ABSTRACT
In this paper, we present a new method for objectively evaluating the quality of stereoscopic 3D videos generated by depth-image-based rendering (DIBR). First, we show how to derive an ideal depth estimate at each pixel value that would constitute a distortion-free rendered video. The ideal depth estimate will then be used to derive three distortion measures to objectify the visual discomfort in the stereoscopic videos. The three measures are temporal outliers (TO), temporal inconsistencies (TI), and spatial outliers (SO). The combination of the three measures will constitute a vision-based quality measure for 3D DIBR-based videos, 3VQM. Finally, 3VQM will be presented and verified against a fully conducted subjective evaluation. The results show that our proposed measure is significantly accurate, coherent and consistent with the subjective scores.

Index Terms— Stereoscopic, Video Quality, Quality Assessment, Depth-Based Rendering , 3DTV

1. INTRODUCTION
The last decade has witnessed a surge in 3D video popularity. Increasingly, a larger number of movies are being recorded in a stereoscopic format. This trend is accompanied by advances in capturing and rendering complex scenes as well as advances in the auto/stereoscopic display technologies. One such rendering technique is depth-image-based rendering (DIBR) [1]. Using DIBR, the left or the right view can be estimated from a single 2D color view and its corresponding gray level depth map (Figure 1). Two main advantages of DIBR-based 3D video are bandwidth efficiency and viewpoint selection interactivity. However, the perceived quality of DIBR-based 3D videos is affected by several factors including (Figure 2):

- accuracy of the estimated depth maps,
- quality of the 3D wrapping process in DIBR,
- quality of the hole-filling algorithm applied to cover the disoccluded areas in the generated frames,
- artifacts caused by compressing the 2D video plus depth map,
- transmission errors and losses, and
- scaling and formatting algorithms in 3D displays.

Fig. 1. Depth image-based rendering (DIBR). Using DIBR, the left or the right view can be estimated from a single 2D color view and its corresponding gray level depth map by 3D wrapping.

Most of the work in the literature concerning objective 3D video quality measures is based on applying 2D video quality measures on the left and the right views and then finding the combination of these measures that best correlates with the subjective scores [2][3][4]. In such methods, a major assumption is made by considering the perceived depth distortions as less significant than color distortions. In addition to its poor correlation with perceived 3D video quality, this approach has been proved to be non-robust [5]. Other works in the literature include a no-reference measure based on evaluating the blockiness and disparity temporally, and then finding the best combination of parameters using particle swam...
optimization [6]. No depth information is considered in the aforementioned measure and it suffers from the same robustness and poor correlation problem of 2D video quality based techniques. While all the aforementioned techniques ignored the depth information, authors in [7] used a combination of a depth map error based comparison function and 2D quality measure for colored images to predict the 3D image quality. The addition of the depth information to the combination did not result in a significant improvement to the prediction of 3D video quality, because the authors did not take into consideration the visual discomfort in analyzing depth information.

In DIBR, errors in depth map are possibly caused by inaccurate estimation, numerical rounding, or compression artifacts. Such inaccuracies lead to errors in the relative pixel location and in the magnitude of the pixels, which are a function of position and focal length. The visual effect of these errors on the synthesized view is spatially noticeable around texture areas in the form of significant intensity changes and temporally noticeable around flat regions in the form of flickering [8]. Visual discomfort in DIBR-based 3D video can also result from several factors including excessive disparities, fast changing disparities, geometric distortions, and inconsistencies between various depth cues such as unmatched object colors [7]. Hence, in order to capture the visual quality of 3D video generated through DIBR all these factors need to be considered.

In our work, we first estimate an ideal depth map that would generate a distortion free DIBR-based 3D video. Then, we derive three distortion measures by quantitatively comparing the ideal depth and the given depth maps in terms of visual discomfort. We package all these distortion measures into a new visual quality measure for DIBR-based 3D videos: 3VQM: 3D Video Quality Measure.

