Objective quality metric for 3D virtual views

This is the author's manuscript

Original Citation:

Availability:
This version is available http://hdl.handle.net/2318/1557627 since 2016-03-09T13:39:10Z

Publisher:
IEEE Computer Society

Published version:
DOI:10.1109/ICIP.2015.7351499

Terms of use:
Open Access
Anyone can freely access the full text of works made available as "Open Access". Works made available under a Creative Commons license can be used according to the terms and conditions of said license. Use of all other works requires consent of the right holder (author or publisher) if not exempted from copyright protection by the applicable law.

(Article begins on next page)
This copy represents the peer reviewed and accepted version of paper: 

Farid, M.S., Lucenteforte, M., Grangetto, M.

doi: 10.1109/ICIP.2015.7351499

The published version is available at
http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7351499

IEEE Copyright. This material is presented to ensure timely dissemination of scholarly and technical work. Copyright and all rights therein are retained by authors or by other copyright holders. All persons copying this information are expected to adhere to the terms and constraints invoked by each author's copyright. In most cases, these works may not be reposted without the explicit permission of the copyright holder. Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE.
ABSTRACT
In free-viewpoint television (FTV) framework, due to hardware and bandwidth constraints, only a limited number of viewpoints are generally captured, coded and transmitted; therefore, a large number of views needs to be synthesized at the receiver to grant a really immersive 3D experience. It is thus evident that the estimation of the quality of the synthesized views is of paramount importance. Moreover, quality assessment of the synthesized view is very challenging since the corresponding original views are generally not available either on the encoder (not captured) or the decoder side (not transmitted). To tackle the mentioned issues, this paper presents an algorithm to estimate the quality of the synthesized images in the absence of the corresponding reference images. The algorithm is based upon the cyclopean eye theory. The statistical characteristics of an estimated cyclopean image are compared with the synthesized image to measure its quality. The prediction accuracy and reliability of the proposed technique are tested on standard video dataset compressed with HEVC showing excellent correlation results with respect to state-of-the-art full reference image and video quality metrics.

Index Terms— Quality assessment, depth image based rendering, view synthesis, FTV, HEVC

1. INTRODUCTION
Depth perception in 3D television (3DTV) is achieved by rendering two views of the scene captured at slightly different viewpoints. The latest 3D display technologies - free-viewpoint television (FTV) [1] and future Super Multiview (SMV) displays [2], are capable to provide hundreds of high resolution views with base line distance smaller than interocular distance [3]. Both FTV and SMV require huge number of views to provide the viewer the freedom to roam around a scene. Due to various hardware, economic and bandwidth constraints acquisition and transmission of such huge number of views is not possible. Therefore, only few views are captured and transmitted; the rest are synthesized on the receiver side.

High quality immersion requires efficient 3D content representation and compression as well as computationally viable view synthesis techniques to generate good quality novel views. Multiview videos plus depth (MVD) format [4] has gained widespread acceptability due to its provision for intermediate virtual views generation and efficient compression. MVD has been adopted for future FTV and SMV technologies for both compression and display [5]. In MVD format in addition to texture images, gray scale depth maps representing the per pixel depth value are also available which permit the generation of novel views through Depth Image Based Rendering (DIBR) techniques [6].

Efficient compression of MVD data is central to 3D television processing chain and a number of compression friendly MVD data representations e.g., [7–9] have been proposed resulting in the developments of novel codecs. With the introduction of the novel state-of-the-art High Efficiency Video Coding (HEVC) [10] the Joint Collaborative Team on 3D Video Coding Extension Development (JCT-3V) has recently developed extensions of HEVC: Multiview-HEVC (MV-HEVC) and 3D-HEVC [11]. 3D-HEVC exploits additional coding tools and achieves the best compression ratio for MVD data [11] and it also guarantees promising quality virtual views with View Synthesis Optimization (VSO) coding tool. To enable autostereoscopy additional views on the receiver side are generated through DIBR. Since there are only few original coded views, the overall user experience largely depends on the synthesized views.

