Using region-of-interest for quality evaluation of DIBR-based view synthesis methods

Andrei Purica*, Giuseppe Valenzise*, Beatrice Pesquet-Popescu* and Frederic Dufaux*
*LTGI, CNRS, Telecom ParisTech, Universite Paris-Saclay, 75013, Paris, France
†University Politehnica of Bucharest, 061071, Romania
Email: {purica, valenzise, pesquet, dufaux}@telecom-paristech.fr

Abstract—As 3D media became more and more popular over the last years, new technologies are needed in the transmission, compression and creation of 3D content. One of the most commonly used techniques for aiding with the compression and creation of 3D content is known as view synthesis. The most effective class of view synthesis algorithms are using Depth-Image-Based-Rendering techniques, which use explicit scene geometry to render new views. However, these methods may produce geometrical distortions and localized artifacts which are difficult to evaluate as they are inherently different from encoding errors and they are perceived differently by human subjects. In this paper, we propose a region-of-interest evaluation technique for view synthesis based on DIBR methods. Based on the assumption that certain areas determined by the geometrical properties of the scene are prone to distortions, we select a ROI by analyzing the multiple DIBR methods together with the ground truth. The approach is tested using a subjective evaluation view synthesis database and show that our method improves the SSIM correlation with subjective scores. We also test another similar method and traditional metrics.

Keywords—View Synthesis; Visual quality assessment; Multi-View Video; SSIM; Depth-Image-Based-Rendering

I. INTRODUCTION

In the past years, 3D content has become more and more popular. As display technology advances, 3D content is now available to the large public and can be enjoyed on most modern television sets. This created a need for efficient compression and transmission systems for 3D information and new algorithms for the creation of 3D scenes or 2D to 3D conversion. Some of the most common applications that involve 3D information are free view point television (FTV) [1], immersive teleconference systems, medical applications and gaming [2].

Several new formats for 3D information representation exist. Some of the most popular include stereo video, MultiView Video (MVV) and Multiview-Video-plus-Depth (MVD) [3]. While stereo video allows a user to experience the sensation of depth, MVV and MVD formats provide additional options such as changing the point of view on the scene or varying the perception of depth. The latter two formats both consist of a number of video sequences acquired in parallel at different points of view of the same scene. In the case of MVD, these texture sequences are also accompanied by depth information in the form of depth maps. Because of this, MVD format supports the creation of virtual view points of the scene by means of Depth-Image-Based-Rendering (DIBR) techniques [4].

The process of generating a video sequence or an image as if acquired from a new point of view from existing sequences or images is known as view synthesis. Several methods exist in the literature and can be mainly divided into three categories based on the use of geometrical information: i) Methods that do not require geometrical information and use interpolation and filtering to synthesize new views. Some of the most popular ones include light field rendering [5], concentric mosaics [6] or lumigraph [7]. ii) Methods that use implicit geometry such as pixel correspondences computed with optical flow or any other motion estimation technique [8], [9], [10] and finally, methods that use explicit scene geometry in the form of depth maps, to warp pixels from one view into a virtual one [11], [12], [13]. The latter category received great interest as it provides a fast and efficient way of generating multiple views.

In the past years, the Moving Picture Experts Group (MPEG) began developing a 3D extension of the High Efficiency Video Coding (HEVC) standard [14], to meet the need for an MVD coding standard. An experimental framework for 3D video was developed [15] and a 3D-HEVC test model (3D-HTM) [16] was build. The model also incorporates a View Synthesis Reference Software (VSRS-1DFast), which uses DIBR techniques to render new views from the texture information and associated depth maps.

However, the quality of the virtual views is greatly affected by multiple factors. A first issue is related to areas in the virtual view which are not visible in the reference views, as no information is available and they manifest as holes in the synthesized image. These areas are also known as disocclusions. They can be divided in two types based on their location: border or non-border disocclusions [17]. The first category are produced by the displacement of the field of view and are located on the sides of the images. The second category appears around foreground object edges. In order to avoid non-border disocclusions it is usually preferred to merge two synthesized views from a left and a right reference view. However, parts of the non-border disocclusions may coincide in the merged views. Traditionally this problem is resolved using inpainting algorithms such as [18], [19], [20]. Other methods propose a preprocessing of the depth maps in order to reduce the size of disocclusions [21], [22]. When working on video sequences, temporal correlations can also be exploited to retrieve information on disoccluded areas. A background extraction can be performed [23].

