 0.095 |
| **S-MOVIE** | 0.705 | 0.696 | 0.091 |
| **T-MOVIE** | 0.518 | 0.461 | 0.110 |
| **MOVIE** | 0.679 | 0.660 | 0.094 |
| **Proposed** | **0.820** | **0.809** | **0.074** |

The Figure 9 illustrates the performance assessment of each 2D objective quality metric, for the subsets of the dataset: UU: uncompressed texture and uncompressed depth; UC: uncompressed texture and compressed depth; CU: compressed texture and uncompressed depth; CC: compressed texture and compressed depth.

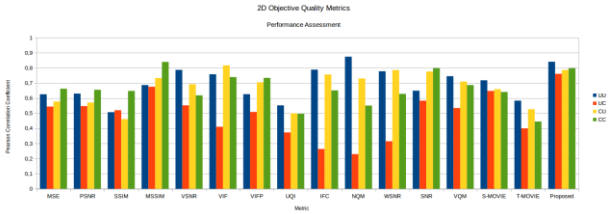



*Figure 9 – Pearson correlation coefficient for each 2D image metric.*

As illustrated in Figure 9, it is noticeable that the subset for which the 2D metrics has lower performance is the case of synthesized videos where the texture is uncompressed and the depth is compressed, notably where the flickering distortions have a greater impact over the video quality. This effect is smaller for the pure signal-based approaches, such as MSE and SNR, because they essentially rely upon the signals' difference between the reference and the test sequences.

There are metrics which show a great difference in different subsets, for example, NQM shows a good performance when the depth and texture is uncompressed, much higher than the MSE; however, NQM performance is poor when depth and texture are both compressed, resulting in the lowest performance than the consistent MSE for the whole dataset.

The scatter plots of DMOS versus $DMOS_p$ for each 2D objective quality metrics are shown in Figure 10, marking the performance of each subset as the previous figure.

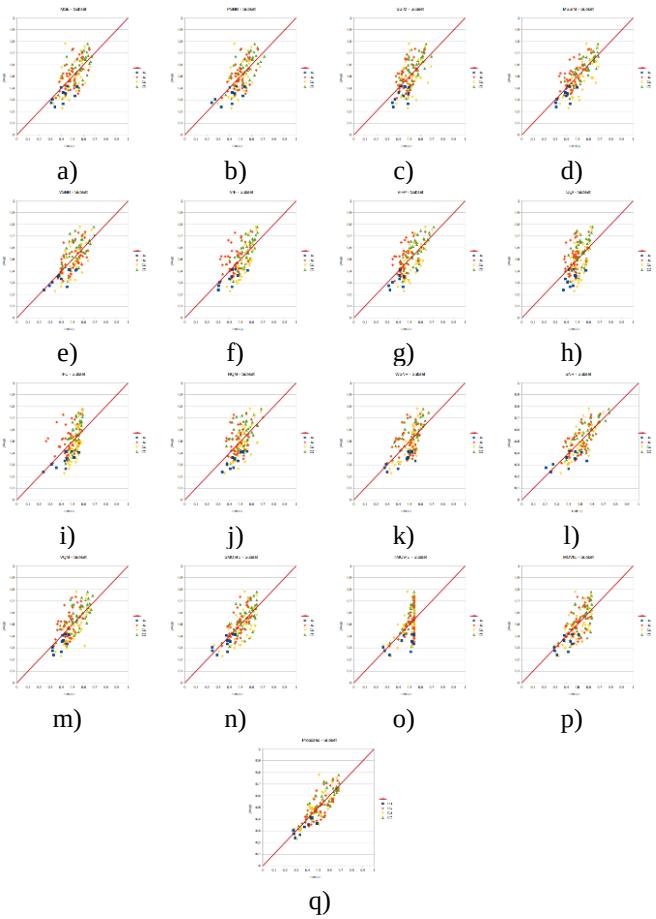



*Figure 10 – DMOS versus DMOSp for different 2D objective quality metrics.*

Among all the compared 2D QA metrics, the SNR has the highest performance across all subsets and for the entire dataset. This reveals that although new artefacts emerge during view synthesis process, such as geometric distortions, these spatial distortions can be measured somewhat efficiently by 2D QA metrics. A special note for MSSIM and Spatial MOVIE that correlated well with DMOS on the entire dataset, and the proposed metric had the highest linear correlation.