Quality assessment of synthesized images is a challenging task as the corresponding reference images are not available. Furthermore, the generated images suffer from various types of artifacts e.g., compression artifacts [12], synthesis distortion [13], textural and structural distortions due to poor quality depth maps [14–16]. Recent studies [17, 18] tested various existing 2D image quality metrics to assess the quality of stereoscopic and synthesized images and concluded that none of them is suitable in this context. In the recent years a number of quality metrics for stereoscopic images have been proposed [19] which mostly extends the existing 2D quality metrics. However, little attention has been paid to the quality evaluation of synthesized images. In [20] quality of the virtual view is assessed by comparing the structures e.g. edges of the original and the warped images. It is limited to struc-
tural distortion estimation and cannot be used to represent the overall quality of the virtual image as it does not compute the color related artifacts. CSED (Color and Sharpness of Edge Distortion) [21] is another full reference quality metric to assess the quality of the virtual image. It targets the hole regions to assess the color distortion and uses the edge sharpness of the reference and virtual images to assess structural distortion. Battisti et al. [22] proposed to weight more the distortion created around the human body that largely affects the quality of the whole synthesized view. They proposed 3DSwIM metric which assesses the virtual image quality by detecting the skin regions and aligning them with the reference images to determine the distortion. View Synthesis Quality Assessment (VSQA) metric [23] combines SSIM with weighting functions derived from contrast, orientation and texture maps of the reference and synthesized views to assess the quality of the virtual pictures.

The previous discussion shows that existing techniques for quality evaluation of synthesized views are either full reference or reduced reference. However, in the FTV scenario a large number of views are synthesized whose references are not available. Therefore, the existing techniques become ineffective. This paper proposes a solution to the problem thanks to the following contributions:

- definition of a novel Synthesized Image Quality Evaluator (SIQE) to assess the quality of depth based rendered images in absence of corresponding reference images;
- exploitation of the cyclopean eye theory and divisive normalization transform to infer the quality of a picture rendered from a stereo pair (video plus depth);
- statistical analysis of the quality prediction accuracy obtained on a set of video plus depth sequences compressed with HEVC: the SIQE performance is compared versus other state-of-the-art full reference image and video quality metrics showing competitive results.

The rest of the paper is organized as follows: the proposed quality metric is described in Sect. 2, followed by experimental evaluation in Sect. 3. The conclusions are drawn in Sect. 4.

2. PROPOSED SYNTHESIZED IMAGE QUALITY EVALUATOR

The proposed quality metric is build around the Béla Julesz’s Cyclopean Perception theory [24] and Divisive Normalization transform [25]. The Cyclopean Perception refers to the formation of a virtual image in our mind from the stimuli received from the left and the right eye. The cyclopean image (or the mental image), is a view obtained by fusing the left and the right views as if it was captured by a virtual eye (usually referred to as the cyclopean eye) placed in between the two eyes. This process is graphically shown in Fig. 1.

Divisive normalization (DN) is based upon the standard psychophysical and physiological model and it has been used to study the nonlinear behaviors of cortical neuron in biological vision. The use of DN model in image quality assessment was pioneered by Teo and Heeger in [25]. Moreover, it is shown that DN model achieves statistical independence [26] and can be represented by Gaussian scale mixture (GSM) model. Using GSM model a number of quality assessment techniques have been proposed using various spatial and frequency based DN transforms e.g., [27, 28]. In the proposed quality model we exploit the divisive normalization to estimate the statistical characteristics of the uncompressed stereo images and fuse them together to obtain the statistical model of the cyclopean image. Depth image based rendering is used to generate the intermediate virtual image from the coded stereopair and the corresponding depth maps. The statistical model of this synthesized image is also computed by using the divisive normalization. This model is compared to the reference cyclopean image model to estimate the quality degradation in the synthesized image due to compression and 3D warping. Fig. 2 shows the block diagram of the whole algorithm.