Other types of artifacts specific to DIBR methods are caused by the depth maps quality. In addition to coding artifacts, depth maps are generally not perfect and may contain noise. This can lead to different types of artifacts in the
synthesized image. A common problem is the texture-depth alignment which may lead to pixels belonging to a foreground object to be warped as if they are part of the background or vice-versa. Another issue is related to the precision of the depth which may cause small "cracks" in the synthesized image due to incorrect displacement of pixels. Finally, depths maps are also subjected to a quantization process as real depth values are not actually stored, they are usually quantized to 256 levels. In general, this problems appear in areas where depth maps are not uniform (i.e. foreground/background separation).

Because the artifacts produced by synthesis are inherently different from those of encoding, evaluating the quality of synthesis in systems using DIBR rendering is not a trivial matter. Especially, considering the final goal of such systems is to provide a 3D experience, The Video Quality Expert Group (VQEG) created the 3DTV Work Group, which is now part of the Immersive Media Group [23], to conduct experiments on the quality of 3D media. Numerous studies were made to address the problem of synthesized video evaluation. Tikamaki et al. [24] studied the assessment of 3D encoded video, the authors also considered the synthesized view quality. Bosc et al. [25] studied the quality of DIBR synthesis and proposed two approaches based on a region of interest (ROI) evaluation. A first method analyzes the contours shifts in the synthesized view and a second one focuses on evaluating the mean SSIM score over disoccluded areas. Purica et al. [26] study the difference between encoding and synthesis artifacts and propose a ROI based SSIM by separating between encoding errors coming from the reference view and distortion caused by the DIBR warping process. In this paper we extend the ideas presented in [25] and propose a new ROI generation technique for SSIM evaluation of synthesized videos. Next, we perform a study of the results using a subjective evaluation database in order to validate our assumptions.

The rest of this paper is organized as follows. Section II motivates and shows the proposed evaluation technique. Section III-A describes the subjective evaluation database used for validating this technique. In Sections III-B and III-C we describe our testing methodology and report our experimental results. Finally, Section IV concludes the paper.

II. TOWARDS A REGION OF INTEREST EVALUATION

As discussed in Sec. II, synthesized videos can have multiple types of artifacts which affect the quality of the image in different ways. DIBR synthesis methods compute pixel disparities from depth map sequences, and then warp the images from the reference view into a new view. Depth maps are usually stored as video sequences and the values are inversely quantized to 256 levels with respect to real scene depth. In the case of aligned camera systems the disparity is easily obtained using the following equation:

\[
d(k) = f \cdot B \left[ \frac{Z(k)}{255} \left( \frac{1}{Z_{\text{min}}} - \frac{1}{Z_{\text{max}}} \right) + \frac{1}{Z_{\text{max}}} \right]
\] (1)

where \( f \) is the focal length of the camera, \( B \) is the distance between view points, \( Z_{\text{min}}, Z_{\text{max}} \) are the minimum and maximum depth of the scene and \( k = (x, y) \) is a position in the image. Because depth maps are subjected to distortions from the acquiring device or transmission systems, the synthesized image can be subjected to geometrical distortion of foreground objects and also poor reproduction of complex textures. As noted in other studies [26] [25] [27], traditional metrics such as PSNR or SSIM may not be the best way to assess the quality of synthesized images. This behavior can be explained by the strong correlation between scene geometry and position of highly distorted areas. In Figure I we depict a gray scale representation of the absolute errors of frame 93 of Newspaper sequence synthesized using [4]. Black indicates an absolute error higher than 50 while white represents an absolute error of 0. It is easily noticeable that the absolute errors are not uniformly distributed throughout the image and are concentrated in certain critical areas. In this example view 6 was synthesized from view 4. We can see a large concentration of high errors on the left side of the image. This is consistent to a border disocclusion which was filled with an inpainting algorithm. Furthermore, highest errors are concentrated around foreground objects and there exists a high correlation between scene geometry and high distortions. Areas that have the same depth and uniform textures are usually represented without distortions, while foreground object edges and more complex textures have a high distortion. Also, we can notice that not all contours are equally distorted. In this example right most edges of objects tend to have a higher distortion. This behavior can be attributed to the direction of the synthesis from view 4 to view 6, which results in holes on one side of the foreground objects. This type of spatial error distribution is usually similar in most DIBR methods. Because of this, using a ROI when evaluating the quality of synthesis methods may provide a better indication of a method's performance.