The three distortions measures that we will introduce evaluate the temporal and spatial variation of the depth errors for the depth values that would lead to inconsistencies between the left and right view, fast changing disparities, and geometric distortions. These measures are the spatial outliers (SO), temporal outliers (TO), and temporal inconsistencies (TI).

The rest of the paper is organized as follows. DIBR will be reviewed in Section II. Then the concept of ideal depth estimation will be presented in Section III. In Section IV we will present the three distortion measures to measure visual discomfort in stereoscopic videos. The combination of three measures will constitute a vision-based quality measure for 3D DIBR-based videos referred to as 3VQM which we will introduce in Section V. Finally, the experimental results and conclusion will be presented in Sections V and VI, respectively.

2. DEPTH IMAGE BASED RENDERING (DIBR)

In DIBR, virtual views are generated by first projecting the pixels in the reference image back to the world coordinates using depth map and camera information. The resulting pixels in the world coordinates are then projected back to the estimated virtual image coordinate. This process is known as 3D wrapping [9].

Consider a reference camera \( C_r \) and a virtual camera \( C_v \) as shown in Figure 1. Where \( F_r \) and \( F_v \) are the focal lengths of the reference and the virtual cameras, respectively. \( B \) is the baseline distance that separates the two cameras. \( Z_c \) is the convergence distance of the two cameras. The horizontal coordinates vector \( \bar{X}_v \) of the virtual camera as a function of the horizontal coordinate vector \( \bar{X}_r \) of the reference camera is given by

\[
\bar{X}_v = \bar{X}_r + s \frac{F_v B}{Z} + h,
\]

where \( s = -1 \) when the estimated view is to the left and \( s = +1 \) when the estimated view is to the right, \( Z \) is a vector of the depth values at pixel location \( (x_r, y_r) \), and \( h \) is the horizontal shift in the camera axis.

Errors in the depth map lead to errors in the relative pixel location defined by the wrapping equation. Depending on the
location of these displacements the distortion will be visible spatially and/or temporally. Additional distortions may be introduced by the hole filling algorithm especially around the edges which represent areas of change in depth. Hole filling is required because the 3D wrapping will result in disocclusion. Disocclusion occurs either when a uniform sampling in the reference image becomes non-uniform in the desired image or mostly when occluded areas in the reference image becomes visible in the virtual image. Several approaches have been proposed for hole filling such as depth map smoothing plus interpolation, extrapolation or background mirroring [1], asymmetric smoothing [10], distance dependent smoothing [11], edge-based smoothing [12], layered depth images (LDI) [13], and hierarchical hole filling (HHF) [14]. Hence the nature of distortion around the edges will depend on the algorithm chosen for hole filling.

### 3. IDEAL DEPTH ESTIMATE

![Image](image.png)

**Fig. 3.** *Ideal depth* is the depth map that would generate the distortion-free image given the same reference image and DIBR parameters $(B, s, F_v, h)$.

As mentioned earlier wrong estimations, numerical rounding and compression artifacts lead to errors in the depth map. Inaccurate relocation of pixels lead to visual discomfort in the stereoscopic video. Henceforth, in order to evaluate the quality of the generated videos for visual discomfort we need a reference depth map. A depth map that would constitute an ultimate reference would be the ground truth depth. However, it is impossible to obtain the ground truth depth in real time. Stereo matching is often used to obtain an estimate of the ground truth depth. But the depth generated by stereo matching suffers from high percentage of bad pixels and, thus, it cannot be used as a reference. Furthermore, DIBR-based stereoscopic videos contain distortions that are introduced by stereo matching. Such distortions need to be accounted for in any proposed quality measure.