Let $V_l$, $V_r$ be the uncompressed left and right texture images and let $V_s$ be a synthesized image obtained from the compressed texture and depth images (as shown in Fig. 2). The size of all the views is $M \times N$. Let $T_l$, $T_r$ and $T_s$ are the divisive normalized images of $V_l$, $V_r$ and $V_s$ respectively which are created in the spatial domain similar to [28, 29].
Any \( T_k \) is computed from \( V_k \), \( k \in \{l, r, s\} \) as follows:

\[
T_k(u, v) = \frac{V_k(u, v) - \mu_k(u, v)}{\sigma_k(u, v) + \epsilon}
\]

(1)

where \( \epsilon \) is a small constant used to avoid division by zero. \( \mu_k(u, v) \) and \( \sigma_k(u, v) \) are local average and standard deviation computed over a block of size \( m \times n \) centered at \((u, v)\). These are computed as:

\[
\mu_k(u, v) = \frac{1}{m \cdot n} \sum_{i=-m}^{m} \sum_{j=-n}^{n} w(i, j) V_k(u + i, v + j)
\]

\[
\sigma_k(u, v) = \left[ \sum_{i=-m}^{m} \sum_{j=-n}^{n} w(i, j) [V_k(u + i, v + j) - \mu_k(u, v)]^2 \right]^{1/2}
\]

where \( w \) is 2D symmetric Gaussian weight function computed as:

\[
w(i, j) = \frac{1}{2\pi\sigma^2} e^{-\frac{i^2 + j^2}{2\sigma^2}}
\]

(2)

The statistical characteristics of the cyclopean image are estimated from the DN representations \( T_l, T_r \) of the left and the right views respectively. To this end, we propose to exploit the histogram of the DN image \( T_c = [T_l|T_r] \), obtained by simple concatenation of the left and right images. In particular, from the histogram of \( T_c \) we compute the normalized probability distribution \( \mathcal{N}_c \) using \( k \) equally spaced bins; it follows that: \( \sum_{i=1}^{k} \mathcal{N}_c(i) = 1 \). Then, the normalized distribution \( \mathcal{N}_s \) of the DN synthesized image \( T_s \) is computed using the same number of bins. Fig. 3 shows the \( \mathcal{N}_c \) and \( \mathcal{N}_s \) curves for a sample image from Poznan_Street video sequence (see Tab. 1). The \( \mathcal{N}_c \) models the distribution of the cyclopean image estimated from the left and the right views, whereas the \( \mathcal{N}_s \) curves show the distribution of the intermediate synthesized image obtained through DIBR from the respective sequence coded at 6 different quality levels. The non-overlapped areas between the \( \mathcal{N}_s \) and \( \mathcal{N}_c \) curves represents the distortion in the respective synthesized image which can be estimated by computing the difference between the two distributions. To this end we propose to exploit the Bhattacharyya coefficient (\( \rho \)) since it has been already shown to be more reliable than other metrics, e.g. the Mahalanobis distance. The Bhattacharyya coefficient is used to estimate the similarity between the two distributions as follows:

\[
\rho(\mathcal{N}_c, \mathcal{N}_s) = \sum_{x \in \mathcal{X}} \sqrt{\mathcal{N}_c(x) \mathcal{N}_s(x)}
\]

(3)

Finally, the proposed SIQE metric is computed as the difference between the two models computed through the Hellinger distance [30], defined as:

\[
SIQE = \sqrt{1 - \rho(\mathcal{N}_c, \mathcal{N}_s)}
\]

(4)

The SIQE measures the distortion in the synthesized image i.e., smaller the value, better the quality of the image.