Given the goal of evaluating multiple synthesis methods the ROI can be selected as discussed in Section II by thresholding the absolute error or analyzing contours. Another possibility which may provide good results is to look at areas that are rendered differently by the methods which we want to evaluate. This is a reasonable assumption as background areas with non complex texture are usually identical in most synthesis methods and do not affect the quality of the image. Also areas that are rendered identically by multiple methods do not provide any differentiation between the tested DIBR algorithms.

Consider a number of distorted images \( I_1^d, I_2^d, \ldots, I_n^d \). Each

![Absolute error color map for frame 93 of Newspaper sequence. View 6 synthesized from view 4 using [4].](image-url)
image is a synthesis of the same view using the same reference and one of $n$ methods. We define $\mathcal{P}$ as:

$$\mathcal{P}(x, y) = \text{std}([I_1^r(x, y), I_2^r(x, y), \ldots, I_n^r(x, y)])$$

(2)

where $(x, y)$ denotes a position in the image and std is the standard deviation.

The binary mask of the ROI can be expressed as:

$$B(x, y) = \begin{cases} 1 & \text{if } \mathcal{P}(x, y) > \tau \cdot \text{mean}(\mathcal{P}) \\ 0 & \text{otherwise} \end{cases}$$

(3)

where $\tau$ is a coefficient used to balance the selection and mean is the average value of $\mathcal{P}$.

As the ground truth is also available when computing the ROI, it is possible to include it in the computation. Including the ground truth in the computation does not provide information towards differentiating the methods. However, it may lead to a more balanced selection of critical areas by taking into account not only regions which differ in the tested methods but also regions that have a relatively high distortion in all methods. This way, the score will also reflect the global quality of a synthesized image instead of only with respect to the tested methods.

$$\mathcal{P}(x, y) = \text{std}([I_1^r(x, y), \ldots, I_n^r(x, y), I_1^m(x, y), \ldots, I_m^m(x, y)])$$

(4)

where $I^r$ is the reference used to compute the metric and $m$ is the number of times we add the ground truth. Due to a variable number of methods that can be evaluated in parallel, the ground truth needs to be weighted. In our experiments we used a weight of $1/6$ (i.e. the ground truth was added once). However, in this case, the mask will have a lot of noise in the form of localized pixels selected for evaluation. Because the artifacts depend on the structure of the scene it is best to remove single pixels and also consider the neighborhood of the critical areas. This can be achieved by performing an erosion and dilation operation on the binary mask. In order to extend the initial ROI, the dilation operation should use a larger morphological structuring element. In our tests we used a $2 \times 2$ square element for the erosion and a $7 \times 7$ square element for the dilation. These values were selected empirically.

### III. EXPERIMENTAL RESULTS

In this section we report our findings using the ROI evaluation technique described in Section III and use a subjective evaluation test database to validate the results. The first part of this section will describe the subjective evaluation database. The second part describes the methodology and finally, the results are discussed in the last part.

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Resolution</th>
<th>Frames per second</th>
<th>Number of frames</th>
<th>Views</th>
</tr>
</thead>
<tbody>
<tr>
<td>Book arrival</td>
<td>1024 × 768</td>
<td>15</td>
<td>100</td>
<td>8 9 10</td>
</tr>
<tr>
<td>Lovebird</td>
<td>1024 × 768</td>
<td>30</td>
<td>150</td>
<td>6 7 8</td>
</tr>
<tr>
<td>Newspaper</td>
<td>1024 × 768</td>
<td>30</td>
<td>200</td>
<td>4 5 6</td>
</tr>
</tbody>
</table>