As a result, the ground truth depth cannot serve as the proper reference for a quality measure. In order to overcome such difficulty, we propose to find the *ideal depth*. As depicted in Figure 3 we define the ideal depth as the depth map that would generate a distortion-free virtual view assuming that same reference image and same DIBR parameters. The ideal depth estimate can be derived as follows:

1. The horizontal coordinate vector $\mathbf{X}_v$ of the synthesized virtual view can be expressed as a function of the horizontal coordinate vector of the the reference view $\mathbf{X}_r$ using the 3D wrapping equation in (1) as follows:

   $$\mathbf{X}_v = \mathbf{X}_r + s \frac{F_v B}{Z} + h.$$  

   (2)

2. Similarly, the horizontal coordinate vector $\mathbf{X}_o$ of an obtained distortion free view can be expressed as a function of the horizontal coordinate vector $\mathbf{X}_r$ of the the reference view as

   $$\mathbf{X}_o = \mathbf{X}_r + s \frac{F_v B}{Z_{\text{IDEAL}}} + h,$$  

   (3)

   where $Z_{\text{IDEAL}}$ is the ideal depth map vector to be estimated and in practice an obtained distortion free view is assumed to be the image captured by a stereo camera.

3. If we subtract (3) from (2) and then preform direct substitution, the ideal depth vector $Z_{\text{IDEAL}}$ can be derived to be as follows:

   $$Z_{\text{IDEAL}} = \frac{s F_v B}{(\mathbf{X}_o - \mathbf{X}_v)} + \frac{s F_v B}{Z}.$$  

   (4)

4. Calculating $\mathbf{X}_o - \mathbf{X}_v$ is non trivial, hence in order to estimate $Z_{\text{IDEAL}}$ the equation in (4) needs to be expressed in terms of intensity variation rather than horizontal shift. In [8] it has been shown that the sum of squared differences (SSD) of the original video frame and its horizontal translations is linear. Hence, a small horizontal shift $\Delta X$ can be estimated in terms of intensity $\Delta I \approx \alpha \Delta X$, where $\alpha$ is a constant. As a result, the ideal depth vector can be estimated from the generated virtual view vector $\mathbf{I}_v$ obtained through DIBR, the original image captured by the camera $\mathbf{I}_o$, the given depth map $Z$ vector, focal length $F_v$, and the baseline $B$ as follows:

   $$Z_{\text{IDEAL}} \approx \frac{s F_v B}{\alpha (\mathbf{I}_o - \mathbf{I}_v)} + \frac{s F_v B}{Z}.$$  

   (5)

Using an ideal depth estimate, we can derive the distortion measures that would quantify different elements of visual discomfort in the DIBR generated 3D videos. These distortion measures will be derived by comparing the estimated ideal depth map against the given depth map. Such comparison will not only capture the visual distortions due to the errors caused by bad pixels in the depth maps from stereo matching and/or
compression but also the errors caused by post processing of the synthesized colored video itself such as hole-filling and colored video compression. The latter is due to the fact that the ideal depth estimate is as a function of the given depth and the difference between the generated virtual image after processing \( I_o \) and its corresponding captured colored image \( I_o \).

4. DISTORTION METRICS

Let us start by defining \( \Delta Z \) as follows:

\[
\Delta Z = |Z_{IDEAL} - Z|, \tag{6}
\]

where \( Z \) is a matrix with \( Z \)'s as the columns. It is important to notice that a non-zero value of \( \Delta Z \) does not always mean a visual distortion. For instance, in the spatial domain a consistent (uniform) error over a specific depth plane will cause the whole plane to be shifted in one direction and the perceptual effect of this shift will be a slight increase or decrease in the perceived depth. This increase or decrease, however, has no effect on the quality of experience of the 3D video. On the other hand, an non-uniform error in depth plane would result in relocation of color pixel/blocks during the wrapping process to an alien position. The visual effect of these errors on the synthesized view is spatially noticeable around texture areas in the form of inconsistent depth cues (unmatched object colors). The temporal variation of \( \Delta Z \) is another indicator of perceptually visual distortion. A temporally inconsistent \( \Delta Z \) indicates random pixel relocation during the wrapping process or inconsistency in the hole filling algorithm which is spatially noticeable around textured areas in the form of significant intensity changes and around flat regions in the form of flickering.