### 3. EXPERIMENTAL EVALUATION AND RESULTS

The performance of the proposed SIQE metric is evaluated on synthesized videos generated from compressed MVD video sequences, by comparing its grading with existing widely used full reference image and video quality metrics. All the results and the software to reproduce them are available at http://www.di.unito.it/~farid/3DQA/SIQE.html. We have used 4 standard video sequences listed in Tab. 1. Each video (texture plus depth), is independently encoded at 6 quality levels using the novel High Efficiency Video Coding (HEVC) [10] HM 11.0 reference software (main profile) with \( QP = \{26,30,34,38,42,46\} \). Using DIBR algorithm [31], 36 different intermediate view sequences are generated using a pair of videos plus depths coded with different QPs. This means a total of 144 synthesized videos are used in evaluation. In all experiments the parameters \( \epsilon = 1, m = 7, n = 7 \) and \( \kappa = 300 \) are used.

The performance of the proposed quality metric is evaluated by comparing it with state of the art full reference video quality metrics as well as with popular 2D image quality metrics. According to Video Quality Expert Group (VQEG) FRTV Phase-I and Phase-II tests the NTIA/ITS video quality model (VQM) [32] performed the best with the highest correlation (0.91) with subjective testing. Due to its best

<table>
<thead>
<tr>
<th>S#</th>
<th>Sequence</th>
<th>#F</th>
<th>V_l</th>
<th>V_r</th>
<th>V_s</th>
<th>Size</th>
<th>FR</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>Poznan_Hall2</td>
<td>200</td>
<td>7</td>
<td>6</td>
<td>5</td>
<td>1920 x 1088</td>
<td>25</td>
</tr>
<tr>
<td>S2</td>
<td>Poznan_Street</td>
<td>250</td>
<td>5</td>
<td>3</td>
<td>4</td>
<td>1920 x 1088</td>
<td>25</td>
</tr>
<tr>
<td>S3</td>
<td>Book_Arrival</td>
<td>100</td>
<td>10</td>
<td>8</td>
<td>9</td>
<td>1024 x 768</td>
<td>16</td>
</tr>
<tr>
<td>S4</td>
<td>Balloons</td>
<td>300</td>
<td>1</td>
<td>5</td>
<td>3</td>
<td>1024 x 768</td>
<td>30</td>
</tr>
</tbody>
</table>
performance we selected VQM for performance evaluation of proposed metric. In experiments NTIA General Model of VQM is used with spatial scaling and temporal registration (1 second) features. Moreover, we compare SIQE with 6 widely used full reference 2D image quality metrics - SSIM [33], FSIM [34], MSSIM [35], iwSSIM [36], Peak Signal to Noise Ratio (PSNR), and UQI [37]. For performance evaluation we use Pearson linear correlation coefficient (PLCC) for prediction accuracy test and root mean square error (RMSE) for prediction error. Before computing these parameters, according to VQEG recommendations [38] the scores are mapped with monotonic nonlinear regression. The following logistic function outlined in [39] is used for regression mapping:

\[ Q_p = \beta_1 \left( \frac{1}{2} - \frac{1}{\exp \beta_2(Q - \beta_3)} \right) + \beta_4Q + \beta_5 \]

where \( Q_p \) is the mapped score and \( \beta_1, \ldots, \beta_5 \) are the regression model parameters.

Tab. 2 lists the Pearson linear correlation coefficient values achieved by SIQE. The proposed metric achieves high correlation with all the reference image and video quality metrics. It has average PLCC of 0.9024 with VQM. It exhibits PLCC larger than 0.87 with all quality metrics except PSNR (that in turn is known to correlate poorly with the actual image quality). The prediction error in terms of root mean square error (RMSE) reported in Tab. 3 also reflects the accurate and reliable of performance of the proposed quality metric.

### 4. CONCLUSIONS

In this paper a quality assessment algorithm is presented to estimate the quality of DIBR synthesized images in the absence of the corresponding references. The algorithm uses the original uncompressed input views to estimate the statistical characteristics of the cyclopean image by using the divisive normalization transform, which are compared to those of the synthesized image to estimate the compression and DIBR warping artifacts in the novel view. The metric is tested in on standard MVD sequences and compared to 7 widely used full reference image video quality metrics to evaluate its performance. The evaluation results show the effectiveness of the proposed quality metric.

### 5. REFERENCES