A. Subjective evaluation database used in our experiments

In order to validate this technique we use a view synthesis subjective evaluation database available at [28]. The tests were performed using Absolute Categorical Rating with Hidden Reference Removal (ACR-HRR) [29] with 32 subjects. Three multiview video sequences were used: Book arrival, Lovebird, Newspaper. Sequence details are reported in Table I. For each sequence there are three views used in the experiments: a left, center and right view indicated in Table I. Four synthesized views are generated for each sequence: left→right, right→left, left→center, right→center. Each synthesis is then performed using the seven methods described below:

A1: based on [30]. Depth map preprocessed by a low pass filter, borders are cropped and the image is resized to the original resolution.
A2: based on [30] with inpainting algorithm proposed by Telea [31].
A4: Muller et al. [33]. Depth aided inpainting.
A5: Ndiki-Nya et al. [34]. Hole-filling using a patch-based texture
A6: Koppel et al. [35]. Synthesis is improved in disoccluded areas using depth temporal information.
A7: the disoccluded areas are not filled

Additional details on the database and an extensive study can be found in [36].

B. Testing methodology

In order to validate the results obtained with the proposed technique we want to evaluate all sequences and views, synthesized with each method. However, as the authors of [28] also notice there are some outliers in the methods. Method A1 has the highest scores in the subjective tests while all objective metrics indicate this method is by far the worst. This is due to the method not using any inpainting algorithms to fill the disoccluded areas. The borders are cropped and the image is rescaled. The non-border disocclusions are avoided by performing a low-pass filtering of the depth map. While the final result is an image with no localized impactful artifacts, it cannot be used for 3D viewing, as the geometry of the scene no longer corresponds to the reference. These results also point out to the subjects inclination to notice localized artifacts more easily than a global change in the frame which further motivates the use of ROI evaluation in synthesis methods. Since we analyze view synthesis for its capability of producing 3D content, we will not use this method in our results.

In our tests we use three quality evaluation metrics: Structural SIMilarity index (SSIM) [37], Peak-Signal-to-Noise-Ratio (PSNR) and Multi-scale SSIM (MSSIM) [38]. For each metric we apply the region of interest we described in Section III and the one proposed by Bosc et al. in [25]. For our method we use multiple variants: proposed mask (P) without erosion/dilation (e/d) or ground truth (GT); P with e/d and P with both e/d and GT. To measure the performance of each metric we compute the average scores across frames for each sequence/view/method ($3 \times 4 \times 6$). In [25] the authors selected four critical points (subjective vs objective results) to evaluate the method. Our tests will be performed on all
Fig. 2. Book arrival sequence view 10 synthesized from view 8 with method A3. Luminance frames and binary masks for the proposed methods and [25]. Black pixels are selected for evaluation.

TABLE II. PCC, SROCC AND RMSE FOR NON-ROI [25] AND OUR PROPOSED METHODS USING SSIM, PSNR AND MSSIM

<table>
<thead>
<tr>
<th>Metric</th>
<th>Non-ROI</th>
<th>[2]</th>
<th>P</th>
<th>P+e/d</th>
<th>P+GT+e/d</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PCC</td>
<td>SROCC</td>
<td>RMSE</td>
<td>PCC</td>
<td>SROCC</td>
</tr>
<tr>
<td>SSIM</td>
<td>60.85</td>
<td>49.94</td>
<td>47.16</td>
<td>61.29</td>
<td>58.64</td>
</tr>
<tr>
<td>PSNR</td>
<td>85.97</td>
<td>77.57</td>
<td>30.36</td>
<td>68.52</td>
<td>32.55</td>
</tr>
<tr>
<td>MSSIM</td>
<td>80.10</td>
<td>65.89</td>
<td>35.58</td>
<td>68.67</td>
<td>38.35</td>
</tr>
</tbody>
</table>

points using the Difference Mean Opinion Score (DMOS). The performance indicators we use are Pearson Correlation Coefficient (PCC), Spearman’s Rank Order Correlation Coefficient (SROCC) and the Root-Mean-Squared-Error (RMSE). Before computing the PCC we will perform a fitting of the results using the recommended nonlinear function from VQEG Phase I final report [39]:

\[ Y = \beta_2 + \frac{\beta_1 - \beta_2}{1 + \exp \left(-\frac{X - \beta_3}{\beta_4}\right)} \]  \hspace{1cm} (5)

where \( \beta_1, \beta_2, \beta_3, \beta_4 \) are parameters, \( Y \) are the predicted values and \( X \) are the objective results.