Based on \( \Delta Z \), we define two measures: Temporal error outliers (TO) and Spatial error outliers (SO). Furthermore, we introduce a third measure as Temporal inconsistency (TI). Next we will explain how we use these measures to quantify elements of visual discomfort in the generated 3D videos.

4.1. Spatial Outliers (SO)

The spatial inconsistencies are quantified through the spatial outliers (SO), which are calculated as the standard deviation of \( \Delta Z \).

\[
SO = STD(\Delta Z). \tag{7}
\]

The standard deviation in this case separates the spatially visible distortions due to non-zero \( \Delta Z \) from the perceptually non-significant \( \Delta Z \)'s. The latter are spatially uniform and originate from inaccuracies in the wrapping equation and inherited approximation in the camera modeling parameters. In Figure 4 a frame from a DIBR generated video is shown. The original stereo video was captured by Point Grey’s Bumblebee2 camera and then the depth map sequence was generated using stereo matching. The depth was then used to obtain a DIBR-based estimate of the right view video. Looking at the chosen frame we can see that there are distortions around the hand, the paper, the head, and the wall in background. These distortions are caused by both the errors in the depth maps as well as by the hole filling algorithm. The SO map of the frame in Figure 4 is shown in Figure 5(a). The spatial distortions were all captured by SO plus the edges where a plan shift occurs. The latter is not a source of visual distortion however it can be filtered using the temporal outliers described next.

4.2. Temporal Outliers (TO)

We define temporal outlier (TO) as the standard deviation of the change in \( \Delta Z \) for two consecutive frames as follows:

\[
TO = STD(\Delta Z_{t+1} - \Delta Z_t). \tag{8}
\]

The error introduced due to depth map noise is temporally inconsistent while an non-zero \( \Delta Z \) around at an edge of plane change will be temporally consistent because the same wrapping parameters were used to generate both frames. By taking the standard deviation, the temporal outliers filters out the edginess in SO and will only keep the visible distortions from depth map errors and hole filling. This can be also observed by looking into the TO map of the Figure 4 is shown in Figure 5(b) where the edginess is no longer part of the captured distortion.

4.3. Temporal Inconsistencies (TI)

Fast changing disparities are another source of visual discomfort and are mainly caused by errors in stereo matching, hole filling algorithms and depth compression. These distortions are usually observed in the form of flickering which is usually observed around smoothly textured areas and noise around highly structured regions. We will refer to this measure as temporal inconsistency (TI) and it can be derived as follows:

\[
TI = STD(Z_{t+1} - Z_t). \tag{9}
\]

The TI map of the frame in Figure 4 is shown in Figure 5(c). Looking into the map you can notice that TI captures all the flickering on the wall in the background. This flickering is caused by inconsistencies in the hole filling algorithm. TI also captures the fast changing noises that were not captured by the spatial outliers earlier.

5. 3VQM

The three distortions measures defined earlier captures all element of visual discomfort that could be derived by comparing
the actual given depth from the ideal depth estimate. We combined the three distortion measures into one the 3D Vision-based Quality Measure (3VQM) for DIBR generated videos defined as follows:

$$3VQM = K(1-\text{SO}(\text{SO} \cap \text{TO}))^a(1-\text{TI})^b(1-\text{TO})^c,$$  \hspace{1cm} (10)

where SO, TO, and TI are normalized to range 0 to 1 and \( a, b, \) and \( c \) are constants which were determined by running a few training sequences. \( K \) is a constant for scaling where 3VQM ranges from 0 for lowest quality to \( K \) for highest quality. \( (\text{SO} \cap \text{TO}) \) is the logical intersection of \( \text{SO} \) and \( \text{TO} \) included in the equation to avoid accounting the outlier distortion more than once. The overall quality measure 3VQM is calculated as the mean of values in the matrix 3VQM. The 3VQM map of the frame in Figure 4 is shown in Figure 5(d).