### *D. Performance Assessment Comparisson of the proposed video quality metric to 3D Objective Quality Metrics*

The overall performance assessment results for the SIAT synthesized video quality database are presented in Table 6.

TABLE 6 – PERFORMANCE COMPARISON OF OBJECTIVE VIDEO QUALITY ASSESSMENT: 3D VQA.

| VQA | ALL DATA | | |
|---|---|---|---|
| | *ρ* | *r* | *RMSE* |
| **SIQE** | 0.140 | 0.139 | 0.127 |
| **3DSwIM** | 0.238 | 0.260 | 0.125 |
| **SIAT** | 0.815 | **0.869** | **0.074** |
| **Proposed** | **0.820** | 0.809 | **0.074** |

The performance assessment results show that 3D quality metrics (Section 3.2), like SIQE and 3DSwIM, exhibit a poor performance when evaluated by the SIAT synthesized video quality database. The lack of good performance for these metrics are consequence of being image oriented, and following a perceptual-based approach which do not consider in anyway the temporal artefacts created by the synthesis process. Hence, as shown in Figure 11, which illustrates the 3DSwIM score versus DMOS by sequence (a) and by subset (b), it can be seen that the 3DSwIM for each sequence evaluates the uncompressed texture and depth ($U_T U_D$) subset as being similar to the uncompressed texture and compressed depth ($U_T C_D$) subset, giving to both subsets an equivalent score, disregarding the flickering distortions.

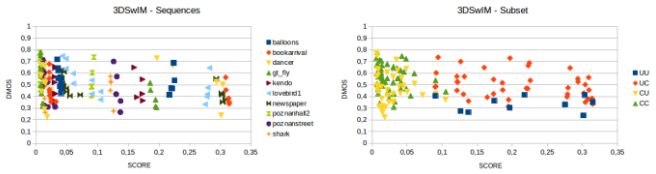
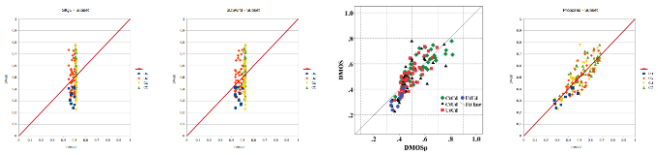



*Figure 11 – 3DSwIM DMOS vs Score per: a) Sequence; b) Subset.*

The proposed metric follows a similar structure as the SIAT 3D video quality metric. However, the performance results are quite different; notably, showing an improvement of the linear correlation (Pearson) but a lower monotonic correlation (Spearman). This indicates that the proposed metric gives better but noisier scores, in terms of monotonic behaviour.

The scatter plots of DMOS versus $DMOS_p$ for each quality assessment metric is shown in Figure 12. The quality assessment metrics that follow the signal-based approach do not have a good correlation regarding the perceptual scores.



a) b) c) d)

*Figure 12 – DMOS versus DMOSp for different 3D objective quality metrics.*

The proposed method has a good performance as suggested in Figure 12, with a better Pearson linear correlation since the points are closer to the ideal line DMOSp=DMOS line.

The proposed method presents higher RMSE for the kendo and lovebird1 sequences, as shown in Table 7.

The kendo sequence has large smooth areas and a considerable percentage of overexposed area, that constituted a major issue for the quality of the MVE. The lovebird1 sequence, is characterized by a large area which has little or no motion at all, thus the observer pays little attention to that areas and much of the attention is devoted to the two human actors in scene. As mentioned by several objective metrics, the HVS is more sensitive to distortions that might occur around humans, thus the proposed method gives a lower distortion score than the subjective one, increasing the RMSE for this sequence.