C. Results and discussion

In Figure 2 we show an example of generated masks for frame 10 of Book arrival sequence, view 10 synthesized from 8. Figures 2(a) and 2(b) show the reference and the synthesized frame with method A3. The filled dissoccluded areas are easy to notice on the left side of foreground objects.
and also on the left border of the image. An additional source of errors which is harder to notice is also a slight displacement of certain textures on foreground object compared to the reference (e.g. the desk). Another source of errors is caused by a slight difference in luminance. This is common with DIBR synthesis methods. While they are able to warp objects to their new position in the virtual view, changes in luminance between views are not accounted. While these types of distortions are not visually impactful, as they are difficult to notice, they can have an impact on the results of objective metrics and are relevant to this study.

Figures 2(c), 2(d), 2(e) and 2(f) show the binary masks for $\text{SSIM}^{\text{-Bosc}}$, $\text{P}^\text{+e/d}$ and $\text{P}^{\text{+GT+e/d}}$, respectively. When comparing 2(c) and 2(d) we can see that our mask is less noisy and better adjusted to the scene geometry. Also, the right side of the image, which corresponds to a border disocclusion is completely selected, as opposed to 25. Furthermore, the texture details of the desk are not selected in our mask, because this area has a uniform depth and is rendered similarly with all DIBR methods. Although there is a slight displacement which will result in high errors, they are hard to notice and are not critical in differentiating the evaluated methods. Performing the e/d operation will reduce the isolated patches/pixels selected in the map while, increasing solid areas. Finally, adding the ground truth in the mask computation will lead to an increased selection. We can notice that additional textures are selected: the desk, the white board and the area surrounding the clock. In this example, the percentages of selected pixels are: 7.5%, 11.44%, 17.21%, 33.2% for Bosc $\text{[25]}$, $\text{P}^{\text{+e/d}}$ and $\text{P}^{\text{+GT+e/d}}$, respectively. This behavior is similar on other sequences/views/methods, however, for brevity reasons we only discuss this example.

In Figure 3 we show the scatter plots for SSIM and ROI SSIM with the binary masks $\text{[25]}$, $\text{P}^{\text{+e/d}}$ and $\text{P}^{\text{+GT+e/d}}$, respectively. Each point represents the DMOS against the average of the objective score over all frames of a sequence/view/method. An improvement can be observed when using our proposed approach. This is also reflected in the numerical results reported in Table III. Our methods outperforms $\text{[25]}$ on all test cases. When compared to the Non-ROI scores, we are able to outperform SSIM with all proposed ROIs, while $\text{P}^{\text{+GT+e/d}}$ show similar performance to PSNR and MSSIM. A loss is observed with PSNR-P and MSSIM-P. This behavior can be explained by the use of e/d and GT. As discussed above the masks will have a larger number of selected pixels. Also, SSIM is already computed using a pixel’s neighborhood, thus performing the e/d operation will allow PSNR-P+GT+e/d to account for the original’s ROI neighborhood. However, the SSIM score will decrease in this case as pixels which are further away from the ROI are evaluated.

Another interesting aspect is the actual implementation for a ROI evaluation with different metrics. For MSSIM the tests were performed by rescaling the ROI. However, it is also possible to recompute the ROI using the rescaled images. Furthermore, additional metrics can be computed with respect to a ROI, though, in the case of perceptual based metrics the way to perform such an evaluation becomes more difficult.

**IV. CONCLUSIONS**

In this paper we presented a study on the use of ROI in the evaluation of DIBR based synthesis methods. We proposed a ROI generation method that can be used with traditional metrics, such as SSIM, PSNR and MSSIM. We validated...
this technique using a publicly available subjective evaluation database, for view synthesis methods, and showed that we can improve the objective results of SSIM, while maintaining similar results for PSNR and MSSIM when compared to subjective scores. Future directions may include finding a better threshold for the ROI selection by taking into account perceptual aspects or finding ways to use a ROI for perceptual metrics. Another study direction is to perform extensive subjective tests for view synthesis using more methods and also encoded reference views and depth maps.

REFERENCES