**Fig. 4.** A single frame chosen from a right view video generated through DIBR. The depth maps were obtained using stereo matching.

**6. SUBJECTIVE RESULTS**

In order to evaluate the performance of 3VQM we conducted a subjective evaluation test. In our experiments the Samsung 2233RZ display with the shutter glass solution from NVIDIA were used. We setup the testing conditions following the new requirements for subjective subjective video quality assessment methodologies for 3DTV described in [15]. The tested videos were a total of 21 video sequences each of 30 seconds in length. Four of these video sequences were actually captured by a stereo camera without any synthesize. The rest were all DIBR synthesized videos for three different applications. These applications included depth map and colored video compression, depth estimation (stereo matching), and depth from 2D to 3D conversion using color information. In these experiments, 20 volunteers were asked to assign each video with a score indicating their assessment of the quality of that video. Most of the participants were engineers with little to no previous experience of 3D video processing. We defined quality as the extent to which the distortions were visible and annoying. The raw scores for each subject was collected and processed to give Mean Opinion Scores (MOS) and a Difference Mean Opinion Score (DMOS) for each distorted videos.

We set \( K = 5 \) in (10) in order to match the MOS range. In MOS the values are discrete and defined as follows: 1-Bad, 2-Poor, 3-Fair, 4-Good and 5-Excellent. This setup will give the values of 3VQM a meaningful representation as well it make it easier to compare against the MOS values. As for the constants \( a, b \) and \( c \) a small training experiment was ran with three video sequences in which three different volunteers were asked to rate the synthesized videos. And based on the training experiment results the constants were set to the following values: \( a = 8, b = 8 \), and \( c = 6 \). In order to make sure that this would not bias our results, the synthesized videos that used in the training experiment were not used in the subjective experiment and the volunteers who evaluated the training sequence were not asked to perform the subjective experiments.

The scatter plot showing 3VQM versus DMOS is shown in Figure 6. By looking at the plot we can see that all of our results are inside the outlier boundary defined by the quality rat-
Fig. 6. Scatter plots for 3VQM. A point is considered an outlier if the distance from the ideal is greater than twice the DMOS standard deviation. In our results the standard deviation of the DMOS values was: $\sigma_{DMOS} = 0.7885$.

ings that are greater that two DMOS standard deviation away from the ideal rating. Also we notice that almost more that 80% of the 3VQM values fall inside the one $\sigma_{DMOS}$ boundary. This indicates that our measure is significantly consistent hence has no outliers. The validation scores for 3VQM are show in Table 1.

<table>
<thead>
<tr>
<th>RMSE</th>
<th>CC</th>
<th>ROCC</th>
<th>MAE</th>
<th>OR</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.6158</td>
<td>0.8942</td>
<td>0.7890</td>
<td>0.5173</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 1. Validation scores for 3VQM. The validation criteria are: root mean squared error (RMSE), Pearson linear correlation coefficient (CC), Spearman rank order correlation coefficient (ROCC), mean absolute error (MAE), and Outlier Ratio (OR).

Looking at the results we can notice that RMSE = 0.6158 $< \sigma_{DMOS} = 0.7885$. This is an indication of high accuracy in our results. Also looking at both the Pearson linear correlation coefficient (CC) and Spearman rank order correlation coefficient (ROCC) we conclude that 3VQM depicts both a high mendicity and a high coherency. Outlier ratio is zero indicating that the 3VQM values are consistent.

7. CONCLUSION

In this paper we presented a new video quality measure, 3VQM, for 3D videos generated through DIBR. We have also introduced the concept of an ideal depth map and showed how to obtain an estimate of that depth. The ideal depth estimate was then used to obtain three distortion measures to evaluate elements of visual discomfort on the synthesized 3D video. 3VQM was compared against subjective rating and the results showed that our quality measure is a coherent, consistent and accurate estimate of DIBR-based 3D videos.

8. REFERENCES