TABLE 6 – RMSE OF THE PROPOSED METRIC PER SEQUENCE.

| Sequence | RMSE |
|---|---|
| **Balloons** | 0.045 |
| **BookArrival** | 0.062 |
| **Dancer** | 0.044 |
| **GT Fly** | 0.054 |
| **Kendo** | 0.101 |
| **Lovebird1** | 0.101 |
| **Newspaper** | 0.063 |
| **PoznanHall2** | 0.097 |
| **PoznanStreet** | 0.087 |
| **Shark** | 0.045 |

As seen in Figure 12 d), there are an obvious outlier that belongs to the $C_T U_D$ subset. This outlier is from the PoznanHall2 sequence with the QP 38 and 0 for the texture and depth, respectively.

## VI. SUMMARY AND FUTURE WORK

In this paper, some relevant quality assessment metrics have been reviewed, which include both image and video objective quality assessment metrics for 2D and 3D synthesized images. The synthesized views quality assessment metrics address new types of distortions introduced by the rendering process. Various 2D and 3D objective quality assessment metrics were studied in this paper, and heavily based on [11] a metric was implemented.

The proposed metric was assessed using the SIAT database, and compared with other objective quality assessment metrics. The experimental results shown that the proposed metric has a good performance compared with the state-of-the-art objective quality assessment metrics on the entire database, and is particularly prominent in the subset that has significant temporal flickering distortions caused by depth compression and by the view synthesis technique.

Despite the good overall performance results, showing that the proposed metric is reliably consistent, some issues need to be addressed to further improve the performance, notably:

1) Improve the MVE approach used to create the tubes, for instance the block matching with adaptive support/weight window penalizing function *[22]*, piecewise rigid scene model *[28]* and instance scene motion flows *[29]*.
2) The use of a visibility prediction model of flicker distortions on natural videos *[30]* to the FDPTM.
3) The use of a region of interest approach instead of using the 10% worst S-T tubes. One suggestion is by using the fast region-based convolutional networks for object detection *[31]*, and quantifying the importance of each region of interest with a scalable visual sensitivity profile estimation *[32]*.
4) The addition of a third score that address geometric distortions by adapting the DeepQA *[33]* to the temporal domain maybe considering using long short-term memory network adaptation.

## VII. REFERENCES


[1] ITU-R, *Rec. BT.500 Methodology for the subjective assessment of the quality of television pictures.*, Geneva, Switzerland: ITU-R, January 2012.

[2] ITU-T, *Rec. P910 Subjective video quality assessment methods for multimedia applications*, Geneva, Switzerland: ITU-T, April 2008.

[3] R. Song, H. Ko and C. C. Kuo, "MCL-3D: a database for stereoscopic image quality assessment using 2D-image-plus-depth source," *Journal of Information Science and Engineering*, vol. 31, no. 5, pp. 1593-1611, March 2014.

[4] Z. Wang, A. C. Bovik and L. Lu, "Why is image quality assessment so difficult?," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2002*, Orlando, FL, USA, May 2002.

[5] A. Srivastava, A. Lee, E. P. Simoncelli and S.-C. Zhu, "On advances in statistical modeling of natural images," *Journal of Mathematical Imaging and Vision*, vol. 18, no. 1, pp. 17-33, January 2003.

[6] M. H. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," *IEEE Transactions on Broadcasting*, vol. 50, no. 3, pp. 312-322, September 2004.

[7] D. J. Fleet and Y. Weiss, "Optical flow estimation," in *Mathematical models for Computer Vision: The Handbook*, Springer, 2005, pp. 239-257.

[8] F. Battisti, E. Bosc, M. Carli, P. Callet and S. Perugia, "Objective image quality assessment of 3D synthesized views," *Signal Processing Image Communication*, vol. 30, no. C, pp. 78-88, January 2015.